# CSE484
# HOMEWORK 2 REPORT

Yağız Hakkı Aydın
1901042612

Firstly I read xml file by using BeautifulSoup library to a single string.

Then replaced unnecessary characters,replaced uppercase letters with lowercase ones.

Then within text, I replaced “.” with “ yyy ” string and “ ” with “ xxx ” strings.So when tokenizing to syllables,dots and spaces can be read.

Then I used TurkishNLP library to tokenize text to syllables but before that, because of library giving server error, I used function from library by modifying to not download turkish words data by using pickle library.

I created syllables 2d list which every sublist of list contains a word’s syllables.

I flattened syllables list,so flattened_syllables kept all syllables as list of strings.

I created a unique_syllables which contains syllables too but removed duplicate ones.

I created unigram table as dictionary,so it keeps table in unigram_table[syllable] format.
I created bigram table with same logic,so it keeps in bigram_table[syllable1][syllable2] format.
I applied same logic also to create trigram table.

Then I applied good-turing smoothing to unigram table by using formula.

I splitted flattened_syllables list.

Then I generated sentences list which includes lists those keep strings as syllables for each sentence.To do this,i read flattened_syllables and everytime current syllable is read as “.”,ending of sentence is detected.

Then I generated random sentences by using unigram and bigram tables.
To do this, I choose most frequent 5 syllable from unigram list and randomly selected one of them as first syllable of generated sentence.

Generated sentences from both bigram and unigram models were meaningless.