GTU Department of Computer Engineering
CSE 554/484
Spring 2017
Homework 03
Due date: Jun 5th 2017 by the Final exam date
No Late Submissions Accepted

Text Categorization using word embeddings

We will use embedding methods for text categorization. First download word2vec (https://github.com/tmikolov/word2vec), compile and run it on a Turkish text that we will provide. The text and the vectors are at
https://www.dropbox.com/sh/umigczctv1y50ss/AADUYl9YXbaqhCnEw4uUZi_5a?dl=0

Word2vec is an unsupervised method that can assign semantic vector representations for each word in the corpora.
After you obtain your word embeddings, run word similarity and analogy demo programs that comes with the word2vec program.

Word2vect assigns only vectors to words. We want to assign vectors to documents so that we can do text categorization using these vectors. Your job will be very similar to HW2.

1. Download news articles from 5 categories from http://www.kemik.yildiz.edu.tr/data/File/1150haber.rar
2. Use 95% of the news set to train your text categorizer for the method of
    - Rocchio
    - K-nearest neighbor (K=3 and K=5)
3. Test your system with the remaining 5% of the set.

In order to represent each document by a vector, you will use these methods

1. AVG: Each document is a vector that is calculated by the average of the word vectors that it contains
2. MAX: Each document is a vector that is calculated by taking the maximum of each dimension of the word vectors that it contains
3. MIN: same as MAX, only takes MIN.

Write a report that lists your findings about the performance of your system on different categorization methods and K values and AVG,MIN, MAX methods. You should write clean tables with results of your work.
One sample run of your HW would be like

Enter the file name for categorization: abc.txt
Enter the classificaiton method: Rocchio
Enter Method: AVG

Assigned class is 3, true class is XXX