Cojocariu Sebastian
Group 507 AI

Information Retrieval - Lucene
Project 1

Dependencies

- Java 17.0.1
- lucene-analyzers-common-8.11.0
- lucene-core-8.11.0
- lucene-queryparser-8.11.0
- tika-app-2.1.0

Observations:

- All jars are in the 'dependencies/' directory!
- The class (binary) files are saved in the 'targets/' directory!
- For information about building/running the application, check the available README (or the last section of this report). There are also some testing commands.
- Inside the root, there is also a stopwords.txt file used to augment the existing stop-words list existing in Lucene.

Implementation details

For this project, three classes were envisioned:

- IRTMAnalyzer: class that extends Lucene's Analyzer. Internally, several custom filters were built, by extending Lucene's TokenFilter:
  - *IRTMReplaceDiacriticsFilter*: filter that replaces, for each token, all romanian diacritics with their base letters.
  - *IRTMSynonymFilter*: filter that creates, for each token, a synonymy relationship with each possible combination of diacritics, while preserving the same index.

  To accommodate the imposed requirements, multiple filters were applied in the following order:
  - First, a lowercase filter (*LowerCaseFilter*).
  - Then, all diacritics are replaced by their base letter (*IRTMReplaceDiacriticsFilter*). For this, a function called replace_diacritics() was implemented.

- All romanian stop-words (without diacritics) are then removed (*StopFilter*). We used the romanian stop-words provided by Lucene. This set is augmented with other stopwords.
- For each remaining token, compute all the possible combinations of diacritics and mark them as synonyms (*IRTMSynonymFilter*).
- Apply the internal Lucene's romanian stemmer to reduce the words to their base form (*SnowballFilter*).
- Remove again the diacritics from the remaining tokens (*IRTMReplaceDiacriticsFilter*).
- Since there may be duplicate 'synonyms', we remove them (*RemoveDuplicatesTokenFilter).*

- Indexer: Class that indexes all the documents specified by -documents_path (default = './documents/') and stores the results in index_path (default = "./index_files/") (there parameters are configurable via command line). A recursive approach is used to get all the files (txt, pdf, doc, docx and more). To read their content, we used Tika built-in parse() method. When saving the documents, we stored two relevant fields: their absolute path ('full_path'), as well as their internal content('contents').

- Searcher: Class that searches a set of indexed files (specified by the location -index_path; default='index_files/'), using a custom query (specified by -query_string command line argument). The matching is based on the field 'contents'. Additionally, one can specify how many hits to show using –max_hits (default = 25). More, by setting -interactive to true, one can enter an interactive session of querying. The matches are shown based on their relevance, using Lucene matching score.

Corpus Collection

To collect a suitable corpus, we used link as starting point. We extracted romanian texts from arta politica (containing diacritics), converting them into various possible formats (txt, pdf, doc, docx). Based on their type, we distinguish:

- 5 doc files.
- 6 docx files.
- 8 pdf files.
- 5 txt files.

Additionally, there are the test_camasa.txt and test_cămașă.txt files used for testing purposes.
All the files are in the ‘documents/’ directory.

Running the application

To run the application, use the following commands:

A) Building

javac -proc:none -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./src/main/java/" -d targets src/main/java/Indexer.java

javac -proc:none -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./src/main/java/" -d targets src/main/java/Searcher.java

B) Running

1) Indexing:

java -Xms1024m -Xmx4096m -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets" Indexer -documents_path documents/ -index_path index_files/

Default: -documents_path=’documents/’, -index_path=’index_files/’

Cojocariu Sebastian
Group 507 AI

2) Searching:

Non-interactive (query is specified through command line):

java -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets/" Searcher -index_path index_files/ -query_string "Carutelor" -interactive false

Interactive:

java -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets/" Searcher -index_path index_files/ -interactive true

Default: -index_path='index_files/', -interactive='false'

Caution: when using a custom -index_path location, one needs to be careful to call both the indexer and the searcher with the same index_path location!

Tests

java -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets/" Searcher -index_path index_files/ -query_string "și si ca că ci" -interactive false

java -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets/" Searcher -index_path index_files/ -query_string "camasa" -interactive false

java -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-

queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets/" Searcher -index_path index_files/ -query_string "camasilor" -interactive false

java -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets/" Searcher -index_path index_files/ -query_string "camașile" -interactive false

java -cp "./dependencies/lucene-analyzers-common-8.11.0.jar:./dependencies/lucene-core-8.11.0.jar:./dependencies/lucene-queryparser-8.11.0.jar:./dependencies/tika-app-2.1.0.jar:./targets/" Searcher -index_path index_files/ -query_string "cămasa" -interactive false