Gebze Technical University
Department of Computer Engineering
CSE 654 / 484
Fall 2022

Homework 03
Due date: Dec 30th 2022

In this homework we will assign vectors for each N-gram of Turkish syllables and measure how well it captures the Turkish morphological derivations.

Follow the steps below for the homework and for your homework report

1. Download the Turkish Wikipedia dump https://www.kaggle.com/mustfkeskin/turkish-wikipedia-dump
2. Separate each word into its syllables using a program that you can find off the net or implement.
3. Calculate the 1-Gram, 2-Gram, and 3-Gram tables for this set using 95% of the set (If the set is too large, you may use a subset).
4. Assign vectors for each of the 1-gram, 2-gram and 3-gram syllables. You can consider a syllable N-gram as a word and run standard word2vect algorithm to assign vectors for each syllable N-gram.
5. Find the word similarity tests that we have used in HW1 for Turkish morphology suffixes. For example, the vectors of 2-grams "la-rı" and "la-rım" should be very similar because they may be observed in words like "o-da-la-rı" and "o-da-la-rım" which might be parsed as **oda<N><pl><p3s>** and **oda<N><pl><p1s>,** respectively**.**
6. Run the morphology analogy tests between the words. A classic word analogy example is "man is to woman as king is to queen" which is shown as "man:woman :: king:queen". Come up with such Turkish morphology analogy examples as "odaları:odalarım :: balonları:balonlarım". List many examples where this system works and list examples where this system fails.

Prepare your report and submit it to the Teams page. You may use any programming language for the implementation. You may also use N-gram library software to calculate the N-Grams efficiently. Please indicate which library you have used.

Notes

1. Convert all the letters to small case letters first. You may convert all Turkish characters to English ones. For example, ş -> s and ğ -> g
2. Do not forget to include punctuation marks (end of sentences and space characters as syllables in your N-grams. Just lower case letters and space character will be enough.