NATURAL LANGUAGE PROCESSING

HOMEWORK 2

REPORT

AHMET MELİH YANALAK

151044044

In order to accomplish the project,first of all a dictionary structure needed to keep n-gram counts and use the values for calculating perplexity.

A function declared in order to create dictionary for Unigram,Bigram,Trigram,Fivegram and Fivegram.

```
def multi_dict(K, type):
    if K == 1:
        return defaultdict(type)
    else:
        return defaultdict(lambda: multi_dict(K-1, type))
```

The dictionary structure also will be used to keep counts of N numbers that will be used in GT smoothing.

dic1 = multi_dict(1,int) -> dictionary for unigram

dic2 = multi_dict(2,int) -> dictionary for bigram

dic3 = multi_dict(3,int) -> dictionary for trigram

dic4 = multi_dict(4,int) -> dictionary for fourgram

dic5 = multi_dict(5,int) -> dictionary for fivegram

n1 = multi_dict(1,int) -> N values for unigram dictionary

n2 = multi_dict(1,int) -> N values for bigram dictionary

n3 = multi_dict(1,int) -> N values for unigram dictionary

n4 = multi_dict(1,int) -> N values for unigram dictionary

n5 = multi_dict(1,int) -> N values for fivegram dictionary

Input data file has been read by 5 letters at a time and in each iteration dictionary values filled.

For example “abcde” is read by program.

Then dic1[‘a’] is incremented by 1,dic2[‘a’][‘b’] is incremented by 1 and so on..

N values are also adjusted in each iteration.

Good Turing Smoothing

In order to make GT Smoothing, N values kept in dictionaries while counting the Ngrams at the same time.After GT Smoothing, c* value is the new count for corresponding ngram.

$$c^{\star} = (c+1)\frac{N_{c+1}}{N_c}$$

$$p(x) = \frac{c(x)}{N}$$

```
cw1 = (float)((dic1[ch1] + 1) * (n1[dic1[ch1] + 1]+1) / (n1[dic1[ch1]]+1))
```

cw1 is the new count value after making GT Smoothing

CALCULATING PERPLEXITY

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

$$equivalently:$$

$$PP(W) = 2^{-l}$$

$$\text{where } l = \frac{1}{N} \log P(w_1 w_2 \ldots w_N)$$

In order to calculate perplexity,the formula above used.

```
pw1 = (float)(pw1 / wordNum)
perpForUniGram = perpForUniGram + math.log2(pw1)
```

```
perpForUniGram = math.pow(2,((float)(perpForUniGram*(-1) / wordNum)))
```

NOTE: A SMALL PART OF THE TRWIKI FILE IS USED AS DATA AND TEST SET SINCE THE COMPUTATION TIME WAS TOO MUCH THAT MADE CALCULATION IMPOSSIBLE

BELOW ARE THE RESULTS OF THE PERPLEXITIES FOR THE SENTENCES THAT CREATED FROM TEST SET AS 25 CHARACTER LONG

uğu uygur kağanlığı ndak	p1 = 8310.029832720757	p2 = 6117.687444448471	p3 = 4546.464380025864	p4 = 3501.158440232277	p5 = 3783.671042203903
i kut yetkisini kaybetmiş	p1 = 7562.889111638069	p2 = 5909.9915199279785	p3 = 4593.023717522621	p4 = 3854.9594472646713	p5 = 3515.0966053009033
tir te vefat eden ur çor	p1 = 7298.792437314987	p2 = 5408.84972012043	p3 = 5115.79047036171	p4 = 4694.813414692879	p5 = 5126.832922697067
kağan nın veliahtı olmadı	p1 = 7607.907404780388	p2 = 5634.349333763123	p3 = 5344.200128674507	p4 = 4753.144155859947	p5 = 4189.285943746567
ğı için onun yerine geçen	p1 = 7478.072470545769	p2 = 4852.647945642471	p3 = 4107.788826823235	p4 = 3301.4853229522705	p5 = 2989.624677181244
ıı kutluk un kimi kaynak	p1 = 7520.521640539169	p2 = 5971.0120849609375	p3 = 5168.067699074745	p4 = 4452.447759389877	p5 = 4435.148121595383
larda bir dokuz oğuz uruğ	p1 = 8038.328714847565	p2 = 5396.93340575695	p3 = 4628.484716415405	p4 = 5175.518150687218	p5 = 4611.188318967819
u olan ediz uruğundan old	p1 = 7706.603366613388	p2 = 5014.640848398209	p3 = 4465.914728879929	p4 = 5476.843891739845	p5 = 4472.820240616798
uğu söylenmekle birlikte	p1 = 7892.4739763736725	p2 = 5457.01711666584	p3 = 4444.23473906517	p4 = 2903.282010793686	p5 = 3404.6691048145294
bu kağanın yaglakar uruğu	p1 = 7538.29394197464	p2 = 5717.54117667675	p3 = 5626.444015145302	p4 = 5932.426569104195	p5 = 6357.249283671379
ndan olduğu da iddia edil	p1 = 7475.1725949049	p2 = 5584.2414700984955	p3 = 4307.29497563839	p4 = 3147.759630918503	p5 = 3109.128803730011
miştir tarih türk kaynakl	p1 = 7697.509699225426	p2 = 5547.579442977905	p3 = 4338.63145339489	p4 = 2897.5732560157776	p5 = 3206.9091362953186
arında yaglakar adı çokça	p1 = 7467.298925638199	p2 = 5414.963357925415	p3 = 4886.781205296516	p4 = 5317.5547132492065	p5 = 5906.143127202988
geçmektedir en hacimli u	p1 = 7949.063754081726	p2 = 5921.425021886826	p3 = 4828.063816547394	p4 = 3500.898682951927	p5 = 3333.1858912706375
ygur bitiglerinden olan t	p1 = 7690.445972084999	p2 = 5125.663497209549	p3 = 4743.9698214530945	p4 = 4636.871188163757	p5 = 4897.9974501132965
aryat yazıtları nın batı	p1 = 7585.754642248154	p2 = 5602.947495937347	p3 = 4636.801197171211	p4 = 3807.09666454792	p5 = 3219.4792696237564
yüzünün yedinci satırında	p1 = 8083.982855558395	p2 = 5698.423858761787	p3 = 4843.718212962151	p4 = 4671.92958343029	p5 = 3632.690640568733
geçen bu kelime boyla ku	p1 = 7880.529505610466	p2 = 5326.915734052658	p3 = 4866.179636359215	p4 = 3572.9700133800507	p5 = 4065.6920219659805
tlug yargan adına tahmine	p1 = 7588.255216360092	p2 = 5939.469568371773	p3 = 5595.773889422417	p4 = 4200.957315802574	p5 = 5375.598244071007
n arasında arasında dikil	p1 = 7073.000237703323	p2 = 4520.789225578308	p3 = 3713.273280620575	p4 = 3314.9158704280853	p5 = 2956.049511194229
en suci bitiği nde şu şek	p1 = 7844.872307300568	p2 = 5518.715446591377	p3 = 5573.831884503365	p4 = 5041.295383095741	p5 = 5681.236914277077
ilde kullanılmaktadır uyg	p1 = 7502.969114422798	p2 = 5136.541328668594	p3 = 4349.191730260849	p4 = 3177.3044695854187	p5 = 2232.3734896183014
ur yirinte yaglakar kanta	p1 = 7281.5745195150375	p2 = 5399.198720812798	p3 = 5806.0995453596115	p4 = 6579.009165644646	p5 = 7347.969462394714
keltim kırkız oglı men u	p1 = 7904.941077470779	p2 = 6457.595999956131	p3 = 5980.494341492653	p4 = 6160.25884950161	p5 = 6714.258935451508
ygur ülkesindeki yaglakar	p1 = 7811.466799855232	p2 = 6007.277269363403	p3 = 5427.005493879318	p4 = 4820.054394721985	p5 = 5487.980001091957
kağanına geldim ben kırg	p1 = 7479.5883004665375	p2 = 5357.8323394060135	p3 = 5035.291196703911	p4 = 4366.333704590797	p5 = 4718.885193705559
ız ım kaynakça türk haned	p1 = 7925.602960586548	p2 = 6302.048278808594	p3 = 4907.988619923592	p4 = 4279.4636372327805	p5 = 4228.893804907799
anları uygur kağanlığı da	p1 = 7937.061000466347	p2 = 5923.016770601273	p3 = 4104.157115936279	p4 = 3883.571310520172	p5 = 3045.1863577365875
phne eurydice zuniga d ek	p1 = 7876.089568257332	p2 = 7184.058292746544	p3 = 7152.445219159126	p4 = 7865.369950413704	p5 = 8059.81672847271
im berkeley kaliforniya a	p1 = 7612.02378487587	p2 = 6130.259867072105	p3 = 5463.541929483414	p4 = 5080.587968587875	p5 = 4204.460259914398
bd li aktris rol aldığı t	p1 = 7189.266751289368	p2 = 6269.138885617256	p3 = 5306.950937390327	p4 = 4385.56080031395	p5 = 4884.62913942337

Since small part of the dataset is used,perplexity values are getting greater.

The perplexity value decreases from starting unigram to 5grams.