INSTITUTO SUPERIOR TÉCNICO

Extracção e Análise de Dados na Web

Extracção e análise de informação contida em *feeds* de notícias. Grupo 11

Pedro Braz 73991

Rui Mangas 70600

1 INTRODUÇÃO

O objectivo deste trabalho consiste na extracção e análise de *feeds* RSS de noticías. Através desta análise serão extraídos diversos nomes de personalidades, as relações entre elas e estatísticas que revelam quais as personalidades mais populares.

Este documento descreve a implementação realizada, os algoritmos usados e as mais diversas decisões tomadas ao longo do desenvolvimento do projecto. Nas secções seguintes serão descritos os seguintes tópicos: recolha e armanezamento de notícias, procura nas notícias, extracção de entidades, relações entre entidades, estatísticas, e por fim uma conclusão que descreve os resultados obtidos.

2 Extração e armanezamento de notícias

As notícias são extraídas de várias fontes jornalísticas e são introduzidas numa plataforma de indexação chamada *whoosh*. O *whoosh* compila as notícias utilizando o algoritmo de classificação *BM25*. Posteriormente, são armazenadas numa colecção de notícias na nossa base de dados não relacional *MongoDB*. Para aumentar a velocidade da extracção, os pedidos às páginas *web* são separados em *threads* com o objectivo de conseguirmos um maior *throughput* visto que a comunicação pela internet é o maior *bottleneck* da aplicação.

Na nossa aplicação o *feedparser*, acede a um conjunto de *links* RSS do qual extrai os endereços das notícias que são acedidos paralelamente. Cada página é analisado com o *beautiful soup* sendo extraídos unicamente o título e o corpo do artigo. Para isto, foram criados três *parsers* que lidam com a estrutura da páginas de maneira diferente dependendo do site da notícia.

As notícias são armazenadas na base de dados com o objectivo de evitar colisões, ou seja, caso uma notícia já tenha sido processada anteriormente, não é indexada novamente. Além disso, o *whoosh* apresenta os resultados com os *links* das notícias, para que depois os elementos de cada notícia possam ser acedidos a partir da base de dados, utilizando o *link* como chave. Por fim, faz-se a extração e armazenamento das entidades. Este último tópico é explicado em detalhe nas secções seguintes.

Na figura seguinte, apresentamos o fluxo de execução descrito anteriormente.

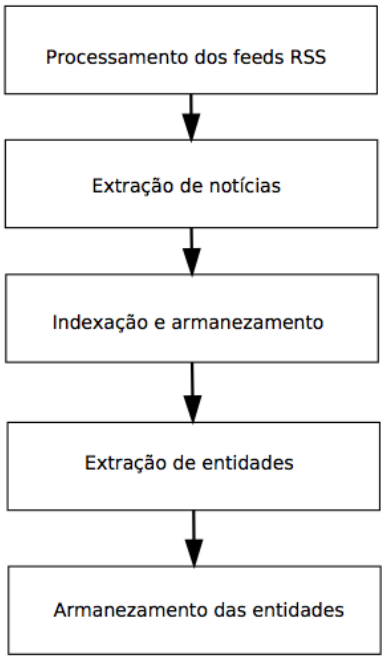

Figure 2.1: Fluxo de execução

3 PROCURA DE NOTÍCIAS

O utilizador, pode fazer procuras, fornecendo uma *query*. Esta é processada pelo *whoosh* que trata de encontrar correspondências devolvendo o *link* da notícia encontrada. Este *link* é depois usado para fazer *queries* à base de dados para ser devolvido os diversos elementos do artigo. Após a procura é devolvido ao utilizador os títulos das notícias, por ordem de relevância, onde a *query* foi encontrada. À frente de cada título, é apresentado as entidades encontradas na noticía. Este último tópico é explicado na secção seguinte.

Os módulos onde esta parte do projecto foi desenvolvida encontram-se na pasta *storageTools*. No ficheiro *whoosh_tools* estão presentes as funções de indexação e procura. No ficheiro *mongo_tools* encontram-se as funções para fazer operações na base de dados. Na secção Anexos, encontram-se vários exemplos de pesquisa de notícias.

4 EXTRAÇÃO DE ENTIDADES

A extração de entidades no nosso sistema é realizada com base numa lista de nomes de personalidades conhecida *a priori*. O objectivo desta abordagem é comparar os resultados obtidos no nosso sistema com os da lista de nomes como forma de validação. Para obter uma a nossa lista de nomes, usámos uma técnica de processamento de linguagem natural (através do módulo *nltk*). O procedimento é o seguinte:

1. Recuperação de todas as notícias da base de dados MongoDB;
2. Para cada frase de cada notícia foram gerados *tokens* através da função *nltk.sent_tokenize*
3. Classificação de cada palavra de cada frase através da função *nltk.pos_tag*;
4. Verificação do nó PERSON para ver se este está contido na lista de entidades descrita anteriormente. Se sim, consideramos como sendo uma entidade. Caso contrário, descartamos esse nome. Estas entidades descobertas são depois colocadas numa nova coleção da base de dados à qual demos o nome de *namesOfPersons*.

Contudo, esta implementação produz alguns problemas. Sendo por exemplo, muito comum o nome do primeiro ministro aparecer nas notícias como 'Passos Coelho' e não como 'Pedro Passos Coelho' o nosso sistema não considera o primeiro caso como sendo uma entidade pois na lista inicial só aparece o segundo caso. Tentámos resolver este problema fazendo procuras por *substrings*. Contudo, apesar desta abordagem funcionar para casos como o descrito acima, produzia inúmeros resultados errados. Como resultado disso, decidimos não comprometer o nosso sistema e mantivémos a abordagem inicial estando cientes que falhava nalguns casos.

As entidades extraídas são apresentadas ao utilizador quando este faz uma procura. Apresentamos os títulos das notícias que contêm uma desterminado *query* escrita pelo utilizador e à frente desta as entidades encontradas na mesma. Em anexo, estão contidas algumas imagens que descrevem a funcionalidade anterior.

Por fim, para efeitos estatísticos fazemos uma contagem de entidades para conseguirmos determinar a personalidade que mais apareceu em todas as notícias extraídas.

O desenvolvimento desta parte do projecto está contido na pasta *entities*. Dentro dessa pasta existem três ficheiros: *namesOfEntities*, *relationships* e *statistics*. O primeiro, serve para extrair as entidades das noticías, o segundo para descobrir as relações entre elas e o último para realizar as estatísticas.

5 Descoberta de relações

Para efectuar a descoberta de relações seguimos o seguinte modelo: se duas personalidades estão contidas na mesma notícia existe relação entre elas. A nossa coleção de base de dados, *namesOfPersons*, contém para cada notícia, uma lista com todas as entidades descobertas na mesma. Para desenvolver esta funcionalidade, percorremos a colecção anterior e contruímos um dicionário que contém como *key* o nome de uma determinada personalidade e como *value* uma lista com todas as entidades relacionadas com a mesma. O utilizador quando desejar descobrir as entidades relacionadas com uma certa personalidade, faz uma procura por esse nome e é-lhe devolvido o *value* do dicionário associado à *key* nome. Mais uma vez, apresentamos em anexo uma imagem que descreve a funcionalidade descrita neste tópico.

6 Estatísticas

Para efeitos estatíscos apresentamos duas opções para o utilizador:

1. Descoberta da personalidade que mais apereceu nas noticias extraídas;
2. Dado um nome de uma personalidade devolvemos o número de vezes que esta apareceu nas noticías.

O procecidemento para desenvolver os tópicos atrás enumerados, foi uma simples contagem dos elementos presentes na base de dados. Em baixo, apresentamos alguns gráficos que representam o número de vezes que uma certa personalidade apareceu nas notícias num determinado dia.

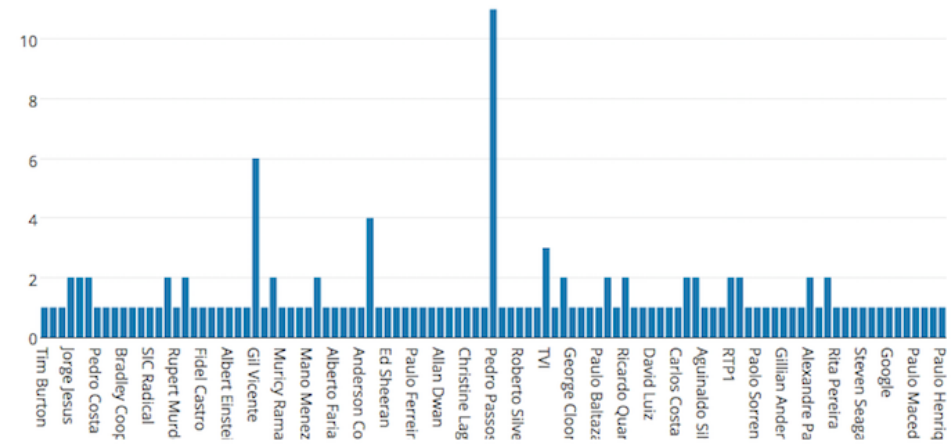

Figure 6.1: Personalidades no dia 9 de maio de 2015

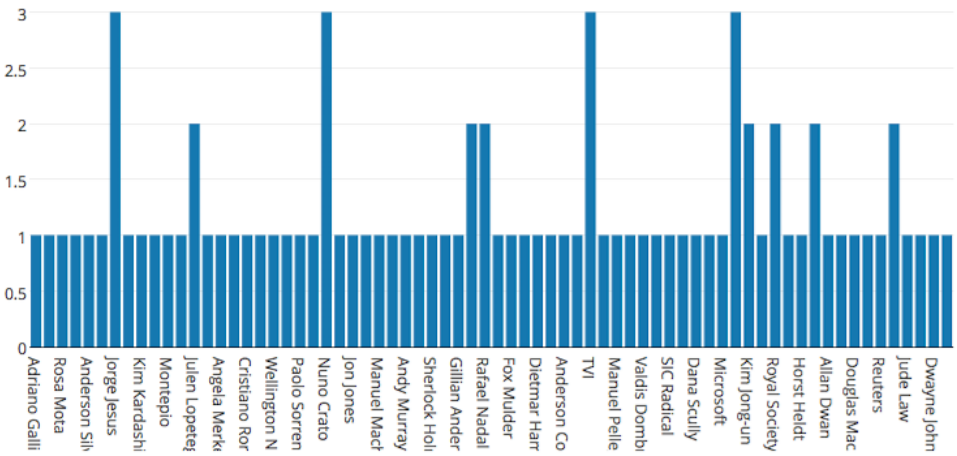

Figure 6.2: Personalidades no dia 13 de maio de 2015

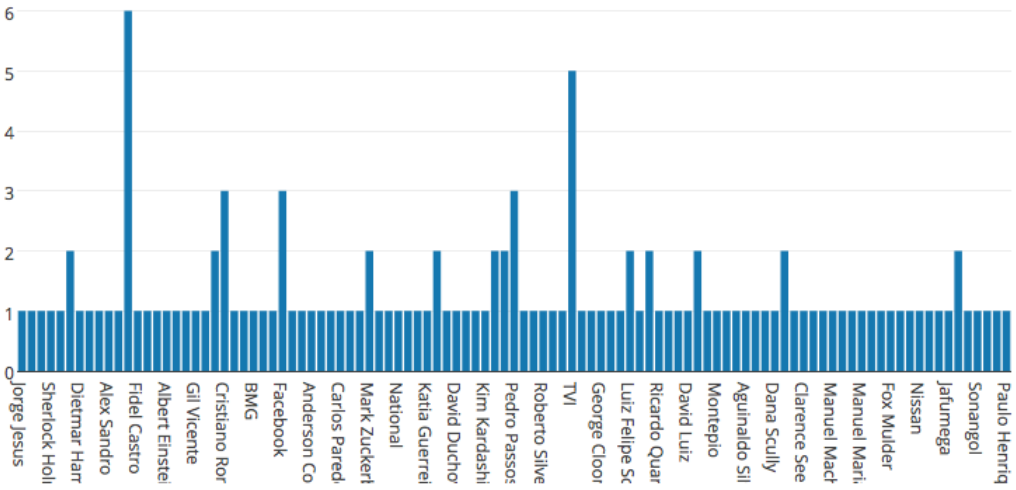

Figure 6.3: Personalidades no dia 14 de maio de 2015

7 Análise Crítica

Com base nos exemplo que apresentamos nos anexos, verificamos que o nosso projecto podia sofrer bastantes melhorias na parte da recolha de entidades. Observa-se por exemplo, que na procura pela *query* 'morte' não foram encontradas quaisquer entidades nas quatro notícias devolvidas. Isto deve-se principalmente ao facto de a ferramenta *nltk* estar preparada para a análise de texto na língua inglesa. Uma melhoria a fazer no futuro seria usar o módulo *floresta* do *nltk* para tentar obter melhores resultados na língua portuguesa. Contudo, com base nos gráficos apresentados na secção estatísticas achamos que os resultados são satisfatórios pois, mesmo assim são encontrados várias entidades portuguesas.

8 Conclusão

Após a realização deste projecto, concluímos que podíamos ter obtido melhores resultados em certas partes do projecto. A linguagem *python* é adequada a este tipo de projectos pois tem disponível um grande conjunto de ferramentas que torna a extração e o processamento de informação um processo mais simples. Além dos melhoramentos explicados atrás, também para trabalho futuro podia ser construída um interface web para o utilizador efectuar as pesquisas e consultar os resultados.

9 ANEXOS

De referir que todos este exemplos são referentes às notícias extraídas no dia 13/05/2015.

9.1 PROCURA E DESCOBERTA DE ENTIDADES

```
1) Fetch News
2) Search news
3) Get All News
4) Get statistics
5) Relationships
0) Quit
>>2
Please enter something to search for: benfica
4 Articles found:
Dragões e o manto do dirigente dos árbitros estendido ao Benfica --> Julen Lopetegui|Jorge Jesus
Lopetegui acusa Benfica de estar sob um manto protetor --> Jorge Jesus|Julen Lopetegui
Diretor de comunicação do Benfica responde a Lopetegui --> Jorge Jesus
Vasco da Gama vence Macaé e lidera Grupo A da Taça Guanabara --> Anderson Costa
```

Figure 9.1: Procura pela query 'benfica'

```
1) Fetch News
2) Search news
3) Get All News
4) Get statistics
5) Relationships
0) Quit
>>2
Please enter something to search for: Nuno Crato
5 Articles found:
Ministro garante dinheiro para obras no Conservatório --> Nuno Crato
Fechadas negociações para descentralizar Educação --> No entities found.
FNE desconvoca greve ao exame de inglês --> Nuno Crato
Mais de 53 mil pessoas no primeiro dia da Festa do Cinema --> No entities found.
Paulo Dentinho é o novo diretor de informação da RTP --> Paulo Baltazar|RTP1|RTP Internacional|RTP2
```

Figure 9.2: Procura pela query 'Nuno Crato'

```
>>2
Please enter something to search for: morte
4 Articles found:
"Striptease" em funerais passa a ser proibido --> No entities found.
Morreu o pintor Rui Pimentel --> No entities found.
Violador diz em documentário que a culpa é das raparigas  --> No entities found.
Peça de Howard Baker no Teatro de Almada --> No entities found.
```

Figure 9.3: Procura pela query 'morte'

9.2 Relações entre entidades

```
Please enter something to search for: Cristiano Ronaldo
1 Articles found:
Bradley Cooper e Irina Shayk apanhados a namorar em Londres  --> Bradley Cooper|Cristiano Ronaldo|Dwayne Johnson
1) Fetch News
2) Search news
3) Get All News
4) Get statistics
5) Relationships
0) Quit
>>5
Entity name: Bradley Cooper
Relationships:
Bradley Cooper --> Dwayne Johnson|Cristiano Ronaldo
```

Figure 9.4: Procura por entidades relacionadas com Bradley Cooper

9.3 Estatísticas

```
1) Fetch News
2) Search news
3) Get All News
4) Get statistics
5) Relationships
0) Quit
>>4
The most famous entities are: Jorge Jesus,Nuno Crato,TVI,SIC
Name to search: Cristiano Ronaldo
Cristiano Ronaldo appeared 1 times in all the news.
```

Figure 9.5: Estatísticas acerca de Cristiano Ronaldo