GTU Department of Computer Engineering
CSE 554/484
Spring 2017
Homework 02
Due date: May 22th 2017

Text Categorization

We will categorize texts into given classes in this homework. We will test a few methods for the categorization tasks.

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.

You will use TF-IDF vectors and cosine similarity for comparing vectors. While counting words, you will take the first 5 characters of the words.

Write a report that lists your findings about the performance of your system on different categorization methods and K values. One sample run of your HW would be like

Enter the file name for categorization: abc.txt
Enter the method: Rocchio
Assigned class is 3, similarity score is XXXX