Sexta práctica de laboratorio de Sistemas de Información para la Web

Motivación

En la sesión anterior se desarrolló código para indexar una colección así como código para poder cargar dicho índice en memoria RAM y acceder a sus contenidos. Hoy se verá cómo acceder a dicho índice para resolver de la manera más eficiente posible consultas textuales.

Descripción del ejercicio y del entregable

Desarrollar un script que demuestre la resolución de consultas mediante el índice. En su versión más sencilla el script recibirá una consulta como parámetro, cargará en RAM el índice, resolverá la consulta con el mismo y retornará los resultados de mayor a menor relevancia.

En una versión ideal el índice se cargará en RAM una sola vez y se ofrecerá un servicio web para la resolución de las consultas, el script que reciba los resultados, además, buscará los documentos en la colección para mostrar un fragmento de los mismos.

Para resolver las consultas se procederá como sigue:

1. La consulta se tokenizará en términos (se aplicará stemming únicamente si también se hizo al indexar la colección) y se pasará a minúsculas.
2. Para cada término de la consulta se obtendrá la lista de documentos que lo contienen.
3. Se calculará la similitud del coseno entre la consulta y cada uno de los documentos candidato.

Para resolver el punto 3 puede procederse de dos formas:

a. Puede adaptarse el código de la tercera práctica y compararse realmente la consulta con el contenido textual de cada documento.
b. Puede usarse únicamente la información recuperada del índice y, en consecuencia, explotar no solo la presencia de los términos sino también los valores TF e IDF de los mismos. En este caso los documentos estarían representados por vectores que incluirían, como máximo, todos los términos de la consulta.[1]

Se entregará en el campus virtual un archivo comprimido que contendrá:

- La versión actualizada del script para indexar la colección.

[1] TF-IDF and cosine similarity

- El código necesario para resolver consultas.
- Una selección de 5 consultas para la colección Cranfield con la lista de documentos resultantes.