UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Interface Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em Big Data

(Versão Final Após a Defesa Pública)

Fátima Joana Dantas Gonçalves Chitongua

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Sebastião Augusto Rodrigues Figueiredo Pais
Co-orientador: Prof. Doutor João Paulo Cordeiro

Covilhã, Abril de 2019

Dedicatória

A minha formação só foi possível pelos grandes sacrifícios feitos pela minha família. Este trabalho é dedicado ao meu Pai (in memoriam) por tudo o que representa na sociedade, pela inteligência e carácter fora do comum e pelo grande homem que foi. À minha mãe (in memoriam) em especial, pois seus ensinamentos guiam minha vida e passados quase vinte anos me apercebo que é tarde para agradecer tudo o que fez por mim e tudo de bom que me deixou. A ti minha mãe a minha profunda gratidão.

Aos meus filhos Erivandro e Catarina, meus anjos de luz, ao meu esposo, irmãos e sobrinhos, por todo apoio incondicional e por toda a compreensão das ausências em várias circunstâncias a que os estudos me fizeram prescindir da partilha e companhia de cada um em particular.

Agradecimentos

À Deus Todo Poderoso, pela vida, à Nossa Senhora de Fátima e a todos os Anjos e Santos, guias e protetores incansáveis, por cuidarem de mim e minha família e propiciarem várias oportunidades de estudos, pela força e perseverança em cada passo da vida e por colocarem em meu caminho pessoas com corações preciosos.

Ao meu esposo Abel por todo apoio incondicional e aos meus filhos Catarina e Erivando pelo amor e compreensão.

À minha irmã Roselinda, que aceitou abrir mão da minha presença e com o coração apertado me largou para o mundo e mesmo distante tem solucionado vários problemas no decurso da minha formação e principalmente durante o desenvolvimento deste trabalho.

À minha avó, Ao mano Dino, à mana Perpétua, à tia Mena por me terem amado, valorizado, protegido e cuidado.

À minha amiga Maria Clara Saraiva, por ter sido a pedra angular para a finalização deste Mestrado, o meu muito obrigado.

Manifesto a minha gratidão ao Prof. Doutor Sebastião Augusto Rodrigues Figueiredo Pais, orientador desta Dissertação de Mestrado, pela sua simpatia e disponibilidade desde o primeiro contacto; agradeço sobretudo pela confiança demonstrada e pelos incentivos, conselhos e ajuda que tornaram possível a concretização da mesma.

Agradeço ao Co-orientador: Prof. Doutor João Paulo Cordeiro, por sua simplicidade, pronta disponibilidade e intervenção em momentos cruciais.

Agradeço à Universidade da Beira Interior e em especial o Departamento de Informática, de modo particular à todos os professores que lecionaram a parte curricular deste mestrado.

Um agradecimento aos colegas do Centro de Tecnologia da Linguagem Humana e Bioinformática, em especial ao Dionísio por sua paciência e disponibilidade.

À minha incansável e querida amiga Joana Raquel, à Doutora Suzana Rodrigues, ao Moser josé, muito obrigada por tudo.

À todos os meus familiares, amigos, colegas e vizinhos que direta ou indiretamente tenham contribuído para a realização desta Tese o meu muito obrigada.

Resumo

Os sistemas de aquisição, armazenamento, processamento, recuperação e divulgação da informação, apresentam uma complexidade indiscutível, existindo por isso, uma grande necessidade académica e científica de criar mecanismos que permitam a pesquisa e o tratamento de dados e informações de forma eficaz. Com o aumento explosivo de dados, o processamento textual torna-se cada vez mais difícil e em alguns casos, onerosos.

Apesar dos avanços quanto a criação de ferramentas para a extração de informações relevantes, há uma clara falta de ferramentas ou Corpora online multilingues para extrair automaticamente tais informações em documentos. Observou-se que o vasto conjunto criado e disponíveis na Web apresenta limitações à nível linguístico, áreas de domínio e às regras de utilização e acesso.

Neste contexto, o estudo realizado no presente trabalho visa desenvolver uma plataforma que disponibilize serviços de PLN em Big Data, sem fins lucrativos.

Para tal foi criado o Hultig-C e desenvolvida a plataforma para a disponibilização dos serviços que o mesmo poderá oferecer, proporcionando o acesso aos mais variados dados de diferentes temáticas e idiomas, o que permite a extração de informações relevantes, descomplexificando a recuperação seletiva da informação e consultas de forma geral. Cujo objetivo é dar suporte ao processamento automático da linguagem humana e providenciar recursos de alto nível para a investigação e desenvolvimento de tecnologias em PLN.

O estudo apresenta como proposta uma nova abordagem não supervisionada e independente da língua para extrair termos relevantes (específicos) em um documento até Trigram e através destes determinar os termos mais gerais de um documento, fazendo uso da abordagem da Implicação Textual por Generalidade.

Vários experimentos foram realizados e com base neles podemos afirmar que o método de extração de termos relevantes proposto na presente Dissertação alcança bons resultados, cujo grau de eficácia revela-se elevado quando comparado com abordagens semelhantes e que fazem uso dos algoritmos mais sofisticados de extração de termos relevantes sem supervisão como o Yake e o Rake.

A abordagem apresentada neste trabalho faz uso dos recursos fornecidos pelo próprio texto, tornando-a independente em relação às técnicas de PLN, acrescido ao facto de ser não supervisionada e independente da língua a torna adequada para outros Corpora dos vários domínios e idiomas ao contrário das abordagens supervisionadas dependentes de um Corpus de treinamento.

Palavras-chave

Corpora, Seleção de características, Extração de termos relevantes

Abstract

The acquisition systems, storage, processing, recovery and popularization of the information, they present an unquestionable complexity, existing for that, a great need academic and scientific of creating mechanisms to allow the research and the treatment of data and information in an effective way. With the explosive increase of data, the textual processing becomes more and more difficult and in some cases, expensive.

Although the advances how much the creation of tools for the extration of relevant information, has a clear lack of tools or Corpora online multilingues to extract such information in documents automatically. It was observed that the vast set bred and available in the Web presents limitations to the linguistic level, areas of domain and to the rules of use and access.

In this context, the study accomplished in the present work seeks to develop a platform that makes available services of PLN in Big Date, without lucrative ends.

For such Hultig-C was created and developed the platform for the disponibilização of the services that the same can offer, providing the access to the most varied data of different themes and languages, what allows the extraction of relevant information, descomplexificando the selective recovery of the information and consultations in a general way. Whose objective is to give support to the automatic processing of the human language and to provide resources of high level for the investigation and development of technologies in PLN.

The study presents as proposal a new approach no supervised and independent of the language to extract relevant (specific) terms even in a document Trigram and through these to determine the most general terms of a document, making use of the approach of the Textual Entailment by Generality.

Several experiments were accomplished and with base in them can affirm that to method of extraction of relevant terms proposed in the present Dissertation reaches good results, whose degree of effectiveness is revealed high when compared with similar approaches and that you/they make use of the most sophisticated algorithms of extraction of relevant terms without supervision as Yake and Rake.

The approach presented in this work does use of the resources supplied by the own text, turning her independent in relation to the techniques of PLN, added to the facto of being not supervised and independent of the language it turns her appropriate for other Corpora of the several domains and languages unlike the approaches supervised dependents of a training Corpus.

Keywords

Corpora, Feature selection, Relevant terms extraction

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AAM *Asymmetric Association Measures*

AIS *Asymmetric InfoSimba Similarity of Term*

BB *Braun-Blanket*

BD Base de Dados

CES Eagles Corpus Encoding Standard

CI Ciências da Informação

Co *Conviction*

COPPA *Corpus Of Parallel Patent Applications*

ECM *Enterprise Content Management*

GED Gerenciamento Eletrónico de Documentos

GI *Gini Index*

HTML *Hypertext Markup Language*

IA Inteligência Artificial

IS Similaridade InfoSimba

KDT *Knowledge Discovery in Text*

LC Linguística de Corpus

LCC *Leipzig Corpora Collection*

LM *Language Model*

LP *Laplace*

PC Probabilidade Condicional

PLN Processamento da Linguagem Natural

SGBD Sistema de Gestão de Base de Dados

UBI Universidade da Beira Interior

QA *Question Answering*

RTE *Recognizing Textual Entailment*

TE *Textual Entailment*

TEI Lite *Text Encoding Initiative Lite*

TF-IDF *Term Frequency-Inverse Document Frequency*

TM *Text Mining*

UFMG Universidade Federal de Mina Gerais

VSM Modelo de Espaço Vetorial

Capítulo 1

Introdução

Desde os primórdios da Humanidade que o Homem viu-se obrigado a comunicar e de alguma forma guardar informações a respeito da apreciação feita em determinada situação do seu quotidiano, a fim de garantir sua subsistência e sobrevivência. Assim, e como resultado da necessidade de fixar os dados produzidos através das experiências vividas, para serem recuperados a curto, médio e longo prazo, não podendo depender somente da capacidade de memória do ser humano, o Homem Primitivo recorria a utilização de objetos e símbolos, que devidamente assinalados, posteriormente permitia a recuperação da informação neles armazenada; facilitando deste modo o controlo da pesca, caça e da criação do gado, que era a fonte de subsistência das populações. Dessa forma, foi desenvolvendo e aperfeiçoando cada vez mais técnicas comunicativas e de armazenamento, catalogação e organização dos dados e informações para uso posterior.

O ser Humano utiliza a linguagem natural como instrumento para comunicar seu conhecimento. Na atualidade, e com os avanços da ciência e da técnica, cresce cada vez mais, e de forma rápida a quantidade de dados e informações de diferentes naturezas, armazenados em formato digital. Surgindo como consequência, uma grande necessidade de processamento da mesma linguagem; pois os volumes de informações arquivadas são imensos e sem um prévio processamento da linguagem natural é difícil aproveitá-la [Med16].

O Processamento da Linguagem Natural (PLN) representa um dos grandes desafios nos dias atuais, pois requer um conjunto de competências variadas para tratar a língua de maneira automática. Nos últimos anos, às contribuições para a ciência no campo de PLN têm conhecido um grande avanço, permitindo o processamento de imensas quantidades de informações em formato textual e não só, com elevado grau de eficácia [SSCN+15].

1.1 Contextualização e Motivação

Os sistemas de aquisição, armazenamento, processamento, recuperação e divulgação da informação, apresentam uma complexidade indiscutível, existindo, por isso, uma grande necessidade académica e científica de criar mecanismos que permitam a pesquisa e o tratamento de dados e informações de forma eficaz, no que diz respeito a recuperação de informações relevantes. Verificando-se um grande aumento na procura de recursos tecnológicos e digitais que possam dar respostas às necessidades dos usuários dos sistemas de recuperação de informação ou Base de Dados (BD), ampliando assim a qualidade das buscas bibliográficas. Sendo que o PLN tem sido amplamente abordado nas diferentes vertentes que os investigadores têm procurado para suavizar a complexidade dos mecanismos de busca e recuperação da informação que atendam à necessidade do usuário final [SSCN+15]. Uma BD é uma coleção de dados inter-relacionados que proporcionam diversificados pontos de acesso a informação e propiciam técnicas de buscas com baixo nível de dificuldade, pois permitem a utilização de vários conceitos na mesma es-

tratégia de busca por termos da linguagem natural. As Informações apresentadas em uma Base de Dados variam de acordo a temática e finalidade da mesma. Existindo às orientadas para um determinado assunto e as institucionais que perseguem dar respostas a missão e visão da instituição que a desenvolve. Sendo que o armazenamento das informações é feito a partir de uma estrutura de BD. E através de um Sistema de Gestão de Base de Dados (SGBD) faz-se a gestão e o processamento dessa informação, tornando-a inteligível aos programas de aplicação [Rod07].

Neste contexto e porque a base do funcionamento das aplicações de PLN são as coleções de escritos para análise linguística, têm vindo a ser desenvolvidos Corpora para dar suporte a base de tais aplicações. Um Corpus, apresentado no plural como Corpora é um conjunto de dados linguísticos reais, criteriosamente coletados e estruturados, utilizado em diversas áreas. Deve ser constituído por dados autênticos e legíveis por máquina [dAG+17].

Para construir um Corpus é necessário ter em conta aspetos como o tipo de escritos, o idioma, tipo de texto, o domínio em que se insere entre outras características. Existindo Corpora anotados manualmente por vários especialistas da área e Corpora anotados recorrendo a ferramentas automáticas. Corpora anotados de forma manual apresentam desvantagens por serem Corpora limitados, pois são pequenos, e fazem cobertura a um número restrito de assuntos. E por outro lado, estes Corpora apresentam implicitamente conhecimentos linguísticos associados a avaliação feita pelos humanos envolvidos no processo [Pai08].

Nas Ciências da Informação (CI) o PLN tem sido amplamente estudado, com especial atenção a Indexação e Recuperação da Informação, pois os softwares desenvolvidos com base no modelo mencionado proporcionam a aquisição de resultados com maior precisão semântica para recuperação da informação em sistemas de busca automatizados [GL03]. Surge deste modo a motivação para a realização do presente trabalho. Onde, pretende-se efetuar uma investigação extensiva da literatura sobre o assunto em questão dando enfoque à aplicação de um novo Modelo de Busca e Recuperação da Informação, que se cinge no desenvolvimento automático de um Corpus que irá retornar respostas às buscas dos utilizadores, independentemente da área geográfica, da diversidade temática e linguística, que irá permitir restringir ou ampliar os resultados das buscas feitas mediante às necessidades dos mesmos, com o objetivo de obter resultados concisos e relevantes; através de uma interface que disponibilizará serviços de PLN em *Big Data*.

1.2 Objetivos

Com o presente projeto e dentro do contexto apresentado, se perspetiva alcançar os seguintes objetivos:

1.2.1 Objetivo Geral

Desenvolvimento de uma Interface Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em *Big Data*, com base no HULTIGCorpus.V1 e sua posterior distribuição sem custos a comunidade científica e académica.

1.2.2 Objetivos Específicos

Para que o objetivo principal fosse alcançado, foram estabelecidos os seguintes objetivos específicos:

- Pesquisa Bibliográfica abrangente a linha de investigação.
- Definição de conceitos referentes a temática abordada.
- Pesquisa sobre interfaces de PLN existentes e trabalhos científicos relacionados.
- Descrição sobre Corpus e plataformas existentes e trabalhos científicos relacionados.
- Descrição sobre ferramentas para compilação, processamento e análise de Corpora existentes e trabalhos científicos relacionados.
- Recolha e indexação dos textos da Web para a criação do Hultig-Corpus.
- A partir das investigações feitas sobre o que já existe disponível no mercado sobre serviços de PLN online, desenvolver uma ferramenta para compilação, processamento e análise de Corpora que permitirá disponibilizar serviços de busca e recuperação da informação, no Corpus, através de uma interface de PLN em *Big Data*.
- Propor uma abordagem metodológica para a extração de termos relevantes em um documento de texto.

1.3 Estrutura da Dissertação

Esta dissertação encontra-se dividida em cinco capítulos. Onde,

Capítulo 1: Introdução– relata as principais temáticas a abordar, fornecendo uma sucinta contextualização do tema e a motivação que esteve na base para a linha de pesquisa desta Dissertação, evidenciando a necessidade existente da implementação de uma Interface Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em *Big Data* e qual a relevância da mesma. São ainda apresentados os objetivos a serem alcançados e a estrutura do documento.

Capítulo 2: Estado da Arte – reflete o levantamento do estado da arte em que se insere a presente Dissertação; apresentando conceitos relacionados com PLN, *Big Data*, *Text Mining*, Corpus, entre outros conceitos, bem como a descrição das ferramentas a utilizar no seu desenvolvimento. É apresentado também um conjunto de plataformas para Serviços de PLN existentes e suas funcionalidades.

Capítulo 3: Metodologia e Trabalho Realizado – apresenta a secção metodológica que descreve a abordagem adotada no presente trabalho, as tecnologias utilizadas e a estratégia para extração dos termos relevantes de um texto, seguida do trabalho realizado em torno da metodologia proposta.

Capítulo 4: Desenvolvimento Experimental e Resultados Obtidos – faz menção ao trabalho desenvolvido no decorrer desta Dissertação, apresentando a implementação prática da abordagem proposta para extração de termos relevantes, com exemplos dos testes e respetivos resultados. É apresentado também neste capítulo a metodologia para a criação do Hultig-Corpus e sua arquitetura, o desenvolvimento da plataforma que permitirá a disponibilização ao usuário dos serviços de PLN criados.

Sendo uma possível solução para a dificuldade existente na comunidade académica e científica para a aquisição/utilização de um Corpus para a execução de serviços requeridos em PLN, sem

custos financeiros. Neste capítulo é evidenciado os problemas encontrados e como se atingiram os objetivos propostos.

capítulo 5: Conclusões e Trabalho Futuro – apresenta a conclusão do trabalho realizado, tendo em vista o alcance das métricas predefinidas e o que se perspetiva desenvolver como trabalho futuro.

Capítulo 2

Estado da Arte

”The isolated man does not develop any intellectual power.

It is necessary for him to be immersed in an environment of other men, whose techniques he absorbs during the first twenty years of his life. He may then perhaps do a little research of his own and make a very few discoveries which are passed on to other men. From this point of view the search for new techniques must be regarded as carried out by the human community as a whole, rather than by individuals.”

Alan Turing

Neste capítulo são apresentados os conceitos fundamentais necessários para o entendimento da presente dissertação e discutidos os trabalhos existentes e plataformas relacionadas à abordagem aqui proposta.

2.1 Processamento da Linguagem Natural (PLN)

Nas sociedades humanas, a comunicação tornou-se cada vez mais necessária, originando dessa forma uma linguagem desenvolvida de forma natural, não envolvendo modelagem e um planeamento prévio. A chamada Linguagem Natural.

Na figura 2.1 é evidenciado um dos exemplos de interação máquina humano, em que o objetivo é testar a capacidade de uma máquina exibir comportamento inteligente equivalente a um ser humano, ou indistinguível deste, o chamado Teste de Turing, onde um interrogador (humano) fará perguntas a duas entidades ocultas; uma delas é um humano e outra é um computador. A comunicação entre o interrogador e as entidades é feita de modo indireto. Um dos humanos é um interrogador que está separado (por uma barreira) do outro humano e do sistema de Inteligência Artificial (IA). Este interrogador entra em uma conversa em linguagem natural (via teclado) com o outro humano e também com a máquina, e caso ele não consiga distinguir se está conversando com a máquina ou com o ser humano é um indicativo de que o sistema é inteligente e passou no Teste de Turing [Zil09].

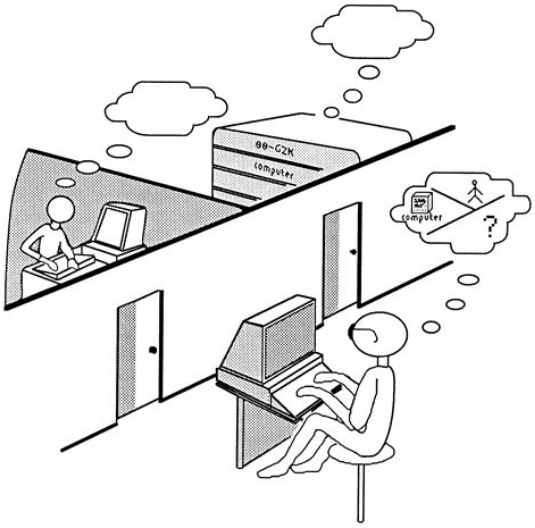

Figura 2.1: Teste de Turing (estraído de iPon Computer Ltd, 2018)

Ao longo do tempo e com a evolução tecnológica que se tem observado, principalmente na disseminação de máquinas capazes de processar dados, e a importância no dia a dia da linguagem, campos de investigação foram surgindo para auxiliar nas atividades humanas no que concerne ao processamento de um número cada vez maior de informações existentes, de forma mais eficiente, e fazendo uso dos variados suportes em que se usam linguagem. Tornando-se cada vez mais necessário a busca por ferramentas e plataformas que sejam capazes de simular ou desempenhar as tarefas realizadas pelo homem.

É nesta vertente que surge a área do Processamento de Linguagem Natural (PLN) [Vie15] e [Fer17].

Com o grande progresso na área, *softwares* capazes de resolver variadas questões relacionadas com a temática linguística foram surgindo, muitas dessas ferramentas apresentam desvantagens, pois tornam-se complexas (a vários níveis) aquando da instalação, utilização e aprendizagem dos utilizadores [Fer17].

Por esse e outros motivos trabalhos têm sido desenvolvidos, com o objetivo de fornecer técnicas e potencializar funcionalidades, com foco na facilidade, simplicidade e eficiência, tirando proveito das múltiplas tecnologias atuais. E neste contexto se insere também a presente Dissertação. Pois nos dias de hoje observamos um conjunto de plataformas capazes de interagir e comunicar com os seres humanos. Sendo cada vez mais comum tal interação e o PLN é uma ferramenta que facilita esse processo. Segundo [Ros11] o PLN mostra-se como uma área de investigação em Ciências da Computação, sendo uma subárea da IA. E pode ser definido como a habilidade que um computador possui em processar a linguagem usada no dia-a-dia pelos seres humanos. Embora também lide com Corpus, seus interesses são essencialmente aplicados. Sendo que, não é objetivo para o PLN descrever problemas, mas sim criar soluções em termos de menor custo e maior benefício, dos problemas pontuais relativamente ao reconhecimento e a reprodução da linguagem humana em alguma escala. Soluções diferentes devem ser comparadas e avaliadas entre si, em termos de precisão, abrangência e margens de erro embutidas.

Vários pesquisadores são de opinião que o PLN seja uma das áreas mais importantes da IA. Evidenciando sua atratividade nas aplicações que facilitem o relacionamento do usuário com a máquina, tornando-o o mais natural possível, facilitando assim, o acesso às informações que o usuário só poderia obter tendo um conhecimento prévio de alguma linguagem de programação.

Em [EF16] é observado que o PLN permite gerar frases escritas em língua natural, através da incorporação de métodos formais para análise textual. Considerando como objetivo final do PLN capacitar computadores a ”entender” e a ”compor” textos em língua natural. Por ”entender”, [EF16] refere-se a possibilidade dos computadores serem ”capazes de reconhecer o contexto, fazerem a análise sintática, semântica, léxica e morfológica, criarem resumos, extraírem informação, gerarem traduções automáticas, interpretarem os sentidos e até ”aprenderem” conceitos, fazendo uso de textos processados”.

Para *Jurafsky e Martin* [JM14] o PLN visa alcançar que computadores executem tarefas relevantes envolvendo a linguagem humana, estas tarefas podem ser: permitir a comunicação homem-máquina, melhorar a comunicação humano-humano ou simplesmente extrair informação relevante de texto ou fala.

No contexto atual as técnicas de PLN têm vindo a ser amplamente empregues na área da IA, particularmente na Aprendizagem de máquina, com enfoque ao processamento automático da linguagem humana e linguística computacional. Estudando múltiplos subtemas como Pesquisa e Recuperação de Informação, Tradução Automática de Texto, Mineração e Extração de Texto, Sumarização Automática, Deteção Automática de Plágio, Análise de Sentimentos em Texto, Reconhecimento de Voz, Análise Morfológica e Sintática, Semântica e Discursiva, Similaridade e Alinhamento Textual, Geração e síntese, Caracterização Estética do Texto, Auxílio a Escrita, Busca de Respostas para Perguntas; entre outras áreas de investigação que vão surgindo gradualmente [Fer17].

2.1.1 *Big Data*

Com as técnicas computacionais para análise automática e representação da linguagem humana, novos desafios têm surgindo nas últimas décadas. A pesquisa do PLN tem conhecido uma evolução, desde a época de cartões perfurados e processamento em lote (em que a análise de uma frase poderia demorar até 7 minutos) [CW14] para a era da internet, tecnologias da informação, *Google* e mais recentemente, *Web* 2.0, redes sociais *(MySpace, Delicious, LinkedIn e Facebook), smartphones* [PdLC]. Quando, anteriormente havia apenas algumas dezenas de *Exabytes* de informações na *Web*; Hoje essa mesma quantidade, é criada semanalmente (em que milhões de páginas podem ser processadas em menos de um segundo) [CW14]. Ao longo das duas últimas décadas, verificou-se um aumento de dados em grande escala nos mais variados campos. Dando origem a um novo termo: *Big Data*. Um estudo realizado, permitiu prever que a quantidade de dados continuará a aumentar muito nos próximos anos, pois grandes quantidades de dados são gerados diariamente e por várias fontes na era digital [ATH16]. O estudo, evidenciado na década do ano 2000, no relatório apresentado pela *International Data Corporation (IDC)*[1] [GR15], indica que o volume total de dados criados e copiados no mundo foi de 1.8 *ZB* (aproximadamente $10^{21}B$), que aumentou quase nove vezes em um período compreendido a cinco anos. Este valor seria redobrado pelo menos todos os outros dois anos seguintes [CML14]. *Richard e Petri*

[1] Provedor de inteligência de mercado, serviços consultivos e eventos para os mercados de tecnologia da informação, telecomunicações e tecnologia de consumo.

[ATH16], evidenciam que o mundo gerou mais de 1ZB de dados em 2010 e, até 2014, 7*ZB* por ano. Em Fevereiro de 2016 uma notícia divulgada pela *Cisco (ComputerWorld, 2016)*, dizia que o tráfego de dados móveis iria crescer oito vezes nos próximos cinco anos, de acordo a previsão feita os dispositivos móveis inteligentes e suas conexões representarão 72 por cento do total de dispositivos e conexões móveis em 2020, 36 por cento maior que ao que foi registado em relação ao ano de 2015. Os dispositivos inteligentes serão responsáveis por 98 por cento do tráfego de dados móveis em 2020 [com18].

Uma estimação feita sobre a produção mundial de conteúdo digital (EMC, 2018) apontam para a cifra de 44 *Zettabytes* gerados no ano de 2020, o que equivale a mais de 5 *Terabytes* para cada ser humano vivo [EMC18].

Com o aumento explosivo da quantidade de dados globais, como consequência de diversos dispositivos utilizados, sendo que esses dados são originários de todos os lados: sensores utilizados para acumular dados de atmosfera, postagens em sites de redes sociais, imagens digitais e vídeos, sinal de *GPS, smartphones*, sistemas informáticos e dispositivos informatizados [SA16]. A expressão *Big Data* é empregue para descrever enormes conjuntos de dados, geralmente não estruturados e que precisam ser analisados em tempo real. Surgindo instigações no que diz respeito às formas e técnicas de organização e gerenciamento desses grandes conjuntos de dados de forma eficaz.

A importância de *Big Data* é amplamente reconhecida, mas uma vez que o seu conceito é relativamente abstrato, opiniões diferentes têm surgindo na sua definição, mediante a forma como se apresentam denotando ou não o grau de importância.

Em geral, *Big Data* (grande volume de dados) devem significar os conjuntos de dados que não podem ser percebidos, adquiridos, gerenciados e processados pelas ferramentas tradicionais de TI e *Software / hardware* dentro de um tempo tolerável. Devido à diferentes preocupações, empresas científicas e tecnológicas, pesquisadores, analistas de dados e técnicos têm diferentes definições para o termo.

Segundo *ERL, Khattak e Buhler* (2016), *Big Data* é um campo da ciência que lida com a análise, processamento e armazenamento de grandes *DataSets*[2]. Os autores defendem que, *Big Data* não é apenas uma tecnologia, mas é também sobre como as tecnologias podem impulsionar uma instituição; visto que a quantidade de dados tem aumentado de forma exponencial, as instituições têm vindo a procurara alternativas para a gestão e organização dos mesmos. Dessa forma e através de algoritmos avançados e técnicas de análise e processamento, os dados podem ser aproveitados. Levando as instituições a descobrir padrões ocultos e usar o conhecimento adquirido para alcançar vantagens competitivas na sociedade atual, onde a competitividade é fator essencial para a sobrevivência das organizações. Conhecimento esse que deve ser fundamentado e apresentado através das contribuições dos principais especialistas em seus respetivos campos, *Big Data*: algoritmos, análises e aplicações bem como os métodos computacionais apropriados para a descoberta científica e social [EKB16].

Em [DMGG16], através de uma análise conjunta das definições existentes e dos principais temas

[2] *DataSets* são coleções ou grupos de dados relacionados, onde cada grupo ou membro do grupo compartilha as mesmas propriedades ou atributos.

de pesquisa na literatura concluem que *Big Data* é o recurso de informação caracterizado por um alto volume, velocidade e variedade, exigindo uma tecnologia específica e métodos analíticos para sua transformação em valor. Sendo que Volume, Velocidade e Variedade, descrevem as características da informação. Para descrever os requisitos necessários para o uso adequado dessas informações é empregue Tecnologia e Métodos analíticos, e o "Valor", refere-se então a transformação da informação em *insights* que possam criar valor económico para as empresas e para a sociedade [DMGG16].

Para o SAS[3] *Big Data* é o termo empregue para descrever o imenso volume de dados, estruturados e não estruturados, presentes nos negócios ou atividades das empresas no dia a dia. Sua análise pode ser feita para a obtenção de *insights* que levam a melhores decisões e direções estratégicas de negócio [Ins]. Apesar do processo de recolha e armazenamento de grandes quantidades de informações para posterior análise de dados ser feito já a muito tempo, o conceito *Big Data* ganhou força somente no início dos anos 2000, quando o analista *Doug Laney*, articulou a definição de *Big Data* como os três Vs:

- Volume: As organizações coletam dados de uma grande variedade de fontes, incluindo transações comerciais, redes sociais e informações de sensores ou dados transmitidos de máquina a máquina. Novas tecnologias como *Hadoop*[4] (plataforma de software em Java de computação distribuída voltada para *clusters* e processamento de grandes volumes de dados, com atenção a tolerância a falhas) têm facilitado o armazenamento de grandes quantidades de informações que no passado constituiria um grande problema.
- Velocidade: Os dados fluem em uma velocidade sem precedentes e devem ser tratados em tempo hábil. *Tags de RFID*, sensores, celulares e contadores inteligentes estão impulsionado a necessidade de lidar com imensas quantidades de dados em tempo real.
- Variedade: Os dados são gerados em todos os tipos de formatos de dados estruturados, dados numéricos em bancos de dados tradicionais, até documentos de texto não estruturados, *email*, vídeo, áudio, dados de cotações da bolsa e transações financeiras [Ins].

O SAS, também considera mais duas dimensões a respeito do conceito *Big Data*: Variabilidade e Complexidade [Ins].

Para [LJYC15], o termo *Big Data* refere-se à um conjunto enorme de dados, com origem em diversas fontes de informação digital, incluindo informação recolhida através de sensores, *scaners*, textos, telefones celulares, Internet, vídeos, emails, redes sociais entre outros; sendo necessário um sistema capaz de lidar com o armazenamento, análise e processamento dessas informações.

Com o exposto pelos diferentes autores, podemos extrair a informação de que *Big Data* não é somente um sistema de gerenciamento de banco de dados robusto capaz de guardar enormes quantidades de dados; é muito mais, pois busca obter relacionamentos e padrões no vasto conjunto e desorganizado de dados e produz informações valiosas. A importância do *Big Data*

[3] SAS (Sistema de Análise Estatística), Disponível em: https://www.sas.com/pt

[4] O *Hadoop* é uma implementação de código aberto do paradigma de programação *Map-Reduce*. *Map-Reduce* é um paradigma de programação introduzido pelo Google para processar e analisar grandes conjuntos de dados.

não é diretamente proporcional a quantidade de dados que este armazena, mas sim para quê esses dados servem. Permitindo a análise e resolução de problemas que envolvam redução de custos, de tempo, desenvolvimento de novos produtos e ofertas otimizadas, gerando deste modo decisões mais inteligentes.

2.1.2 *Text Mining*

No processo de *Text Mining* (TM) são utilizadas técnicas de: Processamento de Linguagem Natural, método que visa melhorar o entendimento da linguagem natural através de técnicas para processar textos rapidamente, tendo como suporte o computador. Recuperação de Informação: através do uso de métodos e medidas estatísticos ou semânticos para o processamento automático do texto de documentos que possuem a resposta para a questão (mas não somente a resposta em si). Extração de Informação: Possui como principal objetivo buscar partes relevantes de um texto em um documento e extrair informações específicas destas partes. Possui um conceito mais limitado da compreensão da linguagem natural(MACHADO et al., 2010) [Pez17].

A TM é uma subárea do *Data Mining* (Mineração de Dados[5]), desenvolvida na procura de técnicas e processos para extração de informação e conhecimento a partir de dados do tipo não estruturados, pois quando utilizados dados textuais é necessário lidar com informação sem uma prévia estruturação, ou seja, ao contrário dos dados estruturados em tabelas de Base de Dados, os dados textuais não estão organizados em campos com tipos de valores, o que é designado como dados não estruturados. Dessa forma, dados em formato textual são de difícil tratamento e análise, quando se perspetiva retirar dos mesmos informação relevante [Cap].

A título de exemplo as tabelas seguintes ilustradas: onde, se pode verificar a mesma informação em formato estruturado[6] na tabela 2.1 e de forma não estruturada[7] na figura 2.2

Tabela 2.1: Informação Estruturada

NumEst	NomeEst	TurmaEst	CustPropina	Curso
M8094	C Ana	A1	1200	Informática
M8056	J João	A2	600	Química
M9050	A Melo	A3	1100	Hidráulica
M9057	B Silva	A4	1200	Matemática
M9082	V Neves	A5	900	Informática

[5]Mineração de Dados é o processo de exploração e análise de uma grande quantidade de dados com a finalidade de encontrar padrões e extrair novas informações. A Mineração de Dados trabalha principalmente com dados estruturados em um SGBD

[6]Dados estruturados são organizados em linhas e colunas, geralmente encontrados em banco de dados relacionais, são eficientes quanto à recuperação e processamento.

[7]Dados não estruturados são geralmente dados de difícil acesso e recuperação, pela forma como se apresentam e muitas vezes não dispõem de componentes necessários para identificação de tipo de processamento e interpretação, tornando o seu uso um desafio.

Olá! Eu sou a Ana Costa, estudante de Mestrado do Curso de Informática, com o número M8094 da Turma A1 e o custo de propina do Curso que frequento é de 1200 euros anuais, pagos na sua totalidade. João Jorge é estudante da Turma A2 do Curso de Mestrado em Química sob o número M8056 e tem paga 600 do total de 1200 euros do custo geral do Curso. Já o Melo António, estudante do Curso de Hidráulica da Turma A3, é número M9050 e pagou de propina 1100 euros. Bruna Silva frequenta o Curso de Mestrado em Matemática na Turma A4 sob o número M9057 e pagou 1200 euros logo no início do ano para fazer cobertura a propina anual do seu Curso. O meu colega Neves Vidal é número M9082 pagou 900 euros de propina, faltando por pagar 300 euros, trato-lhe por colega por frequentar o mesmo Curso que eu, mas faz parte da Turma A5.

Figura 2.2: Informação Não Estruturada

Em [GL09] é defendido que um dos grandes objetivos do TM[8], prende-se com a análise computacional de textos, com suporte à técnicas como recuperação de informação *(information retrieval)*, extração de informação (onde são identificadas frases-chave e a sua relação com o texto), sumarização que consiste na condensação do texto mantendo a informação, a classificação em que são atribuídas classes aos documentos, entre outras técnicas do PLN.

TM pode ser definida assim, como um conjunto de técnicas para obtenção de informações, em grandes quantidades de texto, significativas a partir de um texto não estruturado ou semi-estruturado. A TM pode ser compreendida como um processo que visa descobrir informações por meio da identificação de padrões e relações em dados relevantes. Caracterizando-se pela interação de um usuário com uma coleção de dados (neste caso, texto) ao longo do tempo, por meio de um conjunto de ferramentas de análise [Cap].

Sendo que as aplicações que fazem uso destas técnicas obedecem a algumas etapas: aquisição, pré-processamento, indexação ou transformação, mineração de dados e avaliação [RM05], [col] e [Cap]:

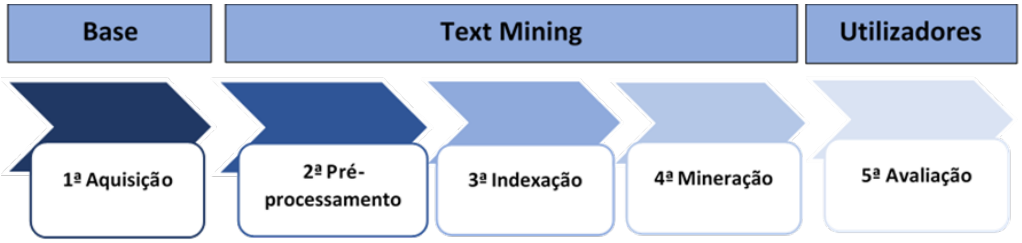

Figura 2.3: Metodologia da Mineração de Textos

1. Aquisição.

   Na etapa inicial do processo, efetua-se a coleta dos textos para a formação do Corpus ou Corpora. podendo ser feita de várias maneiras, como por exemplo através de Robôs de *Crawling* atuando em qualquer ambiente, exigindo por isso um grande esforço, a fim de se conseguir material de qualidade e que sirva de matéria-prima para várias bases de conhecimento.

2. Pré-processamento.

[8] técnica também conhecida como descoberta de conhecimento de textos *Knowledge Discovery in Text* (KDT)

Após o processo de coleta, efetua-se a preparação de dados ou pré-processamento, através de técnicas do PLN, e esta etapa tem como objetivo prover alguma formatação e representação da massa textual para que a mesma esteja apta para as etapas subsequentes.

3. Indexação.

   Processo em que ocorre a seleção e extração de atributos, organizando todos os termos adquiridos das várias fontes de dados, visando gerar a melhor representação dos mesmos, agrupando-os em índices, de forma a facilitar o seu acesso, recuperação e identificação de características, de determinado atributo para um documento. Sendo que uma boa estrutura de indexação garante agilidade no processo de recuperação de informação.

4. Mineração de Dados.

   Também conhecida como etapa de classificação, é responsável pelo desenvolvimento de cálculos, inferências onde serão aplicados os algoritmos de mineração de dados (classificação, regressão, segmentação, associação e análise) e que tem como objetivo a extração de conhecimento, descoberta de padrões e comportamentos. E de acordo a necessidade da aplicação escolhe-se o algoritmo correspondente.

5. Análise, Avaliação ou Interpretação.

   Nesta fase é avaliado o modelo criado a fim de constatar sua eficácia. Tal constatação é dada através da avaliação do algoritmo de Mineração de Dados escolhido para a aplicação em causa. Observando medidas estatísticas, precisão e confiabilidade. A precisão verifica se o algoritmo é correto o suficiente para continuar com a execução, e a confiabilidade se encarrega de analisar qual o nível de sucesso que o algoritmo obteve em relação ao seu conjunto de treinamento. Caso os parâmetros não forem satisfeitos é necessário regressar às etapas anteriores procurando descobrir a origem do problema.

2.1.3 Etapas de Mineração de Textos

Nesta subsecção descrevemos com mais detalhes as etapas a serem observadas na Mineração de Textos.

2.1.3.1 Aquisição

Etapa encarregue de coletar o material para a montagem dos Corpora. É considerada como a etapa principal da Mineração de Textos, pois é nela onde são recolhidos os documentos que constituirão o conjunto de dados, sobre o qual assenta todo o restante processo. Esta etapa pode ser desafiadora e bastante custosa, a começar pela descoberta da localização das fontes de dados. Em [col] é apresentado três ambientes de localização das fontes: pastas de arquivos encontradas no disco rígido de usuários, tabelas de diversos bancos de dados e a Internet. A coleta de documentos no disco rígido de um computador exige bastante cautela, pois faz-se necessário observar a distinção entre arquivos textuais produzidos por pessoas e arquivos binários e de configuração (normalmente interpretados apenas pela máquina). Existindo alguns sistemas que possam facilitar quanto ao gerenciamento de documentos eletrónicos, como é o caso dos sistemas de Gerenciamento Eletrónico de Documentos (GED) ou *Enterprise Content Management* (ECM)

Já a obtenção dos documentos a partir de tabelas de banco de dados dá-se essencialmente através do conteúdo de colunas do tipo string, sem nenhuma restrição a não ser a quantidade máxima de caracteres suportada por registo. Sendo por isso necessário um pré-processamento dos dados, provendo a limpeza dos mesmos a fim de garantir qualidade no conjunto de dados que serão disponibilizados. A Internet constitui o terceiro ambiente de localização de fontes de dados. E devido a sua extensão (constituída por uma infinidade de tipos de página, como notícias de revistas, *bloggers*, anúncios, documentos, artigos técnicos e planilhas), a heterogeneidade é o desafio predominante. Para o processo de coleta na Internet é comum e torna-se imprescindível a utilização de ferramentas de apoio [col]. Essas ferramentas podem ser classificadas em duas categorias: Diretórios de Assunto *(Subject Directories)* e Motores de Busca [V$^+$17]. Com base nessas duas categorias, outros tipos de ferramentas têm surgido, tornando os serviços de busca complexos e volátil [Cen01]. Em função as características específicas de cada ferramenta, pode existir uma variação enorme no que diz respeito ao tipo, número e qualidade dos recursos recuperados. Para a obtenção de melhores resultados no processo de busca e recuperação da informação é necessário entender as peculiaridades dos diferentes tipos de ferramentas de busca na Web para que a ferramenta escolhida possa proporcionar eficiência da busca de informação [V$^+$17]. Sendo que a principal diferença entre os mecanismos de busca é a forma de compilação dos seus bancos de dados.

1. Diretórios de Assunto *(Subject Directories)* De acordo o fundamentado em [Cen01], os Diretórios surgiram num período em que o conteúdo disponível na Web era significativamente pequeno, o que permitia a coleta sem recorrer aos mecanismos automáticos. Teve sua origem nos mecanismos de busca por palavras-chave. Esta ferramenta de busca na Web, constitui a primeira tentativa de solucionar problemas derivados da recuperação da informação, mecanismos de busca por palavras-chave.

   Segundo o *Your Dictionary*, um diretório de assunto é um banco de dados online de sites da Web, cujas informações contidas são organizadas por assunto e categoria [You18].

   Ao contrário dos mecanismos de pesquisa, os Diretórios de Assuntos, são criados e mantidos por editores humanos, e não por *Spiders* ou Robôs. As páginas são rastreadas (visitadas), indexadas e armazenadas por assunto. Cabe aos seus editores, determinar o valor do site para posterior inclusão em seus diretórios com base em critérios de seleção previamente determinados. Seus visitantes podem detalhar a categoria de interesse a pesquisar e suas subcategorias [Cha18]. O usuário digita seus termos de pesquisa e analisa os links das categorias e dos menus retornados, no geral organizados do mais amplo ao restrito.

   De acordo [KQRdC17], diferentes dos Motores de Busca (que permitem a busca e recuperação de qualquer tipo de informação na Internet, e ordena-as segundo o critério de relevância definido pelo usuário), os Diretórios são ferramentas que organizam os seus conteúdos de forma genérica ou temática, sua BD é menor (pois são manipuladas por humanos e não fazem uso de robôs), contendo informações com maior relevância, pois indexam essencialmente as páginas principais e por isso podem ser mais apropriados para busca por tópicos.

   Os Diretórios podem referir-se ainda a uma coleção de índices e BD, organizações ou assuntos, listas alfabéticas ou classificadas organizadas por nomes de arquivos, contendo as informações que possibilitam a recuperação pelo sistema operacional (títulos, endereços, afiliações e outros dados profissionais), de livre acesso [BCB16]. Como são criados e man-

tidos por editores humanos a probabilidade de retornar conteúdo não relacionado ao tema de busca, é muito menor.

Podemos observar que através dos Diretórios é possível aprofundar ou recuar o nível de pesquisa, mediante a necessidade do usuário. É possível buscar somente o conteúdo selecionado. Muitas vezes os termos Motores de Busca e Diretórios são usados para referenciar a mesma coisa, mas como pudemos observar em [Cha18], [KQRdC17] e [BCB16] não o são. Sendo assim os Diretórios podem classificar-se em [BCB16]:

- Diretórios Institucional, compreende a produção científica de uma determinada Instituição;
- Diretórios Temáticos, abrange a produção científica de uma área específica do conhecimento;
- Diretórios Governamentais, regista documentos ou notícias governamentais;
- Diretórios Agregadores, reúne no mesmo local, para melhor visualização pelo usuário, um conjunto de registos atualizados de outros repositórios. Ex: *The Web Directory, HCC LIBRARIES ONLINE*.

Os Motores de Busca muitas vezes recorrem aos Diretórios, incluindo seus links com o objetivo de oferecer opções de maior seletividade de recursos, as informações contidas no banco de dados dos Diretórios são coletados através da busca realizada por seus editores; que visitam vários sites e de acordo o seu interesse os vão incluindo ao banco de dados, acompanhados de uma breve descrição de seus conteúdos; uma outra forma é através de solicitações de inclusão enviada pelo autor interessado em ter seu site catalogado. O autor envia uma breve descrição do conteúdo que deseja, e os editores visitam o site, aceitando ou não sua inclusão. Essas informações são organizadas e classificadas de forma hierárquica em função das categorias temáticas definidas pelos editores. Obedecendo a uma organização que parte das categorias mais amplas para as mais específicas. Os Diretórios auxiliam os Motores de Busca no processo que realizam, pois estes, podem servir de entrada para o algoritmo do *Crawler* [Bra04] e [KQRdC17].

2. Motores de Busca Baseados em Robô *(Robotic Internet Search Engines)*.

   O conjunto de recursos de informação na Web adquiriu proporções inestimáveis, dificultando cada vez mais a busca através da navegação e a coleta de informações de forma manual. Dando assim origem aos chamados Motores de Busca Baseados em Robô. Estes tipos de Motores de Busca são formados por 4 componentes [V$^+$17]:

   (a) Robô (Mecanismo responsável pela localização e busca de documentos na Web. Também chamados de *Spiders* (aranhas), *Crawler* ou *Web Crawler* (Rastreador *Web*) percorrem a *internet* em intervalos regulares, procedendo à leitura dos seus conteúdos e seguindo os links que direcionam à outras páginas de forma recursiva. Os documentos encontrados são então encaminhados para os Indexadores e estes extraem a informação das páginas armazenando-as em uma BD).

   (b) Indexador (Encarrega-se da extração da informação dos documentos e constrói a base).

   (c) Motor de Busca (Motor de Busca propriamente dito utilizado).

   (d) Interface (Responsável pela interação com o usuário).

Importa salientar que *Web Crawler* (também conhecido como Web *Spider, Bots Scutter, Bot Crawler e Automatic Indexer*, é um programa e/ou *script* focado para a *World Wide Web (WWW)*); é o nome dado aos robôs especializados em navegar na Internet, de forma autónoma e exploratória, com o objetivo de realizar a coleta automática de documentos [MPdSK16].

A base de funcionamento de um Rastreador da *Web* consiste em uma lista de *URLs* iniciais a serem visitadas, denominadas de *seeds* (sementes), geralmente definidas de forma manual. Um exemplo de sementes pode ser a página pessoal do autor, *Home Page* de um laboratório ou departamento, embora o último possa não conter nenhuma publicação, mas pode levar à página à publicações.

Um rastreador Web, além de possuir uma boa estratégia de rastreamento, também deve ter uma arquitetura altamente otimizada. Na 2.4 é ilustrada a arquitetura padrão de um Web *Crawler*, cujo funcionamento pode ser resumido nas seguintes etapas [Rod16]:

(a) Parametrização: Nesta etapa o *Crawler*, recebe um conjunto de *URLs* como dados de entrada (*seeds*) e/ou uma descrição do tópico pretendido. Esta descrição normalmente é um conjunto de palavras-chave para *Crawlers* clássicos e semânticos ou um tipo de treino no caso de ser um *Crawler* inteligente.

(b) Download: O download do conteúdo das páginas é feito pelo *Crawler* e os links das mesmas colocados em *queue* (fila), e de acordo a relevância, os *links* são ordenados ou eliminados da fila.

(c) Processamento do Conteúdo: As páginas cujo *download* é feito na etapa anterior, são lexicalmente analisadas e reduzidas a vetores, e o conteúdo é filtrado e guardado (este processamento de conteúdo é realizado por *parsers* (no ramo computacional, são programas informáticos usados para análise de sequência de caracteres, sejam estes caracteres parte de uma estrutura linguística natural ou de uma linguagem informática, e respeitando a língua em questão).

(d) Designação de prioridade: Nesta etapa e de acordo a tipologia do *Crawler* utilizado e das especificações dadas pelo utilizador, os links extraídos anteriormente são colocados em fila ordenada. Estas especificações podem ir deste critérios simples como importância da página ou relevância com o tópico pedido.

(e) Expansão: Os *URLs* escolhidos são então usados para expandir o processo de busca pelo *Crawler*, utilizando-os como parâmetros de entrada e vai repetindo o ciclo, cabe ao utilizador definir o critério de parada (como por exemplo número de páginas limite a transferir) ou mesmo até os recursos do sistema estarem esgotados.

Uma vez que um *Spider* captura informações das páginas, cadastrando os links encontrados, isso facilita a localização de outras páginas e mantém a BD atualizada. Para maior eficiência no processo, chamado de Web *Crawling* ou *Spidering*, realizado pelos *Crawlers*, existem algumas ferramentas que permitem a indexação das páginas de forma mais rápida. Entre elas destacamos:

- *Sitemap.xml*: *Sitemap* é considerado um mapa do seu site que através do qual é indicado ao Robô quais as páginas a serem indexadas e armazenadas nos servidores. *Sitemap.xml* é então o arquivo XML simples que contém a lista de todas as páginas de um site e através do acesso a ela, o *Crawler* identifica as páginas existentes para indexar, garantindo maior eficácia[Far17].

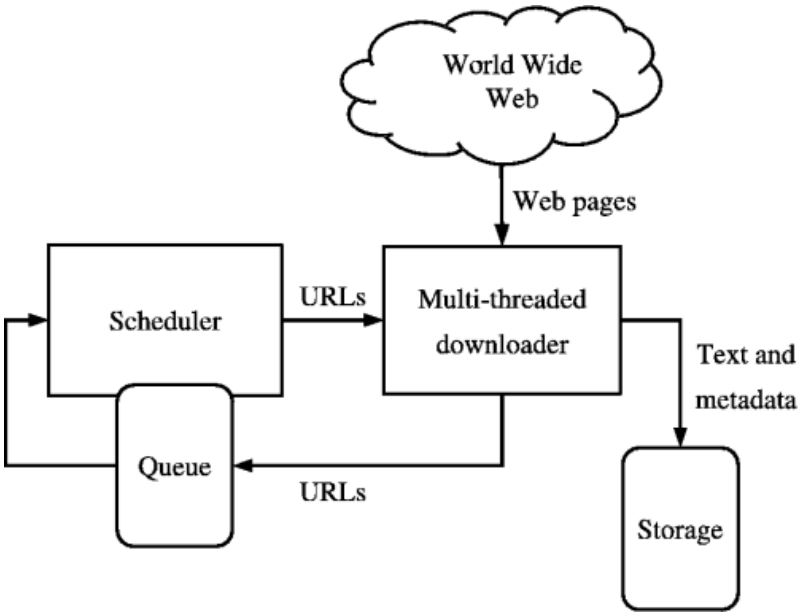

Figura 2.4: *Architecture of a Web crawler*

- *Robots.txt*: O padrão de exclusão de robôs (*robots exclusion standard*), também conhecido como protocolo de exclusão de robôs (*robots exclusion protocol*) ou simplesmente *robots*.txt, é um padrão usado por sites para comunicar-se com os rastreadores Web e outros Robôs da *Web* . Consiste em um padrão que especifica ou informa o Robô da Web sobre quais áreas do site não devem ser processadas ou verificadas. Sendo que nem todos os Robôs implementam o padrão. O arquivo que apresenta-se em formato de texto trabalha como um filtro, avisando aos *Crawlers* quais páginas e diretórios não devem ser indexados [con18a].

Muitas vezes essas ferramentas servem como uma medida de segurança, pois os Rôbos consomem recursos em sistemas visitados e estão sujeitos a visitar sites sem aprovação. Dessa forma é necessário que questões de programação, carga e "polidez" sejam colocadas em prática quando grandes coleções de páginas são acessadas. Por exemplo incluir um arquivo *robots*.txt pode solicitar bots (Agentes de Software) para indexar apenas partes de um site ou nada [con18b].

Os *Crawlers* podem ainda ser usados para tarefas de manutenção automatizadas em um Web Site, como por exemplo checar os links ou validar o código *Hypertext Markup Language* (HTML) e ainda obter informações específicas das Páginas da Web (como por exemplo minerar endereços de email, mais comummente para *spam*).

Um *Spider* normalmente desempenha a função de explorar toda a Web com vista a recolher e classificar conteúdos como páginas web, imagens, vídeos, ficheiros, etc. Começando pela visita a uma lista de *URLs* também chamado de *seeds* (sementes), identificando todos os links na página e os adiciona na lista de *URLs* para posterior visita [con18b].

Entre os *Crawlers* mais conhecidos podemos citar: o *Googlebot* (nome do *Crawler* do *Google*), *Yahoo Slurp* (nome do *Crawler* do *Yahoo*), *DuckDuckBot* (*Web Crawler* do *DuckDuckGo*), *Msnbot* (nome do *Crawler* do *Bing - Microsoft*) [con18b]. Podemos ainda citar entre o vasto conjunto de Rastreadores *Open Source* disponíveis na Web para Mineração de Dados, os apresentados na tabela 2.2.

Tabela 2.2: Alguns Rastreadores disponíveis na *Web* para mineração de dados.

Nome	Linguagem	Plataforma	Licença
Heritrix	Java	Linux/Unixlike	Apache License, version 2.0
Scrapy	Python	Cross-platform	BSD License
GNU Wget	C++	Linux	GNU General Public License, version 3+
JSpider	Java	Cross-platform	GNU Library or LGPLv2
Xepian	C++	Windows	GNU General Public License, version 2 (GPL v2+)
OpenWebSpider	C#, PHP	Cross-platform	MIT License (MIT)
Arachnode.net	C#	Windows	GNU General Public License, version 2 (GPLv2)
Apache Nutch	Java	Cross-platform	Apache License, Version 2.0

Viikmaa A. (2016) [Vii16], defende que quando um Web *Crawler* é usado para extração de dados, este deve identificar e guardar as páginas que contêm os dados procurados para uma futura extração. Tal procedimento deve ser feito através da marcação (através de regras definidas manualmente ou construídas de forma automática) de cada página percorrida como uma página alvo, página com dados a serem extraídos, ou como uma página de navegação ou ainda marcada como uma página cujo acesso pode ser necessário para alcançar as páginas alvo.

Quando o objetivo é acessar conteúdos da Web invisível *(Deep Web) Crawlers* específicos podem ser feitos e utilizados para este fim. Através dos mesmos é então possível acessar, coletar e indexar conteúdos que não são indexados pelos mecanismos de busca padrão. *Crawlers* construídos para esse fim podem localizar também páginas escondidas atrás de formulários e códigos *Javascript*. [AE17]

Para a referida colheita é necessário que as páginasm HTML sejam interpretadas de uma forma correta, com a devida identificação dos links das páginas visitadas. E gerenciar bem o caminho do percurso tomado, que tem a forma de um grafo, para evitar e impedir que o robô visite várias vezes a mesma página ou entre em ciclos eternos. Visto que geralmente um Web *Crawler* captura uma ou várias *URLs* de uma página e enquanto durar a navegação entre as mesmas vai capturando também e de forma recursiva, as *URLs* da fronteira que respeitem os parâmetros definidos, formando assim uma BD [Fer17].

Se o rastreador estiver executando o arquivamento de sites, ele copia e salva as informações conforme elas são enviadas. Os arquivos geralmente são armazenados de forma que possam ser visualizados, lidos e navegados como se estivessem na Web ao vivo, mas preservados como ”instantâneos” [con18b].

Algumas vezes o objetivo perseguido é o de acessar apenas páginas específicas e não fazendo *crawling* a *Web* toda. Para tal é necessário a utilização de *Crawlers*, que façam uma coleta de forma orientada. Os robôs que empregam este tipo de mecanismos são chamados de *Crawlers* Focados. Um *Crawler* Focado coleta páginas da Web que satisfazem alguma propriedade específica definida, priorizando cuidadosamente a fronteira de rastreamento e gerenciando o processo de exploração do *hiperlink*. Sendo altamente efetivo no que diz respeito a construção de coleções de documentos de qualidade com origem na Web e considerados mais eficazes que os *Crawlers* normais, porque tentam direcionar a captura dos dados à páginas de interesse do usuário e através do uso de algoritmos específicos, possibilitam a identificação de documentos similares, agilizando assim a busca e dispensando o uso de grandes recursos de hardware [M+17] e [Fer17].

São considerados como exemplo de Rastreadores Focados os rastreadores académicos, cujo objetivo é rastreiam documentos académicos de acesso livre, como o *citeseerxbot*, que é o rastreador do mecanismo de pesquisa *CiteSeer X*. Outros mecanismos de pesquisa académica são o Google Académico e a Pesquisa Académica da Microsoft entre outros. Como a maioria dos trabalhos académicos é publicada em formatos PDF, esse tipo de rastreador está particularmente interessado em rastrear PDF, arquivos *PostScript, Microsoft Word* e seus formatos compactados [con17].

Os *Crawlers* são classificados não só mediante o método de pesquisa, mas também através da escolha das prioridades de páginas selecionadas, em [Rod16]:

- *Crawlers* clássicos: nos *Crawlers* clássicos, o utilizador fornece a entrada descrevendo o tópico (ou um conjunto de *URLs* de páginas) que guiam o *Crawler* para páginas de interesse. Definindo neste conjunto de entradas, o critério de prioridades para dar a certos links maior prioridade de download baseado na probabilidade desse mesmo link conter dados sobre o tópico de interesse do utilizador.
- *Crawlers* semânticos: é um dos subtipos dos *Crawlers* clássicos. Onde as prioridades de download são concedidas a páginas que sejam semanticamente semelhantes ao critério de entrada.

  Um *Crawler* semântico é considerado como um Rastreador Focado, pois este, faz uso de ontologias de domínio para representar mapas de tópicos e ligar páginas da *Web* com conceitos ontológicos relevantes para os propósitos de seleção e categorização [con17].
- *Crawlers* inteligentes: são *Crawlers* guiados, ou seja é implementado um processo de treinamento para guiar o *crawling* e definir a prioridade das páginas a visitar. O *Crawler* inteligente aprende a identificar páginas relevantes e a seguir links que contenham conteúdo que sejam relevantes no contexto da busca realizada.

  O conjunto de treino para esse tipo de rastreador, é constituído por Páginas Web relevantes e não relevantes. E perante o tópico escolhido, são definidas ordens de prioridade de visita sobre os links extraídos da Web de acordo a relevância dos mesmos. Para tal, é avaliado o conteúdo da página e a correspondente classificação da página (relevante ou não relevante) e também a estrutura de links subjacentes e a probabilidade dessa estrutura ser relevante consoante o número de saltos entre páginas necessários para encontrar o conteúdo pretendido.

Os Rastreadores Focados apresentam-se como ferramentas eficazes para aplicativos que exigem um alto número de páginas pertencentes a um tópico específico, utilizado por indivíduos ou instituições que buscam manter portais da Web ou coleções específicas de documentos da Web localmente. São considerados um tipo de Rastreadores inteligentes (e também semânticos), pois efetuam também um *crawling* guiado (recebem um conjunto de *seeds* e através dos mesmos, em função dos critérios previamente definidos, determinam os links a serem visitados, em caso de satisfação do critério de relevância, são então as páginas baixadas e armazenadas em um repositório, a eficiência do *Crawler* irá depender em grande escala das *seeds* selecionadas) minimizando assim a quantidade de recursos de armazenamento. Sendo por isso apresentada na literatura um conjunto de estratégias para implementar estes *Crawlers*, que visam melhorar a eficiência do rastreamento, aumentando o número de páginas relevantes recuperadas, evitando páginas não relevantes [VBDS+16].

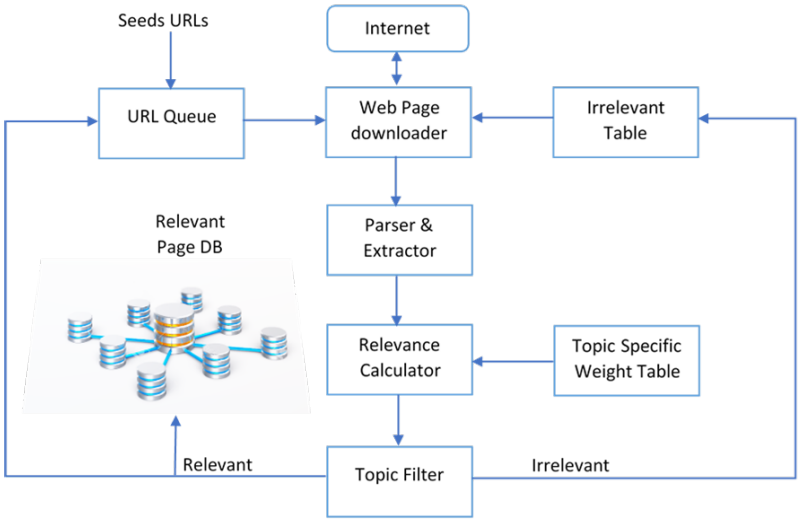

Figura 2.5: *Architecture of focused web crawler* (extraído de [PC15])

Na figura 2.5 é possível observar o funcionamento de um Web *Crawler* Focado. Onde, a fila de *URLs* da arquitetura, contém *URLs* sementes mantidas pelo rastreador e inicializadas com URLs que ainda não foram visitadas. A *Web Page downloader* busca *URLs* da fila de URL e transfere as páginas correspondentes da Internet. O *Parser* e o Extrator são encarregues de extrair informações como o texto e as *URLs* de *hiperlinks* de uma página baixada. A Calculadora de relevância *(Relevance Calculator)* calcula a relevância de uma página em relação ao tópico e atribui pontuação às URLs extraídas da página. O filtro de tópicos *(Topic Filter)* analisa se o conteúdo das páginas resultantes do *Parser* está relacionado a um tópico ou não. Se a página for relevante, as *URLs* extraídas dela serão adicionadas à fila de *URLs*. Caso contrário, será adicionado à tabela irrelevante [PC15].

Os Rastreadores Focalizados devem calcular as prioridades dos links não visitados para que se possam orientar quanto a recuperação das páginas da Web relacionadas a um determinado tópico. As prioridades para os links são afetadas por semelhanças tópicas dos textos completos e pelas características (textos âncora, link-contexto) desses *hiperlinks*. A fórmula é definida como [LZZH16]:

$$\textit{Priority}(l) = \frac{1}{2}.\frac{1}{n}\sum_{p}^{n} Sim(u_p,t) + \frac{1}{2}.Sim(f_l,t), \quad (2.1)$$

Onde $Priority(l)$ é a prioridade do link $l(1 \leq 1 \leq L)$ e L é o número de links. n é o número de páginas da Web recuperadas, incluindo o link l. $Sim(u_p,t)$ é a semelhança entre o tópico t e o texto completo (u_p), que corresponde a página da Web p incluindo o link l. $Sim(f_l,t)$ é a similaridade entre o tópico t e o texto âncora f_l correspondente aos textos âncora, incluindo o link l.

O funcionamento e um Web *Crawler* Focado contempla diferentes abordagens, existindo os: Baseados em prioridade, baseados na estrutura, no contexto e os baseados em aprendizagem [PC15].

Existem ainda outras categorias de classificação dos *Web Crawler* [Fer17]:

Os de Propósito Geral: que realizam a coleta e o processamento dos dados recuperando

um grande número de sites e páginas, independentemente do seu tópico.

Os chamados *Web Crawler* Incremental: estes possuem a capacidade de incrementação, permitindo fazer sucessivas leituras, aproveitando os conteúdos lidos anteriormente que se mantiveram e adicionando novos conteúdos. Sua utilização é recomendável quando os dados são coletados em páginas que sofrem constantes mudanças.

E *Web Crawler* Profundo: específico para a realização de buscas em páginas que não são indexadas por buscadores comuns, como por exemplo as alocadas na segunda camada da *World Wide Web (Internet)*, a chamada de Internet profunda *(Deep Web)*. Um exemplo de uma página da Web localizada na Internet profunda são as contas bancárias online. Nestes casos a página é hospedada na Internet, mas os mecanismos de pesquisa não têm acesso para rastreá-la. *Intranets* de empresas, sites internos de hospitais e clínicas médicas, bancos de dados de instituições governamentais e não governamentais e ainda sites privados que exigem um *login* para acessar o conteúdo constituem outros exemplos.

Ao início da execução, o Rastreador identifica e extrai todas as *URLs* existentes em *hyperlinks* que a página contenha e os adiciona à fila (Queue) de *URLs* a serem visitados (chamada de fronteira de rastreamento) [con18b].

De acordo as estruturas de dados usadas, os *Web Crawlers* são classificados em *Breadth-first*, que utiliza fila do tipo *first in first* out *(FIFO)*, e *Preferential* (ou Focados), que utiliza fila de prioridade.

Um *Breadth-first Web Crawler* percorre as *URLs* por ordem de chegada, já um *Preferential Web Crawler* atribui prioridades às *URLs* através de métricas previamente definidas e por isso podem ser bem mais eficientes em manter atualizados índices ou coleções de páginas com conteúdos dinâmicos e também específicos [AE17].

Breadth-First Crawling é o método de rastreamento mais simples. Este método recupera todas as páginas ao redor do ponto de partida antes de seguir os links mais distantes desde o início. Essa é a abordagem mais comum em que robôs ou rastreadores seguem todos os links. Se o rastreador estiver indexando vários *hosts*, neste tipo de abordagem, a carga é rapidamente distribuída, implementando assim o processamento paralelo [PC15]. A figura 2.6 ilustra uma busca utilizando o método *Breadth First*.

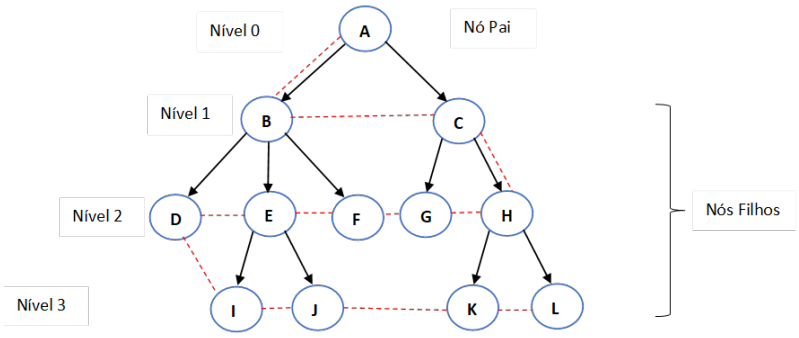

Figura 2.6: *Breadth First Search (BFS)*

Outras abordagens também são consideradas:

- *Fish-Search*

  A abordagem *Fish-Search* é baseada num modelo em que o rastreador ou conjunto de rastreadores são vistos como um cardume, nesse caso, se o peixe encontrar uma

página relevante com base nas palavras-chave especificadas na consulta, ele continuará procurando seguindo mais links dessa página. Se a página não for relevante, os links derivados receberão um valor preferencial baixo [PC15]. Esta abordagem atribui valores de prioridade binários (1 para relevante, 0 para não relevante) para páginas candidatas a download por meio de simples correspondência das referidas palavras-chave. Logo, todas as páginas relevantes recebem o mesmo valor de prioridade [BPM09].

- *Shark-Search*:

  O método *Shark-Search* é uma modificação da abordagem *Fish-Search*, sugere o uso do Modelo de Espaço Vetorial (VSM) [9] para atribuir valores de prioridade não binários às páginas candidatas. Os valores de prioridade são calculados levando em consideração o conteúdo da página, o texto da âncora, o texto em torno dos links e o valor da prioridade das páginas consideradas páginas pai (páginas que apontam para a página que contém os links) [BPM09]. Embora esse algoritmo use a mesma metáfora simples da abordagem *Fish-Search*, ele descobre e recupera informações mais relevantes no mesmo tempo de exploração com habilidades de pesquisa aprimoradas [AXD15]. O *Shark-Search* pode ainda ser visto como uma variante do *Best-First Crawler* com uma função de atribuição de prioridades mais complicada [BPM09]. A abordagem *Shark Search* usa melhores técnicas de pontuação de relevância para as páginas vizinhas antes de acessá-las e analisá-las (*Parsing*). O sistema tem um impacto significativo na eficiência da pesquisa devido à melhoria do sistema de classificação de relevância. Ao invés da avaliação binária de relevância, o método usa uma pontuação entre 0 e 1. Oferecendo por isso melhores resultados quando comparado a classificação binária. A segunda melhoria é o método de herança dos nós filhos. O sistema dá a cada nó uma pontuação herdada que tem um impacto enorme na pontuação de relevância dos nós filhos e dos filhos desses nós. O sistema calcula ainda a relevância dos filhos usando não apenas a herança dos ancestrais, mas também usa meta-dados para analisar sua pontuação de relevância. Sendo esta a melhoria mais significativa e de acordo os resultados de uma experiência realizada em [AXD15] a pesquisa *Shark Search* é mais eficaz na qualidade da informação recuperada e no tempo de operação do que seu ancestral *(Fish-Search)* [AXD15].

- *Depth First Crawling*: O algoritmo *Depth First Crawling*, segue todos os links que contenha a página inicial, e posteriormente segue o primeiro link na segunda página, e esse processo continua. Depois que a primeira página é indexada, segue o primeiro link da segunda página e os links subsequentes [PC15]. Este tipo de *Crawlers* implementam pilhas *(Stack)* do tipo *LIFO (last in, first out)*.

- *The Best-First Search*: *Crawlers* que empregam este tipo de abordagem, concentram-se na recuperação de páginas que são relevantes para um determinado tópico específico. O algoritmo usa um tipo de contagem para definir qual página tem melhor pontuação. *Best-First Crawlers* atribuem valores de prioridade às páginas candidatas, calculando a semelhança de texto com o tópico, aplicando o VSM. O algoritmo em causa emprega uma regra para selecionar a melhor página. Na maioria dos casos, algoritmos de IA como *Naive Bayes*, Similaridade de Cosseno, Vetores Frequência do termo (tf), algoritmo de k vizinhos mais próximos, modelo de mistura de Gauss, entre

[9] VSM, representa documentos e consultas como vetores de termos. Termos são ocorrências únicas nos documentos.

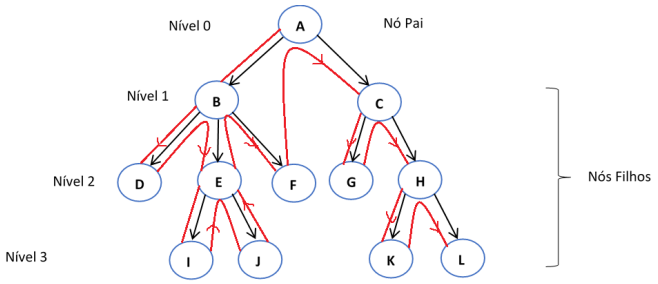

Figura 2.7: *Depth First Search (DFS)*

outros para calcular a relevância do tópico e das páginas e obter melhores resultados [AXD15].

Estes tipos de *Crawlers* efetuam uma busca focada que explora um árvore expandindo o nó mais promissor, este nó é selecionado de acordo com uma regra específica definida pelo utilizador. O princípio do *Best-first search algorithm* está em avaliar o quão promissor é o nó n por uma função de estimação heurística $f(n)$ que, geralmente, depende da especificação de n, da especificação do objetivo, a informação reunida pela busca até aquele ponto e, o mais importante, qualquer conhecimento adicional sobre o domínio do problema [AXD15].

- *N-Best First Crawler*.

  O *N-Best First Crawler* é uma versão generalizada do *Best-First Crawler*: em cada etapa, as N páginas (em vez de apenas uma) com a prioridade mais alta são escolhidas para expansão. E através de uma combinação de conteúdo de página, informações de string de URL, páginas irmãs e estatísticas sobre páginas relevantes ou não relevantes é possível obter melhores resultados, para a atribuição de prioridades a páginas candidatas. Resultando em um algoritmo de rastreamento altamente eficiente que aprende a rastrear sem o treinamento direto do usuário [BPM09].

- *Naive Best First Method*.

  Nesta abordagem é explorado o fato de uma página relevante estar vinculada à outra página relevante. A relevância de uma página A para o tópico específico t, apontada pela página B, é estimada pela relevância da página B para o tópico t. Cada página é representada como um vetor de pesos correspondentes às frequências normalizadas dos termos do documento, de acordo com o modelo *Term Frequency-Inverse Document Frequency* (TF-IDF) (que significa frequência do termo inverso da frequência nos documentos). Neste método, o termo frequência é calculado, que é a frequência do termo dentro de um documento e a frequência inversa do documento, em quantos documentos o termo ocorre [PC15].

- *Info Spiders Algorithm*.

  InfoSpiders é um sistema multiagente que fazem uso de Redes Neurais, no qual cada agente em uma população de pares se adapta ao seu ambiente de informação local aprendendo a estimar o valor dos *hiperlinks*, enquanto a população como um todo tenta cobrir todas as áreas promissoras através de reprodução seletiva. Os *InfoSpiders*

Crawlers complementam mecanismos de pesquisa baseados em índices tradicionais usando agentes no lado do utilizador. Quando o usuário envia uma consulta, os *Info Spiders* obtêm um conjunto de links iniciais que são os resultados da pesquisa de um mecanismo de pesquisa tradicional. O sistema baseado em agente pode navegar online em nome do usuário e evoluir um comportamento inteligente que explora a ligação da Web e as sugestões textuais. Estes, buscam somente o ambiente atual e portanto não retornará informações obsoletas, e terá uma melhor chance de melhorar a regência dos documentos visitados. Um agente é inicializado para cada link e analisa os links da página correspondente procurando o próximo link a seguir. O agente analisa os links calculando a similaridade do texto ao redor o link inicial, com a ajuda de uma Rede Neural. O próximo link a ser seguido é escolhido com uma probabilidade proporcional ao *Score* de similaridade. Os pesos da Rede Neural são ajustados pela relevância do conteúdo da nova página para que o agente atualize seu conhecimento. Através dessas técnicas de IA o *Crawler* adapta-se às características do seu ambiente de informação em rede [PC15] e [AXD15].

- *Page Rank Algorithm*.

  O algoritmo *Page Rank* determina a importância das Páginas Web de qualquer site, fazendo a contagem das citações ou *backlinks* (*links* de entrada) para uma determinada página. Sendo que a classificação da página *(page rank)* é dada por [MK17]:

  (a) Cálculo do *page rank* de todas as páginas através da fórmula:

$$PR(A) = (1-d) + d(PR(T1)/C(T1)) + ... + PR(Tn)/C(Tn) \quad \textbf{(2.2)}$$

  Onde, $PR(A)$ corresponde a *Page Rank* do site A,
  d é o fator de amortecimento, que pode ser definido entre 0 e 1, mas normalmente é ajustado para 0,85
  $PR(Ti)$ é o *PageRank* das páginas Ti que apontam para a página A,
  $C(Ti)$ é o número de links externos na página Ti.

  (b) Repita a etapa 1 até que se faça corresponder os valores de duas iterações consecutivas.

  Quando o número de *backlinks* é alto, a página representa maior interesse em relação a uma outra com o valor inferior. Logo a soma total ponderada de links de entrada define o *page rank* de uma página da Web [GGS17].

2.1.3.2 Pré-processamento

Feita a coleta dos documentos que irão formar o Corpus, é necessário que os mesmos sejam preparados, com o objetivo de transformar a grande quantidade de textos obtidos para que apresentem melhores resultados. Essa etapa é denominada de Pré-processamento [Cap].

O Pré-processamento é uma técnica que constitui um dos principais componentes em muitos algoritmos de TM e PLN e de grande importância, pois através do emprego desta técnica é possível organizar a informação coletada, apresentando os dados de forma estruturada e remover dados desnecessários, oferecendo assim a informação estruturada para as próximas etapas, permitindo um processamento mais eficiente [APA+17] e [San15]. Etapa que consiste na decomposição de cada documento em termos e suas respetivas frequências.

Termos menos significativos podem ser descartados, bem como aqueles que apresentem uma frequência elevada. Esta etapa pode em alguns casos ser muito onerosa, dependendo dos algoritmos utilizados para o efeito, uma vez que não existe uma técnica padronizada que possa ser aplicada para a obtenção de uma representação textual satisfatória em todos os domínios. Na fase de pré-processar textos, é feita a filtragem e a limpeza dos dados contidos nos documentos iniciais e escritos em linguagem natural, com o objetivo de eliminar redundâncias que possam existir e informações irrelevantes do contexto pretendido, mas preservando suas características próprias: Dessa forma é possível extrair os termos que são de grande importância para a aplicação dos algoritmos de extração de conhecimento [DB17] e [Pez17].

Para [APA+17] uma estrutura tradicional de categorização de texto (também conhecida como classificação de texto), é a tarefa de classificar automaticamente um conjunto de documentos em categorias (classes ou tópicos) de um conjunto predefinido. Obedece as seguintes etapas: pré-processamento, extração de recursos, seleção de recursos e etapas de classificação. Sendo a etapa de pré-processamento aquela que consiste em tarefas como tokenização, filtragem, lemmatização e *stemming*.

Para o presente trabalho foram definidas algumas etapas de pré-processar dados e com base as mesmas, é eliminado tudo aquilo que represente conteúdo irrelevante para a pesquisa, através de ações ou técnicas cujo objetivo é melhorar a eficácia do modelo a utilizar [San15], [DB17], [dSPB17] e [Pez17]:

- Correção Ortográfica.

  Permite eliminar os erros ortográficos que possam existir no texto, através do uso de um corretor ortográfico associado a um dicionário de línguas. Neste processo cada termo do texto é comparado com os termos do dicionário e corrigidos se for o caso. Dessa forma palavras que contenham erros ortográficos e que não entrariam nos termos relevantes do documento, passam a fazer parte e são consideradas para as etapas subsequentes. Constituindo o principal objetivo desta técnica, evitar que se percam palavras-chaves (por estarem mal escritas) na classificação [San15].

- *transform cases*.

  Esta técnica é utilizada para transformar os caracteres contidos no documento, de maiúsculas *(upper case)* a minúsculas *(lower case)* e vice-versa. Cujo objetivo é evitar que a mesma palavra quando escrita com maiúsculas seja considerada diferente quando escrita em minúsculas ”exceção” e ”Exceção”. Na bibliografia consultada a abordagem utilizada é a *lower case* (transformação de todos os *tokens* em letra minúscula), tornando assim a representação documental homogénea.

- *Tokenização*.

  A *tokenização* é o processo de dividir um fluxo de texto ou uma sequência de caracteres em pedaços (palavras/ frases), símbolos ou outros elementos significativos chamados *tokens*. Este processo é acompanhado da técnica de Remoção de *Non Letters*, cujo objetivo é eliminar todos os *tokens* que não sejam realmente palavras. Muitas vezes e consoante o *Dataset*, é possível definir uma biblioteca que permita remover outros termos irrelevantes. Esta técnica é que permite remover símbolos, sinais de pontuação, caracteres alfanuméricos e caracteres especiais. O objetivo da *tokenização* é a exploração das palavras em uma frase [VRG14]. A lista de *tokens* é posteriormente usada no processamento [APA+17].

- Remoção de *stopwords*.

  Um texto padrão contém artigos definidos e indefinidos, preposições, pronomes, numerais, conjunções e advérbios. As palavras pertencentes as classes gramaticais referidas são importantes na construção do discurso expresso por meio do texto, mas em tarefas de Mineração de Texto em que o que se persegue é descobrir padrões, essas palavras são consideradas irrelevantes; pois por si só não carregam um sentido semântico e são frequentes em qualquer tipo de texto de um idioma. Dessa forma é constituída uma lista *(stoplist)* composta pelas palavras que não são bons discriminadores e que não induzam a algum sentido semântico, como consequência fornecem pouca informação quanto ao sentido da frase, e por isso devem ser removidas do texto. Palavras com essa característica são denominadas *stopwords* [dSPB17].

  A remoção de *stopwords* pode fornecer um tamanho do documento mais reduzido em relação ao inicial e melhora a ordenação na recuperação, proporcionando uma diminuição quanto a complexidade do conjunto de dados submetido ao processo de mineração. Este processo permite que o documento contenha termos com maior representatividade dentro do contexto e por isso mais discriminatórios. É necessário observar que existem exceções e que devem ser tidas em conta no processo de remoção de *stopwords*, visto que em domínios específicos, muitas fezes torna-se imprescindível a presença de algumas *stopwords*. Razão pela qual engenhos de busca utilizarem representação do texto completo [dSPB17]. Por essa razão a *stoplist*, apesar de incluir as classes gramaticais referidas, pode ser modificada ou constituída de acordo o contexto do problema e os resultados que se pretendam obter [dSPB17].

  A lista de *stopwords* varia de uma língua para outra.

- *Stemming*.

  Um termo em um documento pode sofrer variações de acordo a forma como estiver apresentado (singular, plural, gerúndio, verbos flexionados, aumentativo, diminutivo); para [dSPB17], na análise documental, termos com o mesmo radical não devem ser tratados como diferentes, faz-se então necessário um processo que reduz os termos ao seu radical denominado *stemming*, onde são eliminados os prefixos e sufixos tornando assim os termos uniformizados. O *stemming* permite reduzir a dimensionalidade dos dados a serem indexados. Este método, conhecido também por lematização (em português), é o processo de reduzir palavras que encontram-se em formas derivadas ou flexionadas para o radical *(stem)* de origem (sua forma base) removendo as variações de palavras do tipo (plural, gerúndio, prefixos, sufixos, género e número) de modo que a palavra fique só com a raiz e assim considerada como um só *token* para a representação de todas as variações [San15]. A análise de radicais e semelhança de palavras varia de uma linguagem para outra. Por exemplo, considerando a lematização o processo de deflexionar uma palavra para determinar o seu lema, (as flexões resultantes são denominadas lexemas[10]), considere a palavra gato; os seus lexemas seriam gata, gatos, gatas... De igual modo Cantor, Cantora, Cantores, Cantoras fazem parte do mesmo lexema, pois são de mesma classe (verbos) e se diferem apenas por sufixos (morfema zero,-a,-es,as). A literatura apresenta diferentes técnicas para buscar o radical de um termo como [dSPB17]:

  - *Table look-up*, mantén-se os radicais das palavras em uma tabela que representa um dicionário, é um processo simples, mas que necessita dos dados de todos os

[10]Conjunto de palavras de mesma classe morfológica que diferem entre si por sufixos reflexivos.

radicais da língua;

- *Affix removal*, remove os prefixos e sufixos de um termo;
- *SucessorVariety*, considera os morfemas da língua, *Ngrams*, onde os termos são separados em n *tokens*, procedimento complexo e significativamente dependente da língua.

Algoritmos de *stemming* têm sido propostos, sendo um dos mais utilizados o algoritmo de *Porter*, descrito em [Por80] e usado como base para a implementação de diversos softwares especializados *(stemmers)*.

Thesaurus e *Ngrams* são utilizados para auxiliar o processo. Palavras diferentes podem ter o mesmo significado dentro de um domínio específico de conhecimento e por isso podem ser normalizadas, apoiando-se num dicionário de thesaurus que relacionam palavras com o mesmo significado, o que permite restringir o sistema a um vocabulário controlado para indexação e busca de documentos, bem como ajudar os usuários a localizar termos para a formulação de consulta adequada e fornecer hierarquias classificadas que permitam o alargamento e estreitamento da solicitação de consulta atual de acordo com as necessidades do usuário [KCS+15]. Reduzindo assim o conjunto dos dados, o que diminui significativamente o custo computacional nas etapas subsequentes [Pez17]. Os *Ngrams* consistem na criação de uma subsequência de palavras num comprimento n:

$$Ngrams_k = X - (N - 1) \quad (2.3)$$

Onde K corresponde a frase e X o conjunto de palavras que constituem a frase K.

Quando o *Ngram* tem tamanho 1, é chamado de *Unigram*, em caso de possuir tamanho 2, *Bigram* e assim por diante. Está técnica permite criar um grande número de *tokens* (nº de *tokens*- 1) a somar aos *tokens* individuais contidos no texto [San15].

- *Prune Method*.

  Após a etapa de remoção de *stopwords*, o texto pode ainda apresentar palavras que não acrescem valor ao texto, por serem usadas com muita frequência ou pouca frequência (palavras muito frequentes em um Corpus de texto normalmente serão frequentes em outros Corpora de texto, não agregando por isso valor as análises, e palavras que pouco aparecem têm uma chance pequena de reaparecer em outros textos). O *Prune Method* é aplicado para reduzir o número de termos indexados.

  O operador *Prune Method* especifica se palavras com frequência elevada ou demasiadas infrequentes devem ser ignoradas para a lista de palavras que irão formar o conjunto de dados para a indexação. É uma técnica que permite avaliar dados e frequências diferentes [San15]:

  - Percentagem, define uma determinada percentagem e ignora palavras que estejam presentes em menos ou mais da percentagem escolhida, de entre todos os documentos da coleção.
  - Frequência Absoluta e Relativa. A Frequência Absoluta também conhecida por frequência do termo ou *Term Frequency (TF)*, representa a medida da quantidade de vezes que um determinado termo aparece em um documento. O modelo ignora palavras que estejam presentes em menos ou mais do que X documentos

da coleção. Em alguns casos essa medida não é aconselhável, pois a quando da realização da análise de coleções de documentos, este modelo não é capaz de fazer distinção entre os termos que aparecem em muitos ou em poucos documentos. E também não leva em conta a quantidade de palavras existentes em um documento. Desta feita pode ocorrer que uma palavra pouco frequente em um documento pequeno possa ter a mesma importância de uma palavra que seja muito frequente em um documento grande. Já a análise baseada na Frequência Relativa considera o tamanho do documento (quantidade de palavras que ele possui) e normaliza os pesos de acordo com essa informação. Sendo a frequência relativa (Frel) de uma palavra X em um documento qualquer, calculada da seguinte forma [ea13]:

$$F_rel(X) = \frac{F_abs(X)}{N} \quad (2.4)$$

Onde:
F_abs é a frequência absoluta e N o número total de palavras no mesmo documento.

- *Ranking*, especifica qual a percentagem das palavras mais frequentes ou menos frequentes que serão ignoradas.

O pré processamento geralmente é realizado com o objetivo de transformar o texto em uma estrutura de dados que pode ser utilizada eficientemente por algoritmos de aprendizagem. Os strings são transformados em um vetor de palavras. Em que o valor contido em cada posição do vetor é um fator de ajuste.

Com a etapa de Pré-processamento, o objetivo é aumentar a qualidade dos dados iniciais, e através da aplicação e/ou combinação de diversas técnicas consegue-se obter dados de melhor qualidade e que servirão de entrada para a execução da etapa de Mineração, precedida ou não da de Indexação [DB17].

2.1.3.3 Indexação

Esta etapa é encarregue de armazenar as palavras contidas no texto em uma estrutura de índices a fim de facilitar a pesquisa em documentos de um Corpus através das palavras que o mesmo contenha, evitando que se percorra toda a BD [GP17].

A semelhança do sumário de um livro (lista detalhada com os assuntos abordados e a localização dos mesmos), os índices são usados para otimizar o processo de busca e recuperação de termos ou documentos relevantes. Utilizando no processo de indexação um mecanismo denominado indexador, que é responsável por armazenar as páginas capturadas pelo *Crawler* e suas características em um Índice. Permitindo assim uma fácil localização das páginas através de consultas ao Índice. Muitas fezes os algoritmos utilizados para a busca já incorporam mecanismos de indexação, fornecendo assim resultados organizados [Bil16]. A indexação é o processo encarregue por criar estruturas auxiliares que garantem rapidez e agilidade na recuperação dos documentos e seus termos. Após esse processo os documentos e as características neles contidas são analisadas por algoritmos a fim de extrair conhecimento e identificar padrões úteis [NZ16].

Para auxiliar nesta tarefa, arquivos invertidos são utilizados para a implementação de índices ordenados. Um arquivo invertido pode ser entendido como uma lista ordenada de palavras chaves, em que cada palavra chave contém um apontador para os respetivos documentos em que a mesma ocorre, bem como sua posição. O objetivo do uso de índices invertidos ou arquivos invertidos é o de melhorar o desempenho no processo de busca e recuperação [GP17].

2.1.3.4 Mineração de Dados

Para [DB17], é na etapa de Mineração de Dados que ocorre efetivamente a busca por novos conhecimentos, através da aplicação de técnicas direcionadas ao aprendizado de máquina.

Nesta fase é selecionada a tarefa a realizar de acordo com a tarefa de Mineração de Textos (definida no início do processo) bem como a necessidade do utilizador. Um exemplo de necessidade do utilizador pode prender-se com o facto do mesmo querer verificar o grau de similaridade e a formação de grupos naturais, neste caso a tarefa adequada seria então a clusterização. Em caso dos *Clusters* já estarem formados, seja por conhecimento prévio do especialista ou pela execução de algoritmos, o ideal é empregar algoritmos de classificação e que possam identificar características individuais em um documento ou grupo de documentos.

2.1.3.5 Análise, avaliação ou interpretação

Nesta etapa é avaliado todo o processo realizado desde a coleta, analisando a eficácia dos dados obtidos após à aplicação dos algoritmos para mineração de dados. É nesta fase onde é possível avaliar se o objetivo de extrair novo conhecimento do Corpus foi alcançado [DB17].

No trabalho levado a cabo por [DB17] é ressaltada a importância da análise individual dos resultados de cada etapa, realizando alterações no processo em caso de resultados não satisfatórios. Sendo por isso um processo cíclico [DB17]. O processo de mineração de texto finaliza assim com a etapa de Análise, onde é realizada a avaliação e interpretação de todo o conhecimento alcançado pelo processo [NZ16].

2.1.4 Métodos de Seleção de Características

A quantidade de dados de alta dimensão que existe na Internet e publicamente disponível, aumenta cada vez mais. O que leva pesquisadores a enfrentarem desafios no que concerne aos métodos de Aprendizado de Máquina quanto a extração do grande número de características. Exigindo por isso o pré-processamento dos dados. Uma característica é uma propriedade mensurável individual do processo que está sendo observado. Através do uso de um conjunto de recursos, qualquer algoritmo de aprendizado de máquina pode executar a classificação [CS14]. Dessa forma, a identificação de recursos relevantes tornou-se uma tarefa essencial em Mineração de Dados e Aprendizado de Máquina, tendo sido desenvolvidas técnicas para descobrir automaticamente o conhecimento e reconhecer padrões desses dados, desenvolvendo e empregando algoritmos com eficácia aplicados ao mundo real [KM14] e [TAL14].

Normalmente os dados coletados costumam estar associados a um alto nível de ruído de diferentes fontes, dentre os quais podemos citar a imperfeição nas tecnologias que coletaram os

dados e a fonte dos dados em si. Se tivermos que considerar por exemplo uma imagem médica, qualquer deficiência que tenha o dispositivo de coleta da imagem será refletida como ruído para o processo posterior. Outro exemplo a ter em conta é a qualidade dos dados de mídia social, que varia de dados excelentes para *spam* ou abuso de conteúdo por natureza. E geralmente esses documentos apresentam erros gramaticais, erros de ortografia e pontuação imprópria, uma vez que são escritos de forma informal. Sem dúvida, extrair conhecimento e padrões que se apresentem úteis de tais dados enormes e ruidosos é uma tarefa desafiadora [TAL14].

Neste contexto, recursos irrelevantes são aqueles que não fornecem informações úteis e os recursos redundantes, aqueles que não fornecem mais informações do que os recursos atualmente selecionados [KM14]. Esta redução de recursos visa transformar o conjunto original de recursos em novos recursos, através da aplicação de alguma função de transformação. O novo conjunto de recursos criado contém muito menos recursos ou dimensões que o conjunto original [PK15].

A redução de dimensionalidade é uma das técnicas mais populares para remover recursos ruidosos (isto é, irrelevantes) e redundantes. Técnica essa que pode ser categorizadas principalmente em [TAL14]:

- *Feature Extraction* e;
- *Feature Selection*.

A extração de características *(Feature extraction)* aborda os recursos do projeto em um novo espaço de recursos com menor dimensionalidade e os novos recursos construídos são geralmente combinações de recursos originais [KM14], ou seja, o espaço de recurso original é convertido em um novo espaço mais compacto. Todos os recursos originais são transformados nesse novo espaço reduzido, sem excluí-los, mas substituindo características originais por um conjunto representativo menor. Quando o número de recursos nos dados de entrada é muito grande para ser processado, estes serão transformados em um conjunto de recursos de representação reduzida [ZS15]. Exemplos de técnicas de extração de características incluem Análise de Componentes Principais *(PCA- Principle Component Analysis)*, Análise Linear Discriminante *(LDA- Linear Discriminant Analysis)* e Análise de Correlação Canónico *(CCA- Canonical Correlation Analysis)* [Agg14].

A seleção de características *(Feature Selection)* é uma das técnicas mais frequentes e importante na etapa de pré-processamento de dados, sendo por isso indispensável. Esta técnica reduz o tempo de computação, pois permite realizar o processo de deteção de recursos relevantes e efetua a remoção de dados irrelevantes, redundantes ou ruidosos. Aspetos que aceleram os algoritmos de mineração de dados, melhorando a precisão da previsão e como consequência aumenta a compreensão dos dados em aplicativos de Aprendizado de Máquina ou reconhecimento de padrões [KM14].

É designado *Feature Selection* o processo de seleção de características relevantes ou um subconjunto de recursos candidatos. Critérios de avaliação são usados para obter um subconjunto de características ideais (geralmente envolvem a minimização de uma medida específica de erro preditivo para modelos adequados a diferentes subconjuntos. Os algoritmos buscam um subconjunto de preditores que modelem de forma ideal as respostas medidas, sujeitas a restri-

ções, como recursos relevantes ou não, e o tamanho do subconjunto). As principais vantagens do uso de algoritmos de seleção de características é que estes permitem reduzir a dimensão dos dados, tornam o treinamento mais rápido e podem melhorar a precisão removendo recursos ruidosos [tec18]. Para remover um recurso irrelevante, é necessário ter em conta critérios de seleção de características que possam medir a relevância de cada característica individual com a classe ou o que se perspetiva obter como dados de saída. A partir de um ponto de aprendizado de máquina, se um sistema usar variáveis irrelevantes, incorre a uma generalização imprópria [CS14].

A tarefa de seleção de características foca-se na seleção de um subconjunto de variáveis a partir da entrada que minimizam a redundância e maximizam a relevância para o destino, e que possa descrever eficientemente os dados de entrada enquanto reduz os efeitos de ruído ou variáveis irrelevantes e ainda fornecer bons resultados de previsão [CS14] e [TAL14]. Como consequência, a seleção de características pode ajudar a evitar *overfitting* (o *overfitting* ocorre quando um modelo tenta prever uma tendência em dados que são muito ruidosos). Este tipo de modelo ajusta os dados de treinamento muito bem, mas se mostra ineficaz para prever novos resultados. Quando ocorre o *overfitting* implica que o modelo aprende os detalhes e também o ruído nos dados de treinamento, isso afeta negativamente o desempenho do novo modelo de dados. Significa que o ruído ou as flutuações aleatórias nos dados de treinamento são captados e aprendidos como conceitos pelo modelo. O problema é que esses conceitos não se aplicam a novos dados e afetam negativamente a capacidade de generalização dos modelos. Um modelo que é *overfitted* é impreciso porque a tendência não reflete a realidade dos dados [tec18]). *Feature selection* serve a dois propósitos principais. Primeiro, torna o treinamento e a aplicação de um classificador mais eficiente, diminuindo o tamanho do vocabulário efetivo.

A extração de características difere da seleção de características na medida em que a primeira consiste em transformar dados arbitrários, como texto ou imagens, em recursos numéricos utilizáveis que servirão de entrada para o processo de aprendizado de máquina. Já a segunda é uma técnica de aprendizado de máquina aplicada a esses recursos. Os métodos que aplicam a seleção de características tentam encontrar o menor subconjunto de recursos relevantes, de acordo com um determinado critério, não alterando as características e preservando seu significado original para o usuário. Já os métodos que aplicam a extração de recursos tentam reduzir a dimensionalidade dos conjuntos de dados combinando recursos. Tais métodos tentam minimizar a perda de informações, mas, os recursos originais e seu significado para o usuário geralmente são perdidos [PdAL18].

A seleção de características é um assunto amplamente abordado na literatura e continua sendo objeto de estudo para muitos pesquisadores, cujo objetivo é desenvolver novas técnicas para selecionar características distintas para que a precisão da classificação possa ser melhorada e o tempo de processamento reduzido [HR15].

Dentre os vários métodos eficazes existentes para seleção de características podemos citar os seguintes, agrupados em duas categorias, supervisionados e não supervisionados, existindo ainda algoritmos semi-supervisionados [KM14], [ZS15], [HR15], [APA+17], [PdAL18] :

- Métodos supervisionados, são técnicas de aprendizado de máquina pertencentes a inferir uma função ou aprender um classificador a partir dos dados de treinamento, a fim de

realizar previsões sobre dados não vistos [APA+17]:

- Informação Mútua *(Mutual Information- MI)*; O método da Informação Mútua, é aplicado para representar as relações entre a informação e a medição estatística da correlação de duas variáveis aleatórias. Nesta abordagem, a extração de características baseia-se na hipótese de que as palavras têm grandes frequências em uma determinada classe, mas pequenas em outras, e classes têm informações mútuas relativamente grandes. Normalmente, a informação mútua é usada como a medida entre uma palavra característica e uma classe, e se a palavra característica pertence à classe, então tem maior quantidade de informação mútua. Como esse método não requer hipóteses sobre a propriedade da relação entre palavras e classes de recursos, é extremamente adequado para o registo de características de classificação e classes de texto. A complexidade do tempo de computação de informação mútua é semelhante ao ganho de informação. A deficiência de informação mútua é que a pontuação é extremamente impactada pelas probabilidades marginais das palavras [LSSG17].
- Ganho de Informação *(Information gain- IG)*, Técnica de seleção de características que pode diminuir o tamanho dos recursos computando o valor de cada atributo, classificando-os e em função de uma métrica previamente definida, são mantidos os atributos que estiverem acima desse limiar, cujo objetivo é manter os recursos que se apresentarem como os melhores do *ranking*. Geralmente, o ganho de informações seleciona os recursos por meio de pontuações. O ganho de informação de um termo mede a pontuação do mesmo para previsão de categoria pela presença ou ausência do termo em um documento, o que irá refletir discriminação entre as classes [KM14] e [LSSG17].
- χ^2 *(Chi-Square)*;

  χ^2 é um método estatístico de seleção de características. É usado para medir a associação entre um termo e a categoria na classificação do texto, ou seja, o método avalia os recursos individualmente, calculando as estatísticas de χ^2 em relação às classes. E também é usado para testar se a ocorrência de um termo específico e a ocorrência de uma categoria específica são independentes. Se o termo é independente da classe, então sua pontuação é igual a 0, de outro modo 1. Um termo com maior pontuação *chi-Square* é mais informativo. Assim, é estimada a quantidade para cada termo classificando-os por sua pontuação. Se um termo estiver próximo a mais categorias, a pontuação desse termo será maior. Pontuações altas em χ^2 indicam que a hipótese nula de independência deve ser rejeitada e, portanto, a ocorrência do termo e da categoria são dependentes. E desta forma é selecionado o recurso para a classificação do texto [HR15] e [KM14].

  A medida χ^2 de um termo t para uma categoria c é definida como [HR15]:

$$\chi^2(t,c) = \frac{N * (R_a R_d - R_c R_b)^2}{(R_a + R_c) * (R_b + R_d) * (R_a + R_b) * (R_c + R_d)}. \quad (2.5)$$

  Onde:

  N é o número total de amostras de treinamento, R_a é o número de vezes que t e c co-ocorrem, R_b é o número de vezes que t ocorre sem c, R_c é o número de vezes que c ocorre sem t, R_d é o número de vezes em que não ocorre c nem t. A pontuação de

um termo é calculada para cada categoria de forma individual. Essa pontuação pode ser globalizada em todas as categorias de duas maneiras:

1. Calculando a pontuação média ponderada para todas as categorias.
2. Escolher a pontuação máxima entre todas as categorias.

- Métodos não supervisionados, o aprendizado não supervisionado lida com a descoberta de estruturas ocultas em dados não rotulados não precisando de treinamento, portanto, podem ser aplicados a qualquer dado de texto sem esforço manual:

  - Algoritmos de *Clustering*; A técnica usada para fazer agrupamentos automáticos de dados segundo seu grau de semelhança é denominada *Clustering*. Esta técnica é encarregue de segmentar uma coleção de documentos em partições em que os documentos no mesmo grupo *(cluster)* são mais semelhantes entre si do que aqueles que estejam localizados em outros *Clusters*. Estes algoritmos empregam aprendizado não supervisionado para particionar um conjunto de dados Y em *Clusters* $KS = S_1, S_2, ..., S_K$, de acordo com alguma noção de similaridade. Isso significa que eles são capazes de designar uma entidade y_i para um *Cluster* específico sem precisar de dados rotulados para aprender. O principal objetivo deste tipo de algoritmo é gerar um *clustering* S em que há homogeneidade dentro de *Clusters*, mas heterogeneidade entre *Clusters* [PdAL18].
  - Seleção de características usando similaridade de características *(Feature selection using feature similarity- FSFS)*. A seleção de características usando similaridade, calcula semelhanças entre características emparelhadas para determinar um conjunto maximamente independentes e, em seguida, descarta as que são consideradas redundantes [PdAL18].
  - O valor *TF- IDF (Term Frequency - Inverse Document Frequency)*Para [KM14] o valor TF (frequência do termo) é o número de vezes que um termo ocorre em um documento num conjunto de dados. É o critério mais simples para a seleção de termos e pode facilmente ser dimensionado para um grande conjunto de dados com complexidade computacional linear. É um método simples mas eficaz de seleção de recursos para categorização de texto.

    TF- IDF (Frequência do Termo-Inverso da Frequência nos Documentos) determina a frequência relativa de termos em um documento específico através de uma proporção inversa do termo sobre todo o Corpus. Esta técnica é comummente usada para ponderação de termos no campo de recuperação de informações e o peso (composto por dois termos, o primeiro calcula a frequência do termo normalizado (TF) e o segundo a frequência do documento inverso (IDF)) é usado na classificação de texto. *Term Frequency (TF)* mede o número de vezes que um termo ocorre em um documento e é usado para calcular a capacidade de descrição do termo, *Inverse Document Frequency (IDF)* é uma métrica usada para calcular a capacidade de distinção do termo e também mede importância do mesmo [HR15].

$$TF(t) = \frac{N(t_d n)}{N(T_n)} \quad (2.6)$$

Em que: $N(t_d n)$ é o número de vezes que o termo t aparece em um documento;
$N(T_n)$ corresponde ao número total de termos no documento n.

- Força de Termo *(Term Strength- TS)*, *Term Strength* é uma técnica para seleção de recursos em mineração de texto originalmente proposta e avaliada para redução de vocabulário na recuperação de texto. Está técnica não necessita de uma lista pré-definida de Stop *Words*, ela descobre automaticamente por isso, é uma técnica de redução de vocabulário na recuperação de texto. O método estima a importância do termo com base na frequência com que um termo aparece em documentos relacionados. E é calculado com base na probabilidade condicional de que um termo ocorre na segunda metade de um par de documentos relacionados, dado que ocorre na primeira metade:

  $TS(t) = p(t \in d_j | t \in d_i), d_i, d_j \in D \cap sim(d_i, d_j) > \beta$ onde β é o parâmetro para determinar os pares relacionados. Uma vez que é necessário calcular a similaridade para cada par de documentos, a complexidade temporal do TS é quadrática em relação ao número de documentos [KM14]

- *Ranking accuracy using single feature*

Todos esses métodos atribuem uma pontuação a cada recurso individual e, em seguida, seleciona características maiores que um limite pré-definido [KM14].

A seleção de recursos pode ser feita com vários tipos de ferramentas, incluindo *Weka e Scikit-learn*.

2.1.5 Seleção de Características Relevantes através da técnica de Implicação Textual por Generalidade

O ser humano detém uma capacidade extraordinária de expressar o raciocínio através da linguagem natural. Mas, apesar dessa excelente capacidade necessita cada vez mais de ferramentas que o auxiliem no processamento e na análise de discursos escritos, pois a quantidade de dados cresce diariamente e torna-se imprescindível extrair informações relevantes dos mesmos. Através de um texto, ser capaz de compreender o raciocínio contido, é um passo para entender seu conteúdo. Argumentos convincentes denotam um raciocínio lógico e objetivo, permitindo uma melhor avaliação através da sua interpretação consensual. Desta feita, os textos devem ser escritos de forma persuasiva, o que implica habilidades apropriadas de argumentação. Extrair conclusões através de premissas apropriadas pode levar a argumentos convincentes. Observe que quanto menos suposições forem necessárias para interpretar a premissa, mais provavelmente estaremos na presença de uma relação de Implicação, ou seja, uma inferência que emprega raciocínio lógico e objetivo [RLC18].

Como exemplo podemos analisar um texto fonte onde constem informações como camisa, saia, vestido, casaco, a partir do mesmo, o leitor poderá inferir vestuário?

Obviamente que sim. Nesta perspetiva a implicação textual por generalidade é uma abordagem que consiste na análise textual, a partir de uma regra geral e sucede especializando-a gradualmente até adaptar-se aos dados. O processo pode dar-se também de forma inversa sendo por isso uma abordagem *top-down e bottom-up*, significa que tanto pode efetuar a busca de geral para específico, como do específico para o geral. Demonstrando por isso eficiência ao integrar as duas técnicas.

Em PLN muitos desafios surgem quando o objetivo é determinar através de inferências, conclusões a partir de textos escritos em linguagem natural. Para resolver tal problema algumas soluções têm sido propostas, através de tecnologias, métodos e algoritmos como evidenciado em [DPWW+], [DM], [CWVM17], [SNGB15] [RLC18] e outros.

Nos últimos anos pesquisadores têm despertado um interesse crescente no que diz respeito a *Textual Entailment* (TE). Sendo a TE amplamente abordada como uma solução aos problemas genéricos, útil para melhorar a confiabilidade da informação [HTI16].

Em [SNGB15], TE é definida como a relação que determina se um texto pode ser inferido de outro da seguinte forma:

Dados dois textos, um é chamado de Texto denotado como T e o outro de Hipótese denotado como H, o processo Implicação Textual consiste em decidir se o significado de H pode ou não ser inferido logicamente do significado de T.

Em linguagem natural é comum observar a existência de várias formas de expressar, de forma idêntica ou similar um conceito. Avaliar e observar a relação entre textos ou sentenças é um desafio. Neste contexto a tarefa de Reconhecimento da Implicação Textual (*Recognizing Textual Entailment* (RTE)) é empregada para avaliar se o significado de um texto H pode ser inferido de outro texto T. Apresentando-se como uma tarefa de inferência de lógica pura. Persegue-se que através de T se possa inferir H. Ou seja, se a partir de T se puder inferir que H é uma verdade provável e não que T é condição suficiente para H (Sempre que T é verdade então H é verdade) [FP17]. Ou seja, uma sentença T implica outra sentença H se depois de ler e sabendo que T é verdadeiro, um humano inferiria que H também deve ser verdadeiro [RLC18]. Notemos que essa relação não é expressamente direcional, visto que a especialização é por definição uma classe derivada, possui características próprias e que só dizem respeito a mesma classe. Neste sentido ainda que T infere H seja verdade, o contrário (H infere T) tem uma baixa probabilidade de ocorrer.

Por exemplo: Banana é uma especialização de uma fruta. Por sua vez em fruta contém informação essencial sobre qualquer fruta, não contendo nenhuma característica de uma fruta em particular. Enquanto que em banana obtemos informações que só dizem respeito à banana. Em PLN o RTE consiste na identificação de relações de vinculação (implicação) entre fragmentos de texto. Isso significa que um fragmento de texto implica outro fragmento de texto como explicado acima, isto é, se a partir do significado do primeiro, pode-se inferir o significado do segundo. Caso a relação seja bidirecional, então estamos na presença de uma paráfrase [11][RLC18].

RTE é aplicado em tarefas de PLN como: Pesquisa Semântica, Sistemas de Perguntas Respostas, Resumo de Texto e Extração de Informações. Para tal, métodos baseados em aprendizagem de máquina, programação linear, cálculo probabilístico, otimização e lógica, têm sido empregados [LMMyGJS+18].

[11] A Paráfrase é um texto que procura tornar mais claro e objetivo aquilo que se disse em outro texto. Portanto, é sempre a reescritura de um texto já existente, uma espécie de ”tradução” dentro da própria língua.

2.2 Linguística de Corpus

Estudos comprovados demonstram que os homens de diferentes épocas sempre tiveram especial atenção à linguagem. Mas somente a partir do século XX com o desenvolvimento de uma nova ciência, a linguística, é que o homem passou a ter bases, mediante estudos científicos, para descrever ou explicar a linguagem verbal humana [Orl17].

Com o surgimento dos computadores, o estudo da Linguística de Corpus (LC) foi conquistando espaço. Expansão esta que deveu-se ao facto de poder armazenar grandes quantidades de dados em formato eletrónico e pelo processamento dos dados por meio de programas computacionais de forma automática [Mar14]. A LC serve para explorar estatisticamente elementos lexicais, caracterizar géneros e identificar perfis de práticas textuais, localizar padrões de uso, compreender sentidos, etc.

Em [dAM17] é definida LC, como a área que estuda a linguagem por meio de textos autênticos. Textos esses produzidos naturalmente em um contexto real da língua falada ou escrita. Sendo possível extrair palavras-chave da área de estudo e analisá-las em seu contexto de uso, através de ferramentas eletrónicas.

O desenvolvimento tecnológico e a implementação de softwares específicos para a área da pesquisa de linguagem facilitaram para o avanço dos estudos, fazendo a LC ganhar mais destaque. Mas estudos baseado em Corpus datam a muitas décadas, sendo por isso uma prática já muito antiga, tendo surgido antes da Grécia Antiga com Alexandre, O Grande, com o Corpus helenístico e durante a Idade Média os Corpora eram baseados em textos Bíblicos. [dAM17]. Segundo Sardinha (2004) [Sar04], a LC é uma área da Linguística que se ocupa da coleta e exploração de dados extraídos dos Corpora, que é um conjunto de dados coletados criteriosamente para serem objeto de pesquisa.

SARDINHA (2004), evidencia também que foram os linguistas Boas e *Fries* e o educador *Thorndike* a ganharem destaque nas pesquisas baseadas em Corpus no passado século, estes, dedicaram-se ao estudo e descrição da linguagem por meio de análises linguísticas. Existindo duas grandes diferenças entre esta época e a atual. A primeira, prende-se com o facto de que os corpora não eram eletrónicos, ou seja, eram coletados, mantidos e analisados por especialistas da área de forma manual. Quanto a segunda, é que a ênfase destes trabalhos era em geral o ensino de línguas. Quando atualmente o foco na literatura é a descrição de linguagem e não a pedagogia, apesar de existir um mero interesse na investigação da linguagem [Sar04].

Thorndike (1921), levou a cabo um trabalho que consistia na identificação das palavras mais frequentes da língua inglesa. O levantamento foi feito em um Corpus de aproximadamente 4,5 milhões de palavras de forma manual. Esse facto, depois de publicado impulsionou mudanças no ensino de língua materna e estrangeira nos Estados Unidos e na Europa. As abordagens baseadas no controle do vocabulário, nas quais contam-se em primeiro lugar com as palavras mais frequentes, devem sua inspiração a estudos como o de *Thorndike* [Tho21]. Tendo sido considerado um trabalho fenomenal, em função das condições da época.

Mais tarde, no final dos anos 50 surgiu a obra de *Avram Noam Chomsky* denominada *Syntactic Structures*, que remodelou o estudo científico da linguagem, saía de cena o empirismo e com ele a sustentação dos trabalhos baseados em Corpora, tomando lugar central as teorias racionalis-

tas da linguagem [Cho02]. Linguistas e não linguistas reconheceram *Syntactic Structures* como um dos estudos mais importantes do século XX. Além deste facto foram surgindo crescentes críticas sobre o processamento manual de Corpora, o que viria a pôr em causa a confiabilidade dos Corpora processados de forma manual. Pois o ser humano não é "programado" para tarefas deste tipo. Mesmo que a equipe de analistas seja grande, alias a possibilidade de erro e de falta de consistência podem piorar quando envolvidas na resolução do problema uma vasta equipa. A alternativa seria diminuir o tamanho dos Corpora para facilitar a inspeção manual, mas isto não seria viável, pois iria contra a própria natureza da pesquisa. Exigindo cada vez mais um instrumento que permitisse a análise de grandes quantidades de dados de modo confiável, infelizmente a tecnologia da época não o permitia. Sendo a invenção do computador o impulso para a mudança do quadro. Desta feita e porque Corpora coletados manualmente foram tornando-se muito exaustivos (mas ainda utilizados nos dias atuais), ferramentas computacionais têm sido criadas, pois proporcionam cada vez mais ao estudo, velocidade e confiabilidade, e uma grande quantidade destes Corpora foram e vêm sendo compilados e disponibilizados para pesquisa linguística e para a criação de ferramentas de PLN [Sar04]. Visto que o PLN é extraordinariamente dependente da boa qualidade do chamado Corpus e sem o surgimento das referidas ferramentas a LC e os Corpora de grande tamanho seriam inconcebíveis, uma vez que trabalha-se com grandes quantidades de informação textual e impossíveis de processar manualmente com rapidez. Por isso, são usados programas de análise lexical, para efetuar operações de processamento da linguagem, tais como contagem de palavras, geração de listas de frequência, geração de listas de palavras-chave e exibição de linhas de concordância. Ex: *Wordsmith Tools* [12] *AntConc*[13] [PB].

2.2.1 Corpus

O termo Corpus vem do latim e significa corpo, conjunto. É usado para fazer referência a uma coleção cujos textos são criteriosamente selecionados, com amostras escritas ou faladas, processáveis por computador com um propósito de pesquisa linguística ou ainda para um propósito de investigação concreto, apresentando-se no plural como Corpora. Usados muitas vezes no desenvolvimento de ferramentas no campo de PLN em que suas aplicações incluem recuperação de informações, verificação ortográfica, criação de dicionários linguísticos, reconhecimento de fala, síntese de texto, indexação automática, entre outras [Sil08] e [dAG+17].

Para *Rojo, Guillermo* (2014), Corpus é definido como um conjunto de textos naturais, armazenados em formato eletrónico, representativos em seu conjunto de uma variedade linguística em alguns de seus componentes ou na totalidade e reunidos com o propósito de facilitar o estudo científico para o qual foi desenvolvido [Roj14].

A definição apresentada em [Roj14], evidencia que os textos que compõem um Corpus, devem ser naturais e autênticos, além de que devem estar em formato eletrónico. Têm também que ser representativos da variedade da sua procedência e que permitam o estudo científico e não somente linguístico.

Através de Corpora é possível observar diversos aspetos (morfológicos, sintáticos, semânticos, discursivos, etc.), relevantes para pesquisa linguística. Podem-se ainda explicar a produtividade

[12] programa de análise de corpora que integra dezassete instrumentos de análise, entre os quais, *Wordlist, Concordance, Keywords, Splitter, Text Converter, Dual Text Aligner e Viewer.*

[13] conjunto de ferramentas de análise de corpus distribuído livremente para concordância e análise de texto

bem como o emprego de palavras, expressões e formas gramaticais. Permitindo a descoberta de factos novos na língua, não percetíveis através da intuição, permitindo assim uma descrição objetiva da língua [Sar04].

2.2.2 Tipos de Corpora

Existem muitos Corpora disponíveis tanto livremente como mediante pagamento. A depender da linha de pesquisa e da resolução para o problema em questão, é possível a partir de um Corpus gerar Subcorpus de estudo ou mesmo utilizar o Corpus todo como uma unidade. E outras vezes torna-se necessário a compilação de um Corpus próprio para a questão em causa, como é o caso da presente Dissertação. Sendo que o objetivo de cada Corpus é o que determina o tipo de material a selecionar, que varia mediante a modalidade de produção, modalidade da língua (standard ou dialetal), se o Corpus é Escrito ou Oral, géneros textuais, contexto, etc [MC16]. Assim sendo, e em caso de compilação própria de um Corpus, é necessário seguir três etapas:

- Projeto do Corpus.

  A primeira fase para a compilação de um Corpus é a seleção dos textos pertinentes e que sejam relevantes para a pesquisa. Sendo necessário definir o tipo de Corpus que queremos compilar. É importante avaliar também o tamanho e à sua composição em termos dos textos existentes e os géneros dos quais pertençam.

- Compilação, Manipulação e limpeza do Corpus, Nomeação dos arquivos de textos, e Pedidos de Permissão de Uso.

  A compilação consiste no armazenamento em arquivos predeterminados de todos os textos selecionados. Sendo a fonte de aquisição dos textos variada, podendo ser através da Web ou mesmo textos impressos, nesse caso, é necessário digitalizá-los [AdBA06]. Esta etapa é precedida da fase de desenho da estrutura do Corpus, pois é esta que irá estabelecer os géneros textuais a serem representados bem como sua quantidade. Em um Corpus Monitor a fase de desenho muitas vezes pode não existir ou ser menos significativa [MC16]. Quando utilizada a web para a captura dos textos, a tecnologia oferece algumas opções que facilitem o processo, sendo possível efetuar a busca com o uso de um mecanismos de busca como o Google para pesquisar toda a Web. E ainda utilizar ferramentas que efetuam o processamento dos resultados das buscas feitas. Em seguida as páginas coletadas são organizadas num computador local, e a construção automática de Corpus é feita com ajuda de *offline browsers* ou com auxílio de ferramentas de apoio para a compilação. Posteriormente é feita a Nomeação de arquivos e geração de cabeçalhos. Nesta etapa faz-se a conversão dos textos de "PDF" para "*txt*" e atribui-se um nome. Nomeação esta, deve seguir um padrão a fim de facilitar a recuperação dos textos no futuro. É importante também proteger a identidade dos textos participantes de um do Corpus. Sendo necessário seguir algumas regras legais para a compilação de um Corpus no que diz respeito os direitos de uso do material junto a autores e editores detentores do copyright do texto ou obter consentimento de indivíduos cujos direitos de privacidade devem ser reconhecidos [CdMGJ+08], [MC16].

- Anotação.

  Quanto ao processo de anotação, existem dois níveis para a representação das informações em um Corpus: a anotação estrutural e a linguística. A primeira consiste na marcação dos

dados externos [14] isto é, dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, a tipologia textual e informação sobre a distribuição do Corpus e dos dados internos dos textos, que compreende a anotação de segmentação do texto cru, e esta subdivide-se em:

a) Marcação da estrutura geral, isto é, capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos.

b) Marcação da estrutura de sub-parágrafos, que envolve elementos que são de interesse linguístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e ênfase tipográficas do tipo negrito, itálico, sublinhado, entre outros. Essas informações irão facilitar a recuperação posterior do texto bem como a geração de Subcorpus. Apoiando-se por exemplo em informações como título, autor, época, género, e vários outros dados.

A segunda (anotação linguística) dá-se em qualquer nível como por exemplo nos níveis morfossintático, sintático, semântico, discursivo, e feita de três formas:

a) Manualmente, anotação feita por linguistas;

b) Automaticamente, utilizando ferramentas de PLN;

c) Semi-automática, consiste na correção manual da saída de outras ferramentas, sendo por isso considerada no trabalho apresentado por SM Aluísio, GM de Barcellos Almeida (2006) a mais eficiente e mais rápida gerando dados mais corretos em relação a anotação pela primeira vez [AdBA06].

Na anotação de Corpus para a criação de aplicações de PLN é muito utilizado o *XCES11 (XCES* é um padrão baseado em XML para codificar Corpus de texto, que são usados por linguistas e pesquisadores de linguagem natural. *XCES* é altamente baseado no anterior Eagles Corpus Encoding Standard (CES), mas usa o *XML* como a linguagem de marcação. Suportando vários tipos de Corpus).

Os Corpora podem ser classificados mediante a forma como os mesmos apresentam-se. Diante da pesquisa bibliográfica feita ilustramos na secção seguinte os tipos de Corpora:

2.2.2.1 Corpora Orais, Escritos e Mistos

Quanto a modalidade da língua, os Corpora podem ser Orais, Escritos e Mistos. Corpus Orais *(speech)*, são aqueles em que são selecionados somente amostras da língua falada, recolhida através de gravações, já os Corpus Escritos *(spoken)*, são formados por textos escritos. Por último temos os Corpus Mistos, que agrupam as duas modalidades anteriores, mas inclui maioritariamente amostras da língua escrita, pois está é mais barata e menos trabalhosa no processo de obtenção das amostras.

[14] Entende-se como dados externos a documentação do Corpus na forma de um cabeçalho que inclui os metadados textuais (ou dados estruturados sobre dados).

2.2.2.2 Corpora Monolingues, Bilingues e Multilingues

Os Corpora podem classificar-se também quanto ao Número de Língua. Apresentando-se como:

- Monolingues, Corpus compostos por textos recolhidos em uma só língua e têm como objetivo dar respostas a pesquisas de dita língua.
- Bilingues, formado por textos de duas línguas, sendo uma a tradução e a outra a original.
- Multilingues, os textos neste tipo de Corpus são recolhidos em várias línguas, sendo que não necessariamente são traduções umas das outras e não compartem os mesmos critérios [FDSB14] [Pas17].

Os dois últimos Corpora podem ser Comparáveis (se agrega textos originalmente escritos numa respetiva língua e outros textos similares traduzidos para dita língua a partir de várias línguas diferentes, e contribui para o estudo das diferenças entre as convenções textuais a todos os níveis linguísticos e culturais, sendo por isso muito utilizado em sistemas de tradução automática) e Paralelos (quando alinha o texto fonte de uma língua específica com a sua respetiva tradução em uma ou mais línguas, ou seja, são corpora formado por um conjunto de textos em uma determinada língua de origem e um outro conjunto composto por versões traduzidas destes mesmos textos para um outro idioma).

Dentro dos Corpus Paralelos encontramos os Corpus Alinhados, que são aqueles em que os textos são organizados paralelamente por parágrafos ou frases, de forma a facilitar a extração de semelhanças e equivalências quanto a tradução, pois existem muitas vezes elementos que têm traduções mútuas; facilitando dessa forma a exploração destes Corpora [FDSB14].

Baseado no estudo feito por *Kenny, Dorothy* (1998) [Ken98], em [Day05] é defendido como um dos principais objetivos de um Corpus Paralelo o de possibilitar a identificação de um determinado padrão nas línguas de origem e suas respetivas traduções simultaneamente. E técnicas de alinhamento são normalmente requeridas para possibilitar estabelecer ligações entre os textos de origem e de chegada. Sendo que esses mesmos Corpora servem como uma ferramenta para avaliar o processo de tradução de um determinado par de idiomas, além de apresentarem-se indispensáveis quanto a investigação do relacionamento entre padrões lexicais e sintáticos, nas respetivas línguas. Desempenhando deste modo, um papel importante no treinamento de tradutores, no desenvolvimento de sistemas de tradução automática e na lexicografia bilingue [Day05].

2.2.2.3 Corpora Grandes, Equilibrados, Piramidais e Léxicos

Podemos ainda classificar os Corpora quanto à Quantidade, Proporção e Distribuição em que os Grandes serão aqueles que não têm um limite fixado de palavras. Equilibrados, aqueles que capturam e guardam a mesma quantidade de diferentes tipos de textos. Os Corpus Piramidais, contêm uma distribuição textual nivelada de formas a que em cada nível conste uma variedade temática, mas com muitos textos para cada uma. E os Léxicos, procuram recolher pequenos fragmentos de texto e de longitude constante em cada documento [Pas17].

2.2.2.4 Corpora Gerais, Especializados, Genéricos e Canónicos

De acordo a especificidades de textos, os Corpora que buscam refletir a língua ou variedade linguística da forma mais equilibrada são denominados Corpora Gerais, onde quanto mais géneros e materiais, tipos de textos e modalidades da língua contenham, melhor são. E por esse motivo tendem a ser suficientemente amplos a fim de abranger todas as variedades relevantes de uma língua e seu vocabulário, de modo a que se possa usar como base para a elaboração de Subcorpus ou ainda para a criação de gramáticas, dicionários, tesauros, e incluem uma grande variedade de textos produzidos em situações comunicativas quotidianas [Pas17].

Corpus Especializados são aqueles que na sua criação focalizam em textos que possam servir para a descrição de um tipo particular de língua [Pas17].

Já os Genéricos têm como objetivo caracterizar o género em que se insere o estudo pretendido e a recolha textual é feita para aquele único género, a partir do qual se possa fazer uma caracterização frente a outros. Existem ainda os chamados Corpus Canónico, por ser formado por todos os textos da obra completa de um autor [Pas17].

2.2.2.5 Corpora Cronológicos, Históricos e Sincrónicos

Existe um conjunto de Corpora de acordo ao período temporal e encontram-se subdivididos em:

- Corpus Periódicos ou Cronológicos, agrupam textos de determinados períodos de tempo ou de épocas concretas.
- Corpus Diacrónicos ou Históricos, inclui textos pertencentes a diferentes etapas temporais sucessivas, com o objetivo de observar e estudar a evolução linguística [Pas17].
- Corpus Sincrónicos, este tipo de Corpus, tem como finalidade permitir o estudo de uma língua ou variedade linguística de forma estática, isto é, dedica-se ao estudo da língua no presente e sem preocupar-se com a evolução da mesma, sem descurar ás mudanças rápidas que possam ocorrer no momento do estudo [FDSB14].

2.2.2.6 Corpora de Referência e de Estudo

De acordo a finalidade encontramos Corpus de Referência e de Estudo. Os de Referência, visam representar a língua na sua variedade standard, cujo objetivo é dar conta da diversidade da língua. Os Corpora de referência geralmente são Corpora Mistos e incluem o maior número de géneros textuais e registos possíveis, e servem de termo de comparação para o Corpus de Estudo. Em geral, deve ser três a cinco vezes maior que o Corpus de Estudo. Em função do objetivo deste tipo de Corpus, seu planeamento e compilação obedece a princípios de equilíbrio entre géneros textuais, tornando-os estáticos e fazendo deles Corpora Fechados. Ex: Corpus de Referência do Português Contemporâneo (CRPC) [MC16].

GC Pastor (2017), define Corpus de Referência, como sendo aquele que não incorpora documen-

tos inteiros, mas sim fragmentos, uma vez que o interesse não é no texto, mas sim no estado da língua nele representada [MC16]. Um Corpus de Estudo é aquele em que se baseia a pesquisa a ser desenvolvida.

2.2.2.7 Corpora Abertos e Fechados

Alguns Corpora possuem um número finito de palavras, que é estabelecido previamente durante a recompilação do mesmo e uma vez alcançado esse número, o Corpus é finalizado. Esses são os chamados Corpus Fechado (ou Estáticos). E ainda na mesma classificação quanto aos limites estabelecidos, encontramos outro tipo denominado Corpus Aberto (Monitor), que são Corpus dinâmicos e em constante crescimento, os textos são agregados de forma periódica e mediante a capacidade de armazenamento, e quando esta não mais permita os textos mais antigos são substituídos por novos. Sendo por isso muito utilizados em estudos diacrónicos, pois permitem observar a evolução linguística e tendências de uso, mudanças de significado entre outras questões [MC16]. Ex: *Bank of English*.

Na secção seguinte é apresentado mais uma tipologia de Corpora, baseada no processo de aquisição, armazenamento e disposição dos mesmos.

2.2.2.8 Corpora Simples, Verticais e Anotados

Um Corpus Simples é aquele em que os textos são guardados sem formato e sem acrescer nenhum tipo de informação, códigos ou anotações. E por isso oferecem possibilidades muito limitadas para estudos linguísticos. No Corpus Vertical, as palavras de um texto são dispostas em colunas e ordenadas segundo critérios de frequência ou alfabéticos. Considerando as palavras de forma isolada e sem contexto. Corpus Codificado ou Anotado é aquele que agrega informação adicional aos textos, isto é, são Corpora cujos documentos tenham sido etiquetados linguística ou meta textualmente de forma manual ou automática [MC16]. Aumentando as possibilidades de exploração devido os dados agregados. Mediante o tratamento do Corpus, este pode ser: Corpus Analisado Morfologicamente, Corpus Patentizado e os Analisados. Ainda nesta categoria encontramos também Corpus Não Anotado, que é aquele que não tenha sido etiquetado de nenhuma forma, disposto somente em formato de texto ou ASCII e apresenta um elevado grau de simplicidade.

Em [MC16], é possível observar ainda outros tipos de Corpora nomeados mediante sua característica.

Com o apresentado, podemos observar que existe um vasto conjunto de Corpora, e sua tipologia varia em função dos objetivos e das características dos mesmos.

Este objetivo, previamente definido, é que irá determinar o tipo de material a selecionar, que pode variar de acordo a: Modalidade de produção ou linguística, géneros textuais, recorte sobre a língua, perfil dos autores/informantes, registo formal ou informal dos textos e das gravações, etc.

2.2.3 Ferramentas para compilação, processamento e análise de Corpora

Existe na literatura um vasto conjunto de ferramentas desenvolvidas e empregues na LC. Desta feita apresentamos abaixo algumas ferramentas utilizadas no desenvolvimento ou utilização de Corpus.

2.2.3.1 *W*ordSmith Tools

O programa *WordSmith Tools* é um artefacto eletrónico utilizado para realizar análise linguística de Corpora. Desenvolvido por *Mike Scott* em 1996, *WordSmith Tools* não faz a seleção do material a ser trabalhado, cabe ao pesquisador ter o material organizado para a análise que deseja realizar. Sendo por isso todo o processo metodológico e documental da responsabilidade do pesquisador, incluindo a delimitação do material e a organização temática dos Subcorpora. Por este facto quando utilizado *WordSmith Tools* é necessário em um primeiro momento da pesquisa, utilizar um outro programa para seleção dos documentos [Cra16].

O referido artefacto consiste de três ferramentas principais: *Concord* (concordanciador), WordList (gerador de lista de palavras) e *Keywords* (extrator de palavras-chave).

A ferramenta *WordList* permite, através da contagem de palavras, a criação de listas de palavras. Por sua vez as listas mostram a frequência com que cada palavra foi encontrada nos textos e em quantos textos foi encontrada. É possível ainda organizar as listas por ordem alfabética ou por ordem de frequência. Com a *KeyWords*, obtém-se listas de palavras-chave de textos através da comparação da lista de frequências de ocorrências de palavras dos textos em estudo com a lista de frequências de ocorrências de palavras em um conjunto de textos de referência. Já a ferramenta *Concord* permite a produção de concordâncias em que todas as ocorrências de uma palavra ou de um conjunto de palavras são listadas. Demonstrando-se possível produzir concordâncias diretamente a partir das ferramentas *WordList* e *KeyWords*, selecionando palavras das listagens geradas pelas duas ferramentas [Sil08].

Fazendo parte também outros utilitários como *Collocates* (apresenta os colocados da palavra de busca), *Clusters* (relaciona os agrupamentos em que aparece a palavra de busca); *Aligner* (alinha dois textos, dentre outros) [Cra16].

2.2.3.2 *U*nitex

Unitex é um sistema *open-source* de processamento de Corpus, desenvolvido inicialmente na Universidade *Paris-Est Marne-la-Vallée* (França), e baseado na teoria dos autómatos. Sendo muito utilizado em aplicações de PLN. A ferramenta *Unitex* foi desenvolvida em Java, em conjunto com os programas em C permite que seja altamente portável, sem perdas significativas de desempenho durante o processamento do Corpus. Esta ferramenta oferece recursos que são agrupados em quatro funcionalidades principais [CJVS+15]:

1. Autómatos: usados para criação de dicionários,buscas e transformações nos textos.

2. Dicionários de Apoio: utilizados, entre outras tarefas, para flexionar palavras automaticamente (alguns dos quais utilizados no Portal Min@s).

3. listagem de Frequências.

4. Concordanciador: baseado em dicionários e autómatos.

5. Gerenciador de gramáticas.

O Unitex oferece algumas limitações: Efetua buscas baseadas em lemas e classes gramaticais, porém sem a eliminação de ambiguidade. E apenas um texto ou Corpus pode ser aberto de cada vez [Pau15].

Entre os recursos linguísticos oferecidos estão dicionários e tabelas do léxico-gramática (matrizes binárias nas quais as linhas são ocupadas por entradas do léxico e nas colunas são explicitadas as propriedades sintático-semânticas de cada entrada lexical).

2.2.3.3 *Philologic*

Philologic é um conjunto de ferramentas para processamento de Corpus, desenvolvida pelo projeto *ARTFL (American and French Research on the Treasury of the French Language)* na universidade de Chicago. A ferramenta suporta anotações que são usadas em buscas por critérios bibliográficos, tais como: título, autor e data de publicação. O *Philologic* dispõe de uma interface Web que facilita sua utilização e a criação de Subcorpora. Requer a instalação de um servidor web e software adicionais em um ambiente Linux, requisitos, tornam a instalação complexa e de difícil execução para muitos usuários [Sil08] e [Alu].

A ferramenta Web ou o conjunto de ferramentas, tornam-no capaz de atender a diversos usuários simultaneamente. O *Philologic* contém algumas funcionalidades que podem ser agrupadas em três grandes grupos:

1. Concordâncias.

2. Frequências e Colocações.

3. Gerenciamento de Subcorpus.

A ferramenta oferece ainda recursos para Corpus multimodais. Textos obedecem ao padrão *Text Encoding Initiative Lite* (TEI Lite), mas podem ser personalizados até um certo limite. Empregando também um recurso para normalização ortográfica e utilizado em Corpus históricos ou em Corpus com erros de grafia através da ferramenta *AGREP* [15] permitindo de igual modo que as concordâncias sejam refinadas por parâmetros bibliográficos, fornecidos pelo cabeçalho *TEI* em cada texto. Sendo de difícil instalação por requerer um servidor Web e possuir diversas dependências [CJVS+15].

[15] Disponível em: https://www.rdocumentation.org/packages/base/versions/3.4.3/topics/agrep.

2.2.3.4 *Datumbox*

Datumbox é uma plataforma de aprendizagem de máquinas, desenvolvida e mantida por *Vasilis Vryniotis*, que se concentra em PLN. A plataforma *Datumbox* possui uma variedade de funções acessíveis através da *API REST*, incluindo análise de sentimentos, análise de sentimento de *Twitter*, deteção de idioma, deteção Comercial e educacional, extração de palavras-chave, similaridade documental [FMMG+16].

A parte central do projeto consiste em cerca de 30000 linhas de código. O código está licenciado sob a Licença Apache, Versão 2.0, facilitando a clonagem do repositório para testes. O *Datumbox Machine Learning Framework* é uma estrutura de código aberto escrita em Java que permite o rápido desenvolvimento das aplicações de Aprendizado de Máquinas e Estatísticas. O foco principal da estrutura é incluir uma grande quantidade de algoritmos de aprendizado de máquinas e testes estatísticos e ser capaz de lidar com conjuntos de dados de grande porte [Fra15]. Quanto aos detalhes técnicos, a *Datumbox* API é um serviço Web que permite usar suas ferramentas no site, software ou aplicativo móvel do utilizador. Fornecendo acesso a todas as funções suportadas pelo serviço. É disponibilizada na página *Datumbox* todas as informações necessárias para usar a API, amostras de código totalmente implementadas e a mais recente documentação da API. Sendo a versão atual da API a 1.0v. E para fazer uso da mesma o usuário deve registrar-se, criando uma conta *Datumbox* a fim de obter a chave API correspondente ao perfil criado [Fra15]. A API permite que o usuário crie aplicativos que façam uso de técnicas de Análise de Texto e Processamento de Linguagem Natural, como Ferramentas de Marketing Online, Ferramentas de *SEO*, Serviços de Monitoramento de Mídia Social, Filtros Anti-*Spam* e outros aplicativos de Classificação de Texto. As funções da API atualmente suportadas são: Análise do Sentimento, Análise do Sentimento do *Twitter*, Análise de Subjetividade, Classificação de Tópicos, Deteção de *Spam*, Deteção de Conteúdo para Adultos, Avaliação de Legibilidade, Deteção de Idioma, Deteção Comercial, Deteção Educacional, Deteção de Género, Extração de Texto e Similaridade de Documentos [Fra15].

2.2.4 Corpora criados e disponíveis na Web para fins de pesquisa

Nos últimos anos o desenvolvimento de Corpus, tem conhecido um vasto e importante aumento, com auxílio a técnicas das novas tecnologias. Facilitando desta forma o armazenamento, de modo a que se possa aceder aos mais variados dados e informações; permitindo extrair informações pertinentes, facilitando a recuperação seletiva da informação e consultas no geral.

Dessa forma tem sido crucial a criação de interfaces que facilitem o processo de busca e recuperação de informações nos respetivos Corpus. Assim, pesquisadores e instituições vêm empregado esforços neste sentido. Uma busca bibliográfica, permitiu-nos trazer para o presente documento alguns corpus e plataformas existentes. Sendo o *Brown University Standard Corpus of Present-Day American English (Brown Corpus)*, considerado o primeiro Corpus linguístico eletrónico, compilado na década de 1960 por *Henry Kucera* e *W. Nelson Francis* na *Brown University , Providence , Rhode Island* e lançado em 1964. O Brown Corpus contém 500 amostras de texto em língua inglesa, totalizando cerca de um milhão de palavras, compiladas a partir de obras publicadas nos Estados Unidos em 1961. A cerca de 35 anos atrás, para informatizar um conjunto de textos eram observadas inúmeras dificuldades, e dessa forma o Brown Corpus na época, detinha quantidades invejável de dados. Impulsionando assim o desenvolvimento da

área conhecida atualmente por LC.

Desde então, vários pesquisadores têm vindo a apresentar um conjunto de abordagens no que concerne à propostas para a construção de Corpora de propósito geral ou específico, tanto para o inglês quanto para outras línguas. Como podemos observar no trabalho desenvolvido por *S Sharoff*, 2006 [Sha06]; *Baroni et al.*, 2009 [BBFZ09]; *M Davies*, 2014 [Dav14]; e *koeva*, 2016 [KST+16]. Bem como a implementação e o desenvolvimento de métodos para a construção de Corpus específicos nas mais variadas áreas de domínio e posteriormente ferramentas de lexicografia, e Corpus paralelos também foram propostos. Existindo ainda estudos (em menor percentagem) centrados na construção de Corpus de variedades ou dialetos específicos. Como evidenciado em [CB17] e [Dwy14].

Em [dAG+17] foi construído um Corpus de textos, o qual é formado por artigos científicos na língua portuguesa e de domínio educacional, é demonstrado também as estatísticas que o compõem. O trabalho desenvolvido por LHG de Aguiar, MVC *Guelpeli* - PLURAIS-Revista Multidisciplinar, 2017 [dAG+17], tem como objetivo obter um Corpus que torne possíveis diversas pesquisas na área de Processamento de Linguagem Natural e especificamente na área de Sumarização Automática, para possibilitar análise da performance de sumarizadores na Língua Portuguesa. Para a pesquisa, utilizaram-se repositórios académicos tendo sido adotado um critério, com base no qual os arquivos deveriam possuir resumos e palavras-chave formados por seus autores. uma vez que o objetivo é que o Corpus permita realizar futuras pesquisas na área de Sumarização Automática. Para o desenvolvimento do referido Corpus, os autores adotaram a metodologia referida por Aluísio e Almeida (2006) [AdBA06], que divide em três estágios a compilação de um Corpus próprio:

- Projeto: inclui a seleção, captura, manipulação dos textos, nomeação dos arquivos e, por fim, a anotação. A manipulação dos textos, segue a metodologia proposta por *Guelpeli* (2012) [Gue12]. A seleção dos textos, foi efetuada observando os critérios: gratuitidade; possibilidade de reprodução dos arquivos originais; classificação das bases para o domínio e subáreas escolhidas para a pesquisa; resumo do texto original, denominado sumário de referência, elaborado pelo autor. Tendo sido selecionados artigos científicos (em sua maioria, do repositório *Scientific Electronic Libary Online* (SCIELO[16])), por possuírem resumo e palavras-chave, posteriormente utilizados para testes com sumarizadores automáticos de texto. As categorias em que não houve preenchimento total pelos artigos do repositório da *SCIELO*, foram preenchidas por artigos encontrados no repositório Buscador Coruja[17] e pelo Google. A compilação do Corpus foi feita no período de dezembro de 2015 a março de 2016 e para tal foi utilizada a tabela de áreas de conhecimento da grande área da Educação, para a definição de dez subáreas que compõem as categorias do Corpus (disponível no site da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES[18]), nas categorias de Educação Especial; Educação Permanente; Educação Pré-escolar; Ensino-aprendizagem; Filosofia da Educação; História da Educação; Política Educacional; Psicologia Educacional; Sociologia da Educação e Tecnologia Educacional.

- Limpeza e Formatação do Corpus para o Processamento Computacional: Seguindo o pro-

[16]Disponível em: http://www.scielo.br/

[17]Disponível em: https://buscadorcoruja.com/

[18]Disponível em: http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/tabela-de-áreas-do-conhecimento-avaliação.

cedimento realizado por [Gue12], retiraram tudo que não fazia parte do texto (gráficos, tabelas, figuras e números de páginas) e os arquivos fontes foram convertidos do formato em PDF para o *TXT*, compatível para o processamento. Para cada uma das 10 categorias que compõem o domínio, criou-se uma subpasta e dentro de cada categoria, várias subpastas. Sendo o Corpus divido em cinco pastas, e nomeadas mediante sua característica. As informações estatísticas do Corpus foram coletadas através do *Software FineCount* 2.6 free[19] No total o Corpus é formado por 2.999.646 palavras, cuja distribuição se dá nos 500 artigos selecionados.

- Anotação: Retirada dos métodos que, segundo Aloísio e Almeida (2006) [AdBA06], "são dados estruturados sobre dados, isto é, dados bibliográficos comuns, de catalogação, tais como tamanho do arquivo, tipo da autoria, a tipologia textual e informação sobre a distribuição do Corpus, neste caso específico, a retirada dos dados estatísticos do Corpus gerado" [dAG+17]. Na pasta criada na segunda etapa, denominada "Anotação" são encontradas as referências externas deste Corpus.

Dessa forma o referido Corpus tem como finalidade contribuir com a continuidade dos estudos da eficiência de sumarizadores, com possibilidade de utilização das palavras-chave no processo de sumarização dos textos, além de viabilizar estudos da influência destas palavras nos conteúdos dos textos do domínio [dAG+17].

As interfaces Web são úteis e importantes para facilitar a busca e recuperação de informações pelos utilizadores. Desta feita, têm surgido várias iniciativas quanto ao desenvolvimento de plataformas online que permitam o acesso às ferramentas e serviços para o PLN. Neste sentido fez-se uma busca de algumas plataformas existentes (algumas com objetivos distintos), como espelhado a seguir:

2.2.4.1 Portal Min@s

Como apresentado por Candido Junior, et al (2015) [CJVS+15], a ferramenta Portal Min@as, foi criada para apoiar tarefas de processamento de Corpus de propósito geral; projeto vinculado ao Programa de Pós-Graduação em Estudos Linguísticos (POSLIN), da FALE/Universidade Federal de Mina Gerais (UFMG), disponibilizando Bancos de Dados e Corpora de textos, empregando uma vasta gama de funcionalidades de consulta [CJVS+15] e [Por].

O Portal Min@s é utilizado para armazenar diversos Corpora, sem definir um limite máximo para o tamanho do Corpus. Projetado em linguagem Java para ambiente Web. No processo de importação de Corpus, os *tokens*[20] de cada texto são armazenados em um Banco de Dados *PostgreSQL5*[21] Sendo que as tarefas de *tokenização*, segmentação sentencial e anotação morfossintática são realizadas recorrendo as funcionalidades da biblioteca *OpenNLP*[22]. É usado ainda no Portal Min@s um conjunto de ferramentas com diferentes propósitos e mediante as tarefas e

[19] Ferramenta gratuita que fornece análises e estatísticas de documentos.

[20] *Token* é um segmento de texto ou símbolo que pode ser manipulado por um analisador sintático, que fornece um significado ao texto; ou seja, é um conjunto de caracteres (de um alfabeto, por exemplo) com um significado coletivo.

[21] Disponível em: https://www.postgresql.org/. Acesso: janeiro de 2018.

[22] A biblioteca *OpenNLP* permite aplicar técnicas de processamento de linguagem natural usando apren-

tecnologias que o mesmo emprega, como é o caso da ferramenta *GNU Aspell*[23] usada na correção ortográfica automática. Para a tarefa de otimização das buscas, o Portal Min@s faz uso da estrutura de indexação madura e eficiente oferecida pelos Sistemas Gerenciadores de Banco de Dados [CJVS+15].

São disponibilizadas ao utilizador diversas funcionalidades após a importação de um Corpus. O principal módulo consiste em um concordanciador disponibilizado em duas versões: monolingue e multilingue.

Quanto a anotação do Corpus, estas podem ser efetuadas como fragmentos de lexemas[24], informações morfossintáticas (pronome seguido de flexões do verbo ser) e informações no cabeçalho dos textos (como autor e editora).

Grafia atualizada para textos históricos, também são disponibilizadas, bem como diversos outros módulos são disponibilizados.

Utilizadores do Portal Min@s contam ainda com um conjunto de recursos gerenciais extras para importar e controlar o acesso aos Corpus armazenados na ferramenta. Cabe ao módulo de importação de textos responsabilizar-se por aplicar uma série de preprocessadores (tokenização, lematização, alinhamento). A administração do Corpus, Subcorpus e textos cabe aos módulos de gerenciamento, através deste módulo são definidos os termos de uso e acesso, sendo os Corpora classificados como públicos ou privados.

O submódulo encarregue para gerenciamento de etiquetas permite administrar diferentes categorias, incluindo etiquetas do Corpus, dos textos (Ex: autor), de n gramas (Ex: funções sintáticas), de secções do texto (notas de rodapé) e de formatação. Existindo também um módulo para gerenciamento de usuários que permite o cadastro de usuários para acessar o portal. Sendo definido cinco perfis de usuários, com diferentes níveis de acesso: administrador, coordenador, colaborador, usuário regular e visitante[CJVS+15].

2.2.4.2 *ClueWeb09*

No ano de 2000, o Centro de Recuperação de Informações Inteligentes *(CIIR)* da Universidade de *Massachusetts* em *Amherst*, e o Instituto de Tecnologias Linguísticas *(LTI)* da Universidade *Carnegie Mellon*, deram início a um Projeto denominado *Lemur*[25], fazendo parte do projeto um conjunto de ferramentas de software e motores de busca, mecanismos de busca em larga escala e conjunto de dados [lem]. Dentre os componentes do Projeto *Lemur*, interessa-nos destacar o *ClueWeb09* que é uma coleção de páginas da Web em 10 idiomas coletados pela

dizado de máquina. Entre as técnicas de processamento de linguagem natural suportadas estão: *tokenização*, segmentação de sentenças, extração de entidades nomeadas (nomes de pessoas, locais, etc), *parsing*, etc.

[23] O *GNU Aspell*, geralmente chamado de *Aspell*, é um verificador de correlações de software livre projetado para substituir o *Ispell* . É o verificador ortográfico padrão do sistema operacional GNU

[24] Unidade mínima distintiva do sistema semântico de uma língua que reúne todas as flexões de uma mesma palavra

[25] Projeto que desenvolve motores de busca, barras de ferramentas do navegador, ferramentas de análise de texto e recursos de dados que suportam pesquisa e desenvolvimento de software de recuperação de informações e mineração de texto.

Universidade *Carnegie Mellon* em Janeiro e fevereiro de 2009, criado para apoiar a pesquisa sobre recuperação de informações e tecnologias relacionadas de linguagem humana. O *DataSet ClueWeb09*, está disponível livremente para fins de pesquisa, sendo um recurso valioso e usado por várias faixas da conferência *TREC* [26] [lem].

O conjunto de dados é formado por 1.040.809.705 páginas web, em 10 idiomas; 5 TB, comprimido. (25 TB, descompactado); e o conjunto completo de dados corresponde a:
URLs únicos: 4,780,950,903 (325 GB sem compressão, 105 GB compactados).
Total *Outlinks*: 7,944,351,835 (71 GB sem compressão, 24 GB comprimidos). O *DataSet ClueWeb09* está disponível em vários serviços de "*Cloud Computer*" (por exemplo, *Open Cloud, the Pittsburgh Supercomputer Center)*. Também pode ser acessado usando interfaces de pesquisa fornecidas pelo Projeto *Lemur*. A Universidade *Carnegie Mellon* encarrega-se da distribuição do conjunto de dados *ClueWeb09*, apenas para fins de pesquisa. O *DataSet* pode ser obtido a partir da mesma Universidade, assinando um contrato de licença de dados e pagando uma taxa (em torno de 650 dólares americanos) que cobre o custo da distribuição do *DataSet* [lem].

Faz também parte do Projeto *Lemur*, entre outros, o *ClueWeb12*. Sendo que o *DataSet* consiste em 733.019.372 páginas da Web em inglês, coletadas entre 10 de fevereiro de 2012 e 10 de maio de 2012. Sua distribuição começou em janeiro de 2013 [lem]. O *DataSet ClueWeb*, é usado por muitos pesquisadores.

2.2.4.3 Leipzig Corpora Collection *(LCC)*

LCC é um projeto do Grupo de Processamento de Linguagem Natural do Instituto de Ciência da Computação da Universidade de Leipzig ou Lípsia[27].

O projeto do Corpus Leipzig teve seu início durante a década de 1990, derivado da grande dificuldade verificada naquela época quanto a disponibilidade de recursos para PLN em alemão livremente acessíveis. Desde então, as técnicas para processamento e apresentação de Corpus foram desenvolvidas e as mesmas não dependem das características de um determinado Idioma, pois foram recolhidos recursos de texto em diversas línguas, sendo agora possível fornecer acesso a dados e estatísticas sobre as línguas disponíveis, num formato e em tamanhos padrão. Além disso, o Corpora *Leipizig* fornece serviços linguísticos básicos gratuitos para quem tiver um uso para eles, sem ter que assinar acordos, pagando taxas de envio e afins [RQHB06].

Para o desenvolvimento da tecnologia da linguagem um dos requisitos cruciais é o acesso aberto aos recursos básicos de linguagem, especialmente para idiomas com poucos oradores e recursos escassos. Os Corpus presentes em Leipzig Corpora *Collection*, têm como finalidade fornecer uma base de dados para o desenvolvimento e teste de algoritmos (independentes da linguagem) para várias aplicações de PLN, principalmente para construir modelos de idiomas a partir de dados não rotulados[28].

[26] Série contínua de workshops com foco em uma lista de diferentes áreas de pesquisa de recuperação de informações.

[27] Leipzig é uma universidade da Alemanha, localizada na cidade de Leipzig (Lípsia), fundada em 1409, sendo a segunda universidade mais antiga da Alemanha a funcionar ininterruptamente.

[28] Dados não estruturados e de difícil processamento, pois não possuem uma formatação específica. Por exemplo: Mensagens de email, imagens, documentos de texto, mensagens em redes sociais.

O processo de construção *Leipzig Corpora Collection* obedeceu quatro etapas, sendo: coleta textual, pré-processamento, limpeza e, eventualmente, calcular ou fazer uma aproximação. O pré-processamento é feito removendo *tags*HTML dos textos coletados e separando o conteúdo que contém texto padrão. Em seguida, uma deteção de limite de sentença é realizada e fragmentos de frases mal formadas são removidos, além de frases em línguas estrangeiras e duplicadas [RQHB06]. Antes de trabalhar no Corpus a nível da sentença e reduzir Para tamanhos pré-definidos, é realizada uma limpeza adicional. Processo feito para garantir que realmente haja formação adequada; frases que, obviamente, não possuem linguagem padrão.

As frases fragmentadas ou desorganizadas, imagens e sons dos documentos originais não são inseridos de formas que não possam ser restaurados, o que garante que os textos podem ser distribuídos sem prejudicar a proteção de direitos aurorais, uma vez que frases únicas são muito curtas para serem consideradas como propriedade intelectual. Fazem parte do Corpora vários Corpus que foram coletadas da web e consistem em textos de jornal ou em páginas web coletadas aleatoriamente. Os tamanhos máximos dos Corpus oferecidos são restritos pela disponibilidade atual, em vez de serem arbitrariamente escolhidos. Nesta perspetiva a noção de Corpus está centrada em torno da sentença como a maior unidade. Isto é suficiente para uma grande variedade de aplicações em PLN estatística e lexicografia. Para cada idioma, é calculado um dicionário de formulário completo com informações de frequência para cada palavra. Além disso, a coleção fornece estatísticas de co-ocorrência: palavras que co-ocorrem significativamente com uma determinada palavra. Para o cálculo da significância é aplicada a função de verossimilhança ou função de probabilidade[29]. Dois tipos dos dados de co-ocorrência são pré-computados: as palavras que ocorrem juntas em frases e palavras encontradas como vizinhos imediatos (esquerda ou direita) [RQHB06]. Desta feita e conhecendo um parâmetro B do Corpus em questão, a probabilidade condicional de A é P(A|B), mas se o valor de A é conhecido, então pode-se realizar inferências sobre o valor do parâmetro B apoiando-se ao teorema de Bayes[30], segundo o qual:

$$P(Ai|B) = \frac{P(Ai)P(B|Ai)}{P(B)}$$

Dessa forma, a função de verossimilhança $L(b|A)$ é definida como:

$$L(b|A) = P(A|B = b)$$

Em caso de um número de amostras aleatórias independentes desde $x1, ..., axn$, o log-probabilidade conjunta será a soma de log-probabilidades individuais, e a derivada desta soma será uma soma de derivadas de cada indivíduo de log-verossimilhança[Wik17]:

$$\frac{\partial logL(\alpha, \beta|x1, ..., xn)}{\partial \beta} = \frac{\partial logL(\alpha, \beta|x1)}{\partial \beta} + ... + \frac{\partial logL(\alpha, \beta|xn)}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i.$$

[29]Função dos parâmetros de um modelo estatístico que permite inferir sobre o seu valor a partir de um conjunto de observações.

[30]Em teoria das probabilidades e estatística, o teorema de *Bayes* (alternativamente, a lei de *Bayes* ou a regra de *Bayes*) descreve a probabilidade de um evento, baseado em um conhecimento a priori que pode estar relacionado ao evento. O teorema mostra como alterar as probabilidades a priori tendo em vista novas evidências para obter probabilidades a posteriori

Para estudos de linguagem comparativos, os Corpus de tamanho padrão são ideal para medir e comparar sistematicamente parâmetros de Corpus não-lineares, como taxas de crescimento de vocabulário, distribuições em grande escala e outras características tipológicas [RQHB06].

A LCC recolhe continuamente textos de várias línguas e géneros para a criação de Corpus de texto e seu enriquecimento com diferentes dados adicionais. A coleção apresenta Corpus em diferentes idiomas usando o mesmo formato e fontes comparáveis. Todos os dados estão disponíveis como arquivos de texto simples e podem ser importados para um banco de dados *MySQL* usando o *script* de importação fornecido para uso científico por linguistas de Corpus quanto a aplicações como programas de extração de conhecimento.

Os Corpus são idênticos em formato e similar em tamanho e conteúdo. Eles contêm frases selecionadas aleatoriamente na linguagem do Corpus e estão disponíveis em tamanhos de 10.000 frases até 1 milhão de frases. As fontes são textos de jornal ou textos coletados aleatoriamente da web. Os textos são divididos em frases. Não foram emitidas orações e material de língua estrangeira. Como a informação de co-ocorrência de palavras é útil para muitas aplicações, esses dados são pré-computados e incluídos também. Para cada palavra, as palavras mais significativas que aparecem como vizinho esquerdo ou direito imediato ou que aparecem em qualquer lugar dentro da mesma frase são dadas. Os Corpus da referida coleção são coletados automaticamente de fontes públicas cuidadosamente selecionadas sem considerar em detalhes o conteúdo do texto contido. Não é responsável pelo conteúdo dos dados. Em particular, a visão e opiniões expressas em partes específicas dos dados permanecem exclusivamente com os autores [LCC18].

O uso possível do Corpora como recurso linguístico inclui: Lexicografia monolingue, comparação de diferentes idiomas em uma base estatística, parametrização dos modelos de linguagem, expansão de consultas com palavras estatisticamente similares, extração de termos significativos de documentos por comparação contra um Corpus de referência, seleção de conjuntos de palavras balanceadas para experiências, por exemplo no campo da psicolinguística, indução de sentido de palavras. Os Corpora da LCC fornecem um vasto recurso linguísticos e de tarefas PLN, contudo, e como é característica dos Corpus livres, em oposição aos recursos caros de alta qualidade, podem não cumprir todos os requisitos de qualidade e equilíbrio de texto e não podem fornecer metadados adicionados manualmente ou anotação em grande escala. Quanto a isso, sistemas de consulta de Corpus mais sofisticados tem vindo a ser desenvolvidos e para a sua aquisição é necessário obedecer a um conjunto de requisitos e o possivelmente estar sujeito ao pagamento de uma taxa monetária para sua aquisição.

2.2.4.4 *Google Cloud Platform*

A Google *Cloud Platform* é a plataforma de *cloud computing* oferecida pela *Google*. Funcionando na mesma infraestrutura que a empresa utiliza para outras plataformas como o *Google Searche* e *youtube*. A *Google Cloud Platform* foi criada para permitir a execução de uma série de serviços baseados na nuvem com alta performance, segurança e confiabilidade.

Considerada umas das ferramentas mais poderosas a ferramenta *Cloud Natural Language*, esta revela a estrutura e o significado do texto por meio de avançados modelos de aprendizado de

máquina pré-treinados e uma *API REST* fácil de usar.

Alguns dos recursos disponibilizados *Cloud Natural Language* incluem:

- Análise sintática: através deste recurso é possível extrair *tokens* e frases, identificar classes gramaticais e criar árvores de análise sintática para cada frase;
- Reconhecimento de entidades: permite a identificação de entidades e marcação por tipos, como por exemplo pessoas, organizações, locais, eventos, produtos e mídia;
- Análise de sentimento: possibilita o entendimento do sentimento geral expresso em um bloco de texto;
- Classificação de conteúdo: a classificação de documentos pode ser feita em mais de 700 categorias predefinidas;
- Vários idiomas: é possível através deste recurso analisar facilmente textos em vários idiomas, incluindo inglês, espanhol, japonês, chinês (simplificado e tradicional), francês, alemão, italiano, coreano e português;
- *API REST* integrada: a *API Natural Language* pode ser acessada por meio da *API REST*. Sendo disponibilizado um link para o envio do texto por upload na solicitação ou integrado ao *Google Cloud Storage*.

Algumas restrições e limitações são impostas nos serviços disponibilizados. Por exemplo o serviço só é gratuito na análise de sentimentos entre as 0 e 5 mil unidades por mês.

2.2.4.5 Arquivo.Pt

O Arquivo.pt é uma infraestrutura de investigação gratuita e pública que habilita a pesquisa e o acesso a páginas da web arquivadas desde 1996. Cujo foco é a preservação da informação publicada na Web para fins de investigação. Facilitando assim a recuperação da informação que já não se encontra disponível na web e fornece recursos de investigação nas áreas da História, Sociologia ou Linguística [Arq]. Projeto baseado no *Internet Archive*, teve seu início em 2008, com a pretensão de preservar conteúdo da Web de interesse para a comunidade portuguesa [VGC16].

Sua criação representa um marco histórico e uma aposta de Portugal nas gerações futuras. Cuja perspetiva é disponibilizar os seguintes serviços:

- Pesquisa histórica por termo: permitirá identificar páginas arquivadas ao longo dos anos que contenham determinadas palavras;
- Pesquisa histórica por endereço da web (URL): permitirá identificar várias páginas arquivadas ao longo dos anos referenciadas por um determinado URL;
- Coleções históricas de conteúdos web para fins de investigação: a web contém informação sobre os mais diversos assuntos sendo o reflexo dos nossos dias.

- Infraestrutura para processamento paralelo dos dados arquivados: irá permitir que investigadores, mesmo sem serem especialistas em sistemas informáticos distribuídos, executem os seus programas sobre os dados web arquivados usando várias máquinas do projeto em paralelo [Arq].

O Arquivo.pt funciona de forma semelhante aos outros motores de busca como o Google. Faz a recolha dos dados, através de batedores ou *Crawlers*, a partir de um conjunto inicial de endereços de sítios da Web (raízes), e armazena-o em disco. Extrai endereços para páginas a partir das ligações e insere os novos endereços descobertos para recolha. Terminada a recolha dá-se início ao processo de indexação onde, toda a informação recolhida da Web é processada para construir os índices que permitirão realizar pesquisas rápidas. Após criados os índices, são disponibilizados serviços de pesquisa e acesso à informação recolhida da Web. A principal diferença entre os motores de busca e os arquivos da web é que os arquivos têm a preocupação adicional de preservar a informação para mante-la acessível ao longo do tempo [Arq]. A disponibilização de coleções da web no Arquivo Pt vem permitir que investigadores possam processar informação das mais diversas áreas, que vão da Sociologia à Informática, localmente nos seus computadores sem terem de realizar recolhas da web; usando-o como fonte de informação para os seus estudos.

No arquivo.pt estão disponíveis os seguintes serviços e ferramentas:

- Pesquise no passado: A pesquisa no arquivo permite o acesso a páginas do passado (Acesso via *OpenSearch)*: interface de programação que permite pesquisar no arquivo através do protocolo *OpenSearch*. Os resultados das pesquisas são devolvidos em formato *XML (RSS* 2.0).
- Plataforma para criação de aplicações de processamento sobre a informação arquivada: Para facilitar o desenvolvimento de aplicações que necessitem de realizar análises de larga-escala *(Big Data Analytics)*.
- Sistema de pesquisa do Arquivo da Web Portuguesa desenvolvido com base no projeto *Archive-access*.
- Sistema de sugestões para correção de pesquisas textuais.
- Software para facilitar a utilização da coleção de teste na área de recuperação de informação sobre conteúdos web arquivados (por *Zeynep Pehlivan*).
- *Httrack2Arc* Ferramenta para converter recolhas feitas com o *Httrack* para ficheiros no formato *Arc*.
- Roteiro2Arc Ferramenta usada para converter para formato *ARC* os ficheiros no *CD-rom* do livro ”Novo Roteiro Prático da Internet” por José Magalhães [Arq].

Em setembro de 2013, todo o sistema de arquivo.pt desmoronou e a informação foi completamente perdida. A par dessa perda arquivo.pt teve também que renovar sua equipe. Esta situação causou vários problemas na reconstrução e manutenção do sistema de busca. Durante

o processo de recuperação do serviço, foi identificada uma severa falta de documentação técnica sobre a arquitetura do sistema de busca [VGC16]. Levando pesquisadores a desenvolverem documentos técnicos a respeito da arquitetura de software do sistema de pesquisa arquivo.pt.

Num estudo prévio feito por F Melo, D Bicho, D Gomes (2016) sobre.arquivo.pt, fez-se uma experiência para avaliar a qualidade de repetição e o desempenho do *Software Wayback* (aplicação Java *Open Source* lançada em setembro de 2005 pelo *Internet Archive*, que permite pesquisar e reproduzir páginas da Web arquivadas) [MBG16]. O estudo revelou que o *Software Wayback* do arquivo.pt estava desatualizado e outras alternativas como *OpenWayback* ou *PyWB* melhoraria significativamente a qualidade de repetição dos sites arquivados no mesmo [BMG17]. Com base no estudo feito, o arquivo.pt migrou do *Wayback* 1.2.1 para *PyWB* desenvolvida por *Ilya Kreymer* e utilizada para a reprodução das páginas arquivadas [BMG17]. Todo o software desenvolvido pelo Arquivo.pt está disponível como código-aberto livre [Arq].

2.2.4.6 *BulNC*

O Corpus Nacional Búlgaro *(BulNC)* foi criado por pesquisadores do Departamento de Linguística Computacional e do Departamento de Lexicologia e Lexicografia Búlgara do *Institute for Bulgarian Language* "Prof. L. *Andreychin*". Constituído de vários corpos eletrónicos individuais, desenvolvidos no período 2001-2009 para os propósitos dos dois departamentos; sendo constantemente ampliado com novos textos. É um corpus dinâmico e consiste aproximadamente 1.200.000.000 palavras distribuídas. Sendo uma parte monolingue (búlgara) e 47 corpora paralelos. A parte búlgara inclui cerca de 1,2 biliões de palavras em mais de 240 000 amostras de texto. Os materiais no Corpus refletem o estado da língua búlgara (principalmente na sua forma escrita) de meados do século XX (1945) até ao presente [KST+16].

O Corpus Nacional da Bulgária, através da sua plataforma *BulNC Search* [Bulb], permite uma série de aplicações em várias áreas linguísticas: na linguística computacional; em lexicografia; dentro de estudos teóricos de fenómenos linguísticos específicos; para observações das características dos domínios de linguagem individuais; para extrair sentenças exemplares para a educação em língua búlgara [Bula]. Algumas das aplicações mais específicas do Corpus estão listadas abaixo:

- Extração de Subcorpus específicos ou gerais seguindo critérios particulares (sujeito, autor, ano/ período de publicação, fonte, etc.), que podem ser utilizados como corpos de treinamento para uma série de aplicações - etiquetado gramatical e semântico, entre outros. Como para outros fins de pesquisa.
- Observações sobre a frequência de uso de palavras ou construções linguísticas, geração de listas de frequências, etc.
- Pesquisas no Corpus para casos de fenómenos linguísticos específicos, exemplos lexicográficos ou para fins educacionais na instrução de língua búlgara (disponível para uso na Internet) [Bula] e [KST+16].

O *BulNC* permite acessar a partes específicas do corpus disponíveis para download; permite também o acesso total ao mecanismo de pesquisa e consultas de intercalação, extração de

subcorpora e dicionários de frequência do *BulNC* ou sua subcorpora [Bula]. O seu acesso ao *BulNC* está sujeito aos seguintes termos e condições:

- O acesso é gratuito para fins não comerciais - pesquisa científica, educação, etc.
- O direito de acesso é pessoal e não pode ser ré-atribuído a outras pessoas.
- Os usuários do *BulNC* devem cumprir completamente com a Lei de Direitos Autorais.
- Ao usar dados do *BulNC* em publicações, dicionários, etc., é obrigatório citar o *BulNC* como fonte de dados. É obrigatório incluir também informações sobre o trabalho a partir do qual o exemplo foi extraído.
- Sempre que possível, fornecer detalhes bibliográficos de publicações em que o *BulNC* tenha sido citado, ao administrador (bulnc@dcl.bas.bg).
- O administrador pode encerrar o acesso ao *BulNC* sempre que necessário e sem aviso prévio.

Qualquer subcorpora lançada é distribuída para download sob uma Licença *Creative Commons Attribution-NonCommercial* 3.0[31].

Cada Corpus Paralelo consiste exclusivamente de textos que têm correspondência em búlgaro - seja o original ou a tradução. Ambos os textos podem ser traduções de uma terceira língua. Sua estrutura, formato de dados e descrição seguem de perto o modelo do *BulNC*. Os textos são fornecidos com metadados detalhados, extraídos automaticamente sempre que possível e elaborados manualmente, se necessário. O maior corpus paralelo dentro do *BulNC* é o corpus paralelo búlgaro-inglês que compreende 260,7 milhões de *tokens* para o inglês e 263,1 milhões de *tokens* para o búlgaro [Bula].

[31] licenças públicas que permitem a distribuição gratuita de uma obra protegidas por direitos autorais

Capítulo 3

Metodologia e Trabalho Realizado

Esta secção descreve a metodologia que levou-se a cabo para o desenvolvimento do trabalho proposto.

A compilação do Corpus Hultig-C teve seu início em Janeiro de 2017, e consiste em cerca de milhares de palavras em diferentes idiomas, coletadas com base nos textos brutos (de diferentes naturezas, de diferentes níveis linguísticos e de sofisticação) obtidos através de sites da Web e indexados com *Open Web Spider*.

Foi realizada uma pesquisa qualitativa de Corpora para fins de pesquisa e ferramentas que auxiliam no desenvolvimento de softwares para PLN. Deste modo uma revisão sistemática das ferramentas existentes e suas características, fez-se necessária.

Inicialmente procurou-se identificar as temáticas a abordar bem como os tópicos necessários a serem estudados.

A segunda fase foi composta pela condução da revisão bibliográfica e posterior seleção dos trabalhos a serem analisados. Desta forma foi possível fazer uma síntese do estudo realizado e propor novas abordagens a partir da síntese do mesmo.

Assim, foi adotada a metodologia proposta em [DPWW$^+$], que consiste numa Medida Informativa Assimétrica *Asymmetric InfoSimba Similarity of Term* (AIS), que associada a outras Medidas de Associação Assimétrica, permite determinar a implicância ou o vínculo entre dois textos através da relação Generalização-Especialização. Solução esta, não supervisionada e independente da língua [DPWW$^+$] e [DM].

Foi ainda proposta uma nova abordagem para extração de termos relevantes em um texto até *Trigram*.

Levando-se em conta o que foi observado elaborou-se este trabalho, tendo surgido no âmbito do projeto em curso para a compilação do Hultig-C levado a cabo pelo Centro de Tecnologia da Linguagem Humana e Bioinformática, do Departamento de Informática da UBI. Tendo surgido da necessidade do referido projeto tencionar dar suporte ao processamento automático da linguagem humana e providenciar recursos de alto nível para a investigação em Linguística Computacional e para o desenvolvimento de aplicações e tecnologias no campo de PLN. Cujo objetivo é auxiliar pesquisas Científicas, Académicas e Estudantis.

Para que os objetivos fossem alcançados fez-se necessário a utilização de algumas ferramentas e tecnologias como HTML5 (usado para a criação da Plataforma), *MySQL Workbench* (ferramenta de design de banco de dados visual), *NetBeans IDE* 8.1 e o *Open Web Spider* que é um rastreador

da Web e um mecanismo de pesquisa que permite o rastreamento da Web e integração ao banco de dados. Após a instalação e cumprindo com os requisitos necessários, podemos observar se tivemos êxitos quando aparece a informação espelhada na imagem da figura 3.1.

Dessa forma, acessando a URL do servidor local http://127.0.0.1:9999/ pudemos então dar início ao processo 3.2. As figuras 3.3, 3.4, 3.5 ilustram a a verificação de conexão do *Open Web Spider* ao servidor, a criação da tabela a ser usada pelo servidor, a URL de entrada para a construção do mecanismo de pesquisa e o consequente progresso do processo da indexação.

A semelhança dos demais *Crawlers* o objetivo é visitar sites, mas também criar um índice de entradas para funcionar como um mecanismo de pesquisa 3.7.

Após o rastreio do site cuja URL sirva de semente é então possível pesquisar palavras ou expressões existentes nos sites visitados 3.7. O *Open Web Spider* possibilita a utilização de dois gerenciadores de banco de dados *MySQL Server* ou *MongoDB*. Para este trabalho foi utilizado o *MySQL Server*, pela familiaridade com o mesmo. Desta feita foi utilizado o *Open Web Spider (js)* v0.3.0 com suporte ao *PostgreSQL*.

Um outro requisito para a instalação e utilização dos mesmo é o *Node.js*, que atua junto ao servidor, cuja função é interpretar o código *JavaScript* de maneira a manipular dezenas de milhares de conexões simultaneamente, numa única máquina física. O que permite criar um índice com centenas de milhares de páginas e permitir os usuários façam buscas mais rápidas 3.8.

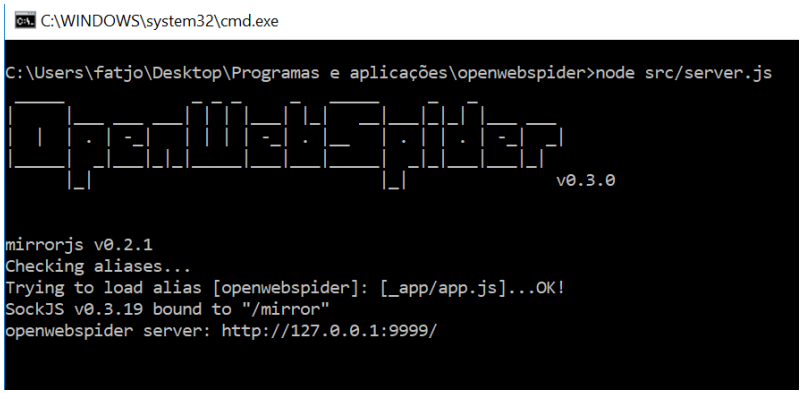

Figura 3.1: *Open Web Spider* Instalado

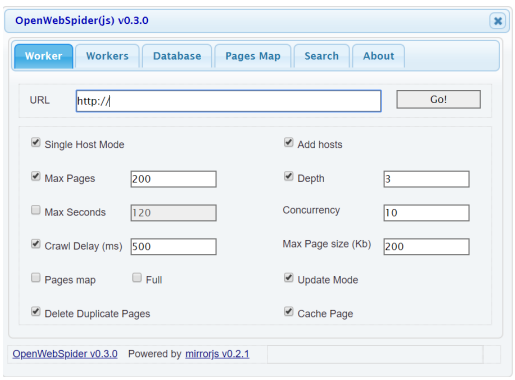

Figura 3.2: Tela inicial mostrada no navegador

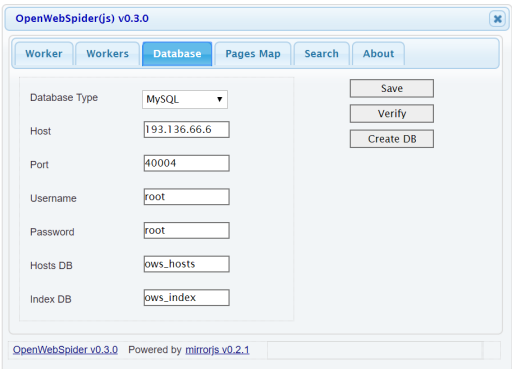

Figura 3.3: Conexão *Open Web Spider Servidor MySQL*

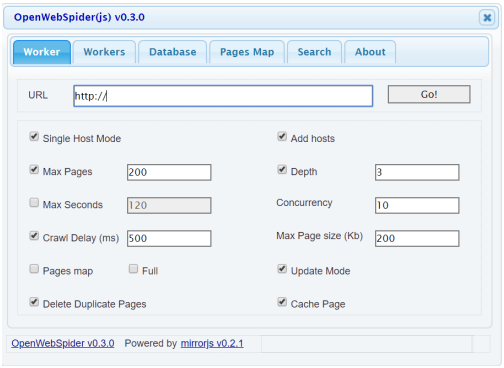

Figura 3.4: Resposta da conexão correta

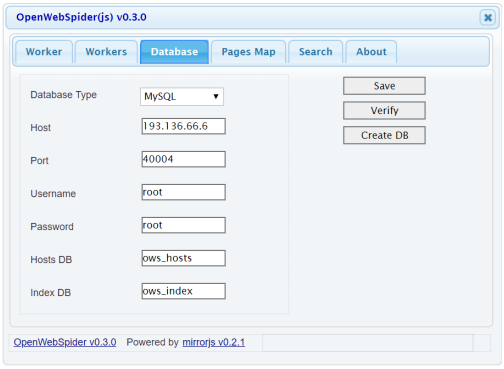

Figura 3.5: Criação da Tabela utilizada pelo Servidor

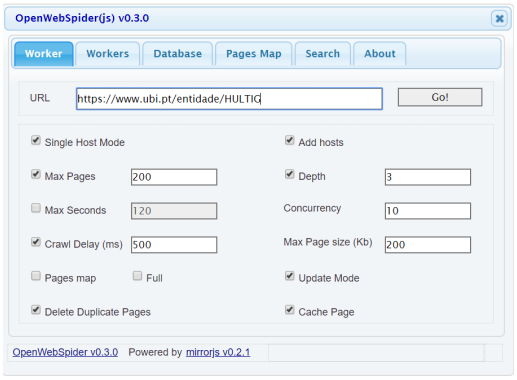

Figura 3.6: Progresso da indexação através da *URL*

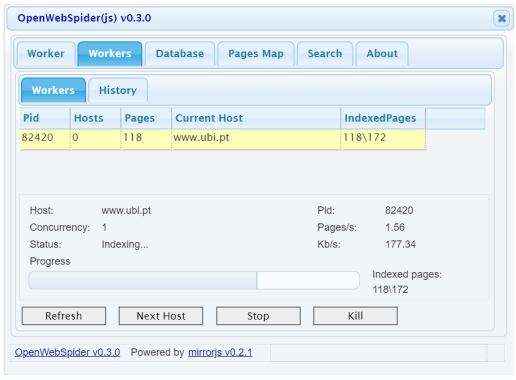

Figura 3.7: Progresso da indexação através da *URL*

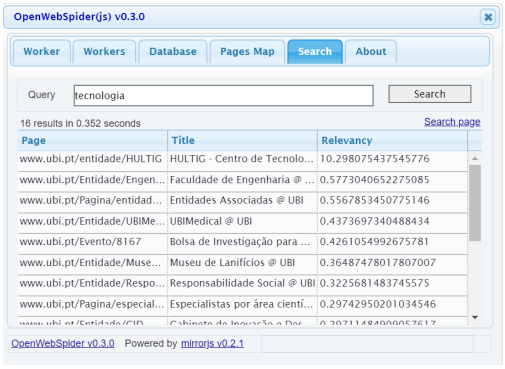

Figura 3.8: Resultados de uma pesquisa realizada

3.1 Projeto do Corpus

A primeira fase para a compilação de um Corpus é a seleção dos textos pertinentes e que sejam relevantes para a pesquisa. Sendo necessário definir o tipo de Corpus que queremos compilar. É importante avaliar também o tamanho e à sua composição em termos dos textos existentes e os géneros dos quais pertençam.

A compilação consiste no armazenamento em arquivos predeterminados de todos os textos selecionados. Sendo a fonte de aquisição dos textos variada, podendo ser através da Web ou mesmo textos impressos, nesse caso, será necessário digitalizá-los [AdBA06]. Esta etapa é precedida da fase de desenho da estrutura do Corpus, pois é esta que irá estabelecer os géneros textuais a serem representados bem como sua quantidade.

Para o Hultig-C os textos estão, numa primeira fase, sendo coletados através da Web, sem restrições de tamanhos, temáticas ou ainda linguísticas, observando os critérios de permissões de uso.

3.2 Abordagem adotada para a extração de termos específicos

Nossa abordagem basea-se na utilizada por [Pai13], desta forma é proposta uma nova Medida denominada AISs(T) que combinada de maneira significativa com diferentes Medidas de Associação Assimétricas *Asymmetric Association Measures* (AAM), visa proporcionar bons resultados no processo de seleção dos termos relevantes.

O Método proposto neste trabalho é Não Supervisionado e Independente da Língua, o que implica que um dos principais requisitos é que tenha a capacidade de extrair conhecimento tanto explícito como implícito de textos em linguagem natural e não estruturados. A depender da unidade textual básica a utilizar, os dois tipos de conhecimento podem ser extraídos [Pai13]. O aprendizado Não Supervisionado pretende extrair informações sem auxílio humano, não existindo por isso uma necessidade de supervisão; diferente do Supervisionado. Por um lado, analisar similaridades de palavras evidencia o conhecimento intrínseco sobre a língua [Pai13]. Entender o significado (semântica) de textos não é uma tarefa trivial, representando um dos principais problemas em tarefas de PLN. É necessário quantificar e decidir quão semelhantes ou não, são o significado de dois textos. Identificar diferentes tipos de semelhanças entre palavras representa um desafio importante em PLN e algumas abordagens têm sido úteis no que diz respeito ao cálculo do grau da similaridade entre termos ou palavras. Na abordagem estatística, uma palavra é representada por um vetor de coocorrência de palavra em que cada entrada corresponde a outra palavra no léxico. O valor de uma entrada especifica a frequência da ocorrência conjunta das duas palavras no Corpus, ou seja, a frequência com a qual as palavras co-ocorrem dentro de alguns relacionamentos particulares no texto. Dessa forma calcula-se o grau de similaridade entre um par de palavras através de alguma medida de similaridade ou distância que é aplicada ao par de vetores correspondente [Pai13].

Um mesmo conceito pode estar presente no texto por palavras e termos diferentes, dessa forma boas consultas de recuperação de texto livre são difíceis de formular. Assim, conceito de semelhança de palavras foi introduzido com base em tesauros. Um tesauro é um recurso lexicográfico que especifica relações semânticas entre palavras, listando para cada palavra as palavras relacionadas a mesma, como sinónimos (Palavras ou termos que possuem o mesmo ou significado

semelhante), hipónimos e hiperónimos (termos usados pela semântica, em estreita relação um com o outro [Pai13]. Um hipónimo é uma palavra cujo significado é hierarquicamente mais específico, do que outra. Exemplos: cenoura ou tomate são hipónimos de legume, Arquiteto ou Professor são hipónimos de profissão. Já um hiperónimo é uma palavra cujo significado é hierarquicamente mais abrangente, do que outra. Exemplos: legume é hiperónimo de cenoura ou tomate profissão é hiperónimo de Arquiteto ou Professor. Os tesauros auxiliam na seleção de palavras e termos apropriados permitindo o enriquecendo vocabular de um texto. E através destes é possível expandir consultas realizadas. A expansão de consulta é uma técnica utilizada, na qual uma consulta é expandida, a partir dos termos originais fornecidos pelo usuário, aos seus termos relacionados. Cujo objetivo é melhorar a qualidade da consulta [Pai13].

As palavras podem ser consideradas como unidades individuais tendo em conta suas relações nos documentos. Mas para efetuar análises de textos interessantes, é necessário ter em conta também as relações entre as palavras, hora examinando quais palavras tendem a suceder ou anteceder outras imediatamente ou que tendem a co-ocorrer dentro dos mesmos documentos. A partir da observação de com que frequência a palavra W é seguida pela palavra $W1$, é possível construir um modelo das relações existentes entre as mesmas. Na literatura diferentes tipos de relações de coocorrência foram examinados. Tanto para a computação de similaridade de palavras, bem como para outras aplicações. Essas relações são classificadas em dois tipos gerais [Pai13]:

- Relações gramaticais, referem à co-ocorrência de palavras dentro de relações sintáticas especificadas.
- Relações não-gramaticais, aquelas que referem-se à ocorrência conjunta de palavras dentro de uma certa distância (Janela) no texto. Este conceito é amplo e captura vários sub-tipos de relações de co-ocorrência como: *Ngrams*, co-ocorrência direcional e não-direcional em pequenas janelas e co-ocorrência em janelas maiores ou dentro de um documento.

Modelos que atribuem probabilidades a sequências de palavras são chamados de modelos de linguagem ou *LMs (Language Models)*.

O modelo mais simples que atribui probabilidades a sentenças e sequências de palavras é o *Ngram (Ngrama)*. Um *Ngram* é uma sequência de N palavras e que aparecem consecutivamente no texto, e por isso são muito usados nos modelos de modelagem de linguagem para sistemas de reconhecimento automático de fala, e também em outras tarefas de reconhecimento e desambiguação. Quando a sequência é de duas palavras, é denotado como *Bigram* ou *2-gram*. Ex: Por favor, da capital, vem cá...

Em caso da sequência ser composta por três palavras como resumo da notícia, dia para pagar, telefone para contacto, é denotado como *Trigram* ou *3-gram*. E assim por diante.

Através dos modelos de *Ngram* e da característica sequencial das palavras (2,3,ou mais) e combinado suas probabilidades através de um Corpus, é possível estimar a probabilidade de interpretações acústicas alternativas do enunciado, a fim de selecionar a mais provável interpretação [JM17] e [Pai13].

Neste trabalho consideramos sequências de palavras do tipo *Trigram*.

Pais (2013) [Pai13], defende que informação capturada por *Ngrams* é, em grande medida, apenas um reflexo indireto das relações lexicais, sintáticas e semânticas na linguagem. E isso ocorre porque a produção de sequências consecutivas de palavras é o resultado de estruturas linguísticas mais complexas, e neste sentido os modelos *Ngrams* demonstraram ter vantagens práticas por ser fácil formular modelos probabilísticos para eles, é possível extrai-los de um Corpus com baixo grau de dificuldade e, acima de tudo, provaram fornecer estimativas de probabilidade úteis para leituras alternativas da entrada.

As similaridades de palavras obtidas por dados de n-grama podem refletir uma mistura de semelhanças sintáticas, semânticas e contextuais. Tais semelhanças são adequadas para melhorar um modelo de linguagem *Ngram*, uma vez que os mesmos abarcam estes tipos de relacionamento.

Uma co-ocorrência de palavras dentro de uma janela relativamente grande no texto sugere que ambas as palavras estão relacionadas ao tópico geral discutido no texto. Essa hipótese geralmente será válida para coocorrências frequentes, isto é, para pares de palavras que frequentemente co-ocorrem no mesmo texto. Um caso especial para esse tipo de relacionamento é coocorrência em todo o documento, o que corresponde a um tamanho máximo de janela. Normalmente, apenas a co-ocorrência de palavras de conteúdo é considerada, uma vez que essas palavras carregam a maioria das informações semânticas [Pai13].

3.2.1 Similaridade Assimétrica de Termos

A medida de similaridade de texto é um problema comum na recuperação de informações, TM, classificação de texto/ *clustering*, deteção de plágio, etc. Na literatura a abordagem mais popular é a baseada em um modelo de esquema de frequências de palavras, que faz uso de um vetor de frequência de palavra para representar um documento. Função cosseno, Produto escalar e Função de proporção, entre outras, são medidas de similaridade regulares de vetor, utilizadas como medidas de similaridade simétrica. Neste trabalho como mencionado, apresentamos um modelo de similaridade assimétrica, desenvolvendo para o contexto em causa (extração de termos relevantes) uma medida assimétrica: AIS_S(T), derivada da medida desenvolvida em [Pai13].

A similaridade apresenta-se como um conceito fundamental na representação de conhecimento vago e raciocínio aproximado, e seu estudo tem vindo a despertar interesse de muitos pesquisadores. A definição de semelhança nem sempre é objetiva e por isso é difícil ou quase impossível definir o conceito [MAM05].

Para [MAM05] e baseado-se no apresentado por [Goo72b] e [Goo72a] sugeriu que dois objetos a e b são mais parecidos com c e d se a importância cumulativa das propriedades compartilhadas por a e b é maior que as propriedades compartilhadas por c e d. Considerando importância como um assunto altamente volátil, variando de acordo ao contexto e interesse. Uma abordagem comum é tratar a similaridade como uma fuzzificação [1] de uma relação de equivalência, generalizando a reflexividade, a simetria e a transitividade.

[1] Fuzzificação, Etapa na qual os valores numéricos são transformados em graus de pertinência para um valor linguístico

Essa abordagem defende que, a similaridade está relacionada a uma métrica de distância e por isso, possui muitas propriedades matemáticas úteis. Mas o autor argumenta que esta não é uma definição apropriada de similaridade para a Mineração de Texto sintático. E assim o axioma da transitividade nem sempre é correto. Desta forma, usar diferentes sentidos de uma palavra pode obviamente violar a transitividade. Por exemplo: Banana é semelhante à *orange* (ambas frutas), por sua vez Orange[2].

A similaridade está intimamente relacionada ao conceito de tipicidade, o que pode levar à assimetria na relação de similaridade. Dizer que o peixinho dourado é semelhante ao animal de estimação implica que a "distância" do peixinho dourado ao animal de estimação é pequena para que se possa substituir a palavra peixinho por animal de estimação em uma frase sem perder muito significado. Vejamos as seguintes frases:

1. Eu mantenho um peixinho dourado no lago do meu jardim.
2. Mantenho um bichinho de estimação no meu lago no jardim.

Já a substituição inversa pode ser considerada imprópria. Vejamos:

1. Eu gosto de acariciar meu animal de estimação.
2. Gosto de acariciar meu peixe dourado.

Nas duas primeiras frases podemos observar que a substituição de um termo por outro não causa perca de sentido na frase. Já nas frases seguintes é óbvio que a substituição de "animal de estimação" por "peixe dourado" é pouco apropriada, pois parece estranho "acariciar meu peixe dourado". Em função de exemplos como os anteriores, o autor defende que na relação de similaridade para Mineração de Texto não procura por transitividade ou simetria. Considerando os argumentos acima informais (até certo ponto) e injustos, pois mistura diferentes significados (*Orange* como empresa de telecomunicações e *orange* como fruta) ou ainda superclasses (animal de estimação) com subclasses mais específicas (peixinho dourado). No entanto um sistema estatístico de Mineração de Texto não leva em conta categorias semânticas ou sentidos de palavras, e tem pouco a dizer além dos símbolos que representam as palavras. Assim é preferível considerar a similaridade como o grau em que uma palavra pode ser substituída por outra em uma dada sentença. O que leva a assimetria em muitos casos [MAM05].

Um recente estudo surgiu quanto à Medidas de Associação Assimétrica *(Asymmetric Association Measures - AAM)*, estudo inspirado no facto de que dentro da mente humana, a associação entre duas palavras ou conceitos nem sempre é simétrica. Para pares como fruta e maçã, pode-se concordar que existe uma forte associação mútua entre as duas palavras. Quando pensando em frutas, não é exagero pensar em maçã também, e vice-versa [Pai13]. Ainda de acordo a analogia feita outros pares também são considerados, no entanto, muitos deles não exibem esse tipo de forte associação em ambas as direções. Por exemplo: Um par composto pelas palavras (fruta e anona). Quando ouvimos a palavra "anona" fazemos uma ligação a "fruta", pois a palavra

[2] A *Orange* S.A. (anteriormente denominada *France Telecom*] é a principal empresa de telecomunicações da França, e a 105ª no ranking mundial

”anona” está fortemente associada ao conceito ”fruta”. Por outro lado se dissermos a alguém que diga o nome de uma ”fruta”, é pouco provável que diga ”anona”.

Um exemplo apresentado por [MES07] e evidenciado em [Pai13] demonstra que existe uma tendência para uma forte associação direta de um termo específico como o adenocarcinoma ao termo mais geral de câncer, enquanto a associação do cancro ao adenocarcinoma é fraca. De acordo a teoria defendida por [MES07], alguns membros da mesma categoria são mais centrais do que outros, tornando-os mais prototípicos da categoria a que pertencem. Por exemplo, o cancro seria mais central que o adenocarcinoma. Mas a principal motivação para a relação de associação está na noção de termos específicos e gerais. Torna-se claro que existe uma forte tendência para uma forte associação direta de um termo específico para o termo mais geral, mas a associação inversa é mais fraca. Nessa perspetiva trabalhos têm surgido, propondo o uso de medidas de similaridade assimétrica [Pai13]. E neste trabalho é também abordado um conjunto de Medidas de Associação Assimétrica (AAM), que possibilitam a extração de termos relevantes, com base no princípio de termos específicos poderem ser associados a termos mais gerais, perspetivando contribuir para melhorias em processos de aquisição de relações semânticas entre palavras.

3.2.1.1 Medidas de Associação Assimétricas *(Asymmetric Association Measures - AAM)*

As medidas de associação são modelos matemáticos que interpretam frequências de coocorrências de termos em um de texto. Para qualquer par de termos, uma pontuação de associação é calculada em uma escala contínua, que indica a quantidade de associação (estatística) entre os dois termos [Dia10].

A definição de medidas é crucial para a correspondência de padrões. Medidas baseadas em padrões podem incorporar a assimetria a partir de relações semânticas. Instanciar e enviar para um mecanismo de busca um número de padrões preenchidos apenas com um possível candidato pode garantir a extração de relações de hiperonímia/ hiponímia (relação hierárquica de inclusão semântica entre duas unidades lexicais, partindo do específico (hipónimo) para o geral (hiperónimo), observando que o primeiro, para além de conservar as propriedades semânticas impostas pelo segundo, possui os seus próprios traços diferenciadores), meronímia/ holonímia (Relação de hierarquia semântica entre duas unidades lexicais; em que uma denota um todo (holónimo) sem impor obrigatoriamente as suas propriedades semânticas à outra, considerada sua parte (merónimo)), que aproveitem ao máximo os padrões assimétricos, caso existam. Apesar do forte suporte intuitivo para a existência generalizada de associação dirigida os estudos de colocação ainda dependem de dados simétricos. E também medidas baseadas em padrões são sensíveis à polissemia da palavra e à ambiguidade do padrão. Técnicas essas dependentes da linguagem que são difíceis de replicar para diferentes idiomas [Pai13].

Em 1975, foi publicado por *Meyer* [M$^+$75], um experimento que buscava avaliar e medir o tempo de resposta de um indivíduo quando confrontado com duas tarefas específicas: (1) Classificar sucessões de letras em palavras e não palavras e (2) pronunciando uma sequência de caracteres. Esse experimento demonstrou que que em ambos os casos a resposta a uma palavra (por exemplo, manteiga) é mais rápida quando precedida por uma palavra associada (como por exemplo, pão) do que quando por uma não associada (por exemplo, enfermeira). *Church* 1990 [CH90], de

acordo ao apresentado por [M+75] defende que as relações semânticas dispersas ou separadas podem ser extraídas quando se conta frequências de co-ocorrência de palavras em janelas de palavras de tamanho fixo, ou n-gramas. Evidencia que o tamanho da janela é o parâmetro que permite observar escalas diferentes, de modo que tamanhos de janela menores identifiquem colocações e outras relações que ocupam intervalos menores, e tamanhos de janela maiores destacam conceitos semânticos e outras relações que se aplicam a escalas maiores. Mas, esclarece que de forma alguma, as medidas de coocorrência podem identificar com exatidão duas palavras relacionadas semanticamente. Neste sentido, as medidas de associação são usadas principalmente no contexto de extração de unidades de palavras múltiplas e para critérios de seleção de palavras [Dia10].

Assim como em [Pai13], a abordagem considerada neste trabalho é inserida no contexto de metodologias não supervisionadas e independentes da língua, propondo Medidas de Associação Assimétricas, avaliadas no contexto de similaridade assimétrica de termos em textos, descritas a seguir:

1. Probabilidade Condicional dada por :

$$P(T_x|T_y) = \frac{P(T_x,T_y)}{P(T_y)} \qquad P(T_y|T_x) = \frac{P(T_y,T_x)}{P(T_x)}. \tag{3.1}$$

2. Valor Adicionado :

$$AV(T_x \parallel T_y) = P(T_x|T_y) - P(T_x). \tag{3.2}$$

3. Braun-Blanket :

$$BB(T_x \parallel T_y) = \frac{f(T_x,T_y)}{f(T_x,T_y) + f(\bar{T}_x,T_y)}. \tag{3.3}$$

4. Factor de Certeza :

$$CF(T_x \parallel T_y) = \frac{P(T_x|T_y) - P(T_x)}{1 - P(T_x)}. \tag{3.4}$$

5. Convicção :

$$Co(T_x \parallel T_y) = \frac{P(T_x) * P(\bar{T}_y)}{P(x,\bar{T}_y)}. \tag{3.5}$$

6. Índice de Gini :

$$GI(T_x \parallel T_y) = P(T_y) * P(x|y)^2 + P(\bar{T}_x|T_y)^2 - P(T_x)^2 + P(\bar{T}_y) * (P(T_x|\bar{T}_y)^2) + P(\bar{T}_x|\bar{T}_y)^2) - P(\bar{T}_x)^2. \tag{3.6}$$

7. Medida J :

$$JM(T_x \parallel T_y) = P(T_x,T_y) * log\frac{P(T_x|T_y)}{P(T_x)} + P(\bar{T}_x,T_y) * log\frac{P(\bar{T}_x|T_y)}{P(\bar{T}_x)}. \tag{3.7}$$

8. *Laplace* (LP) :

$$Lp(T_x \parallel T_y) = \frac{N * P(T_x, T_y) + 1)}{N * P(T_y)^{+}2}. \quad (3.8)$$

3.2.1.2 Atribuição de similaridade em palavras assimétricas

Nos últimos anos, a hipótese distributiva forneceu a base para a teoria da generalização baseada em similaridade na aprendizagem de línguas, ideia baseada no princípio de que as crianças podem descobrir como usar palavras que raramente encontraram antes, generalizando sobre seu uso a partir de distribuições de palavras semelhantes. A hipótese distributiva sugere que quanto mais semanticamente semelhantes duas palavras forem, mais semelhante será sua distribuição [Dia10].

Em [Pai13] é exposto, com base no apresentado por [Dia10], que não é justificável do ponto de vista linguístico assumir que todas as dimensões de um modelo de espaço vetorial (modelo algébrico para representar palavras como vetores de características de contexto) sejam ortogonais entre si. Como cada dimensão normalmente corresponde a uma palavra de contexto, isso equivale à suposição de que duas palavras denotam significados díspares. Existindo uma falha aparente no modelo, ao considerar adequadamente os contextos que são semelhantes em significado ou sinónimos.

Muitas medidas de atribuição de similaridade de palavras tem sido propostas, mas todas elas baseiam-se no pressuposto de que duas palavras são semanticamente relacionadas se compartilham contextos comuns. O que na prática implica que duas palavras para serem declaradas semanticamente relacionadas, estas devem conter tantas palavras comuns quanto possível. Mas, devido à escassez de dados, geralmente é difícil encontrar contextos sobrepostos, mesmo em grandes Corpora. Observando ainda que as línguas naturais são particularmente ambíguas, o que pode implicar que um dado contexto de palavras incorpore muitos significados.

Com o objetivo de alavancar tais questões foi introduzido em [DSC06] e desenvolvido em [Pai13] a medida de similaridade informativa Similaridade InfoSimba (IS) dentro do contexto da aprendizagem não supervisionada de recursos léxico-semânticos. Desta feita a medida de IS é também empregue no presente contexto, em que a IS visa medir as correlações entre todos os pares de termos em vetores de contexto de termos, em vez de depender apenas de sua correspondência exata, como na medida de semelhança do cosseno 3.9.

$$Cos(X,Y) = \frac{\sum_{i=1}^{N} X_i * Y_i}{\sqrt{\sum_{i=1,}^{N} X_i^2} * \sqrt{\sum_{i=1}^{N} X_i^2}}. \quad (3.9)$$

Além do exposto, a IS garante a captura de similaridade entre pares de termos, quando eles não compartilham contextos, devido à dispersão de dados, por exemplo; apesar de terem contextos semelhantes. Na equação 3.10, é definida a IS em que $S(\ ,\ )$ é qualquer medida de similaridade simétrica e cada T_xy corresponde ao atributo do termo em i^th (que ocorre na posição i em uma

sequência) do vetor Ti, e n é o comprimento do vetor Ti:

$$IS(Tx,Ty) = \frac{\sum_{i=1}^{N}\sum_{i=1}^{N} Tx_i * Ty_i * S(Tx_i,Ty_i)}{\begin{pmatrix}\sum_{i=1}^{N}\sum_{i=1}^{N} Tx_i * Tx_l * S(Tx_i,Tx_l)+\\ \sum_{i=1}^{N}\sum_{i=1}^{N} Ty_i * Ty_i * S(Ty_i,Ty_l)-\\ \sum_{i=1}^{N}\sum_{i=1}^{N} Tx_i * Ty_i * S(Tx_i,Ty_i)\end{pmatrix}}. \quad (3.10)$$

Em pesquisas de atribuição de similaridade de palavras assimétrica, as direções das co-ocorrências são notadas e exploradas, não existindo um estudo em profundidade nem um relato teórico deste fenómeno. Os esforços envidados são direcionados para o desenvolvimento de medidas de similaridade assimétricas *distribucional*, como a divergência de *Kullback-Leibler* [KL51] definida na equação 3.11 [Pai13]:

$$KL(x \parallel y) = \sum_{\langle z,r\rangle \in A} logP(z|x) * \frac{logP(z|x)}{logP(z|y)}. \quad (3.11)$$

Onde $A = \{\langle z,r\rangle \,|\exists(x,z,r) \wedge \langle z,r\rangle \,|\exists(y,z,r)\}$.

Embora existam muitas medidas de similaridade assimétrica, elas evidenciam problemas que reduzem sua utilidade. Outro aspeto é que, as medidas de associação assimétrica só podem avaliar a relação generalidade/ especificidade entre palavras ou termos que são conhecidas por estarem em uma relação semântica. Geralmente capturam a direção da associação entre termos com base em contextos de documentos e levam em conta apenas uma proximidade semântica entre os mesmos. Por exemplo, é altamente provável que a Apple seja mais genérica que *iPad*, que não pode ser assimilado a uma relação de hiperonímia/ hiponímia, meronímia/ holonímia. Ainda, as medidas de atribuição de similaridade em palavras assimétricas levam em conta contextos comuns para avaliar o grau de similaridade assimétrica entre dois termos.

A fim de impulsionar questões semelhantes, na medida $AIS_S(T)$, a ideia é que um termo T_x é semanticamente relacionado à outro (T_y) e T_x é mais geral que T_y, se T_x e T_y compartilham tantos termos relevantes quanto possível e cada termo do contexto de T_y é provável que seja mais geral do que a maioria dos termos do contexto em que o termo T_y estiver inserido.

$AIS(T)$ definido na equação 3.12, em que $AS(\parallel)$ é qualquer Medida de Similaridade Assimétrica aplicada aos termos (T e Ti), como visto na medida IS ilustrada na equação 3.10, onde $S(\ ,\ )$ corresponde a qualquer medida de similaridade simétrica. Sua versão simplificada ($AIS_S(T)$) é apresentada na equação 3.13.

$$AIS(T) = \frac{\sum_{i=1}^{N} AS(Tx_i \parallel Ty_l)}{\begin{pmatrix}\sum_{i=1}^{N} AS(Tx_i \parallel Tx_l)+\\ \sum_{i=1}^{N} AS(Ty_i \parallel Ty_l)-\\ \sum_{i=1}^{N} AS(Tx_i \parallel Ty_l)\end{pmatrix}}. \quad (3.12)$$

Apresentada na forma simplificada como:

$$AIS_s(T) = \sum_{i=1}^{N} AS(Tx_i \parallel Ty_l) \Leftrightarrow Tx_i \neq Ty_l. \quad (3.13)$$

Como exemplo apresentamos um cálculo amostral da $AIS(T)$ simplificada, vide equação 3.13 para um texto **X** qualquer: Onde, fazendo uso da Medida de Similaridade Assimétrica demonstrada na equação 3.1, temos que:

$$\begin{aligned} AS(Tx_i \parallel Ty_l) \Leftrightarrow Tx_i \neq Ty_l &= P(Tx_i|Ty_l) \\ &= \frac{P(Tx_i), P(Ty_l)}{P(Ty_l)} \\ &= \frac{P(Tx_i) \cap P(Ty_l)}{P(Ty_l)} \\ &= \frac{P(Tx_i) * P(Ty_l)}{P(Ty_l)}. \end{aligned} \quad (3.14)$$

Assim, o $AS(Tx_i)$ é dado por:

$$AS(Tx_i) = \frac{P(Tx_i) * P(Ty_i)}{P(Ty_i)} + \frac{P(Tx_i) * P(Ty_j)}{P(Ty_j)} + \frac{P(Tx_i) * P(Ty_l)}{P(Ty_l)}, \ldots, \frac{P(Tx_i) * P(Ty_n)}{P(Ty_n)} \quad (3.15)$$

Calculando o $AS(T)$ para todos os termos *(Unigram, Bigram e Trigram)*, estaríamos então em condições de demonstrar o *Ranking*, que neste trabalho é ordenado de forma descendente, e de calcular o $AIS_S(T)$ (3.13), bem como a média e o desvio padrão. Da mesma forma se procede para as demais *AAM*.

Capítulo 4

Desenvolvimento Experimental e Resultados Obtidos

Como forma de validação da abordagem apresentada, foram selecionados alguns textos aleatoriamente e independente da língua. Textos esses que retratam também temas aleatórios, e a partir dos mesmos, os testes ocorreram da seguinte forma:

4.1 Medidas de Associação Assimétrica validadas

Em função dos textos selecionados, foram avaliadas as 8 medidas propostas, a fim de poder observar os resultados apresentados por cada uma. Dentre as quais somente foi possível validar 5 (cinco), visto que as restantes 3 (três) resultam em valores nulos (zero). Assim, dado o texto 1:

Presidente Português elogia o Presidente de Angola, por este ter recebido o Ministro da Defesa Português, mesmo não constando da agenda oficial do Ministério da Defesa. Da agenda divulgada constava a reunião com o Ministro da Defesa de Angola.

Onde constam 39 palavras, num total de 114 termos, dos quais correspondem aos *Unigram* = 39, de 25 tipos; *Bigram* = 38, de 32 tipos; *Trigram* = 37, de 35 tipos. Após a seleção dos *Ngram*, é então calculada a probabilidade individual de cada um dos termos do texto (P. termo), como ilustrado na tabela 4.1 onde constam também as respetivas ocorrências.E através destes dados é então possível dar sequência aos cálculos subsequentes, cujo resultado deverá proporcionar um intervalo de termos candidatos a relevantes. Assim, para o conjunto de entrada (termos que compõem o documento) obtêm-se:

Sendo T um termo em um documento D, $n(T)$ corresponde ao número total de termos T em D, $n(D)$, o conjunto de termos que formam o documento D, temos que a probabilidade de $P(T)$ é igual a:

$$P(T) = \frac{n_T}{n(D)} \quad (4.1)$$

Assim, temos que:

D = { Presidente, Português, elogia, o, Presidente, de, Angola, por, este, ter, recebido, o, Ministro, da, Defesa, Português, mesmo, não, constando, da, agenda, oficial, do, Ministério, da, Defesa, da agenda, divulgada, constava, a, reunião, com, o, Ministro, da, Defesa, de,

Angola }

Fazendo uso da AAM demonstrada na equação 3.1, são formados conjuntos de termos em função do valor da Média e do Desvio Padrão (ver tabela 4.2). fazendo o mesmo para as restantes AAM.

Substituindo na equação 4.1:

$$P(presidente) = \frac{2}{39} = 0.05128;$$
$$P(portugus) = \frac{2}{39} = 0.05128;$$
$$P(o) = \frac{3}{39} = 0.07692;$$

E assim sucessivamente (resultados apresentados nas tabelas 4.1, 4.4 e 4.7.

Desta feita o $AS(presidente)$ para a AAM da Probabilidade Condicional (PC)é dado por: $AS(presidente) = P(Presidente/Portugus) + P(Presidente/de) + P(Presidente/Angola) + P(Presidente/Ministro) + P(Presidente/agenda) + P(Presidente/elogia) + P(Presidente/por) + P(Presidente/este) + P(Presidente/ter) + P(Presidente/recebido) + P(Presidente/mesmo) + P(Presidente/no) + P(Presidente/constando) + P(Presidente/oficial) + P(Presidente/do) + P(Presidente/Ministrio) + P(Presidente/divulgada) + P(Presidente/constava) + P(Presidente/a) + P(Presidente/reunio) + P(Presidente/com) + P(Presidente/o) + P(Presidente/Defesa) + P(Presidente/da).$

$AS(presidente) = P(0.05128/0.05128) + P(0.05128/0.05128) + P(0.05128/0.05128) + P(0.05128/0.05128) + P(0.05128/0.05128) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.02564) + P(0.05128/0.07692) + P(0.05128/0.07692) + P(0.05128/0.12820) = 1.23076.$

Conforme a equação 3.15.

Da mesma forma para as restantes AAM. A tabela 4.1 ilustra os resultados obtidos para os termos *Unigram*.

Tabela 4.1: AS (T) para Unigram do texto 1

Types	Freq	P. termo	AS(T)-PC	AS(T)-BB	AS(T)-Co	AS(T)-GI	AS(T)-LP
presidente	2	0.05128	1.23076	2.00092	1.12220	0.94622	21.37857
português	2	0.05128	1.23076	2.00092	1.12220	0.94622	21.37857
elogia	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
o	3	0.07692	1.84615	2.55524	1.64366	0.91761	28.70476
de	2	0.05128	1.23076	2.00092	1.12220	0.94622	21.37857
angola	2	0.05128	1.23076	2.00092	1.12220	0.94622	21.37857
por	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
este	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
ter	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
recebido	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
ministro	2	0.05128	1.23076	2.00092	1.12220	0.94622	21.37857
da	5	0.12820	3.07692	3.67376	2.61700	0.85746	43.4
defesa	3	0.07692	1.84615	2.55524	1.64366	0.91761	28.70476
mesmo	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
não	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
constando	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
agenda	2	0.05128	1.23076	2.00092	1.12220	0.94622	21.37857
oficial	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
do	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
ministério	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
divulgada	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
constava	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
a	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
reunião	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
com	1	0.02564	0.61538	1.45053	0.57503	0.97371	14.08571
AIS (T)			24.00000	43.99844	21.83806	23.94952	454.45238
Ranking			3.07692	3.67376	2.61700	0.97371	43.4
			1.84615	2.55524	1.64366	0.94622	28.70476
			1.23076	2.00092	1.12220	0.91761	21.37857
			0.61538	1.45053	0.57503	0.85746	14.08571
Média			0.96000	1.75993	0.87352	0.95798	18.17809
Desvio Padrão			0.57937	0.52183	0.49081	0.02694	6.89111

Assim, definimos os intervalos de interesse para todas as medidas, a fim de extrair os candidatos a termos relevantes, conforme ilustra a tabela 4.2:

Tabela 4.2: Granularidade dos conjuntos para Unigram do texto 1

Métricas/AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	Presidente Português elogia o de Angola por este ter recebido Ministro da Defesa mesmo não constando agenda oficial do Ministério divulgada constava a reunião com	Presidente Português elogia o de Angola por este ter recebido Ministro da Defesa mesmo não constando agenda oficial do Ministério divulgada constava a reunião com	Presidente Português elogia o de Angola por este ter recebido Ministro da Defesa mesmo não constando agenda oficial do Ministério divulgada constava a reunião com	Presidente Português elogia de Angola por este ter recebido Ministro mesmo não constando agenda oficial do Ministério divulgada constava a reunião com	Presidente Português elogia o de Angola por este ter recebido Ministro da Defesa mesmo não constando agenda oficial do Ministério divulgada constava a reunião com
Tr= {T ∈ Texto/ T >M}	Presidente Português o de Angola Ministro da Defesa agenda	Presidente Português o de Angola Ministro da Defesa agenda	Presidente Português o de Angola Ministro da Defesa agenda	Presidente Português elogia de Angola por este ter recebido Ministro mesmo não constando agenda oficial do Ministério divulgada constava a reunião com	Presidente Português o de Angola Ministro da Defesa agenda
Tr= {T ∈ Texto/ T >(M+DP)}	o da Defesa	o da Defesa	o da Defesa	elogia por este ter recebido mesmo não constando oficial do Ministério divulgada constava a reunião com	o da Defesa

E, cumprindo os passos anteriores, é procedida a extração da percentagem de stopwords presentes no texto. Cujos resultados (ilustrado na tabela 4.3) correspondem aos termos relevantes dados por cada uma das AAM e as métricas definidas. O mesmo procedimento é feito para Bigrams e Trigrams de diferentes textos. O gráfico 4.1, ilustra a representação dos conjuntos de dados demonstrados na tabela 4.3.

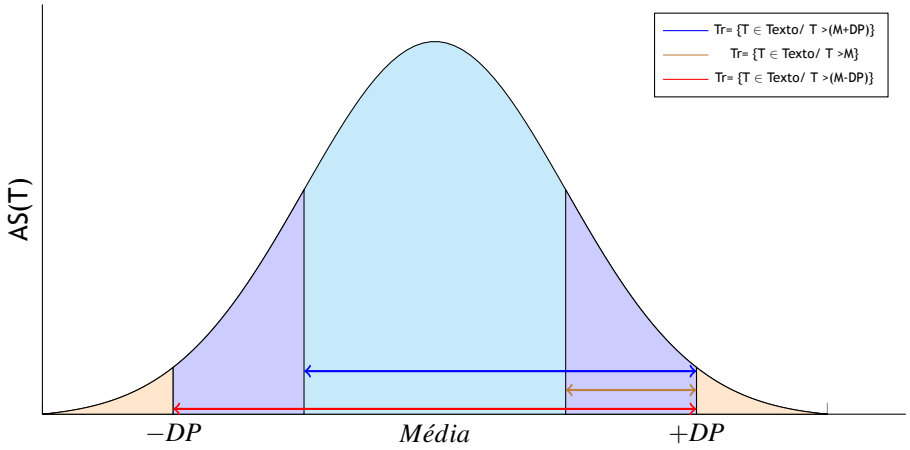

Tabela 4.3: Granularidade dos conjuntos para Unigram do texto 1, sem stopwords

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	Presidente Português elogia Angola recebido Ministro Defesa constando agenda oficial Ministério divulgada constava reunião	Presidente Português elogia Angola recebido Ministro Defesa constando agenda oficial Ministério divulgada constava reunião	Presidente Português elogia Angola recebido Ministro Defesa constando agenda oficial Ministério divulgada constava reunião	Presidente Português Angola recebido ando agenda oficial Ministério divulgada constava reunião	Presidente Português elogia Angola recebido Ministro Defesa constando agenda oficial Ministério divulgada constava reunião
Tr= {T ∈ Texto/ T >M}	Presidente Português Angola Ministro Defesa agenda	Presidente Português Angola Ministro Defesa agenda	Presidente Português Angola Ministro Defesa agenda	Presidente Português elogia Angola recebido Ministro constando agenda oficial Ministério divulgada constava reunião	Presidente Português Angola Ministro Defesa agenda
Tr= {T ∈ Texto/ T >(M+DP)}	Defesa	Defesa	Defesa	elogia recebido constando oficial Ministério divulgada constava reunião	Defesa

Para o conjunto Bigram foram obtidos os resultados apresentados nas respectivas tabelas (4.4, 4.5 e em 4.6)

Tabela 4.4: AS (T) para Bigram do texto 1

Types	Freq	P. termo	AS(T)-PC	AS(T)-BB	AS(T)-Co	AS(T)-GI	AS(T)-LP
presidente português	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
português elogia	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
elogia o	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
o presidente	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
presidente de	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
de angola	2	0.05263	1.63157	2.42435	1.50026	0.94474	29.85
angola por	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
por este	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
este ter	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
ter recebido	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
recebido o	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
o ministro	2	0.05263	1.63157	2.42435	1.50026	0.94474	29.85
ministro da	2	0.05263	1.63157	2.42435	1.50026	0.94474	29.85
da defesa	3	0.07894	2.44736	3.17619	2.19597	0.91531	40.0
defesa português	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
português mesmo	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
mesmo não	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
não constando	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
constando da	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
da agenda	2	0.05263	1.63157	2.42435	1.50026	0.94474	29.85
agenda oficial	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
oficial do	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
do ministério	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
ministério da	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
defesa da	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
agenda divulgada	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
divulgada constava	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
constava a	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
a reunião	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
reunião com	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
com o	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
defesa de	1	0.02631	0.81578	1.67662	0.76926	0.97300	19.73333
AISs(T)			30.99999	58.14242	28.96723	30.96555	692.20000
Ranking			2.44736	3.17619	2.19597	0.97300	40.0
			1.63157	2.42435	1.50026	0.94474	29.85
			0.81578	1.67662	0.76926	0.91531	19.73333
Média			0.96874	1.81695	0.90522	0.96767	21.63125
Desvio Padrão			0.37812	0.34708	0.33454	0.01324	4.69327

Tabela 4.5: Granularidade dos conjuntos para Bigram do texto 1

Métricas /AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da da Defesa Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de

Tabela 4.5 Granularidade dos conjuntos para Bigram do texto 1 (continuação)

Métricas /AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >M}	Presidente Português Português elogia elogia o o Presidente Presidente de Angola por por este este ter ter recebido recebido o Defesa Português Português mesmo mesmo não não constando constando da agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da da Defesa Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro y Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de

Tabela 4.5 Granularidade dos conjuntos para Bigram do texto 1 (continuação)

Métricas /AAM	**PC**	**BB**	**Co**	**GI**	**LP**
Tr= {T $\in$ Texto/ T >(M+DP)}	de Angola o Ministro Ministro da da Defesa	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da da Defesa Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da da Defesa Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por por este este ter ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo mesmo não não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com com o defesa de

Tabela 4.6: Granularidade dos conjuntos para Bigram do texto 1, sem stopwords

Métricas /AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da da Defesa Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de
Tr= {T ∈ Texto/ T >M}	Presidente Português Português elogia elogia o o Presidente Presidente de Angola por ter recebido recebido o Defesa Português Português mesmo não constando constando da agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da da Defesa Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de

Tabela 4.6 Granularidade dos conjuntos para Bigram do texto 1, sem stopwords (continuação)

Métricas /AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M+DP)}	de Angola o Ministro Ministro da da Defesa	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido y recebido o o Ministro Ministro da da Defsa Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da da Defesa Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	Presidente Português Português elogia elogia o o Presidente Presidente de de Angola Angola por ter recebido recebido o o Ministro Ministro da Defesa Português Português mesmo não constando constando da da agenda agenda oficial oficial do do Ministério Ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de

Tabela 4.7: AS (T) para Trigram do texto 1

Types	Freq	P. termo	AS(T)-PC	AS(T)-BB	AS(T)-Co	AS(T)-GI	AS(T)-LP
presidente português elogia	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
português elogia o	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
elogia o presidente	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
o presidente de	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
presidente de angola	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
de angola por	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
angola por este	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
por este ter	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
este ter recebido	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
ter recebido o	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
recebido o ministro	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
o ministro da	2	0.05405	1.83783	2.63602	1.69259	0.94318	33.75
ministro da defesa	2	0.05405	1.83783	2.63602	1.69259	0.94318	33.75
da defesa português	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
defesa português mesmo	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
português mesmo não	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
mesmo não constando	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
não constando da	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
constando da agenda	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
da agenda oficial	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
agenda oficial do	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
oficial do ministério	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
do ministério da	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
ministério da defesa	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
da defesa da	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
defesa da agenda	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
da agenda divulgada	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
agenda divulgada constava	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
divulgada constava a	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
constava a reunião	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
a reunião com	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
reunião com o	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
com o ministro	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
da defesa de	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
defesa de angola	1	0.02702	0.91891	1.78736	0.86845	0.97226	22.33333
AISs(T)			33.99999	64.25515	32.04433	33.97101	804.50000
Ranking			1.83783	2.63602	1.69259	0.97226	33.75
			0.91891	1.78736	0.86845	0.94318	22.33333
Média			0.97142	1.83586	0.91555	0.97060	22.98571
Desvio Padrão			0.21329	0.19698	0.19129	0.00674	2.64998

Tabela 4.8: Granularidade dos conjuntos para Trigram do texto 1, sem stopwords

Métricas /AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este por este ter este ter recebido ter recebido o recebido o ministro da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola

Tabela 4.8 Granularidade dos conjuntos para Trigram do texto 1, sem stopwords (continuação)

Métricas /AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >M}	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente portuguê elogia português elogia o elogia o presidente o presidente de presidente de angol de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa portuguê defesa português me português mesmo nã mesmo não constanc não constando da constando da agend da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola

Tabela 4.8 Granularidade dos conjuntos para Trigram do texto 1, sem stopwords (continuação)

Métricas /AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M+DP)}	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola

Este documento ilustra mais dois dos testes realizados, um em Inglês e o outro em Espanhol. Testes esses ilustrados nos anexos, são apresentados os termos relevantes (após cumprir escrupulosamente com os passos definidos e demonstrados para o texto 1) correspondentes aos textos 2 e 3.

No decurso deste trabalho, no que diz respeito a extração de termos relevantes vários testes foram realizados, com o objetivo de avaliar as Medidas de Associação Assimétrica e as métricas propostas, com os resultados alcançados o trabalho pode servir como base para o desenvolvimento de novos estudos na área. As abordagens defendidas e propostas têm aplicações práticas na vida quotidiana de pesquisadores, estudantes, desenvolvedores ou qualquer outro usuário que tenham interesses em LC, PLN ou ainda em IA. O projeto do Hultig-C visa contribuir de forma significativa para a comunidade Científica, Académica e Estudantil. Os recursos disponibilizados pelo Corpus visam apoiar a educação, pesquisa e desenvolvimento de tecnologias relacionadas à linguística computacional, compartilhando recursos, tais como dados, ferramentas e padrões. Destinando-se assim a todos quantos tenham interesses por áreas afins, ou ainda os que desenvolvem ou pretendam desenvolver programas multilingues, e que consequentemente precisam de matéria prima para dar suporte ao trabalho que tencionem desenvolver.

Numa era em que observa-se uma crescente evolução e renovação em termos tecnológicos, e com o constante lançamento de novas tecnologias, é necessário focar em soluções cada vez mais maleáveis, adequadas aos problemas propostos e expansíveis. Com o exposto e de acordo ao problema a que nos propomos resolver neste trabalho, foram obtidos bons resultados.

4.2 Plataforma

A Plataforma Web foi desenhada com o objetivo de integrar os diversos elementos da infraestrutura do Hultig-C, cuja interface busca facilitar a interação com o usuário, sem necessidade de instalar qualquer programa. A Plataforma serve ainda como base para o desenvolvimento de novas aplicações. O Servidor *Web Apache / 2.2.15 (Oracle)* é encarregue da hospedagem da Plataforma Web, que pode ser acessada através do endereço Web hultigcorpus.di.ubi.pt, Porta 80, que pode ser acessado pelo *browser* instalado no computador.

4.2.1 Hultig-Corpus

O Hultig-Corpus ou simplesmente Hultig-C, é um corpus multilingue, desenvolvido e mantido pelo Centro de Tecnologia da Linguagem Humana e Bioinformática (HULTIG) fundada em 2003 por Gael Harry Dias e atualmente dirigido pelo Dr. João Paulo Cordeiro, Docente do Departamento de Informática da Universidade da Beira Interior (UBI), Covilhã, Portugal. Cujo principal objetivo é apoiar a pesquisa sobre a recuperação de informações e tecnologias relacionadas da linguagem humana. É constituído por um conjunto de textos eletrónicos, cuja coleta teve início em Janeiro de 2017, e consiste em cerca de milhares de palavras em diferentes temáticas e idiomas. O Hultig-C é um corpus dinâmico e contém aproximadamente 4, 943, 857 Páginas Web distribuídas. Através da sua Plataforma Hultig-C disponível na web, visa permitir o acesso total ao mecanismo de pesquisa e uma série de aplicações em várias áreas linguísticas mediante os serviços disponibilizados, cujo acesso é gratuito para fins não comerciais.

O Projeto obedeceu a etapa de seleção, captura e manipulação dos textos pertinentes relevantes para a pesquisa. E definiu-se o tipo de Corpus que se pretendia compilar. Para a seleção dos textos que fazem parte do Corpus, observou-se os critérios: gratuitidade; possibilidade de reprodução dos arquivos originais; classificação das bases para o domínio e subáreas escolhidas para a pesquisa. Tendo sido selecionados textos aleatórios de diferentes contextos e idiomas. Uma vez que a aquisição e captura dos textos definido foi a web, recorreu-se à tecnologias que facilitam o processo, e assim foi possível efetuar a busca com o uso de um mecanismos de busca,

o *Open Web Spider* (também conhecido como *Tracker* ou *Web Robot*) para pesquisar toda a Web e processar os resultados das buscas feitas.

4.2.1.1 Compilação e Indexação

A compilação consiste no armazenamento em arquivos predeterminados de todos os textos selecionados. A compilação do Corpus Hultig-C teve seu início em Janeiro de 2017, e consiste em cerca de milhares de palavras em diferentes idiomas, coletadas com base nos textos brutos (de diferentes naturezas, de diferentes níveis linguísticos e de sofisticação) obtidos através de sites da web com suporte às funcionalidades do *Open Web Spider* e da ferramenta *MySQL Workbench*.

Assim, por defeito retirou-se tudo que não fazia parte do texto (gráficos, tabelas, figuras e números de páginas) e os arquivos fontes foram convertidos do formato em PDF para o *TXT*, compatível para o processamento e em seguida as páginas coletadas são organizadas em uma base de dados criada no *MySQL* e alocada num computador local disponibilizado pelo HULTIG. A construção automática do Corpus é então feita com recurso as ferramentas e tecnologias mencionadas e que permitirão manter a BD atualizada.

Dessa forma o referido Corpus tem a como finalidade contribuir para a Comunidade Científica, Académica e Estudantil, disponibilizando serviços de PLN através de sua interface Web disponível, que facilitará o processo de busca e recuperação de informações, viabilizando deste modo estudos na área.

4.2.1.2 Design da Plataforma Hultig-C

Procuramos no presente trabalho desenhar uma plataforma que permitisse uma interação com o usuário de forma prática e simples. Dentre os design da plataforma podemos apresentar imagem da figura 4.1 ilustrando a página principal do Hultig-C e através dela podemos navegar pelas demais páginas, como é o caso da página de boas vindas, que contém a informação base do Hultig Corpus; bem como as páginas de aquisição e serviços.

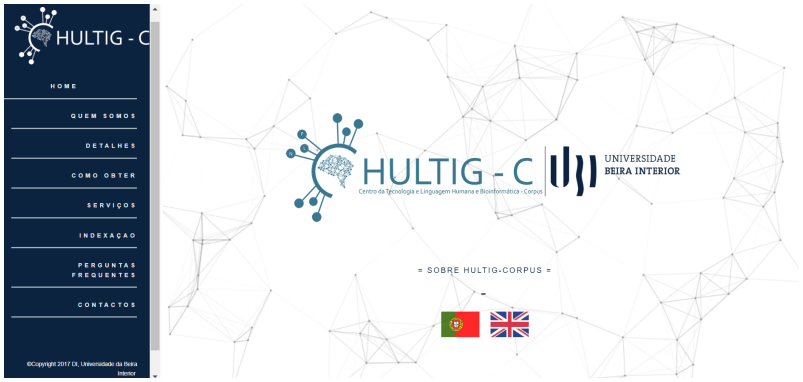

Figura 4.1: Página de rosto Hultig-C

Na ilustração (4.2) podemos encontrar uma descrição breve do Hultig-C clara e concisa, seu desenvolvimento e o grupo encarregue do mesmo.

Figura 4.2: Página de Boas Vindas do Hultig-C

Para o processo de aquisição a figura 4.3

Figura 4.3: Página como obter o Hultig-C

Os Serviços a disponibilizar na página estão em construção e por isso nenhum dos serviços descritos a seguir encontra-se disponível online. Seu acesso será feito através dos respetivos links que deverão estar presentes na página serviços.

4.3 Serviços

Hultig-C visa fornecer um conjunto de serviços para o processamento automático da linguagem humana, identificando padrões nas coleções de informações armazenadas de forma desorganizada. O acesso aos serviços disponibilizados é gratuito para fins não comerciais - Pesquisas Científicas, Académicas e Estudantis.

Dentre os quais podemos citar:

Figura 4.4: Página Serviços do Hultig-C

- Número de sites do domínio .pt
- Lista de sites do domínio .pt
- Número de sites do domínio .es
- Lista de sites do domínio .es
- Número de sites do domínio .fr
- Lista de sites do domínio .fr
- Número de sites do domínio .net
- Lista de sites do domínio .net
- *Textual Entailment* (TE); Neste serviço o Hultig-C prevê alcançar que a partir de uma sentença ou par de sentenças, se possa deduzir se os fatos no primeiro par implicam necessariamente os fatos no segundo. O conjuntos de ferramentas visa fornecer uma visualização da Implicação Textual (probabilidade percentual) a partir de um conjunto de entradas (premissa e hipótese), em que parte-se dos termos mais específicos para os gerais e ainda demonstrar o percentual da contradição e neutralidade.
- Extração de termos relevantes até *Trigram*;
- Destacar termos relevantes; O objetivo é poder visualizar o arquivo de texto de entrada, destacando as palavras relevantes.
- Extração da frequência de uma palavra em um mesmo texto ou em diferentes arquivos textuais.
- Determinar a frequência de sequências de termos em um texto. Com que frequência determinados termos aparecem em uma sequência.O objetivo poderia ser verificar quais palavras são mais frequentes, por exemplo, na posição inicial do termo, no meio ou no final.
- Localização de documentos relevantes;

4.4 Resultados

Mediante os testes feitos com os diferentes textos e as diferentes medidas, os resultados obtidos, foram muito bons. Quando comparados com métodos alternativos existentes na literatura (observe a tabela 4.9) é possível verificar o afirmado. Foi considerada como melhor Medida de Associação Assimétrica o Índice de Gini 3.6 e os dados cujos resultados são mais satisfatórios correspondem ao intervalo da Média, mas sem deixar acurar que as duas outras métricas definidas também apresentam bons resultados tanto para *Unigram, Bigram e Trigram*.

Tabela 4.9: Termos relevantes dados por diferentes metodologias para o texto 1

Ngram	**Y**ake	**H**ultig-C	**R**ake
Unigram	português defesa angola presidente ministro ministério mesmo agenda elogia recebido constando oficial divulgada constava reunião	presidente português angola ministro agenda elogia recebido constando oficial ministério divulgada constava reunião	ministério reunião presidente angola ministro constando defesa
Bigram	presidente português defesa português português elogia agenda oficial agenda divulgada divulgada constava	presidente português português elogia elogia o o presidente presidente de de angola angola por ter recebido recebido o o ministro ministro da da defesa defesa português português mesmo não constando constando da da agenda agenda oficial oficial do do ministério ministério da Defesa da agenda divulgada divulgada constava constava a a reunião reunião com defesa de	defesa português ter recebido agenda oficial
Trigram	presidente português elogia ministro da defesa presidente português oficial do ministério mesmo não constando recebido o ministro presidente de angola defesa de angola elogia o presidente constando da agenda	presidente português elogia português elogia o elogia o presidente o presidente de presidente de angola de angola por angola por este este ter recebido ter recebido o recebido o ministro o ministro da ministro da defesa da defesa português defesa português mesmo português mesmo não mesmo não constando não constando da constando da agenda da agenda oficial agenda oficial do oficial do ministério do ministério da ministério da defesa da defesa da defesa da agenda da agenda divulgada agenda divulgada constava divulgada constava a constava a reunião a reunião com reunião com o com o ministro da defesa de defesa de angola	presidente português elogia agenda divulgada constava

No decurso deste trabalho, no que diz respeito a extração de termos relevantes vários testes foram realizados, com o objetivo de avaliar as Medidas de Associação Assimétrica e as métricas propostas, com os resultados alcançados o trabalho pode servir como base para o desenvolvimento de novos estudos na área. As abordagens defendidas e propostas têm aplicações práticas na vida quotidiana de pesquisadores, estudantes, desenvolvedores ou qualquer outro usuário que tenham interesses em LC, PLN ou ainda em IA. O projeto do Hultig-C visa contribuir de forma significativa para a comunidade Científica, Académica e Estudantil. Os recursos disponibilizados pelo Corpus visam apoiar a educação, pesquisa e desenvolvimento de tecnologias relacionadas à linguística computacional, compartilhando recursos, tais como dados, ferramentas e padrões. Destinando-se assim a todos quantos tenham interesses por áreas afins, ou ainda os que desenvolvem ou pretendam desenvolver programas multilingues, e que consequentemente precisam de matéria prima para dar suporte ao trabalho que tencionem desenvolver.

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Conclusões

Esta dissertação de mestrado assumiu como objetivo o desenvolvimento de uma Plataforma Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em Big Data e sua posterior distribuição sem custos a comunidade científica e académica. Para tal o estudo baseio-se na necessidade da existência de um Corpus que disponibilize seus recursos sem custos monetários e sem restrições de uso ou limitações, um Corpus multi-disciplinar e multilingue e com novas abordagens para extração de termos relevantes, que neste trabalho são considerados como termos relevantes os termos específicos de um documento.

Realizou-se em primeiro lugar uma revisão da literatura buscando por Corpora e interfaces de PLN existentes e trabalhos científicos relacionados e ainda por ferramentas para compilação, processamento e análise de Corpora, a fim de situarmos no tempo e espaço o objetivo pretendido.

Desta feita, procedeu-se a recolha e indexação dos textos da Web para a criação do Hultig-Corpus com recurso às funcionalidades do *Open Web Spider*, e em simultâneo deu-se início ao desenvolvimento da plataforma Web. A maior percentagem do tempo foi dedicado a nova abordagem proposta para extração de termos relevantes em um documento de texto até T*rigram*. Onde os resultados são consideramos excelentes de acordo ao desenvolvimento experimental realizado.

Algumas dificuldades foram observadas quanto ao cumprimento das métricas definidas, o que se almejava alcançar e o que se alcançou. Dificuldades essas tanto materiais quanto de conhecimento que o trabalho em si assim o exige.

Apesar das limitações identificadas, e de outras que podem ser apontadas, considera-se que o estudo realizado permitiu adquirir maiores conhecimentos sobre a área em que se insere.

Foram testadas 8 medidas, das quais validaram-se 5; e por seus resultados consideramos como melhor, a Medida do*Gini Index* (GI), no conjunto de dados no intervalo $Tr = \{T \in Texto / T > M\}$.

Em vista dos argumentos apresentados consideramos que a $AIS_S(T)$ *(Asymmetric InfoSimba Similarity of Term)* com a Medida de Associação Assimétrica do Índice de Gini proposta para o Hultig-C é uma abordagem interessante em PLN e ao contrário das abordagens supervisionadas, dependentes de um Corpus de treinamento, a abordagem aqui proposta é totalmente não supervisionada, cujos recursos são retirados do próprio texto e sua independência em relação às técnicas de PLN a torna adequado para outros Corpora, incluindo domínios e idiomas diferentes. Podendo por isso ser uma mais valia para a comunidade Científica e Académica no campo do

Processamento da Linguagem Natural.

5.2 Trabalho Futuro

Como trabalho futuro, planeamos melhorar o design da página,

Observando que quando utilizada a web para a captura dos textos, a tecnologia oferece algumas opções que facilitem o processo, sendo possível efetuar a busca com o uso de mecanismos de busca para pesquisar toda a Web, textos esses completamente em bruto. Deste modo é nossa pretensão para trabalhos futuros fornecer textos ”limpos”, quer para extração de informações relevantes ou mesmo criação de Subcorpora ou Corpus.

Logo, é necessário que se faça a observação de aspetos como a Nomeação de arquivos e geração de cabeçalhos, nesta etapa pretendemos efetuar também a conversão dos textos do formato original a um formato de fácil processamento.

Posteriormente proceder a leitura do arquivo de textos, a conversão dos textos para letras minúsculas e a remoção de toda a informação considerada irrelevante, como: imagens, caracteres especiais, sinais de pontuação e números, remoção das *stopwords*, correção ortográfica e redução das palavras aos seus radicais. A nomeação dos arquivos deverá obedecer algum padrão a ser definido para facilitar a recuperação dos textos.

Perspetivamos aumentar serviços e disponibilizá-los online. E ainda investigar e avaliar como nosso método funciona em comparação com as abordagens supervisionadas.

Através dos serviços que pretendemos vir a disponibilizar, buscar desenvolver novas abordagens ou ainda investigar formas de melhoramento da abordagem ora proposta.

Bibliografia

[AdBA06] Sandra Maria Aluísio and Gladis Maria de Barcellos Almeida. O que é e como se constrói um corpus? lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio*, 4(3):156–178, 2006. 37, 38, 45, 46, 58

[AE17] Gabriel Assis Erbetta. Meuhorário 2: uma aplicação web para simulação de matrícula. 2017. 17, 20

[Agg14] Charu C Aggarwal. *Data classification: algorithms and applications*. CRC Press, 2014. 29

[Alu] Sandra Maria Aluísio. Recursos léxicos e ferramentas para a tarefa de criação de dicionários. 43

[APA+17] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017. 23, 24, 30, 31

[Arq] Arquivo.pt: pesquise páginas do passado. 51, 52, 53

[ATH16] Richard Addo-Tenkorang and Petri T Helo. Big data applications in operations/supply-chain management: A literature review. *Computers & Industrial Engineering*, 101:528–543, 2016. 7, 8

[AXD15] Andas Amrin, Chunlei Xia, and Shuguang Dai. Focused web crawling algorithms. *JCP*, 10(4):245–251, 2015. 21, 22, 23

[BBFZ09] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226, 2009. 45

[BCB16] Claudia Carmem Baggio, Heloisa Costa, and Ursula Blattmann. Seleção de tipos de fontes de informação. *Perspectivas em Gestão & Conhecimento*, 6(2):32–47, 2016. 13, 14

[Bil16] Cristhian Willrich Bilhalva. Ogd search-uma ferramenta de busca para dados governamentais. 2016. 27

[BMG17] Daniel Bicho, Fernando Melo, and Daniel Gomes. An evaluation of replay quality for web-archived pages. 2017. 53

[BPM09] Sotiris Batsakis, Euripides GM Petrakis, and Evangelos Milios. Improving the performance of focused web crawlers. *Data & Knowledge Engineering*, 68(10):1001–1013, 2009. 21, 22

[Bra04] Regina Meyer Branski. Recuperação de informações na web. *Perspectivas em ciência da informação*, 9(1):70–87, 2004. 14

[Bula] Bulnc: official website. 53, 54

[Bulb] Bulnc search: official website. [Online; accessed 2-Janeiro-2018]. Available from: http://search.dcl.bas.bg/. 53

[Cap] Kelvin Ramires Capobianco. Avaliação da etapa de pré-processamento na mineração de texto em redes sociais digitais. 10, 11, 23

[CB17] Paul Cook and Laurel J Brinton. Building and evaluating web corpora representing national varieties of english. *Language Resources and Evaluation*, 51(3):643–662, 2017. 45

[CdMGJ+08] Joel Sossai Coleti, Daniela Ferreira de Mattos, Luiz Carlos Genoves Jr, Arnaldo Candido Jr, Ariani Di Felippo, Gladis Maria de Barcelos Almeida, Sandra Maria Aluísio, and Osvaldo Novais de Oliveira Jr. Compilação de corpus em língua portuguesa na área de nanociência/nanotecnologia: problemas e soluções. *Avanços da linguística de Corpus no Brasil*, page 167, 2008. 37

[Cen01] Beatriz Valadares Cendón. Ferramentas de busca na web. *Ciência da Informação*, 30(1):39–49, 2001. 13

[CH90] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990. 62

[Cha18] E. Chamberlain. Subject directories. disponível, Acesso em Março de 2018. Available from: http://www.sc.edu/beaufort/library/pages/bones/lesson3.shtml. 13, 14

[Cho02] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002. 36

[CJVS+15] Arnaldo Candido Junior, Thiago Lima Vieira, Marcel Serikawa, Matheus Antônio Ribeiro Silva, Régis Zangirolami, Sandra Maria Aluisio, et al. Portal min@ s: uma ferramenta de apoio ao processamento de córpus de propósito geral. In *Brazilian Symposium in Information and Human Language Technology, X*. Universidade Federal do Rio Grande do Norte-UFRN, 2015. 42, 43, 46, 47

[CML14] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014. 7

[col] Etapas da metodologia de mineração de textos. 11, 12, 13

[com18] computerworld. disponível, Acesso em Janeiro de 2018. Available from: https://computerworld.com.br/2016/02/03/cisco-projeta-expansao-intensa-dos-dados-moveis-afetando-custos-de-ti/. 8

[con17] Wikipedia contributors. Focused crawler — wikipedia, the free encyclopedia, 2017. [Online; accessed 8-April-2018]. Available from: https://en.wikipedia.org/w/index.php?title=Focused_crawler&oldid=811585337. 18

[con18a] Wikipedia contributors. Robots exclusion standard — wikipedia, the free encyclopedia, 2018. [Online; accessed 8-April-2018]. Available from: https://en.wikipedia.org/w/index.php?title=Robots_exclusion_standard&oldid=832658245. 16

[con18b] Wikipedia contributors. Web crawler — wikipedia, the free encyclopedia, 2018. [Online; accessed 8-April-2018]. Available from: https://en.wikipedia.org/w/index.php?title=Web_crawler&oldid=833201574. 16, 17, 20

[Cra16] Clarissa Bastos Craveiro. A utilização do wordsmith nas pesquisas de educação e de ensino. *Revista de Educação, Ciências e Matemática*, 6(2), 2016. 42

[CS14] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014. 28, 30

[CW14] Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014. 7

[CWVM17] Haw-Shiuan Chang, ZiYun Wang, Luke Vilnis, and Andrew McCallum. Distributional inclusion vector embedding for unsupervised hypernymy detection. 2017. 34

[dAG+17] Luís Henrique G de Aguiar, Marcus Vinícius C Guelpeli, et al. Uma coleção de artigos científicos de português compondo um corpus no domínio educacional. *PLURAIS-Revista Multidisciplinar*, 2(1):60–74, 2017. 2, 36, 45, 46

[dAM17] Cintia Ferreira de Almeida and Jane Marian. Um olhar para a linguagem na produção de artigos científicos. *Caderno PAIC*, 18(1):478–503, 2017. 35

[Dav14] Mark Davies. The corpus of contemporary american english (coca) and the british national corpus (bnc), 2014. 45

[Day05] Carmen Dayrell. O uso de corpora para o estudo da tradução: objetivos e pressupostos. *Tradução em revista: Intervenções*, 2:87–102, 2005. 39

[DB17] Edeleon Marcelo Nunes DE Brito. Mineração de textos: detecção automática de sentimentos em co-mentários nas mídias sociais. *Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento*, 6(1), 2017. 24, 27, 28

[Dia10] Gaël Dias. *Information Digestion*. PhD thesis, Université d'Orléans, 2010. 62, 63, 64

[DM] Sebastifio Pais Gaél Dias and Rumen Moraliyski. Unsupervised and language-independent method to recognize textual entailment by generality. 34, 55

[DMGG16] Andrea De Mauro, Marco Greco, and Michele Grimaldi. A formal definition of big data based on its essential features. *Library Review*, 65(3):122–135, 2016. 8, 9

[DPWW+] Gaël Diasa, Sebastiao Paisa, Katarzyna Wegrzyn-Wolskac, Robert Mahld, and Pierre Jouvelotd. Textual entailment by generality. 34, 55

[DSC06] Gaël Dias, Cláudia Santos, and Guillaume Cleuziou. Automatic knowledge representation using a graph-based algorithm for language-independent lexical chaining. In *Proceedings of the Workshop on Information Extraction Beyond The Document*, pages 36–47. Association for Computational Linguistics, 2006. 64

[dSPB17] Leandro Augusto da Silva, Sarajane Marques Peres, and Clodis Boscarioli. *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil, 2017. 24, 25

[Dwy14] Gareth Dwyer. Building a corpus of south african english: Literature review. 2014. 45

[ea13] Hack et al. Text mining. data mining. 2013. 27

[EF16] Aline Evers and Maria José Bocorny Finatto. Linguística de corpus, léxico-estatística textual e processamento de linguagem natural: perspectiva para estudos de vocabulário em produções textuais. *Revista GTLex*, 1(2):271–295, 2016. 7

[EKB16] Thomas Erl, Wajid Khattak, and Paul Buhler. *Big Data Fundamentals: Concepts, Drivers & Techniques*. Prentice Hall Press, Upper Saddle River, NJ, USA, 1st edition, 2016. 8

[EMC18] The digital universe of oportunities. disponível, Acesso em Janeiro de 2018. Available from: `https://www.emc.com/infographics/digital-universe-2014.htm`. 8

[Far17] Flaubi Farias. Sitemap xml: tudo o que você precisa saber., 2017. Available from: `https://resultadosdigitais.com.br/blog/sitemap-xml/`. 15

[FDSB14] Lincoln Paulo Fernandes, Ana Paula de Carvalho Demétrio, Danielle Amanda Raimundo da Silva, and Mara Gonzalez Bezerra. Exemplos de corpora online: Aprendendo a classificar um corpus. 2014. 39, 40

[Fer17] Renato César Borges Ferreira. Uma abordagem semiautomática para identificação de elementos de processo de negócio em texto de linguagem natural. 2017. 6, 7, 17, 19

[FMMG+16] Fernandez Franco, Andrey Mauricio, Mauricio Muñoz Guzmán, et al. Caracterización de la información mediante la extracción de metadatos utilizando recuperación de información sobre convocatorias. 2016. 44

[FP17] David Feitosa and Vladia Pinheiro. Análise de medidas de similaridade semântica na tarefa de reconhecimento de implicação textual (analysis of semantic similarity measures in the recognition of textual entailment task)[in portuguese]. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 161–170, 2017. 34

[Fra15] Datumbox Machine Learning Framework. Datumbox: official website, 2015. 44

[GGS17] Abhinav Garg, Kratika Gupta, and Abhijeet Singh. Survey of web crawler algorithms. *International Journal*, 8(5), 2017. 23

[GL03] Marco Gonzalez and Vera Lúcia Strube Lima. Recuperação de informação e processamento da linguagem natural. In *XXIII Congresso da Sociedade Brasileira de Computaçao*, volume 3, pages 347–395, 2003. 2

[GL09] Vishal Gupta and Gurpreet S Lehal. and gurpreet s lehal. a survey of text mining techniques and applications. In *Journal of Emerging Technologies in Web Intelligence*. Citeseer, 2009. 11

[Goo72a] Nelson Goodman. Problems and projects. 1972. 60

[Goo72b] Nelson Goodman. Seven strictures on similarity. 1972. 60

[GP17] Ronaldo Goldschmidt and Emmanuel Passos. *Data Mining*. Elsevier Brasil, 2017. 27, 28

[GR11] John Gantz and David Reinsel. Extracting value from chaos. *IDC iview*, 1142(2011):1–12, 2011.

[GR15] J Gantz and D Reinsel. Extracting value from chaos. idc iview (2011), 2015. 7

[Gue12] Marcus Vinicius Carvalho Guelpeli. *Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização*. PhD thesis, Tese (doutorado)–Universidade Federal Fluminense. Programa de Pós-graduação em Computação, Niteroi, BR-RJ, Brasil, 2012. 45, 46

[HR15] BS Harish and MB Revanasiddappa. A quantitative evaluation of text feature selection methods. *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering*, 2(8), 2015. 30, 31, 32

[HTI16] Daiki Hayakawa, Masatoshi Tsuchiya, and Hitoshi Isahara. Developing corpus of japanese-english singular sentence textual entailment. In *Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On*, pages 1–6. IEEE, 2016. 34

[Ins] SAS Insights. Big data insights. 9

[JM14] Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London:, 2014. 7

[JM17] Daniel Jurafsky and James H. Martin. Speech and language processing: An introduction to natural language processing. *Computational Linguistics, and Speech Recognition (3rd ed draft chapter 4)*, 2017. 59

[KCS+15] Nikita P Katariya, MS Chaudhari, B Subhani, G Laxminarayana, Kalyani Matey, Ms Archana Nikose, Sonali A Tinkhede, and SP Deshpande. Text preprocessing for text mining using side information. *International Journal of Computer Science and Mobile Applications*, 3(1):01–05, 2015. 26

[Ken98] Dorothy Kenny. Corpora in translation studies. *Routledge encyclopedia of translation studies*, pages 50–53, 1998. 39

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 65

[KM14] Vipin Kumar and Sonajharia Minz. Feature selection. *SmartCR*, 4(3):211–229, 2014. 28, 29, 30, 31, 32, 33

[KQRdC17] Carlos Mamori Kono, Luc Quoniam, Leonel Cezar Rodrigues, and Hérmani Magalhães Olivense do Carmo. O uso criativo dos mecanismos de busca da web 2.0 para pesquisar invenções e criar inovações frugais. *Future Studies Research Journal: Trends and Strategies*, 9(2):30–60, 2017. 13, 14

[KST+16] Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetozara Leseva, and Tsvetana Dimitrova. Metadata extraction, representation and management within the bulgarian national corpus. In *4 th Workshop on Challenges in the Management of Large Corpora Workshop Programme*, page 33, 2016. 45, 53

[LCC18] Leipzig corpora collection download page. disponível„ Acesso em Março de 2018. Available from: `http://wortschatz.uni-leipzig.de/en/download/`. 50

[lem] Clueweb: official website. 47, 48

[LJYC15] Kuan-Ching Li, Hai Jiang, Laurence T Yang, and Alfredo Cuzzocrea. *Big data: Algorithms, analytics, and applications*. CRC Press, 2015. 9

[LMMyGJS+18] José de Jesús Lavalle Martínez, Manuel Montes y Gómez, Héctor Jiménez Salazar, Luis Villaseñor Pineda, and Beatriz Beltrán Martínez. Automatic theorem proving for natural logic: a case study on textual entailment. *Computación y Sistemas*, 22(1), 2018. 34

[LSSG17] Hong Liang, Xiao Sun, Yunlei Sun, and Yuan Gao. Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking*, 2017(1):211, 2017. 31

[LZZH16] Houqing Lu, Donghui Zhan, Lei Zhou, and Dengchao He. An improved focused crawler: using web page classification and link priority evaluation. *Mathematical Problems in Engineering*, 2016, 2016. 19

[M+75] David Meyer et al. Loci of contextual effects on visual word recognition. *Attention and performance vol. V.*, 1975. 62, 63

[M+17] Gilney Nathanael Mathias et al. qfex: um crawler para busca e extração de questionários de pesquisa em documentos html. 2017. 17

[MAM05] Trevor P Martin and Masrah Azmi-Murad. An incremental algorithm to find asymmetric word similarities for fuzzy text mining. In *Soft Computing as Transdisciplinary Science and Technology*, pages 838–847. Springer, 2005. 60, 61

[Mar14] Jane Marian. O estudo da linguística de corpus para a tradução especializada: Elaboração de um glossário da área da informática: Manutenção de computadores. *Cultura e Tradução*, 3(1), 2014. 35

[MBG16] Fernando Melo, Daniel Bicho, and Daniel Gomes. A comparison between the performance of wayback machines, 2016. 53

[MC16] Ana Maria Martins and Ernestina Carrilho. *Manual de linguística portuguesa*, volume 16. Walter de Gruyter GmbH & Co KG, 2016. 37, 40, 41

[Med16] José Medina. *Linguagem: conceitos-chave em filosofia*. Artmed Editora, 2016. 1

[MES07] Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. Asymmetric association measures. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007)*, 2007. 62

[MK17] Kuber Mohan and Jitendra Kurmi. A technique to improved page rank algorithm in perspective to optimized normalization technique. *International Journal of Advanced Research in Computer Science*, 8(3), 2017. 23

[MPdSK16] Cristian Cleder Machado, Daniel Heler Pohlmann, Eduardo Germano da Silva, and Luís Augusto Dias Knob. Um web crawler para projeções e análise de vulnerabilidades de segurança e consistência estrutural de páginas web. *Revista de Empreendedorismo, Inovação e Tecnologia*, 2(2):3–12, 2016. 15

[NZ16] Ronald Gonçalves das Neves and Thiago Chagas Zaccaro. Mineração de texto para análise de interações em redes sociais acadêmicas. 2016. 27, 28

[Orl17] Eni Puccinelli Orlandi. *O que é linguística*. Brasiliense, 2017. 35

[Pai08] Sebastião Pais. *Classification of opinionated texts by analogy*. PhD thesis, 2008. 2

[Pai13] Sebastião Pais. *Asymmetric Distributional Similarity Measures to Recognize Textual Entailment by Generality*. PhD thesis, Ecole Nationale Supérieure des Mines de Paris, 2013. 58, 59, 60, 61, 62, 63, 64, 65

[Pas17] Gloria Corpas Pastor. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS. Revista de traductología*, (5):155–184, 2017. 39, 40

[Pau15] Sébastien Paumier. Unitex 3.1. beta user manual. université paris-est marne-la-vallée, 2015. 43

[PB] Lucas Maciel Peixoto and Luiz Fernando Afra Brito. Procedimentos para compilação de um corpus composto por legendas e construção de uma ferramenta de corpus on-line: o corpus of english language videos. *Domínios de Linguagem*, 9(3):275–299. 36

[PC15] Ayar Pranav and Sandip Chauhan. Efficient focused web crawling approach for search engine. *Proc. International Journal of Computer Science and Mobile Computing*, 4(5), 2015. xiii, 19, 20, 21, 22, 23

[PdAL18] Deepak Panday, Renato Cordeiro de Amorim, and Peter Lane. Feature weighting as a tool for unsupervised feature selection. *Information Processing Letters*, 129:44–52, 2018. 30, 32

[PdLC] Thiago AS Pardo and Núcleo Interinstitucional de Linguística Computacional. Introdução ao processamento de línguas naturais. 7

[Pez17] Anderson Pezzini. Mineração de textos: Conceito, processo e aplicações. *REAVI-Revista Eletrônica do Alto Vale do Itajaí*, 5(8):58–61, 2017. 10, 24, 26

[PK15] Divya P and Nanda Kumar. Study on feature selection methods for text mining. 2015. 29

[Por] Portal min@s: official website. [Online; accessed 6-Janeiro-2018]. Available from: `http://portalminas.letras.ufmg.br/`. 46

[Por80] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. 26

[RLC18] Gil Rocha and Henrique Lopes Cardoso. Recognizing textual entailment: Challenges in the portuguese language. *Information*, 9(4):76, 2018. 33, 34

[RM05] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005. 11

[Rod07] S Rodrigues. Conceitos básicos de bd, sbd e sgbd, 2007. 2

[Rod16] Iuri de Vilhena Gonzalez Rodrigues. *Desenvolvimento de um motor de busca e comparação na Web*. PhD thesis, 2016. 15, 18

[Roj14] Guillermo Rojo. Hispanic corpus linguistics. *The Routledge Handbook of Hispanic Applied Linguistics. Nueva York: Routledge*, pages 371–387, 2014. 36

[Ros11] João Luís Garcia Rosa. Fundamentos da inteligência artificial. *Rio de Janeiro: LTC*, 2011. 6

[RQHB06] Matthias Richter, Uwe Quasthoff, Erla Hallsteinsdóttir, and Chris Biemann. Exploiting the leipzig corpora collection. *Proceedings of the IS-LTC*, pages 68–73, 2006. 48, 49, 50

[SA16] PSG Aruna Sri and M Anusha. Big data-survey. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 4(1):74–80, 2016. 8

[San15] Cedric Michael dos Santos. *Classificação de documentos com processamento de linguagem natural*. PhD thesis, 2015. 23, 24, 25, 26

[Sar04] Tony Berber Sardinha. *Lingüística de corpus*. Editora Manole Ltda, 2004. 35, 36, 37

[Sha06] Serge Sharoff. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98, 2006. 45

[Sil08] Filipe Pereira da Silveira. Integração de ferramentas para compilação e exploração de corpora. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 2008. 36, 42, 43

[SNGB15] Tanik Saikh, Sudip Kumar Naskar, Chandan Giri, and Sivaji Bandyopadhyay. Textual entailment using different similarity metrics. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 491–501. Springer, 2015. 34

[SSCN+15] Ronnie ES Santos, Ellen PR Souza, Jorge S Correia-Neto, Cleyton VC Magalhães, and Guilherme Vilar. Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: Resultados preliminares de um mapeamento sistemático. *Revista de Sistemas e Computação-RSC*, 4(2), 2015. 1

[TAL14] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014. 28, 29, 30

[tec18] techopedia.com. disponível, Acesso em Março de 2018. 30

[Tho21] Edward L Thorndike. The teachers word book. 1921. 35

[V+17] Augusto Vinhaes et al. Busca de informação na internet. rev. e atual. 2017. 13, 14

[VBDS+16] Karane Vieira, Luciano Barbosa, Altigran Soares Da Silva, Juliana Freire, and Edleno Moura. Finding seeds to bootstrap focused crawlers. *World Wide Web*, 19(3):449–474, 2016. 18

[VGC16] Hugo Viana, Daniel Gomes, and Miguel Costa. Architecture of the portuguese web archive search system. 2016. 51, 53

[Vie15] Nuno Miguel Leal Gonçalves Vieira. Splineapi, uma api rest para serviços de processamento de linguagem natural. 2015. 6

[Vii16] Andres Viikmaa. Web data extraction for content aggregation from e-commerce websites. 2016. 17

[VRG14] Tanu Verma, R Renu, and D Gaur. Tokenization and filtering process in rapidminer. *International Journal of Applied Information Systems*, 7(2):16–18, 2014. 24

[Wik17] Wikipédia. Função de verossimilhança — wikipédia, a enciclopédia livre, 2017. [Online; accessed 21-Fevereiro-2018]. Available from: `https://pt.wikipedia.org/w/index.php?title=Fun%C3%A7%C3%A3o_de_verossimilhan%C3%A7a&oldid=49361901`. 49

[You18] Subject directories. disponível, Acesso 19 April 2018. Available from: `http://www.yourdictionary.com/subject-directory`. 13

[Zil09] Diego Zilio. Inteligência artificial e pensamento: redefinindo os parâmetros da questão primordial de turing. *Ciências & Cognição*, 14(1):208–218, 2009. 5

[ZS15] Masoumeh Zareapoor and KR Seeja. Feature extraction or feature selection for text classification: A case study on phishing email detection. *International Journal of Information Engineering and Electronic Business*, 7(2):60, 2015. 29, 30

Apêndice A

Anexos

Assim, para o Texto 2, temos como termos relevantes os apresentados:

Is Trump Heading Back to ‘Fire and Fury’ With Kim?

On Sunday, President Donald Trump took to Twitter to savage the critics of his summit in Singapore on June 12 with Kim Jong Un.

“The denuclearization deal with North Korea is being praised and celebrated all over Asia,” he tweeted. “They are so happy! Over here, in our country, some people would rather see this historic deal fail than give Trump a win, even if it does save potentially millions & millions of lives!”

Trump, like Secretary of State Mike Pompeo, is feeling the heat of criticism over the joint statement signed in Singapore. Many think that vaguely worded document will not lead to the “amazing deal” the president promised the world in his June 12 interview with Voice of America’s Greta Van Susteren.

There is, despite Trump’s vigorous defense, much to be concerned about.

The American president, until three weeks ago, had brilliantly outmaneuvered Kim. Then, surprisingly, Trump began to make what looked like rookie negotiating mistakes, squandering hard-won advantage.

His most recent tactics have been so ineffective-and the break from smart tactics to poor ones so clear-that it appears Trump may have shifted goals from trying to disarm Kim to winning him over instead.

Yet if Trump is still trying to take away Kim’s missiles and nukes, the forecast is a return to the tension that marked most of last year.

Tabela A.1: Granularidade dos conjuntos para Unigram do texto 2, sem stopwords

Métricas / AAM	PC	*Braun-Blanket* (BB)	*Conviction* (Co)	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	sunday	sunday	sunday	sunday	sunday
	president	president	president	president	president
	donald	donald	donald	donald	donald
	twitter	twitter	twitter	twitter	twitter
	savage	savage	savage	savage	savage
	critics	critics	critics	critics	critics
	summit	summit	summit	summit	summit
	singapore	singapore	singapore	singapore	singapore
	june	june	june	june	june
	kim	kim	kim	kim	kim
	jong	jong	jong	jong	jong
	denuclearization	denuclearization	denuclearization	denuclearization	denuclearization
	deal	deal	deal	deal	deal
	north	north	north	north	north
	korea	korea	korea	korea	korea
	praised	praised	praised	praised	praised
	celebrated	celebrated	celebrated	celebrated	celebrated
	asia	asia	asia	asia	asia
	tweeted	tweeted	tweeted	tweeted	tweeted
	happy	happy	happy	happy	happy
	country	country	country	country	country
	people	people	people	people	people
	historic	historic	historic	historic	historic
	fail	fail	fail	fail	fail
	win	win	win	win	win
	save	save	save	save	save
	potentially	potentially	potentially	potentially	potentially
	millions	millions	millions	millions	millions
	secretary	secretary	secretary	secretary	secretary
	mike	state	mike	mike	mike
	pompeo	mike	pompeo	pompeo	pompeo
	feeling	pompeo	feeling	feeling	feeling
	heat	feeling	heat	heat	heat
	criticism	heat	criticism	criticism	criticism
	joint	criticism	joint	joint	joint
	statement	joint	statement	statement	statement
	signed	statement	signed	signed	signed
	vaguely	signed	vaguely	vaguely	vaguely
	worded	vaguely	worded	worded	worded
	document	worded	document	document	document
	lead	document	lead	lead	lead
	promised	lead	promised	promised	promised
	world	promised	world	world	world
	interview	world	interview	interview	interview
	voice	interview	voice	voice	voice
	greta	voice	greta	greta	greta
	van	greta	van	van	van
	despite	van	despite	despite	despite
	vigorous	despite	vigorous	vigorous	vigorous
	defense	vigorous	defense	defense	defense
	concerned	defense	concerned	concerned	concerned
	american	concerned	american	american	american
	weeks	american	weeks	weeks	weeks
	ago	weeks	ago	ago	ago
	brilliantly	brilliantly	brilliantly	brilliantly	brilliantly
	outmaneuvered	outmaneuvered	outmaneuvered	outmaneuvered	outmaneuvered
	surprisingly	surprisingly	surprisingly	surprisingly	surprisingly
	looked	looked	looked	looked	looked
	rookie	rookie	rookie	rookie	rookie
	negotiating	negotiating	negotiating	negotiating	negotiating
	mistakes	mistakes	mistakes	mistakes	mistakes
	squandering	squandering	squandering	squandering	squandering
	recent	recent	recent	recent	recent
	tactics	tactics	tactics	tactics	tactics
	break	break	break	break	break
	smart	smart	smart	smart	smart
	poor	poor	poor	poor	poor
	appears	appears	appears	appears	appears
	shifted	shifted	shifted	shifted	shifted
	goals	goals	goals	goals	goals
	trying	trying	trying	trying	trying
	disarm	disarm	disarm	disarm	disarm
	winning	winning	winning	winning	winning
	missiles	missiles	missiles	missiles	missiles
	nukes	nukes	nukes	nukes	nukes
	forecast	forecast	forecast	forecast	forecast
	return	return	return	return	return
	tension	tension	tension	tension	tension
	marked	marked	marked	marked	marked

$Tr = \{T \in Texto / T > M\}$					
	sunday	sunday	sunday	sunday president	sunday
	donald	donald	donald	donald	donald
	twitter	twitter	twitter	twitter	twitter
	savage	savage	savage	savage	savage
	critics	critics	critics	critics	critics
	summit	summit	summit	summit	summit
	singapore	singapore	singapore	singapore	singapore
	june	june	june	june	june
				kim	
	jong	jong	jong	jong	jong
	denuclearization	denuclearization	denuclearization	denuclearization	denuclearization
				deal	
	north	north	north	north	north
	korea	korea	korea	korea	korea
	praised	praised	praised	praised	praised
	celebrated	celebrated	celebrated	celebrated	celebrated
	asia	asia	asia	asia	asia
	tweeted	tweeted	tweeted	tweeted	tweeted
	happy	happy	happy	happy	happy
	country	country	country	country	country
	people	people	people	people	people
	historic	historic	historic	historic	historic
	fail	fail	fail	fail	fail
	win	win	win	win	win
	save	save	save	save	save
	potentially	potentially	potentially	potentially	potentially
	millions	millions	millions	millions	millions secretary
	secretary	secretary	secretary	secretary	mike
	mike	mike	mike	mike	pompeo
	pompeo	pompeo	pompeo	pompeo	feeling
	feeling	feeling	feeling	feeling	heat
	heat	heat	heat	heat	criticism
	criticism	criticism	criticism	criticism	joint
	joint	joint	joint	joint	statement
	statement	statement	statement	statement	signed
	signed	signed	signed	signed	vaguely
	vaguely	vaguely	vaguely	vaguely	worded
	worded	worded	worded	worded	document
	document	document	document	document	lead
	lead	lead	lead	lead	promised
	promised	promised	promised	promised	world
	world	world	world	world	interview
	interview	interview	interview	interview	voice
	voice	voice	voice	voice	greta
	greta	greta	greta	greta	van
	van	van	van	van	despite
	despite	despite	despite	despite	vigorous
	vigorous	vigorous	vigorous	vigorous	defense
	defense	defense	defense	defense	concerned
	concerned	concerned	concerned	concerned	american
	american	american	american	american	weeks
	weeks	weeks	weeks	weeks	ago
	ago	ago	ago	ago	brilliantly
	brilliantly	brilliantly	brilliantly	brilliantly	outmaneuvered
	outmaneuvered	outmaneuvered	outmaneuvered	outmaneuvered	surprisingly
	surprisingly	surprisingly	surprisingly	surprisingly	looked
	looked	looked	looked	looked	rookie
	rookie	rookie	rookie	rookie	negotiating
	negotiating	negotiating	negotiating	negotiating	mistakes
	mistakes	mistakes	mistakes	mistakes	squandering
	squandering	squandering	squandering	squandering	recent
	recent	recent	recent	recent	tactics
	tactics	tactics	tactics	tactics	break
	break	break	break	break	smart
	smart	smart	smart	smart	poor
	poor	poor	poor	poor	appears
	appears	appears	appears	appears	shifted
	shifted	shifted	shifted	shifted	goals
	goals	goals	goals	goals	trying
	trying	trying	trying	trying	disarm
	disarm	disarm	disarm	disarm	winning
	winning	winning	winning	winning	missiles
	missiles	missiles	missiles	missiles	nukes
	nukes	nukes	nukes	nukes	forecast
	forecast	forecast	forecast	forecast	return
	return	return	return	return	tension
	tension	tension	tension	tension	marked
	marked	marked	marked	marked	

Tr= {T ∈ Texto/ T >(M+DP)}	president trump kim deal	president trump kim deal	president trump kim deal	sunday president donald twitter savage critics summit singapore june kim jong denuclearization deal north korea praised celebrated asia tweeted happy country people historic fail win save potentially millions secretary mike pompeo feeling heat criticism joint statement signed vaguely worded document lead promised world interview voice greta van despite vigorous defense concerned american weeks ago brilliantly outmaneuvered surprisingly looked rookie negotiating mistakes squandering recent tactics break smart poor appears shifted goals trying disarm winning missiles nukes forecast return tension marked	president trump kim deal

Tabela A.2: Granularidade dos conjuntos para Bigram do texto 2, sem stopwords

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	on sunday sunday president president donald donald trump trump took to twitter twitter to to savage savage the the critics critics of his summit summit in singapore on on june june with with kim kim jong jong denuclearization denuclearization deal deal with with north north korea korea is being praised praised and and celebrated celebrated all over asia asia he he tweeted tweeted are so happy happy over our country country some some people people would this historic historic deal deal fail fail than give trump trump a a win win even does save save potentially potentially millions millions millions millions of like secretary secretary of state mike mike pompeo pompeo is is feeling feeling the the heat heat of of criticism criticism over the joint joint statement statement signed signed in singapore many that vaguely vaguely worded worded document document will not lead lead to	on sunday sunday president president donald donald trump trump took to twitter twitter to to savage savage the the critics critics of his summit summit in singapore on on june june with with kim kim jong jong denuclearization denuclearization deal deal with with north north korea korea is being praised praised and and celebrated celebrated all over asia asia he he tweeted tweeted are so happy happy over our country country some some people people would this historic historic deal deal fail fail than give trump trump a a win win even does save save potentially potentially millions millions millions millions of like secretary secretary of state mike mike pompeo pompeo is is feeling feeling the the heat heat of of criticism criticism over the joint joint statement statement signed signed in singapore many that vaguely vaguely worded worded document document will not lead lead to	on sunday sunday president president donald donald trump trump took to twitter twitter to to savage savage the the critics critics of his summit summit in singapore on on june june with with kim kim jong jong denuclearization denuclearization deal deal with with north north korea korea is being praised praised and and celebrated celebrated all over asia asia he he tweeted tweeted are so happy happy over our country country some some people people would this historic historic deal deal fail fail than give trump trump a a win win even does save save potentially potentially millions millions millions millions of like secretary secretary of state mike mike pompeo pompeo is is feeling feeling the the heat heat of of criticism criticism over the joint joint statement statement signed signed in singapore many that vaguely vaguely worded worded document document will not lead lead to	on sunday sunday president president donald donald trump trump took to twitter twitter to to savage savage the the critics critics of his summit summit in in singapore singapore on on june june with with kim kim jong jong denuclearization denuclearization deal deal with with north north korea korea is being praised praised and and celebrated celebrated all over asia asia he he tweeted tweeted are so happy happy over our country country some some people people would this historic historic deal deal fail fail than give trump trump a a win win even does save save potentially potentially millions millions millions millions of like secretary secretary of state mike mike pompeo pompeo is is feeling feeling the the heat heat of of criticism criticism over the joint joint statement statement signed signed in singapore many that vaguely vaguely worded worded document document will not lead lead to	on sunday sunday president president donald donald trump trump took to twitter twitter to to savage savage the the critics critics of his summit summit in singapore on on june june with with kim kim jong jong denuclearization denuclearization deal deal with with north north korea korea is being praised praised and and celebrated celebrated all over asia asia he he tweeted tweeted are so happy happy over our country country some some people this historic historic deal deal fail fail than give trump trump a a win win even does save save potentially potentially millions millions millions millions of like secretary secretary of state mike mike pompeo pompeo is is feeling feeling the the heat heat of of criticism criticism over the joint joint statement statement signed signed in singapore many that vaguely vaguely worded worded document document will not lead lead to

Tabela A.2 Granularidade dos conjuntos para Bigram do texto 2, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M-DP)}	the deal deal the the president president promised promised the the world world in his june june interview interview with with voice voice of of greta greta van van is is despite despite vigorous vigorous defense defense much be concerned concerned american american president president until three weeks weeks ago ago had had brilliantly brilliantly outmaneuvered outmaneuvered kim kim then then surprisingly surprisingly trump trump began what looked looked like like rookie rookie negotiating negotiating mistakes mistakes squandering squandering most most recent recent tactics tactics have the break break from from smart smart tactics tactics to to poor poor ones it appears appears trump trump may have shifted shifted goals goals from from trying to disarm disarm kim kim to to winning winning him if trump trump is still trying away missiles missiles and and nukes nukes the the forecast forecast is a return return to the tension tension that that marked marked most	the deal deal the the president president promised promised the the world world in his june june interview interview with with voice voice of of greta greta van van is is despite despite vigorous vigorous defense defense much be concerned concerned american american president president until three weeks weeks ago ago had had brilliantly brilliantly outmaneuvered outmaneuvered kim kim then then surprisingly surprisingly trump trump began what looked looked like like rookie rookie negotiating negotiating mistakes mistakes squandering squandering most most recent recent tactics tactics have the break break from from smart smart tactics tactics to to poor poor ones it appears appears trump trump may have shifted shifted goals goals from from trying to disarm disarm kim kim to to winning winning him if trump trump is still trying away missiles missiles and and nukes nukes the the forecast forecast is a return return to the tension tension that that marked marked most	the deal deal the the president president promised promised the the world world in his june june interview interview with with voice voice of of greta greta van van is is despite despite vigorous vigorous defense defense much be concerned concerned american american president president until three weeks weeks ago ago had had brilliantly brilliantly outmaneuvered outmaneuvered kim kim then then surprisingly surprisingly trump trump began what looked looked like like rookie rookie negotiating negotiating mistakes mistakes squandering squandering most most recent recent tactics tactics have the break break from from smart smart tactics tactics to to poor poor ones it appears appears trump trump may have shifted shifted goals goals from from trying to disarm disarm kim kim to to winning winning him if trump trump is still trying away missiles missiles and and nukes nukes the the forecast forecast is a return return to the tension tension that that marked marked most	the deal deal the the president president promised promised the the world world in his june june interview interview with with voice voice of of greta greta van van is is despite despite vigorous vigorous defense defense much be concerned concerned american american president president until three weeks weeks ago ago had had brilliantly brilliantly outmaneuvered outmaneuvered kim kim then then surprisingly surprisingly trump trump began what looked looked like like rookie rookie negotiating negotiating mistakes mistakes squandering squandering most most recent recent tactics tactics have the break break from from smart smart tactics tactics to to poor poor ones it appears appears trump trump may have shifted shifted goals goals from from trying trying to to disarm disarm kim kim to to winning winning him if trump trump is still trying away missiles missiles and and nukes nukes the the forecast forecast is a return return to the tension tension that that marked marked most	the deal deal the the president president promised promised the the world world in his june june interview interview with with voice voice of of greta greta van van is is despite despite vigorous vigorous defense defense much be concerned concerned american american president president until three weeks weeks ago ago had had brilliantly brilliantly outmaneuvered outmaneuvered kim kim then then surprisingly surprisingly trump trump began what looked looked like like rookie rookie negotiating negotiating mistakes mistakes squandering squandering most most recent recent tactics tactics have the break break from from smart smart tactics tactics to to poor poor ones it appears appears trump trump may have shifted shifted goals goals from from trying to disarm disarm kim kim to to winning winning him if trump trump is still trying away missiles missiles and and nukes nukes the the forecast forecast is a return return to the tension tension that that marked marked most

Tabela A.2 Granularidade dos conjuntos para Bigram do texto 2, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
	on sunday	on sunday	on sunday	on sunday	on sunday
	sunday president	sunday president	sunday president	sunday president	sunday president
	president donald	president donald	president donald	president donald	president donald
	donald trump	donald trump	donald trump	donald trump	donald trump
	trump took	trump took	trump took	trump took	trump took
	to twitter	to twitter	to twitter	to twitter	to twitter
	twitter to	twitter to	twitter to	twitter to	twitter to
	to savage	to savage	to savage	to savage	to savage
	savage the	savage the	savage the	savage the	savage the
	the critics	the critics	the critics	the critics	the critics
	critics of	critics of	critics of	critics of	critics of
	his summit	his summit	his summit	his summit	his summit
	summit in	summit in	summit in	summit in	summit in
				in singapore	
	singapore on	singapore on	singapore on	singapore on	singapore on
	on june	on june	on june	on june	on june
	june with	june with	june with	june with	june with
	with kim	with kim	with kim	with kim	with kim
	kim jong	kim jong	kim jong	kim jong	kim jong
	jong denuclearization	jong denuclearization	jong denuclearization	jong denuclearization	jong denuclearization
	denuclearization deal	denuclearization deal	denuclearization deal	denuclearization deal	denuclearization deal
	deal with	deal with	deal with	deal with	deal with
	with north	with north	with north	with north	with north
	north korea	north korea	north korea	north korea	north korea
	korea is	korea is	korea is	korea is	korea is
	being praised	being praised	being praised	being praised	being praised
	praised and	praised and	praised and	praised and	praised and
	and celebrated	and celebrated	and celebrated	and celebrated	and celebrated
	celebrated all	celebrated all	celebrated all	celebrated all	celebrated all
	over asia	over asia	over asia	over asia	over asia
	asia he	asia he	asia he	asia he	asia he
	he tweeted	he tweeted	he tweeted	he tweeted	he tweeted
	tweeted are	tweeted are	tweeted are	tweeted are	tweeted are
	so happy	so happy	so happy	so happy	so happy
	happy over	happy over	happy over	happy over	happy over
	our country	our country	our country	our country	our country
Tr= {T ∈ Texto/ T >M}	country some	country some	country some	country some	country some
	some people	some people	some people	some people	some people
	people would	people would	people would	people would	people would
	this historic	this historic	this historic	this historic	this historic
	historic deal	historic deal	historic deal	historic deal	historic deal
	deal fail	deal fail	deal fail	deal fail	deal fail
	fail than	fail than	fail than	fail than	fail than
	give trump	give trump	give trump	give trump	give trump
	trump a	trump a	trump a	trump a	trump a
	a win	a win	a win	a win	a win
	win even	win even	win even	win even	win even
	does save	does save	does save	does save	does save
	save potentially	save potentially	save potentially	save potentially	save potentially
	potentially millions	potentially millions	potentially millions	potentially millions	potentially millions
	millions millions	millions millions	millions millions	millions millions	millions millions
	millions of	millions of	millions of	millions of	millions of
	like secretary	like secretary	like secretary	like secretary	like secretary
	secretary of	secretary of	secretary of	secretary of	secretary of
	state mike	state mike	state mike	state mike	state mike
	mike pompeo	mike pompeo	mike pompeo	mike pompeo	mike pompeo
	pompeo is	pompeo is	pompeo is	pompeo is	pompeo is
	is feeling	is feeling	is feeling	is feeling	is feeling
	feeling the	feeling the	feeling the	feeling the	feeling the
	the heat	the heat	the heat	the heat	the heat
	heat of	heat of	heat of	heat of	heat of
	of criticism	of criticism	of criticism	of criticism	of criticism
	criticism over	criticism over	criticism over	criticism over	criticism over
	the joint	the joint	the joint	the joint	the joint
	joint statement	joint statement	joint statement	joint statement	joint statement
	statement signed	statement signed	statement signed	statement signed	statement signed
	signed in	signed in	signed in	signed in	signed in
	singapore many	singapore many	singapore many	singapore many	singapore many
	that vaguely	that vaguely	that vaguely	that vaguely	that vaguely
	vaguely worded	vaguely worded	vaguely worded	vaguely worded	vaguely worded
	worded document	worded document	worded document	worded document	worded document
	document will	document will	document will	document will	document will
	not lead	not lead	not lead	not lead	not lead
	lead to	lead to	lead to	lead to	lead to

Tabela A.2 Granularidade dos conjuntos para Bigram do texto 2, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
	the deal	the deal	the deal	the deal	the deal
	deal the	deal the	deal the	deal the	deal the
	the president	the president	the president	the president	the president
	president promised	president promised	president promised	president promised	president promised
	promised the	promised the	promised the	promised the	promised the
	the world	the world	the world	the world	the world
	world in	world in	world in	world in	world in
	his june	his june	his june	his june	his june
	june interview	june interview	june interview	june interview	june interview
	interview with	interview with	interview with	interview with	interview with
	with voice	with voice	with voice	with voice	with voice
	voice of	voice of	voice of	voice of	voice of
	of greta	of greta	of greta	of greta	of greta
	greta van	greta van	greta van	greta van	greta van
	van is	van is	van is	van is	van is
	is despite	is despite	is despite	is despite	is despite
	despite vigorous	despite vigorous	despite vigorous	despite vigorous	despite vigorous
	vigorous defense	vigorous defense	vigorous defense	vigorous defense	vigorous defense
	defense much	defense much	defense much	defense much	defense much
	be concerned	be concerned	be concerned	be concerned	be concerned
	concerned american	concerned american	concerned american	concerned american	concerned american
	american president	american president	american president	american president	american president
	president until	president until	president until	president until	president until
	three weeks	three weeks	three weeks	three weeks	three weeks
	weeks ago	weeks ago	weeks ago	weeks ago	weeks ago
	ago had	ago had	ago had	ago had	ago had
	had brilliantly	had brilliantly	had brilliantly	had brilliantly	had brilliantly
	brilliantly outmaneuvered	brilliantly outmaneuvered	brilliantly outmaneuvered	brilliantly outmaneuvered	brilliantly outmaneuvered
	outmaneuvered kim	outmaneuvered kim	outmaneuvered kim	outmaneuvered kim	outmaneuvered kim
	kim then	kim then	kim then	kim then	kim then
	then surprisingly	then surprisingly	then surprisingly	then surprisingly	then surprisingly
	surprisingly trump	surprisingly trump	surprisingly trump	surprisingly trump	surprisingly trump
	trump began	trump began	trump began	trump began	trump began
	what looked	what looked	what looked	what looked	what looked
	looked like	looked like	looked like	looked like	looked like
	like rookie	like rookie	like rookie	like rookie	like rookie
	rookie negotiating	rookie negotiating	rookie negotiating	rookie negotiating	rookie negotiating
	negotiating mistakes	negotiating mistakes	negotiating mistakes	negotiating mistakes	negotiating mistakes
Tr= {T ∈ Texto/ T >M}	mistakes squandering	mistakes squandering	mistakes squandering	mistakes squandering	mistakes squandering
	squandering most	squandering most	squandering most	squandering most	squandering most
	most recent	most recent	most recent	most recent	most recent
	recent tactics	recent tactics	recent tactics	recent tactics	recent tactics
	tactics have	tactics have	tactics have	tactics have	tactics have
	the break	the break	the break	the break	the break
	break from	break from	break from	break from	break from
	from smart	from smart	from smart	from smart	from smart
	smart tactics	smart tactics	smart tactics	smart tactics	smart tactics
	tactics to	tactics to	tactics to	tactics to	tactics to
	to poor	to poor	to poor	to poor	to poor
	poor ones	poor ones	poor ones	poor ones	poor ones
	it appears	it appears	it appears	it appears	it appears
	appears trump	appears trump	appears trump	appears trump	appears trump
	trump may	trump may	trump may	trump may	trump may
	have shifted	have shifted	have shifted	have shifted	have shifted
	shifted goals	shifted goals	shifted goals	shifted goals	shifted goals
	goals from	goals from	goals from	goals from	goals from
	from trying	from trying	from trying	from trying	from trying
				trying to	
	to disarm	to disarm	to disarm	to disarm	to disarm
	disarm kim	disarm kim	disarm kim	disarm kim	disarm kim
	kim to	kim to	kim to	kim to	kim to
	to winning	to winning	to winning	to winning	to winning
	winning him	winning him	winning him	winning him	winning him
	if trump	if trump	if trump	if trump	if trump
	trump is	trump is	trump is	trump is	trump is
	still trying	still trying	still trying	still trying	still trying
	away missiles	away missiles	away missiles	away missiles	away missiles
	missiles and	missiles and	missiles and	missiles and	missiles and
	and nukes	and nukes	and nukes	and nukes	and nukes
	nukes the	nukes the	nukes the	nukes the	nukes the
	the forecast	the forecast	the forecast	the forecast	the forecast
	forecast is	forecast is	forecast is	forecast is	forecast is
	a return	a return	a return	a return	a return
	return to	return to	return to	return to	return to
	the tension	the tension	the tension	the tension	the tension
	tension that	tension that	tension that	tension that	tension that
	that marked	that marked	that marked	that marked	that marked
	marked most	marked most	marked most	marked most	marked most

Tabela A.2 Granularidade dos conjuntos para Bigram do texto 2, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M+DP)}	in singapore	in singapore	in singapore	on sunday sunday president president donald donald trump trump took to twitter twitter to to savage savage the the critics critics of his summit summit in in singapore singapore on on june june with with kim kim jong jong denuclearization denuclearization deal deal with with north north korea korea is being praised praised and and celebrated celebrated all over asia asia he he tweeted tweeted are so happy happy over our country country some some people people would this historic historic deal deal fail fail than give trump trump a a win win even does save save potentially potentially millions millions millions millions of like secretary secretary of state mike mike pompeo pompeo is is feeling feeling the the heat heat of of criticism criticism over the joint joint statement statement signed signed in singapore many that vaguely vaguely worded worded document document will not lead lead to	in singapore

Tabela A.2 Granularidade dos conjuntos para Bigram do texto 2, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M+DP)}	trying to	trying to	trying to	the deal deal the the president president promised promised the the world world in his june june interview interview with with voice voice of of greta greta van van is is despite despite vigorous vigorous defense defense much be concerned concerned american american president president until three weeks weeks ago ago had had brilliantly brilliantly outmaneuvered outmaneuvered kim kim then then surprisingly surprisingly trump trump began what looked looked like like rookie rookie negotiating negotiating mistakes mistakes squandering squandering most most recent recent tactics tactics have the break break from from smart smart tactics tactics to to poor poor ones it appears appears trump trump may have shifted shifted goals goals from from trying trying to to disarm disarm kim kim to to winning winning him if trump trump is still trying away missiles missiles and and nukes nukes the the forecast forecast is a return return to the tension tension that that marked marked most	trying to

Tabela A.2 Granularidade dos conjuntos para Bigram do texto 2, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M+DP)}			to the the deal deal the the president president promised promised the the world world in in his his june june interview interview with with voice voice of of greta greta van van is is despite despite vigorous vigorous defense defense much much to to be be concerned concerned american american president president until until three three weeks weeks ago ago had had brilliantly brilliantly outmaneuvered outmaneuvered kim kim then then surprisingly surprisingly trump trump began began to to make make what what looked looked like like rookie rookie negotiating negotiating mistakes mistakes squandering squandering most most recent recent tactics tactics have have been been so so the the break break from from smart smart tactics tactics to to poor poor ones ones so so it it appears appears trump trump may may have have shifted shifted goals goals from from trying trying to to disarm disarm kim kim to to winning winning him him over over if if trump trump is is still still trying to take take away away missiles missiles and and nukes nukes the the forecast forecast is is a a return return to the tension tension that that marked marked most most of of last last year		

Tabela A.3: Granularidade dos conjuntos para Trigram do texto 2, sem stopwords

Métricas/ AAM	PC	BB	Co	GI	LP
	on sunday president	on sunday president	on sunday president on sunday president	on sunday president	on sunday president
	on sunday president	on sunday president	sunday president donald	on sunday president	on sunday president
	sunday president donald	sunday president donald	president donald trump	sunday president donald	sunday president donald
	president donald trump	president donald trump	donald trump took	president donald trump	president donald trump
	donald trump took	donald trump took	trump took to	donald trump took	donald trump took
	trump took to	trump took to	took to twitter	trump took to	trump took to
	took to twitter	took to twitter	to twitter to	took to twitter	took to twitter
	to twitter to	to twitter to	twitter to savage	to twitter to	to twitter to
	twitter to savage	twitter to savage	to savage the	twitter to savage	twitter to savage
	to savage the	to savage the	savage the critics	to savage the	to savage the
	savage the critics	savage the critics	the critics of	savage the critics	savage the critics
	the critics of	the critics of	critics of his	the critics of	the critics of
	critics of his	critics of his	of his summit	critics of his	critics of his
	of his summit	of his summit	his summit in	of his summit	of his summit
	his summit in	his summit in	summit in singapore	his summit in summit in singapore	his summit in
	summit in singapore	summit in singapore	in singapore on	in singapore on	summit in singapore
	in singapore on	in singapore on	singapore on june	singapore on june	in singapore on
	singapore on june	singapore on june	on june with	on june with	singapore on june
	on june with	on june with	june with kim	june with kim	on june with
	june with kim	june with kim	with kim jong	with kim jong	june with kim
	with kim jong	with kim jong	kim jong denuclearization	kim jong denuclearization	with kim jong
	kim jong denuclearization	kim jong denuclearization	jong denuclearization deal	jong denuclearization deal	kim jong denuclearization
	jong denuclearization deal	jong denuclearization deal	denuclearization deal with	denuclearization deal with	jong denuclearization deal
	denuclearization deal with	denuclearization deal with	deal with north	deal with north	denuclearization deal with
	deal with north	deal with north	with north korea	with north korea	deal with north
	with north korea	with north korea	north korea is	north korea is	with north korea
	north korea is	north korea is	korea is being	korea is being	north korea is
	korea is being	korea is being	is being praised	is being praised	korea is being
	is being praised	is being praised	being praised and	being praised and	is being praised
	being praised and	being praised and	praised and celebrated	praised and celebrated	being praised and
	praised and celebrated	praised and celebrated	and celebrated all	and celebrated all	praised and celebrated
	and celebrated all	and celebrated all	celebrated all over	celebrated all over	and celebrated all
	celebrated all over	celebrated all over	all over asia	all over asia	celebrated all over
	all over asia	all over asia	over asia he	over asia he	all over asia
	over asia he	over asia he	asia he tweeted	asia he tweeted	over asia he
	asia he tweeted	sia he tweeted	he tweeted are	he tweeted are	asia he tweeted
	he tweeted are	he tweeted are	tweeted are so	tweeted are so	he tweeted are tweeted are so
	tweeted are so	tweeted are so	are so happy	are so happy	are so happy
	are so happy	are so happy	so happy over	so happy over	so happy over
	so happy over	so happy over	happy over here	happy over here	happy over here
	happy over here	happy over here	in our country	in our country	in our country
	in our country	in our country	our country some	our country some	our country some
	our country some	our country some	country some people	country some people	country some people
	country some people	country some people	people would rather	people would rather	people would rather
	people would rather	people would rather	see this historic	see this historic	see this historic
	see this historic	see this historic	this historic deal	this historic deal	this historic deal
	this historic deal	this historic deal	historic deal fail	historic deal fail	historic deal fail
	historic deal fail	historic deal fail	deal fail than	deal fail than	deal fail than
	deal fail than	deal fail than	fail than give	fail than give	fail than give
	fail than give than give trump	fail than give	than give trump	than give trump	than give trump
	give trump a	than give trump	give trump a	give trump a	give trump a
	trump a win	give trump a	trump a win	trump a win	trump a win
	a win even	trump a win	a win even	a win even	a win even
	win even if	a win even	win even if	win even if	win even if
	it does save	win even if	it does save	it does save	it does save
	does save potentially	it does save	does save potentially	does save potentially	does save potentially
	save potentially millions	does save potentially	save potentially millions	save potentially millions	save potentially millions
	potentially millions millions	save potentially millions	potentially millions millions	potentially millions millions	potentially millions millions
	millions millions of	potentially millions millions	millions millions of	millions millions of	millions millions of
	millions of like	millions millions of	millions of like	millions of like	millions of like
	of like secretary	millions of like	of like secretary	of like secretary	of like secretary
	like secretary of	of like secretary	like secretary of	like secretary of	like secretary of
	secretary of state	like secretary of	secretary of state	secretary of state	secretary of state
	of state mike	secretary of state	of state mike	of state mike	of state mike
	state mike pompeo	of state mike	state mike pompeo	state mike pompeo	state mike pompeo
	mike pompeo is	state mike pompeo	mike pompeo is	mike pompeo is	mike pompeo is
	pompeo is feeling	mike pompeo is	pompeo is feeling	pompeo is feeling	pompeo is feeling
	is feeling the	pompeo is feeling	is feeling the	is feeling the	is feeling the
	feeling the heat	is feeling the	feeling the heat	feeling the heat	feeling the heat
	the heat of	feeling the heat	the heat of	the heat of	the heat of
	heat of criticism	the heat of	heat of criticism	heat of criticism	heat of criticism
	of criticism over	heat of criticism	of criticism over	of criticism over	of criticism over
	criticism over the	of criticism over	criticism over the	criticism over the	criticism over the
	over the joint	criticism over the	over the joint	over the joint	over the joint
	the joint statement	over the joint	the joint statement	the joint statement	the joint statement
	joint statement signed	the joint statement	joint statement signed	joint statement signed	joint statement signed
	statement signed in	joint statement signed	statement signed in	statement signed in	statement signed in
	signed in singapore	statement signed in	signed in singapore	signed in singapore	signed in singapore
	in singapore many	signed in singapore	in singapore many	in singapore many	in singapore many
	singapore many think	in singapore many	singapore many think	singapore many think	singapore many think
	think that vaguely	singapore many think	think that vaguely	think that vaguely	think that vaguely
	that vaguely worded	think that vaguely	that vaguely worded	that vaguely worded	that vaguely worded
	vaguely worded document	that vaguely worded	vaguely worded document	vaguely worded document	vaguely worded document
	worded document will	vaguely worded document	worded document will	worded document will	worded document will
	document will not	worded document will	document will not	document will not	document will not
	will not lead	document will not	will not lead	will not lead	will not lead
	not lead to	will not lead	not lead to	not lead to	not lead to
Tr= {T ∈ Texto/ T >(M-DP)}	lead to the	not lead to	lead to the	lead to the	lead to the
Tr= {T ∈ Texto/ T >M}	to the deal	lead to the	to the deal	to the deal	to the deal
Tr= {T ∈ Texto/ T >(M+DP)}	the deal the	to the deal	the deal the	the deal the	the deal the
	deal the president	the deal the	deal the president	deal the president	deal the president
	the president promised	deal the president	the president promised	the president promised	the president promised
	president promised the	the president promised	president promised the	president promised the	president promised the
	promised the world	president promised the	promised the world	promised the world	promised the world
	the world in	promised the world	the world in	the world in	the world in
	world in his	the world in	world in his	world in his	world in his
	in his june	world in his	in his june	in his june	in his june
	his june interview	in his june	his june interview	his june interview	his june interview
	june interview with	his june interview	june interview with	june interview with	june interview with
	interview with voice	june interview with	interview with voice	interview with voice	interview with voice
	with voice of	interview with voice	with voice of	with voice of	with voice of
	voice of greta	with voice of	voice of greta	voice of greta	voice of greta
	of greta van	voice of greta	of greta van	of greta van	of greta van
	greta van is	of greta van	greta van is	greta van is	greta van is
	van is despite	greta van is	van is despite	van is despite	van is despite
	is despite vigorous	van is despite	is despite vigorous	is despite vigorous	is despite vigorous
	despite vigorous defense	is despite vigorous	despite vigorous defense	despite vigorous defense	despite vigorous defense
	vigorous defense much	despite vigorous defense	vigorous defense much	vigorous defense much	vigorous defense much
	defense much to	vigorous defense much	defense much to	defense much to	defense much to
	to be concerned	defense much to	to be concerned	to be concerned	to be concerned
	be concerned american	to be concerned	be concerned american	be concerned american	be concerned american
	concerned american president	be concerned american	concerned american president	concerned american president	concerned american president
	american president until	concerned american president	american president until	american president until	american president until
	president until three	american president until	president until three	president until three	president until three
	until three weeks	president until three	until three weeks	until three weeks	until three weeks
	three weeks ago	until three weeks	three weeks ago	three weeks ago	three weeks ago
	weeks ago had	three weeks ago	weeks ago had	weeks ago had	weeks ago had
	ago had brilliantly	weeks ago had	ago had brilliantly	ago had brilliantly	ago had brilliantly
	had brilliantly outmaneuvered	ago had brilliantly	had brilliantly outmaneuvered	had brilliantly outmaneuvered	had brilliantly outmaneuvered
	brilliantly outmaneuvered kim	had brilliantly outmaneuvered	brilliantly outmaneuvered kim	brilliantly outmaneuvered kim	brilliantly outmaneuvered kim
	outmaneuvered kim then	brilliantly outmaneuvered kim	outmaneuvered kim then	outmaneuvered kim then	outmaneuvered kim then
	kim then surprisingly	outmaneuvered kim then	kim then surprisingly	kim then surprisingly	kim then surprisingly
	then surprisingly trump	kim then surprisingly	then surprisingly trump	then surprisingly trump	then surprisingly trump
	surprisingly trump began	then surprisingly trump	surprisingly trump began	surprisingly trump began	surprisingly trump began
	trump began to	surprisingly trump began	trump began to	trump began to	trump began to
	make what looked	trump began to	make what looked	make what looked	make what looked
	what looked like	make what looked	what looked like	what looked like	what looked like
	looked like rookie	what looked like	looked like rookie	looked like rookie	looked like rookie
	like rookie negotiating	looked like rookie	like rookie negotiating	like rookie negotiating	like rookie negotiating
	rookie negotiating mistakes	like rookie negotiating	rookie negotiating mistakes	rookie negotiating mistakes	rookie negotiating mistakes
	negotiating mistakes squandering	rookie negotiating mistakes	negotiating mistakes squandering	negotiating mistakes squandering	negotiating mistakes squandering
	mistakes squandering most	negotiating mistakes squandering	mistakes squandering most	mistakes squandering most	mistakes squandering most
	squandering most recent	mistakes squandering most	squandering most recent	squandering most recent	squandering most recent
	most recent tactics	squandering most recent	most recent tactics	most recent tactics	most recent tactics
	recent tactics have	most recent tactics	recent tactics have	recent tactics have	recent tactics have
	tactics have been	recent tactics have	tactics have been	tactics have been	tactics have been
	so the break	tactics have been	so the break	so the break	so the break
	the break from	so the break	the break from	the break from	the break from
	from smart tactics	the break from	from smart tactics	from smart tactics	from smart tactics
	smart tactics to	from smart tactics	smart tactics to	smart tactics to	smart tactics to
	tactics to poor	smart tactics to	tactics to poor	tactics to poor	tactics to poor
	to poor ones	tactics to poor	to poor ones	to poor ones	to poor ones
	poor ones so	to poor ones	poor ones so	poor ones so	poor ones so
	so it appears	poor ones so	so it appears	so it appears	so it appears
	it appears trump	so it appears	it appears trump	it appears trump	it appears trump
	appears trump may	it appears trump	appears trump may	appears trump may	appears trump may
	trump may have	appears trump may	trump may have	trump may have	trump may have
	may have shifted	trump may have	may have shifted	may have shifted	may have shifted
	have shifted goals	may have shifted	have shifted goals	have shifted goals	have shifted goals
	shifted goals from	have shifted goals	shifted goals from	shifted goals from	shifted goals from
	goals from trying	shifted goals from	goals from trying	goals from trying	goals from trying
	from trying to	goals from trying	from trying to	from trying to	from trying to
	trying to disarm	from trying to	trying to disarm	trying to disarm	trying to disarm
	to disarm kim	trying to disarm	to disarm kim	to disarm kim	to disarm kim
	disarm kim to	to disarm kim	disarm kim to	disarm kim to	disarm kim to
	kim to winning	disarm kim to	kim to winning	kim to winning	kim to winning
	to winning him	kim to winning	to winning him	to winning him	to winning him
	winning him over	to winning him	winning him over	winning him over	winning him over
	over if trump	winning him over	over if trump	over if trump	over if trump
	if trump is	over if trump	if trump is	if trump is	if trump is
	trump is still	if trump is	trump is still	trump is still	trump is still
	is still trying	trump is still	is still trying	is still trying	is still trying
	still trying to	is still trying	still trying to	still trying to	still trying to
	trying to take	still trying to	trying to take	trying to take	trying to take
	take away missiles	trying to take	take away missiles	take away missiles	take away missiles
	away missiles and	take away missiles	away missiles and	away missiles and	away missiles and
	missiles and nukes	away missiles and	missiles and nukes	missiles and nukes	missiles and nukes
	and nukes the	missiles and nukes	and nukes the	and nukes the	and nukes the
	nukes the forecast	and nukes the	nukes the forecast	nukes the forecast	nukes the forecast
	the forecast is	nukes the forecast	the forecast is	the forecast is	the forecast is
	forecast is a	the forecast is	forecast is a	forecast is a	forecast is a
	is a return	forecast is a	is a return	is a return	is a return
	a return to	is a return	a return to	a return to	a return to
	return to the	a return to	return to the	return to the	return to the
	to the tension	return to the	to the tension	to the tension	to the tension
	the tension that	to the tension	the tension that	the tension that	the tension that
	tension that marked	the tension that	tension that marked	tension that marked	tension that marked
	that marked most	tension that marked	that marked most	that marked most	that marked most
	marked most of	that marked most	marked most of	marked most of	marked most of
		marked most of			

E para o Texto 3, os apresentados nas tabelas A.4, A.5 e A.6 , correspondendo a Unigram, Bigram e Trigram:

La competición, que reúne a los ocho mejores equipos de Europa en categoría masculina y femenina, se disputará en Gaetà Huguet.

La Copa de Europa de clubes de atletismo se disputará en 2019 en Castellón, según ha dado a conocer la Asociación Europea de Atletismo, y tendrá al Playas de Castellón como entidad organizadora. Las instalaciones de Gaeta Huguet serán el escenario en el que competirán los mejores ocho clubes del continente tanto en categoría masculina como femenina, con el Playas como anfitrión y uno de los aspirantes al título, al igual que el Valencia Esports. En la competición femenina. La Copa de Europa se disputará los días 25 y 26 de mayo y ofrecerá la oportunidad para que los amantes del atletismo puedan ver de cerca a los mejores atletas del continente.

Tabela A.4: Granularidade dos conjuntos para Unigram do texto 3, sem stopwords

Métricas / AAM	**PC**	**BB**	**Co**	**GI**	**LP**
Tr= {T ∈ Texto/ T >(M-DP)}	competición reúne mejores equipos europa categoría masculina femenina disputará gaetà copa clubes atletismo castellón conocer asociación europea playas entidad organizadora instalaciones gaeta huguet escenario competirán continente anfitrión aspirantes título valencia esports días mayo ofrecerá oportunidad amantes atletas	competición reúne mejores equipos europa categoría masculina femenina disputará gaetà copa clubes atletismo castellón conocer asociación europea playas entidad organizadora instalaciones gaeta huguet escenario competirán continente anfitrión aspirantes título valencia esports días mayo ofrecerá oportunidad amantes atletas	competición reúne mejores equipos europa categoría masculina femenina disputará gaetà copa clubes atletismo castellón conocer asociación europea playas entidad organizadora instalaciones gaeta huguet escenario competirán continente anfitrión aspirantes título valencia esports días mayo ofrecerá oportunidad amantes atletas	competición reúne mejores equipos europa categoría masculina femenina disputará gaetà copa clubes atletismo castellón conocer asociación europea playas entidad organizadora instalaciones gaeta huguet escenario competirán continente anfitrión aspirantes título valencia esports días mayo ofrecerá oportunidad amantes y atletas	competición reúne mejores equipos europa categoría masculina femenina disputará gaetà copa clubes atletismo castellón conocer asociación europea playas entidad organizadora instalaciones gaeta huguet escenario competirán continente anfitrión aspirantes título valencia esports días mayo ofrecerá oportunidad amantes atletas

Tabela A.4 Granularidade dos conjuntos para Unigram do texto 3, sem stopwords (continuação)

Métricas / AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >M}	competición mejores europa categoría masculina femenina disputará copa clubes atletismo castellón playas continente	competición mejores europa categoría masculina femenina disputará copa clubes atletismo castellón playas continente	competición mejores europa categoría masculina femenina disputará copa clubes atletismo castellón playas continente	competición reúne mejores equipos europa categoría masculina femenina disputará gaetà copa clubes atletismo castellón conocer asociación europea playas entidad organizadora instalaciones gaeta huguet escenario competirán continente anfitrión aspirantes título valencia esports días mayo ofrecerá oportunidad amantes atletas	competición mejores europa categoría masculina femenina disputará copa clubes atletismo castellón playas continente
Tr= {T ∈ Texto/ T >(M+DP)}				competición reúne equipos categoría masculina femenina gaetà copa clubes castellón conocer asociación europea playas entidad organizadora instalaciones gaeta huguet escenario competirán continente anfitrión aspirantes título valencia esports días mayo ofrecerá oportunidad amantes atletas	

Tabela A.5: Granularidade dos conjuntos para Bigram do texto 3, sem stopwords

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T $\in$ Texto/ T >(M-DP)}	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del

Tabela A.5 Granularidade dos conjuntos para Bigram do texto 3, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >M}	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del

Tabela A.5 Granularidade dos conjuntos para Bigram do texto 3, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
Tr= {T ∈ Texto/ T >(M+DP)}	la competición de europa en categoría categoría masculina se disputará disputará en copa de de atletismo los mejores del continente	la competición de europa en categoría categoría masculina se disputará disputará en copa de de atletismo los mejores del continente	la competición de europa en categoría categoría masculina se disputará disputará en copa de de atletismo los mejores del continente	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de de europa europa en en categoría categoría masculina masculina y y femenina femenina se se disputará disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del	la competición competición que que reúne reúne a ocho mejores mejores equipos equipos de europa en en categoría categoría masculina masculina y y femenina femenina se disputará en en gaetà gaetà copa copa de europa de de clubes clubes de de atletismo atletismo se en castellón castellón según a conocer conocer la la asociación asociación europea europea de atletismo y al playas playas de de castellón castellón como como entidad entidad organizadora organizadora las las instalaciones instalaciones de de gaeta gaeta huguet huguet serán el escenario escenario en que competirán competirán los los mejores mejores ocho ocho clubes clubes del del continente continente tanto masculina como como femenina femenina con el playas playas como como anfitrión anfitrión y los aspirantes aspirantes al al título título al el valencia valencia esports esports en competición copa europa se disputará los los días días y de mayo mayo y y ofrecerá ofrecerá la la oportunidad oportunidad para los amantes amantes del del atletismo atletismo puedan mejores atletas atletas del

Tabela A.6: Granularidade dos conjuntos para Trigram do texto 3, sem stopwords

Métricas/ AAM	PC	BB	Co	GI	LP
	la competición que	la competición que	la competición que	la competición que	la competición que
	competición que reúne	competición que reúne	competición que reúne	competición que reúne	competición que reúne
	que reúne a	que reúne a	que reúne a	que reúne a	que reúne a
	reúne a los	reúne a los	reúne a los	reúne a los	reúne a los
	los ocho mejores	los ocho mejores	los ocho mejores	los ocho mejores	los ocho mejores
	ocho mejores equipos	ocho mejores equipos	ocho mejores equipos	ocho mejores equipos	ocho mejores equipos
	mejores equipos de	mejores equipos de	mejores equipos de	mejores equipos de	mejores equipos de
	equipos de europa	equipos de europa	equipos de europa	equipos de europa	equipos de europa
	de europa en	de europa en	de europa en	de europa en	de europa en
	europa en categoría	europa en categoría	europa en categoría	europa en categoría	europa en categoría
				en categoría masculina	
	categoría masculina y	categoría masculina y	categoría masculina y	categoría masculina y	categoría masculina y
	masculina y femenina	masculina y femenina	masculina y femenina	masculina y femenina	masculina y femenina
	y femenina se	y femenina se	y femenina se	y femenina se	y femenina se
	femenina se disputará	femenina se disputará	femenina se disputará	femenina se disputará	femenina se disputará
				se disputará en	
	disputará en gaetà	disputará en gaetà	disputará en gaetà	disputará en gaetà	disputará en gaetà
	en gaetà copa	en gaetà copa	en gaetà copa	en gaetà copa	en gaetà copa
	gaetà copa de	gaetà copa de	gaetà copa de	gaetà copa de	gaetà copa de
				copa de europa	
	de europa de	de europa de	de europa de	de europa de	de europa de
	europa de clubes	europa de clubes	europa de clubes	europa de clubes	europa de clubes
	de clubes de	de clubes de	de clubes de	de clubes de	de clubes de
	clubes de atletismo	clubes de atletismo	clubes de atletismo	clubes de atletismo	clubes de atletismo
	de atletismo se	de atletismo se	de atletismo se	de atletismo se	de atletismo se
	atletismo se disputargaetá	atletismo se disputargaetá	atletismo se disputargaetá	atletismo se disputargaetá	atletismo se disputargaetá
	disputará en en	disputará en en	disputará en en	disputará en en	disputará en en
	en en castellón	en en castellón	en en castellón	en en castellón	en en castellón
	en castellón según	en castellón según	en castellón según	en castellón según	en castellón según
	castellón según ha	castellón según ha	castellón según ha	castellón según ha	castellón según ha
	dado a conocer	dado a conocer	dado a conocer	dado a conocer	dado a conocer
	a conocer la	a conocer la	a conocer la	a conocer la	a conocer la
	conocer la asociación	conocer la asociación	conocer la asociación	conocer la asociación	conocer la asociación
	la asociación europea	la asociación europea	la asociación europea	la asociación europea	la asociación europea
	asociación europea de	asociación europea de	asociación europea de	asociación europea de	asociación europea de
	europea de atletismo	europea de atletismo	europea de atletismo	europea de atletismo	europea de atletismo
	de atletismo y	de atletismo y	de atletismo y	de atletismo y	de atletismo y
	atletismo y tendrá	atletismo y tendrá	atletismo y tendrá	atletismo y tendrá	atletismo y tendrá
	tendrá al playas	tendrá al playas	tendrá al playas	tendrá al playas	tendrá al playas
	al playas de	al playas de	al playas de	al playas de	al playas de
	playas de castellón	playas de castellón	playas de castellón	playas de castellón	playas de castellón
	de castellón como	de castellón como	de castellón como	de castellón como	de castellón como
	castellón como entidad	castellón como entidad	castellón como entidad	castellón como entidad	castellón como entidad
	como entidad organizadora	como entidad organizadora	como entidad organizadora	como entidad organizadora	como entidad organizadora
	entidad organizadora las	entidad organizadora las	entidad organizadora las	entidad organizadora las	entidad organizadora las
	organizadora las instalaciones	organizadora las instalaciones	organizadora las instalaciones	organizadora las instalaciones	organizadora las instalaciones
	las instalaciones de	las instalaciones de	las instalaciones de	las instalaciones de	las instalaciones de
	instalaciones de gaeta	instalaciones de gaeta	instalaciones de gaeta	instalaciones de gaeta	instalaciones de gaeta
	de gaeta huguet	de gaeta huguet	de gaeta huguet	de gaeta huguet	de gaeta huguet
	gaeta huguet serán	gaeta huguet serán	gaeta huguet serán	gaeta huguet serán	gaeta huguet serán
	huguet serán el	huguet serán el	huguet serán el	huguet serán el	huguet serán el
	serán el escenario	serán el escenario	serán el escenario	serán el escenario	serán el escenario
	el escenario en	el escenario en	el escenario en	el escenario en	el escenario en
{T >(M-DP)}	escenario en el	escenario en el	escenario en el	escenario en el	escenario en el
	el que competirán	el que competirán	el que competirán	el que competirán	el que competirán
	que competirán los	que competirán los	que competirán los	que competirán los	que competirán los
	competirán los mejores	competirán los mejores	competirán los mejores	competirán los mejores	competirán los mejores
	los mejores ocho	los mejores ocho	los mejores ocho	los mejores ocho	los mejores ocho
	mejores ocho clubes	mejores ocho clubes	mejores ocho clubes	mejores ocho clubes	mejores ocho clubes
	ocho clubes del	ocho clubes del	ocho clubes del	ocho clubes del	ocho clubes del
	clubes del continente	clubes del continente	clubes del continente	clubes del continente	clubes del continente
	del continente tanto	del continente tanto	del continente tanto	del continente tanto	del continente tanto
	continente tanto en	continente tanto en	continente tanto en	continente tanto en	continente tanto en
	tanto en categoría	tanto en categoría	tanto en categoría	tanto en categoría	tanto en categoría
	categoría masculina como	categoría masculina como	categoría masculina como	categoría masculina como	categoría masculina como
	masculina como femenina	masculina como femenina	masculina como femenina	masculina como femenina	masculina como femenina
	como femenina con	como femenina con	como femenina con	como femenina con	como femenina con
	femenina con el	femenina con el	femenina con el	femenina con el	femenina con el
	con el playas	con el playas	con el playas	con el playas	con el playas
	el playas como	el playas como	el playas como	el playas como	el playas como
	playas como anfitrión	playas como anfitrión	playas como anfitrión	playas como anfitrión	playas como anfitrión
	como anfitrión y	como anfitrión y	como anfitrión y	como anfitrión y	como anfitrión y
	anfitrión y uno	anfitrión y uno	anfitrión y uno	anfitrión y uno	anfitrión y uno
	de los aspirantes	de los aspirantes	de los aspirantes	de los aspirantes	de los aspirantes
	los aspirantes al	los aspirantes al	los aspirantes al	los aspirantes al	los aspirantes al
	aspirantes al título	aspirantes al título	aspirantes al título	aspirantes al título	aspirantes al título
	al título al	al título al	al título al	al título al	al título al
	título al igual	título al igual	título al igual	título al igual	título al igual
	que el valencia	que el valencia	que el valencia	que el valencia	que el valencia
	el valencia esports	el valencia esports	el valencia esports	el valencia esports	el valencia esports
	valencia esports en	valencia esports en	valencia esports en	valencia esports en	valencia esports en
	esports en la	esports en la	esports en la	esports en la	esports en la
	en la competición	en la competición	en la competición	en la competición	en la competición
	la competición copa	la competición copa	la competición copa	la competición copa	la competición copa
	competición copa de	competición copa de	competición copa de	competición copa de	competición copa de
	de europa se	de europa se	de europa se	de europa se	de europa se
	europa se disputará	europa se disputará	europa se disputará	europa se disputará	europa se disputará
	se disputará los	se disputará los	se disputará los	se disputará los	se disputará los
	disputará los días	disputará los días	disputará los días	disputará los días	disputará los días
	los días y	los días y	los días y	los días y	los días y
	días y de	días y de	días y de	días y de	días y de
	y de mayo	y de mayo	y de mayo	y de mayo	y de mayo
	de mayo y	de mayo y	de mayo y	de mayo y	de mayo y
	mayo y ofrecerá	mayo y ofrecerá	mayo y ofrecerá	mayo y ofrecerá	mayo y ofrecerá
	y ofrecerá la	y ofrecerá la	y ofrecerá la	y ofrecerá la	y ofrecerá la
	ofrecerá la oportunidad	ofrecerá la oportunidad	ofrecerá la oportunidad	ofrecerá la oportunidad	ofrecerá la oportunidad
	la oportunidad para	la oportunidad para	la oportunidad para	la oportunidad para	la oportunidad para
	oportunidad para que	oportunidad para que	oportunidad para que	oportunidad para que	oportunidad para que
	que los amantes	que los amantes	que los amantes	que los amantes	que los amantes
	los amantes del	los amantes del	los amantes del	los amantes del	los amantes del
	amantes del atletismo	amantes del atletismo	amantes del atletismo	amantes del atletismo	amantes del atletismo
	del atletismo puedan	del atletismo puedan	del atletismo puedan	del atletismo puedan	del atletismo puedan
	atletismo puedan ver	atletismo puedan ver	atletismo puedan ver	atletismo puedan ver	atletismo puedan ver
	a los mejores	a los mejores	a los mejores	a los mejores	a los mejores
	los mejores atletas	los mejores atletas	los mejores atletas	los mejores atletas	los mejores atletas
	mejores atletas del	mejores atletas del	mejores atletas del	mejores atletas del	mejores atletas del
	atletas del continente	atletas del continente	atletas del continente	atletas del continente	atletas del continente

Tabela A.6 Granularidade dos conjuntos para Trigram do texto 3, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
	la competición que	la competición que	la competición que	la competición que	la competición que
	competición que reúne	competición que reúne	competición que reúne	competición que reúne	competición que reúne
	que reúne a	que reúne a	que reúne a	que reúne a	que reúne a
	reúne a los	reúne a los	reúne a los	reúne a los	reúne a los
	los ocho mejores	los ocho mejores	los ocho mejores	los ocho mejores	los ocho mejores
	ocho mejores equipos	ocho mejores equipos	ocho mejores equipos	ocho mejores equipos	ocho mejores equipos
	mejores equipos de	mejores equipos de	mejores equipos de	mejores equipos de	mejores equipos de
	equipos de europa	equipos de europa	equipos de europa	equipos de europa	equipos de europa
	de europa en	de europa en	de europa en	de europa en	de europa en
	europa en categoría	europa en categoría	europa en categoría	europa en categoría	europa en categoría
				en categoría masculina	
	categoría masculina y	categoría masculina y	categoría masculina y	categoría masculina y	categoría masculina y
	masculina y femenina	masculina y femenina	masculina y femenina	masculina y femenina	masculina y femenina
	y femenina se	y femenina se	y femenina se	y femenina se	y femenina se
	femenina se disputará	femenina se disputará	femenina se disputará	femenina se disputará	femenina se disputará
				se disputará en	
	disputará en gaetà	disputará en gaetà	disputará en gaetà	disputará en gaetà	disputará en gaetà
	en gaetà copa	en gaetà copa	en gaetà copa	en gaetà copa	en gaetà copa
	gaetà copa de	gaetà copa de	gaetà copa de	gaetà copa de	gaetà copa de
				copa de europa	
	de europa de	de europa de	de europa de	de europa de	de europa de
	europa de clubes	europa de clubes	europa de clubes	europa de clubes	europa de clubes
	de clubes de	de clubes de	de clubes de	de clubes de	de clubes de
	clubes de atletismo	clubes de atletismo	clubes de atletismo	clubes de atletismo	clubes de atletismo
	de atletismo se	de atletismo se	de atletismo se	de atletismo se	de atletismo se
	atletismo se disputargaetá	atletismo se disputargaetá	atletismo se disputargaetá	atletismo se disputargaetá	atletismo se disputargaetá
	disputará en en	disputará en en	disputará en en	disputará en en	disputará en en
	en en castellón	en en castellón	en en castellón	en en castellón	en en castellón
	en castellón segón	en castellón segón	en castellón segón	en castellón segón	en castellón segón
	castellón según ha	castellón según ha	castellón según ha	castellón según ha	castellón según ha
	dado a conocer	dado a conocer	dado a conocer	dado a conocer	dado a conocer
	a conocer la	a conocer la	a conocer la	a conocer la	a conocer la
	conocer la asociación	conocer la asociación	conocer la asociación	conocer la asociación	conocer la asociación
	la asociación europea	la asociación europea	la asociación europea	la asociación europea	la asociación europea
	asociación europea de	asociación europea de	asociación europea de	asociación europea de	asociación europea de
	europea de atletismo	europea de atletismo	europea de atletismo	europea de atletismo	europea de atletismo
	de atletismo y	de atletismo y	de atletismo y	de atletismo y	de atletismo y
	atletismo y tendrá	atletismo y tendrá	atletismo y tendrá	atletismo y tendrá	atletismo y tendrá
	tendrá al playas	tendrá al playas	tendrá al playas	tendrá al playas	tendrá al playas
	al playas de	al playas de	al playas de	al playas de	al playas de
	playas de castellón	playas de castellón	playas de castellón	playas de castellón	playas de castellón
	de castellón como	de castellón como	de castellón como	de castellón como	de castellón como
	castellón como entidad	castellón como entidad	castellón como entidad	castellón como entidad	castellón como entidad
	como entidad organizadora	como entidad organizadora	como entidad organizadora	como entidad organizadora	como entidad organizadora
	entidad organizadora las	entidad organizadora las	entidad organizadora las	entidad organizadora las	entidad organizadora las
	organizadora las instalaciones	organizadora las instalaciones	organizadora las instalaciones	organizadora las instalaciones	organizadora las instalaciones
	las instalaciones de	las instalaciones de	las instalaciones de	las instalaciones de	las instalaciones de
	instalaciones de gaeta	instalaciones de gaeta	instalaciones de gaeta	instalaciones de gaeta	instalaciones de gaeta
	de gaeta huguet	de gaeta huguet	de gaeta huguet	de gaeta huguet	de gaeta huguet
	gaeta huguet serán	gaeta huguet serán	gaeta huguet serán	gaeta huguet serán	gaeta huguet serán
	huguet serán el	huguet serán el	huguet serán el	huguet serán el	huguet serán el
	serán el escenario	serán el escenario	serán el escenario	serán el escenario	serán el escenario
	el escenario en	el escenario en	el escenario en	el escenario en	el escenario en
{ T >M}	escenario en el	escenario en el	escenario en el	escenario en el	escenario en el
	el que competirán	el que competirán	el que competirán	el que competirán	el que competirán
	que competirán los	que competirán los	que competirán los	que competirán los	que competirán los
	competirán los mejores	competirán los mejores	competirán los mejores	competirán los mejores	competirán los mejores
	los mejores ocho	los mejores ocho	los mejores ocho	los mejores ocho	los mejores ocho
	mejores ocho clubes	mejores ocho clubes	mejores ocho clubes	mejores ocho clubes	mejores ocho clubes
	ocho clubes del	ocho clubes del	ocho clubes del	ocho clubes del	ocho clubes del
	clubes del continente	clubes del continente	clubes del continente	clubes del continente	clubes del continente
	del continente tanto	del continente tanto	del continente tanto	del continente tanto	del continente tanto
	continente tanto en	continente tanto en	continente tanto en	continente tanto en	continente tanto en
	tanto en categoría	tanto en categoría	tanto en categoría	tanto en categoría	tanto en categoría
	categoría masculina como	categoría masculina como	categoría masculina como	categoría masculina como	categoría masculina como
	masculina como femenina	masculina como femenina	masculina como femenina	masculina como femenina	masculina como femenina
	como femenina con	como femenina con	como femenina con	como femenina con	como femenina con
	femenina con el	femenina con el	femenina con el	femenina con el	femenina con el
	con el playas	con el playas	con el playas	con el playas	con el playas
	el playas como	el playas como	el playas como	el playas como	el playas como
	playas como anfitrión	playas como anfitrión	playas como anfitrión	playas como anfitrión	playas como anfitrión
	como anfitrión y	como anfitrión y	como anfitrión y	como anfitrión y	como anfitrión y
	anfitrión y uno	anfitrión y uno	anfitrión y uno	anfitrión y uno	anfitrión y uno
	de los aspirantes	de los aspirantes	de los aspirantes	de los aspirantes	de los aspirantes
	los aspirantes al	los aspirantes al	los aspirantes al	los aspirantes al	los aspirantes al
	aspirantes al título	aspirantes al título	aspirantes al título	aspirantes al título	aspirantes al título
	al título al	al título al	al título al	al título al	al título al
	título al igual	título al igual	título al igual	título al igual	título al igual
	que el valencia	que el valencia	que el valencia	que el valencia	que el valencia
	el valencia esports	el valencia esports	el valencia esports	el valencia esports	el valencia esports
	valencia esports en	valencia esports en	valencia esports en	valencia esports en	valencia esports en
	esports en la	esports en la	esports en la	esports en la	esports en la
	en la competición	en la competición	en la competición	en la competición	en la competición
	la competición copa	la competición copa	la competición copa	la competición copa	la competición copa
	competición copa de	competición copa de	competición copa de	competición copa de	competición copa de
	de europa se	de europa se	de europa se	de europa se	de europa se
	europa se disputará	europa se disputará	europa se disputará	europa se disputará	europa se disputará
	se disputará los	se disputará los	se disputará los	se disputará los	se disputará los
	disputará los días	disputará los días	disputará los días	disputará los días	disputará los días
	los días y	los días y	los días y	los días y	los días y
	días y de	días y de	días y de	días y de	días y de
	y de mayo	y de mayo	y de mayo	y de mayo	y de mayo
	de mayo y	de mayo y	de mayo y	de mayo y	de mayo y
	mayo y ofrecerá	mayo y ofrecerá	mayo y ofrecerá	mayo y ofrecerá	mayo y ofrecerá
	y ofrecerá la	y ofrecerá la	y ofrecerá la	y ofrecerá la	y ofrecerá la
	ofrecerá la oportunidad	ofrecerá la oportunidad	ofrecerá la oportunidad	ofrecerá la oportunidad	ofrecerá la oportunidad
	la oportunidad para	la oportunidad para	la oportunidad para	la oportunidad para	la oportunidad para
	oportunidad para que	oportunidad para que	oportunidad para que	oportunidad para que	oportunidad para que
	que los amantes	que los amantes	que los amantes	que los amantes	que los amantes
	los amantes del	los amantes del	los amantes del	los amantes del	los amantes del
	amantes del atletismo	amantes del atletismo	amantes del atletismo	amantes del atletismo	amantes del atletismo
	del atletismo puedan	del atletismo puedan	del atletismo puedan	del atletismo puedan	del atletismo puedan
	atletismo puedan ver	atletismo puedan ver	atletismo puedan ver	atletismo puedan ver	atletismo puedan ver
	a los mejores	a los mejores	a los mejores	a los mejores	a los mejores
	los mejores atletas	los mejores atletas	los mejores atletas	los mejores atletas	los mejores atletas
	mejores atletas del	mejores atletas del	mejores atletas del	mejores atletas del	mejores atletas del
	atletas del continente	atletas del continente	atletas del continente	atletas del continente	atletas del continente

Tabela A.6 Granularidade dos conjuntos para Trigram do texto 3, sem stopwords (continuação)

Métricas/ AAM	PC	BB	Co	GI	LP
{T >(M+DP)}	la competición que competición que reúne que reúne a reúne a los los ocho mejores ocho mejores equipos mejores equipos de equipos de europa de europa en europa en categoría categoría masculina y masculina y femenina y femenina se femenina se disputará disputará en gaetà en gaetà copa gaetà copa de de europa de europa de clubes de clubes de clubes de atletismo de atletismo se atletismo se disputargaetá disputará en en en en castellón en castellón segón castellón según ha dado a conocer a conocer la conocer la asociación la asociación europea asociación europea de europea de atletismo de atletismo y atletismo y tendrá tendrá al playas al playas de playas de castellón de castellón como castellón como entidad como entidad organizadora entidad organizadora las organizadora las instalaciones las instalaciones de instalaciones de gaeta de gaeta huguet gaeta huguet serán huguet serán el serán el escenario el escenario en escenario en el el que competirán que competirán los competirán los mejores los mejores ocho mejores ocho clubes ocho clubes del clubes del continente del continente tanto continente tanto en tanto en categoría categoría masculina como masculina como femenina como femenina con femenina con el con el playas el playas como playas como anfitrión como anfitrión y anfitrión y uno de los aspirantes los aspirantes al aspirantes al título al título al título al igual que el valencia el valencia esports valencia esports en esports en la en la competición la competición copa competición copa de de europa se europa se disputará se disputará los disputará los días los días y días y de y de mayo de mayo y mayo y ofrecerá y ofrecerá la ofrecerá la oportunidad la oportunidad para oportunidad para que que los amantes los amantes del amantes del atletismo del atletismo puedan atletismo puedan ver a los mejores los mejores atletas mejores atletas del atletas del continente	la competición que competición que reúne que reúne a reúne a los los ocho mejores ocho mejores equipos mejores equipos de equipos de europa de europa en europa en categoría categoría masculina y masculina y femenina y femenina se femenina se disputará disputará en gaetà en gaetà copa gaetà copa de de europa de europa de clubes de clubes de clubes de atletismo de atletismo se atletismo se disputargaetá disputará en en en en castellón en castellón segón castellón según ha dado a conocer a conocer la conocer la asociación la asociación europea asociación europea de europea de atletismo de atletismo y atletismo y tendrá tendrá al playas al playas de playas de castellón de castellón como castellón como entidad como entidad organizadora entidad organizadora las organizadora las instalaciones las instalaciones de instalaciones de gaeta de gaeta huguet gaeta huguet serán huguet serán el serán el escenario el escenario en escenario en el el que competirán que competirán los competirán los mejores los mejores ocho mejores ocho clubes ocho clubes del clubes del continente del continente tanto continente tanto en tanto en categoría categoría masculina como masculina como femenina como femenina con femenina con el con el playas el playas como playas como anfitrión como anfitrión y anfitrión y uno de los aspirantes los aspirantes al aspirantes al título al título al título al igual que el valencia el valencia esports valencia esports en esports en la en la competición la competición copa competición copa de de europa se europa se disputará se disputará los disputará los días los días y días y de y de mayo de mayo y mayo y ofrecerá y ofrecerá la ofrecerá la oportunidad la oportunidad para oportunidad para que que los amantes los amantes del amantes del atletismo del atletismo puedan atletismo puedan ver a los mejores los mejores atletas mejores atletas del atletas del continente	la competición que competición que reúne que reúne a reúne a los los ocho mejores ocho mejores equipos mejores equipos de equipos de europa de europa en europa en categoría categoría masculina y masculina y femenina y femenina se femenina se disputará disputará en gaetà en gaetà copa gaetà copa de de europa de europa de clubes de clubes de clubes de atletismo de atletismo se atletismo se disputargaetá disputará en en en en castellón en castellón segón castellón según ha dado a conocer a conocer la conocer la asociación la asociación europea asociación europea de europea de atletismo de atletismo y atletismo y tendrá tendrá al playas al playas de playas de castellón de castellón como castellón como entidad como entidad organizadora entidad organizadora las organizadora las instalaciones las instalaciones de instalaciones de gaeta de gaeta huguet gaeta huguet serán huguet serán el serán el escenario el escenario en escenario en el el que competirán que competirán los competirán los mejores los mejores ocho mejores ocho clubes ocho clubes del clubes del continente del continente tanto continente tanto en tanto en categoría categoría masculina como masculina como femenina como femenina con femenina con el con el playas el playas como playas como anfitrión como anfitrión y anfitrión y uno de los aspirantes los aspirantes al aspirantes al título al título al título al igual que el valencia el valencia esports valencia esports en esports en la en la competición la competición copa competición copa de de europa se europa se disputará se disputará los disputará los días los días y días y de y de mayo de mayo y mayo y ofrecerá y ofrecerá la ofrecerá la oportunidad la oportunidad para oportunidad para que que los amantes los amantes del amantes del atletismo del atletismo puedan atletismo puedan ver a los mejores los mejores atletas mejores atletas del atletas del continente	la competición que competición que reúne que reúne a reúne a los los ocho mejores ocho mejores equipos mejores equipos de equipos de europa de europa en europa en categoría en categoría masculina categoría masculina y masculina y femenina y femenina se femenina se disputará se disputará en disputará en gaetà en gaetà copa gaetà copa de copa de europa de europa de europa de clubes de clubes de clubes de atletismo de atletismo se atletismo se disputargaetá disputará en en en en castellón en castellón segón castellón según ha dado a conocer a conocer la conocer la asociación la asociación europea asociación europea de europea de atletismo de atletismo y atletismo y tendrá tendrá al playas al playas de playas de castellón de castellón como castellón como entidad como entidad organizadora entidad organizadora las organizadora las instalaciones las instalaciones de instalaciones de gaeta de gaeta huguet gaeta huguet serán huguet serán el serán el escenario el escenario en escenario en el el que competirán que competirán los competirán los mejores los mejores ocho mejores ocho clubes ocho clubes del clubes del continente del continente tanto continente tanto en tanto en categoría categoría masculina como masculina como femenina como femenina con femenina con el con el playas el playas como playas como anfitrión como anfitrión y anfitrión y uno de los aspirantes los aspirantes al aspirantes al título al título al título al igual que el valencia el valencia esports valencia esports en esports en la en la competición la competición copa competición copa de de europa se europa se disputará se disputará los disputará los días los días y días y de y de mayo de mayo y mayo y ofrecerá y ofrecerá la ofrecerá la oportunidad la oportunidad para oportunidad para que que los amantes los amantes del amantes del atletismo del atletismo puedan atletismo puedan ver a los mejores los mejores atletas mejores atletas del atletas del continente	la competición que competición que reúne que reúne a reúne a los los ocho mejores ocho mejores equipos mejores equipos de equipos de europa de europa en europa en categoría categoría masculina y masculina y femenina y femenina se femenina se disputará disputará en gaetà en gaetà copa gaetà copa de de europa de europa de clubes de clubes de clubes de atletismo de atletismo se atletismo se disputargaetá disputará en en en en castellón en castellón segón castellón según ha dado a conocer a conocer la conocer la asociación la asociación europea asociación europea de europea de atletismo de atletismo y atletismo y tendrá tendrá al playas al playas de playas de castellón de castellón como castellón como entidad como entidad organizadora entidad organizadora las organizadora las instalaciones las instalaciones de instalaciones de gaeta de gaeta huguet gaeta huguet serán huguet serán el serán el escenario el escenario en escenario en el el que competirán que competirán los competirán los mejores los mejores ocho mejores ocho clubes ocho clubes del clubes del continente del continente tanto continente tanto en tanto en categoría categoría masculina como masculina como femenina como femenina con femenina con el con el playas el playas como playas como anfitrión como anfitrión y anfitrión y uno de los aspirantes los aspirantes al aspirantes al título al título al título al igual que el valencia el valencia esports valencia esports en esports en la en la competición la competición copa competición copa de de europa se europa se disputará se disputará los disputará los días los días y días y de y de mayo de mayo y mayo y ofrecerá y ofrecerá la ofrecerá la oportunidad la oportunidad para oportunidad para que que los amantes los amantes del amantes del atletismo del atletismo puedan atletismo puedan ver a los mejores los mejores atletas mejores atletas del atletas del continente