Gebze Technical University
Department of Computer Engineering
CSE 654 / 484
Fall 2021

Homework 01
Due date: Oct 5th 2021

In this homework we will use vector representation of words. Here are the steps of the homework

1. Download the standard textbooks from the Ministry of Education (http://aok.meb.gov.tr/kitap/) for at least 20 textbooks (literature, history, sociology, etc.) Convert them to text documents. You should have at least 2000 pages of Turkish text. More is better.
2. Download the word2vect source code https://github.com/tmikolov/word2vec . Compile and run it on the whole textbook set. Run a few distance and analogy demos to see if the vectors are fine.
3. As we know from the lectures, word2Vect produces vectors only for the words seen before. We will try to come up with a new way of producing vectors. Here are the steps for your new algorithm
    a. Find an open source program to divide Turkish words into syllables. If we are given Turkish words "okula gitmek", the syllables are (o, ku, la, git, mek).
    b. Convert all the words in your data sets into syllables and run word2Vect on this new set. Each syllable will have its own vector.
    c. To calculate the vector for a given Turkish word, divide the word into syllables and then add the vectors of each word. This way you can find the word representation for any word.
4. Run analogy and similarity experiments for both vector sets (word2vect and syllable based) for at least 30 Turkish examples. Measure the accuracy using formulas we learned in the lectures.

Prepare your report and submit it to the Teams page. You may use any programming language for the implementation. The word2vec source code in C language but you will use it to produce only the word vectors.

Notes

1. You will demo your homework result online
2. Your report should be very clear about the formula for testing calculation
3. Your report is as important as the homework itself and it is not optional