Gebze Technical University
Department of Computer Engineering
CSE 654 / 484
Fall 2023

Homework 02
Due date: Dec 11th 2023

In this homework we will develop a statistical language model of Turkish that will use N-grams of Turkish syllables.

Follow the steps below for the first part of 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). Note that your N-gram tables will be mostly empty, so you need to use smart ways of storing this information. You also need to use smoothing, which will be GT smoothing that we have learned in the class.
4. Calculate perplexity of the 1-Gram to 3-Gram models using the chain rule with the Markov assumption for each sentence. You will use the remaining 5% of the set for these calculations. Make a table of your findings in your report and explain your results.
5. Produce random sentences for each N-Gram model. You should pick one of the best 5 syllables randomly. Include these random sentences in your report and discuss the produced sentences.

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. Do not forget to use logarithm of the multiplication of the chain rule formula
2. 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
3. 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.