INSTITUTO SUPERIOR TÉCNICO

# Extracção e Análise de Dados na Web

**Extracção e análise de informação contida em *feeds* de noticías.**

Pedro Braz 73991

Rui Mangas 70600

## 1 INTRODUÇÃO

O objectivo deste trabalho consiste na extracção e análise de *feeds* RSS de notícias. 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 noticías, procura nas noticías, extracção de entidades, relações entre entidades, estatísticas, e por fim uma conclusão que descreve os resultados obtidos.

# 2 EXTRACÇÃO E ARMANEZAMENTO DE NOTICÍAS

# 3 PROCURA DE NOTICÍAS

# 4 Extracção de entidades

A extracção 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 noticía 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. Estás entidades descobertas são depois colocadas numa nova colecção 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 noticías 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.

# 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 colecção 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 noticías extraídas;
2. Dado um nome de uma personalidade devolvemos o número de vezes que esta apareceu nas noticías.