

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. INTRODUCTION

2. FUNCTIONAL REQUIREMENTS

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

2.1 News Collection and Storage

2.2 News Search

2.3 Extraction of Named Entities

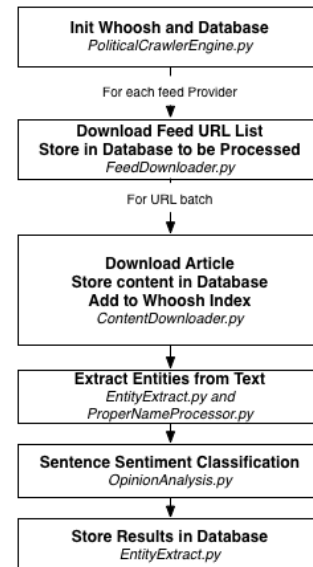


Figure 1: Arquitectura do Sistema

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

Com base numa lista de entidades previamente conhecida, geramos 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 identifica como potencial nome próprio, verificamos:

- Se o nome não pertence à lista de nomes próprios conhecidos, é adicionado à lista de nomes próprios conhecidos ("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 conhecidos.

- 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.

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

2.5 Other Functionalities

3. EXPERIMENTAL RESULTS

4. CONCLUSIONS AND FUTURE WORK

O download e processamento da notícia 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.

4.1 References

4.2 References