TÉCNICO LISBOA

Processamento e Recuperação de Informação

Course Project - Part 2

G13 - T

Introdução

O objetivo geral deste trabalho consiste em extrair automaticamente as palavras mais relevantes de um determinado documento (keyphrases). Para tal, foi-nos proposto que utilizássemos diferentes técnicas, de modo a comparar quais as técnicas que oferecem uma melhor cobertura sobre os temas dos diferentes documentos.

1. An approach based on graph ranking

Neste exercício tinhamos como objetivo classficar os diferentes candidatos de um documento à nossa escolha, tendo utilizado o mesmo documento usado na primeira parte do projeto, referente ao tema Baseball.

Primeiramente procedemos à sua divisão em frases, tendo a cada uma dessas frases aplicado steming, pré-processamento (*lowercasing*, remoção de e-mails, remoção de pontuação) e retirámos stopwords (tendo utilizado a biblioteca nltk).

Para classificar os candidatos (unigramas, bigramas e trigramas) utilizámos o algoritmo PageRank, sendo que neste exercício o grafo que construímos não era pesado, ou seja, as ligações entre nós (candidatos) não apresentam peso.

2. Improving the graph-ranking method

Este exercício consiste em extender o algoritmo do exercício anterior considerando uma probabilidade não uniforme para cada candidato assim como uma pesagem não uniforme das ligações entre nós.

O dataset utilizado foi o mesmo que no segundo exercício do primeiro projeto. Neste caso, o objetivo foi utilizar diferentes funções de probabilidade e pesagem uniformes, sabendo de antemão quais as expressões relevantes de cada documento. Os documentos do dataset estão no formato XML, foi, portanto, necessário extrair as palavras que se encontravam dentro da tag word. Decidimos extrair as words em vez dos lemmas pois as keyphrases dos documentos estavam apenas disponíveis no formato word ou stem. Desta forma, ao obter as palavras extraídas dos documentos XML foi necessário aplicar as técnicas de processamento já referidas anteriormente para tornar mais simples a comparação com as keyphrases.

No que toca às funções de probabilidade (priors) testámos as sugestões referidas no enunciado, das quais: funções de probabilidade baseadas na posição do candidato no documento e no seu tamanho, assumindo que candidatos com maior comprimento e que aparecem nas primeiras frases do documento têm maior probabilidade de serem bons candidatos; função de probabilidade baseada na pontuação BM25 do candidato;

Os pesos das arestas entre candidatos foram calculados utilizando dois métodos diferentes: função de pesagem de ligações relacionada com as coocorrências entre candidatos na mesma frase (com uma janela igual a 5), ou seja, caso o candidato apareça próximo de outro candidato na mesma frase, mais peso terá essa ligação, e função de pesagem baseada em *word embeddings*, utilizando o modelo Word2Vec da biblioteca **gensim**.

Discussão de resultados (Os resultados podem ser consultados no anexo A)**:**

Em primeiro lugar, como baseline, considerámos a abordagem mais simples, ou seja, a aplicação do algoritmo sem atribuição de pesos nem priors tendo obtido uma mean average precision de 26%. Todas abordagens testadas posteriormente obtiveram melhores resultados em relação à **baseline**.

Os melhores resultados obtidos foram quando utilizámos o tamanho dos candidatos como função de probabilidade, consideramos que isto se deva ao facto das keyphrases "reais" serem maioritariamente compostas por duas ou mais palavras, ou seja o seu tamanho será provavelmente maior. Para além disso, obtivemos resultados muito semelhantes ao usar word embeddings (skipgrams) como função de pesos relativamente a usar coocorrências entre keyphrases, tal se deve ao facto dos words embeddings apenas melhorarem a performance da pesagem (uma vez que reduzem a dimensionalidade) e não a precisão.

3. An unsupervised rank aggregation approach

Neste exercício calculámos o *rank* de cada candidato de acordo com os scores das diferentes técnicas: BM25, TF-IDF, tamanho dos candidatos, número de palavras num candidato e frequência do candidato.

O método não supervisionado de combinação dos diferentes *scores* escolhido foi o Reciprocal Rank Fusion, como sugerido no enunciado.

Discussão de resultados (Os resultados podem ser consultados no anexo A)**:**

Em primeiro lugar, comparámos os resultados desta abordagem com a supervised approach desenvolvida no exercício 4 do primeiro projeto, usando o mesmo conjunto de teste e as mesmas features (Frequência do candidato, comprimento do candidato e número de n-grams). Ao contrário do esperado, não existiu uma diferença significativa entre os resultados da supervised/unsupervised approach, sendo que os melhores resultados foram obtidos com a abordagem não supervisionada. Consideramos que estes resultados estejam relacionados com facto de na abordagem não supervisionada ser dado um peso superior a candidatos com maior comprimento, o que, tal como referido no exercício 2 tem um impacto significativo na descoberta de keyphrases.

Testámos depois para todo o conjunto de ficheiros diversas abordagens, tendo obtido os melhores resultados usando BM25, número de n-grams e PageRank.

4. A practical application

Neste exercício tínhamos como objetivo apresentar, de uma forma criativa, as keyphrases mais representativas de um conjunto de notícias do jornal **The New York Times**, tendo escolhido para esse efeito notícias à cerca da Europa. Para isso utilizámos a biblioteca d3, através da qual criámos uma word cloud.

Para classificar as keyphrases mais relevantes utilizámos a abordagem apresentada no exercício 1, com a diferença que os arcos gerados representavam, não a ligação entre candidatos da mesma frase, mas sim a ligação entre os candidatos presentes no título e na descrição da mesma notícia.

Neste exercício geramos o ficheiro wordCloud.html onde é possível ver as 20 melhores keyphrases. Na representação da word cloud, palavras com maior score, terão maior dimensão. Ao fazer *hover* sobre uma keyphrase da word cloud é ainda possível visualizar os títulos das notícias em que essa keyphrase aparece, bem como o número de ocorrências da mesma.

TÉCNICO LISBOA

Anexo A – Resultados obtidos

Ex 2:

Priors			Weights		MAP
Embeddings	Co-occurrences	BM25	Len	Pos	-
X			X		0.318
X				X	0.259
X		X			0.313
	X		X		0.311
	X			X	0.255
	X	X			0.304
					0.261

Ex 3:

Supervised vs Unsupervised Approach (Using the same test set)

	TF	CandidateLen	NGram Count	MAP
Supervised	X	X	X	0.279
Unsupervised	X	X	X	0.304

Unsupervised Approach

TF	CandidateLen	NGram Count	TF-IDF	BM25	PageRank	MAP
			X	X		0.280
	X		X		X	0.280
	X		X	X	X	0.325
X	X			X		0.312
		X		X	X	0.333
X	X	X	X	X	X	0.292
	X	X			X	0.331
	X			X		0.328