Data Mining 11.10.2017 – Piotr Mrowczynski – Assignment 1

Finding textually similar documents based on Jaccard similarity using the shingling, minhashing, locality-sensitive hashing (LSH) techniques and corresponding algorithms.

Project structure

Projects consists of commented code, in which all the steps of preparing dataset, shingling, compressing shingles,

```
finding-similar-docs
├── src/main/java/mrow4a/spark/java/
│   ├── FindingSimilarDocsBasicJob .java        :  Version using just shingles and similarity
│   │
│   ├── FindingSimilarDocsMinHashJob.java     :  Version using MinHash algorithm
│   │
│   └── alg
│       ├── EmailParser.java                  :  Cleans the email input data
│       ├── Shingling.java                    :  Shingling library
│       ├── MinHash.java                      :  Class responsible for MinHash algorithm
│       ├── Signature.java                    :  Class responsible for creating signatures
│       └── JaccardSimilarity.java            :  Jaccard Similarity library
│
├── build.sh                                  : Building project
│
├── run.sh                                    : Running classes
│
└── SimilarDocs.odt
```

Building (Docker required for maven single assembly)

chmod +x finding-similar-docs/build.sh
chmod +x finding-similar-docs/run.sh
./finding-similar-docs/build.sh

Running basic version

finding-similar-docs/run.sh FindingSimilarDocsBasicJob

Running MinHash version

finding-similar-docs/run.sh FindingSimilarDocsMinHashJob

Project results – basic algorithm

In the first measurement, I used basic version of finding similar items (without MinHash) and used similarity threshold for emails 0.3.

Basic, Threshold: 0.3

(file:mini_newsgroups/alt.atheism/51170,file:mini_newsgroups/alt.atheism/51203): 0.34260178
(file:mini_newsgroups/alt.atheism/53222,file:/mini_newsgroups/alt.atheism/53235): 0.45844224
(file:mini_newsgroups/alt.atheism/53062,file:mini_newsgroups/alt.atheism/53759): 0.4419643
(file:mini_newsgroups/alt.atheism/53190,file:mini_newsgroups/alt.atheism/54201): 0.42916915
(file:mini_newsgroups/alt.atheism/53633,file:mini_newsgroups/alt.atheism/54237): 0.37470996

Found similar document pairs [5/4950] with similarity threshold [0.3] and shingle lenght [5]
Time: 4264 milliseconds

Found similar document pairs [80/1999000] with similarity threshold [0.3] and shingle lenght [5]
Time: 122649 milliseconds

Increasing number of documents from 100 to 2000, running time of algorithm increased from 4s to 122s

Project results – MinHash algorithm

In the first measurement, I used basic version of finding similar items (without MinHash) and used similarity threshold for emails 0.2. It was found, that for MinHash, decreasing a threshold, allowed to retrieve nearly the same result as for basic version of algorithm.

MinHash, Threshold: 0.2

(file:mini_newsgroups/alt.atheism/51170,file:mini_newsgroups/alt.atheism/51203): 0.24223602
(file:mini_newsgroups/alt.atheism/53222,file:mini_newsgroups/alt.atheism/53235): 0.3605442
(file:mini_newsgroups/alt.atheism/53062,file:mini_newsgroups/alt.atheism/53759): 0.30718955
(file:mini_newsgroups/alt.atheism/53190,file:mini_newsgroups/alt.atheism/54201): 0.24223602
(file:mini_newsgroups/alt.atheism/53633,file:mini_newsgroups/alt.atheism/54237): 0.24223602

Time: 3283 milliseconds
Found similar document pairs [5/4950] with similarity threshold [0.2] and shingle lenght [5] and signature lenght [10]

Time: 18397 milliseconds
Found similar document pairs [422/1997001] with similarity threshold [0.2] and shingle lenght [5] and signature lenght [10]

Furthermore, icreasing number of documents from 100 to 2000, increased running time only from 3s to 18s (instead of 122s)

Time: 26232 milliseconds
Found similar document pairs [70/1997001] with similarity threshold [0.2] and shingle lenght [5] and signature lenght [100]

Time: 108654 milliseconds
Found similar document pairs [237/1997001] with similarity threshold [0.2] and shingle lenght [5] and signature lenght [1000]

Additionaly, increasing signature length from 10 to 100 and 1000, increased running time from 3s to 26s and 108s, however this allowed to increase the precision.