Homework #1: Finding Similar Items - Textually Similar Documents

Davide Anghileri ~ *davidean@kth.se*
Nathan Consuegra ~ *nacon@kth.se*

Short description of the solution

- First we load the text inside the array *docs* and we parallelize it using spark,
- Then we map each document to a set of shingles obtained using the MD5 hash function on each substring of size *shingleSize,*
- Then, using the minHash, we compute the signature for each document. We apply *k* different hash functions of the form h(x) = (a*x +b)%c where a,b and c are random integers to all the shingles of each document and select the shingle that has the smallest hash value,
- Using LSH we select the candidate pairs of documents (documents that have at least one hashed band element equal in the same bucket),
- We check that candidate pairs have similar signature with threshold *(1/band)^(1/r)*, otherwise we remove the candidate pair,
- Then we also check that the remained pairs have similar sets ("shingles") using the Jaccard similarity,
- Finally we print the results.

Dataset

We have use the Opinosis Dataset from UCI and more specifically we have compared 100 reviews of the sound of the ipod nano written into "/topics/sound_ipod_nano_8gb.txt.data". This file was split into multiple files for every new line.

How to build and run

For the project to work, spark must be installed on the computer running the code. Moreover, Jupyter Notebook needs to also be installed to be able to open the file "*source/TextualSimilarity.ipynb*" and run it. Once the file has been opened, go to **Cell > Run All**.

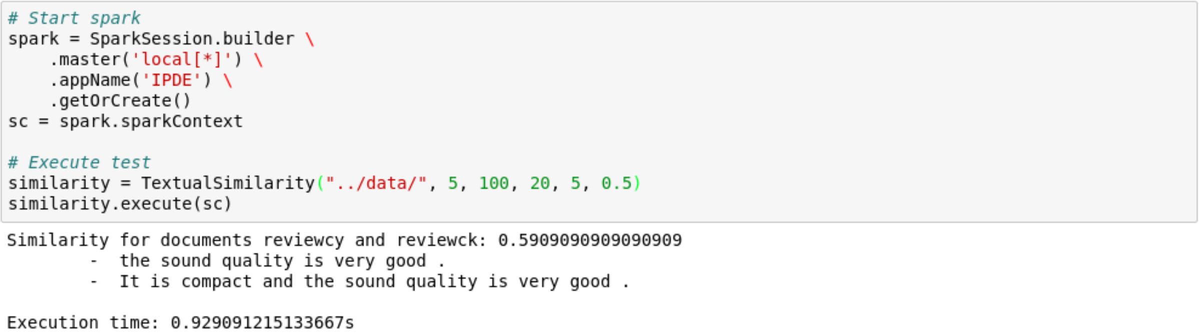
Command-line parameters

No need for command-line parameters since everything can be executed with Jupyter Notebook.

Default Parameters:

- ❑ shingleSize = 5 , is the length of each substring the document is splitted in.
- ❑ band = 20 , the band for the LSH
- ❑ r = 5 , number of rows for the LSH
- ❑ k = 100 , length of each signature produced with k different hash functions during minHashing
- ❑ threshold = (1/band)^(1/r) ,used to check that candidate pairs have similar signatures
- ❑ jaccardThreshold = 0.5 , used to check the similarity with shingles

Screenshots



```
# Start spark
spark = SparkSession.builder \
    .master('local[*]') \
    .appName('IPDE') \
    .getOrCreate()
sc = spark.sparkContext

# Execute test
similarity = TextualSimilarity("../data/", 5, 100, 20, 5, 0.5)
similarity.execute(sc)
```

```
Similarity for documents reviewcy and reviewck: 0.5909090909090909
        -  the sound quality is very good .
        -  It is compact and the sound quality is very good .

Execution time: 0.929091215133667s
```