

Mestrado Integrado em Engenharia Biomédica

2º Teste de Bibliotecas Digitais

I

1. Compare o Modelo do Espaço Vectorial com o Modelo Booleano, apresentando vantagens e desvantagens.
2. Considere uma colecção constituída pelos seguintes 4 documentos:
D1: João deu um livro à Maria
D2: O João que lê um livro e tem amor por Maria
D3: O que pensa o João do amor de Maria?
D4: O João pensa que o livro é um bom presente
Suponha que a colecção é indexada usando um dicionário negativo (stop words) e um stemmer.
 - a. Utilize uma representação gráfica ou textual para o índice da colecção resultante.
 - b. Vamos focar-nos em três termos do dicionário nomeadamente livro, amor e Maria. Calcule a representação dos 4 documentos usando o tf-idf (os documentos estão normalizados usando a normalização euclidiana)
 - c. Considere a interrogação amor Maria. Como ordenaria os documentos, usando a fórmula de similaridade do cosseno.
3. Comente a seguinte afirmação: “O MEV também serve para construir aglomerados (clusters) de documentos”.

II

1. Defina relevância. Contraste com importância.
2. Dois motores de busca A e B indexam a mesma colecção. Cada motor devolve os 30 primeiros documentos para a mesma interrogação. A lista está ordenada por ordem decrescente de pontuação. As duas listas seguintes apresentam os documentos relevantes com um sinal + e os não relevantes com um sinal -. Existem 12 documentos relevantes para a interrogação na colecção.
A: + + - + - - - + - - - + - - - - - - - + - - - - - - -
- +
B: - + + - - - + - - + + - + - - - + - - + - - + - + - - -
- -
 - a. Apresente Precisão para o Rank 10 para o sistema A.
 - b. Apresente um gráfico de precisão versus cobertura para cada engenho. Explique como chegou ao resultado.
 - c. Apresente os valores MAP e R-Precision para cada um dos motores.
 - d. Qual dos engenhos apresenta maior desempenho? Justifique a sua resposta.
3. O que é mais importante para um motor de busca na Web? Precisão ou Cobertura? Explique porquê.

4. A relevância de documentos devolvidos foi avaliada por dois juízes com os seguintes resultados (+ significa relevante, - irrelevante):

Judge a: 1.+ 2.+ 3.+ 4.-

Judge b: 1.+ 2.- 3.+ 4.+

Existe uma boa concordância entre os juízes? Apresente uma medida.

III

1. Dê uma definição de PageRank e o que significa.
2. Considere o seguinte grafo de páginas Web
A página A aponta para as páginas C e D; a página B aponta para A e C;
a página C aponta para B; a página D aponta para E; e a página E aponta para A.
 - a. Calcule a matriz de adjacências subjacente
 - b. Calcule a matriz de probabilidades subjacente partindo do princípio que a probabilidade de teleporte é 0,2.
 - c. Calcule o valor aproximado do PageRank das páginas usando 3 iterações.
3. Apresente um trecho de pseudo-código para calcular a pontuação de autoridade e de hub para um conjunto base de páginas. Para tornar o seu código mais modular defina rotinas para calcular a pontuação de hub e de autoridade.
4. Compare os dois métodos para pontuação de páginas,(PageRank e Hubs+Autoridades) apresentando vantagens e desvantagens.