Homework 1: Finding Similar Items: Textually Similar Documents

Konstantin Sozinov, sozinov@kth.se
Kim Hammar, kimham@kth.se

November 7, 2017

1 Solution

The implementation is done in pure Scala without any big data processing framework. The functionality is split into different class files: `Shingling.scala`, `MinHashing.scala`, `CompareSets.scala`, `CompareSignatures.scala`, `LSH.scala`, `Dataset.scala` and `Main.scala`. The first four classes have the functionality as described in the problem description. Dataset is a class with functionality for reading the dataset used for evaluation [1]. Main is a class for orchestrating the different steps of the pipeline: $Shinglinng \rightarrow MinHashing \rightarrow LSH \rightarrow Filter(CompareSignatures) \rightarrow Evaluation$.

2 How to run

Clone this repository and navigate to *similar_items* project. Then use:

```
sbt compile //compile
sbt test //test
sbt run //run
sbt assembly //generate fat jar
```

3 Evaluation and results

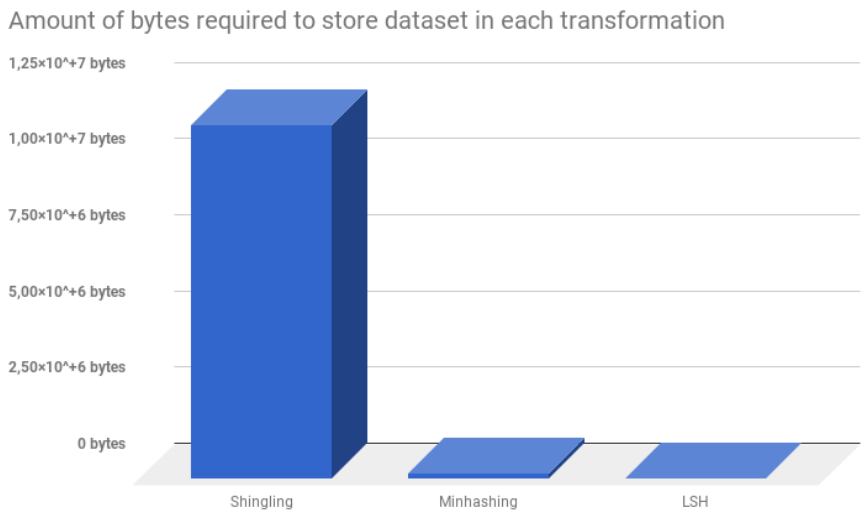

Figure 1: Memory analysis for different stages in comparing the documents

[1] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

Example output (t=0.8, b=10, r=10, n=100)

```
[info] Running kth.se.id2222.Main
Shingles size: 11624856 bytes
Size after minhashing: 153112 bytes
Number of candidates pre LSH is approx: 10816.0
Number of candidates after LSH: 5
Size after LSH: 15656 bytes
Similar items: 4
Similar pair:
 src/resources/mini_newsgroups/alt.atheism/54485_copy,
 src/resources/mini_newsgroups/alt.atheism/54485
 similarity: 0.99
Similar pair:
 src/resources/mini_newsgroups/alt.atheism/51131,
 src/resources/mini_newsgroups/alt.atheism/51131copy
 similarity: 0.96
Similar pair:
 src/resources/mini_newsgroups/alt.atheism/54244,
 src/resources/mini_newsgroups/alt.atheism/54244_copy
 similarity: 0.99
Similar pair:
 src/resources/mini_newsgroups/alt.atheism/53653_copy,
 src/resources/mini_newsgroups/alt.atheism/53653
 similarity: 0.98
Time to compute similar items: 7.256831409 seconds, number of similar items found: 4
[success] Total time: 8 s, completed 2017-nov-07 10:41:05
```