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. INTRODUCTION

2. FUNCTIONAL REQUIREMENTS

A nossa solução é modular afim permitir a combinação de funcionalidades e a rápida extensabilidade. A 1 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.

2.1 News Collection and Storage

A recolha de noticias é iníciada descarregando os feeds dos prinicípais jornais politicos portugueses: Diário de Notícias, Jornal de Notícias, Diário Económico, Sol. De cada um destes feeds, extraimos o URL da notícia original e a data do feed. Estes dados são guardados na base de dados.

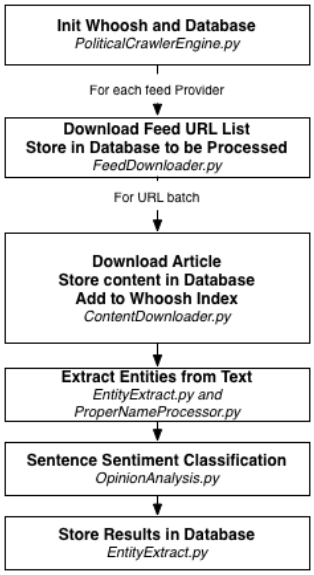

Figure 1: Arquitectura do Sistema

Imediatamente a seguir ou à posteriori, descarregamos o conteúdo da notícia através do URL. Para tal construímos um HTML parser para cada um dos websites. O contéudo é separado em Título, Sumário e Article de modo a atribuírmos relevância distinta a cada palavra. 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.

2.2 News Search

2.3 Extraction of Named Entities

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.

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.

2.4 Sentiment Analysis

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.
A análise de sentimento foi realizada ao nível de ———— —- A nossa análise de sentimento consiste em várias fases:

1º - Pré-processamos o texto e extraímos as entidades de cada frase, como explicado na secção anterior.

2.4.1 Sentence-Level Sentiment Analysis

Através do mecanismo da secção anterior, extraimos as entidades de cada frase. Assumimos que existe apenas 1 opinião por cada frase. No entanto, caso exista mais do que uma entidade, tentámos dividir a frase em subfrases.
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 difíceis de analisar. 1º - A frase é objectiva ou subjectiva, isto é, a frase contém alguma opinião, visão ou crença subjacente? Apenas considerámos frases subjectivas porque frases objectivas tem uma análise muito mais complexa.

VER O MÉTODO DO Hai e o Pang and Lee (32 e 26) Ler o 35 sobre sarcasmo

TENTAR INDENTIFICAR SARCARMOS

2º

2.4.2 Document-Level Sentiment Analysis

Admitimos um método de análise de ıSupervied Learning em que classificamos o documento em 3 classes: Positivo, Negativo e Neutro. Ao nível do documento USAR UM MÉTODO COMO O SVM, BAYES, Logistic Regression o KNN USAR TFIDF, POS e Sentiment Lexicon para analisar

2.4.3 Sentiment Lexicon Acquisition

Existem 3 formas de obter o léxico de sentimentos: ıManualmente, ıDictionary-based e ıCorpus-based. Uma hipótese possível de criação do léxico seria criar uma pequena lista de adjectivos manualmente. Utilizando um dicionário online, como a infopedia [**?**], verificar que se trata de um adjectivo, extrair os seus sinónimos e antónimos e realizar a mesma análise iterativamente para cada um. Um processo semelhante a um url-crawler mas com palavras num dicionário. Este processo geraria um léxico classificado como palavras positivas ou palavras negativas.
Em alternativa, utilizámos o léxico Sentilex [**?**]. Processámos este léxico para uma lista de expressão:Part-of-Speech:Gênero:Opinião
A opinião das expressões foi dividida em Positiva, Negativa ou Neutra.

2.5 Other Functionalities

3. EXPERIMENTAL RESULTS

Precision e recal do sistema

4. CONCLUSIONS AND FUTURE WORK

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 contéudo da notícia.

4.1 References