Quinta práctica de laboratorio de Sistemas de Información para la Web

Motivación

Como ha quedado claro en varias de las prácticas anteriores no es viable comparar cada consulta recibida por un buscador con todos los documentos de una colección. Para acelerar el proceso de matching entre consultas y documentos se empleará un índice (o fichero invertido). Se trata de una estructura de datos que permite encontrar un número limitado de documentos que contienen algún término de una consulta dada; será sobre ese subconjunto limitado de documentos sobre los que se determinará la similitud con la consulta. En la práctica de hoy se trata de construir ese índice, en la próxima se procederá a resolver consultas usándolo.

Descripción del ejercicio y del entregable

Consultar y estudiar las transparencias disponibles aquí, desde *"El fichero invertido"* hasta *"¿Cómo de grande es el fichero invertido?"*

El objetivo del ejercicio de esta semana es crear un script en Python que permita indexar una colección. Esto es, crear, a partir de los documentos de texto plano almacenados en un directorio, un índice (que se almacenará en disco pero que debe poder cargarse completamente en memoria RAM) con la siguiente información:

- Un diccionario de términos que tendrán asociada la siguiente información: el valor *IDF* del término en la colección, así como la correspondiente *post-list*.
- La *post-list* de cada término consistente en una lista (o un diccionario) que contendrá los identificadores de los documentos que contienen el término así como el valor *TF* del término en cada uno de los documentos.

Queda a elección del estudiante cómo se extraerán los términos de los documentos pero se valorará positivamente el uso de tokenizadores y estematizadores, así como el paso a minúsculas de todos los términos.

Para la entrega se usará la colección Cranfield empleada en la tercera práctica (disponible en el este archivo).

Se subirá al campus virtual un archivo comprimido que contendrá: el script para indexar la colección, un documento explicando las decisiones tomadas para su implementación, además del índice correspondiente a la colección junto con un script que demuestre cómo se puede cargar en memoria RAM y acceder a algunos de sus contenidos.

¡Atención! En este ejercicio no es preciso resolver consultas, ése será el objetivo de la siguiente práctica.