Mestrado Integrado em Engenharia Biomédica

Bibliotecas Digitais - Ano lectivo 2010/2011

Duração: 3 horas

Notas Importantes: Mostre trabalho suficiente em todas as respostas para lhe poder ser atribuída pontuação. As perguntas em texto normal devem ser respondidas por todos alunos. As perguntas em *itálico* só devem ser respondidas pelos *alunos que fazem o exame*. As perguntas a **negrito** apenas pelos **alunos que fazem a segunda frequência**.

I

(Cotação: 6 valores)

Perguntas para respostas curtas:

1. *Em que medida e como as stop-words e a radicalização ou lematização reduzem o tamanho do vocabulário?*
2. *Que importância pode ter a ordenação dos documentos por DOCID?*
3. *Que vantagens e desvantagens tem a técnica das biwords relativamente ao índice com informação de posição para responder a interrogações do tipo frase?*
4. *Como deve ser interpretada uma interrogação do tipo s1 /k s2 em que s1,s2 são palavras e k é um inteiro positivo?*
5. *Como especifica em bigramas a interrogação a*mal s*gem?*
6. *A que distancia de edição está a palavra inverso da palavra universo? A distancia seria mais curta se fosse possível a transposição?*
7. **Porquê que a frequência dos termos e a frequência inversa dos documentos são tão utilizados para pontuação dos documentos na Recuperação de Informação?**
8. **Que importância pode ter a normalização na pontuação dos documentos?**
9. **Porque se usa a interpolação no cálculo da precisão?**
10. **Qual a diferença entre relevância e importância duma página Web?**
11. **Porque se usa a noção de relevância marginal?**
12. **Discuta o conceito e a noção de passeio aleatório.**

II

(Cotação: 4 valores)

Considere a pequena colecção

D1: o menino vai à escola com a bola menino

D2: a menina brinca no recreio da escola recreio

D3: no recreio os meninos brincam e à bola bola

D4: No recreio há bola e gelados

Suponha que a colecção é indexada usando um dicionário negativo (*stop words*) e um lematizador.

1. **Utilize uma representação gráfica ou textual para o índice da colecção resultante. (1 valor)**
2. *Apresente o índice posicional da colecção* (1 valor).
3. Vamos focar-nos em três termos do dicionário nomeadamente escola, recreio e bola. Calcule a representação dos 3 documentos usando o tf-idf (os documentos estão normalizados usando a normalização euclidiana) (2 valores)
4. *Considere a interrogação recreio /2 bola . Que documentos devem ser devolvidos?* (0,5 valores)
5. *Apresente o índice de permuterm para uma entradas do índice (0,5 valor)*
6. **Devolva a resposta seriada para a interrogação *recreio bola*. (1 valor)**

III

(Cotação: 5 valores)

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 15 documentos relevantes para a interrogação na colecção.

A: + + – + – – – + – – – + – – – – – – – + – – – – – – – – – +
B: – + + – – – + – – + + – + – – – + – – + – – + – + – – – – –

1. Apresente precisão, cobertura e F1 ao rank 15 para os dois motores. (1 valor)
2. Apresente um gráfico de precisão interpolada versus cobertura para cada engenho. Explique como chegou ao resultado. (2 valores)
3. Apresente os valores MAP e para cada um dos motores. (0,5 valor)
4. Para alem da eficácia medida através do precisão e cobertura, que outras métricas utilizaria para avaliar um Motor de Busca Web? (0,5 valor)?

IV

(Cotação: 5 valores)

1. **Descreva sucintamente os conceitos subjacentes e a estratégia usada para atribuir uma pontuação às páginas usando o algoritmo de Hubs e Autoridades.**
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 (1 valor)

b. Calcule o valor aproximado de Hub e de Autoridade das páginas usando 3 iterações. (2 valores)

3. Utiliza certamente uma aplicação como o FB ou similar. Que utilidade pode ter um algoritmo derivado do PageRank nesse contexto?

Algumas Formulas (apenas para recordar quem já saiba....)

$$tf - idf = \log(1 + tf).\frac{N}{df}$$

M*= (1-d) (M+Z) + d K $\quad R_{i+1} = M^* \times R_i$

$$a(p) = \sum_{q \mid q,p \in E} h(q) \quad h(p) = \sum_{q \mid q,p \in E} a(q)$$

Joaquim Macedo
28-6-2011