Projecto

Parte 2

Processamento e Recuperação de Informação

Grupo 24

Bernardo Casaleiro	87827
Rui Oliveira	70604

9 de Dezembro de 2016

1 An approach based on graph ranking

No primeiro exercício foi implementado o algoritmo pedido de forma muito simples usando apenas a biblioteca nltk por forma a fazer download das stopwords.

O algoritmo começa por ler o ficheiro desejado, gerando em seguida um grafo onde cada nó corresponde a um n-grama com 1 <= n <= 3, ignorando as stopwords, e as arestas entre estes representam a co-ocorrência na mesma frase, sendo assim não direccionados.

O grafo gerado é então usado para calcular o ranking de acordo com o algoritmo page rank num processo iterativo com 50 iterações.

Por fim foi então aplicado o algoritmo ao ficheiro '*document.txt*' sendo os resultados:

gram dry gland	2.4984416512e-07
rat parotid gland	2.49844199144e-07
milliliters perfused	2.498442989e-07
quantity saliva	2.498442989e-07
perfused gram dry	2.50039652151e-07

Tabela 1: Resultados

2 Improving the graph-ranking method

O exercicio 2 consistiu num upgrade ao exercicio 1, atribuindo o valor Prior a cada nó e atribuir um peso a cada aresta.

Foram assim implementados como possiveis Prior o **número de ocorrências**, **posição na texto** e **tf-idf score** do n-grama, bem como para peso das arestas o **número de ocorrências** da co-ocorrência de dois n-gramas.

Após vários testes efetuados foi claro para nós que a variação com melhores resultados foi usando **tf-idf** como Prior e o **número de ocorrências** da co-ocorrência de dois n-gramas como peso das arestas. Sendo com estes valores calculados os resultados apresentados. Foi também necessário reescrever o código por forma a melhorar a performance, reduzindo significativamente o tempo de execução. No entanto, mesmo assim, de modo a aplicar o algoritmo ao dataset apresentado foram necessárias 67 horas.

Aplicando ambos os exercicios ao dataset escolhido, que se encontra em anexo, obteve-se uma precisão média de **0.0066667** para o exercicio 1 e **0,1416667** para o exercicio 2. Provando assim uma melhoria significativa nos resultados.

Encontram-se em anexo os ficheiros '*results-1.txt*' e '*results-2.txt*' com o output resultante da análise do dataset usando, respectivamente, o exercicio 1 e 2, pois apesar da precisão média ser um indicador da eficácia destes algoritmos, a nosso ver não é um bom identificador pois existem casos em que os algoritmos implementados escolhem bons ngramas ou, por exemplo, apenas uma palavra do n-grama sugerido nos ficheiros key.

3 A supervised learning-to-rank approach

O algoritmo que optamos por implementar para realizar este exercício foi o Perceptrão.

De modo a comparar a solução desenvolvida neste exercício com a desenvolvida no exercicio 2 foi usado como ficheiro de teste o mesmo dataset que no exercicio 2, usando outros dos datasets [1] para treinar o algoritmo.

Foram consideradas todos os n-gramas com 1 <= n <= 3 de cada documento após a remoção das stopwords. Sendo as propriedades consideradas para a classificação de cada n-grama:

- Se todos os termos do n-grama se encontram entre as primeiras 250 palavras do documento. Consideramos que um abstract tem cerca de 200 palavras, portanto com o limite de 250 esperamos na maior parte dos casos abranger o titulo, abstract e as primeiras frases do documento.
- O valor do n-grama de acordo com a função de classificação BM25 recorrendo a código desenvolvido na primeira parte do projecto.
- O quão o n-grama se aproxima duma frase, recorrendo à função pos_tag() da biblioteca nltk e comparando as classes gramaticais obtidas com a expressão regular {(<NN.*>+ <IN>)? <JJ>* <NN.*>+} que pretende representar uma frase.
- O valor da centralidade do n-grama no grafo gerado pelo respectivo documento.

Após o treino estar completo e obtidos os pesos das propriedades e um adicional correspondente ao bias, obtém-se o valor de cada n-grama do documento a classificar multiplicando cada propriedade pelo peso respectivo. De seguida ordenam-se os n-gramas utilizando este valor e devolve-se os 5 com melhor classificação.

Finalmente, com base nas keywords já conhecidas, calcula-se a precisão media do método para se comparar com o valor obtido no exercício anterior.

4 A practical application

Para o exercicio final foi utilizado a implementação do exercicio 2, mas utilizando a posição do n-grama na frase como valor Prior ao invés do tf-idf score desse mesmo n-grama.

O programa começa por ler o ficheiro XML/RSS e dar parse de todos os items e em seguida aplicar o algoritmo implementado no exercicio 2, sendo o resultado impresso no ficheiro '*index.html*'.

Por forma a tornar visualização dos dados mais apelativa foi utilizada a framework Bootstrap e as tabelas fornecidas pelo mesmo.

Encontra-se em anexo um ficheiro '*index.html*' gerado pela nossa implementação a titulo de exemplo.

[1] https://github.com/zelandiya/keyword-extraction-datasets