

Laboratorio 6

COSENO

Instrucciones

1. Utilice la matrices de pesos tf, que representan los documentos y consultas (queries) de la colección CACM guardados en el laboratorio anterior.
2. Realice un programa que para cada query, encuentre los documentos relevantes. Para encontrar los documentos relevantes, calcule el coseno entre la representación tf de cada query y los documentos de la colección CACM. Ordene los resultados obtenidos para cada query, en función del valor del coseno, de manera descendente.

Instrucciones

3. Genere un archivo con los documentos para cada consulta, cuyo coseno sea diferente de 0. El archivo deberá tener el nombre CACM_TF_rels y extensión txt. Por ejemplo para el query 001 el archivo deberá tener los documentos con coseno diferente de 0, cada uno de ellos primero tendrá un número que representará la consulta, después el número del documento y finalmente el valor del coseno obtenido (similitud) Por ejemplo:
4. 001 1220 0.4
Significa que el query 001 y el documento 1220 tienen una similitud de 0.4
5. Repita los pasos anteriores pero utilizando ahora los archivos tf_idf. El archivo deberá llamarse CACM_TFIDF_rels.txt
6. Suba su reporte, el código de su práctica y los archivos generados.