UNIVERSIDADE DA BEIRA INTERIOR
Faculdade de Engenharia

Similaridade Documental e Detecção de Plágio

Henrique da Costa Mendes

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

Orientador: Prof. Doutor João Paulo da Costa Cordeiro

Covilhã, Outubro de 2013

Agradecimentos

Antes demais gostaria de agradecer ao meu orientador e professor Dr. João Paulo da Costa Cordeiro, por todo o apoio, disponibilidade, força e ajuda que me prestou antes e durante o desenvolvimento da dissertação de mestrado.

Agradeço ainda a todos os Docentes que me leccionaram aulas inicialmente na minha licenciatura e posteriormente no primeiro ano de mestrado.

Não podia esquecer a instituição de ensino pública a Universidade da Beira Interior visto que foi nesta instituição que se deu o meu desenvolvimento como pessoa e profissional, assim como ao Centro de Tecnologia da Linguagem Humana e Bioinformática (HULTIG), laboratório no qual tive a oportunidade de desenvolver a minha dissertação.

Por final e não com menos importância agradeço do fundo do meu coração aos meus pais Luís Filipe Dias Mendes e Isabel Maria da Costa Pais Mendes estiveram desde sempre ao meu lado e também à minha namorada e amigos que se tornaram importantes no meu caminho para o sucesso.

A todos os restantes e não mencionados o meu profundo agradecimento.

Henrique Mendes

Resumo

Nesta dissertação de mestrado, apresentamos um estudo em duas áreas, a Similaridade Documental e a Detecção Automática de Plágio. Começamos por fazer uma breve introdução no qual explicamos alguns conceitos importantes, passamos para um estudo do que outros autores fizeram nestas áreas e após este estudo e aproveitando bastantes aspetos, expomos alguns dos caminhos abordados, bons ou maus, até atingirmos o melhor método. Um método que fosse eficaz, rápido e eficiente na detecção de plágio, por fim apresentamos a nossa solução, os testes efectuados, e os resultados obtidos que nos ajudaram a alcançar o objectivo final.

Palavras-Chave

Pesquisa de Informação, Similaridade Documental, Medidas de Similaridade, Detecção Automática de Plágio.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos

PInf — Pesquisa de Informação

SD — Similaridade Documental

DAP — Detecção Automática de Plágio

PAN — Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

TFIDF — Term Frequency Inverse Document Frequency

TF ou tf — Term Frequency

IMA — Importância Mínima Admitida

XML — eXtensible Markup Language

1. Introdução

1.1 Motivação

Esta dissertação tem como motivação fazer um estudo na área de Pesquisa de Informação (PInf) [1] nomeadamente nas áreas da Similaridade Documental (SD) e na Detecção Automática de Plágio (DAP). A SD não difere muito da PInf, pois a PInf preocupa-se em "dar respostas" a consultas efectuadas pelos utilizadores, um exemplo conhecido por todos é o motor de busca Google, mas também podemos considerar o Bing ou Yahoo, outros motores no qual são inseridas palavras-chave e como resposta são filtrados os resultados que mais se aproximam das palavras chave como podemos observar na Figura 1.

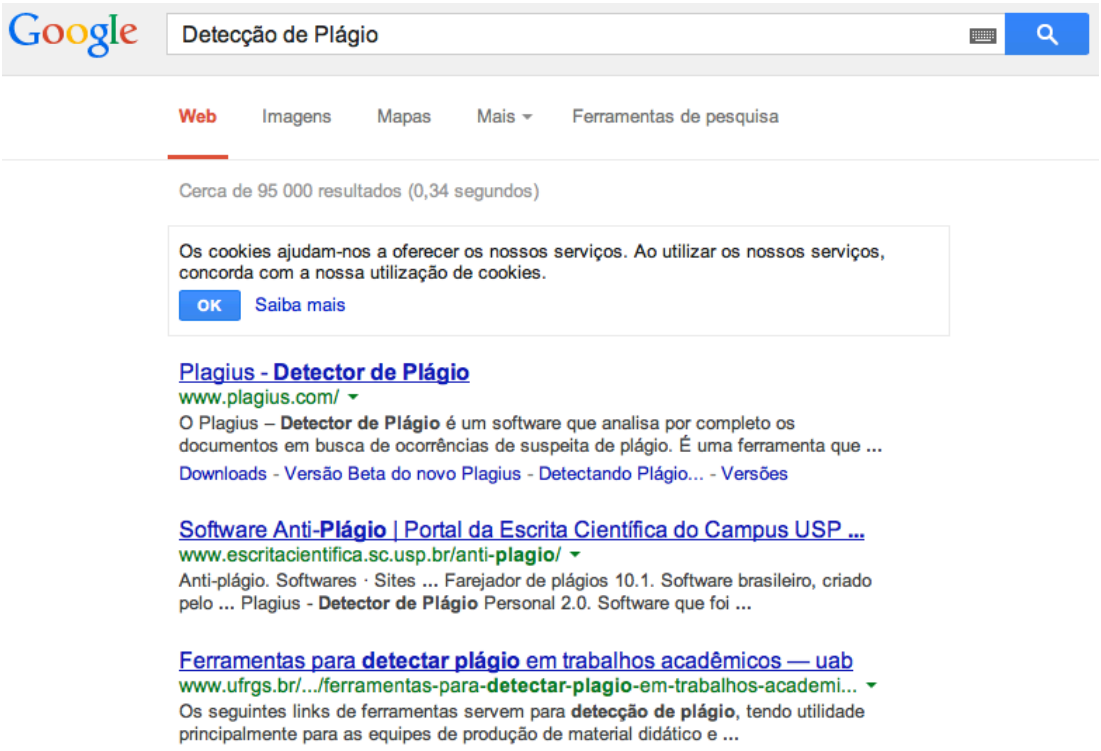

Figura 1 - Exemplo de uma pesquisa no motor de busca Google

De uma forma mais geral pretende-se a partir de uma grande colecção de informação recolher o conteúdo que mais se aproxime ao que se pretende. Se no caso de palavras-chave fosse utilizado um documento para pesquisa então estaríamos a entrar no domínio da SD — Figura 2.

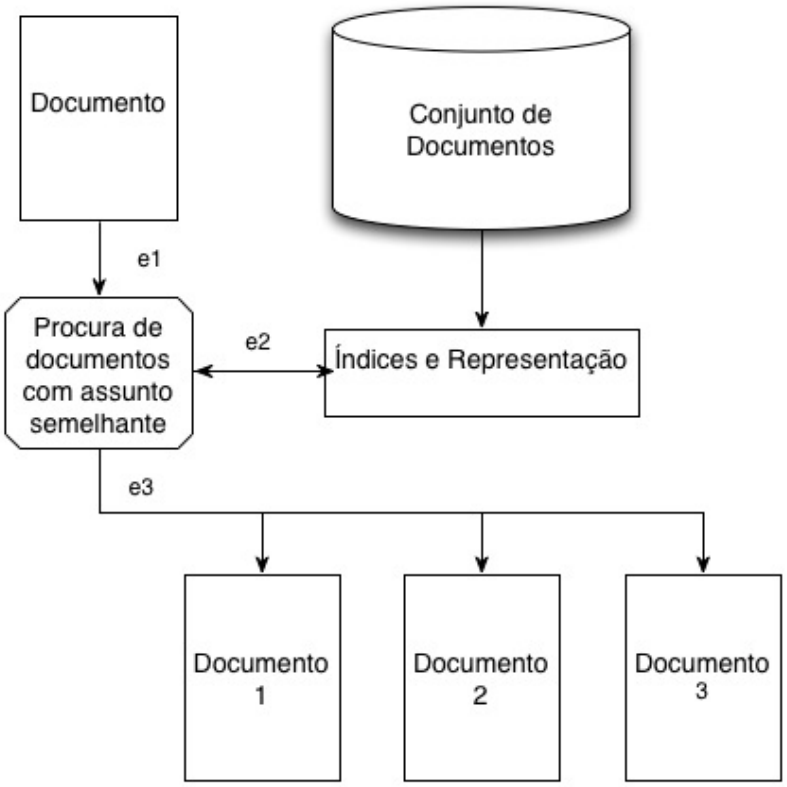

Figura 2 – Esquema geral da pesquisa em Similaridade Documental

Na Figura 2 exemplificamos por etapas no domínio da SD, na etapa e1 o documento é passado ao módulo de pesquisa de documentos que por sua vez interage com o Índice de Representação e este actua sobre o Conjunto de Documentos para receber como resposta à sua consulta a identificação dos documentos com conteúdos semelhantes (etapa e2), por fim, na etapa e3 é apresentado o conjunto de documentos que contêm algum conteúdo semelhante com o documento inserido na etapa 1.

A confrontação de um par de documentos permite recolher várias informações, por exemplo tendo em conta o padrão de escrita, o conteúdo representado no texto, entre outros. Aplicando a SD, permite a classificação de documentos incidindo sobre o conteúdo destes e consequentemente determinar se um par de documentos é sobre o mesmo assunto, ou não.

A principal motivação da dissertação enquadra-se numa área que emergiu recentemente (depois de 2008) e que ultimamente tem estado muito na ordem do dia, sendo até um assunto mediático com alguma frequência — a Detecção Automática de Plágio (DAP). Para podermos compreender melhor a DAP recorremos a dois exemplos constituídos por dois pares de frases:

1º Par de Frase: *"Economia Portuguesa destrói emprego há 16 trimestres consecutivos"* e *"Economia Portuguesa elimina emprego há 16 trimestres consecutivos"* , este seria uma caso flagrante de plágio em que nas frases pouco se alterou.

2º Par de Frases: *"O Presidente da Venezuela morreu terça-feira na sequência de um cancro" e "Hugo Chávez, morreu nesta terça-feira, aos 58 anos, num hospital militar em Caracas,"*, ambas as frases com conteúdo semelhante poderiam ou não apresentar plágio. Neste caso teríamos uma análise mais complexa de DAP.

Ao referir que a DAP é um assunto mediático que tem surgido com alguma frequência nos últimos tempos podemos colocar a nossa atenção em dois casos de anulação de teses de doutoramento devido à existência de plágio por parte do Ministro da Defesa da Alemanha, Karl-Theodor zu Guttenberg[1] e também da Ministra da Educação, Annete Schavan[2].

[1] url: http://terramagazine.terra.com.br/interna/0,,OI4968530-EI6580,00.html, consultado em Março de 2013

[2] url: http://www.tvi24.iol.pt/503/internacional/annette-schavan-plagio-alemanha-ministra-educacao-tvi24/1417102-4073.html, consultado em Março de 2013

Um outro contexto em que podemos observar a aplicação da DAP é junto às instituições educacionais [2] onde podemos verificar preocupações elevadas em combater-se o plágio. Universidades Americanas como Stanford, Yale, Berkeley, MIT [2] e mesmo na Europa, como por exemplo Cambridge e Oxford, aplicam elevadas punições a quem praticar plágio, dentro do meio académico, como relatórios, trabalhos, apresentações, entre outros [2].

A partir de 2009 uma conferencia — PAN[3] — tem vindo a ganhar mais significado tornando-se a mais importante da actualidade no domínio da detecção automática de plágio. Esta conferência é uma competição constituída por equipas, na qual são propostas técnicas e experiências para a DAP. Aos participantes desta conferência é fornecido um Corpus (conjunto de documentos) com elevado número de documentos e é sobre este Corpus que as equipas participantes apresentam as suas propostas, por vezes o Corpus pode conter documentos de diversas línguas, um parâmetro que tem que ser tomado em conta para se conseguirem os resultados pretendidos. O principal objectivo dos participantes é atingirem os melhores resultados segundo alguns parâmetros pré-definidos, como por exemplo a eficácia e eficiência dos algoritmos, quanto ao tempo de processamento e aos valores obtidos em diversas medidas (*Recall*, *Precision*, entre outras). Um dos paradigmas que se torna um dos mais importantes refere-se às questões de eficiência em tempo de execução dos algoritmos, visto que o Corpus é de tamanho elevando, logo requer gastos computacionais muito grandes, obrigando a ser uma das preocupações a redução de recursos computacionais elevados. Com o passar dos anos e ao ganhar uma maior evidência esta conferência tem despertado o interesse cada vez mais ao público alvo e tem crescido substancialmente o número de equipas participantes desempenhando papéis mais importantes na

[3] url: http://pan.webis.de/, consultada em Março de 2013

exploração desta área da detecção automática de plágio [3].

1.2 Definição do Problema

Actualmente, os problemas relacionados com plágio têm vindo a ocorrer com mais frequência, e como se sabe a procura de plágio em documentos é feita manualmente, sem formas automáticas torna-se muito difícil e demorada consoante o tamanho do documento a ser tratado e o tamanho da coleção de documentos que poderá estar envolvida com o documento que se pretende comparar. As tentativas de se conseguir uma forma de ultrapassar este problema recai sobre métodos que sejam automáticos, acabando por se necessitar sempre de meios manuais (avaliador humano) para finalizar este processo de detecção.

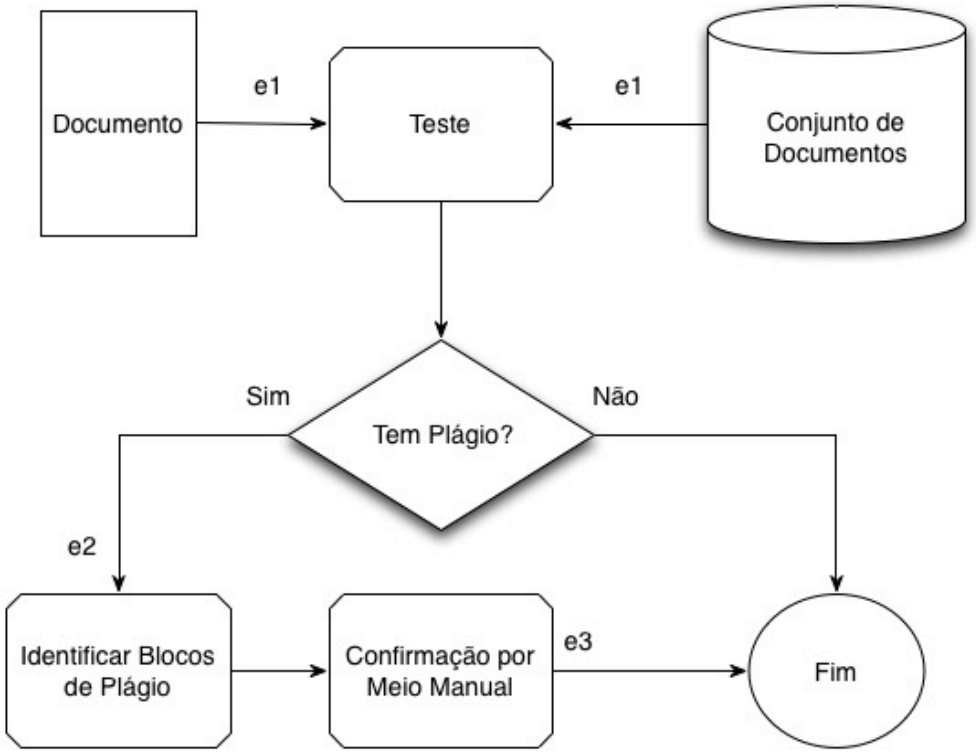

Figura 3 - Esquema da Definição do Problema

A nossa abordagem (Figura 3) numa primeira etapa (e1) pretende identificar se o documento é plagiado, caso se confirme que o documento contém plágio numa segunda etapa (e2) pretendemos identificar as possíveis zonas plagiadas, por fim (e3) pretende-se que o avaliador final confirme se as possíveis zonas são realmente plágio, ou não. Este processo apesar de requerer a influência de um avaliador final pode ser vista como uma vantagem, pois o nosso sistema poderá usar o *feedback* obtido para ser melhorado. O nosso trabalho vai no sentido de propor metodologias automáticas, eficazes e eficientes para atingir estes objectivos em concreto.

1.3 Objectivos

Esta dissertação tem como estudo duas áreas: a Similaridade Documental e a Detecção Automática de Plágio sendo dentro destas áreas que se centram os nossos objectivos, dando uma relevância maior à DAP. Em seguida serão apresentados os objectivos subjacentes às duas áreas propostos para esta dissertação de mestrado, começamos por ver os objectivos referente à SD e finalizamos com os objectivos referentes à DAP.

Similaridade Documental

1. Aprofundar o estudo nesta área e aprender novos conhecimentos acerca dela.
2. Estudar os métodos existentes, recolher informação acerca deles, optar por um método que seja realmente bom e o mais utilizado; se possível desenvolver método que difere dos métodos estudados.
3. Gerar um ranking no qual teremos ordenado por valor de semelhança, pares de textos.
4. Interpretar resultados e tirar conclusões.

Detecção Automática de Plágio

1. Estudar o conceito Detecção Automática de Plágio e aprender novos conhecimentos acerca desta área .
2. Estudar métodos aplicados anteriormente e explorar caminhos que tragam novas contribuições nesta vasta área, se possível contribuir com uma nova metodologia para a determinação de plágio em documentos de forma automática. Obter como resultado final as passagens plagiadas entre documentos.
3. Apresentar os resultados e expor as conclusões obtidas.

1.4 Estrutura da Dissertação

Esta dissertação está dividida em 5 capítulos, neste primeiro capítulo começamos por introduzir o tema que vamos focar, definimos o problema que será tratado e os principais objectivos propostos a atingir, finalizamo-lo com uma breve sumarização do conteúdo presente em cada capítulo que é a própria estrutura da dissertação.

No segundo capítulo debruçamo-nos sobre o trabalho relacionado e efectuado por outros autores em ambas as áreas estudadas — a Similaridade Documental e da Detecção de Plágio. Começamos pela Similaridade Documental, explicamos alguns conceitos base, necessários para mais fácil compreensão e só após a explicação destes conceitos, partimos para o estudo do trabalho relacionado e elaborado por outros autores. Em seguida debruçamo-nos na outra área de estudo da dissertação, a Detecção Automática de Plágio, fazemos uma abordagem semelhante ao que foi feito para a Similaridade Documental e finalizamos com as diferentes perspetivas dos autores.

No terceiro capítulo serão explicadas de forma detalhada as diferentes experiências efectuadas ao longo do desenvolvimento da dissertação.

No quarto capítulo enumeramos todos os resultados obtidos a partir das diferentes experiências, as considerações mais relevantes e importantes face a cada experiência sejam os resultados obtidos os esperados ou não.

Por fim temos o quinto capítulo no qual são tomadas as conclusões relativas a esta dissertação e também propostos trabalhos futuros neste âmbito.

2. Trabalho Relacionado

Neste capítulo apresenta-se o trabalho realizado anteriormente por outros autores nas subáreas da Pesquisa de Informação (PInf) [1], as áreas da Similaridade Documental (SD) e na Detecção Automática de Plágio (DAP). Na secção 2.1 será feito o estudo relativamente à SD, métodos e técnicas efectuadas, na secção 2.2 estarão presentes os mesmos assuntos mas para a DAP, por fim na secção 2.3 irá ser feito a sumarização deste capítulo.

2.1 Similaridade Documental

Antes de partirmos para a Detecção Automática de Plágio (DAP) é importante conseguir separar documentos por assunto, em documentos cujo o assunto seja similar, há mais probabilidade de existir plágio do que em documentos que o assunto seja completamente diferente, ao fazermos a separação dos documentos fará com que seja mais fácil aplicar a DAP.

Então é na área da Similaridade Documental que fazemos a separação ou agrupamento de textos que têm, ou não, assuntos em comum, vários autores [2], [9], [10] já debateram esta área (SD) e utilizaram diferentes medidas para o calculo da similaridade [10].

Antes de avançarmos para as medidas de similaridade propriamente ditas vamos explicar alguns conceitos necessário para uma mais fácil compreensão.

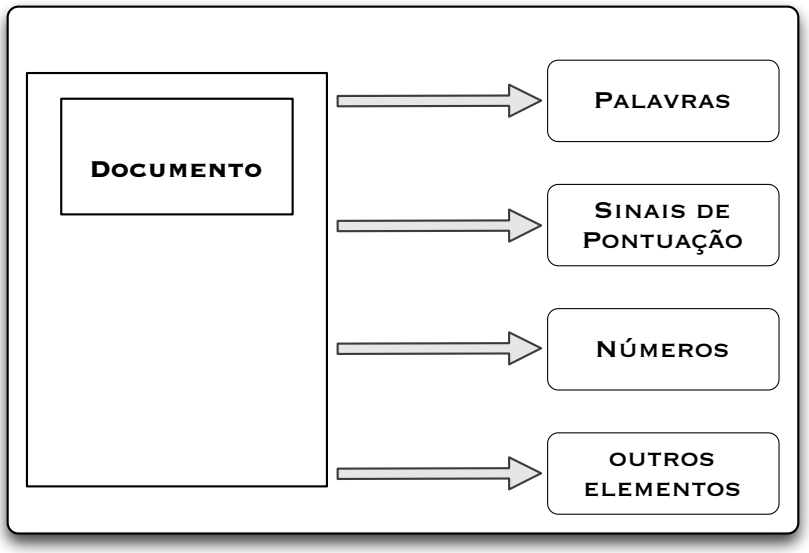

Figura 4 - Constituição de um Documento

Um documento de texto é definido por palavras, números, sinais de pontuação entre outros elementos — Figura 4. Matematicamente ao definirmos um documento este define-se segundo um vector, neste vector não vamos considerar todos os elementos representados na Figura 4, primeiro necessitamos reduzir os nossos elementos ao conjunto de todas as palavras que formam o documento e com estas palavras vamos criar o vocabulário do documento. O vocabulário de um documento é o conjunto de palavras únicas nesse documento.

Utilizando a frase seguinte para exemplificar:

"O Pedro vai às compras com a Mãe e depois vai ao cinema com a Maria às 21h30".

Neste exemplo o vocabulário seria reduzido a:

a|às|ao|cinema|com|compras|depois|e|maria|mãe|o|pedro|vai

Como podemos observar as palavras "vai", "às", "com", "a" não aparecem repetidas no nosso vocabulário. Na sequência do termo vocabulário surge um outro termo o $tf(d,t)$ que pode ser explicado como a frequência de uma palavra t ocorrer num documento de texto d.

Para podermos utilizar as medidas de similaridade necessitamos representar um documento (d) na forma matemática, ou seja, em um vector. Sendo assim um conjunto de documentos será definido por D em que D = $\{d_1, d_2, d_3, \dots, d_n\}$ e T = $\{t_1, t_2, t_3, \dots, t_m\}$ é o conjunto de palavras que ocorrem em um documento mas que não se repetem (o nosso vocabulário anteriormente visto).

Inicialmente um documento é representado por um vetor de m dimensões $\overrightarrow{t_d}$, a frequência de uma palavra t ocorrer num documento d será denotada por $tf(d,t)$, então a representação de um vetor de um documento ficará definida por $\overrightarrow{t_d} = \langle tf(d,t_1), \dots, tf(d,t_m) \rangle$.

Por fim um documento é representado por um vector em que cada componente denota a importância de uma palavra no documento $-$ o $TFIDF$ (term frequency inverse document frequency). Essa importância é calculada pelo produto de dois factores, um deles o tf que dá a frequência relativa da palavra no dado documento d, a outra componente é o idf que dá a importância de uma palavra numa coleção de documentos. [5]. Inicialmente um documento era representado pela frequência de cada palavra mas ao aplicarmos o $TFIDF = tf * idf$ reduzimos o vocabulário excluindo as palavras chamadas de "stop words", que são as palavras que ocorrem em todos os documentos com frequência elevada, como veremos em seguida com auxilio à Figura 5 e à explicação seguinte.

Exemplificando:

Dado o conjunto de documentos em várias áreas, suponhamos quatro tal como na Figura 5.

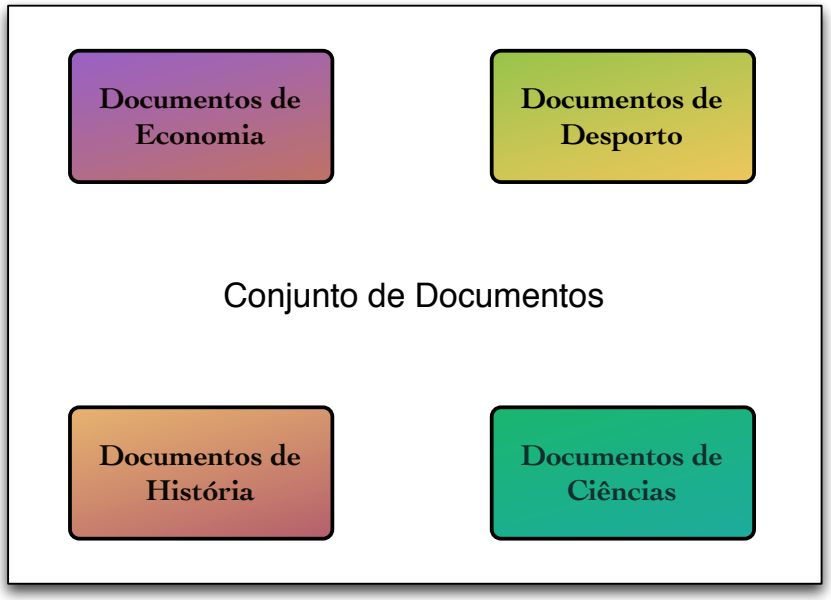

Figura 5 – Conjunto de Documentos em diversas áreas

Dentro de cada área existem por sua vez diversos temas, dentro de Desporto teríamos um documento com conteúdo acerca da final da Liga Europa, e é neste que nos vamos centralizar, palavras como “final”, ”Liga”, “Europa” são palavras que ocorrem com um $TFIDF$ maior neste documento que em qualquer outro do conjunto de Documentos (provavelmente em documentos de Economia, Ciências e História nem apareceriam estas palavras) mas palavras como “de”, ”da”, ”com” são palavras com elevada frequência (tf) em todos os documentos do conjunto, o seu idf é próximo de zero o que levará a ter um $TFIDF$ próximo de zero (as “stop words” que foram referidas anteriormente); as palavras com maior valor de $TFIDF$ seriam as palavras necessárias para definir um documento, estas palavras são consideradas as palavras mais importantes.

Para calcularmos o $TFIDF$ apoiámo-nos no método dos autores Gerard Salton e Christopher Buckley [6] e recorremos à sua fórmula:

$$tfidf = tf(d,t) * \log\left(\frac{|D|}{df(t)}\right), \quad (1)$$

temos que $df(t)$ é o número de documentos em que a palavra t aparece. Futuramente usaremos $w_{t,d}$ para quantificar o peso que uma palavra tem num dado documento d [7], tf foi visto anteriormente e utilizaremos D para o conjunto de Documentos. Em seguida serão apresentadas algumas medidas de similaridade existentes na literatura.

Com esta representação vectorial dos documentos, podemos agora usar fórmulas matemáticas, para o cálculo da similaridade entre vectores. Uma dessas fórmulas é a "similaridade cosseno" (SimCos), isto é, o cosseno do ângulo formado por dois vectores.

Dados estes dois vectores $\overrightarrow{t_a}$ e $\overrightarrow{t_b}$ que são vetores de importâncias de palavras nos documentos a e b definidos por T = $\{t_1, t_2, t_3, \ldots, t_m\}$ com tamanho m. Representamos SimCos através da formula:

$$Sim_C\left(\overrightarrow{t_a}, \overrightarrow{t_b}\right) = \frac{t_a \,.\, t_b}{\left\|\overrightarrow{t_a}\right\| * \left\|\overrightarrow{t_b}\right\|} \quad (2)$$

Um ponto forte desta medida é não dar importância ao tamanho dos documentos, não apresentando problema se um for maior do que outro, actualmente é a medida mais utilizada na comparação de documentos seja para pesquisa de informação ou para criação de clusters [4], os resultados obtidos variam entre 0 e 1, sendo que quanto mais próximo de 1 mais similares são os documentos considerados [8].

Para documentos a medida de similaridade Coeficiente de Jaccard "compara a soma dos pesos dos termos comuns com a soma dos pesos dos termos que estão em qualquer um dos dois documentos, mas que não são

os termos comuns", a similaridade é medida com a divisão da intersecção pela união dos objetos, nesta medida a variação do resultado é entre 0 e 1, sendo que 1 indica que os objetos são iguais e 0 diferentes [5]. Dados dois documentos $\overrightarrow{t_a}$ e $\overrightarrow{t_b}$ que são vetores de importância de palavras definidos por T = $\{w_1, w_2, w_3, \dots, w_n\}$ com tamanho n. Representamos o Coeficiente de Jaccard através da formula:

$$Sim_J(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{t_a \,.\, t_b}{\left\|\overrightarrow{t_a}\right\|^2 + \left\|\overrightarrow{t_b}\right\|^2 - t_a \,.\, t_b} \tag{3}$$

Correlação de Pearson [5] [9] é também uma medida que se rege pelo calculo entre vetores mas apresenta resultados entre a gama de intervalo de -1 a 1, sendo que dois documentos são iguais se o resultado da similaridade for 1. Esta Correlação de Pearson representa-se segundo a seguinte equação:

$$Sim_P(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{m \sum_{t=1}^{m} w_{t,a} * w_{t,b} - TF_a * TF_b}{\sqrt{\left[m \sum_{t=1}^{m} w_{t,a}^2 - TF_a^2\right] * \left[m \sum_{t=1}^{m} w_{t,b}^2 - TF_b^2\right]}}, \tag{4}$$

em que $TF_a = \sum_{t=1}^{m} w_{t,a}$ e $TF_b = \sum_{t=1}^{m} w_{t,b}$.

$w_{t,a}$ é o peso da palavra t ocorrer no documento a, para $w_{t,b}$ é o peso da palavra t ocorrer no documento d.

Distância Euclidiana [5] [9] é a medida vulgarmente utilizada para a resolução de problemas geométricos, pois permite medir a distância entre dois pontos facilmente, relativamente a dois documentos a distância utiliza as palavras de cada documento para calcular esta distância, recorre-se à seguinte formula:

$$DE(\overrightarrow{t_a}, \overrightarrow{t_b}) = \sqrt{\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2}$$

(5)

Averaged Kullback-Leibler Divergence (KL-Divergence) [10] é uma medida que testa o quanto duas distribuições de probabilidades diferem uma da outra, sendo assim nesta medida os documentos são considerados como distribuições de probabilidades. Olhando-se exclusivamente para documentos, dadas duas distribuições de palavras A e B a divergência entre elas define-se pela seguinte formula:

$$D_{KL}(\overrightarrow{t_a}||\overrightarrow{t_b}) = \sum_{t=1}^{m} w_{t,a} * \log\left(\frac{w_{t,a}}{w_{t,b}}\right)$$

(6)

sendo $w_{t,a}$ a importância que uma palavra t tem num dado documento a e $w_{t,a}$ a importância que uma palavra t tem num dado documento b.

Esta medida não é simétrica então usa-se é a *Averaged KL Divergence* que ao se aplicar torna os valores simétricos, temos ainda três expressões necessárias à formula final sendo elas [11]:

$$\pi_1 = \frac{w_{t,a}}{w_{t,a}+w_{t,b}}, \pi_2 = \frac{w_{t,b}}{w_{t,a}+w_{t,b}}, \; w_t = \pi_1 * w_{t,a} + \pi_2 * w_{t,b}$$

como fórmula final para a *Averaged KL Divergence*:

$$D_{Avg\ KL}(\overrightarrow{t_a}||\overrightarrow{t_b}) = \sum_{t=1}^{m}(\pi_1 * D(w_{t,a}||w_t) + \pi_2 * D(w_{t,b}||w_t))$$

(7)

Liu e Guo [12] debruçaram-se sobre uma medida envolvendo a média o desvio padrão da distribuição dos documentos. Os autores explicam que a partir de um documento de texto é gerado um documento definido sobre os parâmetros anteriormente vistos (um vetor de importância de palavras, ou seja cada coordenada é um TFIDF).

Para cada classe de documentos C_j (por exemplo classes de documentos de desporto, economia, culinária, etc.), criam-se dois vetores protótipos: o

vetor da média $u_j = \{u_{j1}, u_{j2}, u_{j3}, \ldots, u_{jm}\}$, e o vetor de desvio padrão $\sigma_j = \sigma_{j1}, \sigma_{j2}, \sigma_{j3}, \ldots, \sigma_{jm}$, ou seja, cada classe dá origem a estes dois vetores. A similaridade entre um documento e uma dada classe é calculada por:

$$Sim(d_{test}, C_j) = -\sum_{i=1}^{m} \frac{\max\{|w_{testi} - u_{ji}| - \sigma_{ji}, 0\}}{\sigma_{ij}^2},$$

(8)

em que d_{test} representa o documento a ser classificado numa das classes C_j, w_{testi} é o TFIDF da palavra do documento, u_{ji} é a coordenada do vetor protótipo de médias, σ_{ji} é a coordenada do desvio padrão.

Após o cálculo da similaridade com todas as classes o valor que for mais elevado classificará o documento como sendo similar com a determinada classe, exemplificando:

Classes: Desporto, Economia, Culinária
Documento de teste: A

$$Sim(A, Desporto) = 0{,}92341$$
$$Sim(A, Economia) = 0{,}00321$$
$$Sim(A, Culinária) = 0{,}00121$$

Neste caso o documento A seria atribuído como um documento cujo conteúdo era de Desporto.

Os autores mostraram que a utilização do desvio padrão ajuda a uma classificação mais precisa ao contrário de outros trabalhos similares que só usavam o vetor médio. Portanto para uma dada distribuição de documentos não interessa só o ponto médio mas também a sua dispersão.

2.2 Detecção Automática de Plágio

A área da Detecção Automática de Plágio, apesar de recente, tem evoluído consideradamente a partir de 2009, ano em que surgiu a conferência PAN, a conferência com objectivo único e exclusivo para a DAP. Autores como Grman e Ravas [13], Gorzea e Popescu [14], Oberreuter [15] , Torrejón e Ramos [16], Ghosh [17], entre outros têm vindo a participar nas PAN's e vindo a dar o seu contributo. Como veremos de seguida.

Grman e Ravas [13] desenvolveram um método que se pode dividir em 3 fases, sendo elas o pré-processamento dos dados de entrada, a detecção dos possíveis casos de plágio, ou seja, os documentos candidatos a conterem plágio e o pós-processamento que inclui a remoção de passagens que se repetem, a fusão de passagens e a remoção das passagens nas quais à partida não existira plágio (passagens incertas).

No pré-processamento estes trabalham ao nível das palavras individuais dando importância aos sinónimos, variações, abreviaturas, etc de forma a reduzir a quantidade de palavras e permitindo a comparação de forma eficaz, com o objectivo de guardar a informação processada numa estrutura de dados mais eficiente, para este processo é feita a tradução do texto para o inglês (se necessário), a extração de texto (caracteres, deslocamento e comprimento), normalização do texto, por fim o texto que era constituído por palavras acaba por ser transformado em valores numéricos (códigos) e por ser guardado como ficheiro binário. Numa segunda fase é feita a detecção das passagens suspeitas que poderão ser plágio. O método utilizado recai sobre a concordância das passagens, isto é, quantifica-se a intersecção dos conjuntos de palavras existentes nas diferentes passagens, evitando assim a ordem em que possam estar as palavras das passagens, exemplificando recorrendo a duas frases:

1ª Frase:

"Atentado contra mercado em Bagdá deixa 11 mortos e 35 feridos"

2ªFrase:

"11 pessoas mortas e 35 feridos num atentado contra um mercado em Bagdá"

Como podemos observar, independentemente da ordem das palavras, ao fazer-se a intersecção das palavras de ambas as frases vão resultar o conjunto de palavras {atentado, bagdá, contra, em, feridos, mercado, mortos}.

Para finalizar a metodologia na fase de pós processamento os autores apenas tratam de remover as passagens que se encontram sobrepostas. Este método revelou-se como sendo o que obteve melhores resultados na conferencia PAN2012 [3].

Torrejón e Ramos [18] , [16], [19] debruçaram-se sobre um método no qual explicam a forma de detectar plágio do sistema CoReMo 1.9 Plagiarism Detector, CoReMo foi inicialmente criado com o objectivo de participar na PAN a fim de obter resultados eficazes em termos de performance.

Torrejón e Ramos [18], [10], [16], [19] basearam o seu sistema no modelo apresentado na Figura 6, pertencente a Potthast [3].

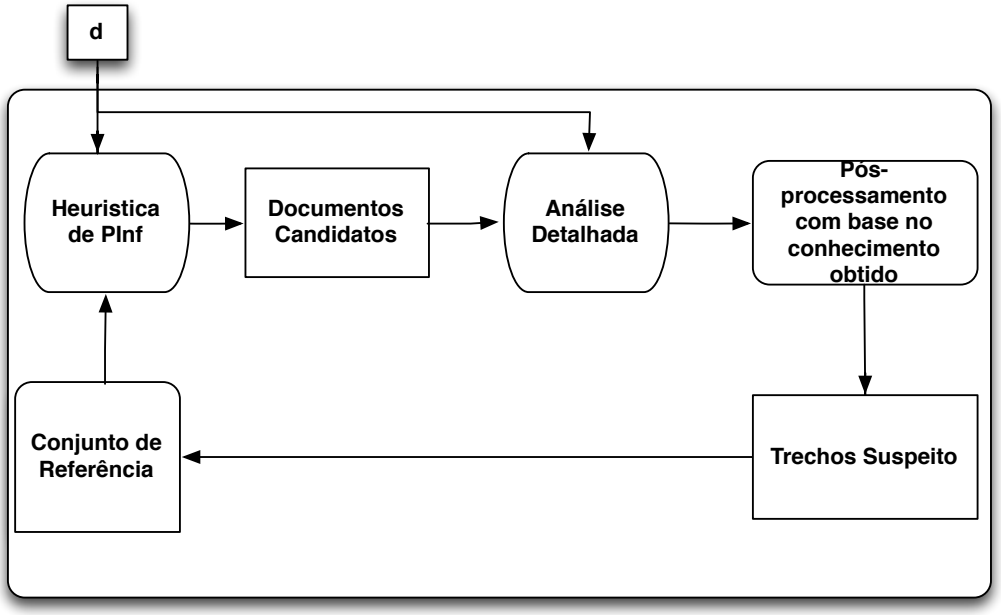

Figura 6 – Arquitectura para a Detecção de Plágio de Potthast [18]

É recebido um documento d e através de uma heurística selecciona-se um conjunto de documentos que são os documentos como fonte de plágio. Com uma análise detalhada entre os pares de documentos fonte e o suspeito de plágio d, aplicava-se um pós-processamento o qual classificaria o documento como um documento em que potencialmente existe plágio [18]. Um dos aspetos fortes no qual estes autores se basearam foi na construção melhorada dos n-gramas (n-gramas são conjuntos de ocorrências de palavras [19]), já Barron e Rosso [20] tinham tido a mesma preocupação outrora; visto que o sistema CoReMo é um sistema baseado na detecção dos n-gramas Torrejón e Ramos debruçaram-se sobre uma forma de fazer o tratamento dos n-grames detalhadamente dividindo este processo em seis passos sendo eles enumerados pela ordem a seguir:

1. Passar todo o texto para letras minúsculas;
2. Eliminação das “stop words”;
3. Eliminar palavras de um carácter;
4. Aplicar a técnica de redução de palavras ao seu radical (Stemming);

5. Ordenação alfabética das palavras internas ao N-Grama;
6. Utilização da sobreposição de palavras n – 1 entre N-Gramas contextuais consecutivos;

Nota:

Exemplos de Stemming: weakness → weak e temptation → temptat

Com o tratamento dos N-Gramas conseguiam superar o problema da ofuscação (forma de esconder o plágio) conseguindo obter melhores resultados.

Oberreuter [15] divide o seu método em duas fases, numa primeira fase pretende reduzir o espaço de procura de plágio, pretendendo identificar os pares de textos nos quais pode existir plágio (pares suspeitos).Tal como em outros autores anteriormente vistos, um dos passos deste método é a exclusão de "stop words", se dois documentos contiverem pelo menos dois conjuntos de 4-gramas de palavras no mesmo parágrafo então estes documentos de texto passam ambos à segunda fase do teste, se não são excluídos [21]. Numa segunda fase aplica-se uma pesquisa exaustiva para encontrar os segmentos de plágio, de modo semelhante na qual se utilizam 3-Gramas de palavras mas neste caso as "stop words" não são excluídas [21].

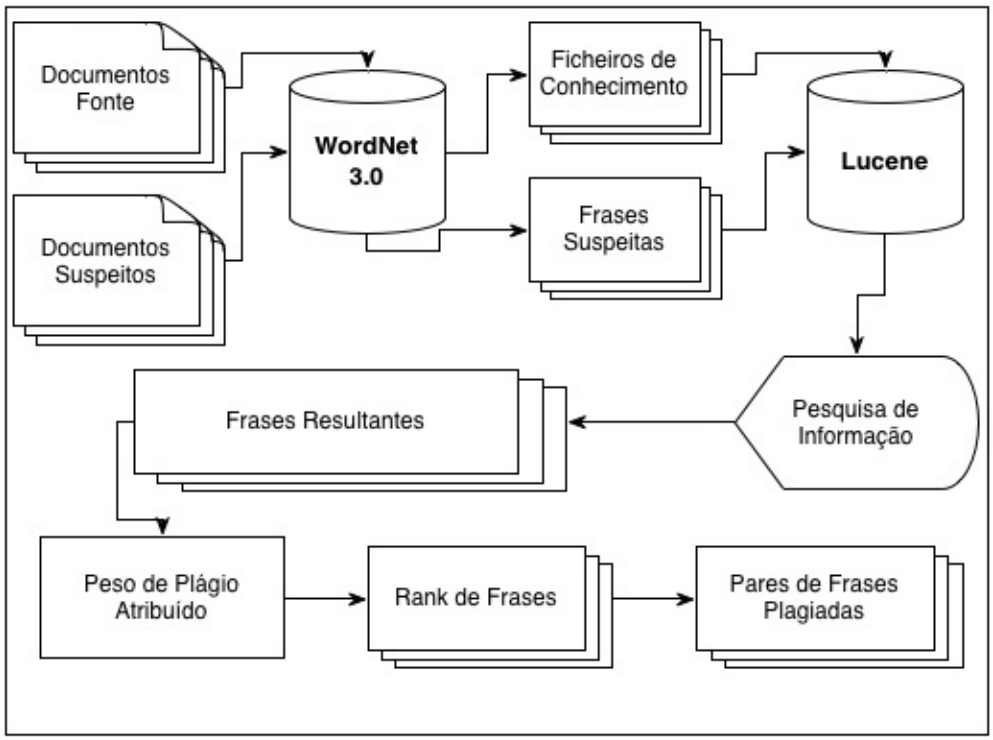

Figura 7 - Arquitectura do Método de A.Ghosh [17]

Ghosh [17] dividiu a sua abordagem em três fases: preparação do conhecimento, recolha de pares candidatos e a detecção de plágio. Começando pela preparação do conhecimento, nesta fase são tratados os documentos de texto fonte, para cada texto são extraídas e analisadas as frases, após a análise das frase são criados os ficheiros de conhecimento que permitem fácil acesso às frases no documento fonte através de um índice. No ficheiro de conhecimento é colocado o vocabulário de cada frase segundo alguns parâmetros (sinónimos, homónimos, hiperónimos, é também feito o stemming, entre outras operações de transformação de palavras) recorrendo ao WordNet 3.0[4], posteriormente à criação dos ficheiros de conhecimento é utilizado o Lucene[5] [23] para os indexar.

[4] http://wordnet.princeton.edu

[5] http://lucene.apache.org/

Numa segunda fase Ghosh debruça-se sob a recolha dos pares de textos candidatos, começando pela extracção das frases dos documentos suspeitos, remoção de "stop words" e aplicando um stemmer transformando as palavras no seu radical, com estas frases são criadas querys que posteriormente são passadas ao Nutch[6] (é um "Web Crawler" open source, um "Web Crawler" é um programa que vai seguindo as hiperligações na web e descarregando as páginas que vai encontrando, este pode seguir algum critério, como por exemplo procurar páginas relativas a um determinado tema, ou então que tenham um mínimo volume de texto), este retornará as frases do documento fonte que são próximas às frases do documento suspeito que foi passado como query, estas frases provenientes do Nutch serão as frases do documento fonte são as candidatas a serem plagiadas.

Por fim numa ultima fase é feita a detecção de plágio em si, recorrendo ao algoritmo proposto por Kešelj et al. [22] calculando o valor dissimilar entre as frases candidatas do documento fonte e as frases do documento suspeito, ao valor de similaridade é retirado o valor dissimilar anteriormente calculado e criado um ranking (listagem de valores de forma ordenada, crescentemente ou decrescentemente) ordenado do maior para o menor valor e identificado com a frase de um dado documento fonte e a frase do documento suspeito.

Supondo que o Documento Suspeito é constituído por frases de A — F e o Documento Fonte é constituído por fases de 1 — 6, exemplificando como é dado o ranking tal como a Tabela 1.

[6] http://nutch.apache.org/

Ranking de Frases		
Similaridade	**Frases Documento Suspeito**	**Frases Documento Fonte**
1.0	A	2
0.9	F	3
0.6	E	6
0.5	D	1
0.4	B	5
0.1	C	4

Tabela 1 - Ranking de Valores de Similaridade entre frases do Método de Ghosh

2.3 Sumário

Neste capítulo apresentamos o trabalho relacionado que tem sido desenvolvido no âmbito das duas áreas de estudo desta dissertação: SD e DAP. Primeiramente observamos algumas medidas de calculo de similaridade entre documentos e finalizamos com os métodos expostos por alguns autores que trabalharam na área e contribuíram significativamente para o avanço destas áreas.

3. Metodologias e Experimentação

Este Capítulo está dividido em 4 secções, descrevendo as metodologias utilizadas no desenvolvimento da dissertação, assim como o conjunto de experiências realizadas. A maioria dos métodos implementados e experiências realizadas foram feitas em linguagem Java. Na secção 3.1 é descrita a primeira abordagem relativa a detecção de documentos similares de forma não eficiente, na secção 3.2 debruçamo-nos sobre um método muito semelhante mas eficiente combatendo alguns dos problemas apresentados no método Base, na secção 3.3 é explicada uma experiência na qual damos importância à média e desvio padrão dos valores de similaridade dos documentos com e sem plágio, a secção 3.4 explicamos duas abordagens "Overlap" e "CosSim" em que obtemos resultados de medidas que se pretendem obter na PAN tais como a *Precision* o *Recall* e a *F-Measure*, para finalizar as experiências na secção 3.5 descrevemos outro método a Determinação das Zonas de Plágio. Por fim na secção 3.5 fazemos o sumário do capítulo.

3.1 Similaridade Documental — *"Método Base"*

No inicio do nosso trabalho fizemos uma primeira experiencia no âmbito da Similaridade Documental tendo como objectivo desenvolver uma abordagem que fosse eficaz no cálculo da similaridade entre documentos. Sendo assim, e partindo do inicio em que tínhamos um conjunto de documentos de texto reduzido (102 documentos), não houve uma preocupação muito grande com os tempos de processamento e gastos de memória, visto que era o nosso primeiro teste. A principal preocupação foi

analisar o tipo de valores obtidos no cálculo de similaridade. Nesta fase o conjunto de documentos utilizado foi uma coleção de noticias em Português recolhidos e provenientes da web, especialmente do Google Notícias. Nesta coleção 10 pares de textos são muito idênticos (basicamente plágios uns dos outros), 10 pares de textos são similares, sendo o conteúdo da noticia muito semelhante e os restantes textos apresentam conteúdo diferente entre eles. Como corpus[7] mais geral para estimar frequências de palavras usámos uma coleção de noticias do jornal Público, da década de 90, disponível eletronicamente — o CETEMPúblico.

Para o calculo da Similaridade entre documentos propusemos uma nova fórmula que nos dá como resultado a importância de uma palavra num documento. Esta fórmula é uma adaptação do TFIDF de Salton e Buckley [6], para calcular a importância ou relevância das palavras num documento. A fórmula — ***Função de Importância do Termo*** — proposta e utiliza foi a seguinte:

$$I(w|R) = comprimento\ palavra\ * log_2\left(\frac{P(w|R)}{P(w)}\right) \tag{9}$$

sendo $P(w) = \frac{freq(w|C)}{|C|}$, $P(w|R) = \frac{freq(w|R)}{|R|}$

em que $I(w|R)$ é o valor da importância de cada palavra do vocabulário, $P(w|R)$ é a probabilidade da palavra w ocorrer em um região (R), $P(w)$ é a probabilidade da palavra aparecer no Corpus[7], $freq(w|C)$ é a frequência da palavra w no Corpus e $|C|$ é o tamanho do Corpus[7], $freq(w|R)$ é a frequência da palavra na w na região e |R| é o tamanho da Região. Nas Figuras 8 e 9 podemos ver uma representação da Região e do Corpus e no exemplo seguinte a aplicação da nossa fórmula com o auxilio de um exemplo com valores numéricos.

[7] No presente caso o Corpus é um Corpus Português

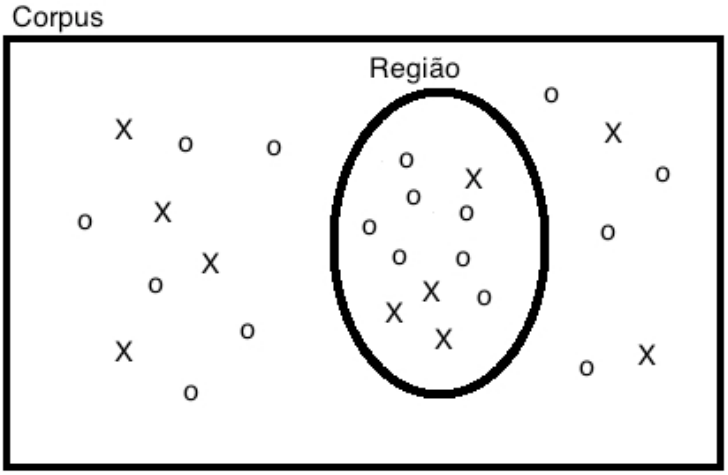

Figura 8 - Representação do Corpus e da Região no espaço

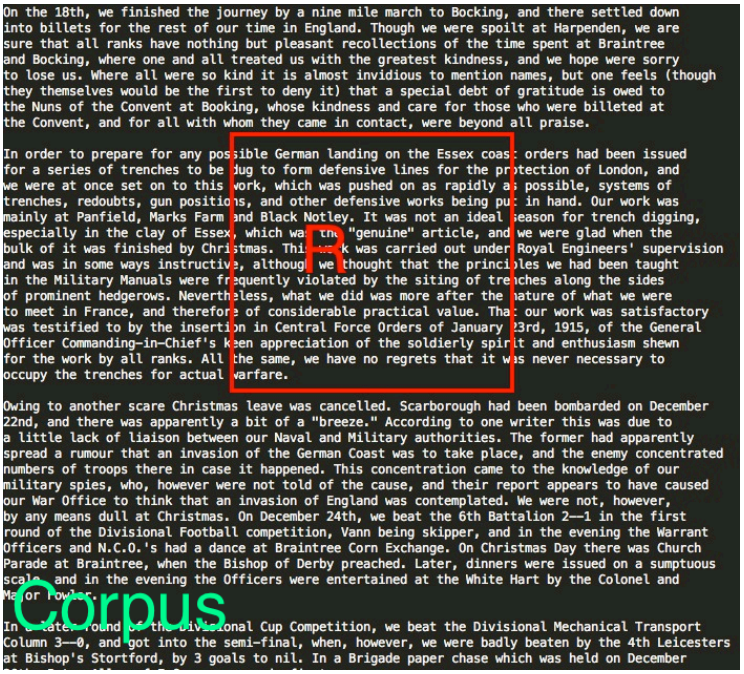

Figura 9 - Representação do Corpus e da Região em um Documento

Exemplificando a nossa fórmula e recorrendo a valores numéricos:

Comprimento da palavra 1 (w1) = 8 TF(w1|R) = 5 TF(w1|C) = 10

Comprimento da palavra 2 (w2) = 3 TF(w2|R)= 20 TF(w2|C) = 300

|C| = 1000 termos |R| = 100 termos

$$I(w1|R) = 8 * log_2\left(\frac{\frac{5}{100}}{\frac{10}{1000}}\right) <=> I(w1|R) = 8{,}9657$$

$$I(w2|R) = 3 * log_2\left(\frac{\frac{20}{100}}{\frac{300}{1000}}\right) <=> I(w2|R) = -0{,}7369$$

Através do exemplo podemos concluir que a palavra mais relevante é a w1, w1 ocorre menos vezes na região mas também ocorrem menos vezes no Corpus.

Como foi visto anteriormente o vocabulário de um documento é formado por todas as palavras únicas nesse documento, a nossa *função de Importância do Termo* foi gerada para ser aplicada em cada uma das palavras do vocabulário do documento dando como resultado a importância de cada palavra e dependendo desse valor decidimos se essa palavra realmente era uma palavra pertencente às mais importantes ou não. Para escolhermos as palavras mais importantes de um documento que o representassem, efetuámos alguns testes de procura da melhor percentagem. Portanto, estas percentagens diziam-nos quantas palavras seriam necessárias para posteriormente se obter um valor de similaridade pretendido.

Em seguida (Figura 10) explicaremos o método utilizado.

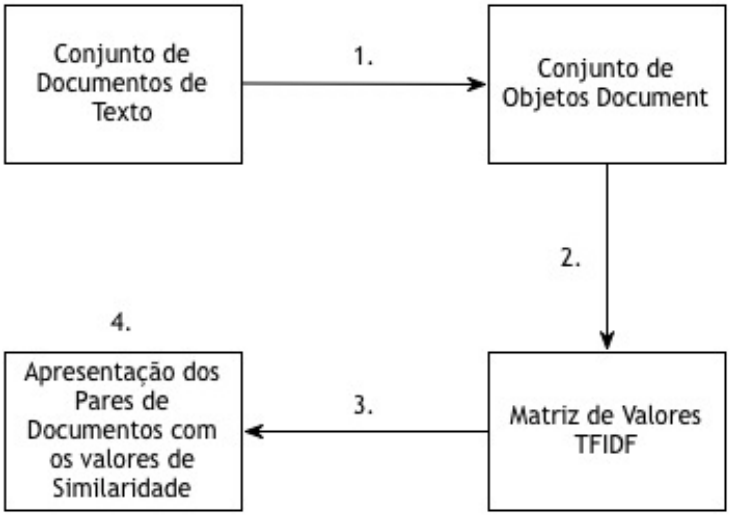

Figura 10 - Fases do Método Base

Começamos por receber um conjunto de Documentos de Texto.

1) Processamos o conjunto de documentos de texto em objetos Document. Este Objeto Document surge no contexto em que necessitamos guardar algumas informações relativas a um documento, como por exemplo o seu nome, ou a localização para o ficheiro de texto, as palavras mais importantes, entre outros elementos.
   a) Processamento do vocabulário dos documentos de texto do conjunto de teste.
   b) Cálculo de 20% de palavras mais importantes para cada documento de texto. Os 20% foi um valor decidido através de vários testes que são apresentados na Secção 4.1.
2) Processamento da matriz que contem os valores de TFIDF.
   a) Cálculo das normas de cada Documento.
3) Cálculo da Similaridade entre documentos baseada num método anteriormente visto — Similaridade Cosseno.
4) Apresentação dos pares de documentos e valores dos resultados de Similaridade entre eles para uma conclusão.

Este método preliminar revelou naturalmente uma dificuldade muito pertinente, em termos de eficiência computacional. Nomeadamente, no passo 2 da Figura 10 gerámos uma matriz que continha todos os valores TFIDF a qual ficava guardada em memória para posteriormente ser calculada a norma dos documentos e assim o cálculo da similaridade. Para um conjunto reduzidos de documentos de textos, que era o nosso caso, funcionava perfeitamente visto que a quantidade de memória utilizada era baixa, mas ao aumentarmos este conjunto de documentos os problemas de memória apareceram, tal como o esperado. Um dos principais problemas nesta área é precisamente a eficiência computacional, devido aos gastos acrescidos de recursos (neste caso memória da máquina). Quanto mais recursos tivermos ao dispor mais eficientemente conseguimos resultados. Portanto, esta nossa estratégia de redução do vocabulário, através da escolha das palavras mais significativas/informativas, prende-se com a necessidade de utilizar recursos minimalistas para a representação dos documentos.

3.2 Similaridade Documental — *"Método Eficiente"*

Este método surgiu com o propósito de optimizarmos o nosso Método Base, pois o método base estava preparado para receber um conjunto reduzido de documentos de texto (102 testados inicialmente), no qual não existiam dificuldades a nível da eficiência dos recursos. Um dos caminhos que decidimos tomar foi colocar de parte a matriz de TFIDF que era guardada em memória e fazer o cálculo quando necessário em tempo de processamento, desta forma poupando-se a memória que estava a ser ocupada pela matriz. Recorrendo à Figura 11 será explicado o método de forma a compreendermos melhor.

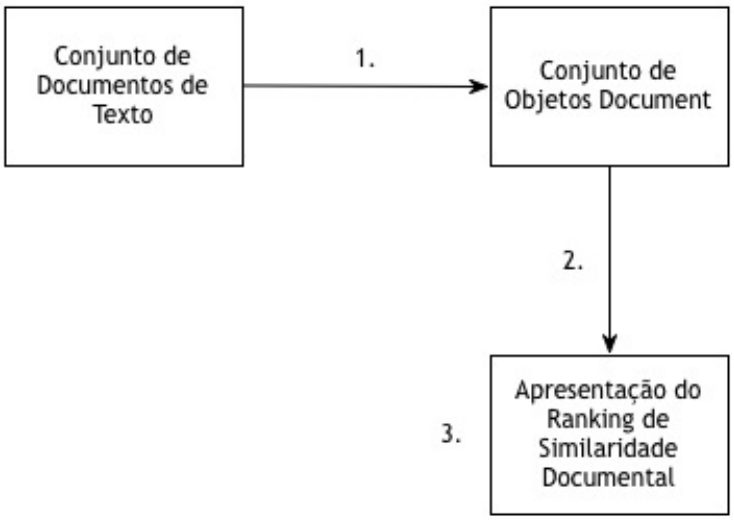

Figura 11 - Esquema com as fases do Método Eficiente

Neste método começávamos novamente por receber o conjunto de documentos de texto

1) Processamos o conjunto de documentos de texto em objetos Document.
   a) Processamento do vocabulário dos documentos de texto do conjunto de teste.
   b) Cálculo de 20% de palavras mais importantes para cada documento de texto e também da norma que seria guardada no objecto Document.
2) Cálculo da similaridade entre documentos baseada num método anteriormente visto - Similaridade Cosseno.
3) Apresentação dos resultados sobre a forma de um ranking de similaridade entre documentos.

Como podemos ver o algoritmo é semelhante ao efectuado no método base mas com a particularidade de evitarmos a matriz de TFIDF pois a norma é calculada no passo 1b) como podemos constatar. Ao olharmos para a explicação do algoritmo e superado o problema da matriz como passo 3 temos os nossos resultados que foram apresentados na forma de um Ranking.

Neste caso um Ranking é uma listagem ordenada decrescente segundo o valor de similaridade entre pares de documentos de texto, como ilustra a Tabela 2, na qual podemos observar que a primeira coluna contém o valor de similaridade, entre os documentos da coluna 2 com os documentos da coluna 3.

Ranking		
Similaridade	**Documento**	**Documento**
1,0000	A	B
0,9999	A	F
0,9111	E	D
0,8976	C	X
0,4321	C	M
0,1230	O	U

Tabela 2 - Ranking gerado pelo método Eficiente

Através do método Base os resultados eram gerados no formato de um ranking no qual os valores de similaridade entre documentos eram calculados a partir da totalidade de palavras, ou seja, 100% das palavras existentes em cada documento, sem ser necessário recorrer à filtragem de palavras mais importantes, este ranking identificámo-lo como "Ranking Perfeito". Ao obtermos os rankings de ambos os métodos (Base e Eficiente) fizemos um teste no qual era calculada a qualidade dos rankings.

Posteriormente, na secção 4.3 são apresentados os resultados obtidos para o cálculo da qualidade dos rankings. Deste modo, pudemos concluir que não perdemos qualidade quando trabalhamos só com uma pequena percentagem das palavras, como representantes de cada documento. Isto torna-se pertinente no aumento da eficiência da representação documental.

3.3 Similaridade em Grandes Colecções de Documentos — *"Método PAN11"*

Como foi visto nos métodos anteriores, o conjunto de documentos de texto era um conjunto reduzido, então decidimos aumentar o nosso conjunto substancialmente e dos 102 documentos de texto em Português passámos para um conjunto utilizado na PAN11 contendo 22186 documentos de texto em Inglês e outras línguas. Pretendemos passar para um conjunto maior porque é nestes conjuntos que surgem os maiores problemas de eficiência, o caso da quantidade de informação a processar ser muito maior o que torna mais difícil a identificação de plágio.

Neste método os documentos utilizados estão divididos em dois grupos, os "*source-documents*" são os documentos fonte, os documentos dos quais foram retirados excertos ou de algum modo foi feito o plágio e os "*suspicious-documents*" correspondendo aos documentos que são suspeitos e podem conter os segmentos de plágio. Cada grupo contém 23 partes, cada parte contem 1000 ficheiros, sendo 500 ficheiros .txt e os outros 500 .xml à excepção da ultima parte que contem 186 ficheiros 93 deles .txt e os outros 93 .xml. No grupo referente aos suspicious-documents os ficheiros .xml apresentam, informação acerca do nome do documento, autor, linguagem, entre outros mas mais importante ainda é conter informação indicadora de plágio, ou seja, aparecer a informação acerca do próprio suspicious-documento .txt, inicio do segmento de plágio e tamanho do segmento, assim como a informação referente ao source-document, onde está o possível segmento do mesmo, a posição em que começa o plágio e o tamanho do segmento. Outra informação é a classificação de ofuscação dos ditos segmentos de plágio, na Figura 12 podemos observar as informações referidas.

```
<?xml version="1.0" encoding="UTF-8"?>
<document reference="suspicious-document00005.txt">
  <feature name="about" authors="Saint-Simon, Louis de Rouvroy, duc de"
  title="Memoirs of Louis XIV and His Court and of the Regency - Volume 06" lang="en" />
  <feature name="md5Hash" value="ae971a081edc198d6a1132420419020f" />
  <feature name="plagiarism" type="artificial" obfuscation="low" this_language="en" this_offset="19254" this_length="1557"
  source_reference="source-document00178.txt" source_language="en" source_offset="3835" source_length="1560" />
</document>
```

Figura 12 - Exemplo da Constituição de um Ficheiro XML

A metodologia utilizada no método PAN11 baseou-se nas experiências anteriores mas neste caso sofreu algumas alterações como poderemos ver em seguida, visto o conjunto ter aumentado. Começamos com a primeira de duas abordagens diferentes.

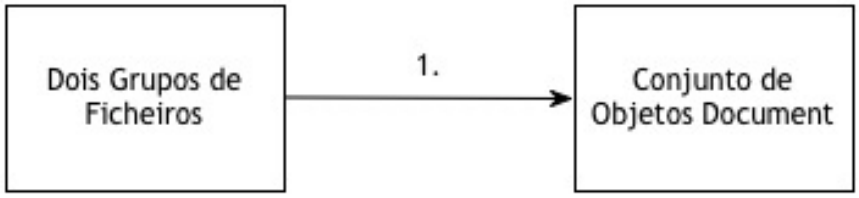

Figura 13 - Primeira etapa do método PAN11

Numa primeira etapa da primeira abordagem (Figura 13) temos os nossos dois conjuntos de documentos, os source-documents e os suspicious-documents. Como estes conjuntos contêm muitos documentos nesta primeira etapa necessitámos de processar todos os documentos e após o processamento é que partíamos para o nosso objectivo identificar documentos que continham plágio. Em seguida explicamos esta primeira etapa.

Começando pelos dois grupos de ficheiros.

1) Processamos o conjunto de documentos de texto em objetos Document.
   a) Processamento do vocabulário dos documentos de texto do conjunto de teste.

b) Aqui calculamos as 20% de palavras mais relevantes para cada documento de texto. Com estas formamos os vectores que representam os documentos e calculamos as respectivas normas destes mesmo 20% de palavras que foram guardados no objecto Document, após vários testes efectuados nos métodos anteriores chegámos à conclusão que os 20% de palavras mais relevantes em documentos de texto grandes eram demasiadas palavras e muitas das palavras já não apresentavam importância que fosse relevante, então consideramos apenas as palavras acima de um certo valor de importância — a Importância Mínima Admitida (IMA) — este valor é definido à priori. O valor que definimos para o tamanho da nossa estrutura de palavras importantes é o valor mais baixo entre a IMA e os 20% de palavras.

c) Todos os ficheiros .txt são guardados em objetos Document e toda a estrutura é guardada em um ficheiro binário, isto para garantir que apenas é feito uma vez este processamento demorado e sempre que necessário a sua utilização fazemos o carregamento do ficheiro binário para a memoria.

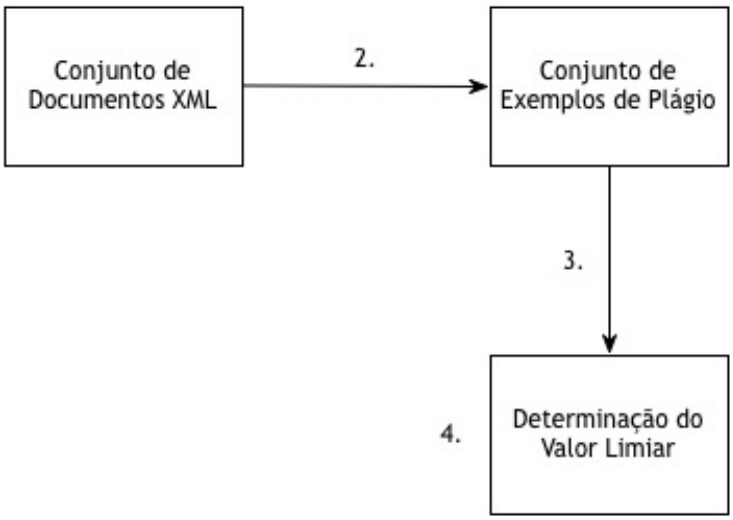

Figura 14 - Segunda etapa do Método PAN11

Nesta 2ª etapa (Figura 14) e já com os documentos de texto processados em Objetos Document, recebemos o grupo suspicious-document que contem a informação anteriormente vista no seus ficheiros XML. Nesta fase necessitamos da informação proveniente dos ficheiros XML de modo que começamos por processar estes documentos, sendo assim:

2) Passámos ao processamento dos ficheiros XML salvamos toda a informação numa estrutura que denominámos por Exemplos, estes exemplos contêm um identificador dos pares de documentos e a informação dos plágios provenientes dos documentos .xml.
3) Ao chegarmos a este passo, já temos os Objetos Document e os Exemplos de plágios, então passamos aos cálculos.
   a) Cálculo da Similaridade entre documentos baseada num método anteriormente visto — "Similaridade Cosseno". Neste passo fizemos vários testes individuais, testes em que:
      i) os pares testados contêm plágio.
      ii) os pares testados não contêm plágio.
      iii) os pares testados são escolhidos aleatoriamente podendo conter ou não plágio.
   b) Cálculo da Média e do Desvio Padrão dos valores de similaridade para os pares de documentos.
4) Apresentação dos resultados referentes à Média e Desvio Padrão.

Através dos testes em que passamos pelos passos 3a i) e 3a ii) obtemos a média e o desvio padrão dos valores de similaridade para pares de documentos com plágio e documentos sem plágio. A soma da média com o desvio padrão para ambos os casos (documentos com e sem plágio) resultam em dois valores, a média destes dois valores dá o valor limiar necessário para que quando fazemos o teste que passa pelo passo 3a iii) consigamos classificar o par de documentos como havendo ou não plágio.

Apesar desta experiência aproveitar algum do trabalho efetuado em outras experiencias, anteriormente vistas, este incidiu sobre um conjunto de textos completamente diferente, o Corpus utilizado foi em Inglês, o tamanho do nosso conjunto de teste passou de 102 documentos de texto para 22186 documentos de texto, pretendeu-se obter além da média e do desvio padrão da similaridade documental para os documentos que continham plágio e para os documentos que não continham. Outros valores que obtivemos foram os valores das medidas que são habitualmente utilizados para a determinação da qualidade de sistemas de classificação, tais como a *Precision*, o *Recall*, e a *F-Measure* [23].

	Plágio	not Plágio
Plágio^	TP	FN
not Plágio	FP	TN

Tabela 3 – Tabela de Confusão

$$Recall = \frac{TP}{TP + FP} \tag{10}$$

$$Precision = \frac{TP}{TP + FN} \tag{11}$$

$$F - Measure = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{12}$$

Numa segunda etapa (Figura 14) surgiu uma segunda abordagem, esta abordagem difere no processamento dos documentos em objetos Document. Nesta fase o vocabulário é processado normalmente, o que vai mudar é a escolha das palavras relevantes como será observado em seguida.

1) Processamos o conjunto de documentos de texto em objetos Document.
    a) Processamento do vocabulário dos documentos de texto do conjunto de teste.
    b) Um documento de texto é dividido por frases e em cada frase não são calculadas as importâncias através da nossa função de "Importância do termo" de todas as palavras, apenas as três palavras mais importantes foram copiadas para a nossa estrutura de dados de palavras mais importantes, no final do processo tivemos uma estrutura de dados com as três palavras mais importantes de cada frase sem ocorrerem repetições de palavras. Neste ponto também calculámos a norma documental e guardamos essa informação no objeto Document.
    c) Todos os ficheiros .txt são guardados em objectos Document e salvos em um ficheiro binário, isto para garantir que apenas é feito uma vez este processamento e sempre que necessário a sua utilização faz-se o carregamento do ficheiro binário para a memoria.

A segunda etapa da segunda abordagem processa-se da mesma forma que a da primeira abordagem por isso não voltamos a descrever o processo.

Nesta metodologia realçamos que através dos testes efetuados nesta conseguimos o valor limiar para decidir se um par de documentos contém ou não plágio. Com esta abordagem das três palavras por frase obtivemos os seguintes valores limiar:

- Para documentos com plágio: 0,07317±0,06906
- Para os documentos sem plágio: 0,01841±0,01212

O valor limiar de decisão que resolvemos adotar foi de 0,02.

3.4 Similaridade em Grandes Coleções — *"Overlap"* e *"CosSim"*

Esta experiência utiliza o mesmo conjunto de documentos que a experiencia anterior utilizou, mas o processo para atingir os fins é diferente, passamos por fazer algumas optimizações ao nosso método anterior. Apresentamos duas abordagens diferentes no cálculo da Similaridade Documental a "Overlap" e a "CosSim" como vamos observar de seguida.

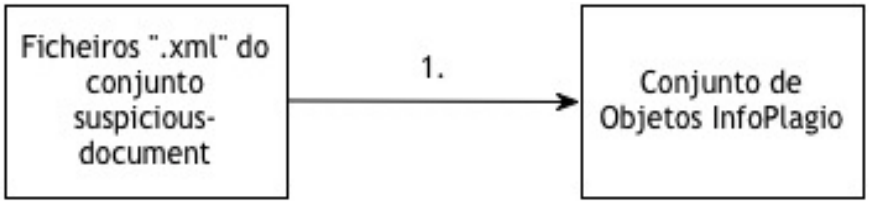

Figura 15 - Primeira etapa do Método "Overlap" e "CosSim"

1) Na primeira etapa (Figura 15) vamos processar todos os ficheiro ".xml" que se encontrem no conjunto suspicious-Documents transformando-os em objetos InfoPlagio, estes objetos iriam conter a informação que foi vista na experiencia anterior (nome do documento suspeito, inicio do possível segmento de plágio e tamanho do segmento, assim como a informação referente ao source-document onde estaria o possível segmento de plágio, o caracter em que começava o possível plágio e o tamanho do segmento).

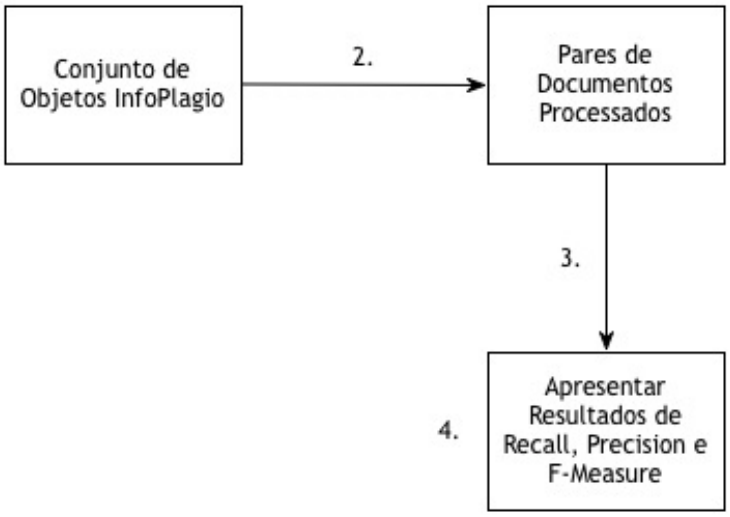

Figura 16 - Segunda etapa do método "Overlap" e "CosSim"

2) Na segunda etapa (Figura 16) fizemos um teste com 2000 pares de documentos, 1000 pares de plágio e 1000 pares sem conterem plágio.
   a) Em tempo de processamento, processamos o par de documentos em objetos Document, ou seja, para cada objeto Document processamos as palavras mais importantes segundo a nossa *função de Importância do Termo* proposta na Experiencia Base (explicada na secção 3.1) e guardamo-las na nossa estrutura de palavras mais importantes.
3) Neste passo apresentamos duas abordagens a) e b) de calculo de similaridade entre dois documentos. Consequentemente a *Precision*, *Recall* e *F1-Measure* terão valores finais diferentes, mostrados no secção 4.4 .
   a) A primeira abordagem para o cálculo da similaridade entre pares de documentos denominamos por “CosSim”. Neste passo intermédio a similaridade entre dois documentos era calculada segundo o critério: se as palavras se encontrassem nos dois documentos, então o seu valor de importância era somado e acumulado, após percorridas todas as palavras importantes da nossa estrutura, este valor

acumulado era divido pelo produto das normas de similaridade dos dois documentos.

b) "Overlap" foi como denominamos a segunda abordagem e o critério utilizado foi: se as palavras mais importantes estivessem em ambos os documentos contávamos como uma "colisão" ou um "overlap", após percorrer todo um documento temos o número total de colisões entre palavras importantes dos dois documentos. Este valor dividíamos por o número de palavras mais importantes do documento que contivesse menor número de palavras importantes.
c) Cálculo da performance, em termos de *Precision*, o *Recall* e a *F-Measure*.

4) Apresentamos os resultados das medidas de *Precision*, *Recall* e *F-Measure*.

No capítulo 4 apresentamos e interpretamos os resultados obtidos aplicando a abordagem "Overlap" e a "CosSim" para o cálculo da Similaridade. Estes resultados são importantes porque demonstram a resposta que as nossas abordagens dão como classificação dos ficheiros, assim como um contributo para o objectivo final — encontrar segmentos de plágio em documentos.

3.5 Determinação das Zonas de Plágio

Como foi visto na experiência anteriormente explicada, neste método utilizamos o mesmo conjunto de documentos — Grandes Coleções — e damos enfâse a alguns dos aspectos contidos nessa mesma experiência, pois são parte fulcral para esta experiência, sendo assim, a abordagem tem em conta os documentos .xml como vemos de seguida.

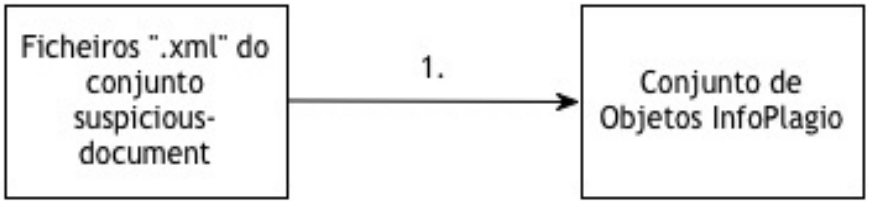

Figura 17 - Primeira etapa do método de Determinação das Zonas de Plágio

A primeira etapa (Figura 17) é igual à experiencia anterior, ou seja, processamos os ficheiros .xml do conjunto de suspicious-documents em um conjunto de objetos InfoPlagio. Neste método atingimos o objectivo pretendido, além de identificar os documentos que contêm plágio, obtemos as zonas onde existe plágio, estas zonas são necessárias para o avaliador final (humano) as poder classificar.

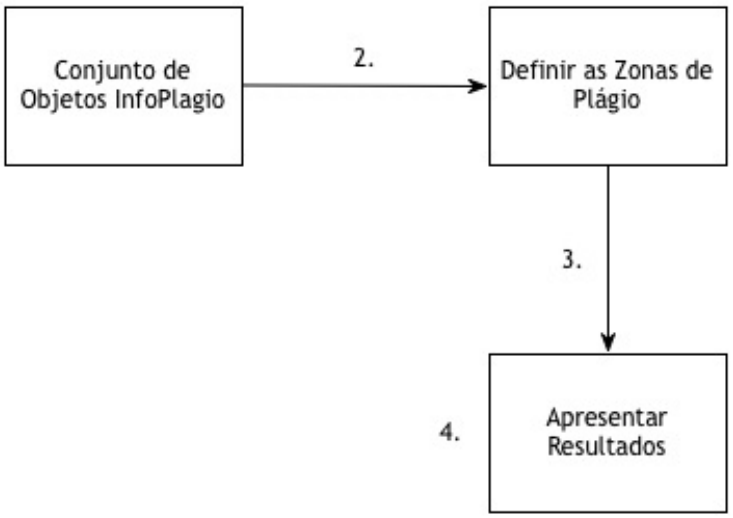

Figura 18 - Segunda etapa do Método de Determinação das Zonas de Plágio

Esta segunda etapa (Figura 18) do método difere bastante da experiencia anterior. O objectivo passa por classificar se um par de documentos contêm plágio e depois identificar os segmentos plagiados. Aqui vamos focar-nos só

na determinação dos segmentos de plágio, dentro de documentos que já sabemos à partida que contém plágio, de um determinado documento.

2) Numa primeira fase tínhamos obtido o conjunto de objetos com a informação dos documentos que continham os possíveis exemplos de plágio, ao termos o suspicious-document e o source-document (pares de plágios), transformávamos o texto de cada documento de modo a ficar todo dividido por frases.

   No passo 2. da Figura 18 vamos definir as zonas de plágio mas necessitamos fazer alguns passos intermédios.

   a) Começamos por criar um valor de threshold, este valor é um valor de referência necessário para filtrar as frases que são possíveis plágios nos documentos. Para gerarmos o valor de threshold aplicamos uma função que calcula a média e o desvio padrão da similaridade entre as frases do suspicious-document com o source-document. Após obtermos estes valores (média e desvio padrão) o valor de threshold será definido por:

$$threshold = média + (factor * desvio\ padrão) \tag{13}$$

   a soma da média com o produto de um factor com o desvio padrão, o factor é um valor previamente definido. Em seguida aplicamos a medida de similaridade que é calculada a partir da sobreposição exclusiva entre as palavras das duas frases:

$$simFrases(s_1, s_2) = e^{\frac{1}{4}\log(p_1) + \frac{3}{4}\log(p_2)} \tag{14}$$

   sendo $p_1 = \frac{NL}{\min\{|s_1|,|s_2|\}}$ e $p_2 = \frac{NL}{\max\{|s_1|,|s_2|\}}$

em que s_1 *e* s_2 são duas frases, $|s_1|$ *e* $|s_2|$ tamanho de cada frase, em número de palavras, e NL é o número de ligações excusivas que existe entre as palavras das duas frases, tal como exemplificado no exemplo a seguir.

Exemplificando, dadas duas frases:

- *"Apple vende duas vezes mais iPhone 5S que 5C nos EUA"*

- *"Nos EUA o iPhone 5S é vendido mais vezes que o 5C"*

A 1ª frase (s_1) tem 11 palavras;

A 2ª frase (s_2) tem 12 palavras;

NL = 8, existem 8 ligações exclusivas de palavras;

logo,

$$p_1 = \frac{7}{11} \text{ e } p_2 = \frac{8}{12}$$

$$simFrases(s_1, s_2) = e^{\frac{1}{4}\log\left(\frac{8}{11}\right) + \frac{3}{4}\log\left(\frac{8}{12}\right)}$$

$$\Leftrightarrow simFrases(s_1, s_2) = e^{\frac{1}{4}\log(0{,}727272) + \frac{3}{4}\log(0{,}666666)}$$

$$\Leftrightarrow simFrases(s_1, s_2) \approx 0{,}8$$

Após obtermos o valor de threshold vamos processar os pontos que vão ser necessários para definir as zonas de plágio, numa matriz de pontos relativos às similaridades entre as frases do suspicious e as do

source. Um ponto é constituído pelo índice da frase do suspicious-document (por exemplo para o ponto A, a coordenada de índice 2 — Figura 19), pelo índice da frase do source-document (por exemplo para o ponto A, a coordenada de índice 3 — Figura 19) ponto A de coordenadas (2,3) e pelo valor de similaridade entre as duas frases que não está representado mas é calculado como explicado anteriormente através de (número da fórmula).

As zonas de plágio são então definidas por um conjuntos de pontos que se encontram na área de uma zona de plágio dada por o dx e dy, a partir dos pontos extremos A(2,3) e B(8,4) conseguimos saber dx (8 – 3 + 1 = 6) e dy (4 – 2 +1 = 3) . Recorrendo à Figura 19 podemos observar o que acabámos de explicar.

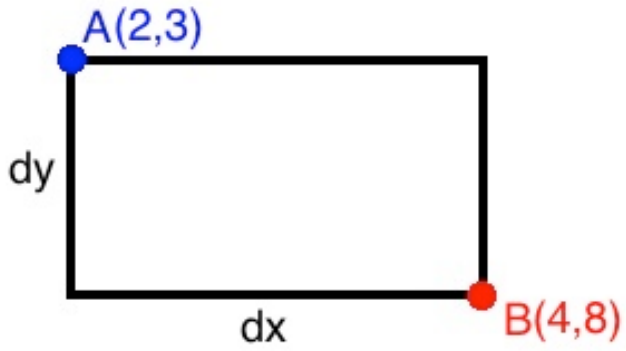

Figura 19 - Definição de Zona de Plágio

Voltando ao método, como vimos o ponto era definido por um valor de similaridade, se este valor de similaridade for maior que o threshold calculado anteriormente, o ponto constituído pelo valor de similaridade e pelos índices das frases do suspicious e source document é adicionado à nossa estrutura de pontos para definir as zonas de plágio.

Após termos processado todos os pontos dos documentos (source-document e suspicious-document) que estão a ser testados necessitamos definir uma ou várias zonas de plágio.

b) Vamos aglomerar os pontos em zonas de plágio seguindo um determinado critério. No nosso caso para verificarmos se um ponto pertencia a uma zona de plágio usámos a distância euclidiana, se a distância euclidiana entre o ponto que está a ser testado e o ponto que tem menor distância pertencente a uma determinada zona de plágio já construída anteriormente for menor que a distância previamente definida por nós, então esse ponto é adicionado à zona de plágio, caso não seja, verificamos as zonas de plágio seguintes.
   Ao percorrermos todos as zonas e se chegarmos à situação em que o ponto não foi aglomerado em alguma zona, criamos uma nova zona e adicionamos o ponto a essa nova zona.
   Ao aglomerarmos todos os pontos a zonas o resultado final deste passo intermédio é uma listagem de zonas de plágio.

Após a criação das zonas de plágio passamos à eliminação das zonas que não eram relevantes ou que não continham plágio. Usámos dois critérios para excluir zonas irrelevantes. Se a zona tivesse menos que dois pontos seria eliminada e se os valores de similaridade dos pontos fossem menor que um valor pré-definido (exemplo: 0,25) então também seriam eliminadas estas zonas.

3) Por fim testávamos se as zonas geradas pelo nosso sistema eram iguais ou parecidos ao resultado dado pela PAN11. Para atingir o resultado intersectamos a área das nossos zonas com a área das zonas fornecidos pela PAN11. A área de uma zona era dada como foi visto anteriormente. Caso houvesse intersecção era contabilizado como um resultado positivo. Após a execução dos n pares de plágio teríamos como resultado a eficiência do nosso sistema, na determinação destas zonas críticas.

Ao definirmos zonas de plágio pretendemos ajudar o avaliador (humano) final a verificar com mais facilidade se realmente é uma zona de plágio. Em documentos pequenos não aparenta grande problema mas suponhamos que temos um documento com muitas frases, por exemplo 30 mil frases, e neste documento existem 5 zonas de plágio. Através do nosso método, com estes resultados, vamos encaminhar o avaliador a verificar as zonas identificadas de modo prático, rápido e seguro, não necessitando de olhar para todo o documento mas restringindo-se apenas a estas zonas identificadas. Na literatura revista, não encontramos metodologias, que rastreassem documentos de modo a apresentar resultados de acordo conforme apresentamos, utilizando uma metodologia simples, sem necessitar de grande maquinaria para o processamento.

No capítulo 4 apresentamos os resultados obtidos nos testes efetuados por esta experiencia.

3.6 Sumário

Ao longo deste capítulo tivemos a oportunidade de descrever os métodos, começando pela abordagem mais simples face a um conjunto reduzido de documentos, passando por métodos intermédios no qual falhava algum pormenor até ao método no qual atingimos os objectivos que pretendíamos, foi feito um sumário global do trabalho mais prático que elaboramos ao longo do desenvolvimento da dissertação de mestrado assim como os caminhos que decidimos tomar na tentativa de detectar plágio entre documentos.

4. Resultados Obtidos

Neste capítulo iremos apresentar nos resultados obtidos, nas diferentes etapas e testes efectuados ao longo do desenvolvimento da dissertação. Na secção 4.1 e 4.2 estão presentes os resultados obtidos na Experiencia Base explicada na secção 3.1. Na 4.3 os resultados da Experiencia Eficiente explicada na secção 3.2, e mantendo esta ordem numérica nas diversas experiencias, em cada secção serão explicados os resultados por fim na ultima secção é feita a sumarização do capítulo 4.

Todos estas experiencias e medições que são enunciados foram efectuadas numa máquina MacBook Pro 13 polegadas do final de 2011, contendo um processador a 2,4 GHz core i5, 4GB de memória a 1333MHz DDR3 e uma gráfica Intel HD 3000 384MB. Portanto uma máquina convencional, contrariamente à grande maioria que em muitos casos usam super computadores e paralelismo para atacar o problema [13 — 16]

4.1 Resultados referentes à "Experiencia Base" — Parte 1

Como primeira experiência usámos o pequeno conjunto de documentos de texto (102) no qual apenas verificamos os valores de similaridade obtidos pelos supostos pares de documentos que seriam plágios. Na Tabela 4 apenas estão algumas percentagens de palavras utilizadas no teste de cálculo de Similaridade, na Figuras 20 e 21 estão representados o Gráficos dos Resultados dos testes.

Textos	5% Palavras	20% Palavras	50% Palavras	Totalidade
PLG01.txt|PLG02.txt	0,99999	0,99999	1,00000	0,99999
PLG03.txt|PLG04.txt	0,99999	0,98805	0,99471	0,99286
PLG05.txt|PLG06.txt	0,99431	0,99407	0,98844	0,98692
PLG07.txt|PLG08.txt	1,00000	1,00000	0,99993	1,00000
PLG09.txt|PLG10.txt	1,00000	0,99792	0,99855	0,99814
PLG11.txt|PLG12.txt	0,99999	0,99623	0,99623	0,99535
PLG13.txt|PLG14.txt	1,00000	0,99805	0,99647	0,99502
PLG15.txt|PLG16.txt	0,97633	0,95736	0,96573	0,95777
PLG17.txt|PLG18.txt	1,00000	0,98249	0,98049	0,97549
PLG19.txt|PLG20.txt	0,99663	0,99002	0,99195	0,98831

Tabela 4 – Resultados dos valores de Similaridade de Documentos Plagiados

Na Tabela 4 estão os resultados detalhados de 10 pares de documentos que foram testados, estes documentos são praticamente cópias, por exemplo o PLG01.txt é plagiado do PLG02.txt, o PLG03 é plagiado do PLG04, etc. A primeira coluna contém os identificadores dos pares de documentos, da segunda à ultima coluna estão os valores de similaridade entre os documentos consoante as percentagens de palavras utilizadas, indicadas na primeira linha. Pegando na segunda linha como exemplo:

Textos	5% Palavras	20% Palavras	50% Palavras	Totalidade
PLG01.txt|PLG02.txt	0,99999	0,99999	1,00000	0,99999

PLG01.txt|PLG02.txt é o identificador. Neste caso foi testado o documento PLG01.txt com o documento PLG02.txt

Na célula da segunda linha e segunda coluna está o valor de 0,99999 correspondente à similaridade documental entre os dois documentos

utilizando 5% das palavras mais importantes/relevantes para os dados documentos. Na célula definida pela segunda linha e terceira coluna temos o valor de 0,9999 de similaridade documental utilizando 20% de palavras mais importantes/relevantes para os dados documentos e assim sucessivamente.

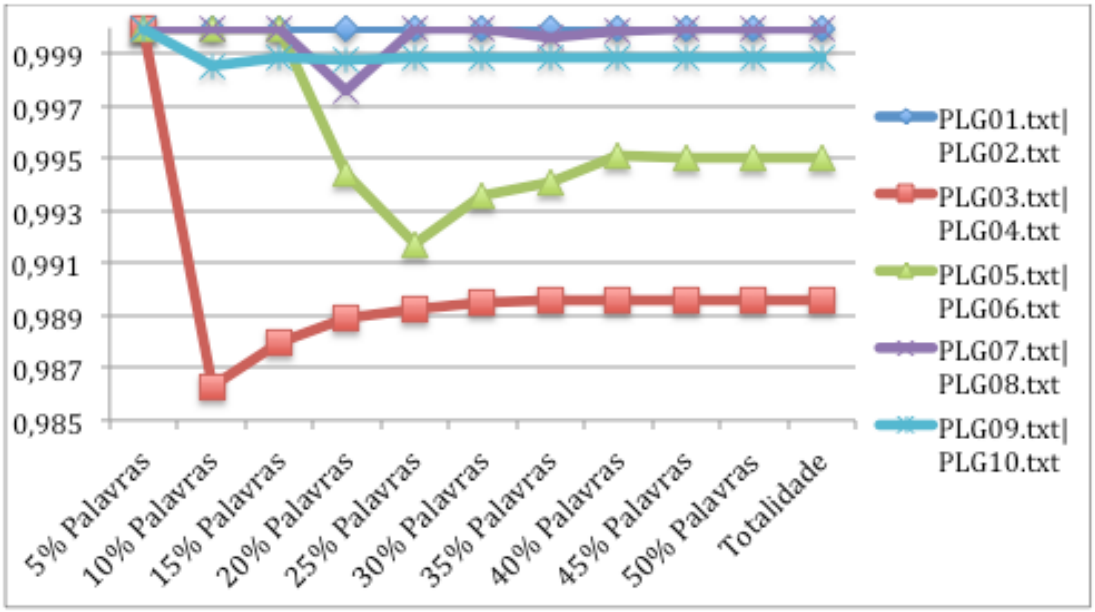

Figura 20 - Gráfico de Resultados dos valores de Similaridade de Documentos Plagiados 1 — 10

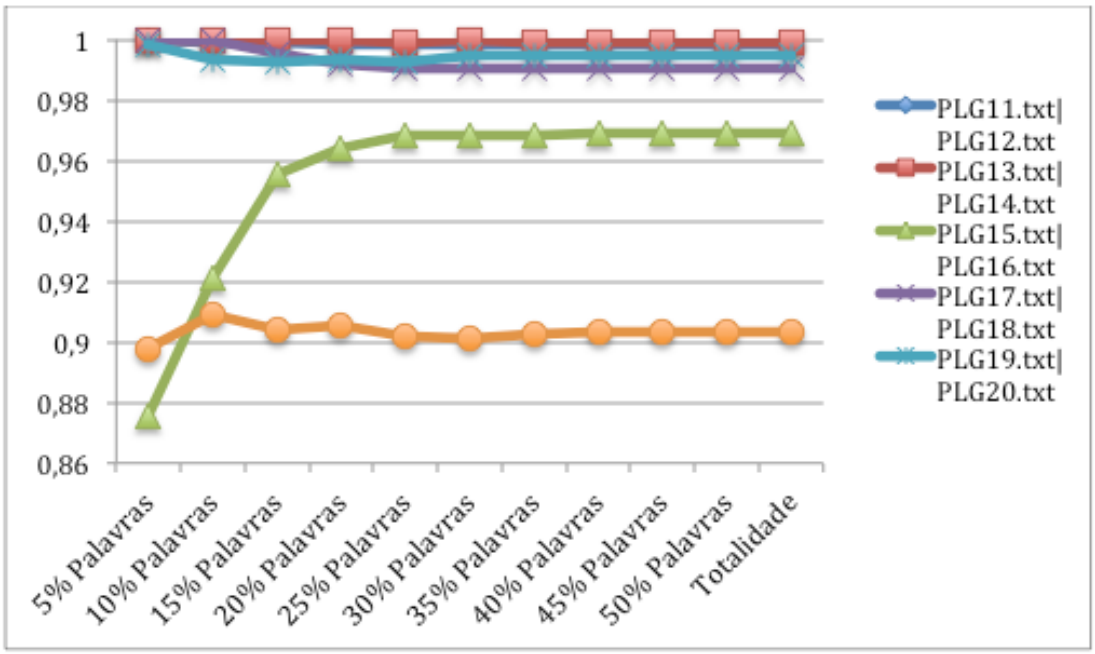

Figura 21 - Gráfico de Resultados dos valores de Similaridade de Documentos Plagiados 11 — 20

Relativamente aos gráficos, no eixo do y temos o valor de similaridade documental e no eixo do x a percentagem de palavras mais importantes utilizadas, os gráficos presentes na Figura 20 e 21 representam o comportamento dos valores da Tabela 4, com maior detalhe.

Ao observarmos a Tabela 4 e os gráficos das Figuras 20 e 21 as nossas suspeitas revelaram-se certas, a partir de certo valor de percentagem de palavras importantes/relevantes utilizadas no processamento do cálculo de similaridade documental a tendência destes valores de similaridade tende a estabilizar-se.

4.2 Resultados Referente à "Experiência Base" — Parte 2

Num segundo teste fizemos uma tentativa para chegarmos a um valor de referencia para o número ou percentagem de palavras a utilizar num futuro conjunto de Documentos com um tamanho bastante maior que o

tamanho actual (102 Documentos) , efetuámos testes com documentos que à partida seriam semelhantes, e obtivemos a Tabela 5 e os gráficos das Figuras 22 e 23.

Textos	5% Palavras	20% Palavras	50% Palavras	Totalidade
SIM01.txt|SIM02.txt	0,02538	0,24560	0,23598	0,23597
SIM03.txt|SIM04.txt	0,65266	0,57089	0,54288	0,54289
SIM05.txt|SIM06.txt	0,76068	0,60323	0,56078	0,56069
SIM07.txt|SIM08.txt	0,64964	0,59685	0,60483	0,60483
SIM09.txt|SIM10.txt	0,88622	0,87462	0,87635	0,87635
SIM11.txt|SIM12.txt	0,48359	0,43370	0,43728	0,43728
SIM13.txt|SIM14.txt	0,76565	0,77001	0,77972	0,77972
SIM15.txt|SIM16.txt	0,72687	0,69335	0,69142	0,69142
SIM17.txt|SIM18.txt	0,49167	0,35796	0,34229	0,34229
SIM19.txt|SIM20.txt	0,59194	0,51643	0,53405	0,53406

Tabela 5 - Resultados dos valores de Similaridade dos Documentos Similares

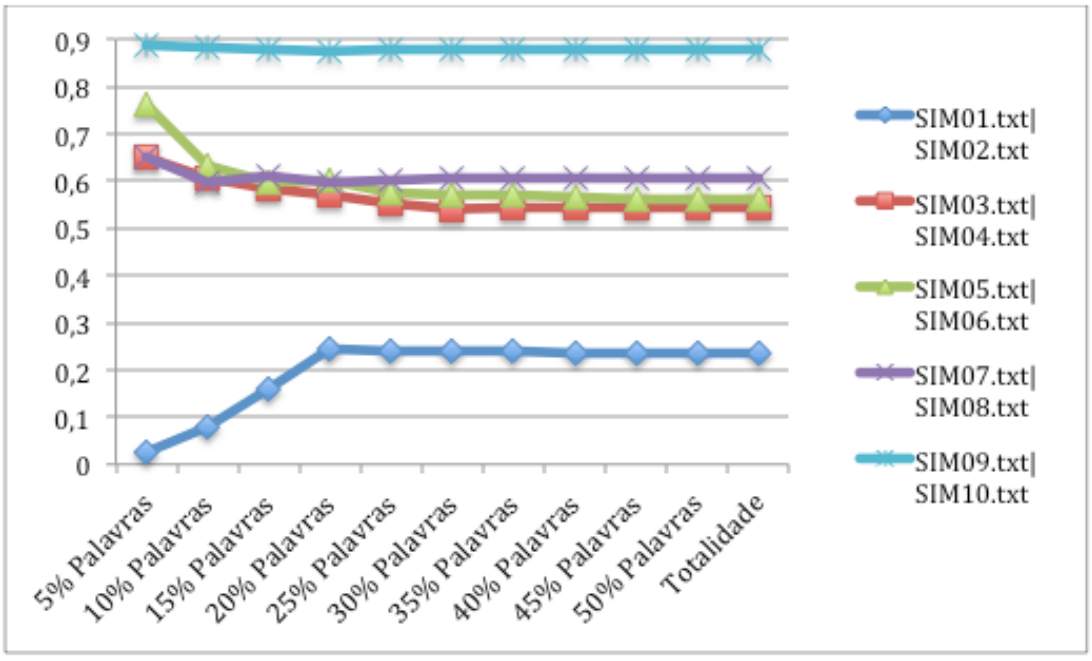

Figura 22 - Gráfico de Resultados dos valores de Similaridade dos Documentos Similares 1 — 10

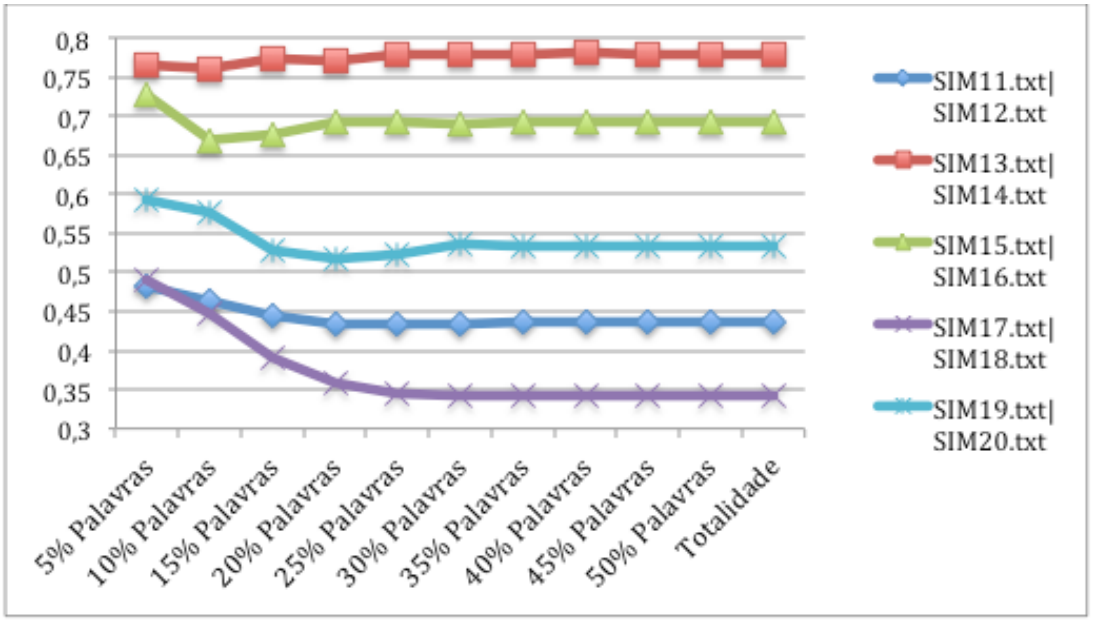

Figura 23 - Gráfico de Resultados dos valores de Similaridade dos Documentos Similares 11 — 20

Os resultados desta experiencia foram semelhantes aos da experiencia anterior, apenas difere nos valores de similaridade visto que o teste foi feito com pares de documentos em que o conteúdo era semelhante e a experiencia anterior com documentos plagiados. Mas o que se pretendeu verificar foi a tendência dos valores de similaridade documental a estabilizar.

Olhando por exemplo para o par de documentos SIM15.txt e SIM16.txt, a partir dos 20% — 25% de palavras importantes utilizadas para o cálculo da similaridade o valor tende a estabilizar por volta do 0,69.

Após cuidada análise concluímos então que o nosso valor de referencia para o número de palavras representativo por documento, necessárias para o cálculo da similaridade documental, eram os 20% de palavras. Concluímos que esta era uma percentagem que nos permitia atingir resultados relativamente corretos face às diferentes percentagens de palavras, ajudando a combater o problema de "custo computacional" no caso se conjunto de Documentos de texto aumentar substancialmente.

Na Tabela 6 apresentamos apenas os valores da Similaridade Documental Simétrica. Similaridade Documental Simétrica baseia-se em testar se ao aplicarmos o método que nos dá a similaridade documental, os resultados são iguais, este resultado era esperado e com os resultados confirmámos o pretendido.

Similaridade Simétrica	
Textos	**10% Palavras**
SIM03.txt|SIM04.txt	0,60547
SIM04.txt|SIM03.txt	0,60547

Tabela 6 - Exemplo de Similaridade Documental Simétrica

Por fim ainda fizemos testes entre documentos que eram diferentes apenas para verificarmos os valores de similaridade. Como esperado estes deram valores bastante baixos — Tabela 7.

Textos	5% Palavras	20% Palavras	40% Palavras	Totalidade
DIF05.txt	0,00000	0,00000	0,00844	0,00844
DIF10.txt	0,00000	0,00000	0,00188	0,00193
DIF15.txt	0,00000	0,00000	0,00000	0,00016
DIF20.txt	0,00000	0,00000	0,00013	0,00016
DIF25.txt	0,00000	0,00000	0,00396	0,00400
DIF30.txt	0,01166	0,01021	0,01216	0,01219
DIF35.txt	0,00000	0,00000	0,00064	0,00080
DIF40.txt	0,00000	0,00561	0,00584	0,00596
DIF45.txt	0,00000	0,00000	0,00290	0,00306
DIF50.txt	0,00000	0,00000	0,00212	0,00216
DIF55.txt	0,01461	0,01244	0,01373	0,01374
DIF60.txt	0,00000	0,00000	0,00309	0,00636

Tabela 7 - Valores relativos ao Cálculo da Similaridade no Conjunto de Documentos Diferentes

Na experiencia relativa à Tabela 7, pegámos no documento DIF01.txt e testámos este documento com os restantes documentos do conjunto de documentos diferentes. Através dos resultados obtidos podemos ver que os valores de similaridade documental são valores muito baixos, o que veio comprovar que os documentos que são realmente de conteúdos que não têm a ver uns com os outros.

Este conjunto de experiências levaram-nos a obter um valor de referencia para a escolha do número de palavras mais importantes a considerar nos documentos, assim como podemos observar comportamento dos valores de pares de documentos que são plágios, documentos que o seu conteúdo é similar e documentos em que o conteúdo difere praticamente na totalidade

4.3 Resultados face à "Experiência Base" e à "Experiencia Eficiente" — Rank

O teste desta experiência baseou-se na classificação da qualidade de um ranking, como enunciámos anteriormente na secção 3.2. Um ranking está ordenado decrescentemente consoante o valor de similaridade de um par de documentos de textos e os nossos resultados vão classificar a qualidade de um ranking face ao ranking perfeito. Relembrar que "ranking perfeito" é o ranking gerado utilizando a totalidade das palavras de cada Documento. Os rankings que foram testados foram rankings gerados a partir de uma determinada percentagem de palavras relevantes usadas. No caso de 5% de palavras o ranking seria denominado por Rank05, no caso de 10% de palavras relevantes Rank10 e seguindo o mesmo formato para outras percentagens de palavras relevantes. Na Tabela 8 estão representados o ranking perfeito e o ranking com 20% de palavras relevantes por documento de texto.

RankingPerfeito	Ranking20
PLG01.txt|PLG02.txt = 1,00000	PLG01.txt|PLG02.txt = 0,99999
PLG07.txt|PLG08.txt = 1,00000	PLG13.txt|PLG14.txt = 0,99999
PLG09.txt|PLG10.txt = 0,99855	PLG07.txt|PLG08.txt = 0,99999
PLG13.txt|PLG14.txt = 0,99687	PLG09.txt|PLG10.txt = 0,99876
PLG11.txt|PLG12.txt = 0,99595	PLG11.txt|PLG12.txt = 0,99840
PLG03.txt|PLG04.txt = 0,99167	PLG05.txt|PLG06.txt = 0,99461
PLG05.txt|PLG06.txt = 0,99003	PLG19.txt|PLG20.txt = 0,99370
PLG19.txt|PLG20.txt = 0,98989	PLG17.txt|PLG18.txt = 0,99191
PLG17.txt|PLG18.txt = 0,98142	PLG03.txt|PLG04.txt = 0,98880
PLG15.txt|PLG16.txt = 0,96372	PLG15.txt|PLG16.txt = 0,96433
PLG21.txt|PLG22.txt = 0,84422	PLG21.txt|PLG22.txt = 0,90745
SIM09.txt|SIM10.txt = 0,78936	SIM09.txt|SIM10.txt = 0,87384
SIM13.txt|SIM14.txt = 0,65442	SIM13.txt|SIM14.txt = 0,76081
SIM15.txt|SIM16.txt = 0,58427	SIM15.txt|SIM16.txt = 0,69467
SIM17.txt|SIM18.txt = 0,53853	SIM17.txt|SIM18.txt = 0,61791
SIM07.txt|SIM08.txt = 0,53506	SIM07.txt|SIM08.txt = 0,59503
SIM05.txt|SIM06.txt = 0,39669	SIM05.txt|SIM06.txt = 0,58333
SIM03.txt|SIM04.txt = 0,37255	SIM03.txt|SIM04.txt = 0,57004

SIM19.txt|SIM20.txt = 0,36134	SIM01.txt|SIM02.txt = 0,54508
SIM01.txt|SIM02.txt = 0,35305	SIM19.txt|SIM20.txt = 0,51305
SIM11.txt|SIM12.txt = 0,35056	SIM11.txt|SIM12.txt = 0,43528

Tabela 8 - RankingPerfeito e Ranking20

Para esta representação demos 3 resultados possíveis:

1. Desvio Total = 0.18561759
2. Média dos desvios linha a linha = 0.00883893
3. Qualidade do RankingPerfeito|Ranking20 = 0.09401560

O desvio total refere-se a todos os elementos que se encontram no ranking, visto que o nosso ranking contém 100 pares de documentos de texto, a média dos desvios linha a linha é a que realmente nos interessa porque os pares aos quais estamos a dar importância são os que se encontram nas primeiras 21 posições, o que nós esperámos é que se encontrem os 11 pares que contêm plágio mais os 10 pares que são documentos similares, como podemos observar os resultados são os que pretendemos.

Os desvios são calculados através da diferença entre os valores do Ranking Principal com o Ranking20. Utilizámos como exemplo os primeiros três pares da Tabela 8:

PLG01.txt|PLG02.txt (1)

PLG07.txt|PLG08.txt (2)

PLG09.txt|PLG10.txt (3)

o cálculo efetuado para se calcular a qualidade do Ranking é:

$$QR = ((|RankingPerfeito_{(1)} - Ranking20_{(1)}|) + (|RankingPerfeito_{(2)} - Ranking20_{(2)}|) + (|RankingPerfeito_{(3)} - Ranking20_{(3)}|)) / np$$
$$= (|1{,}00000 - 0{,}99999|) + (|1{,}00000 - 0{,}99999|) + (|0{,}99855 - 0{,}99876|)$$
$$= 0{,}00023$$

Deste modo podemos confirmar que o desvio de qualidade é quase insignificante, se utilizarmos uma percentagem reduzida (20%) de palavras de cada texto, quando comparado com a totalidade das palavras.

4.4 Qualidade na Detecção de Pares de Plágio

Nesta secção apresentamos os resultados efectuados na experiencia da secção 3.4 através das duas abordagens efectuadas no cálculo da similaridade: a Overlap e a CosSim.

Para esta experiencia utilizámos um conjunto de 2000 documentos, 1000 documentos continham plágio os outros 1000 não continham. Para os documentos de plágio denominámos o conjunto por **conjunto R** e relativamente a este conjunto calculámos o *Recall*. Para os documentos sem plágio denominámos o conjunto por **conjunto P** e calculámos a *Precision* ao ter o valor destas duas medidas calculámos também a *F-Measure*.

Os testes foram ainda efetuados com diferentes valores de IMA (ver secção 3.4) relativa às palavras importantes que definiam um documento.

Nas Tabelas 9, 10 e 11 apresentamos os valores obtidos para as duas abordagens e com as diferentes IMAs utilizadas.

IMA > 7,0			
Abordagem	**Precision**	**Recall**	**F1-Measure**
CosSim	61,6	90,8	73,4
Overlap	67,4	78,5	72,5

Tabela 9 - Resultados das duas abordagens com IMA > 7,0

IMA > 11,0			
Abordagem	**Precision**	**Recall**	**F1-Measure**
CosSim	66	85,9	76,8
Overlap	66,7	81,7	73,4

Tabela 10 - Resultados das duas abordagens com IMA > 11,0

IMA > 15,0			
Abordagem	**Precision**	**Recall**	**F1-Measure**
CosSim	82,5	81,9	81,3
Overlap	78,4	72,6	75,4

Tabela 11 - Resultados das duas abordagens com IMA > 15,0

Como podemos observar independentemente qual seja a IMA a abordagem CosSim acaba por ter sempre melhor valor de Recall, esta era a medida à qual demos mais importância. Por exemplo se em 1000 pares de documentos existissem 100 que tinham plágio e com o nosso sistema conseguíssemos encontrar entre 80% a 95% dos possíveis documentos que continham plágio já era um resultado relativamente bom. Através da Tabela 11 podemos observar que a percentagem obtida de Recall foi uma percentagem relativamente interessante. Este valor é fundamental para a identificação se um documento contêm plágio. Só após esta fase de identificação de plágio em um documento podemos passa à identificação das zonas de plágio.

4.5 Qualidade na Detecção de Zonas de Plágio

Nesta secção apresentamos os resultados do método apresentado na secção 3.5, para a determinação das zonas de plágio, num documento suspeito, isto é entre que frases é que está localizado o plágio.

Para esta medição utilizamos 100 pares de documentos que contêm plágio e o nosso objectivo divide-se em dois. O nosso objectivo passa por encontrar quais as zonas em que realmente existia plágio. Uma "zona" é formada por um conjunto de frases classificadas como frases plagiadas dentro de um mesmo documento. Sendo assim um resultado é apresentar as frases do suspicious-document e as frases do source-documento, com o intuito que após esta seleção de frases passasse por um avaliador humano de modo a confirmar se estas zonas contêm realmente plágio. O plágio nos documentos testados pode ser classificado em vários tipos: plágio de fácil identificação, plágio de identificação intermédia e plágio de difícil identificação. O nosso sistema revelou-se apropriado a encontrar o plágio de fácil e média identificação.

Em 100 pares de documentos encontrámos 75% das zonas plagiadas. Na Figura 24 podemos observar um exemplo de zonas de plágio, em que a similaridade das frases toma valores elevados, também podemos observar na Figura 26 um exemplo de frases plagiadas em um documento suspeito a partir de um documento fonte.

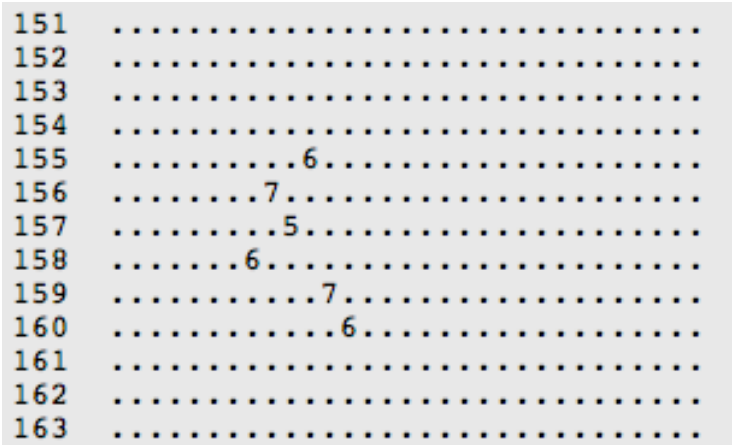

Figure 24 - Exemplo de uma Zona em que existe Plágio

Na Figura 24 apresentamos um exemplo em que podemos ver uma zona onde existe plágio. Nos números que se encontra no lado esquerdo estão representadas os índices das linhas correspondentes ao suspicious-document, os pontos são os índices das linhas do source-document, não representamos estes índices porque como podemos ver a zona começa na linha 155 do suspicious-document o que não nos permite ver os índices face ao source-document. Por exemplo, o valor da matriz da Figura 24, na posição (155,10) é 6. Este número indica a pontuação obtida pela semelhança entre as frases (neste caso será 0,6.. mas para se entender melhor mudámos a escala de 0 a 1 para 0 a 10). Basicamente este é um esquema onde podemos visualizar as possíveis zonas de plágio mais facilmente. Desenhando um retângulo imaginário (ou formando um cluster — Figura 25 — exemplo de um cluster gerado pelo sistema) podemos ver que nesta zona há fortes indícios de plágio. Através da Figuras 26 explicamos melhor como entender os resultados.

```
CLUSTER 0
    (160,  12,   0,61422)
    (159,  11,   0,73133)
    (158,   7,   0,67973)
    (157,   9,   0,55639)
    (156,   8,   0,73091)

    (155,  10,   0,65090)
```

Figure 25 - Exemplo de um conjunto de pontos onde Existe Plágio

Neste exemplo, a zona de plágio seria definida pelas frases de índice 155 a 160 do suspicious-document e das frases de índice 7 a 12 do source-document. os valores na escala entre 0 e 1 são os valores de similaridade entre as frases.

```
156     8     0,7309117

He had replaced ______ Burntwood , who was ___ indistinctly
He had replaced Magney _________ , who was now ___________

whirling away to the nearest railway point , Bowenville ,
whirling away to the nearest railway point , Bowenville ,

thirty-five miles distant.
thirty-five miles distant.
```

Figura 26 - Exemplo de frases Plagiadas

Por exemplo na Figura 26 o primeiro número (o 156) diz respeito à frase do suspicious-document (frase número 156) e o segundo número (o 8) ao source-document (frase número 8), o terceiro número (0,7309117) é o valor de semelhança das frases, portanto um valor alto, visto que este valor varia entre 0 e 1. Na Figura 24 transformámos os valores numa escala de 0 a 10 como foi dito anteriormente. Na Figura 26 a amarelo identificamos a frase que corresponde ao suspicious-document e a azul a frase que corresponde ao source-documente. Podemos ainda observar a correspondência entre as palavras alinhando as duas frases, apenas os espaços " _____ ", são palavras que não têm correspondente nas frases. Podemos assim perceber

que as zonas determinadas pelo sistema são zonas de frases plagiadas. Estas zonas que são dadas como resultado pretendem ser direcionadas a um avaliador que as avalie como plagiadas ou não, para uma confirmação mais fidedigna, visto que nem todo o tipo de plágio é igual e por vezes está mais escondido. O nosso sistema visa facilitar a identificação das zonas plagiadas, dando uma fácil percepção ao avaliador final — o utilizador.

4.6 Sumário

Neste capítulo debruçamo-nos sobre os resultados dos diferentes trabalhos efectuados ao longo da concepção desta dissertação de mestrado, apresentámos os resultados obtidos e explicámos o que pretendíamos.

5. Conclusões e Trabalho Futuro

Neste Capítulo final veremos algumas conclusões acerca do trabalho efectuado, dos testes e resultados, assim como no decorrer do desenvolvimento e escrita desta dissertação e também o trabalho futuro relativo às duas áreas no qual esta dissertação está focada. Na secção 5.1 são mencionadas as conclusões retiradas ao longo de um ano lectivo a trabalhar sobre as áreas da Similaridade Documental e Detecção de Plágio. Para finalizar teremos a secção 5.2 destinada a possíveis trabalhos futuros nestas duas áreas.

5.1 Conclusão

Este capítulo reflecte o trabalho explorado ao longo do desenvolvimento desta dissertação de mestrado assim como uma visão de algum trabalho que poderá ser efectuado em um futuro próximo. Como principal objectivo da dissertação pretendeu-se fazer um estudo das áreas da Similaridade Documental e da Detecção Automática de Plágio, com um principal foco nesta ultima, pois para um par de documentos pretendia-se verificar se havia plágio e caso houvesse, encontrar as zonas onde este estaria localizado, no interior dos documentos considerados.

Uma dificuldade actual na detecção automática de plágio situa-se nos conjuntos alargados de informação, ou seja, em grandes corpora de documentos de texto, por isso é uma área que ainda requer muita atenção e tende a ser muito explorada nos próximos anos. Prevê-se que nos próximos 10 a 15 anos esta área da detecção automática de plágio atinja a sua maturidade.

Como sabemos a detecção de plágio é uma área que actualmente está na ordem do dia, tendo sido relatados vários casos de pessoas cujas teses de doutoramento vieram a ser anulados por alegada fraude de conteúdos.

Detectar plágios em documentos não passa "só" por identificar se um documento é ou não de plágio, na nossa abordagem, aprofundámos esta questão e além de classificarmos se um documento continha plágio, obtivemos as zonas onde realmente havia plágio. Ao obtermos as zonas de plágio, dentro de um documento, estamos a facilitar o trabalho a um avaliador humano. Quando dois documentos com poucas frases contêm plágio, normalmente não é difícil o avaliador encontrar o plágio existente nos documentos, o problema principal é em documentos de alargado conteúdo, com muitas centenas e até milhares de frases. Como o nosso sistema dá como "output" as diversas zonas que contêm plágio, o avaliador necessita apena de verificar as zonas propostas como zonas de plágio e fazer a sua avaliação final. Apesar dos resultados encorajadores obtidos, com cerca de 75% de detecção de zonas de plágio, há aqui ainda espaço para se tentar melhorar esta detecção, deixando assim desde já essa indicação como trabalho futuro. Salientamos ainda que na literatura, existente até à data, não encontrámos nenhuma descrição deste tipo de procedimento.

Alguns pontos que são uma mais valia para o nosso trabalho podem ser referidos nos seguintes 4 parágrafos que descrevem sucintamente o nosso trabalho elaborado.

Com as experiências realizadas no desenvolver da dissertação pretendemos dar um maior enfâse a uma função que propusemos — função de Importância de Termos (ver secção 3.1) — que nos calculava a importância de um termo numa região, esta função tornou-se essencial para atingirmos o resultado final.

Foram efectuados vários testes no nosso trabalho, começando pelas tentativas de obter um valor "confortável" de palavras importantes por documento, sob a premissa de que palavras importantes/relevantes são bons representantes de documentos. Os 20% de palavras importantes para cada documento foi o valor encontrado. Após uma sucessão de experiências, concluímos que trabalhando apenas com 20% das palavras por documento conseguimos processar de forma eficaz a similaridade documental, reduzindo assim a carga computacional, quer a nível de espaço, quer a nível de tempo de processamento. Estes 20% era um valor referencia para maior parte dos documentos apenas detectámos uma dificuldade, que acabámos por resolver, num determinado tipo de documentos — os muito grandes. Nestes casos, a barreira dos 20% continuava a comportar um elevado número. Foi então que decidimos acrescentar uma filtragem extra, por importância mínima admitida (Secção 3.3). Com a combinação destes dois factores pudemos reduzir o número de palavras representativas em documentos grandes, para valores mais comportáveis.

Propusemos duas abordagem simples e eficientes para determinar se documentos continham ou não plágio — a "Overlap" e a "CosSim" Secção 3.4. Estas abordagens baseavam-se num valor de referencia gerado a partir de dois conjuntos de documentos — documentos com plágio e sem plágio — ou seja, para posteriormente podermos classificar qualquer outro documento a ser testado o valor de similaridade documental era considerado. Para cada conjunto de documentos calculávamos a média de similaridade e o desvio padrão. O valor de referência seria a média da soma da média de similaridade com o desvio padrão. Com este método tivemos a possibilidade de calcular o *Recall*, a *Precision* e a *F-Measure*, medidas que habitualmente são utilizadas para a determinação da qualidade de um sistema de classificação. Assim conseguimos obter uma avaliação do sistema em termos de detecção de plágio em documentos.

Quando desenvolvemos a abordagem da secção 3.5 na qual determinamos a qualidade ao determinar zonas de plágio aplicámos uma fórmula para o cálculo da similaridade entre frases utilizando uma média geométrica ponderada. A forma como encontramos as zonas de plágio não se encontra na literatura o que é uma mais valia para o nosso trabalho.

5.2 Trabalho Futuro

Ao longo do ano curricular a desenvolver a dissertação de mestrado o trabalho incidiu sobre as áreas da Similaridade Documenta e da Detecção Automática de Plágio com especial enfâse nesta ultima. O nosso objectivo passou por criar um método simples que numa primeira fase fosse capaz de identificar a existência de plágio em documentos e numa segunda fase fizesse a pesquisa por zonas de plágio. Estes objectivos foram atingidos e os resultados acabaram por serem bons, mas acreditamos que a nossa metodologia ainda possa sofrer algumas alterações, de modo a optimizarmos o algoritmo tornando-o mais eficaz e eficiente, mas mantendo a simplicidade de modo a ser possível a utilização de máquinas convencionais.

Uma abordagem que poderia ser interessante implementar era uma abordagem que utilizasse a indexação de documentos. Através do motor de indexação, por exemplo o Lucene [24], numa primeira fase da abordagem, indexaríamos os documentos com potencial plágio relativamente a um determinado documento suspeito em análise e após esta indexação, numa segunda fase aplicaríamos o nosso método de seleção de zonas de plágio.

Ao longo da dissertação tentámos desenvolver uma técnica de detecção de plágio, as nossas experiências abordaram sempre a análise extrínseca, baseada na coocorrência na detecção, algo que se poderá fazer seria a

conjugação da análise extrínseca com a análise intrínseca. Esta tenta determinar zonas suspeitas de plágio mediante a alteração abrupta dos estilos de escrita.

Em termos de aplicações deste trabalho, uma possibilidade interessante seria o desenvolvimento de aplicações destinadas às instituições académicas, por exemplo uma plataforma de submissão de trabalhos escritos. No qual existiria uma aplicação que fizesse o trabalho de detecção de plágio e das passagens plagiadas e como output fornecesse os segmentos plagiados, esta ferramenta serviria de apoio aos docentes pois iria facilitar os docentes quando necessitam corrigir determinados trabalhos. Como sabemos o plágio é uma parte fulcral na correção de trabalhos e que tem vindo a ser muito explorado pelos alunos, como foi visto anteriormente.

Uma outra plataforma que poderia ser viável e estar inserida no mesmo contexto académico, seria uma ferramenta semelhante ao que foi referido no parágrafo anterior mas colocada ao dispor dos alunos com o intuito de ajudar a controlar a forma como estes estão a desenvolver os seus trabalhos, mostrando percentagens de plágio que se encontram no seu trabalho, visto que há certas "fronteiras" e mesmo valores convencionados, a partir dos quais se considera plágio.

Uma outra tentativa futura poderia passar por explorar técnicas de aprendizagem automática conjugada com a detecção automática de plágio. Através das técnicas de aprendizagem automática os sistemas seriam treinados de modo a reconhecer pares de documentos plagiados e posteriormente seria feita uma pesquisa exaustiva para encontrar as zonas de plágio.

Por último, mas não menos importante, deverá ser escrito um artigo de modo a apresentar o trabalho realizado, incluindo as experiências feitas e respectivos resultados (positivos e negativos), tendo também o objectivo de prevenir que outras pessoas que trabalham na área não sigam por

caminhos já explorados e com resultados menos positivos, dando um especial enfâse à nossa abordagem que nos deu resultados relativamente bons, através de métodos simples e inovadores, face às metodologias propostas por outros autores.

Referências

[1] P. Arvola, M. Junkkari, and J. Kekäläinen, "Generalized contextualization method for XML information retrieval," presented at the CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, 2005.

[2] H. Maurer, F. Kappe, and B. Zaka, "Plagiarism - A survey," *Journal of Universal Computer Science*, vol. 12, pp. 1050–1084.

[3] M. Potthast, A. Eiselt, and A. Barrón-Cedeno, "Overview of the 3rd international competition on plagiarism detection," Notebook Papers of CLEF 2011 Labs and Workshops, 19-22 September, Amsterdam, The Netherlands. ISBN 978-88-904810-1-7, 2011

[4] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," presented at the the fifth ACM SIGKDD international conference, New York, New York, USA, 1999, pp. 16–22.

[5] A. Huang, "Similarity measures for text document clustering," presented at the Proceedings of the Sixth New Zealand Computer Science Research Student Conference 2008, Christchurch, 2008.

[6] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[7] A. Aizawa, "The feature quantity," presented at the the 23rd annual international ACM SIGIR conference, New York, New York, USA, 2000, pp. 104–111.

[8] R. A. B. Yates and B. R. Neto, "Modern Information Retrieval", 1999.

[9] N. Sandhya, Y. S. Lalitha, A. Govardhan, and K. Anuradha, "Analysis of Similarity Measures for Text Clustering," *cscjournals.org*

[10] B. Bigi, "Using Kullback-Leibler distance for text categorization," presented at the ECIR'03: Proceedings of the 25th European conference on IR research, 2003.

[11] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," Apr. 2000.

[12] T. Liu and J. Guo, "Text similarity computing based on standard deviation," *Advances in Intelligent Computing*, 2005.

[13] J. Grman and R. Ravas, “Improved implementation for finding text similarities in large collections of data,” presented at the Proceedings of PAN, 2011.

[14] C. Grozea and M. Popescu, “The encoplot similarity measure for automatic detection of plagiarism,” *Notebook for PAN at CLEF*, 2011.

[15] G. Oberreuter, G. L'Huillier, and S. A. Rıos, “Approaches for intrinsic and external plagiarism detection,” presented at the Proceedings of the Notebook for PAN at CLEF 2011, 2011.

[16] D. Torrejón and J. Ramos, “Detailed Comparison Module In CoReMo 1.9 Plagiarism Detector,” *uni-weimar.de*

[17] A. Ghosh, P. Bhaskar, S. Pal, and S. Bandyopadhyay, “Rule Based Plagiarism Detection using Information Retrieval,” *Petras et al*, 2011.

[18] D. A. Rodríguez Torrejón and J. M. Martín Ramos, *Detección de plagio en documentos: sistema externo monolingüe de altas prestaciones basado en n-gramas contextuales*. Sociedad Española para el Procesamiento del Lenguaje Natural, 2010.

[19] D. A. Rodríguez-Torrejón and J. M. Martín-Ramos, “N-gramas de Contexto Cercano para mejorar la Detección de Plagio,” *users.dsic.upv.es*

[20] A. Barrón-Cedeño and P. Rosso, “On Automatic Plagiarism Detection Based on n-Grams Comparison,” in *Advances in Information Retrieval*, vol. 5478, no. 69, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 696–700.

[21] G. Oberreuter, S. A. Ríos, and J. D. Velásquez, “ FASTDOCODE: Finding Approximated Segments of N-Grams for Document Copy Detection Lab Report for PAN at CLEF 2010,” 2010.

[22] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, “N-gram-based author profiles for authorship attribution,” presented at the Proceedings of the Conference PACLING’03, Halifax, Canada, 2003.

[23] D. Powers, “Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, 2011.

[24] Lucene, "url: http://lucene.apache.org/core/"