Extração Frases Chave

Processamento e Recuperação de Informação

Gonçalo Bustorff Silva
Instituto Superior Técnico
82050

António Santos
Instituto Superior Técnico
87632

Gonçalo Adolfo
Instituto Superior Técnico
97090

1 INTRODUÇÃO

O trabalho realizado apresentava como objectivo, tal como descrito na primeira parte, a extracção de termos chave de documentos. Na primeira parte, foram implementados e avaliados modelos com base na relevância do conteúdo (TF-IDF, BM25, BM25-F). Ao longo deste documento, irão ser explorados abordagens com base em análise de grafos e uma abordagem baseada em rank aggregation. Por fim, foi implementada uma aplicação prática dos modelos desenvolvidos através de uma aplicação para a WWW (World Wide Web).

2 ABORDAGEM GRAPH RANKING

O problema de extracção de frases chave pode ser decomposto em uma problema de análise de ligações (estrutura). Deste modo, é necessário transformar um documento sob a forma de um grafo. Para tal, são extraídos n-gramas com n entre 1 e 3 (termos candidatos) em que cada candidato será representado por um nó no grafo. Dois nós estarão interligados se estiverem presentes na mesma frase. O sistema correspondente a esta abordagem está ilustrado na figura 1. No caso do exercício 1, o documento é lido do sistema de ficheiros, enquanto que para o exercício 2 os documentos do dataset (em formato XML) são processados e armazenados em um ficheiro pickle para evitar constantemente este processamento. Para cada documento, a sequência de blocos é o mesmo em cada um dos exercícios:

- extrai-se os termos candidatos;
- constrói-se o grafo correspondente ao documento;
- aplica-se o algoritmo PageRank;
- obtêm-se os nós (candidatos) com maior prestígio.

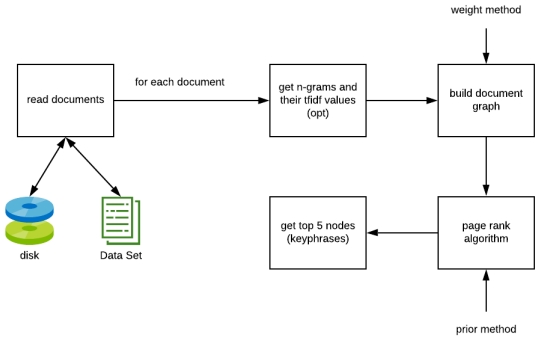

Figure 1: Sistema abordagem Graph Ranking

Note-se que a construção do grafo pode apresentar diferentes tipos de métodos para estimar o peso das ligações: unitários, co-ocorrência em uma background collection, semelhança de strings. É dado pelo seguinte pseudo-code:

```
split doc into sentences
calculate weight matrix
for each sentence do
    reset nodes list
    for each candidate do
        if candidate in sentence: then
            create node if not already exists
            update node edges
            add node to sentence nodes list
```

Após a construção do grafo, este é passado ao algoritmo PageRank, que de forma resumida, trata-se de um algoritmo iterativo que estima actualiza prestígio com uma fórmula baseada nas ligações e o prestígio dos nós ao qual está conectado. Esta fórmula também inclui uma dada probabilidade de um determinado nó não "saltar" directamente para um nó para o qual está interligado (Priors). São implementados três métodos para a função probabilidade dos Priors: função uniforme, função baseada no valor TF-IDF do candidato ou com base na posição do candidato no documento.

O script implementado para o exercício 1 permite ler um documento do sistema de ficheiros e aplicar o sistema descrito. No exercício 2, o sistema é avaliado em um dataset composto por abstracts com métricas já extraídas na primeira parte do trabalho (tabela 1): precision, recall, mean average precision e tempo de processamento. Nesta parte do trabalho, foi considerada correta uma determinada keyphrase se esta for uma substring de pelo menos uma das true keyphrases ou vice-versa.

weight	priors	prec	rec	map	t
unitary	uniform	0.68	0.42	0.37	244s
unitary	sentence pos	0.70	0.43	0.39	280s
unitary	tfidf	0.72	0.45	0.40	271s
str_sim	tfidf	0.72	0.44	0.40	3614s

Table 1: Resultados Graph Ranking (resumido)

3 ABORDAGEM RANK AGGREGATION

Durante ambas as partes do trabalho, foram atribuídos diferentes scores a cada candidato para um dado documento: TF, IDF, TF-IDF, BM25, BM25-F, PageRank. A abordagem Rank Aggregation tem como objetivo agregar os scores mencionados de modo a que cada

candidato seja apenas descrito por um único valor. Para tal, cada candidato é representado por um vetor de características, calculados recorrendo à biblioteca sklearn e ao sistema descrito na secção anterior. Para avaliar esta abordagem, são previamente calculados os vectores para cada candidato e para cada documento, armazenados num ficheiro pickle (evitar tempo de processamento). Foram experimentados diferentes sets de características obtendo os seguintes resultados:

features	prec	rec	map
tf, idf	0.21	0.14	0.07
tf, idf, tf-idf	0.20	0.13	0.06
tf-idf, pr	0.17	0.10	0.05
tf, idf, pr	0.20	0.13	0.06

Tal como se pode observar, não se verificou um impacto significativo relativamente às características a utilizar no vetor. No entanto, existiu uma ligeira descida ao serem utilizadas somente as características tf-idf e prestigio proveniente do algoritmo PageRank.

4 APLICAÇÃO PRÁTICA

De modo a ilustrar a aplicação do sistema desenvolvido, foi implementada uma aplicação web. Esta ilustra os termos chave para os artigos no RSS (Really Simple Syndication) da revista New York Times na categoria de desporto. A figura 2 ilustra os blocos do sistema. O servidor web foi implementado com auxílio ao package tornado, sendo possível a geração de páginas HTML (tornado template) com código python embebido. O servidor, ao receber um pedido HTTP, executa os seguintes blocos:

- faz um pedido HTTP para obter o ficheiro XML (RSS) e faz o seu processamento;
- extrai os termos chave para cada artigo, sendo este composto pelo título e pela descrição, adoptando o sistema da primeira secção (Graph Ranking);
- renderiza a página html (tornado template) passando-lhe os artigos e os respectivos termos chave sob a forma de um dicionário, de modo a que seja possível o encode para o formato JSON.

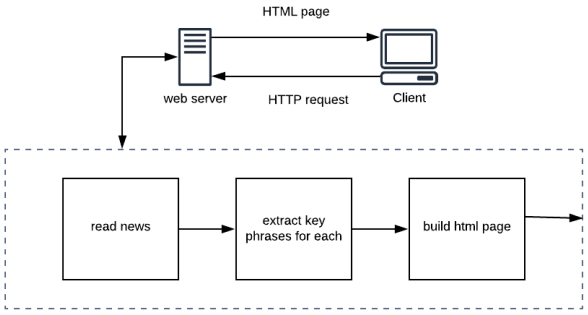

Figure 2: Aplicação web

A interface gráfica é composta por duas secções (figuras 3 e 4): lista de termos chave/artigos e uma word cloud. Na lista da esquerda, estão ilustradas todos os termos chave, com o número de artigos onde o termo se verificou. Esta permite o click por parte do utilizador, actualizando a lista de artigos do lado direito. Na lista de artigos, estes são representados pelo seu título e pelos termos chave que o representam. Para o desenho do idioma word cloud, todos artigos (título e descrição) são concatenados num único documento, sendo os termos chave extraídos deste. A proporção entre palavras é dada pelo valor do prestigio.

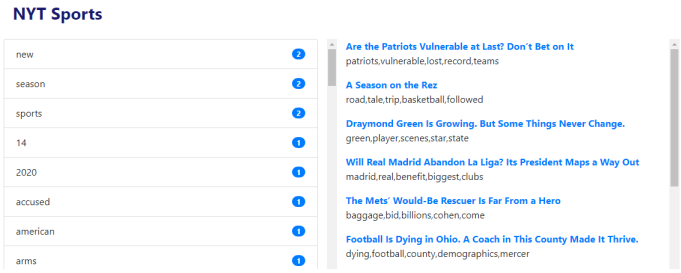

Figure 3: Listas de termos chave/artigos

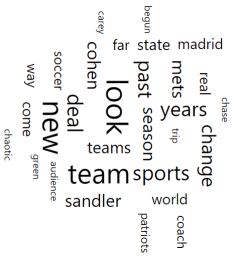

Figure 4: Word cloud

5 CONCLUSÕES

Durante o projeto, foram avaliadas diferentes abordagens para o problema de extração de termos chave de documentos: TF-IDF, BM25, BM25-F, abordagem supervisionada, graph ranking, rank aggregation. Durante a primeira parte do trabalho, foi verificado que a avaliação foi prejudicada, no sentido em que, apenas se considerava um termo chave correto se estivesse contido na lista de termos verdadeiros (strings iguais). Desse modo, tal como referido, foi incorporado a opção de apenas ser verificado se a string termo chave está contida em algum termo verdadeiro. Relativamente à análise de grafos, os diferentes tipos de cálculo dos pesos apresentam maior complexidade temporal, e não melhoram a performance de forma considerável. Adicionalmente, os priors baseados nos valores TF-IDF aumentaram ligeiramente a performance. Por fim, foi posto em prática o sistema implementado, dando uma ideia mais concisa da sua utilidade para os utilizadores.