

EADW Political Personalities

Grupo 3

Artur Balanuta - 68206
Instituto Superior Técnico
artur.balanuta@ist.utl.pt

Dário Nascimento - 68210
Instituto Superior Técnico
dario.nascimento@ist.utl.pt

ABSTRACT

A extracção e análise de feeds de notícias online permite inferir inúmeros aspectos de uma determinada sociedade ou amostra populacional. Caracterizando aspectos como a relevância de determinado acontecimento, qual o sentimento geral da população relativamente a um dado acontecimento ou personalidade. Este projecto permite-nos obter informações como a opinião sobre os políticos portugueses, sobre o respectivo partido e qual a evolução da sua notoriedade ao longo do tempo. Analisamos ainda quais os países que têm mais relevância no nosso paradigma político e quais as personalidades mais citadas. Deste modo podemos inferir aspectos como a opinião que os média propagam sobre um determinado político ou o partido com melhor reputação ou que acontecimentos melhoram a opinião dos políticos nos média.

Categories and Subject Descriptors

Information Retrieval, Sentiment Analysis [

Keywords

]: Information Retrieval, Sentiment Analysis, Politics, Opinion

1. INTRODUÇÃO

O objectivo deste trabalho consiste na construção de um sistema de recolha, extração e análise de feeds (RSS) de notícias. Vamo-nos focar na recolha de personalidades Políticas Portuguesas. Como também a análise e classificação dessas opiniões. Vamos também criar um motor de busca onde iremos poder efectuar pesquisas sobre essas personalidades.

2. REQUISITOS FUNCIONAIS

A nossa solução é modular afim permitir a combinação de funcionalidades e a rápida extensibilidade. A Figura 1 explicita a arquitectura da nossa solução. Consultamos vários sites de feeds, extraímos os links para a notícia original, descarregamos a notícia, extraímos as entidades presentes

e analisamos o sentimento de cada uma das frases da notícia e associamos esse sentimento à entidade de cada frase. Em todos os casos a informação é guardada em várias tabelas utilizando uma base de dados local (**sqlite3**). As tabelas foram desenhadas por forma a poderemos efectuar pesquisas eficientes e complexas sobre a base de dados. A linguagem de programação utilizada foi o **Python**. Utilizamos também algumas ferramentas para processar e indexar a informação como é o caso do **Whoosh**, o **Feedparser**, o **Beautiful Soup** e o **NTLK**.

2.1 Colecção e Armazenamento de Notícias

A recolha de notícias é iniciada descarregando os feeds dos principais jornais políticos portugueses: Diário de Notícias, Jornal de Notícias, Diário Económico, Sol. Para cada um destes feeds utilizando a ferramenta **feedparser**, extraímos o URL da notícia original e a data do feed. Estes dados são guardados na base de dados.

Imediatamente a seguir ou à posteriori, descarregamos o conteúdo da notícia através do URL. Para tal utilizando a ferramenta **Beautiful Soup**. Construímos um HTML parser para cada um dos websites. O conteúdo é separado em Título, Sumário e Artigo de modo a atribuímos relevância distinta a cada parte do texto. A entrada na base de dados é actualizada com o conteúdo e marcado como processado. Estes dados são então utilizados para realizar as pesquisas e para extração de informação, assuntos abordados nas próximas secções.

Criamos também um sistema de caching das páginas descarregadas. Ao longo do projecto verificamos que existia um bottleneck na extração da informação da Web, logo para acelerar o processo da aquisição da informação guardamos as páginas em disco utilizando várias threads e depois efectuamos o seu processamento. Para distinguir os ficheiros, o seu nome é dado em função de uma função de resumo (SHA-1) do URL original.

2.2 Pesquisa de Notícias

Depois da obtenção do artigo do seu sumário e título procedemos a indexação desta informação. A indexação é efectuada utilizando a ferramenta **Whoosh** utilizando o modelo **BM25**. Para diferenciar e modificar o peso de cada parte da mensagem por forma a obter melhores resultados utilizamos a seguinte metodologia: O ID será o URL original do Artigo que é o nosso identificador único no Sistema. E para cada par (**ID**, **TEXTO**) que introduzimos no Whoosh a parte do texto será composta pelas três partes do artigo de seguinte modo.

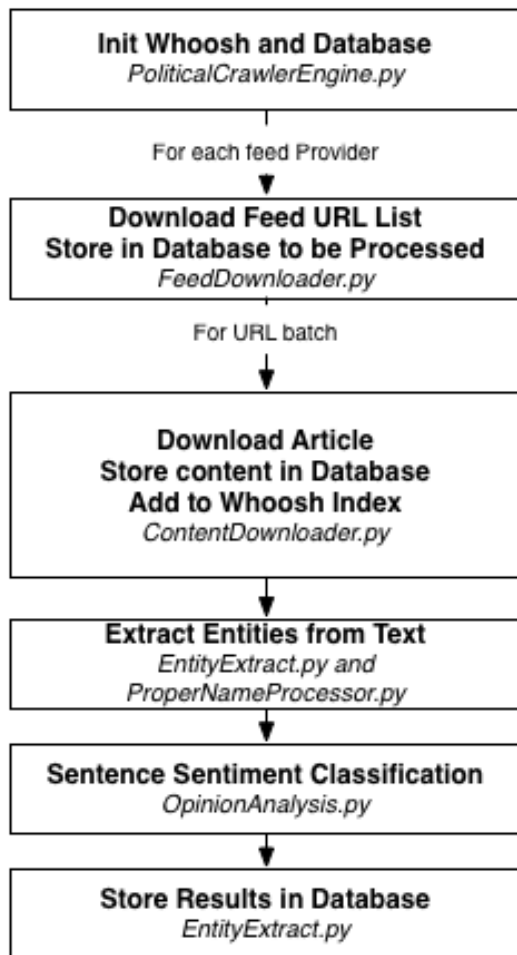


Figure 1: Arquitectura do Sistema

$$\text{TEXTO} = 10 \times \text{Título} + 2 \times \text{Sumário} + \text{Artigo}$$

Isto permite-nos obter melhores resultados pois em testes efectuados verificamos que a informação relevante vem incluída no Título ou no Sumário do Artigo.

O Resultado de uma pesquisa é um conjunto de links por ordem crescente de relevância. Para cada link efectuamos varias *queries* a base de dados por forma a encontrar toda a informação estruturada que indexamos anteriormente. Logo como temos as entidades presentes na base de dados para cada notícia apresentamos os 7 Entidades mais relevantes. A relevância das entidades é dada em função de dois factores: o numero de vezes que a entidade aparece na notícia e a relevância deste na lista de entidades fornecida para o projecto. Para cada entidade apresentamos o seu sentimento em relação a notícia, como também o sentimento geral da notícia.

2.3 Extração de Entidades

A extracção de nomes de entidades no nosso sistema é suportada por uma lista de nomes de personalidades previamente conhecida. O objectivo é que novos nomes sejam adiciona-

dos a esta lista e que nomes já existentes sejam reconhecidos com boa confiança.

Com base numa lista de entidades previamente conhecida, gerámos uma tabela de nomes próprios designada de "ProperNounTable" e uma tabela de Entidades com uma popularidade pré-definida e a popularidade adquirida nas notícias políticas designada de "EntitiesTable".

Para cada nome que o NLTK indentifica como potencial nome próprio, verificamos:

- Se o nome não pertence à lista de nomes próprios errados conhecida ("O", "A", "Desde", entre outros)
- Se o nome pertence à lista de nomes próprios conhecidos. Caso pertença, é concatenado com o nome em formação. Caso contrário é definido como candidato a nome próprio. Se o nome seguinte for também um nome próprio, então ambos são adicionados à "ProperNounTable", expandindo a lista de nomes próprios conhecida.
- Se os nomes próprios constituintes do nome da entidade desconhecida correspondem em mais de 60% ao nome de uma das entidades conhecidas. De todas as entidades conhecidas candidatas a este nome, é seleccionada a que tem maior reputação nas notícias e a que tem maior reputação pré-definida. Caso não haja nenhuma entidade conhecida com o nome, este nome é adicionado como nova entidade.

Esta abordagem permite determinar novas entidades e acrescentá-las a nossa lista.

Ao nível do processamento de texto, obtámos por unificar todos os nomes sem acentos graves ou agudos de modo a que palavras como: "Luís" ou "Luis" fossem equivalentes. Esta situação permitiu aumentar em 3 terços o número de entidades detectadas. Em particular o número de ocorrências do ministro "Vitor Gaspar" aumentou imenso.

No caso se novos nomes, o NLTK não foi capaz de identificar com precisão se um determinado nome é ou não um nome próprio, o que provocava um grande numero de falsos positivos. Isso acontece pelo facto de o NLTK classificar as palavras como Inglezas, o que resultava em resultados de pouca precisão. Para ultrapassar este problema criamos a nossa propria base de dados de Nomes Proprios. A Base de dados é formada por uma combinação de nomes extraídas de tres fontes: A lista de Politicos fornecida para o projecto, NLTK Corpus Floresta TreeBank (Portugues) e NLTK Corpus MacMorpho TreeBank (Brasileiro/Portugues). Apesar de conseguiremos resultados muito bons tinhamos muitos nomes proprios que apareciam muitas vezes e não se adequavam ao nosso caso. Mais a frente vamos explicar como conseguimos melhorar os resultados retirando as palavras desnecessarias atravez de um filtro.

2.4 Analise do Sentimento

A análise de sentimento permite determinar a reputação e ter um feedback em tempo real do panorama político nacional. A opinião que os média publicam sobre um determinado político ou partido político é extremamente relevante e decisiva para o futuro político-partidário de um país. Em 2013, as eleições em Itália sofreram uma forte influência por

parte dos média. A maioria dos média televisivos italianos são propriedade de Silvio Berlusconi, então candidato à liderança, fizeram campanhas de opinião positiva ao político e fazendo com que a opinião de um político envolvido em vários escândalos se altera-se e deste modo obtivesse quase o mesmo número de votos que Luigi Bersani.

2.4.1 *Análise do Sentimento ao nível da Frase*

Ao fim de definir um sentimento para cada entidade nos iremos classificar as frase que contem essa entidade. A frase é classificada em função dos adjectivos que esta contem. Para identificar o sentimento dos adjectivos utilizamos um lexicon (**SentiLex-PT.03**) fornecido pelo projecto *DMIR* do *indesc-id*. Após a extração dos adjectivos inserimos estes na nossa base de dados *sqlite* para rapido acesso. Para classificar uma determinada frase verificamos todas as palavras da frase contra a nossa base de dados e classificamos a quantidade de adjectivos positivos e negativos obtidos. Através do mecanismo da secção anterior, extraímos as entidades de cada frase. Assumimos que existe apenas 1 opinião por cada frase.

Uma das maiores dificuldades, especialmente em notícias políticas, é a análise de opiniões contidas nas citações de outros políticos. Estas citações são muitas vezes sarcásticas e difíceis de analisar. A frase é objectiva ou subjectiva, isto é, a frase contém alguma opinião, visão ou crença subjacente? Apenas considerámos frases objectivas porque frases subjectivas tem uma análise muito mais complexa que não vamos abordar neste projecto.

2.4.2 *Análise do Sentimento ao nível do Documento*

Admitimos um método de análise de *Supervised Learning* em que classificamos o documento em 3 classes: Positivo, Negativo e Neutro. Para definir o sentimento do texto procedemos a soma dos sentimentos de todas as frases, obtendo a classificação do documento. Esta informação é útil para o caso de pesquisas mais especificas em que indicamos se queremos obter textos com sentimento positivo ou negativo.

2.5 *Otras Funcionalidades*

Realizámos a recolha de nomes de cada um dos ministros do actual governo e marcámos as entidades como membros do governo. Além disso, fizemos parsing ao site da assembleia da republica e recolhemos todos os nomes de deputados e respectivos partidos e associamos às respectivas entidades. Com base nesta contextualização das entidades, realizámos a análises entre partidos e da avaliação do governo.

2.5.1 *Análise Partidária*

2.5.2 *Avaliação do Governo*

2.5.3 *Caracterização das entidades*

Recolhemos em cada frase os adjectivos que caracterizam as entidades dessa frase e associámos estes adjectivos à entidade. Deste modo, não só sabemos a opinião como também quais as características mais faladas a cada uma das entidades.

2.5.4 *Evolução da opinião*

Para cada entidade, podemos consultar o número de notícias positivas e o nº de notícias negativas e deste modo

saber como é que a popularidade da entidade evoluiu ao longo do tempo.

2.5.5 *Interface de Consulta*

A percepção sobre os dados recolhidos aumenta quando analisados graficamente. Para tal, criámos uma interface web de pesquisa. Esta interface foi criada no servidor Python Bottle [?].

2.5.6 *Mecanismo de caching*

Durante a execução dos testes encontramos um *bottle-neck* no caso da ligação de Internet. O problema consistia no tempo que cada pagina demorava a descarregar. Para tal criamos mecanismo de caching de paginas e de descarregamento em paralelo. A ideia baseia-se em descarregar todas as paginas com um processo em multithread para um cache local. Isso diminuiu em um terço o tempo necessario para processar a nossa base de dados. Por outro lado como todas as paginas permanciam em cache as seguintes leituras seriam locais. O nome dos ficheiros é dado utilizando o url original e usando a função de resumo SHA-1 o que assegura a unicidade dos ficheiros e evita mecanismos complexos para a atribuição dos nomes num sistema de cache.

2.5.7 *Garbage Collector de Entidades*

Como referido em 2.3 a nossa solução utiliza uma base de dados de nomes que provem de varias fontes, por outro lado o nosso sistema contém um mecanismo que tenta determinar entidades novas. Este tipo de abordagem gera muitos falsos positivos. No Nosso caso queremos nos focar nas entidades políticas que podem nem sempre ser asmais predominantes num artigo. Para resolver este problema criamos um mecanismo de treino. Este consiste em processar todos os artigos obtidos por forma a encontrar todas as entidades. Para cada entidade encontrada vamos manter um contador. No final para as cem entidades mais referenciadas vamos perguntar ao utilizador quais destas são realmente Entidades Politicas. As decisões tomadas pelo utilizador são guardadas em duas listas uma '**White List**' e uma '**Black List**'. A Black List é utilizada em conjunto com 'Stop Words' para filtrar o conteudo, desta forma estas palavras não são classificadas e não contam para as estatisticas das entidades. Por outro lado a White List é utilizada como filtro para o *trainingSet* desta forma utilizador ira sempre receber novas palavras para classificar.

Esta técnica permitiu retirar muitas palavras que não representavam entidades e os resultados podiam ser visto nas novas pesquisas efectuadas melhorando a precisão das nossas pesquisas.

3. **RESULTADOS EXPERIMENTAIS**

Nos Anexos na figura 2 e 3 podemos verificar resultados relativos ao Ministro Miguel Relvas. Os resultados do grafico são bastante interessantes pois demonstram o impacto das acções do ministro durante o intervalo de tempo apresentado. Em especial a data de 5/4/2013 demonstra uma grande concentração de artigos com sentimentos Negativos face ao Ministro, este dia foi representativo pois foi o dia em que este apresentou a sua demissão.

4. ANÁLISE CRÍTICA

O nosso projecto podia sofrer uma melhoria utilizando as ferramentas das ultimas aulas de laboratorio que ocorreram depois da data de entrega. O classificador de Bayes seria uma boa alternativa a utilizar no caso da análise do sentimento dos artigos. Seria interessante verificar quais os *outputs* das soluções. E verificar qual é o mais preciso e eficiente em termos de recursos.

5. CONCLUSÃO E TRABALHO FUTURO

O download e processamento da noticia é um processo bastante dispendioso a nível computacional e de largura de banda. Uma das melhorias sugeridas seria descarregar apenas feeds cujo o título se refira claramente a uma notícia política. Esta feature não é trivial visto que o texto do título é bastante curto e nem sempre reflecte o conteúdo da notícia.

Em geral o trabalho efectuado demonstra resultados bastante satisfatorios tendo em conta o curto tempo do projecto. A linguagem Python tem sido bastante util para este projecto pois tem um excelente leque de ferramentas que facilitam a extração e processamento da informação.

6. ANEXOS

EADW Political Searcher

Artur Balanuta, Dário Nascimento

Top Words

Top Countries

Political Parties Positive

Political Parties Neutral

Political Parties Negative

Political Parties Opinion

Miguel Relvas

Search

Frases de dois anos de Miguel Relvas no executivo (1)
Seleção de frases de Miguel Relvas como ministro Adjunto e dos Assuntos Parlamentares ...
[Miguel Relvas : 2](#) [Governo : 2](#) [Banco de Portugal : 0](#) [Miguel Pais do Amaral : -1](#) [PSD : -1](#) [Diário : 0](#) [Diário Notícias : -1](#)

Miguel Relvas demite-se do Governo (8)
O ministro da Presidência e dos Assuntos Parlamentares pediu esta quinta-feira a sua demissão do Governo de Pedro Passos Coelho, já confirmou oficialmente o gabinete do primeiro-ministro. ...
[Miguel Relvas : 2](#) [Pedro Passos Coelho : 2](#) [Governo : 2](#) [Miguel Pais do Amaral : 0](#) [Álvaro Santos Pereira : 0](#) [Diogo Feio : 0](#) [Parlamentares : 2](#)

CDS limita-se a respeitar demissão de Miguel Relvas (2)
O líder parlamentar democrata-cristão Nuno Magalhães limitou-se hoje a dizer que a bancada centrista "respeita" a decisão de pedir a demissão por parte do ministro-Adjunto e dos Assuntos Parlamentares, Miguel Relvas. ...
[Miguel Relvas : 1](#) [Pedro Passos Coelho : 0](#) [Governo : 0](#) [Nuno Magalhães : 1](#) [Diogo Feio : 0](#) [PSD : 0](#) [Executivo : 0](#)

Crato não informou Miguel Relvas do relatório da inspeção (70)
(ATUALIZADA) O ministro Nuno Crato garantiu esta quinta-feira à noite que não informou Miguel Relvas sobre os resultados da auditoria da Inspeção-Geral da Educação e Ciência (IGEC). "Nunca falámos com o senhor ministro Miguel Relvas sobre este assunto", disse o ministro da Educação, em entrevista à SIC Notícias. ...
[Miguel Relvas : 10](#) [Pedro Passos Coelho : 10](#) [Nuno Crato : 10](#) [Público : 10](#) [Ciência : 10](#) [SIC Notícias : 10](#) [Tribunal Administrativo Lisboa : 10](#)

Passos diz que Miguel Relvas "não cometeu nenhum abuso" (-10)
O primeiro-ministro desvalorizou hoje os efeitos da demissão do ministro adjunto Miguel Relvas, considerando, em contrapartida, que a moção de censura apresentada quarta-feira pelo PS criou muito mais "instabilidade" no país. ...
[Miguel Relvas : -3](#) [António José Seguro : -1](#) [Pedro Passos Coelho : -3](#) [Governo : -1](#) [Jerónimo de Sousa : 0](#) [Nuno Crato : -2](#) [José Sócrates : 0](#)

PSD respeita e compreende motivos invocados por Relvas (9)
O PSD manifestou hoje "respeito" e "compreensão" pelos motivos invocados por Miguel Relvas para abandonar o Governo e considerou "prematurados" os pedidos para que o ministro da Educação explique os processos de licenciatura na Universidade Lusófona. ...
[Miguel Relvas : 2](#) [Pedro Passos Coelho : -1](#) [Governo : 2](#) [Luís Montenegro : 1](#) [PSD : 2](#) [Parlamentares : 1](#) [Universidade : 2](#)

Figure 2: Resultado de uma pesquisa pelo Ministro Miguel Relvas

Miguel Relvas

Governo: Ministro Adjunto e dos Assuntos Parlamentares

Partido: PSD

Reputation: 629

Adjectives

adjunto,16,própria,8,são,8,partido,7,profissional,6,direito,6,grande,6,anterior,6,atual,5,fragilizado,5,político,5,Administrativa,5,tardia,4,antigo,4,vão,4,aceite,3,negativa,3,alma,3,preciso,3,longo,3

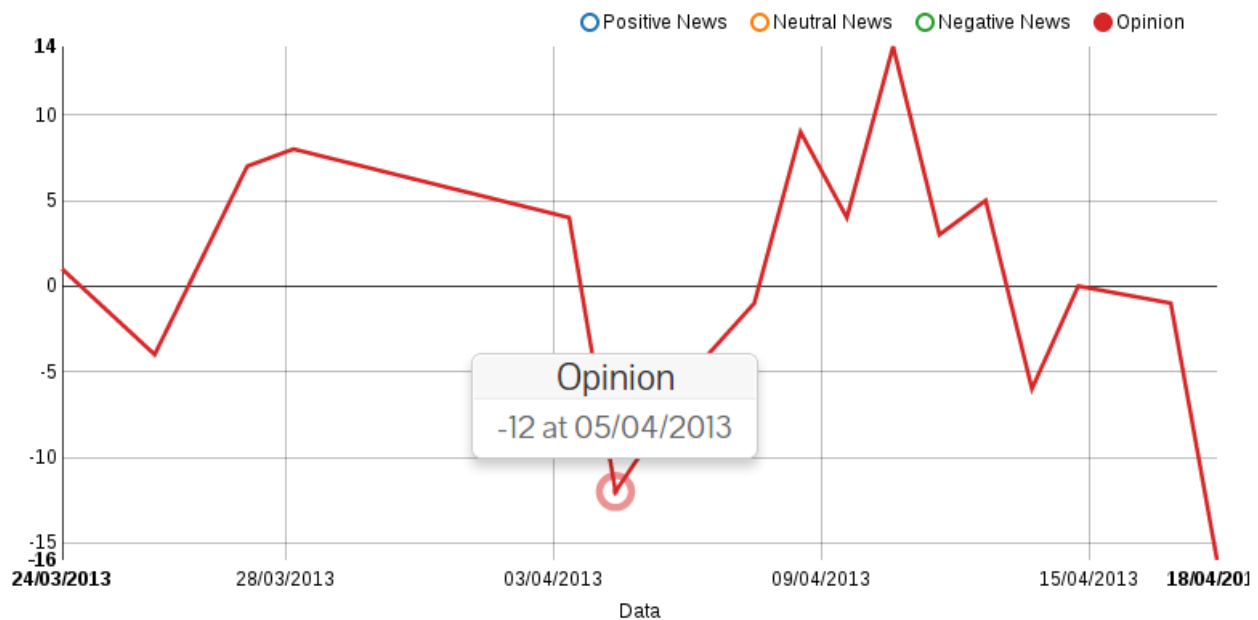


Figure 3: Opinião sobre o Ministro Miguel Relvas no dia 5/4/2013