Mestrado Integrado em Engenharia Biomédica

Bibliotecas Digitais

Frequência de Avaliação - 19/5/2011

Duração 10h -13h

Importante: Por favor , mostre trabalho suficiente para se poder determinar a sua compreensão das diversas matérias. Só assim será possível atribuir pontuação parcial em algumas respostas. Boa sorte!

1. Considere os seguintes documentos:
   Doc 1: ajuda financeira ajuda internacional fundo fundo crise
   Doc 2: Portugal ajuda critério ajuda financeira Maio financeira interesse
   Doc 3: situação ajuda monetária internacional complicada Maio
   Doc 4: Portugal situação crise Portugal financeira crise Maio

   a. Construa a matriz de incidência termo-documento para esta colecção (1 valor)
   b. Usando a matriz obtida na alínea anterior determine e justifique a resposta às interrogações: (2 valores)
      i. financeira AND NOT Portugal
      ii. (ajuda AND NOT Maio) OR Portugal
   c. Construa a matriz de ocorrência para esta colecção (1 valor)
   d. Construa a matriz de pesos tf.idf para os termos ajuda e financeira. (1,5 valores)
   e. Calcule e justifique a ordenação dos documentos para a interrogação ajuda financeira (1 valor)
   f. Apresente o índice invertido da colecção com informação de posição. (1,5 valores)
   g. Com base no índice invertido construído na aliena anterior, justifique a resposta para a interrogação crise /2 Portugal (1 valor)

2. Qual a diferença entre a lematização e o stemming? Apresente vantagens e desvantagens destes dois métodos. Sumarize os seus efeitos nos sistema de Recuperação de Informação. (2 valores)
3. Suponha que um sistema de Recuperação de Informação combina as stopwords com índices com informação de posição. Quais são os potenciais problemas e como podem ser ultrapassados? (1 valor)
4. Identifique as principais desvantagens do modelo booleano para recuperação de informação. (1,5 valores)
5. Construa o índice Permuterm para a palavra Portugal. Que entrada deste índice é adequada para a interrogação Por*gal? Justifique a sua resposta. (1,5 valores)
6. Suponha que para as entradas de dicionário da colecção apresentada na primeira pergunta construímos um índice de bigramas. Qual seria a

interrogação a realizar para a frase “crise Portugal”? Seriam devolvidos alguns falsos positivos? (1,5 valores)

7. Qual o papel da função log(x) no peso tf.idf atribuído aos termos? (1,5 valores)
8. Suponha que pretende construir um sistema de RI que suporte simultaneamente busca tolerante e interrogações com operadores de proximidade. Como pode ser construído os respectivos índices? Especifique os componentes necessários e concretize com um desenho, se necessário. (2 valores)