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

Manuel Orantes Taboada

Indexador

```java
package indexer;

/**
 *
 * @author manuel-aircury
 */
public class LuceneConstants {
   public static final String CONTENTS = "contents";
   public static final String FILE_NAME = "filename";
   public static final String FILE_PATH = "filepath";
   public static final int MAX_SEARCH = 10;
}
```

```java
package indexer;

/**
 *
 * @author manuel-aircury
 */
import java.io.File;
import java.io.FileFilter;
public class TextFileFilter implements FileFilter {
   @Override
   public boolean accept(File pathname) {
      return pathname.getName().toLowerCase().endsWith(".txt");
   }
}
```

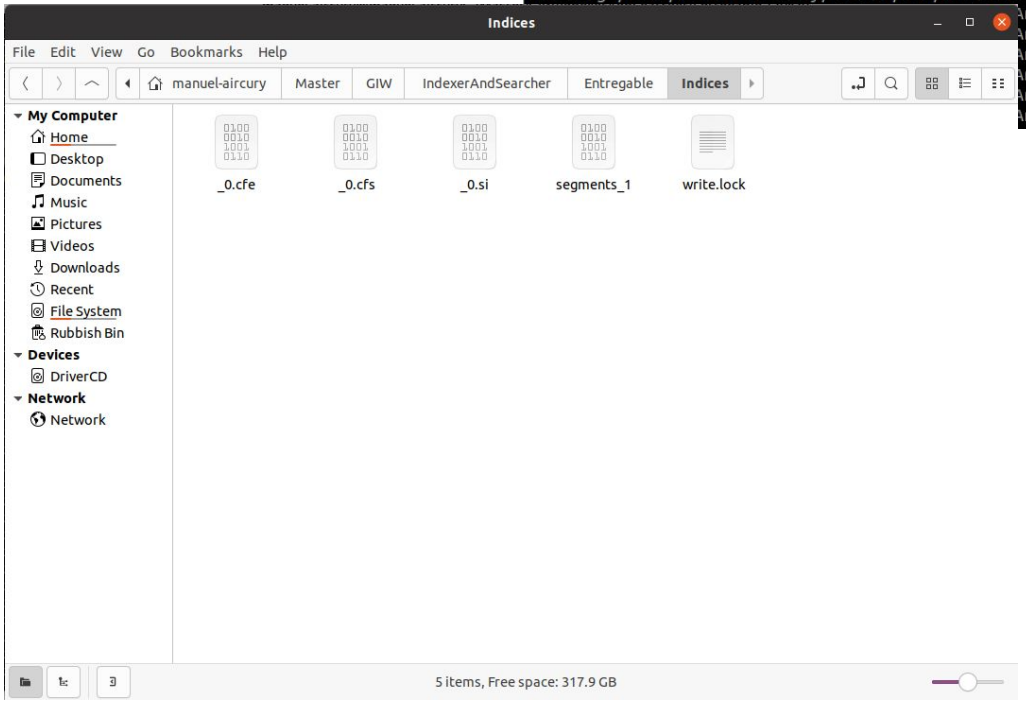
Indexador

```java
public class IndexerEngine {
   private final IndexWriter writer;

   public IndexerEngine(String indexDirectoryPath) throws IOException {
      Directory indexDirectory = FSDirectory.open(Paths.get(indexDirectoryPath));
      List<String> stopWords = new ArrayList<>();
      String cadena;
      FileReader f = new FileReader("PalabrasVaciasIngles.txt");
      BufferedReader b = new BufferedReader(f);
      while((cadena = b.readLine())!=null) {
         stopWords.add(cadena);
      }
      b.close();
      CharArraySet stopSet = new CharArraySet(stopWords, true);
      StandardAnalyzer analyzer = new StandardAnalyzer(stopSet);
      IndexWriterConfig iwc = new IndexWriterConfig(analyzer);
      writer = new IndexWriter(indexDirectory, iwc);
      if(DirectoryReader.indexExists(indexDirectory)){
        writer.deleteAll();
      }
   }

   public void close() throws CorruptIndexException, IOException {
      writer.close();
   }
```

```java
private Document getDocument(File file) throws IOException {
   Document document = new Document();
   TextField contentField = new TextField(LuceneConstants.CONTENTS,
         new FileReader(file));
   TextField fileNameField = new TextField(LuceneConstants.FILE_NAME,
         file.getName(), TextField.Store.YES);
   TextField filePathField = new TextField(LuceneConstants.FILE_PATH,
         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("Indexing "+file.getCanonicalPath());
   Document document = getDocument(file);
   writer.addDocument(document);
}

public int createIndex(String dataDirPath, FileFilter filter)
   throws IOException {
   File[] files = new File(dataDirPath).listFiles();
   for (File file : files) {
      if(!file.isDirectory()
         && !file.isHidden()
         && file.exists()
         && file.canRead()
         && filter.accept(file)
      ){
         indexFile(file);
      }
   }
   return writer.numRamDocs();
}
```

Indexador

```
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
ndSearcher/awd_1994_61/a9461554.txt
ndSearcher/awd_1994_61/a9461065.txt
ndSearcher/awd_1994_61/a9461868.txt
ndSearcher/awd_1994_61/a9461701.txt
ndSearcher/awd_1994_61/a9461773.txt
ndSearcher/awd_1994_61/a9461757.txt
```

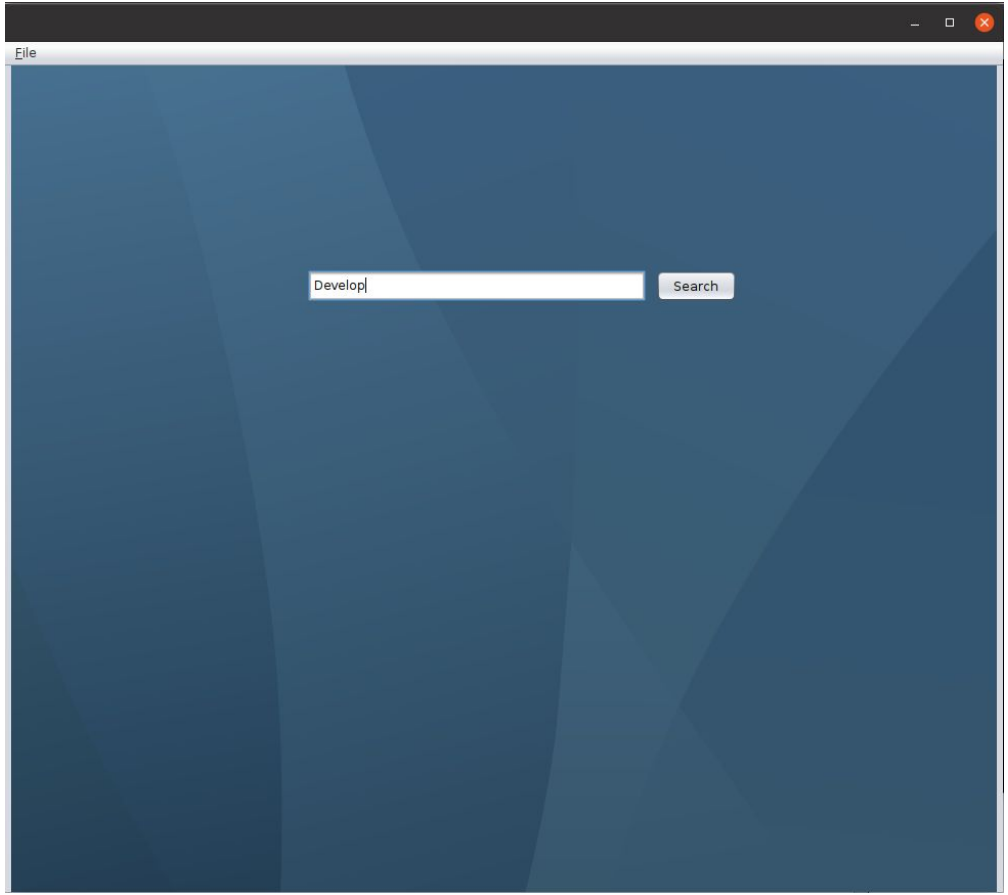


Buscador

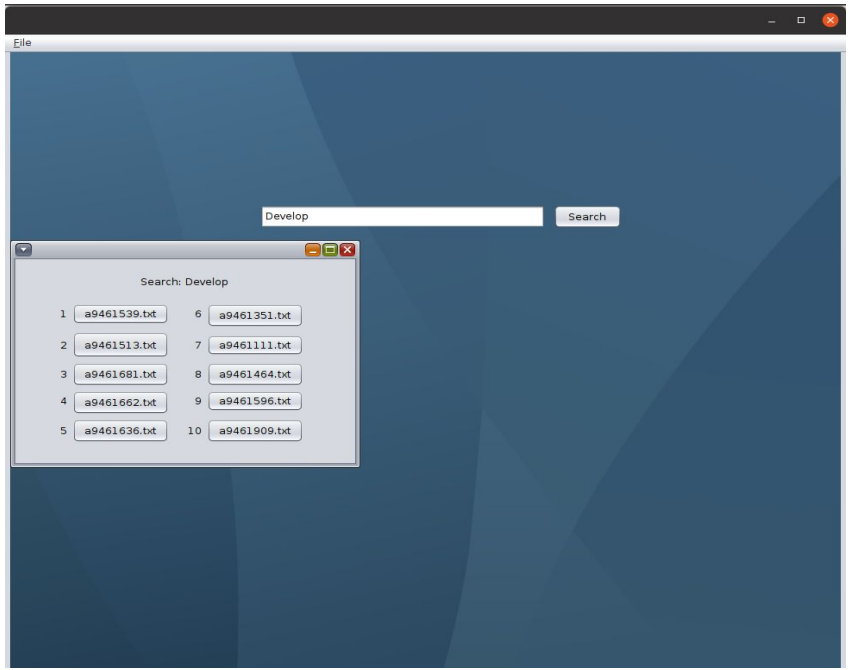
```java
public class SearcherEngine {

   IndexSearcher indexSearcher;
   QueryParser queryParser;
   Query query;

   public SearcherEngine(String indexDirectoryPath)
      throws IOException {
      Directory indexDirectory = FSDirectory.open(Paths.get(indexDirectoryPath));
      IndexReader reader = DirectoryReader.open(indexDirectory);
      indexSearcher = new IndexSearcher(reader);
      queryParser = new QueryParser(LuceneConstants.CONTENTS, new StandardAnalyzer());
   }

   public TopDocs search(String searchQuery)
      throws IOException, ParseException {
      query = queryParser.parse(searchQuery);
      return indexSearcher.search(query, LuceneConstants.MAX_SEARCH);
   }

   public Document getDocument(ScoreDoc scoreDoc)
      throws CorruptIndexException, IOException {
      return indexSearcher.doc(scoreDoc.doc);
   }

}
```

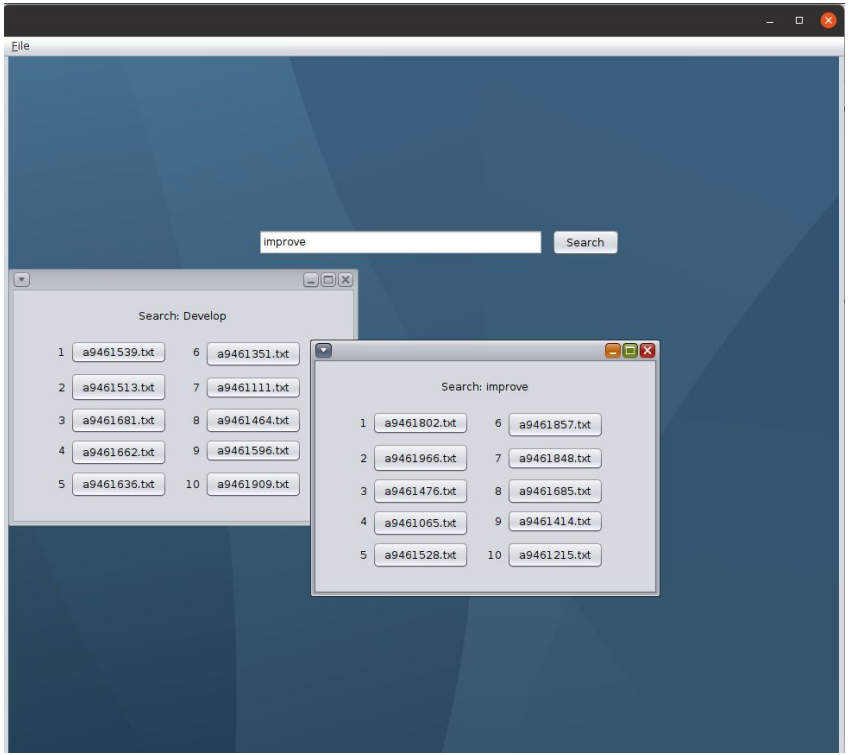
Buscador

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

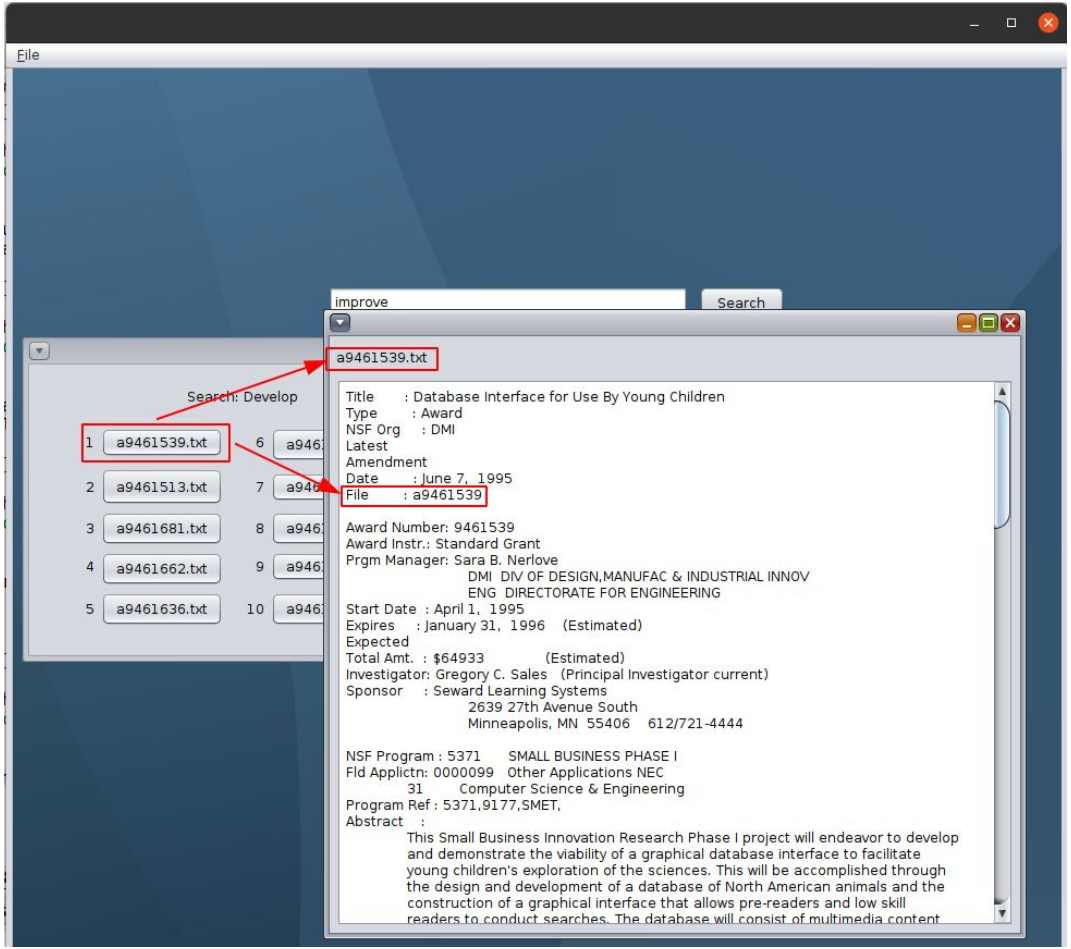


Buscador


File
Develop
Search
Search: Develop
1 a9461539.txt
2 a9461513.txt
3 a9461681.txt
4 a9461662.txt
5 a9461636.txt
6 a9461351.txt
7 a9461111.txt
8 a9461464.txt
9 a9461596.txt
10 a9461909.txt


File
improve
Search
Search: Develop
1 a9461539.txt
2 a9461513.txt
3 a9461681.txt
4 a9461662.txt
5 a9461636.txt
6 a9461351.txt
7 a9461111.txt
8 a9461464.txt
9 a9461596.txt
10 a9461909.txt
Search: improve
1 a9461802.txt
2 a9461966.txt
3 a9461476.txt
4 a9461065.txt
5 a9461528.txt
6 a9461857.txt
7 a9461848.txt
8 a9461685.txt
9 a9461414.txt
10 a9461215.txt

Buscador


File
improve
Search
Search: Develop
1 a9461539.txt
2 a9461513.txt
3 a9461681.txt
4 a9461662.txt
5 a9461636.txt
a9461539.txt
Title : Database Interface for Use By Young Children
Type : Award
NSF Org : DMI
Latest
Amendment
Date : June 7, 1995
File : a9461539
Award Number: 9461539
Award Instr.: Standard Grant
Prgm Manager: Sara B. Nerlove
DMI DIV OF DESIGN,MANUFAC & INDUSTRIAL INNOV
ENG DIRECTORATE FOR ENGINEERING
Start Date : April 1, 1995
Expires : January 31, 1996 (Estimated)
Expected
Total Amt. : $64933 (Estimated)
Investigator: Gregory C. Sales (Principal Investigator current)
Sponsor : Seward Learning Systems
2639 27th Avenue South
Minneapolis, MN 55406 612/721-4444
NSF Program : 5371 SMALL BUSINESS PHASE I
Fld Applictn: 0000099 Other Applications NEC
31 Computer Science & Engineering
Program Ref : 5371,9177,SMET,
Abstract :
This Small Business Innovation Research Phase I project will endeavor to develop and demonstrate the viability of a graphical database interface to facilitate young children's exploration of the sciences. This will be accomplished through the design and development of a database of North American animals and the construction of a graphical interface that allows pre-readers and low skill readers to conduct searches. The database will consist of multimedia content