UNIVERSIDAD DE GRANADA
E.T.S.I. INFORMÁTICA Y TELECOMUNICACIÓN

Departamento de Ciencias de la
Computación e Inteligencia Artificial

Gestión de Información en la Web

Guión de Prácticas

Práctica 3:
Desarrollo de un Sistema de Recuperación de Información con Lucene

Curso 2016-2017

Máster en Ingeniería Informática

Práctica 3

Desarrollo de un Sistema de Recuperación de Información con Lucene

1. Objetivos

1. Conocer las partes principales que tiene un sistema de recuperación de información y qué funcionalidad tiene cada una.
2. Implementar un sistema de recuperación de información.
3. Emplear la biblioteca Lucene para facilitar dicha implementación.

2. Trabajo a Realizar

En algún momento nos puede surgir la necesidad de desarrollar un sistema de recuperación de información. Aunque podemos partir de cero, existe un gran número de bibliotecas en diferentes lenguajes de programación que nos pueden permitir montar el software de búsqueda de una forma rápida y sencilla.

Este es el caso de Lucene, el cual es un conjunto de bibliotecas, escritas en Java (nativo), C++, C#, PHP, Python, Ruby o Perl, que nos facilitará la confección de aplicaciones para realizar la indexación de grandes volúmenes de documentos y recuperación a partir de consultas suministradas por los usuarios del sistema.

En esta práctica se construirá un sistema de recuperación de información empleando la biblioteca Lucene, compuesto de dos programas:

1. Un **indexador**, el cual recibirá como argumentos la ruta de la colección documental a indexar, el fichero de palabras vacías a emplear y la ruta donde alojar los índices, y llevará a cabo la indexación, creando los índices oportunos y ficheros auxiliares necesarios para la recuperación. Esta aplicación se ejecutará en la línea de mandatos y no tendrá ningún componente gráfico. Este software realizará las tareas de tokenización, eliminación de palabras vacías y extracción de raíces antes de crear el índice.
2. Un **motor de búsqueda**, que al ejecutarse recibirá como argumento la ruta donde está alojado el índice de la colección y permitirá que un usuario realice una consulta de texto y obtenga el conjunto de documentos relevantes a dicha consulta. En este caso, el programa sí será gráfico. Sobre la consulta se realizarán los mismos procesos que sobre los documentos en el indexador.

3. Documentación y Ficheros a Entregar

La práctica se subirá a la plataforma docente Prado2 en forma de fichero comprimido (en formato zip o rar), con el nombre "practica3_apellidos_nombre.zip". Este fichero contendrá a su vez todo el código fuente, dos ficheros jar, uno para el indexador y otro para el recuperador, y por último, un archivo pdf con la documentación.

La **documentación** de la práctica será un fichero *pdf* que deberá incluir, al menos, el siguiente contenido:

a) Portada con el número y título de la práctica, el curso académico y el nombre, DNI y dirección e-mail del alumno.

b) La documentación de la práctica contendrá una breve descripción de la misma, explicando cómo se ha hecho, los elementos innovadores y un pequeño manual de usuario.

c) Referencias bibliográficas u otro tipo de material distinto del proporcionado en la asignatura que se haya consultado para realizar la práctica (en caso de haberlo hecho).

La fecha de entrega será el 3 de mayo, como máximo justo antes de clase (15:30h). En esa clase de prácticas procederéis a mostrar vuestras aplicaciones al resto de compañeros y a explicar cómo la habéis desarrollado.

Aunque lo esencial es el contenido, también debe cuidarse la presentación y la redacción.

4. Evaluación

La evaluación de esta práctica tendrá en cuenta la calidad de los siguientes aspectos:

- Adecuación a los requerimientos de las aplicaciones propuestas.
- Correcto uso de las bibliotecas de Lucene.
- Inclusión de aspectos innovadores no vistos en clase.
- Documentación de la práctica.

5. Logística

Colección documental

La colección documental que emplearemos será la formada por noticias en español de la agencia EFE publicadas en los años 1994. Os la podéis descargar del enlace *https://consigna.ugr.es/g/NVUThaH1Q69zSopM/efe94.tgz*.

A continuación se muestra un ejemplo:

```
<DOC>
<DOCNO>EFE19940101-00022</DOCNO>
<DOCID>EFE19940101-00022</DOCID>
<DATE>19940101</DATE>
<TIME>09.13</TIME>
<SCATE>VAR</SCATE>
<FICHEROS>94F.JPG</FICHEROS>
<DESTINO>MUN EXG</DESTINO>
<CATEGORY>VARIOS</CATEGORY>
<CLAVE>DP2456</CLAVE>
<NUM>159</NUM>
<PRIORIDAD>R</PRIORIDAD>
<TITLE>  JAPON-SIDA  CIENTIFICOS DESARROLLAN VACUNA CONTRA SIDA, SEGÚN PRENSA </TITLE>
<TEXT>  Tokio, 1 ene (EFE).- Científicos japoneses han obtenido una vacuna  contra el virus HIV, causante del Síndrome de Inmunodeficiencia  Adquirida (SIDA), utilizando la vacuna BCG contra la tuberculosis,  afirma hoy, sábado la prensa local.  El rotativo de difusión nacional "Mainichi" indica que los dos  equipos que han estado trabajando en el mismo proyecto, el Instituto  Nacional de Salud y un instituto de investigación de una empresa de alimentación japonesa, confirmaron la validez de la vacuna en el rechazo de la enfermedad en ensayos con ratas de laboratorio.  El diario añade que ambos centros harán este mes experimentos inmunológicos con monos en el instituto de investigación de la ciudad de las Ciencias, Tsukuba, próxima a Tokio. Según el periódico, la vacuna estará diseñada para prevenir el HIV  entre los tailandeses y japoneses, pues se ha desarrollado con genes para la vacuna antituberculosa de estas poblaciones. EFE  cd/mar/ha  01/01/09-13/94
</TEXT>
</DOC>
```

En nuestro caso, sólo indexaremos en los campos de Lucene el contenido de las marcas TITLE y TEXT.

Una lista de palabras vacías que se puede emplear está situada en la sección de la *Practica 3* de Prado2.

Material de referencia sobre Lucene

- https://lucene.apache.org
- Lucene in Action.
- http://en.wikipedia.org/wiki/Lucene

Apuntes sobre Lucene

Al descargarnos del sitio web de Lucene la aplicación, nos encontramos tres ficheros que tenemos que incluirlos en el classpath de las aplicaciones que los usen (XXX es la versión de Lucene):

- lucene-core-XXX.jar
- lucene-analyzer-XXX.jar
- lucene-queryparser-XXX.jar

Para usar Lucene, una aplicación de indexación debería:

- Crear Documents añadiéndole Fields.
- Crear IndexWriter y añadir documentos con addDocument(); (eliminando palabras vacías y haciendo stemming, en su caso).

Un código simplificado de la indexación sería el siguiente:

```
Analyzer analyzer = new   StandardAnalyzer(Version.LUCENE_43);
// Store the index in memory:
Directory directory = new RAMDirectory();
//  en disco ... //Directory directory = FSDirectory.open("/tmp/testindex");
IndexWriterConfig config = new  IndexWriterConfig(Version.LUCENE_43, analyzer);
IndexWriter iwriter = new IndexWriter(directory, config);
Document doc = new Document();
String text = "This is the TEXT to be indexed (and searched)!! .";
doc.add(new Field("name", text, TextField.TYPE_STORED));
iwriter.addDocument(doc);
iwriter.close();
```

Lucene ofrece una amplia variedad de analizadores, que son los encargados de realizar toda la gestión de tokens del texto original. En el caso de esta práctica, tendréis que emplear el SpanishAnalyzer, que realiza stemming del español (https://lucene.apache.org/core/4_3_0/analyzers-common/org/apache/lucene/analysis/es/SpanishAnalyzer.html).

Una vez tengamos los índices creados, podemos emplear la herramienta Luke (https://code.google.com/p/luke) para acceder a un índice existente, poder visualizarlo y comprobar que el proceso se ha realizado correctamente.

Y en el caso de una aplicación de consulta:

- Llamar a QueryParser.parse() para construir una consulta desde una cadena de caracteres.
- Crear un IndexSearcher y pasarle la consulta con el método search();

El código simplificado para la consulta y la recuperación sería el siguiente:

```
DirectoryReader ireader = DirectoryReader.open(directory);
IndexSearcher isearcher = new IndexSearcher(ireader);
// Parse a simple query that searches for "text":
QueryParser parser = new QueryParser(Version.LUCENE_43, "fieldToSearch", analyzer);
```

```
Query query = parser.parse("text");
ScoreDoc[] hits = isearcher.search(query, null, 1000).scoreDocs;
// Iterate through the results:
for (int i = 0; i < hits.length; i++) {
    Document hitDoc = isearcher.doc(hits[i].doc);
     System.out.println("salida "+hitDoc.get("name").toString());
     System.out.println("salida "+hitDoc.toString());
    }
ireader.close();
directory.close();
```

Las clases que se han empleado en este ejemplo son:

```
import org.apache.lucene.analysis.Analyzer;
import org.apache.lucene.analysis.standard.StandardAnalyzer;
import org.apache.lucene.store.Directory;
import org.apache.lucene.store.RAMDirectory;
import org.apache.lucene.index.DirectoryReader;
import org.apache.lucene.search.IndexSearcher;
import org.apache.lucene.queryparser.classic.QueryParser;
import org.apache.lucene.queryparser.classic.ParseException;
import org.apache.lucene.search.Query;
import org.apache.lucene.document.Document;
import org.apache.lucene.document.Field;
import org.apache.lucene.search.ScoreDoc;
import org.apache.lucene.document.TextField;
import org.apache.lucene.index.IndexWriter;
import org.apache.lucene.index.IndexWriterConfig;
import org.apache.lucene.store.FSDirectory;
import org.apache.lucene.util.Version;
```