Homework 1
Finding Similar Items: Textually Similar Documents

Denys Tykhoglo

Solution

For this homework, I decided to compare similarity between 20 emails from the Ernon email dataset. I took 10 first emails written by John Arnold and 10 first emails by Phillip Allen (both people were Ernon employees). The emails are attached to the homework and are located in the *emails* directory. Each email file is titled as *<sender>-<number>*.

The solution computes Jaccard similarity for each pair of emails (it is easy to count that there are 190 pairs in total). For this purpose, a sorted set of hashed shingles is created for each email. After that, for each email the solution prints out a list of similar emails in terms of Jaccard similarity and the corresponding similarity values.

Then the solution computes the signature-based similarities. For this purpose, it creates a signature matrix representation that for each shingle stores a list of emails where it appears. After that, several random hash functions are created in format $|(a * x + b)| \: \% \: number\_of\_shingles$, where a and b are random integers and x is the sequential number of a shingle. The signature matrix is calculated using the algorithm provided during the lecture. After all min-hash-based similarity values are determined, the similar pairs are printed out the same way as for the Jaccard similarity case.

How to run

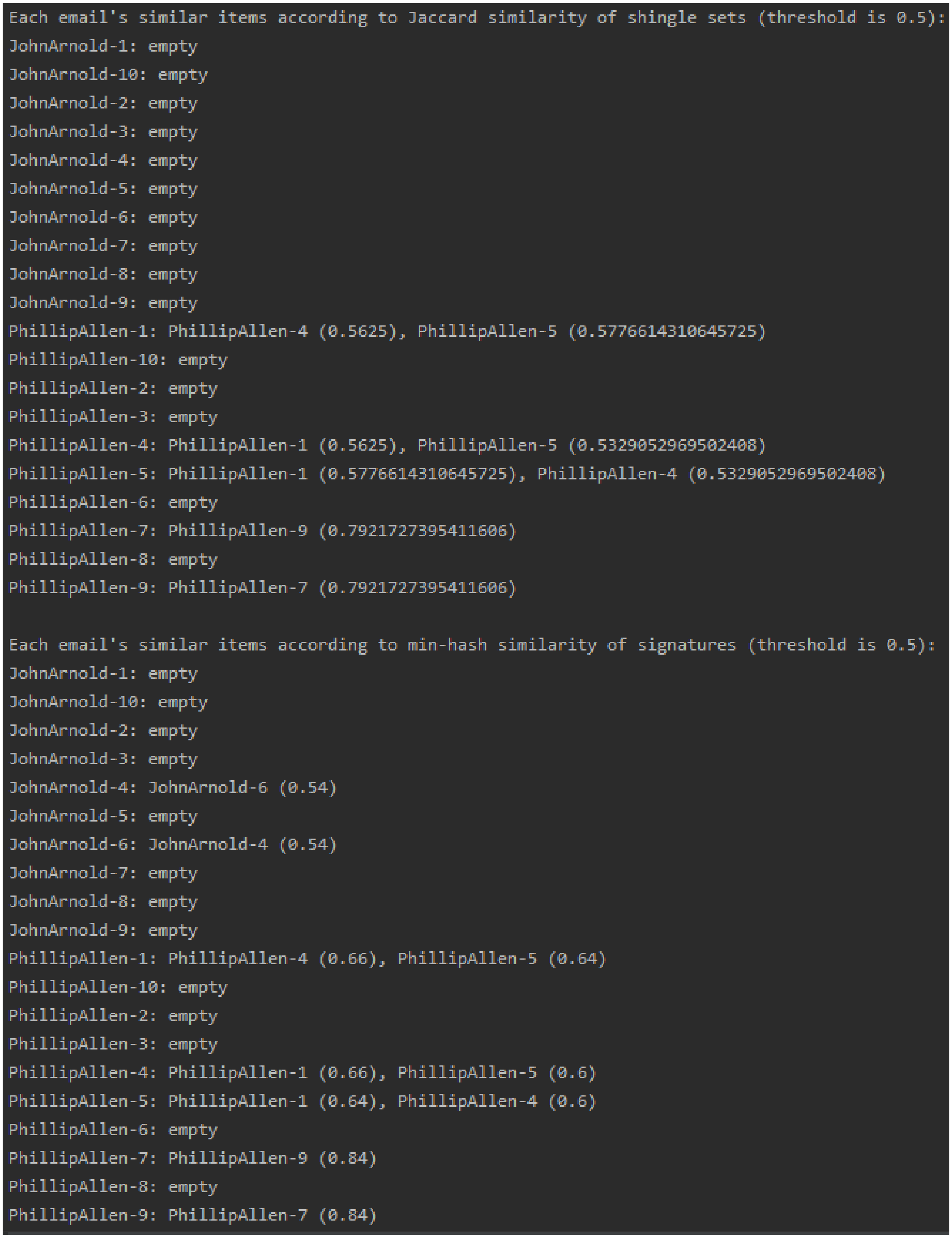
In order to run the solution it is only necessary to compile the attached *TextSimilarity.java* file. The required console parameters are the following:

1. The path to directory with email files (string). The default path is *src\main\resources.*
2. The shingle length (integer number). The default value is *5* as it is suitable for dealing with emails.
3. The threshold for Jaccard similarity between two sets of shingles (decimal number). The default value is *0.5*.
4. The threshold for min-hash similarity between two sets signature vectors (decimal number). The default value is *0.5*.
5. The length of signature that is produced after min-hashing (integer number). The default value is *100*.

Results

Figure 1 demonstrates the screenshot of the output when the solution is executed with the default parameters. We see that the outcomes are almost equivalent for Jaccard similarity and for min-hash signature similarity. It is also visible that all pairs of similar documents belong to the same author. This fact is even clearer when decreasing the threshold.

Note that while Jaccard similarity produces the same result during each run, the result of min-hash signature comparison is varying from run to run due to the random hash functions used to compute the signatures.



```
Each email's similar items according to Jaccard similarity of shingle sets (threshold is 0.5):
JohnArnold-1: empty
JohnArnold-10: empty
JohnArnold-2: empty
JohnArnold-3: empty
JohnArnold-4: empty
JohnArnold-5: empty
JohnArnold-6: empty
JohnArnold-7: empty
JohnArnold-8: empty
JohnArnold-9: empty
PhillipAllen-1: PhillipAllen-4 (0.5625), PhillipAllen-5 (0.5776614310645725)
PhillipAllen-10: empty
PhillipAllen-2: empty
PhillipAllen-3: empty
PhillipAllen-4: PhillipAllen-1 (0.5625), PhillipAllen-5 (0.5329052969502408)
PhillipAllen-5: PhillipAllen-1 (0.5776614310645725), PhillipAllen-4 (0.5329052969502408)
PhillipAllen-6: empty
PhillipAllen-7: PhillipAllen-9 (0.7921727395411606)
PhillipAllen-8: empty
PhillipAllen-9: PhillipAllen-7 (0.7921727395411606)

Each email's similar items according to min-hash similarity of signatures (threshold is 0.5):
JohnArnold-1: empty
JohnArnold-10: empty
JohnArnold-2: empty
JohnArnold-3: empty
JohnArnold-4: JohnArnold-6 (0.54)
JohnArnold-5: empty
JohnArnold-6: JohnArnold-4 (0.54)
JohnArnold-7: empty
JohnArnold-8: empty
JohnArnold-9: empty
PhillipAllen-1: PhillipAllen-4 (0.66), PhillipAllen-5 (0.64)
PhillipAllen-10: empty
PhillipAllen-2: empty
PhillipAllen-3: empty
PhillipAllen-4: PhillipAllen-1 (0.66), PhillipAllen-5 (0.6)
PhillipAllen-5: PhillipAllen-1 (0.64), PhillipAllen-4 (0.6)
PhillipAllen-6: empty
PhillipAllen-7: PhillipAllen-9 (0.84)
PhillipAllen-8: empty
PhillipAllen-9: PhillipAllen-7 (0.84)
```

Figure 1. Printed result of the execution