CSE 484 HW02 REPORT

SPRING

Eyyüp AYDIN – 131044038

For document categorization, Rocchio and knn method developed.

In the news data set, there are roughly 15000 different words (word length >= 2 and <= 5).

In training phase, tfidf vectors are created for each training article and saved.

In classification phase, given article's vector is compared with training vectors with cosine similarity.

While training, Java's parallel streams used to decrease the training time. TextCategorizator class serialized to be used as a model file later.

For this data set, average training time is 190 seconds, 3.1 minutes.

To test the model, a test set generated randomly from the news article data set. In this test set, there are 58 different articles (%5 of original data set).

The results are shown below.

	Rocchio	**knn 3**	**knn 5**
Test 1	45	46	42
Test 2	34	33	32
Test 3	36	40	32

Here, the numbers represent the correctly classified articles (out of 58) by the corresponding method.

Sample usage of the program is below.

```
Enter file path: sample/ekonomi/1.txt
Enter method (r or k): k
Enter k: 3
Assigned class: magazin
Number of matches in neighborhood is: 2
Continue? > y

Enter file path: sample/ekonomi/1.txt
Enter method (r or k): k
Enter k: 5
Assigned class: ekonomi
Number of matches in neighborhood is: 4
Continue? > y

Enter file path: sample/saglik/a.txt
Enter method (r or k): r
Assigned class: ekonomi
Similarity score is: 0.2996215447142985
Continue? > y

Enter file path: sample/saglik/303.txt
Enter method (r or k): r
Assigned class: siyasi
Similarity score is: 0.12470923216559016
Continue? > y

Enter file path: sample/siyasi/a.txt
Enter method (r or k): r
Assigned class: siyasi
Similarity score is: 0.09985912202684412
Continue? > n

Process finished with exit code 0
```