CSE 484 HW03 REPORT

2017 SPRING

Eyyüp AYDIN – 131044038

For document categorization, Turkish Wikipedia archive trained with word2vec and vectors for each word obtained. Then, Rocchio and knn methods are developed using cosine similarity metric with the vectors that created with the aforementioned word vectors. To create a vector for document, I implemented the three different approaches that mentioned in the homework PDF (average, minimum and maximum).

In training phase, I calculated the AVG, MIN and MAX vectors of each document with word vectors. The calculated vectors, their assigned classes and word vectors are serialized to file as a model file.

In classification phase, the given document's vector (created with the user choice method) is compared with training vectors with cosine similarity.

While training, Java's parallel streams used to decrease the training time. For this data set, average training time is 12 seconds.

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). For this report, I generated 6 different training and test sets randomly.

The results are shown below.

Table 1 Performance results in terms of number of correctly classified documents out of 58.

Tests	**Rocchio**			**knn 3**			**knn 5**		
	avg	***max***	***min***	***avg***	***max***	***min***	***avg***	***max***	***min***
Test 1	49	37	31	48	29	33	47	37	36
Test 2	51	34	25	52	38	29	52	36	30
Test 3	52	31	29	51	31	29	52	33	34
Test 4	53	32	34	54	33	31	54	39	37
Test 5	55	25	25	55	30	27	56	30	33
Test 6	53	41	30	53	44	36	53	46	35
AVG	52,17	33,33	29,00	52,17	34,17	30,83	52,33	36,83	34,17

Table 2 Performance results in terms of percentage.

Tests	Rocchio			knn 3			knn 5		
	avg	***max***	***min***	***avg***	***max***	***min***	***avg***	***max***	***min***
Test 1	84,48	63,79	53,45	82,76	50,00	56,90	81,03	63,79	62,07
Test 2	87,93	58,62	43,10	89,66	65,52	50,00	89,66	62,07	51,72
Test 3	89,66	53,45	50,00	87,93	53,45	50,00	89,66	56,90	58,62
Test 4	91,38	55,17	58,62	93,10	56,90	53,45	93,10	67,24	63,79
Test 5	94,83	43,10	43,10	94,83	51,72	46,55	96,55	51,72	56,90
Test 6	91,38	70,69	51,72	91,38	75,86	62,07	91,38	79,31	60,34
AVG	89,94	57,47	50,00	89,94	58,91	53,16	90,23	63,51	58,91

The trained model file is too big (175 MB) to add moodle. Therefore I added the model file to the my drive:
https://drive.google.com/file/d/0BzM0IGrAiafJRVZNd19Uc0xFdVE/view?usp=sharing