Informe de la quinta práctica

Ángel García Menéndez

Universidad de Oviedo

Introducción

El script implementa el concepto de fichero invertido, esto es, un índice en el que las claves son los propios términos. Asimismo, se ofrece un script auxiliar para demostrar su correcto funcionamiento. Para ver la ayuda puede emplearse la orden *python test.py -h*. Un posible listado de términos que están presentes en el índice son:

- heat
- flow
- dimension
- linear

Se puede emplear también términos no presentes para comprobar que el error menester aparece.

Estructura

Se han implementado 2 clases:

- InvertedIndex: el índice propiamente dicho.
- IndexEntry: cada una de las entradas del índice.

La primera funciona mediante un diccionario donde las claves son las cadenas textuales de cada término, y los calores la entrada correspondiente.

Las entradas por su parte contienen 3 elementos:

- El término en sí.
- El IDF del término.
- La post_list del término.

Cada una de las clases tienen sus propios métodos y propiedades, las cuales se encuentran suficientemente documentadas.

Consideraciones

Se han tomado algunas decisiones de implementación que merecen mención.

Por un lado, en caso de que se trate de acceder a un término que no esté contenido en el índice, se lanzará una excepción.

El IDF no se calcula durante la creación del índice. Este es calculado y almacenado la primera vez que se solicita para un término concreto. El valor de IDF en el constructor de las entradas es por defecto 0.

Finalmente, es necesario proporcionarle a las entradas el número de documentos de la colección. La razón de esto es la necesidad de este dato para el cálculo del idf de cada término. No es la solución más elegante, pero se asume que al crear el índice se aporta toda la colección, y de verse esta ampliada, debería crearse el índice de nuevo.

Finalmente, es necesario proporcionarle a las entradas el número de documentos de la colección. La razón de esto es la necesidad de este dato para el cálculo del idf de cada término. No es la solución más elegante, pero se asume que al crear el índice se aporta toda la colección, y de verse esta ampliada, debería crearse el índice de nuevo.