2017

Homework#1

Professor: Markov Nikolay Vladimirovich

BDS Master Program

Author: Rashid Ali

Text similarity task

The four documents were taken related to topic “Pakistan and Pakistani famous people” from different Wikipedia and put into folder “text”. Each text file contains almost 1000 words. I tried to perform this task through: textresue package and tm package of R.

The steps for finding similarity using textresue package are:

1. The function “minhash_generator (n =200)” is used to generate random Minhash values.
2. The function “TextReuseCorpus” is used to tokenize the documents into shingles and the value of k=3. And Minhash key value pairs of these shingles are generated.
3. The comparison of every documents Minhash keys are done using function pairwise comparison and results are shown on Console.
4. The results are following when k=3 and n=200 (parameter of random number generated).

doc1 doc2 doc3 doc4

```
                    Jinnah 2 Muhammad Ali Jinnah Pakistan_text_1 sharmeen obaid
Jinnah 2                  NA          0.01852573      0.01083172    0.002751452
Muhammad Ali Jinnah       NA                  NA      0.01276850    0.002017756
Pakistan_text_1           NA                  NA              NA    0.002797594
sharmeen obaid            NA                  NA              NA             NA
```

The k=3 is chosen because according [1], k should be equal to 3 or 4 for small documents and results shows that documents do have much similar text. But, the doc1 and doc2 are relatively more similar. When k=4 and k=5 are chosen, the decrease in similarity of doc1 and doc2 is noticed.

Reference:

[1] Phillips J. (December 25, 2016) Jaccard Similarity and Shingling. Retrieved from https://www.cs.utah.edu/~jeffp/teaching/cs5955/L4-Jaccard+Shingle.pdf