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 notiaridade 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 Retrial, Sentiment Analysis [

Keywords

]: Information Retrial, Sentiment Analysis, Politics,Opinion

1. INTRODUÇÃO

O objectivo deste trabalho consiste na construção de um sistema de recolha, extração e analise de feeds (RSS) de notícias. Vamo-nos focar na recolha de personalidades Politicas Portuguesas. Como tambem a analise e clasificação dessas opinioes. Vamos tambem criar um motor de busca onde iremos poder efectuar pesquisas sobre esas personalidades.

2. REQUISITOS FUNCIONAIS

A nossa solução é modular afim permitir a combinação de funcionalidades e a rápida extensabilidade. A Figura1 explicita a arquitectura da nossa solução. Consultamos vários sites de feeds, extraimos os links para a notícia original, descarregamos a noticia, extraimos as entidades presentes e analisamos o sentimento de cada uma das frases da noticia e associamos esse sentimento à entidade de cada frase. Em todos os casos a informação é guardada em varias tabelas utilizando uma base de dados local **(sqlite3)**. As tabelas foram desenhadas por foram a poderemos efectuar pesquisas eficientes e complexas sobrea a base de dados. A linguagem de programação utilizada foi o **Python**. Utilizamos tambem algumas faramentas 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 noticias é iníciada descarregando os feeds dos prinicípais jornais politicos portugueses: Diário de Notícias, Jornal de Notícias, Diário Económico, Sol. Para cada um destes feeds utilizando a faramenta **feedparser**, extraimos 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 farramenta **Beautiful Soup**. Construímos um HTML parser para cada um dos websites. O contéudo é separado em Título, Sumário e Artigo de modo a atribuírmos relevância distinta a cada parte do texto. A entrada na base de dados é actualizada com o contéudo e marcado como processado.
Estes dados são então utilizados para realizar as pesquisas e para extracção de informação, assuntos abordados nas próximas secções.
Criamos tambem um sistema de caching das paginas descaregadas. Ao longo do projecto verificamos que exestia um bottleneck na extração da informação da Web, logo para acelerar o processo da aquisição da informação guardamos as paginas em disco utilizando varias threads e depois efectuamos o seu processamento. Para distinguir os ficheros, o seu nome é dado em função de uma função de resumo (SHA-1) do URL original.

2.2 Pesquisa de Notícias

Despois da obtenção do artigo do seu sumario e titulo procedemos a indexaçao desta informação. A indexação é efectuada utilizando a ferramenta **Whoosh** utilizando o modelo **BM25**. Para diferenciar e modificar o pesso de cada parte da messagem por forma a obter melhores resultados utilizamos a seguinte metodologia: O ID sera o URL original do Artigo que é o nosso identificador unico no Sistema. E para cada par **(ID, TEXTO)** que introduzimos no Whoosh a parte do texto sera composta pelas tres partes do artigo de seguinte modo.

Figure 1: Arquitectura do Sistema

TEXTO = 10 x Título + 2 x Sumário + Artigo

Isto permite-nos obter melhores resultados pois em testes efectuados verificamos que a informação relevante vem incluida no Título ou no Sumário do Artigo.
O Resultado de uma pesquisa é um conjunto de links por ordem crescente de relevancia. Para cada link efectuamos varias *queries* a base de dados por forma a encontrar toda a informação estruturada que indexamos anteriormente. Logo como temos o as entidades presentes na base de dados para cada noticia apresentamos os 7 Entidades mais relevantes. A relevancia das entidades e dada em função de dois factores: o numero de vezes que a entidade aparece na noticia e a relevancia deste na lista de entidades fornecida para o projecto. Para cada entidade apresentamos o seu sentimento em releção a noticia, como tambem o sentimento geral da noticia.

2.3 Extração de Entidades

A extracção de nomes de entidades no nosso sistema é suportada por uma lista nomes de personalidades previamente conhecida. O objectivo é que novos nomes sejam adicionados 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 noticías politicas designada de "EntitiesTable".
Para cada nome que o NLTK indentifica como potêncial nome próprio, verificamos:

- Se o nome não pertence à lista de nomes próprios errados conhecida ("O","A", "Desde", entre outros)
- Se o nome pertente 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 noticias 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 acrescentalas 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 "Vítor 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 proprio, o que provocava um grande numero de falsos positivos. Isso acontece pelo facto de o NLTK classificar as plavaras 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 extraidas de tres fotes: A lista de Politicos fornecida para o projecto, NLTK Corpus Floresta TreeBank (Portugues) e NLTK Corpus MacMorpho TreeBank (Brasileiro/Portugues). Apesar de conseguiremos reultados 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 fitro.

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 posítiva ao político e fazendo com que a opinião de um político envolvido em vários escandalos se altera-se e deste modo obtivesse quase o mesmo número de votos que Luigi Bersani.

2.4.1 Análise do Sentimento ao nivel da Frase

Ao fim de defenir um sentimento para cada entidade nos iremos classificar as frase que contem essa entidade. A frase é clasificada 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 clasificamos a quantidade de adjectivos positivos e negativos obtidos. Através do mecanismo da secção anterior, extraimos as entidades de cada frase. Assumimos que existe apenas 1 opinião por cada frase.
Uma das maiores difículdades, especialmente em notícias políticas, é a análise de opiniões contídas nas citações de outros políticos. Estas citações são muitas vezes sarcásticas e dificeis 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 nivel do Documento

Admitimos um método de análise de *Supervied Learning* em que classificamos o documento em 3 classes: Positivo, Negativo e Neutro. Para defenir o sentimento do texto procedomos a soma dos sentimentos de todas as frases, obtendo a classificação do documento. Esta informação é util 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 caracteristicas mais faladas a cada uma das entidades.

2.5.4 Evolução da opinião

Para cada entidade, podemos consultar o número de noticias positivas e o n^{o} de noticias 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 analisádos gráficamente. 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 *bottleneck* no caso da ligção de Internet. O problema consistia no tempo que cada pagina demorava a descaregar. Para tal criamos mecanismo de caching de paginas e de descaregamento em paralelo. A ideia baseia-se em descaregar todas as paginas com um processo em multithread para um cache local. Isso diminuiu em um terço o tempo nessesario 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 ficheros 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 preguntar 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 *triningSet* 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 demostra 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 CRÍTCA

O nosso projecto podia sofrer uma melhoria utilizando as faramentas das ultimas aulas de laboratorio que ocoreram 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* duas soluções. E verificar qual é o mais preciso e eficiente em termos de recursos.

5. ANEXOS

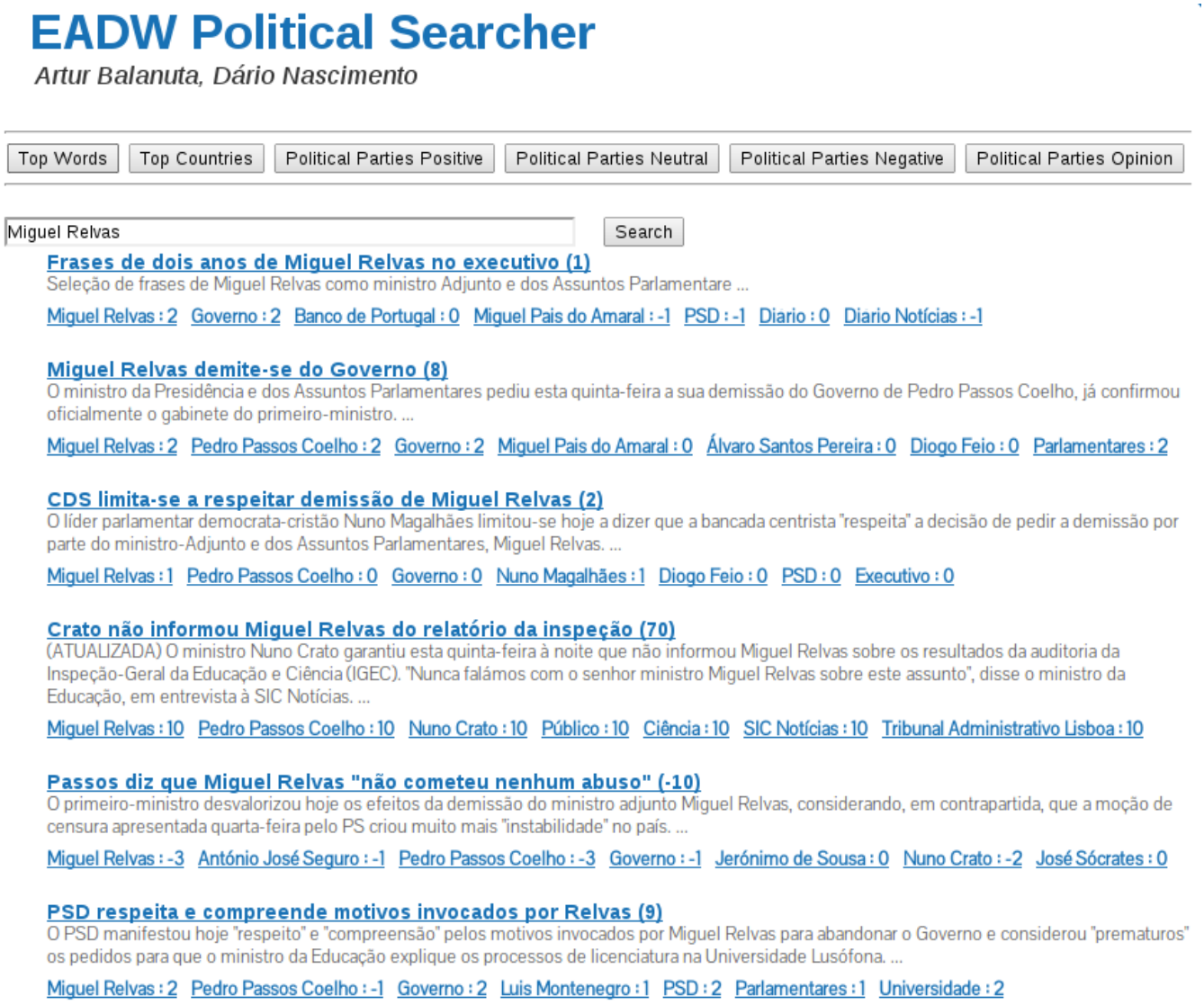

Figure 2: Resultado de uma pesquisa pelo Ministro Miguel Relvas

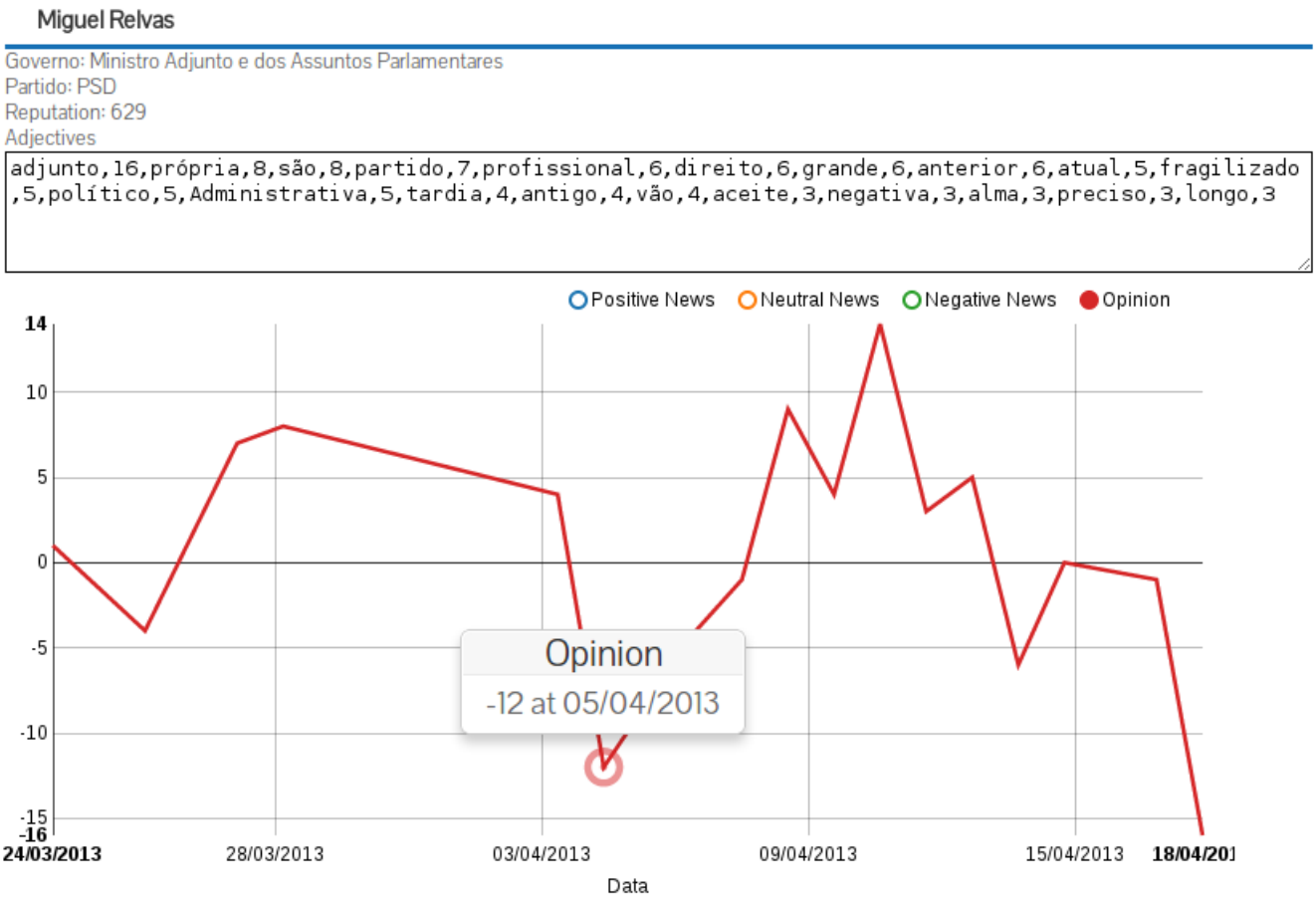

Figure 3: Opinião sobre o Ministro Miguel Relvas no dia 5/4/2013