Sistema de recuperación de información

José Alberto Gómez García

Índice de contenidos

01

Herramientas empleadas

Herramientas empleadas

02

Arquitectura del software

Arquitectura del software

- INDEXADOR
  - Clase que abstrae el funcionamiento.
  - Programa principa
- BUSCADOR
  - Clase que abstrae el funcionamiento.
  - Programa principal.
- UTILIDADES
  - Configuración y su modificación.
  - Gestión de ficheros y directorios.
  - Scrapper para obtener documentos.

03

Código desarrolado

Indexador I

```java
public Indexador() throws IOException {
    // Palabras vacías vienen de la configuración
    List<String> words = StopWordsReader.readStopWords();
    CharArraySet stopWords = StopFilter.makeStopSet(words, true);
    Analyzer analyzer;  // Especializado en función del idioma
    if (Configuration.LANGUAGE.equalsIgnoreCase("ES"))
        analyzer = new SpanishAnalyzer(stopWords);
    else if (Configuration.LANGUAGE.equalsIgnoreCase("EN"))
        analyzer = new EnglishAnalyzer(stopWords);
    else
        analyzer = new StandardAnalyzer(stopWords);

    // Gestión del directorio del índice. De existir se sobreescribe, si no se crea.
    File directorio = new File(Configuration.INDEX_DIR);
    if (directorio.exists() && directorio.isDirectory()) {
        System.out.println("Borrando directorio de índice existente: " + Configuration.INDEX_DIR);
        borrarDirectorio(directorio);
    }
    if (!directorio.exists()) {
        directorio.mkdir();
        System.out.println("Directorio para índice creado: " + Configuration.INDEX_DIR);
    }
    Directory indexDirectory = FSDirectory.open(Paths.get(Configuration.INDEX_DIR));
    IndexWriterConfig config = new IndexWriterConfig(analyzer);
    writer = new IndexWriter(indexDirectory, config);
}
```

Indexador II

```java
/**...
private Document getDocumentToIndex(File file) throws IOException {
    Document document = new Document();
    TextField contentField = new TextField("contents", new FileReader(file));
    TextField fileNameField = new TextField("filename", file.getName(), TextField.Store.YES);
    TextField filePathField = new TextField("filepath", file.getCanonicalPath(), TextField.Store.YES);
    document.add(contentField);
    document.add(fileNameField);
    document.add(filePathField);
    return document;
}

/**...
private void indexFile(File file) throws IOException {
    System.out.println("Indexando " + file.getCanonicalPath());
    try {
        Document document = getDocumentToIndex(file);
        writer.addDocument(document);
    } catch (IOException e) {
        String errorWhere = "Error durante la indexación " + file.getCanonicalPath() + "\n";
        System.out.println(errorWhere + e.getMessage());
    }
}

/**...
public Long doIndexing(FileFilter filter) throws IOException {
    File[] files = new File(Configuration.DOCUMENTS_DIR).listFiles();
    assert files ≠ null;
    for (File file : files) {
        if (!file.isDirectory() && !file.isHidden() && file.exists() && file.canRead() && filter.accept(file))
            indexFile(file);
        else
            System.out.println("No se ha indexado " + file.getCanonicalPath());
    }
    return writer.getMaxCompletedSequenceNumber();
}
```

Buscador

```java
public Buscador() throws IOException {
    // Obtener directorio del índice del sistema de archivos
    Directory indexDirectory = FSDirectory.open(Paths.get(Configuration.INDEX_DIR));
    try {
        IndexReader indexReader = DirectoryReader.open(indexDirectory);
        indexSearcher = new IndexSearcher(indexReader);

        // El fichero de las palabras vacías viene de la configuración (ya sea del fichero o del usuario)
        List<String> words = StopWordsReader.readStopWords();
        CharArraySet stopWords = StopFilter.makeStopSet(words, true);

        Analyzer analyzer;  // Especializaremos en función del idioma
        if (Configuration.LANGUAGE.equalsIgnoreCase("ES"))
            analyzer = new SpanishAnalyzer(stopWords);
        else if (Configuration.LANGUAGE.equalsIgnoreCase("EN"))
            analyzer = new EnglishAnalyzer(stopWords);
        else
            analyzer = new StandardAnalyzer(stopWords);

        queryParser = new QueryParser("contents", analyzer);

    } catch (IOException e) {
        System.out.println("Error en la creación del buscador. \n" + e.getMessage());
    }
}

/**...
public TopDocs search(String searchQuery) throws ParseException, IOException {
    query = queryParser.parse(searchQuery);
    return indexSearcher.search(query, Configuration.MAX_SEARCH_RESULTS);
}

/**...
public Document getDocument(ScoreDoc scoreDoc) throws IOException {
    return indexSearcher.doc(scoreDoc.doc);
}
```

Configuración

```java
public class Configuration {
    public static String INDEX_DIR = "./proyecto/index";
    public static String DOCUMENTS_DIR = "./proyecto/collection";
    public static String LANGUAGE = "EN";    // EN or ES
    public static String STOPWORDS_FILE = "./proyecto/data/stopwords_en.txt";    // ./proyecto/data/stopwords_es.txt
    public static int MAX_SEARCH_RESULTS = 15;
    public static String SEARCH_BY = "filename";    // filename or filepath
    public static Boolean USE_THRESHOLD = false;
    public static Float THRESHOLD = 2.0f;  // [0; 100]
}
```

```java
/**...
public void configurarBuscador() {...
}

/**...
public void configurarIndexador() {...
}

/**...
private void configurarDirectorios(String mode) {...
}

/**...
private void configurarIdioma() {...
}
```

```java
/**...
private void configurarFicheroStopWords() {...
}

/**...
private void configurarNumResultadosBusqueda() {...
}

/**...
private void configurarFormatoSalidaBusqueda() {...
}

/**...
private void configurarFiltroSimilitud() {...
}
```

04

Demostración

Uso del indexador

```
¿Desea utilizar la configuración por defecto? (S/N)
S
Directorio para índice creado: ./proyecto/index
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10001.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10002.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10003.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10004.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10005.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\G-Philemon.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\G-Philippians.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\G-Revelation.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\G-Romans.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\G-Titus.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\Introduction_and_Copyright.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\phys-0.txt
Tiempo de indexación de 227 documentos: 6056 milisegundos
```

```
¿Desea utilizar la configuración por defecto? (S/N)
N
Configuración de los parámetros del indexador:

Introduzca el directorio donde crear el índice:
./proyecto/index
Introduzca el directorio con los documentos:
./proyecto/collection
Introduzca el idioma de los documentos (EN o ES):
EN
Introduzca el fichero de stopwords (txt):
./proyecto/data/stopwords_en.txt
Borrando directorio de índice existente: C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\.\proyecto\index
Directorio para índice creado: C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\.\proyecto\index
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10001.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10002.txt
Indexando C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10003.txt
```

Uso del buscador

```
¿Desea utilizar la configuración por defecto? (S/N)
N
Configuración de los parámetros de la búsqueda:

Introduzca el directorio del índice:
./proyecto/index
Introduzca el directorio con los documentos:
./proyecto/collection
Introduzca el idioma de los documentos (EN o ES):
EN
Introduzca el fichero de stopwords (txt):
./proyecto/data/stopwords_en.txt
Introduzca el número máximo de resultados a mostrar:
12
¿Desea buscar por nombre de fichero o por ruta de fichero? (filename/filepath)
filepath
¿Desea utilizar un umbral de similitud? (S/N)
S
Introduzca el umbral de similitud [0-100]:
1.5
Introduzca una consulta para realizar la búsqueda:
matthew

Tiempo de búsqueda: 42 milisegundos
Número de resultados: 4

Documento: C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10100.txt  --> Score: 1.7259666
Documento: C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\12372.txt  --> Score: 1.7130105
Documento: C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\G-Matthew.txt  --> Score: 1.6518188
Documento: C:\Users\Usuario\Desktop\MASTER\GIW\Practicas\Desarrollo\PracticaLucene\proyecto\collection\10116.txt  --> Score: 1.572145
```

Gracias

¿Alguna pregunta o comentario?

José Alberto Gómez García
modej@correo.ugr.es