Proyecto del bloque de Recuperación de información:

Diseño y desarrollo de un sistema de recuperación de información

El objetivo principal del proyecto es la implementación de un sistema de recuperación de información básico basado en el modelo Espacio Vectorial y empleando el esquema TF-IDF para la ponderación de los términos.

Para ello, el programa tendrá que indexar un conjunto de documentos. Posteriormente, se permitirá la realización de consultas de modo que el sistema devuelva los documentos más relevantes de acuerdo con la consulta introducida por el usuario.

Por lo tanto, el sistema estará formado por dos bloques: indexación y búsqueda.

En la indexación, se tienen realizar los siguientes pasos:

- Preprocesar cada uno de los ficheros a indexar. Para ello tendremos que aplicar: a) filtros de caracteres que eliminarán aquellos caracteres que no necesitamos (. , ¿ ? ¡, ¡ =), b) filtros de palabras como por ejemplo el filtro de palabras vacías, stemming (opcional), filtro de palabras por debajo de un tamaño determinado (opcional). La entrada este proceso será el texto contenido en el documento. Finalmente, la salida será la lista de los términos de indexación del documento.
- Calcular el TF de los términos de los documentos. Cada término en un documento tendrá un peso específico. Recordar que el peso de un término se calcula utilizando el TF-IDF, por lo que el peso final del término en el documento sólo lo podremos obtener cuando tengamos el IDF.
- Construir un índice invertido en el que para cada término contenga su valor IDF (el IDF es el peso del término en el corpus), y la lista de documentos en los que aparece ese término. Para simplificar el proceso de búsqueda en el otro bloque, lo ideal es que junto al documento se guarde el peso (TF-IDF) de dicho término en ese documento. De forma esquemática, el índice podría diseñarse como sigue:

  Término 1 IDF (doc1-peso doc2-peso doc3-peso)

  Término 2 IDF (doc2-peso doc4-peso)

- Para cada documento calcular su longitud, entendida como el módulo del vector formado por los términos del documento. Es decir, $\sqrt{\sum w^2}$.

De forma resumida, la entrada del proceso de indexación será el conjunto de ficheros a indexar. Por otro lado, la salida será el índice invertido y el peso de cada documento. Ambas salidas deberían guardarse en un fichero propio.

Finalmente, en el proceso de búsqueda se tendrían que realizar los siguientes pasos:

- Cargar en memoria tanto el índice invertido, como la longitud de los documentos de la colección.
- Pedir al usuario que introduzca una consulta.
- Aplicar exactamente la misma secuencia de preprocesado a la consulta.
- Recuperar los documentos que contienen los términos de la consulta.
- Calcular el ranking de cada documento.
- Devolver al usuario el listado de documentos ordenados por ranking.