Universidade Católica de Angola

Licenciatura em Engenharia Informática

Bibliotecas Digitais – Exame

Importante: Por favor mostre trabalho suficiente para se poder determinar a sua compreensão das diversas matérias. Só assim poderá obter pontuação parcial em algumas respostas. Boa Sorte!

Grupo I

Para as seguintes questões diga quais são verdadeiras e falsas justificando a sua resposta de forma sucinta. (9 valores)

1. O processamento duma interrogação booleana X AND Y é mais complexa que o processamento de X OR NOT Y
2. Ao contrario da lematização, o *stemming* reduz os termos para um forma gramatical correcta
3. Os *skip pointers* não são úteis para interrogações do tipo X OR Y
4. Na correcção de erros ortográficos, os sistemas de RI calculam a distancia de edição entre o termo da interrogação e todos termos do dicionário.
5. De acordo com o ficheiro de ocorrência com informação posicional, os documentos que podem conter a frase *"to be or not to be"* são o 4 e o 5
   <***be***: 993427; *1*: 7, 18, 33, 72, 86, 231;*2*: 3, 149;*4*: 17, 191, 291, 430, 434;*5*: 363, 367, …>
6. A lei de *Zipf* especifica a distribuição dos diversos termos na colecção, relacionando a frequência de ocorrência dos termos
7. Para alem de atribuir a cada termo uma pontuação dependente da interrogação faz sentido atribuir uma pontuação constante.
8. Tal como a precisão a acurácia pode ser usada para medir a eficácia dum sistema de RI
9. Num sistema de RI normalmente a Precisão e a Cobertura[1] estão inversamente correlacionados
10. Não há nenhuma diferença entre o mecanismo de realimentação de relevância e a expansão de interrogações.

Grupo II

1. Explique como utilizaria o algoritmo MAP-REDUCE para o problema de contagem da ocorrência de palavras num conjunto de termos. Especifique a operação *map* e *reduce* para esta tarefa. Se necessário faça diagramas para ilustrar a sua resposta. (2 valores)

[1] *Recall*, no original em inglês

2. De acordo com a lei de *Heap* o número de termos (o tamanho do vocabulário) pode duma colecção pode ser estimado usando a expressão $M = kT^b$ em que T é o número de *tokens* na colecção. Suponha que 33000 é o número de termos para 25% dos documentos da colecção. Poderia estimar o número total de termos? A constante b é aproximadamente 0,5 (0,5 valores)

3. Considere a seguinte a colecção de documentos (5,5 valores)
   Doc 1: Interest in real estate speculation
   Doc 2: Interest rates and rising home costs
   Doc 3: Kids do not have an interest in banking
   Doc 4: Lower interest rates, hotter real estate market
   Doc 5: Feds' interest in raising interest rates rising

   Considere as seguintes *stop words*: an, and, do, in, not .
   Apresente sempre nas matrizes os termos por ordem lexicográfica.

   a. Construa a matriz termo-documento para ser usada pelo modelo booleano de recuperação de informação
   b. Que respostas devem ser devolvidas para as seguintes interrogações
      interest NOT rates
      (interest AND rates) NOT (rising OR kids)
      ((real AND estate) OR home) AND (interest AND rates)
   c. Construa a matriz termo-documento para ser usado pelo modelo vectorial usando a pontuação para os termos tf.idf. Na construção poderá usar várias matrizes. Indique com clareza qual é a matriz final.
   d. Como determinaria a similaridade entre os termos nos documentos?

4. Suponha que um sistema de RI devolve a seguinte resposta a uma Interrogação. Sabe-se que existem 6 documentos relevantes na colecção. Na resposta apresentada os documentos estão ordenados da esquerda para a direita e (R) indica documento relevante e (I) documento irrelevante . Justifique cuidadosamente as suas respostas. (3 valores)

   R I R I I R I I R I

   a. Qual o valor do MAP (Mean Average Precision)?
   b. Qual o valor do R-Precision ?
   c. Qual o valor da precisão aos 5?
   d. Construa a curva de Precisão versus Cobertura interpolada e não interpolada.