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 and 0.366 MB test data (5% of the set).

2.2) Dividing Turkish words into syllables

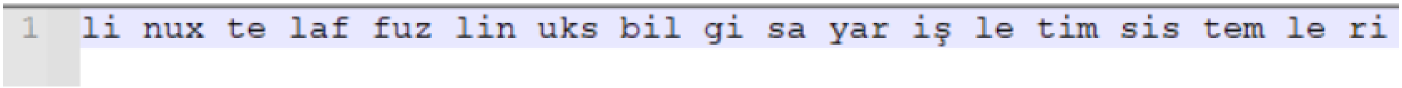
First, I convert all the letters to small case letters.

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

https://github.com/MeteHanC/turkishnlp

Output:

1-line syllabled corpus text;

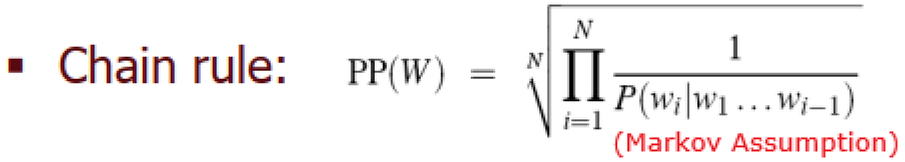


```
li nux te laf fuz lin uks bil gi sa yar iş le tim sis tem le ri
```

2.3) Calculating N-grams

2.4) Calculating perplexity with the Markov assumption

Perperlixty formula with the Markov assumption;



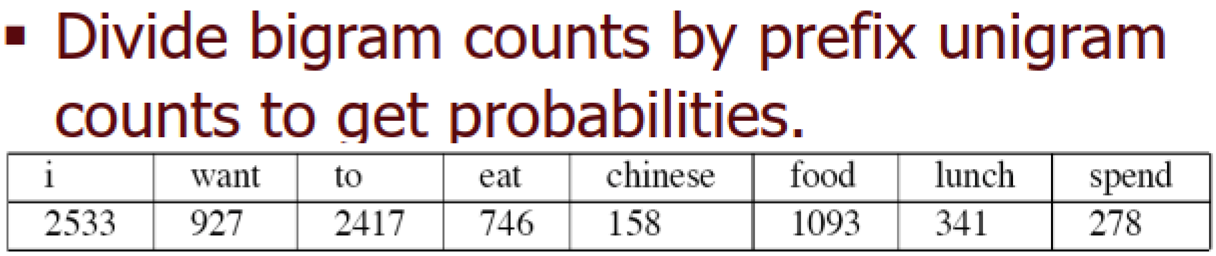
- Chain rule: $\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$

(Markov Assumption)

Markov assumption and calculating probabilities

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1}) \quad \text{==>} \quad P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}$$

Calculating probabilities



- Divide bigram counts by prefix unigram counts to get probabilities.

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

Using logarithm of the multiplication of the chain rule formula

Following formula will be used while calculating probabilities

$$p_1 \times p_2 \times p_3 \times p_4 = \exp(\log p_1 + \log p_2 + \log p_3 + \log p_4)$$

Putting all of these together and getting result