Merve Horuz

1801042651

First, I downloaded the Turkish Wikipedia dump https://www.kaggle.com/mustfkeskin/turkish-wikipedia-dump . But I used subset of the wiki because the set is too large.

I separated each word into its syllables:

```
boz kır ge le ne ğin den ge len on lu teş ki la tı kul la na rak me ri tok ra t
han ın bü yük bir as ker o la rak ün ka zan ma sı nın te me lin de, kur du ğu p
tı na ver di ği bü yük de ğer ö nem li bir yer tu tar se fer le ri so nu cun da
da kat le dil miş ti an cak cen giz han ya sa sı a dı i le me tin leş ti ri len
lar, dok tor lar, bel li bil gi be ce ri si o lan e ği tim li ki şi ler ve her
sun a ra la rın da bir ay rım ya pıl mak sı zın say gı gös te ril me si ve ver
giz han, hal kı nın ya zı ya sa hip ol ma sı nı sağ la mak i çin uy gur lar dan
mış ve mo ğol ca i çin uy gur al fa be si ni u yar la ta rak bu nu ço cuk la rı
de ki tan gut lar ü ze ri ne çık tı ğı se fer es na sın da ra hat sız la na rak
gü nü müz de rus ya ha riç tüm ül ke ler den da ha ge niş top rak lar ü ze ri n
la rı ve to run la rı dö ne min de da ha da ge niş le ye rek, in san lık ta ri
im pa ra tor luk ha li ne gel di cen giz han ın şe ce re si ya rı mi to lo jik
rı nın ve ken di si nin do ğuş ef sa ne si, mo ğol mi to lo ji si nin ö nem li
cen giz han ın ya şa dı ğı dö ne me en ya kın o la nı şa ma nizm et ki le ri ni
lın mış o lan mo ğol la rın giz li ta ri hi ad lı e se re gö re cen giz han dan
ef sa ne vi bü yük an ne si o la rak ka bul e dil miş tir mo ğol la rın giz li
kal dık tan son ra ev len me di ği hal de üç oğ lu da ha ol muş tur cen giz han
dan bo don car ad lı en kü çük o la nı nın so yun dan gel mek te dir a lan go y
lik te ni run ya ni ı şı ğın ço cuk la rı a dı ve ri len bir yı ğın bo yu il gi
o la rak ka bul e dil miş tir ni run boy la rın dan ön ce lik le cen giz han ın
lar, der ben ler, sal c iut lar ve baş ka bir kaç boy da ha sa yı la bi lir yı
ka bul, bü tün mo ğol la rın ilk li de ri o la rak han un va nı nı al mış tır c
dur ka bul han ve o nun ha le fi am ba kay han za ma nın da mo ğol lar, çin de
dar kuv vet len se ler de ta tar lar, çinl le ri hoş nut et mek i çin am ba kay
bir şe kil de, tah ta e şek şek li de nen bir du ru ma so ku la rak çar mı ha g
ku t ua, bu ha ka re te çin ü ze ri ne ve ta tar la ra bir di zi sal dı rı dü z
ğol her kü lü un va nı nı ka zan dı fa kat, yı lın da, de tay la rı bi lin me y
ha ne da nı, mo ğol la rı he zi me te uğ rat tı mo ğol lar bir sü re kar ma şa
şık hal de ki mo ğol lar ın i çe ri sin de ki ö nem siz li der ler den bi ri o
lar ku ra rak mo ğol la rı güç len dir me ye ça lış tı mo ğol la rın ba tı kom
i di ke r ait ler yıl dan be ri nas tu ri hı ris ti yan dı hı ris ti yan ke r a
lın da iç me se le ler so nu cu tah tı nı kay bet miş ti mo ğol la rın li de ri
ni den e le ge çir me si i çin yar dım et ti tuğ rul ve ye sü gey an da i le ka
```

Unigrams

```
zeroday@zeroday-Lenovo-V330-15IKB:~/Desktop/cse654_nlp/hw1$ python3 n_grams_table.py
################## N-GRAMS #####################

### uni-grams ###
('boz',)
('kır',)
('ge',)
('le',)
('ne',)
('ğin',)
('den',)
('ge',)
('len',)
('on',)
('lu',)
('teş',)
('ki',)
('la',)
('tı',)
('kul',)
('la',)
('na',)
('rak',)
('me',)
('ri',)
('tok',)
('ra',)
```

Bigrams

```
● zeroday@zeroday-Lenovo-V330-15IKB:~/Desktop/cse654_nlp/hw1$ python3 n_grams_table.py
################## N-GRAMS #####################

### bi-grams ###
('boz', 'kır')
('kır', 'ge')
('ge', 'le')
('le', 'ne')
('ne', 'ğin')
('ğin', 'den')
('den', 'ge')
('ge', 'len')
('len', 'on')
('on', 'lu')
('lu', 'teş')
('teş', 'ki')
('ki', 'la')
('la', 'tı')
('tı', 'kul')
('kul', 'la')
('la', 'na')
('na', 'rak')
('rak', 'me')
('me', 'ri')
('ri', 'tok')
('tok', 'ra')
```

Trigrams

```
zeroday@zeroday-Lenovo-V330-15IKB:~/Desktop/cse654_nlp/hw1$ python3 n_grams_table.py
################## N-GRAMS #####################

### tri-grams ###
('boz', 'kır', 'ge')
('kır', 'ge', 'le')
('ge', 'le', 'ne')
('le', 'ne', 'ğin')
('ne', 'ğin', 'den')
('ğin', 'den', 'ge')
('den', 'ge', 'len')
('ge', 'len', 'on')
('len', 'on', 'lu')
('on', 'lu', 'teş')
('lu', 'teş', 'ki')
('teş', 'ki', 'la')
('ki', 'la', 'tı')
('la', 'tı', 'kul')
('tı', 'kul', 'la')
('kul', 'la', 'na')
('la', 'na', 'rak')
('na', 'rak', 'me')
('rak', 'me', 'ri')
('me', 'ri', 'tok')
('ri', 'tok', 'ra')
('tok', 'ra', 'tik')
('ra', 'tik', 'liy')
('tik', 'liy', 'ka')
('liy', 'ka', 'ta')
('ka', 'ta', 'bağ')
('ta', 'bağ', 'lı')
```

Now, calculated frequencies for each of the syllables. And stored in python dictionary.

Frequencies for unigrams

```
### frequencies each of the grams ###
('boz',) 1
('kır',) 1
('ge',) 4
('le',) 6
('ne',) 3
('ğin',) 1
('den',) 5
('len',) 2
('on',) 1
('lu',) 1
('teş',) 2
('ki',) 4
('la',) 13
('tı',) 5
('kul',) 1
('na',) 7
('rak',) 5
('me',) 5
('ri',) 6
('tok',) 1
('ra',) 5
('tik',) 1
('liy',) 1
('ka',) 4
('ta',) 3
('bağ',) 1
('lı',) 1
('bir',) 4
('or',) 1
```

Frequencies for bigrams

```
### frequencies each of the grams ###
('boz', 'kır') 1
('kır', 'ge') 1
('ge', 'le') 1
('le', 'ne') 1
('ne', 'ğin') 1
('ğin', 'den') 1
('den', 'ge') 1
('ge', 'len') 1
('len', 'on') 1
('on', 'lu') 1
('lu', 'teş') 1
('teş', 'ki') 2
('ki', 'la') 2
('la', 'tı') 2
('tı', 'kul') 1
('kul', 'la') 1
('la', 'na') 2
('na', 'rak') 2
('rak', 'me') 1
('me', 'ri') 1
('ri', 'tok') 1
('tok', 'ra') 1
('ra', 'tik') 1
('tik', 'liy') 1
('liy', 'ka') 1
('ka', 'ta') 1
('ta', 'bağ') 1
('bağ', 'lı') 1
('lı', 'bir') 1
('bir', 'or') 1
('or', 'du') 1
('du', 'mey') 1
('mey', 'da') 1
```

Frequencies for trigrams

```
### frequencies each of the grams ###
('boz', 'kır', 'ge') 1
('kır', 'ge', 'le') 1
('ge', 'le', 'ne') 1
('le', 'ne', 'ğin') 1
('ne', 'ğin', 'den') 1
('ğin', 'den', 'ge') 1
('den', 'ge', 'len') 1
('ge', 'len', 'on') 1
('len', 'on', 'lu') 1
('on', 'lu', 'teş') 1
('lu', 'teş', 'ki') 1
('teş', 'ki', 'la') 2
('ki', 'la', 'tı') 2
('la', 'tı', 'kul') 1
('tı', 'kul', 'la') 1
('kul', 'la', 'na') 1
('la', 'na', 'rak') 2
('na', 'rak', 'me') 1
('rak', 'me', 'ri') 1
('me', 'ri', 'tok') 1
('ri', 'tok', 'ra') 1
('tok', 'ra', 'tik') 1
('ra', 'tik', 'liy') 1
('tik', 'liy', 'ka') 1
('liy', 'ka', 'ta') 1
('ka', 'ta', 'bağ') 1
('ta', 'bağ', 'lı') 1
('bağ', 'lı', 'bir') 1
('lı', 'bir', 'or') 1
('bir', 'or', 'du') 1
```

Then, inserted each of the frequencies to matrix.

Unigram matrix

```
zeroday@zeroday-Lenovo-V330-15IKB:~/Desktop/cse654_nlp/hw1$ python3 n_grams_table.py
1.0  1.0  4.0  6.0  3.0  1.0  5.0  2.0  1.0  1.0  2.0  4.0  13.0  5.0  1.0  7.0  5.0  5.0  6.0  1.0  5.0  1.0  1.0  4.0  3.0  1.0  1.0  4.0  1.0  3.0  1.0  7.0  5.0  1.0  3.0  3.0
3.0  1.0  2.0  2.0  1.0  1.0  5.0  1.0  1.0  3.0  4.0  2.0  4.0  1.0  1.0  2.0  2.0  1.0  5.0  3.0  1.0  5.0  2.0  5.0  2.0  1.0  1.0  1.0  4.0  2.0  1.0  2.0  3.0  1.0  3.0  2.0  4
.0  1.0  2.0  1.0  2.0  2.0  1.0  2.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  4.0  2.0  2.0  1.0  5.0  1.0  1.0  1.0  1.0  2.0  1.0  1.0  8.0  1.0  1.0  1.0  3.0  1.0  1.0  2.0  2.
0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  2.0  1.0  2.0  1.0  1.0  3.0  2.0  4.0  1.0  1.0  1.0  1.0  2.0  1.0  2.0  1.0  4.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  2.0  1.0  1.0
 1.0  1.0  1.0  1.0  1.0  1.0  3.0  1.0  1.0  2.0  3.0  1.0  1.0  1.0  1.0  1.0  1.0  3.0  2.0  2.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  2.0  1.0  1.0  1.0  1.0
 1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  2.0  2.0  1.0  1.0  1.0  1.0  1.0  3.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0
1.0  1.0  1.0  1.0  1.0  1.0

zeroday@zeroday-Lenovo-V330-15IKB:~/Desktop/cse654_nlp/hw1$
```

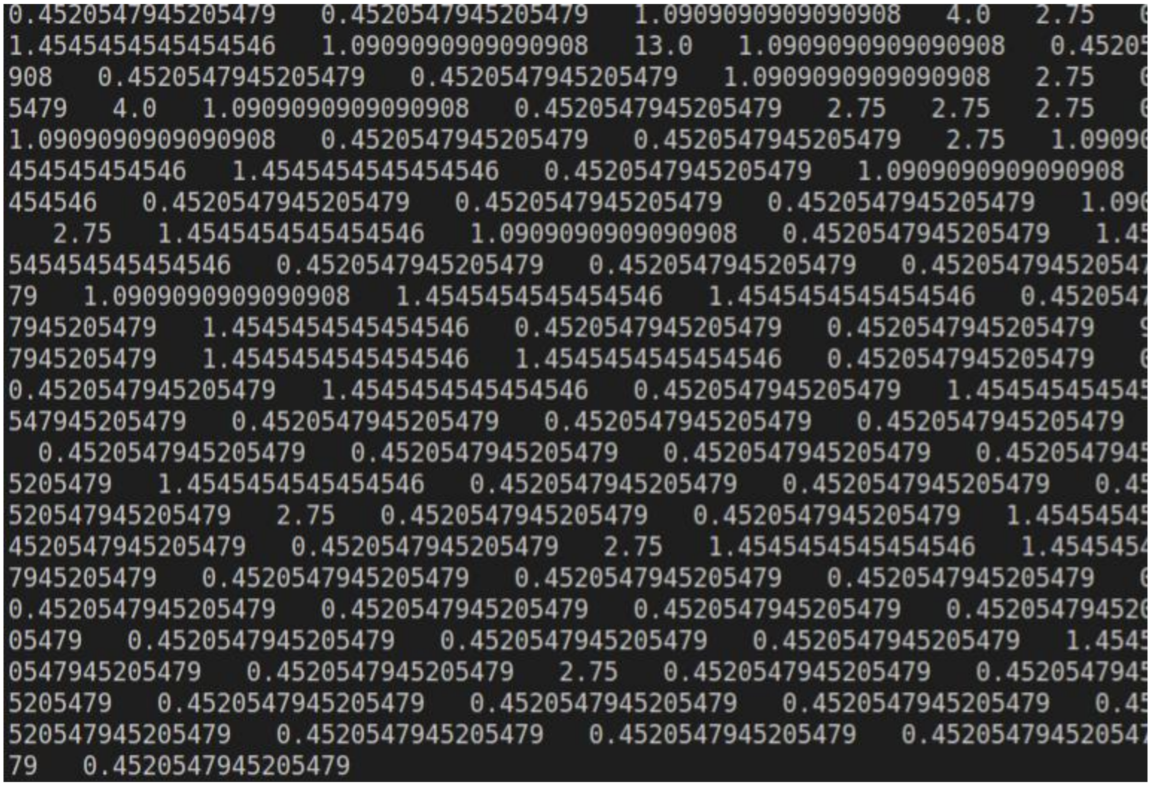
Little piece

```
0  7.0  5.0  5.0  6.0  1.0  5.0  1.0  1
)  2.0  2.0  1.0  5.0  3.0  1.0  5.0  2.
 1.0  4.0  2.0  2.0  1.0  5.0  1.0  1.0
4.0  1.0  1.0  1.0  1.0  2.0  1.0  2.0
.0  1.0  1.0  3.0  2.0  2.0  1.0  1.0
0  1.0  3.0  1.0  1.0  1.0  1.0  1.0  1
```

Then, applied the Good-Turing smoothing (each element is separated with 3 space)

```
zeroday@zeroday-Lenovo-V330-15IKB:~/Desktop/cse654_nlp/hw1$ python3 n_grams_table.py
0.4520547945205479   0.4520547945205479   1.0909090909090908   4.0   2.75   0.4520547945205479   1.0909090909090908   1.4545454545454546   0.4520547945205479   0.4520547945205479
1.4545454545454546   1.0909090909090908   13.0   1.0909090909090908   0.4520547945205479   4.0   1.0909090909090908   1.0909090909090908   4.0   0.4520547945205479   1.0909090909090
908   0.4520547945205479   0.4520547945205479   1.0909090909090908   2.75   0.4520547945205479   0.4520547945205479   1.0909090909090908   0.4520547945205479   2.75   0.452054794520
5479   4.0   1.0909090909090908   0.4520547945205479   2.75   2.75   2.75   0.4520547945205479   1.4545454545454546   1.4545454545454546   0.4520547945205479   0.4520547945205479
1.0909090909090908   0.4520547945205479   0.4520547945205479   2.75   1.0909090909090908   1.4545454545454546   1.0909090909090908   0.4520547945205479   0.4520547945205479   1.4545
454545454546   1.4545454545454546   0.4520547945205479   1.0909090909090908   2.75   0.4520547945205479   1.0909090909090908   1.4545454545454546   1.0909090909090908   1.454545454
454546   0.4520547945205479   0.4520547945205479   0.4520547945205479   1.0909090909090908   1.4545454545454546   0.4520547945205479   1.4545454545454546   2.75   0.4520547945205479
   2.75   1.4545454545454546   1.0909090909090908   0.4520547945205479   1.4545454545454546   0.4520547945205479   1.4545454545454546   1.4545454545454546   0.4520547945205479   1.4
545454545454546   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.45205479452054
79   1.0909090909090908   1.4545454545454546   1.4545454545454546   0.4520547945205479   1.0909090909090908   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.452054
7945205479   1.4545454545454546   0.4520547945205479   0.4520547945205479   9.0   0.4520547945205479   0.4520547945205479   0.4520547945205479   2.75   0.4520547945205479   0.452054
7945205479   1.4545454545454546   1.4545454545454546   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479
0.4520547945205479   1.4545454545454546   0.4520547945205479   1.4545454545454546   0.4520547945205479   0.4520547945205479   2.75   1.4545454545454546   1.0909090909090908   0.4520
547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   1.4545454545454546   0.4520547945205479   1.4545454545454546   0.4520547945205479   1.0909090909090908
  0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.452054794
5205479   1.4545454545454546   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4
520547945205479   2.75   0.4520547945205479   0.4520547945205479   1.4545454545454546   2.75   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.
4520547945205479   0.4520547945205479   2.75   1.4545454545454546   1.4545454545454546   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.452054
7945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   1.4545454545454546   0.4520547945205479
0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.45205479452
05479   0.4520547945205479   0.4520547945205479   0.4520547945205479   1.4545454545454546   1.4545454545454546   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.452
0547945205479   0.4520547945205479   2.75   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.452054794
5205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4
520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.45205479452054
79   0.4520547945205479
```

Little piece



```
0.4520547945205479   0.4520547945205479   1.0909090909090908   4.0   2.75
1.4545454545454546   1.0909090909090908   13.0   1.0909090909090908   0.4520
908   0.4520547945205479   0.4520547945205479   1.0909090909090908   2.75
5479   4.0   1.0909090909090908   0.4520547945205479   2.75   2.75   2.75
1.0909090909090908   0.4520547945205479   0.4520547945205479   2.75   1.0909
454545454546   1.4545454545454546   0.4520547945205479   1.0909090909090908
454546   0.4520547945205479   0.4520547945205479   0.4520547945205479   1.09
   2.75   1.4545454545454546   1.0909090909090908   0.4520547945205479   1.4
545454545454546   0.4520547945205479   0.4520547945205479   0.452054794520547
79   1.0909090909090908   1.4545454545454546   1.4545454545454546   0.452054
7945205479   1.4545454545454546   0.4520547945205479   0.4520547945205479
7945205479   1.4545454545454546   1.4545454545454546   0.4520547945205479
0.4520547945205479   1.4545454545454546   0.4520547945205479   1.4545454545
547945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479
  0.4520547945205479   0.4520547945205479   0.4520547945205479   0.452054794
5205479   1.4545454545454546   0.4520547945205479   0.4520547945205479   0.4
520547945205479   2.75   0.4520547945205479   0.4520547945205479   1.4545454
4520547945205479   0.4520547945205479   2.75   1.4545454545454546   1.454545
7945205479   0.4520547945205479   0.4520547945205479   0.4520547945205479
0.4520547945205479   0.4520547945205479   0.4520547945205479   0.45205479452
05479   0.4520547945205479   0.4520547945205479   0.4520547945205479   1.454
0547945205479   0.4520547945205479   2.75   0.4520547945205479   0.452054794
5205479   0.4520547945205479   0.4520547945205479   0.4520547945205479   0.4
520547945205479   0.4520547945205479   0.4520547945205479   0.452054794520547
79   0.4520547945205479
```

Bigram matrix

A sparse matrix is a matrix in which most of the elements are zero. Sparsity can be a useful property when storing and working with large matrix because it can significantly reduce the amount of memory and computational resources required to represent and manipulate the matrix.

To implement a sparse matrix, one common approach is to use a data structure that only stores the non-zero elements of the matrix, along with their corresponding row and column indices. This allows for efficient storage and manipulation of the matrix, because only the non-zero elements need to be considered.

In bigram matrix, most of the entries are zeros. So, applied the sparse matrix implementation to bigram matrix.

Good-Turing smoothing is a technique used to smooth the probabilities estimated from a dataset in natural language processing and other fields. It is a type of smoothing that considers the fact that the observed frequencies of events in a dataset may not accurately reflect their true underlying probabilities.

Good-Turing smoothing is based on the idea that if an event has been observed a certain number of times in a dataset, then the probability of that event can be estimated by the number of times the event has been observed plus one, divided by the total number of events plus the number of possible events. This formula can be used to smooth the probabilities of events that have been observed, as well as events that have not been observed.

For example, consider a dataset containing the words "the", "cat", "sat", and "on". The observed frequencies of these words in the dataset are 2, 1, 1, and 1, respectively. Using Good-Turing smoothing, we can estimate the probabilities of these words as follows:

The probability of "the" is (2 + 1) / (4 + 4) = 0.375

The probability of "cat" is (1 + 1) / (4 + 4) = 0.125

The probability of "sat" is (1 + 1) / (4 + 4) = 0.125

The probability of "on" is (1 + 1) / (4 + 4) = 0.125

In this example, Good-Turing smoothing has adjusted the probabilities of the words based on their observed frequencies, as well as the number of possible events (in this case, the number of unique words in the dataset). This can help to provide more accurate estimates of the probabilities of events in the dataset and can be particularly useful when working with small or imbalanced datasets.

Before applying the GT smoothing:

```
zeroday@zeroday-Lenovo-V330-15IKB:~/Desktop/cse654_nlp/hw1$ python3 n_grams_table.py
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 2.0, 1.0, 1.0, 2.0, 2.0, 1.0, 1.0, 1.0, 4.0, 1.0,
1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1
.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.
0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 1.0,
 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0,
1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1
.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.
0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]
```

After applying the GT smoothing:


```
[0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1253561253
5612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12
535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.272727272727272
7, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 4.0, 0.125356
12535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536,
0.12535612535612536, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356125356
12536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.125356
12535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536,
0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535
612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.1253
5612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536
, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 2.0, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535
612535612536, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536,
0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.125356125356
12536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1253
5612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536
, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356125
35612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 2.0, 0.2727272727272727, 0.12535612535612536,
 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1253561253
5612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.27
27272727272727, 0.12535612535612536, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536
, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1253561253
5612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12
535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356125356125
36, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612
535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1
2535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612
536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356
12535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536,
0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.27272727272
72727, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1253
5612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536
, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356125
35612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1
2535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612
```



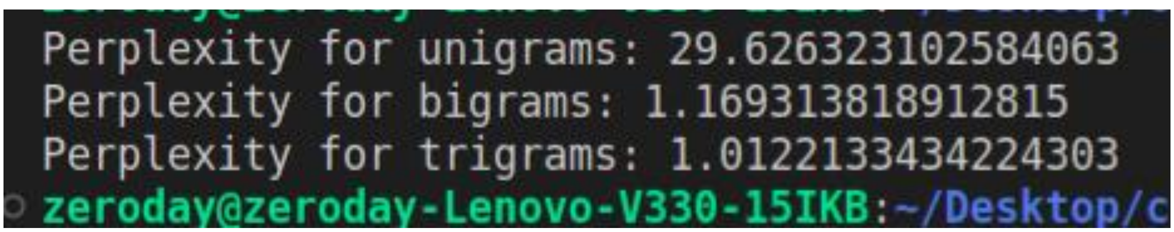
```
[0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12
5612536, 0.12535612535612536, 0.12535612535612536, 0.125356125356125
535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1253561
7, 0.2727272727272727, 0.12535612535612536, 0.12535612535612536, 0.2
12535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535
0.12535612535612536, 0.2727272727272727, 0.12535612535612536, 0.1253
12536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536
12535612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535
0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125
612536, 0.12535612535612536, 0.12535612535612536, 0.1253561253561253
5612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356125
, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1
612535612536, 0.12535612535612536, 0.12535612535612536, 0.2727272727
0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125
12536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536
5612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356125
, 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1
35612536, 0.12535612535612536, 0.12535612535612536, 0.12535612535612
 0.12535612535612536, 0.12535612535612536, 0.12535612535612536, 0.12
5612536, 0.12535612535612536, 0.12535612535612536, 0.125356125356125
27272727272727, 0.12535612535612536, 0.2727272727272727, 0.125356125
, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.12
5612536, 0.12535612535612536, 0.12535612535612536, 0.125356125356125
535612535612536, 0.12535612535612536, 0.12535612535612536, 0.1253561
36, 0.12535612535612536, 0.12535612535612536, 0.2727272727272727, 0.
535612536, 0.12535612535612536, 0.12535612535612536, 0.1253561253561
2535612535612536, 0.12535612535612536, 0.12535612535612536, 0.125356
```

Perplexity

First, a new sentence that was not used in the training set was taken from the wiki and separated into its syllables, then unigram, bigram and trigram models were created. For this sentence, the perplexity was calculated using the Markov assumption chain rule.

Sentence that calculated perplexity:

al tın or da nın par ça lan ma sın dan son ra ce ti su neh ri kı yı la rın da yı lın da ku ru lan ve gü nü müz de ka zak la rın kö ke ni ni o luş tu ran ka zak han lı ğı da cen giz han ın o ğul la rın dan cu ci nin u lu su na bağ lı to ka te mür nes lin den ca ni beg ve ke rey ta ra fın dan ku rul muş tur gü nü müz de ka za kis tan cen giz so yu nun ha la de ğer li ol du ğu ül ke ler den bi ri dir ka zak bi lim a dam la rı yap tık la rı ge ne tik a raş tır ma lar so nu cun da ka za kis tan cum hur baş ka nı nur sul tan na zar ba yev in de soy o la rak cen giz han ın to ru nu ol du ğu nu id d ia et miş ler dir



Perplexity is a measure of how well a probabilistic model can predict a sample. It is commonly used in natural language processing and other fields to evaluate the performance of language models.

Perplexity is calculated as the exponentiated average log-likelihood of a sample and is often used as a measure of the uncertainty of a model's predictions. Specifically, it is defined as follows:

Perplexity = 2 ^ (-1/N * ∑log(P(w)))

A low perplexity indicates that the model has high confidence in its predictions, while a high perplexity indicates that the model is uncertain or confused about the words in the sample. Therefore, perplexity is often used as a metric for evaluating the performance of language models, with lower perplexity indicating better performance.

Using larger n-grams (e.g., 5-grams or 6-grams) in a language model can result in a lower perplexity, because larger n-grams provide the model with more context and information about the words in the sample. This can help the model make more accurate predictions, leading to a lower perplexity.

On the other hand, using smaller n-grams (e.g., 1-grams or 2-grams) in a language model can result in a higher perplexity, because smaller n-grams provide the model with less context and information about the words in the sample. This can make it more difficult for the model to make accurate predictions, leading to a higher perplexity.