Organización de Datos – Curso Servetto

Evaluación Sistemas de recuperación total de textos (FTRS), 6 de noviembre del 2006

Tema 1

Para probar es necesario tener 10/16 puntos. Además no debe haber ningún error conceptual grave.

1. Considerar que cada palabra es un término y cada línea un documento.

   a) Resuelva la consulta rankeada "P y Q" para los siguientes documentos, utilizando la distancia euclidea sobre la representación TF-IDF.(3)

      D1: P Q Q Q R
      D2: P P Q
      D3: O P
      D4: Q Q Q R
      D5: O O O R

   b) Represente los documentos D1 y D4 utilizando representación vectorial booleana y calcule el índice de similitud de Jaccard.(1)
   c) Escriba la estructura del índice invertido para representar los términos de estos documentos(3)

2. Un término aparece en los siguientes documentos:

   Doc 9, Doc 12, Doc 17, Doc 23, Doc 35, Doc 39 y Doc 42

   Dicho término representa bastante bien al resto de los términos que existen en el sistema, por lo que se quiere analizar en base a él si conviene utilizar códigos unarios, delta o gamma. Calcule la forma de almacenamiento utilizando los 3 códigos y determine cual es la mejor.(3)

3. Explicar cómo puede resolverse la siguiente consulta (*ATA) utilizando léxico rotado suponiendo que se tienen los siguientes términos:

   CASA
   GATA
   GATO
   MATA
   MATO
   TASA(3)

4. Mostrar el ahorro que podría efectuarse en un índice si se utiliza Front Coding para el almacenamiento de los términos del punto anterior. Calcular el ahorro que se produciría si se usara front coding parcial de 3 en 4. (3)