

## Práctica 1: Desarrollo de un Sistema de Recuperación de Información con Lucene

---

Manuel Orantes Taboada

manuelorantes96@gmail.com

morantes96@correo.ugr.es

77150692K

25 de marzo de 2021

## Índice

<b>1</b>	<b>Descripción</b>	<b>3</b>
1.1	Indexador . . . . .	3
1.2	Buscador . . . . .	5
<b>2</b>	<b>Bibliografía consultada</b>	<b>7</b>
<b>3</b>	<b>Enlace al fichero con los documentos</b>	<b>7</b>

# 1. Descripción

En la práctica, el objetivo era realizar un sistema de recuperación con Lucene. Para ello se han desarrollado dos aplicaciones. La primera aplicación se trata del indexador, que crea a partir de una colección de documentos un conjunto de índices que permitan más tarde realizar una búsqueda a través de ellos. El segundo programa es un buscador, que recibe como argumento el conjunto de índices creado anteriormente. Además, es una aplicación con interfaz gráfica, por lo que no es necesario pasar por terminal la búsqueda a realizar.

En primer lugar, comentar los archivos que elegí como documentos de mi sistema. Es un conjunto de 157 documentos obtenidos de los enlaces que se disponían desde la práctica. La colección es: <https://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html> - Parte1 - Premios 1994 - Awd\_1994\_61 . La colección hace referencia a abstract premiados por distintos tipos de búsquedas (búsqueda básica, bolsa de palabras, búsqueda indexada...).

## 1.1. Indexador

Para realizar la primera aplicación, he seguido tanto los pasos del esbozo de la práctica como unos tutoriales por internet, que al fin y al cabo, solo distan entre sí del hecho de que unos son más actuales y con versiones más recientes de lucene. Los pasos a seguir fueron:

- Crear un filtro para los archivos. El conjunto de los documentos que elegí, tiene dentro de este un archivo .html con la dirección de cada uno de ellos, y esto no resultaba interesante para el proyecto. Con el filtro, solo proceden a ser indexados los archivos .txt .
- Uso de constantes globales para definir donde se encuentran los archivos.
- Crear el indexador en sí, y para ello:
  - Creamos un indexador con un Analizador al que previamente le hemos introducido nuestra serie de palabras vacías.
  - Abrimos el fichero en donde están los documentos.
  - Pasamos el indexador documento por documento.

A continuación vemos una ejecución por terminal del indexador junto a los archivos que genera:

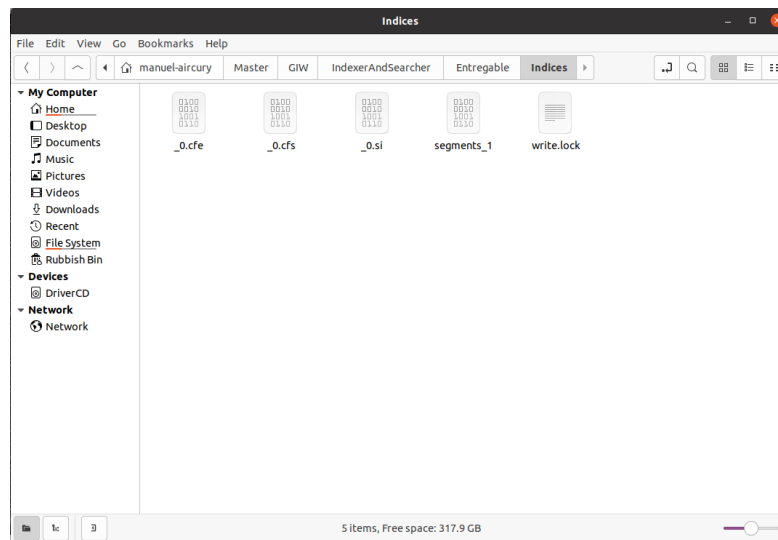
Para usarlo, simplemente hay que ejecutar el .jar junto con un argumento, que será el directorio en el cual están los documentos:

```
java -jar Indexer.jar ../awd_1994_61
```

En este caso, mi directorio estaba en la carpeta previa a donde estaba el .jar . Una vez se ejecuta, podemos ver como empieza a cargar los ficheros del directorio y comienza a indexarlos uno a uno:

```
manuel-aircury@manuel-aircury ~/Master/GIW/IndexerAndSearcher/Entregable$ java -jar Indexer.jar ../awd_1994_61
Cargando ficheros de: ../awd_1994_61
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461668.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461664.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461243.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461541.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461708.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461625.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461031.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461859.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461528.txt
Indexing /home/manuel-aircury/Master/GIW/IndexerAndSearcher/awd_1994_61/a9461554.txt
```

Como resultado, obtenemos en el directorio en el que se encuentra nuestro .jar otra carpeta llamada Indices, que muestra el siguiente contenido:



Ya estarían los índices preparados para que más tarde el buscador los utilice.

Hay que subrayar que intenté un visualizado de estos índices mediante Luke pero me salía que las versiones no eran compatibles, así que no he podido adjuntar imágenes de estos índices.

## 1.2. Buscador

Para la segunda aplicación, al igual que en la primera, seguí los pasos tanto de la práctica como de otros tutoriales. Esta vez, los pasos a seguir son los siguientes:

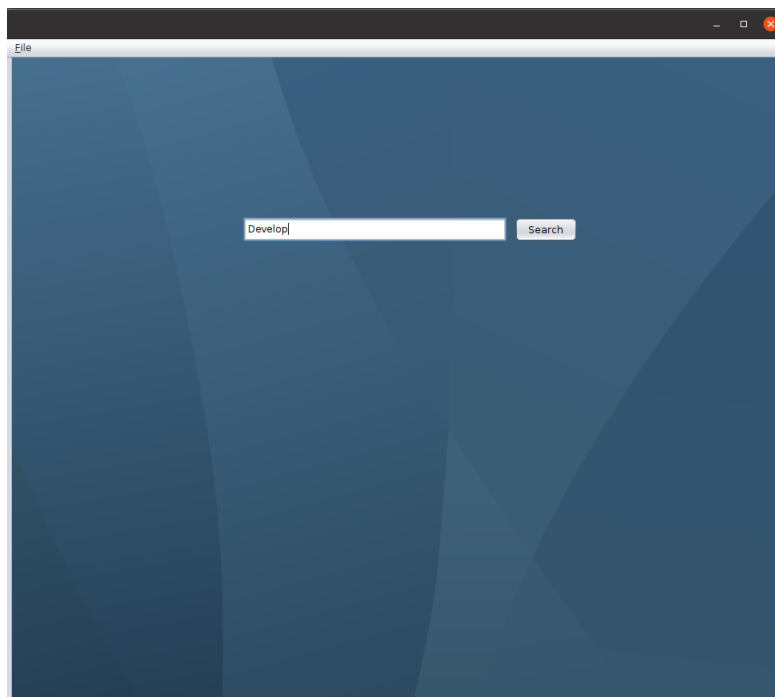
- Obtener un buscador a través de índices con Lucene.
- Aplicar las medidas pertinentes a la búsqueda introducida.
- Obtener los resultados del buscador, que serán los archivos encontrados.

Además, también se vieron diferentes vídeos para recordar como era crear una aplicación con Java, para poder buscar por ella, obtener los resultados y poder abrir los ficheros.

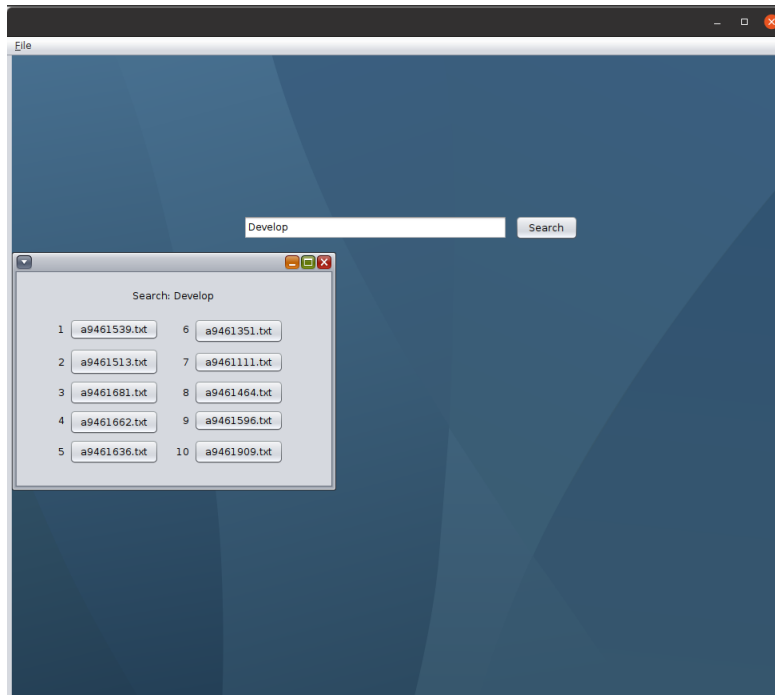
Para usar la aplicación, al igual que antes, ejecutaremos el .jar desde la línea de comandos junto con el conjunto de índices a usar. En este caso, usaremos los índices que hemos obtenido a través de la primera aplicación:

```
java -jar Searcher.jar Indices
```

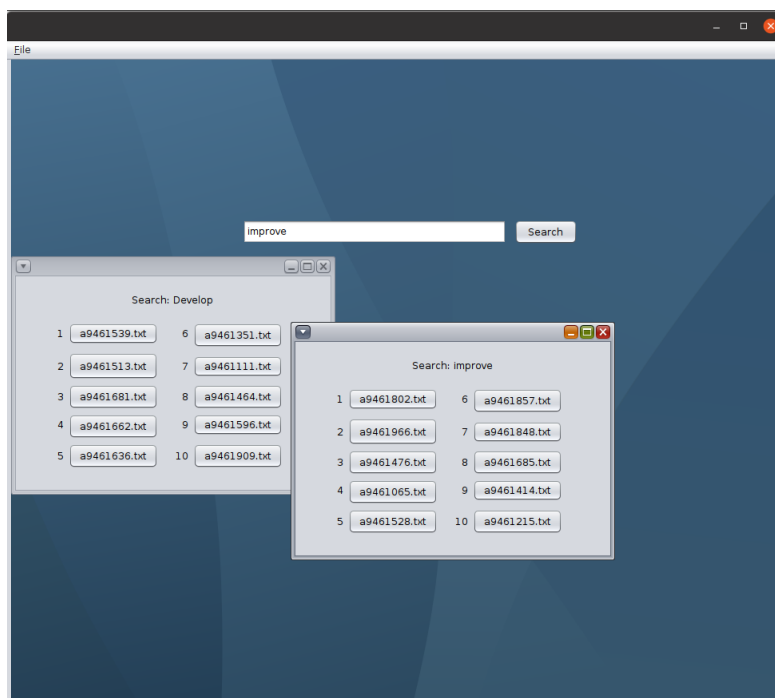
Esto nos abrirá una interfaz, en la que encontramos un campo de texto en el cual podemos escribir junto con un botón para realizar la búsqueda:



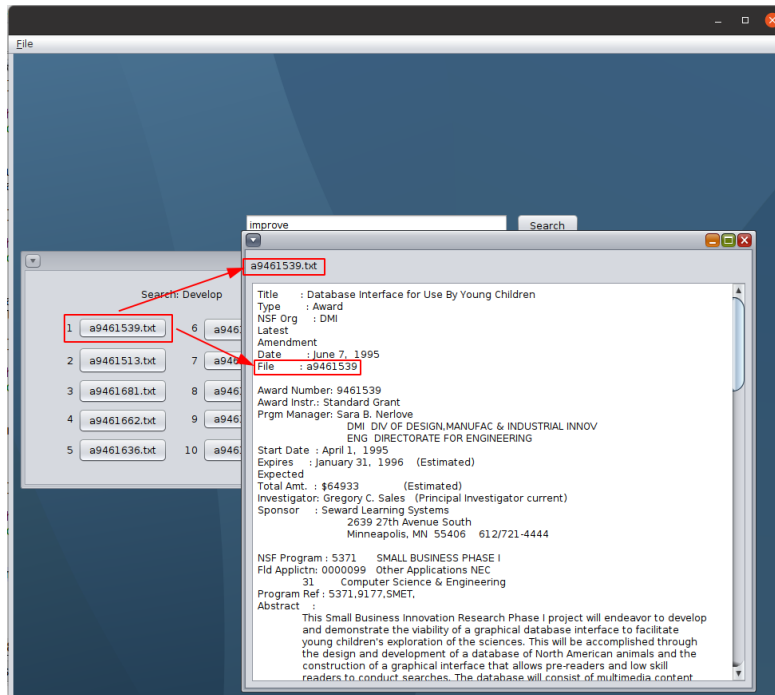
Cuando le demos al botón de buscar, se nos abrirá otra pestaña en la cual encontraremos 10 botones junto con un número del 1 al 10, indicando dicho número qué documento se adecua más a nuestra búsqueda:



Además, podemos realizar varias búsquedas sin cerrar esta pestaña, y podremos diferenciarlas gracias a un etiqueta que nos indica a qué búsqueda pertenecen los archivos:



Por último, podemos acceder al documento deseado pinchando en el botón de su nombre:



Y con esto, llegamos al fin de la práctica realizada.

## 2. Bibliografía consultada

Además de basarme en la propia práctica, he usados estos dos tutoriales:

<https://www.ionos.es/digitalguide/servidores/configuracion/apache-lucene/>

[https://www.tutorialspoint.com/lucene/lucene\\_sorting.htm](https://www.tutorialspoint.com/lucene/lucene_sorting.htm)

[https://lucene.apache.org/core/3\\_5\\_0/api/core/org/apache/lucene/search/IndexSearcher.html](https://lucene.apache.org/core/3_5_0/api/core/org/apache/lucene/search/IndexSearcher.html)

<https://www.jitendrazaa.com/blog/java/generating-executable-jar-file-with-all-dependencies/>

## 3. Enlace al fichero con los documentos

<https://drive.google.com/drive/folders/1S91bNau6TiDt-Xv2UnfUlgE94jQ3FzTh?usp=sharing>