**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;

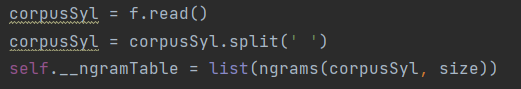


**2.3) Calculating N-grams**

**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**

**Things with zero frequency (N1/N)**

Text

Description automatically generated

**Full Formula**

Text

Description automatically generated with low confidence

Here, I will take k = c

**Directly implemented code from the formula**

Text

Description automatically generated

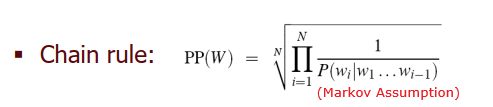
**After GT smoothing probability table**



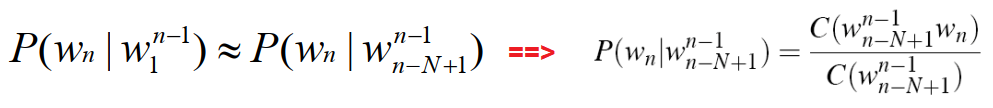


**2.4) Calculating perplexity with the Markov assumption**

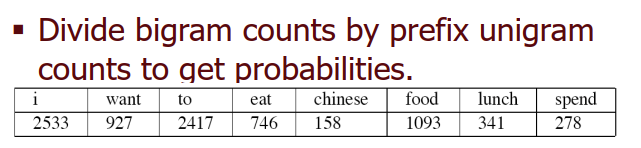
Perperlixty formula with the Markov assumption;

****

Markov assumption and calculating probabilities

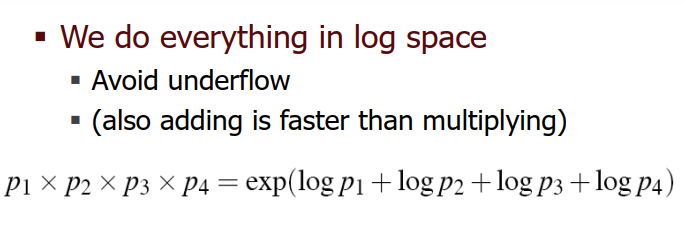


Calculating probabilities



**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**

Text

Description automatically generated



**Calculating perplexity**

Text

Description automatically generated

**2.5) NGrams Tables**

**2.5.1) 1 Grams**

|  |  |
| --- | --- |
| Test Sentence | Perplexity |
| yosif visaryonoviç 18 aralık 1878'de gori’de dünyaya geldi. | 7.065709660608951e+98 |
| 7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı. | 2.5957782331037305e+156 |
| 10 yaşında rahip okuluna devam etti. | 5.552477505431367e+68 |
| burada gürcü çocuklar rusça eğitim alırlardı. | 1.0425720484528676e+85 |
| 12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi. | 1.7372929233687402e+232 |
| 16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı. | 1.5322849499040614e+273 |
| bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi. | 3.5906697780679196e+180 |
| tiflis'teki rsdi̇p örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi. | 1.1688630402690117e+247 |
| buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı. | 2.413738470586876e+139 |
| mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu. | 1.0024355201329885e+115 |
| 1903 yılında bolşeviklere katıldı. | 4.218030878155611e+52 |
| rusya sosyal demokrat i̇şçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti. | 1.9230752032688765e+229 |

**2.5.2) 2 Grams**

|  |  |
| --- | --- |
| Test Sentence | Perplexity |
| yosif visaryonoviç 18 aralık 1878'de gori’de dünyaya geldi. | 5.659786582756472e+33 |
| 7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı. | 6.664675190450381e+53 |
| 10 yaşında rahip okuluna devam etti. | 4.353922120739199e+22 |
| burada gürcü çocuklar rusça eğitim alırlardı. | 2.766973447261439e+27 |
| 12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi. | 1.1530125316524255e+82 |
| 16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı. | 5.238041515469019e+95 |
| bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi. | 8.193032740278134e+65 |
| tiflis'teki rsdi̇p örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi. | 3.210685671345573e+90 |
| buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı. | 9.854330464856806e+49 |
| mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu. | 3.617244047164363e+40 |
| 1903 yılında bolşeviklere katıldı. | 5.871233906520236e+18 |
| rusya sosyal demokrat i̇şçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti. | 5.406426522546649e+75 |

**2.5.3) 3 Grams**

|  |  |
| --- | --- |
| Test Sentence | Perplexity |
| yosif visaryonoviç 18 aralık 1878'de gori’de dünyaya geldi. | 70806777204506.81 |
| 7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı. | 5.88432535752316e+26 |
| 10 yaşında rahip okuluna devam etti. | 86494051930.01068 |
| burada gürcü çocuklar rusça eğitim alırlardı. | 5103410870826.092 |
| 12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi. | 2.3591159623276247e+40 |
| 16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı. | 1.0923797784619242e+49 |
| bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi. | 1.6885774075992623e+32 |
| tiflis'teki rsdi̇p 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. | 6.100887835016103e+23 |
| mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu. | 2.0406375269165003e+19 |
| 1903 yılında bolşeviklere katıldı. | 568312532.8489039 |
| rusya sosyal demokrat i̇şçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti. | 3.4921633676705826e+39 |

**2.5.4) 4 Grams**

|  |  |
| --- | --- |
| Test Sentence | Perplexity |
| yosif visaryonoviç 18 aralık 1878'de gori’de dünyaya geldi. | 20213203.289745927 |
| 7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı. | 7.923030588099656e+19 |
| 10 yaşında rahip okuluna devam etti. | 1282588.9277863775 |
| burada gürcü çocuklar rusça eğitim alırlardı. | 338214.0152171493 |
| 12 yaşına geldiğinde geçirdiği iki at arabası kazası sonucu sol kolu sakatlandı ve hayatı boyunca tam iyileşmedi. | 2.1162830806705702e+26 |
| 16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı. | 3.8737588180021527e+30 |
| bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi. | 4.6318292782281285e+22 |
| tiflis'teki rsdi̇p örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi. | 1.7295999177749725e+28 |
| buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı. | 4319478875703111.5 |
| mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu. | 9107426322.263557 |
| 1903 yılında bolşeviklere katıldı. | 51534.75909111291 |
| rusya sosyal demokrat i̇şçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti. | 4.152569104761506e+25 |

**2.5.5) 5 Grams**

|  |  |
| --- | --- |
| Test Sentence | Perplexity |
| yosif visaryonoviç 18 aralık 1878'de gori’de dünyaya geldi. | 82220.45272261639 |
| 7 yaşında çiçek hastalığına yakalandı ve bu hastalık yüzünde kalıcı izler bıraktı. | 1624696913150.871 |
| 10 yaşında rahip okuluna devam etti. | 4558.254885916331 |
| 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. | 1.08962537401734e+16 |
| 16 yaşında gürcü ortodoks rahip okuluna gitmeye hak kazansa da, burada otoriteye karşı başkaldırıp huzursuzluk çıkardığı için 1899 yılında okuldan atıldı. | 3241310811854632.5 |
| bu dönemde stalin, lenin'in eserlerini okudu ve marksist bir devrimci olmaya karar verdi. | 54038464958216.9 |
| tiflis'teki rsdi̇p örgütüne katıldı ve 1901 yılında tiflis'te çarlık askerleri tarafından bastırılan 1 mayıs gösterilerini örgütledi. | 4721516269880569.0 |
| buradan batum'a geçti ve petrol işçilerinin örgütlenmesinde görev aldı. | 14254039313.187462 |
| mart 1902'de petrol işçilerinin greve gitmesinde etkili oldu. | 1894809.1471574623 |
| 1903 yılında bolşeviklere katıldı. | 327.67645992011484 |
| rusya sosyal demokrat i̇şçi partisi 2. kongresi'nde kararlı ve devrimciliğe destek veren tavrıyla lenin'in dikkatini çekti. | 4.666355030142255e+17 |

**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.