Data Management : Lab 1

Antoine Friant

14 octobre 2018

1 Understanding the Lucene API

1.1 Does the command line demo use stopword removal ?

Yes, because the query "a an or and at query response" returned the exact same results as "query response". If it weren't using stopword removal, we'd get less results when using stopwords.

1.2 "Does the command line demo use stemming ?"

No, because the queries "query" and "queries" don't give the same number of results.

1.3 "Is the search of the command line demo case insensitive ?"

Yes, because "QUERY" and "query" give the same results.

1.4 "Does it matter whether stemming occurs before or after stopword removal ?"

It only matters if stopwords find themselves grouped with non stopwords after the stemming step. If such a thing is possible, then we need to remove stopwords before stemming. If we don't, words from the query could be ignored or stopwords could change the result of the query.

2 Indexing and Searching the CACM collection

2.1 Explain which field type can be used for id, title, summary and author

We need an integer for "id" so LongPoint or IntPoint. The others need to be TextField to perform full-text searches. We chose LongPoint since the id is type long. In order to store it, we also create an StoreField "id".

2.2 Using different Analyzers

2.2.1 StandardAnalyzer

- Indexed documents : 3'203
- Indexed terms : 26'634

- Indexed terms in the summary field : 19’972
- The top 10 frequent terms of the summary field in the index : which, system, computer, paper, can, described, given, presented, time, from
- Size of the index on disk : 2.2 MB
- Required time for indexing : 965 ms

2.2.2 WhitespaceAnalyzer

- Indexed documents : 3’203
- Indexed terms : 34’443
- Indexed terms in the summary field : 26’821
- The top 10 frequent terms of the summary field in the index : of, the, is, a, and, to, in, for, The, are
- Size of the index on disk : 2.6 MB
- Required time for indexing : 971 ms

2.2.3 EnglishAnalyzer

- Indexed documents : 3’203
- Indexed terms : 22’513
- Indexed terms in the summary field : 16’724
- The top 10 frequent terms of the summary field in the index : us, which, comput, program, system, present, describ, paper, method, can
- Size of the index on disk : 2.1 MB
- Required time for indexing : 1’129 ms

2.2.4 ShingleAnalyzerWrapper (shingle size 2)

- Indexed documents : 3’203
- Indexed terms : 105’802
- Indexed terms in the summary field : 85’610
- The top 10 frequent terms of the summary field in the index : which, system, paper, computer, can, paper, described, given, presented, time
- Size of the index on disk : 4.8 MB
- Required time for indexing : 1’682 ms

2.2.5 ShingleAnalyzerWrapper (shingle size 3)

- Indexed documents : 3'203
- Indexed terms : 147'205
- Indexed terms in the summary field : 125'776
- The top 10 frequent terms of the summary field in the index : which, system, computer, paper, can, described, time, given, presented, from
- Size of the index on disk : 6.3 MB
- Required time for indexing : 1'751 ms

2.2.6 StopAnalyzer

- Indexed documents : 3'203
- Indexed terms : 24'507
- Indexed terms in the summary field : 18'658
- The top 10 frequent terms of the summary field in the index : which, system, computer, paper, described, can, presented, given, time, from
- Size of the index on disk : 2.2 MB
- Required time for indexing : 812 ms

2.3 Reading Index

```
   Most frequent terms in authors :
38 : Thacher Jr., H. C.
   Most frequent terms in title :
983 : algorithm
262 : computer
172 : system
154 : programming
124 : method
112 : data
109 : systems
101 : language
93 : program
84 : time
```

```
// file Main.java
private static Analyzer getAnalyzer() {
        // treat the author field as a single word
        Map<String, Analyzer> analyzerMap = new HashMap<>();
        analyzerMap.put("authors", new KeywordAnalyzer());
        return new PerFieldAnalyzerWrapper(
        new StandardAnalyzer(), analyzerMap);
}
```

```
// file QueriesPerformer.java
public void printTopRankingTerms(String field, int numTerms) {
        // This methods print the top ranking term for a field.
        try {
                TermStats[] foundTerms = HighFreqTerms.getHighFreqTerms(
                    indexReader, numTerms, field,
                new HighFreqTerms.TotalTermFreqComparator());

                System.out.println("\nMost frequent terms in " + field + " :
                    ");

                for (TermStats term : foundTerms) {
                        System.out.println(term.totalTermFreq + " : " + term
                            .termtext.utf8ToString());
                }
        } catch (Exception e) {
                e.printStackTrace();
        }
}
```

2.4 Searching

```
// file Main.java
private static void searching(QueriesPerformer queriesPerformer) {
        queriesPerformer.query("\"Information Retrieval\"");
        queriesPerformer.query("Information Retrieval");
        queriesPerformer.query("Information +Retrieval -Database");
        queriesPerformer.query("Info*");
        queriesPerformer.query("\"Information Retrieval\"~5");
}

// file QueriesPerformer.java
public void query(String q) {
        final int LIMIT = 10;

        QueryParser parser = new QueryParser("summary", analyzer);

        try {
                Query query = parser.parse(q);
                IndexSearcher searcher = new IndexSearcher(indexReader);
                searcher.search(query, 100);

                TopDocs results = searcher.search(query, LIMIT);
                ScoreDoc[] hits = results.scoreDocs;

                System.out.println("\nSearching for: " + q + " (" + results.
                    totalHits + " results)");

                for (int i = 0; i < LIMIT && i < results.totalHits; ++i) {
                        Document doc = searcher.doc(hits[i].doc);
                        System.out.println(doc.get("id") + ": " + doc.get("
                            title"));
                }

                System.out.println();
        } catch (Exception e) {
                e.printStackTrace();
        }

System.out.println("\n");
}

// Output :
Searching for: "Information Retrieval" (11 results)
891: Everyman's Information Retrieval System
1457: Data Manipulation and Programming Problemsin Automatic Information
    Retrieval
1935: Randomized Binary Search Technique
1699: Experimental Evaluation of InformationRetrieval Through a
    Teletypewriter
2516: Hierarchical Storage in Information Retrieval
2519: On the Problem of Communicating Complex Information
2307: Dynamic Document Processing
2795: Sentence Paraphrasing from a Conceptual Base
2990: Effective Information Retrieval Using Term Accuracy
1652: A Code for Non–numeric Information ProcessingApplications in Online
    Systems

Searching for: Information Retrieval (188 results)
1457: Data Manipulation and Programming Problemsin Automatic Information
    Retrieval
891: Everyman's Information Retrieval System
1699: Experimental Evaluation of InformationRetrieval Through a
    Teletypewriter
2307: Dynamic Document Processing
3134: The Use of Normal Multiplication Tablesfor Information Storage and
    Retrieval
1032: Theoretical Considerations in Information Retrieval Systems
1935: Randomized Binary Search Technique
1681: Easy English,a Language for InformationRetrieval Through a Remote
    Typewriter Console
2990: Effective Information Retrieval Using Term Accuracy
2519: On the Problem of Communicating Complex Information

Searching for: Information +Retrieval −Database (54 results)
1457: Data Manipulation and Programming Problemsin Automatic Information
    Retrieval
891: Everyman's Information Retrieval System
1699: Experimental Evaluation of InformationRetrieval Through a
    Teletypewriter
2307: Dynamic Document Processing
3134: The Use of Normal Multiplication Tablesfor Information Storage and
    Retrieval
1032: Theoretical Considerations in Information Retrieval Systems
1935: Randomized Binary Search Technique
1681: Easy English,a Language for InformationRetrieval Through a Remote
    Typewriter Console
2990: Effective Information Retrieval Using Term Accuracy
2519: On the Problem of Communicating Complex Information

Searching for: Info∗ (193 results)
222: Coding Isomorphisms
272: A Storage Allocation Scheme for ALGOL 60
396: Automation of Program  Debugging
397: A Card Format for Reference Files in Information Processing
```

```
409: CL-1, An Environment for a Compiler
440: Record Linkage
483: On the Nonexistence of a Phrase Structure Grammar for ALGOL 60
616: An Information Algebra - Phase I Report-LanguageStructure Group of the
   CODASYL Development Committee
644: A String Language for Symbol Manipulation Based on ALGOL 60
655: COMIT as an IR Language

Searching for: "Information Retrieval"~5 (15 results)
891: Everyman's Information Retrieval System
1457: Data Manipulation and Programming Problemsin Automatic Information
    Retrieval
1935: Randomized Binary Search Technique
1699: Experimental Evaluation of InformationRetrieval Through a
    Teletypewriter
2307: Dynamic Document Processing
2516: Hierarchical Storage in Information Retrieval
2519: On the Problem of Communicating Complex Information
2795: Sentence Paraphrasing from a Conceptual Base
2990: Effective Information Retrieval Using Term Accuracy
1652: A Code for Non-numeric Information ProcessingApplications in Online
    Systems
```

2.5 Tuning the Lucene Score

```
// MySimilarity.java
package ch.heigvd.iict.dmg.labo1.similarities;

import org.apache.lucene.index.FieldInvertState;
import org.apache.lucene.search.similarities.ClassicSimilarity;

public class MySimilarity extends ClassicSimilarity {
        @Override
        public float tf(float freq) {
                return 1 + (float)Math.log(freq);
        }

        @Override
        public float idf(long docFreq, long numDocs) {
                return (float)Math.log(numDocs/((float)docFreq + 1)) + 1;
        }

        @Override
        public float coord(int overlap, int maxOverlap) {
                return (float)Math.sqrt(overlap/(float)maxOverlap);
        }

        @Override
        public float lengthNorm(FieldInvertState state) {
                return 1;
        }
}

// Output :
Searching for: Information Retrieval (188 results)
1032: Theoretical Considerations in Information Retrieval Systems
```

Lucene classic scoring	TF-IDF scoring
1457	1032
891	1457
1699	3134
2307	891
3134	1699
1032	2307
1935	1527
1681	1681
2990	2990
2519	1652

Table 1 – Comparison of the top 10 results using ClassicSimilarity and MySimilarity (document ids). The number of results are the same, but their ranks are different. Only two documents disappear from the top 10 results, so the change is not dramatic overall.

```
1457: Data Manipulation and Programming Problemsin Automatic Information
    Retrieval
3134: The Use of Normal Multiplication Tablesfor Information Storage and
    Retrieval
891: Everyman's Information Retrieval System
1699: Experimental Evaluation of InformationRetrieval Through a
    Teletypewriter
2307: Dynamic Document Processing
1527: A Grammar Base Question Answering Procedure
1681: Easy English ,a Language for InformationRetrieval Through a Remote
    Typewriter Console
2990: Effective Information Retrieval Using Term Accuracy
1652: A Code for Non–numeric Information ProcessingApplications in Online
    Systems
```