Organización de Datos – Curso Servetto

Evaluación Sistemas de recuperación total de textos (FTRS), 12 de mayo del 2006

Tema 1

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

1. a) Resuelva la consulta rankeada "F o G" para los siguientes documentos, utilizando el método del coseno.(3)

   D1: F F F G I
   D2: G G H H H F
   D3: F F H I
   D4: F F F G H

   b) Represente el documento D2 utilizando el método TF-IDF. (0,5)
   c) Represente los documentos D3 y D4 utilizando representación vectorial booleana y calcule la distancia usando Jaccard.(1)
2. Un término aparece en los siguientes documentos:

   Doc 4, Doc 8, Doc 12, Doc 15, Doc 18, Doc 22, Doc 26 y Doc 29

   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. ¿Cuál de estas tres alternativas conviene?. ¿Cómo se explica que esa alternativa sea mejor que las otras dos?(3)

3. Utilizando una de las alternativas vistas en clase, explicar que cambios estructurales se utilizan para poder resolver una consulta con wildcards (sin recorrer todos los términos uno por uno) y ejemplificar mostrando como se resuelve la consulta "*ASA" para los términos:
   CASA
   CASO
   COSA(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.(3)

Organización de Datos – Curso Servetto

Evaluación Sistemas de recuperación total de textos (FTRS), 12 de mayo del 2006

Tema 2

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

1) a) Resuelva la consulta rankeada "M o P" para los siguientes documentos, utilizando el método del coseno.

    i. D1: M M M N P
    ii. D2: N N Q Q Q M
    iii. D3: M M Q P
    iv. D4: M M M N Q

    b. Represente el documento D2 utilizando el método TF-IDF.
    c. Represente los documentos D3 y D4 utilizando representación vectorial booleana y calcule la distancia usando Jaccard.

2) Un término aparece en los siguientes documentos:

    i. Doc 2, Doc 5, Doc 8, Doc 13, Doc 16, Doc 22, Doc 25 y Doc 28

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. ¿Cuál de estas tres alternativas conviene?. ¿Cómo se explica que esa alternativa sea mejor que las otras dos?

3) Utilizando una de las alternativas vistas en clase, explicar que cambios estructurales se utilizan para poder resolver una consulta con wildcards (sin recorrer todos los términos uno por uno) y ejemplificar mostrando como se resuelve la consulta "*ATA" para los términos:

    i. MATA
    ii. MASO
    iii. MOTA

4) Mostrar el ahorro que podría efectuarse en un índice si se utiliza Front Coding para el almacenamiento de los términos del punto anterior.