Data Mining Homework 1

Finding Similar Items: Textually Similar Documents

Group 22 - Blanca Bastardés Climent, Alice Ciallella

Short description

The implementation of this assignment has been done in Scala using Spark. All the code is provided in the HW1.ipynb file. The file is divided as definitions of the classes at the beginning and the executable code to find the similarity among documents. The following different classes can be found:

- Shingle: it creates shingles from text files with length k=7. Note that when reading the data, it is preprocessed by removing punctuation marks and converting everything to lowercase. The shingles are then hashed.
- CompareSet: it computes the jaccard similarity between documents by comparing their sets of hashed shingles.
- MinHashing: creates a signature for each document using 500 permutations.
- CompareSignatures: it compares the similarity between signatures of two documents.
- LocSensHashing: it implements the LSH technique and shows the candidate pairs satisfying a similarity over a threshold t. This threshold is computed from the band size and the length of the signatures in each. In our application, we select a threshold of 0.8 and to ensure no false negatives a band size of 10 was chosen ($1/b^{1/r}$ = 0.67 and probability of having 0.8 pairs of 99.7%).

The dataset used is 20 documents in a .txt format extracted from:
https://github.com/stonecoldnicole/Plagiarism_Detection/tree/master/data

Instructions to run the code

We provide the following folders and files:

- **src/HW1.ipynb:** A Spark notebook containing code for finding similarity between documents.
- **data/*:** txt files corresponding to the documents to compare.

Results

After using 20 different documents of the set (*_taska.txt) to test our implemented program, the results for document similarity are available on the notebook. Following the highest similarity values:

Pairs	Jaccard similarity	Signatures comparison
(g0pE_taska.txt, g4pC_taska.txt)	0.8371958285052143	0.842

(g0pE_taska.txt, orig_taska.txt)	0.8897228637413395	0.9
(g4pC_taska.txt, orig_taska.txt)	0.899188876013905	0.916

From the LSH function the following pairs were returned:

1. (g0pE_taska.txt,g4pC_taska.txt)
2. (g0pE_taska.txt,orig_taska.txt)
3. (g3pC_taska.txt,g0pD_taska.txt) FP
4. (g4pC_taska.txt,orig_taska.txt)

The output contains all the pairs with similarity above 0.8 and one FP.