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. **Only, I didn't create random sentences.** 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;

```
bil gi sa yar ,iş le tim ,sis tem le ri nin ,en ,te mel
```

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:

```
corpusSyl = f.read()
corpusSyl = corpusSyl.split(' ')
self.__ngramTable = list(ngrams(corpusSyl, size))
```

2.3.2 Counting Words

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

Simple Code:

```
count = collections.Counter(self.__ngramTable)
f1.write(str(dict(count)))
```

Output (2-gram counts):

```
{('bil', 'gi'): 1088, ('gi', 'sa'): 523, ('sa', 'yar'): 481, ('yar', ' '): 580, (' ', 'iş'): 1890,
```

2.3.3 GT Smoothing

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

$$P^*_{GT}(\text{things with zero frequency}) = \frac{N_1}{N} \qquad c^* = \frac{(c+1)N_{c+1}}{N_c}$$

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):

```
{1088: 3, 523: 2, 481: 1, 580: 1, 1890: 1, 660: 3, 159: 14, 1006: 1, 1212: 1,
```

Calculating GT Smoothing by using Count Table

Things with zero frequency (N1/N)

```
c0 = self.__n1 / self.__N
if c not in self.__countTable:
    self.__GtTable[i] = c0
```

GT Smoothing(c*)

```
else:
    nc1 = self.__countTable[c + 1]
    nc = self.__countTable[c]

    res = (((c + 1) * nc1) / nc)
    self.__GtTable[i] = res
```

Before smoothing

```
{('bil', 'gi'): 1088, ('gi', 'sa'): 523,
('sa', 'yar'): 481, ('yar', ' '): 580, (' ', 'iş'): 1890,
```

After GT smoothing new table

```
{('bil', 'gi'): 3.151513570179881, ('gi', 'sa'): 1048.0,
('sa', 'yar'): 964.0, ('yar', ' '): 1162.0, (' ', 'iş'): 16.417388
```

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

```
if self.__GtTable[i] / self.__N <= 0:
    logSum += 0
else:
    logSum += math.log10(self.__GtTable[i] / self.__N)

return math.exp(logSum)
```

Calculating perplexity

```
res = self.__chainWithMarkovAssumption(sentence)
if res != 0:
    root = 1 / res
    return math.pow(root, 1 / self.__ngr)
else:
    return 0
```

2.5) NGrams Tables

2.5.1) 1 Grams

Test Sentence	Perplexity
yosif visaryonoviç 18 aralık 1878'de gori'de dünyaya geldi.	2.8512983712486357e+54
7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı.	1.4500640573963515e+85
10 yaşında rahip okuluna devam etti.	3.5564457859083457e+36
burada gürcü çocuklar rusça eğitim alırlardı.	3.754755388096676e+47
12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi.	4.652762923090376e+125
16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı.	1.2862655963382328e+153
bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi.	4.392662842555145e+96
tiflis'teki rsdip örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi.	2.600540609582902e+136
buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı.	4.65855590647716e+76
mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu.	5.772294507964312e+62
1903 yılında bolşeviklere katıldı.	3.863112666374002e+29
rusya sosyal demokrat işçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti.	3.2358898793866144e+126

2.5.2) 2 Grams

Test Sentence	Perplexity
yosif visaryonoviç 18 aralık 1878'de gori'de dünyaya geldi.	4.985829467378515e+23
7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı.	1.0987990782549322e+39
10 yaşında rahip okuluna devam etti.	2.3126893769021604e+16
burada gürcü çocuklar rusça eğitim alırlardı.	1.2346242945323125e+21
12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi.	2.5318505827018804e+58
16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı.	1.0321184701340856e+45
bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi.	1.0321184701340856e+45
tiflis'teki rsdip örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi.	1.0783229109168015e+64
buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı.	1.0985596467735538e+35
mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu.	4.0381686428472923e+27
1903 yılında bolşeviklere katıldı.	13008562284222.695
rusya sosyal demokrat işçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti.	5.880451890705958e+56

2.5.3) 3 Grams

Test Sentence	Perplexity
yosif visaryonoviç 18 aralık 1878'de gori'de dünyaya geldi.	5308793975562.04
7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı.	1.3120714859497252e+26
10 yaşında rahip okuluna devam etti.	42651036772.641136
burada gürcü çocuklar rusça eğitim alırlardı.	429647142558557.9
12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi.	1.0526585889279279e+37
16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı.	8.053117738493229e+48
bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi.	5.291130302533982e+28
tiflis'teki rsdip örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi.	7.658907009826172e+45
buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı.	5.726410990847328e+21
mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu.	1.1702906354321944e+16
1903 yılında bolşeviklere katıldı.	187468970.43766424
rusya sosyal demokrat işçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti.	2.3370918504405104e+37

2.5.4) 4 Grams

Test Sentence	Perplexity
yosif visaryonoviç 18 aralık 1878'de gori'de dünyaya geldi.	13454547.26640562
7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı.	2.94943062681401e+21
10 yaşında rahip okuluna devam etti.	6881674.54007085
burada gürcü çocuklar rusça eğitim alırlardı.	187161843.46239105
12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi.	2.533889460275275e+26
16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı.	6.290590049226733e+36
bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi.	5.5648896074680115e+22
tiflis'teki rsdip örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi.	5.909894487975789e+26
buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı.	1.0280968746994618e+16
mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu.	10968803168.327368
1903 yılında bolşeviklere katıldı.	651915.0055560797
rusya sosyal demokrat işçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti.	2.910896886930841e+27

2.5.5) 5 Grams

Test Sentence	Perplexity
yosif visaryonoviç 18 aralık 1878'de gori'de dünyaya geldi.	80737.7701062663
7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı.	6203938569794092.0
10 yaşında rahip okuluna devam etti.	10943.20082728885
burada gürcü çocuklar rusça eğitim alırlardı.	61.42212396412726
12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi.	3907.444569351241
16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı.	1.5976621035862554e+17
bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi.	1.2368753043332512e+16
tiflis'teki rsdip örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi.	4205132069612987.0
buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı.	210834088902.4126
mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu.	1832646.9380168936
1903 yılında bolşeviklere katıldı.	5180.735723374886
rusya sosyal demokrat işçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti.	9.906211405669561e+19

*Here, as you can see **perplexity results are very high** and I don't know whether they are correct or not. I just implemented formulas.

References

- Course Slide "slp04 LM and Ngrans.pdf"
- https://github.com/MeteHanC/turkishnlp
- https://www.nltk.org/

How to run

Just install needed python libraries and run 171044098.py file. Check output files after running.