Gebze Technical University
Department of Computer Engineering
CSE 654 / 484
Fall 2020

Homework 02
Due date: Nov 30th 2020

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

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

1. Download the Turkish Wikipedia dump https://www.dropbox.com/sh/umigczctv1y50ss/AADUYl9YXbaqhCnEw4uUZi_5a?dl=0
2. Calculate the 1-Gram, 2-Gram, 3-Gram, 4-Gram and 5-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.
3. Calculate perplexity of the 1-Gram to 5-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.
4. Produce random sentences for each N-Gram model. You should pick one of the best 5 letters randomly. Include these random sentences in your report and discuss the produced sentences.

Prepare your report and submit it to the moodle 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.
3. Do not include any punctuation marks in your N-grams. Just lower case letters and space character will be enough.
4. You will demo your homework result online