PRACTICA 4: INDEXACIÓN Y BÚSQUEDA EN FICHEROS CON LUCENE

Diego Santolaya Martínez
Abel Chils Trabanco
Alexandru Oarga Hategan

1. Introducción

En esta práctica se realizan las dos aplicaciones utilizando la librería Lucene tras haber analizado adecuadamente los ejemplos proporcionados junto al enunciado y también se ha integrado esta librería en la página web desarrollada en sesiones de prácticas anteriores.

2. Sección Principal

2.1. Bloque 1 (Introducción a la librería Lucene)

Después de modificar el código de ejemplo y ejecutar nuevamente las consultas, se puede comprobar cómo mediante un analizador de tipo StandardAnalyzer todas las consultas excepto “of” reciben puntuaciones superiores o iguales respecto a las mismas consultas ejecutadas mediante un analizador de tipo SimpleAnalyzer. Esto es debido a que el analizador de tipo StandardAnalyzer elimina las stopwords (palabras sin significado como artículos, determinantes ...) del inglés. Como para obtener la puntuación se tiene en cuenta el número de términos totales, y en el caso del StandardAnalyzer se reduce, la puntuación es mayor. Por otro lado,“of” al ser una stopword, no cuenta sus coincidencias, por ello obtiene una puntuación de 0 coincidencias.

2.2. Indexación y búsqueda de documentos de un determinado directorio

Para implementar la primera aplicación, primeramente se crea una clase llamada Indexador, la cual se encarga de la gestión del índice. Por otro lado se crea la clase Apartado22_CreadorIndice la cual se encargara de la interacción con el usuario.
En esta se implementa una función auxiliar, que dado un directorio devuelve una lista de todos los ficheros contenidos en este. Esta función se invoca cuando el usuario proporciona por entrada estándar un directorio y su resultado serán los ficheros a indexar. Para indexar los ficheros se usa un indice en disco que se guardará en un directorio llamado indice_disco en el directorio desde el que se ejecute la aplicación. Por otro lado se utiliza un indexador de tipo SpanishAnalyzer, el cual elimina las stopwords del castellano.
Para la segunda aplicación, se crea una clase auxiliar llamada Buscador, la cual se encargara de realizar la búsqueda sobre el indice creado con la aplicación anterior. Por otro lado se ha creado una clase llamada Apartado22_Buscador, la cual se encarga de interactuar con el usuario.

2.3 Búsqueda con palabras clave

Se ha incluido la busqueda con restricciones en la Web desarrollada hasta ahora. Para ello se ha indexado la base de datos. Esto se ha realizado obteniendo con una query todos los datos de los libros de la base de datos y añadiendo los campos uno a uno. Para escribir los campos se ha utilizado la clase IndexWriter. Para cada libro, se ha creado un Document con los campos: isbn, título, precio, descripción, idioma y

país (palabras clave). Se crea una clase Field para cada campo con su valor y la clase IndexWriter se encarga de escribir el Document en el índice.
Para la búsqueda se ha utilizado un SpanishAnalizer y la clase IndexSearcher. La búsqueda según palabras claves se ha hecho comprobando antes de la búsqueda si los términos introducidos coinciden con una de las palabras clave. En caso de ser así, en la query se incluye el campo antes del termino introducido para limitar la búsqueda a dicho campo. Por ejemplo, la query resultante de introducir "precio 15" seria "precio:15".
Por último, se ha añadido la funcionalidad a la Web. Para cada uno de los hits obtenidos, se consulta la información del libro y se muestra en la página de resultado de búsquedas.