PRÁCTICA 5: BÚSQUEDA DE INFORMACIÓN MEDIANTE LA LIBRERÍA LUCENE

Sistemas de Información

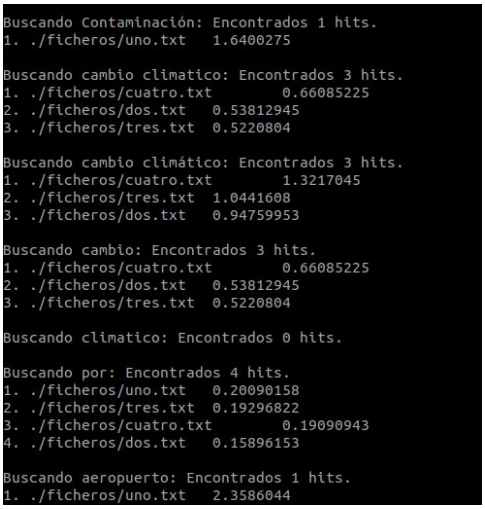
Curso 2019-2020

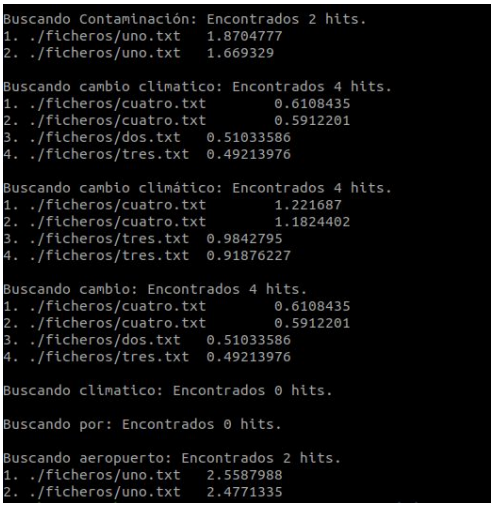
Pedro Tamargo Allué - 758267

Juan José Tambo Tambo - 755742

Raúl Rustarazo Carmona - 715657

1. ¿Qué pasa si utilizamos el “StandardAnalyzer” en lugar del “SimpleAnalizer”? ¿Qué función tiene el fichero “stopwords.txt”?

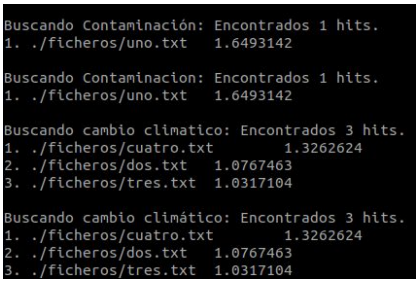


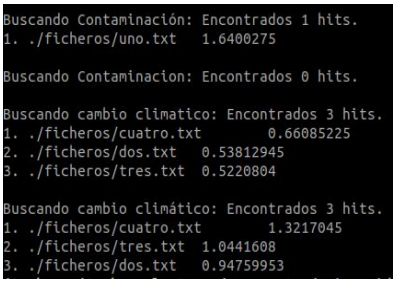


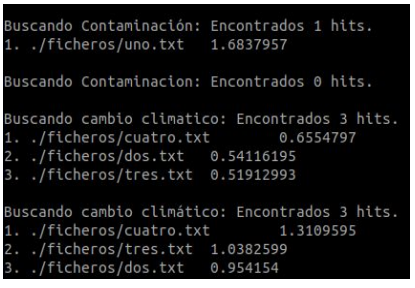
Resultado utilizando SimpleAnalizer Resultado StandardAnalizer

Usando StandardAnalizer, por defecto la búsqueda se realiza eliminando todas palabras “stopword” del idioma inglés. Opcionalmente se puede indicar un fichero donde se almacenan los “stopwords” deseados. En este caso, utilizamos el fichero stopwords.txt.
Podemos observar que utilizando el StandarAnalizer la palabra “por” no se encuentra ya que es una de las “stopwords” indicadas en “stopwrods.txt”.

2. Qué ocurre si en la búsqueda ponemos “contaminacion” o “cambio climatico” (sin tildes)?







SpanishAnalyzer SimpleAnalyzer StandardAnalyzer(stopwords.txt)

Se puede observar que los resultados son iguales excepto en “contaminacion”, la cual sólo es encontrada con SpanishAnalyzer. Esto es debido a que el Spanish Analyzer es capaz de encontrar palabras sin tilde aunque en el texto sí que la lleven, debido a la codificación. En cuanto “cambio climatico” sí que encuentra hits con la

palabra sin tilde pero con una puntuación menor. Esto se debe a que es una palabra compuesta, ya que la palabra “cambio” sí que es capaz de encontrarla.

3.¿Qué ocurre si re-indexamos todos los ficheros cada vez que ejecutamos el programa, en lugar de, simplemente, reabrir el índice creado previamente?

```
Buscando Contaminación: Encontrados 1 hits.
1. ./ficheros/uno.txt   1.6837957

Buscando Contaminacion: Encontrados 0 hits.

Buscando cambio climatico: Encontrados 3 hits.
1. ./ficheros/cuatro.txt        0.6554797
2. ./ficheros/dos.txt   0.54116195
3. ./ficheros/tres.txt  0.51912993

Buscando cambio climático: Encontrados 3 hits.
1. ./ficheros/cuatro.txt        1.3109595
2. ./ficheros/tres.txt  1.0382599
3. ./ficheros/dos.txt   0.954154
```

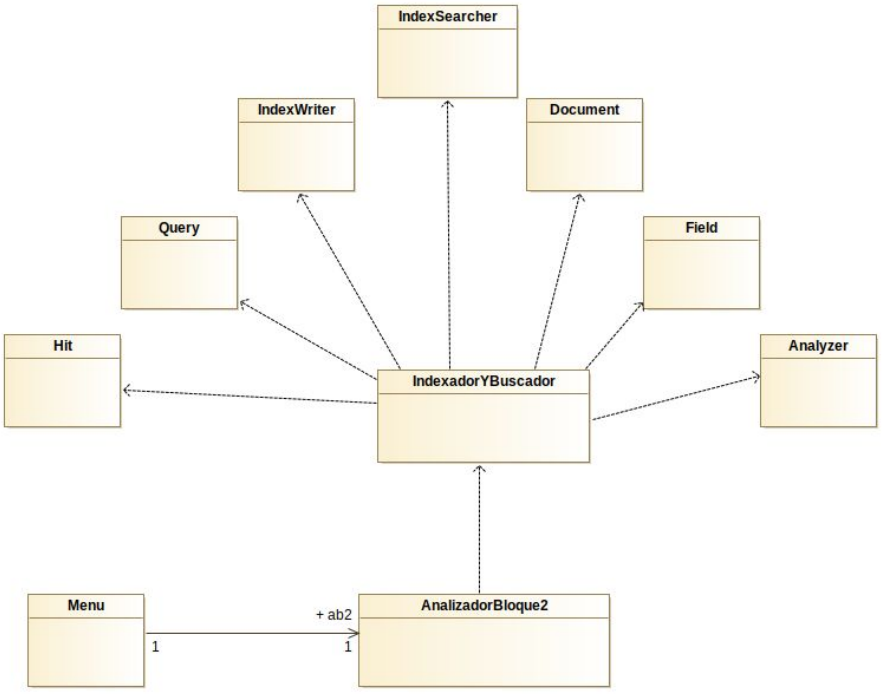
Re-indexando

```
Buscando Contaminación: Encontrados 2 hits.
1. ./ficheros/uno.txt   1.7914282
2. ./ficheros/uno.txt   1.7914282

Buscando Contaminacion: Encontrados 0 hits.

Buscando cambio climatico: Encontrados 4 hits.
1. ./ficheros/cuatro.txt        0.59804535
2. ./ficheros/cuatro.txt        0.59804535
3. ./ficheros/dos.txt   0.49374434
4. ./ficheros/dos.txt   0.49374434

Buscando cambio climático: Encontrados 4 hits.
1. ./ficheros/cuatro.txt        1.1960907
2. ./ficheros/cuatro.txt        1.1960907
3. ./ficheros/tres.txt  0.9472856
4. ./ficheros/tres.txt  0.9472856
```

Reabriendo el índice creado

Si Re-indexamos cada vez que ejecutamos, las búsquedas consiguen menos hits que reutilizando los índices. Esto se debe a que los índices son auto-incrementales, es decir, que una vez han sido creados, se pueden añadir documentos susceptibles a ser indexados. Por ello, reutilizando los índices se encuentran más resultados en otros documentos.

Diagrama de clases



Metodología de trabajo

Para la realización de esta práctica, se dividió el trabajo entre los integrantes del grupo de la siguiente manera:

- **Raúl**: Realización de pruebas.
- **Pedro y Juanjo**: Modificación del código, desarrollo del apartado de Menú y memoria.

Se ha tenido que buscar documentación de la librería Lucene para familiarizarse más con la misma.
Las herramientas utilizadas han sido: GitHub, GitAhead, Eclipse, Google Drive.

La distribución en horas del trabajo de los integrantes ha sido:

- Raúl: 3 horas.
- Pedro: 4 horas.
- Juanjo: 5 horas.

No ha aparecido ninguna dificultad en la realización de la práctica ya que ha sido de gran ayuda el código proporcionado y, en cuanto al apartado del menú, ya se había realizado algo similar anteriormente.