PRACTICA 3: DESARROLLO DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN CON LUCENE

Manuel Blanco Rienda

GESTIÓN DE INFORMACIÓN EN LA WEB (2016-2017)

Memoria de la práctica 3 de GIW por Manuel Blanco Rienda

¿Qué se ha hecho?

La práctica tres de la asignatura Gestión de Información Web consta de dos partes diferenciadas que han sido desarrolladas de forma separa cada una. Por un lado tenemos un indexador de documentos de texto plano y por otro, un buscador de esos documentos en función de un texto especificado. Cada uno de los programas funciona independientemente, aunque el buscador requiere que previamente se haya creado un índice en el que buscar.

¿Cómo se ha hecho?

El proyecto se ha desarrollado usando Lucene como biblioteca base y Netbeans como IDE de desarrollo. Cada una de las partes que tal y como se ha mencionado arriba, componen la práctica, ha sido desarrollada de forma independiente, por lo que a la hora de explicar dicho proceso es necesario dividirlo en dos secciones:

Indexador

El indexador tiene que recibir desde el inicio los parámetros: ruta del archivo de palabras vacías, ruta donde almacenar el índice a crear y ruta donde se encuentran los documentos a indexar. Teniendo estos datos, hace uso del analizador de español de Lucene que se menciona en el guión de prácticas, el cuál utiliza el fichero de palabras vacías para analizar cada documento que se indexará más tarde. A partir de la información de los directorios que se le especifican de inicio al programa, leemos primero los documentos a indexar con la siguiente heurística:

- Teniendo la ruta del directorio donde se encuentran los documentos, se itera a través de cada uno de ellos, leyéndolos como string (con las pertinentes configuraciones para obtener los datos en codificación UTF-8).
- Los string se modifican para que tengan un formato adecuado de cara a crear objetos tipo DOM a partir de ellos, que es lo siguiente que se hace.
- Se obtiene la lista de títulos y textos de los objetos DOM creados y con cada par de estos elementos se crea un objeto “Noticia” (diseñado para almacenar estos dos campos) y se añade a una colección que los almacene.

Tras ello, se procede a pasar la información de los objetos noticia a documentos con la información clasificada en función de los campos: “titulo” y “texto” y se escriben dichos documentos con el escritor de índices creado anteriormente, en el directorio que le especificamos inicialmente al programa.

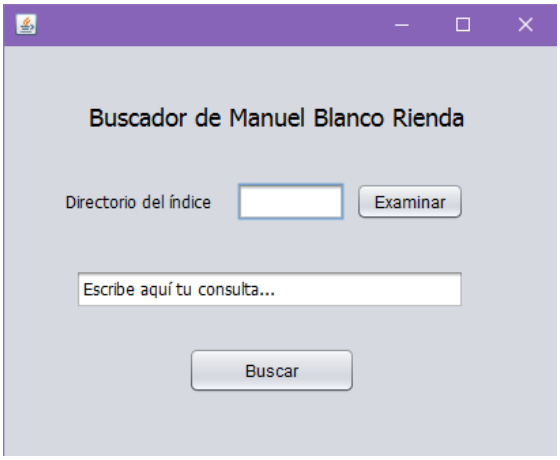
Buscador

El buscador consta de varias partes diferenciadas dentro de su código: Tres ventanas de interfaz gráfica y una clase controladora que realiza las funciones de búsqueda dentro del directorio de índice que especifiquemos. Las ventanas de interfaz gráfica se dividen en:

- Ventana principal del buscador.
- Ventana de búsqueda del directorio del índice.
- Ventana de muestra de resultados de búsqueda.

Ventana principal

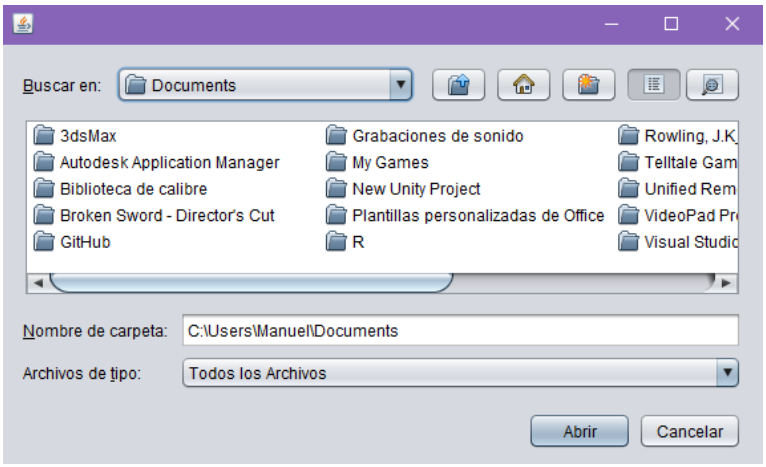
Es la que contiene la lógica gráfica principal: dar opción a especificar el directorio donde se encuentra el índice y determinar el texto a partir del cual buscar documentos. A partir de esta pantalla se llaman a otras dos: la de búsqueda de directorio índice y la de muestra de resultados. El aspecto gráfico de esta ventana es el que sigue:



Si el directorio que especificamos no contiene un índice, no se realiza el proceso de búsqueda y por tanto no se abrirá la ventana de muestra de resultados. Una vez que tiene todos los datos que son requeridos en la interfaz, los pasa al controlador de búsqueda.

Ventana de búsqueda del directorio de índice

Se trata de una ventana cuya funcionalidad es recoger de forma gráfica un directorio que le especificamos. Una vez ha sido determinado, lo manda a la ventana principal, para que se muestre en el recuadro destinado para ello (al lado del botón “examinar”). El aspecto gráfico de esta ventana es el siguiente:



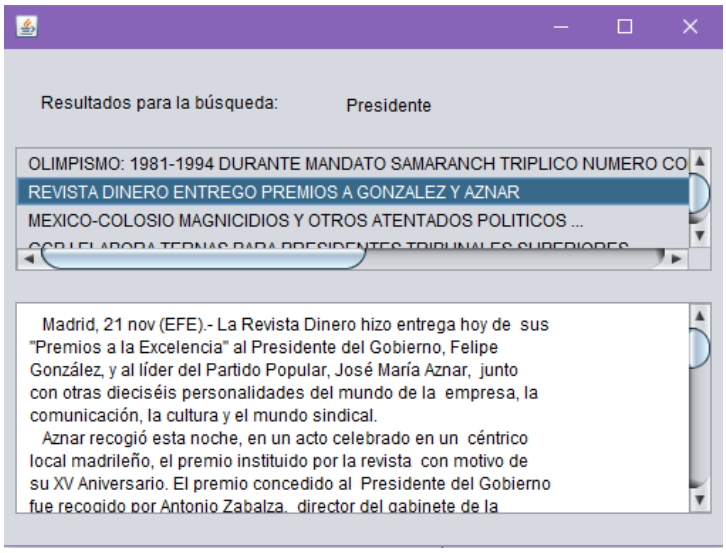
Controlador de búsqueda

Si el directorio de almacenamiento del índice es correcto y verdaderamente almacena un índice, se procede a buscar el texto especificado. Para ello, al igual que en la indexación se crea un analizador de español con las palabras vacías (que deben haber sido leídas de su archivo previamente). Se abre el directorio que se ha especificado en la ventana de búsqueda de índice y se crea un objeto buscador con él. Además de ello, se utiliza un "parseador" para obtener la consulta que vamos a realizar al índice en función de un campo determinado.

El texto que contiene la palabra clave que queremos buscar es sometido a "steamming" (para obtener documentos con palabras derivadas de la raíz de la palabra) y comienza el proceso de búsqueda en el índice a partir de él. Con cada coincidencia, se obtiene el documento completo con el titular de la noticia y su texto correspondiente, el cual es almacenado en una colección que será enviada a la ventana de muestra de resultados, junto al término que originó la búsqueda.

Ventana de muestra de resultados

Una vez que se recibe el conjunto de documentos que han coincidido con el término buscado, se procede a mostrar la información que contienen de la siguiente forma: En el recuadro superior se muestra una lista con los títulos de todas las noticias de todos estos documentos. Cuando se selecciona alguno de estos títulos, en el recuadro de abajo puede observarse el contenido de la noticia. El aspecto gráfico de la ventana de muestra de resultados es el siguiente:



En caso de querer realizar una consulta después de otra, se hace en la ventana de búsqueda y obtendremos otra ventana de resultados paralela a la primera que conseguimos, con sus propios documentos. Esta ventana puede ser vista como un elemento innovador, ya que es muy cómodo poder visualizar los documentos buscados de una forma tan directa y sencilla, a la par de que se pueden tener varias de estas ventanas abiertas concurrentemente.

Manual de usuario

Indexador

Para arrancar el indexador es preciso hacerlo a través de una terminal, haciendo uso de java (versión superior a la 1.8) con el siguiente comando: “java -Xmx1024m -jar ruta_de_localizacion\Indexador.jar ruta_fichero_palabras_vacias ruta_alojamiento_indice/ ruta_directorio_documentos/”. Si los parámetros se han proporcionado de forma correcta obtendremos la siguiente salida (habiendo un espacio de tiempo entre que comienza y finaliza el proceso de indexación: un minuto y cincuenta segundos en un core i7-3553U):

```
C:\Users\Manuel\Desktop>java -Xmx1024m -jar .\Indexador.jar palabras_vacias_utf8
.txt index/ efe/
Comienza el proceso de indexación...
¡Proceso de indexación completado!

C:\Users\Manuel\Desktop>_
```

Buscador

Para arrancar el buscador es preciso hacerlo a través de una terminal, haciendo uso de java (versión superior a la 1.8) con el siguiente comando: “java -Xmx1024m -jar ruta_de_localizacion\Buscador.jar ruta_fichero_palabras_vacias”. Inmediatamente aparecerá la ventana gráfica del buscador a la cual deberemos proporcionar: la ruta

del directorio del índice que creamos con el indexador (paso anterior) y un texto a partir del cual realizar la búsqueda. Si pulsamos “Buscar” y el directorio es el correcto, el programa se quedará procesando unos cinco segundos, tras los cuales mostrará la pantalla de muestra de resultados. En esta pantalla podremos ver: el texto por el cual se han buscado los documentos, el titular de las noticias de los documentos encontrados y un recuadro cuyo contenido se rellenará con el texto de una noticia cuyo titular esté seleccionado.

NOTA: Para probar el indexador con la colección de documentos de la agencia efe es necesario tener en cuenta que hay que eliminar el archivo “efe.dtd” de la carpeta de almacenamiento de los documentos a indexar, dado que no es un documento a tener en cuenta y el programa recorre todos los elementos de la carpeta suponiendo que son documentos SGML.