Construção automática de histórias a partir de notícias

Building up a temporal storyline from the News

Carla Abreu
Faculdade de Engenharia da Universidade do Porto - LIACC - UP/Sapo Labs
cfma@fe.up.pt

Jorge Teixeira
Faculdade de Engenharia da Universidade do Porto - LIACC - UP/Sapo Labs
jft@fe.up.pt

Eugénio Oliveira
Faculdade de Engenharia da Universidade do Porto - LIACC - UP/Sapo Labs
eco@fe.up.pt

Resumo

Todos os dias são publicadas grandes quantidades de notícias *online*. Em particular, para que o leitor esteja ao corrente de um determinado acontecimento que ocorreu num determinado dia, este depara-se com o problema de selecionar entre um vasto conjunto de publicações. A situação agrava-se quando o mesmo pretende saber mais detalhes de uma história noticiosa particular que decorreu num intervalo temporal longo (p.ex. um mês). O trabalho desenvolvido e aqui descrito, surge para proporcionar ao leitor uma nova forma de "navegação"em histórias noticiosas compostas por notícias que aparecem dispersas no tempo e que se referem a um mesmo assunto. Mais propriamente, o objetivo deste trabalho é permitir a compreensão de sequências de notícias, através da construção automática de cadeias temporais de notícias relacionadas. A abordagem seguida no nosso trabalho é composta por três passos: (i) deteção de notícias similares; (ii) extração de termos chave; (iii) e criação de ligações entre notícias para a construção automática de histórias. A abordagem usada baseia-se na utilização de métodos de Processamento de Linguagem Natural, Extração de Informação, Reconhecimento de Entidades Mencionadas e na utilização de algoritmos supervisionados de aprendizagem automática. Foi realizado e analisado um elevado número de experiências descritas nas secções 4 e 5. Os resultados obtidos pela abordagem proposta na identificação de notícias duplicadas foi de 93.8%; e na construção de cadeias noticiosas de 93.1 %.

Foi ainda desenvolvida uma interface web para a navegação e exploração de cadeias noticiosas.

Palavras chave

Extracção de Informação, Aprendizagem Máquina, Processamento Linguagem Natural, Reconhecimento de Entidades Mencionadas, Jornalismo Computacional, Relacionamento Temporal de Informação

1 Introdução

Diariamente são publicadas grandes quantidades de notícias *online*, o que pode conduzir a uma sobrecarga de informação. Para estar ao corrente de uma determinada notícia, o leitor depara-se com um vasto conjunto de artigos noticiosos, artigos esses que, em muitos casos, descrevem um mesmo evento, podendo apresentar ou não variações textuais. A situação agrava-se quando o leitor pretende saber mais sobre uma dada história ou sequência de eventos. Um exemplo concreto é o desaparecimento do avião da *Malaysia airlines* a 8 de março de 2014. Para o dia 6 de outubro de 2014 a pergunta (*query*) *"avião Malaysia"* apresentada ao *Google News (news.google.pt)*, retorna uma lista com mais de 50 notícias relacionadas. Da leitura às notícias desse dia retira-se a informação de que as buscas pelo avião foram retomadas. Como é possível observar pelos seguintes títulos: *Retomadas buscas pelo avião da Malaysia Airlines* (Renascença, 06/10/2014) e *Recomeçam as buscas pelo avião desaparecido da Malaysia Airlines* (Jornal de Notícias, 06/10/2014) o evento noticiado é o mesmo, mas pelo facto das notícias serem provenientes de fontes noticiosas diferentes apresentam variações textuais.

Quando o leitor quer perceber a história do desaparecimento do avião como um todo, e informar-se sobre todos os eventos que se passaram relativamente a este acontecimento, a pergunta (*query*) *"desaparecimento Malaysia airlines"* sem delimitações temporais ao Google News apresenta mais de 4.500 resultados. Neste conjunto de resultados torna-se complicado ou até mesmo humanamente impossível a deteção de todos os eventos subjacentes a este acontecimento,

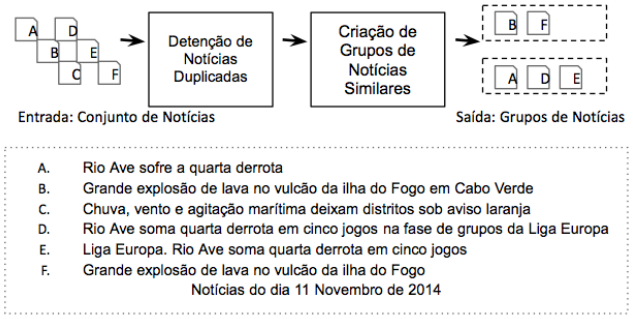

Figura 1: Deteção e agrupamento de notícias similares

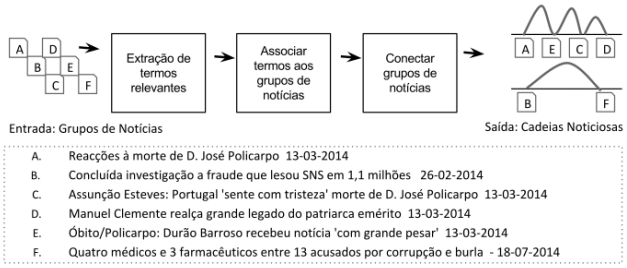

Figura 2: Construção de cadeias noticiosas

e por consequente, o leitor não consegue ter a perceção de toda a história, descrita em 4.500 notícias diferentes.

O objetivo deste trabalho é o de automaticamente detetar e agrupar notícias similares e o de automaticamente criar cadeias noticiosas temporais, proporcionando desta forma ao leitor uma nova forma de navegação entre eventos relativos a um mesmo acontecimento.

Com base na metodologia proposta, pretendemos numa primeira fase, detetar e agrupar notícias duplicadas (ver Figura 1). Para a realização desta tarefa foram utilizados: métodos de processamento de linguagem natural; algoritmos de medição de distância entre *strings*[1], para o cálculo da proximidade; e algoritmos supervisionados de aprendizagem automática, para a determinação da similaridade. Com vista à formação automática de cadeias noticiosas a nossa abordagem consistiu em extrair termos relevantes das notícias, que correspondem a palavras que nos sugerem entre outros: o tópico principal da notícia, entidades, locais e personalidades; ligar grupos de notícias pela medição da distância entre os mesmos e pela utilização de algoritmos de aprendizagem supervisionada. As etapas relativas a este segundo objetivo estão representadas na Figura 2.

Na secção 2 apresentaremos o essencial sobre trabalhos relacionados. Na secção 3 exporemos detalhadamente todos os passos da metodologia aplicada. Seguem-se a apresentação e discussão dos resultados. Por fim são apresentadas as conclusões e o trabalho futuro.

[1]Sequência de carateres

2 Trabalhos Relacionados

2.1 Detetar Notícias Duplicadas

No domínio da imprensa é muito frequente a existência de notícias duplicadas ou quase duplicadas. Isto acontece, porque diferentes fontes noticiosas publicam notícias idênticas para dar a conhecer ao leitor um mesmo acontecimento, um exemplo de notícias quase duplicadas pode ser observado pelos seguintes títulos: *Retomadas buscas pelo avião da Malaysia Airlines* (Renascença, 06/10/2014) e *Recomeçam as buscas pelo avião desaparecido da Malaysia Airlines* (Jornal de Notícias, 06/10/2014); como é possível constatar ambas as notícias querem-nos transmitir a mesma informação, mas por terem sido publicadas por diferentes fontes noticiosas, aparecem escritas de uma forma distinta.

O surgimento de notícias duplicadas, notícias que se referem ao mesmo acontecimento, é um problema muito comum no domínio da imprensa. Notícias duplicadas não adicionam nenhum conhecimento ao leitor e o seu armazenamento tem elevados custos computacionais, como: espaço de armazenamento e diminuição do desempenho do sistema a nível de pesquisas. Devido a estes constrangimentos torna-se necessário a deteção de notícias duplicadas (Kumar e Govindarajulu, 2009). A deteção de notícias duplicadas é o problema de se encontrarem documentos onde o grau de semelhança entre si é superior a um determinado *threshold* e que, convencionamos, indica quando as notícias não introduzem nenhum informação adicional relevante.

São várias as abordagem propostas para a resolução do problema de deteção de notícias duplicadas, entre elas encontram-se: a abordagem baseada no léxico, a abordagem baseada no URL e a abordagem baseada na semântica. A abordagem baseada no léxico não requer nenhum conhecimento linguístico. O objetivo da mesma é perceber a existência de termos em comum entre documentos. A abordagem baseada no URL visa detetar notícias duplicadas pela comparação do endereço URL. Porém esta abordagem continua a não ser suficiente. Isto porque, não existe um padrão estabelecido pelas diversas fontes noticiosas de como criar um URL e, portanto, podendo este conter ou não informação útil. A abordagem semântica é uma abordagem mais completa,

esta inclui a necessidade de pré-processamento implicando: *tokenization*, *stemming* e remoção das *stop-words*. Após o pré-processamento do texto, as notícias são comparadas através de uma função de similaridade. Esta função tem como objetivo medir o grau de semelhança entre pares de notícias. O valor retornado por esta função varia entre [0,1], e é tanto maior quanto maior for a semelhança existente entre as notícias.

No trabalho intitulado *Duplicate Record Detection: A Survey* (Elmagarmid, Ipeirotis e Verykios, 2007), os autores explicam todo o fluxo necessário à deteção de documentos duplicados. Este trabalho refere-se à abordagem semântica. As notícias são inicialmente processadas, seguidamente são determinados os campos a comparar; é, depois, medido o grau de semelhança entre pares de notícias; e por fim, com base no resultado obtido é determinado se os documentos são ou não similares. O autor ilustra quatro métricas de medição de similaridade, são elas: similaridade de *strings* baseada em carateres; similaridade baseada em *tokens*; similaridade fonética e similaridade numérica.

A similaridade baseada em caracteres foi desenvolvida para detetar erros tipográficos, alguns exemplos dessas métricas são: algoritmos de edição de distância (Hamming(He, Petoukhov e Ricci, 2004) e Levenshtein (Levenshtein, 1965)) que visam calcular o número de adições, substituições e remoções necessárias para converter uma *string* numa outra, como por exemplo 'futebol' e 'futbol'; distância Affine Gap (Waterman, Smith e Beyer, 1976) que consiste em abrir ou estender um espaço, para transformar uma *string* noutra, como: 'C Ronaldo' e 'Cristiano Ronaldo'; a métrica de distância Jaro (Bilenko et al., 2003) que mede a semelhança entre duas *strings* tendo em conta o comprimento das mesmas, o número de carateres em comum e o número de transposições necessárias; e a métrica Q-grams (Ullmann, 1977) que consiste na divisão das *strings* iniciais em *substrings* de tamanho q, a medição de similaridade entre documentos consiste na medição de *substrings* em comum entre as duas notícias.

Após o cálculo da similaridade entre pares de notícias, a fim de determinar se duas notícias são ou não similares, são utilizados algoritmos de aprendizagem supervisionada.

Infelizmente, existem poucos estudos desenvolvidos no sentido de verificar a eficiência da utilização de métricas de distância (Elmagarmid, Ipeirotis e Verykios, 2007). Existem, por exemplo, alguns estudos que mencionam a eficiência da métrica de distância Jaro (Bilenko et al., 2003) (Yancey, 2005) na comparação de nomes.

O nosso contributo, na parte da deteção de notícias duplicadas, diz respeito ao estudo da eficiência de alguns algoritmos de edição de distância para textos estruturados de dimensão variável.

2.2 Geração Automática de Histórias

Diversos trabalhos tem sido conduzidos com o objetivo de criarem histórias a partir de vários documentos como: notícias (Shahaf e Guestrin, 2010)(Mei e Zhai, 2005), blogs (Lin et al., 2012)(Qamra, Tseng e Chang, 2006) e resultados de pesquisas (Kumar, Mahadevan e Sivakumar, 2004). Em alguns trabalhos, antes da criação da história noticiosa o leitor tem que indicar o tema de pesquisa (Shahaf e Guestrin, 2010)(Mei e Zhai, 2005)(Lin et al., 2012). Outros trabalhos porém, visam ser mais abrangentes, e determinar dentro do seu conjunto de dados todas as histórias existentes (Allan, Papka e Lavrenko, 1998)(McKeown et al., 2002). A primeira abordagem é utilizada em estudos relacionados com o tema 'Geração da História' sendo que a segunda abordagem é mais popular em estudos de 'Deteção de Tópicos e Rastreamento'. Em relação a estes dois tópicos, é de notar que existem poucos estudos sobre o primeiro, mas, no entanto, o segundo tópico tem vindo a ser extensivamente estudado (Lin e Liang, 2008). Segundo (Allan, Papka e Lavrenko, 1998), o conhecimento inicial dado ao sistema para a criação das histórias pode não ser adequado ao rastreamento das mesmas uma vez que o tema de discussão associado a um evento muda frequentemente.

Outra área que visa organizar a informação é a classificação hierárquica (Sun e Lim, 2001)(Lawrie e Croft, 2000)(Yang et al., 2000)(Li, Zhu e Ogihara, 2007). A estrutura hierarquia impõe uma estrutura no conjunto de dados, porém, nenhum estudo foi realizado de forma a perceber se essa estrutura reflete as relações existentes entre os diversos documentos (Nallapati et al., 2004).

2.2.1 Geração da História

O trabalho intitulado *Connecting the Dots Between News* (Shahaf e Guestrin, 2010) visa encontrar uma história coerente num conjunto de artigos noticiosos a partir de um conhecimento inicial. O método utilizado neste trabalho é aplicável a outros domínios como: *emails*, artigos científicos e inteligência militar. Neste trabalho os autores introduziram a noção de coerência, e *feedback* do utilizador, tendo avaliado a utili-

dade do sistema desenvolvido via *user studies*. A abordagem proposta por estes autores consistiu na medição da ligação entre notícias, tendo em conta: palavras omissas, palavras que estão relacionadas com as palavras do texto embora não apareçam no mesmo, e a importância das palavras. O problema da formação das cadeias de notícias foi solucionado recorrendo a uma abordagem de programação linear.

Outro trabalho desenvolvido com o propósito de gerar uma linha temporal de uma história é o *A Graph Teoretic Approach to Extract Storylines from Serach Results* (Kumar, Mahadevan e Sivakumar, 2004). Neste trabalho os resultados de pesquisa são representados numa estrutura de grafos. Sobre uma estrutura de grafos, onde cada documento tem a si associada informação, e entre si, os documentos tem um peso de ligação, para a elaboração das cadeias, os autores recorrem à utilização de um algoritmo de pesquisa local.

2.2.2 Deteção de Tópicos e Rastreamento

Existem três tarefas associadas a deteção de tópicos e rastreamento, são elas: rastreamento de eventos conhecidos, deteção de eventos desconhecidos, e segmentação das noticias em histórias (Allan et al., 1998). O grande objetivo dos estudos de deteção de tópicos e rastreamento é o de identificar todas e quaisquer notícias relacionadas com um dado evento (Allan et al., 1998).

Para o nosso trabalho em particular, a parte mais interessante deste estudo é a forma de fazer o rastreamento de uma história nas notícias. A abordagem de rastreamento utilizada em 'Online News event detection and tracking' (Allan, Papka e Lavrenko, 1998) começa por reduzir o conteúdo noticioso a um conjunto de entre 10 a 20 *features*. Os autores acreditam que poucas *features* são necessárias para o rastreamento de notícias uma vez que o essencial de uma história tende a ser descrito por um conjunto pequeno de palavras ou frases. Neste trabalho, as cadeias são obtidas pelo cálculo de semelhança entre as *queries* que caracterizam cada notícia.

3 Metodologia

3.1 Similaridade

No trabalho que desenvolvemos, a fim de detetar notícias similares foi utilizada a abordagem semântica. Esta abordagem pode ser descrita em quatro passos distintos: (i) Normalização do conteúdo noticioso;(ii) Determinação dos campos a serem comparados; (iii) Comparação entre pares de notícias; (iv) Decisão sobre a similaridade entre notícias.

Tabela 1: Exemplo do fluxo da normalização

Operação	**Exemplo**
Notícia	Nova Deli, 02 jan (Lusa) - A Índia anunciou que vai permitir a cidadãos estrangeiros investirem no seu mercado de ações.
1- **Pontuação**	Nova Deli 02 jan Lusa A Índia anunciou que vai permitir a cidadãos estrangeiros investirem no seu mercado de ações.
2- **Padrões**	A Índia anunciou que vai permitir a cidadãos estrangeiros investirem no seu mercado de ações.
3- **Stop-words**	Índia anunciou vai permitir cidadãos estrangeiros investirem mercado ações.
4- **Stemm**	Índi anunc va permit cidadã estrangeir invest merc açõ.

3.1.1 Normalização

Este passo tem como objetivo melhorar a qualidade dos dados de entrada e tornar esses dados mais comparáveis e mais usáveis (Elmagarmid, Ipeirotis e Verykios, 2007). A normalização inclui as seguintes ações:

1) Remoção de carateres de pontuação, como: ¡, ¿, '/', ',' , '(', ')', '-';

2) Remoção de padrões obtidos por inspeção manual, que são redundantes e, no âmbito deste trabalho, não adicionam informação ao conteúdo da notícia, como é o case de: *"Lusa - Esta notícia foi escrita nos termos do Acordo Ortográfico"*;

3) Remoção de *stop-words*, através da utilização da lista específica para a língua portuguesa disponibilizada pela *snowball* [2];

4) Redução das palavras à sua raiz através da utilização do *'Porter Stemmer'* para língua portuguesa, disponibilizado pelo *PTStemmer* (Oliveira, 2008).

Na Tabela 1 apresentamos um exemplo de uma notícia e o resultado da aplicação das diversas fases.

[2]https://snowball.tartarus.org

Tabela 2: Exemplos de Urls

1	Al Qaeda revindica atentados em quartel militar do Iêmen http://visao.sapo.pt/al-qaeda-revindica-atentados-em-quartel-militar-do-iemen=f803958
2	Plantel empenhado na vitória em Barcelos http://www.record.xl.pt/Futebol/Nacional/1a_liga/academica/interior.aspx?content_id=919169
3	Cidade chinesa gera energia com queima de notas de banco http://diariodigital.sapo.pt/news.asp?id_news=750321

Figura 3: Campos da notícia a serem comparados

3.1.2 Determinação dos campos a serem comparados

Os artigos noticiosos publicados em formato digital tem normalmente cinco campos associados ao texto da notícia propriamente dita, são eles: título, conteúdo, data de publicação, *tags* e o URL. Antes de se proceder à comparação das notícias é necessário perceber que influência tem cada campo na determinação de notícias similares.

Urls provenientes de diferentes dominios têm uma composição distinta. A Tabela 2 apresenta três pares de títulos com os respetivos Urls. Como é possível observar na Tabela 2 o primeiro Url é composto pelo título da notícia; já o segundo dá-nos a indicação das áreas a que a notícia está associada, não explicitando em concreto o acontecimento presente; já o terceiro exemplo, não nos consegue transmirir nada, uma vez que o Url é formado apenas por um identificador numérico.

Observando notícias referentes a um mesmo evento publicadas por fontes noticiosas diferentes, é possível observar, que um campo isolado, como o título ou conteúdo, não são suficientes para a determinação da similaridade entre notícias. Existe uma vasta gama de variações possíveis. De forma a cobrir a maior gama de variações que notícias duplicadas podem assumir, consideramos para a comparação três campos: o título da notícia, o conteúdo, e ainda um campo obtido pelo processamento da notícia que se refere ao "foco"da mesma (baseado no primeiro parágrafo do conteúdo). Os campos considerados podem ser observados na Figura 3.

É de notar que as notícias correspondem a informação temporal, pelo que o fator tempo, assume neste contexto um importância de extrema relevância. Acreditamos que existirá um intervalo de tempo restrito dentro do qual há uma maior tendência para o aparecimento de notícias duplicadas.

3.1.3 Comparação de Notícias

Diferentes métricas de distância podem ser utilizadas como função de similaridade. Neste estudo iremos considerar as seguintes métricas para cálculo de distâncias: Hamming (He, Petoukhov e Ricci, 2004), Levensthein (Levenshtein, 1965) e Jaro (Bilenko et al., 2003).

De forma a termos resultados equipáraveis é necessário proceder à normalização dos mesmos. De forma a obter um resultado entre [0,1] foi aplicada a fórmula seguinte (Expressão 1) aos resultados retornados pelos métodos de edição de distância.

$$D'_{Alg}(s,t) = 1 - \frac{D_{Alg}(s,t)}{max(|s|,|t|)} \quad (1)$$

Onde:
$D_{Alg}(s,t)$: Distância obtida pela métrica de edição de distância entre a string s e t;
$max(|s|,|t|)$: Comprimento da string de maior dimensão entre s e t;
$D'_{Alg}(s,t)$: Distância normalizada entre s e t.

3.1.4 Decisão da similaridade entre notícias

A deteção de notícias duplicadas é o problema de encontrar documentos onde o seu grau de similaridade é maior ou igual a um determinado *th-*

reshold. A definição dos *thresholds* feita de forma manual por nós definido ou de forma automática. Neste trabalho estudamos o comportamento de diversos algoritmos de aprendizagem supervisionada na determinação de notícias duplicadas. Os algoritmos testados foram: Support Vector Classifier (SVC), SVC Linear, Decision Tree e Random Forest. Estes métodos estão disponíveis no *scikit learn* (Pedregosa et al., 2011).

3.2 Agrupamento de Notícias

Este módulo é responsável pela criação de grupos de notícias duplicadas usando os resultados obtidos do módulo que o precede (deteção de notícias duplicadas).

3.3 Extração de Termos Chave

Sintetizar a informação contida nos grupos de notícias é uma tarefa essencial para a formação de cadeias noticiosas. Vários estudos reduzem o conteúdo noticioso numa frase ou num conjunto de *features* (Allan, Papka e Lavrenko, 1998).

Na nossa abordagem, vamos representar as noícias por um conjunto de termos relevantes. Os termos relevantes podem ser considerados termos que transmitem informação considerada relevante do texto, como: o tópico da notícia, nomes de personalidades, locais e outros. Consideramos quatro tipos de termos chave: (i) palavras e (ii) expressões relevantes, (iii) entidades e (iv) personalidades.

3.3.1 Palavras e Expressões Relevantes

As palavras e expressões relevantes correspondem a termos que aparecem explicitamente no conteúdo noticioso e que de uma forma simplificada podem transmitir informação relevante contida no texto. A abordagem seguida considerou a existência de palavras relevantes, representadas por *uni-grams*, como: *convocatória*, *equipa*, *treinador*; e expressões relevantes, formadas por *n-grams*, como: *Campeonato da Europa*, *fase de qualificação* entre outros.

Para verificar a frequência dos termos na notícia, é necessário numa primeira fase, indicar ao algoritmo quais os termos presentes na notícia. Nesta fase, existe uma diferença em relação às palavras e às expressões relevantes devido à sua tipologia. Em relação às palavras, formadas por *uni-grams*, são considerados como termos todos os *tokens* existentes, como: *derrota*, *ministro* e *estudantes*. Quanto às expressões, formadas por *n-grams* são indicados como termos todas as sequências de palavras que seguem os seguintes padrões:

- Entidade: "[Nome] [Nome] *"(exemplos: Setembro; Michele Bachmann; Domingos Paciência)
- Entidade e sua caracterização: "[Nome] [Nome]* [Adjetivo]"(exemplos: turista israelita; polícias municipais, homens encapuzados)
- Entidades Compostas: "[Nome] [Nome]* [Preposição+Determinante] [Nome]*"(exemplo: Presidente da República)

São necessários quatro passos para a extração das palavras e expressões das notícias. O *POS Tagger* (i) que visa identificar as categorias gramaticais de todas as palavras que compõem o corpo da notícia. Para esta tarefa é utilizado o *TreeTagger* (TreeTagger, 1996) adaptado para a língua portuguesa, disponibilizado pelo *Pablo Gamallo* (Marcos Garcia, 2013). A normalização (ii) que corresponde à remoção de padrões linguísticos e frases recorrentes do corpo da notícia obtidos por inspeção manual, como: expressões de datas (Porto, 12 Agosto 2014), resultados de futebol (2-1) e padrões jornalísticos (Porto, 12 Agosto 2014 (Lusa)). Análise da frequência da palavra (iii) pela utilização da métrica estatística *Term Frequency-Inverse Document Frequency* (TF-IDF), representada pela Expressão 2. No seu cálculo, esta métrica relaciona o aparecimento de um termo na notícia com o aparecimento do mesmo na coleção permitindo assim detetar a existência de termos relevantes. A atribuição das palavras e expressões às notícias (iv) consiste na associação às notícias de um conjunto de palavras e expressões consideradas como relevantes pela etapa anterior.

3.3.2 Reconhecimento de entidades mencionadas

Para a extração de entidades mencionadas no texto foi utilizado um algoritmo com o objetivo de verificar, numa primeira fase, quais as palavras no texto que se iniciam com um caráter capitalizado. Das palavras encontradas, se a palavra capitalizada estiver posicionada no inicio da frase é verificado se a palavra é ou não uma *stop-word*, e caso seja, então não é considerada. Para as palavras que passarem a fase anterior é verificado se são precedidas de outras palavras capitalizadas, sendo permitido uma palavra de ligação entre termos capitalizados inicializada a minúscula. Um exemplo de entidades extraídas pelo algoritmo é

$$TF - IDF = \frac{o(W, DOC)}{npalavras(DOC) * log(\frac{docs(ALL))}{1+docs(W,ALL))})} \quad (2)$$

Onde:

$o(W, DOC)$: número de ocorrências da palavra W no documento DOC
$npalavras(DOC)$: número de palavras no documento DOC
$docs(ALL)$: número de documentos na coleção
$docs(WORD, ALL)$: número de documentos na coleção que contém a palavra W

dado pelos seguintes termos: "Passos", "Paulo Portas".

De forma a enriquecer a estrutura foi adicionado a cada notícia a lista de personalidades nela contidas. As personalidades foram obtidas através das expressões e entidades extraídas do conteúdo noticioso pela utilização de uma fonte de conhecimento externo, o Verbetes [3].

3.4 Atribuição de termos relevantes aos agrupamentos

Depois da junção de notícias similares em agrupamentos e realizada a extração de termos relevantes de cada notícia, é possível fazer a atribuição dos termos chave aos agrupamentos de notícias.

Os termos chave associados a cada agrupamento correspondem aos termos relevantes que estão associados a todas as notícias do agrupamento. É de referir que cada termo chave tem um peso, que está relacionado com a sua importância no agrupamento. A importância de um termo é dado pela relação entre o número de notícias em que o termo aparece e número total de notícias que compõe o agrupamento. Um exemplo de palavras relevantes associadas a um agrupamento e respetiva importância é dado por:

reclusos[9];presos[9];cárcere[7];
sudoeste[7];representantes[6];
violação[6];cadeia[5];
quilómetros[4];irmãos[4];

Neste agrupamento, o termo *reclusos* é mais representativo do conjunto do que o termo *irmãos*. Isto porque, considerando que o agrupamento em questão tem nove notícias, o primeiro termo aparece associado a todas as notícias do agrupamento, tendo um peso de $\frac{9}{9}$, ou seja 1; enquanto que o segundo termo só se encontra associado a 4 notícias do conjunto, tendo um peso de $\frac{4}{9}$.

[3] https://store.services.sapo.pt/pt/Catalog/other/free-api-information-retrieval-verbetes

3.5 Ligações entre Agrupamentos

Este módulo visa encontrar ligações entre os agrupamentos existentes recebendo para esta tarefa um conjunto de agrupamentos com termos relevantes associados. O objetivo deste módulo é a criação de ligações entre agrupamentos, que corresponde aos arcos existentes na Figura 2.

É de notar que partimos do pressuposto que as cadeias noticiosas só poderiam ser obtidas a partir da mesma categoria. Para isso, fizemos a atribuição das categorias aos grupos de notícias, através de uma fonte de conhecimento externo que mapeia as *tags* atribuídas pelos jornalistas com a categoria a que a notícia fica associada. As categorias indicam de uma forma geral a área a que a notícia pertence como: Desporto, Sociedade, Política, Economia, entre outros.

A abordagem utilizada para o processo de ligação de pontos entre os agrupamentos foi realizado em duas etapas:

1. Cálculo da distância entre termos relevantes;
2. Determinação das ligações entre agrupamentos.

3.5.1 Similaridade de termos relevantes

Começamos por fazer a normalização dos termos relevantes. Para todos os casos, palavras, expressões, entidades e personalidades, o texto é convertido para letra minúscula. Para as palavras relevantes que são constituídas apenas por *uni-grams* também se efetua a redução ao seu radical. Após a normalização do texto, é efetuado o cálculo da similaridade entre agrupamentos. Para esta tarefa é considerado o peso dos termos relevantes pois acreditamos que eles representam bem a informação do agrupamento.

Para a determinação das ligações entre agrupamentos de notícias, é realizado o cálculo da distância entre os seguintes elementos: palavras e expressões relevantes; entidades e personalidades.

A abordagem utilizada para o cálculo da similaridade entre: palavras relevantes, entidades e personalidades, considera o peso de cada palavra individual no agrupamento e é dada pelas Expressões 3 e 4. As distâncias $D_1(a, b)$ e $D_2(a, b)$ têm em conta a percentagem de termos em comum entre os dois agrupamentos e a relação dos pesos que os termos em comum têm nos seus agrupamentos. A diferença entre $D_1(a, b)$ e $D_2(a, b)$ é que a primeira estabelece um peso entre as duas parcelas, dando um maior relevo à parcela que mede o relacionamento dos pesos das palavras em comum; enquanto na segunda não existem pesos associados às parcelas, mas sim, uma relação entre elas.

Para o cálculo da similaridade entre as expressões relevantes a abordagem utilizada foi distinta. Para este tipo de termo, a normalização inclui um passo adicional que consistiu na remoção das *stop-words*. Após esta tarefa foi construída uma *string* com todas as expressões pertencentes a cada agrupamento, não considerando para este tipo de termo relevante o seu peso. Para a realização do cálculo da similaridade entre as expressões foi utilizado um algoritmo de edição de distância o *qgrams* (Ullmann, 1977), com o fator $q = 3$.

3.5.2 Determinação das ligações entre agrupamentos

Esta etapa tem como objetivo determinar a partir dos valores recebidos da comparação entre os diferentes tipos de termos chave, se existe ou não ligação entre os agrupamentos. É a partir das ligações que se formam as cadeias noticiosas.

Para a ligação de agrupamentos, utilizamos algoritmos de aprendizagem supervisionada. Estes algoritmos como referido na *Subsecção 3.2.3* recebem um conjunto de treino sobre o qual vão inferir regras para determinar, neste caso, se existe ou não ligação entre os agrupamentos. Utilizamos como características (*features*) a distância entre as palavras-chave simples, compostas, entidades e personalidades. Os algoritmos utilizados foram: Support Vector Classifier (SVC), SVC Linear, Decision Tree e o Random Forest.

4 Experimentação

Nesta secção é caracterizado o conjunto de dados utilizados neste trabalho, referidas as diferentes métricas de avaliação utilizadas e descrito o conjunto de experiências realizadas.

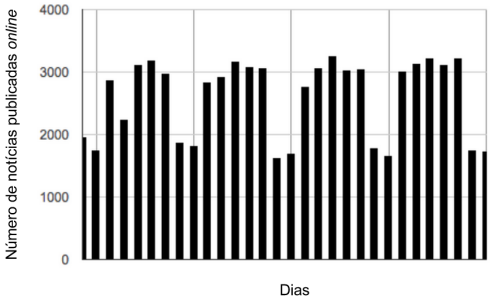

Figura 4: Número de notícias publicadas por dia no mês de Março de 2014

4.1 Caracterização do conjunto de dados

Para a realização deste trabalho foram utilizadas notícias publicadas *online*, escritas na língua portuguesa e provenientes de diversas fontes noticiosas da imprensa portuguesas. O conjunto de dados compreende mais de 4 milhões de notícias publicadas entre 2008 e 2014.

As notícias são provenientes de 73 [4] fontes notíciosas distintas e compostas em média [5] por: 9 palavras no título; 204 palavras no conteúdo; 10 frases no conteúdo.

Na imprensa portuguesa são publicadas diariamente, em formato digital, aproximadamente 2.500 notícias [6]. A Figura 4 representa a distribuição de notícias durante mês de Março de 2014. Através da observação da mesma é possível constatar que tendencialmente são publicadas menos notícias durante o fim-de-semana.

Estima-se que aproximadamente 45% [7] das notícias publicadas diariamente sejam duplicadas ou quase duplicadas. A relação entre o número de notícias publicadas mensalmente com o número de notícias utilizadas para a criação dos agrupamentos pode ser visualizada na Figura 5. Para os primeiros oito meses de 2014 o número médio de notícias por grupo é de 3.8, os dados referentes ao número médio de notícias por grupo relativo a cada mês pode ser observado na Figura 6.

Na Figura 7 podemos constatar que tendencialmente os grupos são constituidos por 2 notícias. É possível observar que o número de grupos existentes é inversamente proporcional ao número de notícias que o compõe.

[4] Número de fontes com mais de 100 notícias publicadas.

[5] Análise de aproximadamente 74000 notícias selecionadas de um mês aleatório de 2014.

[6] Dados relativos às notícias publicadas na imprensa portuguesa, no formato digital, no mês de Março de 2014

[7] Número médio de notícias diárias duplicadas, publicadas na imprensa portuguesa, no formato digital, de 10 a 15 de Março de 2014

$$D_1(a,b) = 0.3 * \frac{|ka| \wedge |kb|}{max(|ka|,|kb|)} + 0.7 * \frac{\sum_{i=1}^{|ka|}(\sum_{j=1 \wedge a_i = b_j}^{|kb|} Wkja * Wkib)}{|ka| \wedge |kb|} \quad (3)$$

$$D_2(a,b) = \frac{|ka| \wedge |kb|}{max(|ka|,|kb|)} * \frac{\sum_{i=1}^{|ka|}(\sum_{j=1 \wedge a_i = b_j}^{|kb|} Wkja * Wkib)}{|ka| \wedge |kb|} \quad (4)$$

Onde:
$Wkja$: Peso da palavra-chave j no agrupamento a.
$Wkjb$: Peso da palavra-chave i no agrupamento b.
$|ka|e|kb|$: número de palavras-chave iguais entre o agrupamento a e b.
$máx.|ka||kb|$: número máximo de palavras-chave distintas.

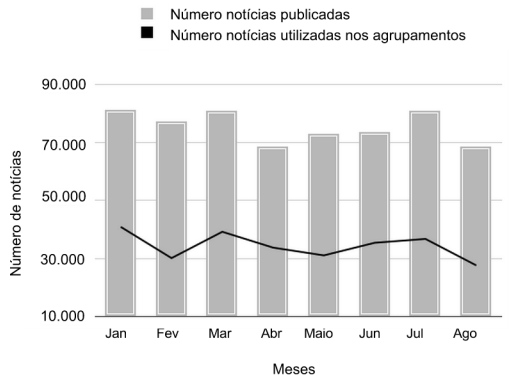

Figura 5: Relação entre o número de notícias publicadas por mês com o número de notícias utilizadas na criação dos agrupamentos (Janeiro a Agosto de 2014)

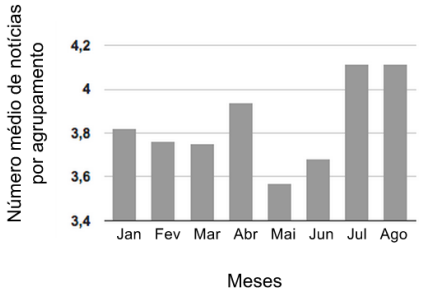

Figura 6: Número médio de notícias por agrupamento (Janeiro a Agosto de 2014)

Definimos oito categorias associadas aos agrupamentos: Política, Economia, Desporto, Saúde, Ciências e Tecnologias, Sociedade, Cultura, Local e Educação. Relativamente aos agrupamentos obtidos, aproximadamente 50% não tem categoria associada ou estão associados a mais do que uma categoria. Dos agrupamentos com apenas uma categoria associada a distribuição dos mesmos por àreas pode ser observado na Figura 8. É possível observar que a categoria com maior expressão é a categoria Desporto (54.4%) e assim

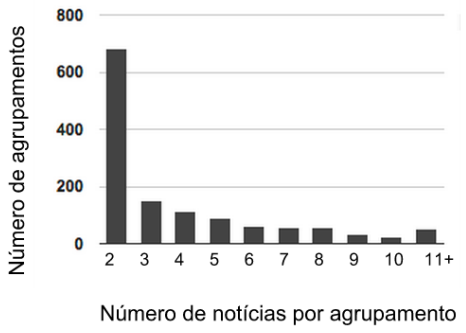

Figura 7: Constituição dos agrupamentos (seleção aleatória de 5 dias de 2014)

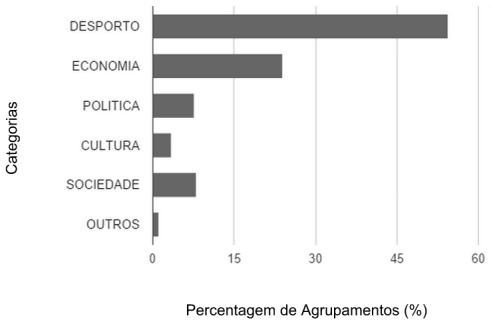

Figura 8: Distribuição dos agrupamentos por categoria

sucessivamente.

4.2 Métricas de Avaliação

Para avaliar o módulo de similaridade e ligações entre agrupamentos, foram utilizadas quatro métricas de avaliação: a precisão (*precision*), a abrangência (*recall*), a *accuracy* e a *F-measure* (F1). No contexto deste trabalho, a precisão indica a taxa de notícias consideradas similares que realmente o são e a taxa de ligações efetuadas entre agrupamentos que realmente existem. A abrangência (*recall*) indica-nos, neste contexto, taxa de notícias duplicadas encontradas. A me-

dida F1 estabelece uma relação entre a precisão e a abrangência. A *accuracy* indica-nos a avaliação geral do sistema.

A avaliação aos termos relevantes consistiu em perceber, dos termos extraídos, quais são de facto realmente representativos da notícia. A avaliação foi realizada usando a Expressão 5. A avaliação geral do sistema, é dada pelo somatório percentagem de termos representativos das notícias analisadas, Expressão 6.

$$E(ni) = \frac{TermosRepresentativos}{TermosAtribuídos}(5)$$

$$Avaliação = \frac{\sum_{i=1}^{||N||}(E(n_i))}{||N||} \quad (6)$$

Onde:

Termos Representativos: corresponde ao número de termos relevantes ou entidades atribuídos pelo método, que realmente representam o conteúdo noticioso;

Termos Atribuídos: corresponde ao número total de termos relevantes ou entidades atribuídas ao documento;

$||N||$: número de notícias da coleção N;

n_i: corresponde à notícia de índice i do conjunto de notícias N.

4.3 Enunciação e definição das experiências

Neste capítulo são apresentadas as diferentes experiências realizadas. A Exp_{ij} representa a j-ésima configuração de parâmetros para a experiência i.

4.3.1 Similaridade - Algoritmos de Edição de Distância

A similaridade entre notícias é obtida através do cálculo da:

Similaridade do título (ST) que corresponde à percentagem de semelhança entre os títulos;

Similaridade do 1º parágrafo (SB) que corresponde ao resultado de comparação entre a parte das notícias que foca o evento em si;

Similaridade de conteúdo noticioso (SC) que corresponde ao resultado da comparação do corpo das respetivas notícias.

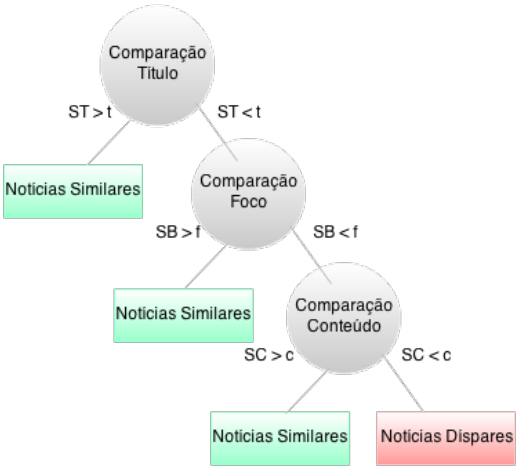

Figura 9: Árvore de decisão elaborada para verificar se um par de notícias é ou não similar

A experiência Exp_1 visou perceber qual o algoritmo com o melhor desempenho para o cálculo da similaridade entre pares de notícias. Esta experiência foi efetuada sobre uma estrutura em forma de árvore de decisão, representada na Figura 9. Esta foi criada manualmente, onde os valores *t, f, c,* correspondem aos valores de decisão para o título, foco e conteúdo da notícia. L, H, J correspondem respetivamente aos algoritmos Leventshein, Hamming e Jaro. A parametrização usada nesta experiência encontra-se enunciada na Tabela 3.

Tabela 3: Parametrização para a experiência do cálculo da similaridade

Exp	Algoritmos	t	f	c
1.1	L H J	0.60	0.60	0.60
1.2	L H J	0,70	0,60	0,60
1.3	L H J	0,70	0,70	0,60
1.4	L H J	0,70	0,70	0,70
1.5	L H J	0,80	0,70	0,70
1.6	L H J	0,80	0,80	0,70
1.7	L H J	0,80	0,80	0,80

Para a realização desta experiência foram comparadas aleatoriamente 124750 notícias, para um dia aleatório de 2014.

4.3.2 Similaridade - Fator Tempo

A experiência sobre o fator tempo tem como objetivo verificar a influência do intervalo temporal no que diz respeito à comparação de notícias. Para tal, foram considerados cinco intervalos de tempo distintos para o cálculo da similaridade entre notícias: 3, 6, 12, 24, 48 horas; e utilizados quatro métodos de classificação para a determinação da similaridade: SVC, SVC Linear, *Decision Tree* e o *Random Forest*. Esta

experiência foi elaborada utilizando uma técnica de avaliação cruzada, o *k-fold cross validation*. Esta técnica pretende avaliar qual a capacidade de generalização de um modelo, para tal, faz a partição do conjunto de dados em conjuntos mutuamente exclusivos utilizando um subconjunto para a criação do modelo e os outros subconjuntos para a validação do mesmo. Esta técnica de avaliação foi utilizada para um $k = 5$, o que significa que se efetuou uma partição do conjunto de dados em 5 subconjuntos distintos. O conjunto de dados utilizado resulta da seleção aleatória de 500 notícias de dois dias distintos e consecutivos. Foram anotadas manualmente a similaridade entre todos os pares de notícias existentes.

4.3.3 Similaridade - Determinação da Semelhança

Foi efetuada uma experiência com o objetivo de perceber qual o algoritmo de aprendizagem supervisionada com o melhor desempenho na determinação da similaridade entre pares de notícias. A experiência foi efetuada em 500 notícias selecionadas de forma aleatória de um dia aleatório de 2014.

4.3.4 Extração de Termos relevantes

Esta experiência tem como objetivo testar a abordagem utilizada para a extração de termos chave (palavras-chave simples, compostas e entidades). Para a realização desta experiência foi selecionado aleatoriamente um dia de cada mês do ano 2012, de cada dia foi selecionado um intervalo de três horas, dessas três horas foram selecionadas aleatoriamente dez notícias sobre as quais se efetuou a inspeção manual das palavras-chave atribuídas.

4.3.5 Ligações entre agrupamentos

Para a determinação das ligações entre agrupamentos de notícias, é realizado o cálculo da distância entre os seguintes elementos: palavras relevantes; expressões relevantes; entidades; personalidades.

A experiência Exp_2 tem como objetivo perceber qual a fórmula mais adequada para o cálculo da similaridade e qual o algoritmo de aprendizagem supervisionada mais eficiente para a determinação das ligações. Todas as experiências consideraram o cálculo distância pelo algoritmo Q-grams, para as expressões. A avaliação resultante das diferentes experiências realizadas entre grupos de notícias ao longo do tempo, para a formação de ligações entre agrupamentos de notícias, encontra-se na Tabela 4. O conjunto de dados é composto por agrupamentos pertencentes aos meses de março e abril de 2014. Desses agrupamentos, foram selecionados aleatoriamente 10 cadeias de notícias com tamanho variável para cada uma das seguintes categorias: Desporto, Economia, Política, Cultura e Sociedade. O conjunto de dados compreende, em média, 317 comparações por categoria.

Tabela 4: Descrição das experiências para o cálculo das ligações

Exp	Palavras	Entidades	Personalidades
2.1	D1	D2	D1
2.2	D2	D2	D1
2.3	D1	D1	D1
2.4	D1	D2	D2

5 Resultados e Análise

5.1 Experiências

5.1.1 Similaridade - Algoritmos de Edição de Distância

Os resultados obtidos nesta experiência aos algoritmos de edição de distância podem ser observados na Tabela 5. Desta tabela foi excluído o resultado obtido pelo algoritmo Jaro. Isto aconteceu devido ao fraco desempenho obtido em todas as experiências.

Ao efetuar uma comparação entre o algoritmo *Levensthein* e o *Hamming*, recorrendo à comparação entre o caso $Exp_{1.1}$ podemos verificar que os valores da precisão são semelhantes, o que indica que a percentagem de notícias consideradas similares que realmente o são é igual. Para o mesmo caso podemos verificar uma melhoria do algoritmo *Levensthein* para a métrica *recall*, o que indica que este algoritmo consegue ter uma maior abrangência.

5.1.2 Similaridade - Fator Tempo

De forma a testar a influência do fator tempo na comparação de notícias, foi testado o comportamento dos diferentes algoritmos considerando diferentes intervalos. O resultado obtido desta análise pode ser observado no gráfico apresentado na Figura 10. Como podemos constatar pela análise do gráfico, o aumento do intervalo de tempo faz com que os valores se tornem constantes. Ao alargar o intervalo de tempo de 24

Tabela 5: Resultados dos testes aos algoritmos de edição de distância

	Levensthein			Hamming		
Exp	P	R	F	P	R	F
1.1	0,941	0,761	0,841	0,941	0,289	0,442
1.2	0,950	0,655	0,775	0,940	0,284	0,436
1.3	0,951	0,645	0,769	0,940	0,284	0,436
1.4	**0,972**	**0,637**	**0,770**	0,940	0,284	0,436
1.5	0,965	0,507	0,665	0,939	0,279	0,430
1.6	0,964	0,483	0,643	0,939	0,279	0,430
1.7	0,962	0,463	0,625	0,938	0,279	0,430

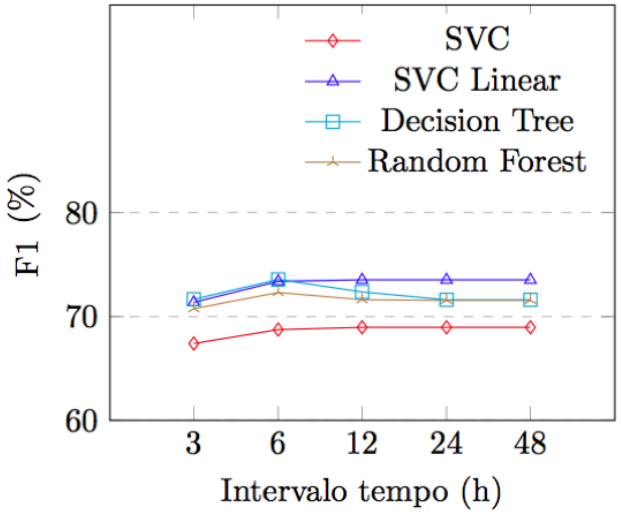

Figura 10: Valor da métrica F1 obtido pelos diferentes algoritmos nos diferentes intervalos de tempo

para 48 horas não há variação nos valores de *precision*, *recall* e da métrica *F1*.

5.1.3 Similaridade - Determinação Semelhança

Os resultados dos algoritmos de aprendizagem supervisionados na determinação da similaridade podem ser observados na Tabela 6. Pela visualização da tabela é possível constatar que apesar dos valor do *recall* ser baixo, o valor obtido pela *precision* é alto, o que garante a qualidade da informação recolhida. O algoritmo que apresenta um melhor desempenho é o SVC Linear.

5.1.4 Extração de Termos Relevantes

Os resultados da extração de termos relevantes pode ser observado na Tabela 7. 73,2% das palavras extraídas, 76,2% das expressões e 80.4% das entidades são representativas da notícia.

Tabela 6: Resultado médio das métricas de avaliação obtidas pelo *k fold cross validation*

	P	R	F1	A
Decision Tree	0,863	0,679	0,760	0,998
SVC	0,931	0,508	0,657	0,997
SVC Linear	**0,938**	**0,561**	**0,702**	**0,998**
Random Forest	0,803	0,542	0,647	0,998

Tabela 7: Avaliação dos termos chave

	Avaliação
Palavras	0,732
Expressões	0,762
Entidades	0,804

5.1.5 Ligações entre agrupamentos

Na Tabela 8 são apresentados os resultados da precisão para as ligações entre agrupamentos. Na mesma tabela é possível observar-se o desempenho dos algoritmos nas principais categorias (D - Desporto; E - Economia; P - Política; C - Cultura; S - Sociedade) bem como o desempenho a nível global. A partir da análise dos resultados podemos verificar que o método com um melhor desempenho é o *SVC Linear*.

5.2 Análise dos resultados obtidos

5.2.1 Similaridade - Algoritmos de Edição de Distância

Para calcular a similaridade entre pares de notícias, recorreu-se à utilização dos seguintes algoritmos de edição de distância: *Hamming*, *Levenshtein* e *Jaro*. Para estes três algoritmos realizaram-se as experiências descritas na Tabela 3, cujos parâmetros de decisão t, f , c indicam a similaridade entre pares de notícias seguindo os testes sugeridos pela estrutura em árvore apresentada na Figura 9. Os resultados obtidos dessas experiências podem ser observados na Tabela 5. O algoritmo *Jaro* é o que apresenta a nível global um pior desempenho. No entanto, segundo estudos realizados, este algoritmo tem um melhor desempenho aquando da comparação de pequenas *strings* (Bilenko et al., 2003), o que não acontece no domínio das notícias. Os valores da precisão entre a utilização do algoritmo *Levensthein* e o *Hamming* são muito próximos, obtendo o algoritmo *Levensthein* ao longo das diferentes experiências um melhor desempenho nesta métrica. Comparando as restantes métricas de avaliação, para estes dois algoritmos, é possível observar que

Tabela 8: Valor da precisão na determinação de ligações entre agrupamentos de notícias.

Exp	Cat	SVC	SVC Linear	Decision Tree	Random Forest
2.1_D	D	-	0.947	0.779	0.858
2.1_E	E	-	1.000	0.952	0.886
2.1_P	P	-	0.947	0.779	0.858
2.1_C	C	1.000	1.000	0.909	0.911
2.1_S	S	-	0.838	0.808	0.855
2.1		-	**0.931**	0.849	0.859
2.2_D	D	-	0.936	0.703	0.789
2.2_E	E	1.000	1.000	0.952	0.917
2.2_P	P	-	0.936	0.703	0.789
2.2_C	C	1.000	1.000	0.909	0.962
2.2_S	S	-	0.802	0.729	0.753
2.2		-	0.921	0.821	0.852
2.3_D	D	-	0.852	0.656	0.772
2.3_E	E	-	0.970	0.799	0.861
2.3_P	P	-	0.852	0.656	0.772
2.3_C	C	-	1.000	0.881	0.901
2.3_S	S	-	0.915	0.776	0.766
2.3		-	0.906	0.764	0.824
2.4_D	D	-	0.931	0.834	0.858
2.4_E	E	-	1.000	0.952	0.914
2.4_P	P	-	0.947	0.708	0.772
2.4_C	C	1.000	1.000	0.909	0.932
2.4_S	S	-	0.838	0.816	0.853
2.4		-	0.931	0.834	0.858

o *Levensthein* obtém uma melhor performance a nível da métrica *recall*, o que significa que consegue detetar mais casos do que o *Hamming*. Uma razão para que isto suceda está relacionado com uma particularidade deste último algoritmo que é a comparação de *strings* do mesmo comprimento; a nível da métrica F1, também o *Levensthein* obtém um melhor resultado. Através da análise efetuada a estes três algoritmos é possível concluir que o *Levensthein* é o algoritmo mais indicado para o cálculo da similaridade entre pares de notícias.

5.2.2 Similaridade - Fator Tempo

Um fator importante para a comparação das notícias é a sua data de ocorrência. A Figura 10 representa o valor da métrica F1 obtida pelos diferentes algoritmos em diferentes intervalos de tempo. A nível da precisão, os algoritmos que apresentam um melhor desempenho são o SVC e o SVC Linear. Sendo que destes dois, o SVC Linear tem um desempenho superior a nível do *recall* e da métrica *F1*. Relativamente à questão temporal, podemos perceber, através da visualização dos gráficos, que todos os algoritmos seguem a mesma tendência a nível da forma do gráfico. Pela análise do grafico é possível verificar que não existem variações dos resultados quando o intervalo de tempo é alargado de 24 para 48 horas. Isto pode indicar que os casos de notícias duplicadas ou quase duplicadas surjam num intervalo inferior ou igual a 24 horas. Com base nos resultados obtidos constatou-se que um intervalo de tempo de 24 horas era o mais adequado para a comparação de notícias.

5.2.3 Similaridade - Determinação Semelhança

Para a determinação da similaridade, os algoritmos que apresentam um melhor desempenho, considerando o $\Delta T = 1$ dia, são: a nível da precisão o SVC Linear (93.8%) e SVC (93.1%) ; em relação à métrica recall e a métrica F1 o *Decision Tree* (67.9% e 76.0%) e SVC Linear (56.1% e 70.2%). Comprando o desempenho dos diferentes algoritmos para as diferentes fases de processamento e tendo em conta as opções escolhidas a nível de algoritmo de cálculo da similaridade e intervalo de tempo considerado, podemos constatar que o algoritmo que apresenta um melhor desempenho a nível global é o SVC Linear.

5.2.4 Extração de Termos Relevantes

Foi efetuada uma avaliação manual à relevância das palavras-chave extraídas. A avaliação consistiu em analisar a representatividade dos termos extraídos do texto em relação ao conteúdo da notícia. O resultado da avaliação efetuada a estes elementos pode ser observada na Tabela 7. Os resultados indicam que 73,2% das palavras, 76,2% das expressões e 80,5% das entidades são representativas do conjunto. Através da análise ao teor dos termos extraídos foi possível observar que as palavras relevantes dizem respeito a palavras que descrevem de uma forma muito genérica o conteúdo da notícia; e, por sua vez, as expressões relevantes já transmitem com mais precisão o assunto da notícia. Consideremos de novo o exemplo das notícias sobre o desaparecimento do Avião da Malaysia airlines, temos como palavra relevante *avião* e como expressão *avião Malaysia Airlines*

5.2.5 Ligações entre agrupamentos

Para a formação de ligações entre agrupamentos de notícias considerou-se que só devemos retêr as ligações entre agrupamentos da mesma categoria. Para testar as ligações entre agrupamentos foram efetuadas quatro experiências (ver Tabela 4). Os resultados obtidos, para a métrica precisão, para

as experiências referidas estão apresentados na Tabela 8.

Da análise aos resultados obtidos pela comparação da $Exp_{2.1}$ com a $Exp_{2.2}$, em que o que foi modificada a fórmula de cálculo da distância entre as palavras relevantes, é possível observar que todos os algoritmos conseguem um melhor desempenho considerando a fórmula de cálculo *D1*; a diferença da precisão entre os algoritmos é de: 0.010 no SVC Linear; 0.028 no *Decision Tree* e 0.007 no *Random Forest* . Estabelecendo uma comparação entre as experiências Exp_{v1} e a $Exp_{2.3}$, que divergem apenas na fórmula de cálculo da distância entre as entidades, temos que: a utilização da fórmula *D2* no cálculo da proximidade de entidades entre dois conjuntos reflete um aumento de desempenho. Confrontando os valores obtidos para a experiência $Exp_{2.1}$ em relação à experiência $Exp_{2.3}$ é possível constatar que independentemente do algoritmo de aprendizagem supervisionada os resultados da $Exp_{2.1}$ são os que apresentam um melhor desempenho. Os valores da precisão obtidos para a experiência $Exp_{2.1}$ e a $Exp_{2.4}$ são bastante próximos. Esta experiência difere da primeira na fórmula de cálculo da distância entre personalidades. A partir dos resultados obtidos conclui-se que as personalidades não têm tanto impacto na formação das ligações como as palavras-chave simples e entidades, uma vez que a mudança de cálculo para este elemento não reflete uma variação considerável no resultado. Podemos ainda observar que o melhor desempenho continua a ser o resultante da experiência $Exp_{2.1}$. Após o estudo dos resultados obtidos, podemos concluir que a fórmula mais apta para cada tipo de palavra-chave é a seguinte: *D1*- personalidades e palavras-chave simples; *D2* - entidades; sendo que esta combinação se refere à experiência $Exp_{2.1}$. Comparando os resultados obtidos pelos diferentes métodos de aprendizagem supervisionada para $Exp_{2.1}$ podemos observar que o método com um melhor desempenho é o SVC Linear (93.1%).

6 Conlcusões e Trabalho Futuro

Este artigo apresenta o estudo experimental realizado para a construção automática de cadeias temporais de notícias relacionadas. A abordagem utilizada para a criação das cadeias baseia-se na elaboração de dois passos chave. São eles: (i) deteção de notícias duplicadas e (ii) a criação de ligações entre notícias relacionadas.

A nossa abordagem para o primeiro passo consistiu na utilização de uma abordagem dita baseada na semântica para o cálculo da similaridade entre notícias. Foi também utilizado um algoritmo de aprendizagem supervisionado na determinação da semelhança entre as mesmas. As notícias incluem informação temporal e, tal com acreditávamos, existe um intervalo onde há uma maior tendência no aparecimento de notícias duplicadas. O nosso estudo indicou que tendencialmente as notícias consideradas duplicadas aparecem num intervalo inferior a 24 horas.

A nossa abordagem, na determinação de notícias duplicadas, num intervalo de tempo de 24 horas, obteve uma precisão de 93.8% para o par Levenshtein, SVC Linear.

Para a criação de ligações entre grupos de notícias similares, a nossa abordagem consistiu na medição do grau de semelhança entre os diferentes grupos. Para esta etapa, sugerimos uma nova forma de medição de distância que tem em conta os termos em comum e a expressão de cada termo nos agrupamentos de notícias similares. Para a determinação das ligações, foram também utilizados algoritmos de aprendizagem supervisionada. A abordagem proposta para a realização desta segunda tarefa apresenta uma precisão de 93.1%.

A nível de trabalho futuro será necessário criar testes mais exaustivos e objetivos para as cadeias de noticias. Tais testes consistirão entre outros melhoramentos, na medição da familiaridade do leitor com um tema em específico antes e depois da utilização da plataforma.

Também como trabalho futuro pretendemos melhorar e incrementar o sistema. A melhoria estará relacionada com a componente de visualização, ou seja, pretendemos criar uma interface mais intuitiva para que o utilizador de uma forma simples aceda ao conteúdo pretendido. Quanto ao incremento ao sistema, poderá consitir:(i) na facilidade de sumarização, cujo objetivo é o de resumir as notícias, para que o leitor perceba de uma forma sucinta o teor das mesmas; (ii) na deteção de novos factos incluidos na história, que consiste em analisar cada novo evento e perceber o que acontece de novo; (iii) e na hierarquização das notícias, que visa organizar hierarquicamente por tópicos e sub-tópicos as notícias (p. ex. Desporto; Futebol; 1ª Liga).

Agradecimentos

Agradecemos a colaboração do UP/SAPO Labs pela disponibilização dos dados utilizados neste trabalho.

Referências

Allan, James, Jaime G Carbonell, George Doddington, Jonathan Yamron, e Yiming Yang. 1998. Topic detection and tracking pilot study final report.

Allan, James, Ron Papka, e Victor Lavrenko. 1998. On-line new event detection and tracking. pp. 37–45.

Bilenko, Mikhail, Raymond Mooney, William Cohen, Pradeep Ravikumar, e Stephen Fienberg. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, September, 2003.

Elmagarmid, Ahmed K, Panagiotis G Ipeirotis, e Vassilios S Verykios. 2007. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16.

He, Matthew X, Sergei V Petoukhov, e Paolo E Ricci. 2004. Genetic code, hamming distance and stochastic matrices. *Bulletin of mathematical biology*, 66(5):1405–1421.

Kumar, J Prasanna e P Govindarajulu. 2009. Duplicate and near duplicate documents detection: A review. *European Journal of Scientific Research*, 32:514–527.

Kumar, Ravi, Uma Mahadevan, e D. Sivakumar. 2004. A graph-theoretic approach to extract storylines from search results. pp. 216–225.

Lawrie, Dawn e W Bruce Croft. 2000. Discovering and comparing topic hierarchies. Em *RIAO*, pp. 314–330.

Levenshtein, Vladimir. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163:845–848. original in Russian—translation in Soviet Physics Doklady, vol. 10, no. 8, pp. 707-710, 1966.

Li, Tao, Shenghuo Zhu, e Mitsunori Ogihara. 2007. Hierarchical document classification using automatically generated hierarchy. *Journal of Intelligent Information Systems*, 29(2):211–230.

Lin, Chen, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, e Tao Li. 2012. Generating event storylines from microblogs. pp. 175–184.

Lin, Fu-ren e Chia-Hao Liang. 2008. Storyline-based summarization for news topic retrospection. *Decision Support Systems*, 45(3):473–490.

Marcos Garcia, Pablo Gamallo. 2013. Freeling e treetagger: um estudo comparativo no âmbito do português.

McKeown, Kathleen R, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, e Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. Em *Proceedings of the second international conference on Human Language Technology Research*, pp. 280–285. Morgan Kaufmann Publishers Inc.

Mei, Qiaozhu e ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. pp. 198–207.

Nallapati, Ramesh, Ao Feng, Fuchun Peng, e James Allan. 2004. Event threading within news topics. Em *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp. 446–453. ACM.

Oliveira, Pedro. 2008. Ptstemmer - a stemming toolkit for the portuguese language. disponível em http://code.google.com/p/ptstemmer, em Maio 2014.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, e E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qamra, Arun, Belle Tseng, e Edward Y Chang. 2006. Mining blog stories using community-based and temporal clustering. Em *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 58–67. ACM.

Shahaf, Dafna e Carlos Guestrin. 2010. Connecting the dots between news articles. pp. 623–632.

Sun, Aixin e Ee-Peng Lim. 2001. Hierarchical text classification and evaluation. Em *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 521–528. IEEE.

TreeTagger. 1996. Treetagger - a language independent part-of-speech tagger. disponível

em http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/, em Dezembro 2013.

Ullmann, Julian R. 1977. A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. *The Computer Journal*, 20(2):141–147.

Waterman, Michael S, Temple F Smith, e William A Beyer. 1976. Some biological sequence metrics. *Advances in Mathematics*, 20(3):367–387.

Yancey, William E. 2005. Evaluating string comparator performance for record linkage. *Statistical Research Division Research Report, http://www. census. gov/srd/papers/pdf/rrs2005-05. pdf*.

Yang, Yiming, Tom Ault, Thomas Pierce, e Charles W Lattimer. 2000. Improving text categorization methods for event tracking. Em *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 65–72. ACM.