**GTU Department of Computer Engineering**

**CSE 484/654 Natural Language Processing**

**Fall 2021 - Homework 2 Report**

**Akif KARTAL**

**171044098**

**1) Problem Definition**

The problem is to develop a statistical language model of Turkish that will use N-grams of Turkish syllables.

**2) Solution**

The homework was finished as expected in homework pdf file. Solution steps are following;

**2.1) Creating Corpus**

The given turkish-wikipedia-dump text was to big(441MB) to test therefore, I used small portion of it which is 6,39 MB of data that 6.065MB(95% of the set) for ngrams and 0. 334MB(5% of the set) test data.

**2.2) Dividing Turkish words into syllables**

First, I convert all the letters to small case letters.

As in hw1, I used following program to divide turkish words into syllables.

<https://github.com/MeteHanC/turkishnlp>

**Output:**

1-line syllabled corpus text;

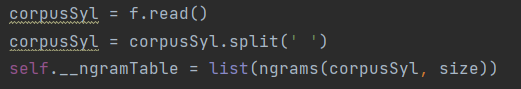


**2.3) Calculating N-grams**

**2.3.1 Creating N-gram tables**

In order to create N gram tables I have used **python ngram library from from nltk.util library.**

**Simple Code:**



**2.3.2 Counting Words**

In order to create count words in N gram table, I have used **collections.Counter library.**

**Simple Code:**

Text

Description automatically generated

**Output (2-gram counts):**

****

**2.3.3 GT Smoothing**

In order to apply Good-Turing Smoothing to the probabilities, we will use following formulas;

Text

Description automatically generated with medium confidence

As you can see in order to calculate probabilities, we need a **count table** that hold number of occurrences of that number. Count table will be implemented by using **1-grams.**

**Count table output (N[x] is the frequency-of-frequency-x):**



**Calculating GT Smoothing by using Count Table**

**Full Formula**

Text

Description automatically generated with low confidence

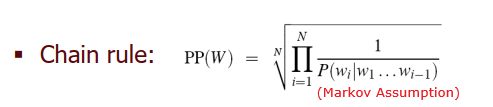
Here, I will take k = c

**After GT smoothing probability table**

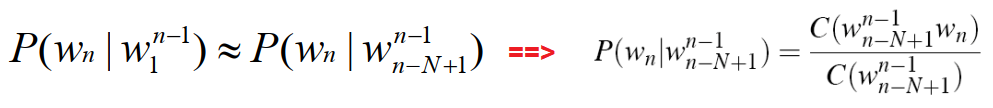


**2.4) Calculating perplexity with the Markov assumption**

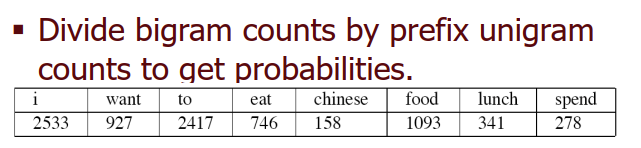
Perperlixty formula with the Markov assumption;

****

Markov assumption and calculating probabilities

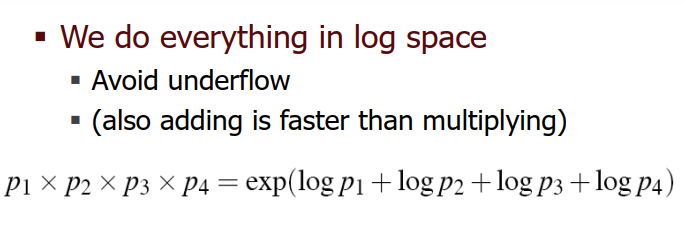


Calculating probabilities



**Using logarithm of the multiplication of the chain rule formula**

Following formula will be used while calculating probabilities



**Putting all of these together and getting result**