CC52D: Recuperación de Información
Tarea 2

Profesores: Ricardo Baeza Yates y Carlos Castillo

Plazo de entrega: Fecha del examen

Para la Tarea 1 usted construyó el vocabulario de una colección de texto. Complete ese trabajo agregando a ese vocabulario la información necesaria para convertirlo en un índice invertido:

- el idf de cada palabra
- la lista de documentos donde aparece cada palabra, con su tf en cada documento.
- a que parte del documento pertenece cada palabra

La segunda parte de la tarea es permitir recuperar documentos relevantes a una determinada consulta. Escriba un programa que permita ingresar una lista de palabras y recupere los 20 documentos más relevantes a ella, del más relevante al menos relevante. La relevancia sera la suma del tf*idf sobre las palabras de la consulta, agregando a su criterio más peso si la palabra pertenece a campos específicos del documento (título, etc.)

Se premiará la eficiencia de la implementación: en primer término, velocidad de respuesta a la consulta; en segundo término, velocidad de indexación. Se probará con la colección de Fibrosis Cística ya usada en la primera tarea, la cual DEBE ser usada en formato XML.

Se espera una velocidad de indexación de no menos de 1 Mb por minuto. Asimismo, una consulta de 10 palabras no debería requerir más de unos pocos segundos. Se debe proveer una interfaz Web para su programa, lo que permitirá probarlo realizando consultas en la colección.