Gebze Technical University Department of Computer Engineering

CSE 654 / 484 Fall 2019

Homework 02

Sevgi Bayansalduz
151044076

NGRAM CLASS

Before creating n-grams, corpus was corrected, and divided into syllables. Later, ngrams were created in the create_ngrams method with the help of nltk library. After the ngrams were created, Count method was found to be how many of each ngram. Ngrams and ther counts store in a python dictionary to avoid waste memory.

After count is done, the Nx (frequeny-of-frequency) is calculated to find good smoothing counts with the following equation:

$$c^* = c \text{ for } c > k$$

That complicates c*, making it:

$$c^* = \frac{(c+1)\frac{N_{c+1}}{N_c} - c\frac{(k+1)N_{k+1}}{N_1}}{1 - \frac{(k+1)N_{k+1}}{N_1}}, \text{ for } 1 \leq c \leq k.$$

Then, probabilities of each gram are calculated with the new counts.

(Random sentences created with the help of the probailities of ngrams. One of the best 5 syllables randomly picked.)

A probabilities of a sentence calculated using chain rule with the markov assumption; formula i s below:

$$\prod_{k=1}^{n} P(w_k | w_{k-K+1}^{k-1})$$

(Next formula is used to calculate to multiplication of probabilties
$p_1 \times p_2 \times p_3 \times p_4 = \exp(\log p_1 + \log p_2 + \log p_3 + \log p_4)$)

After the probabilities were calculated, the perplexity value was calculated with the following formula.

$$\text{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

RESULTS

The below table shows the perplexity of each gram for given sentence. (Corpus size:2986kb)

1-Gram	2-Gram	3-Gram	4-Gram	5-Gram	Sentence
1.52209180 73150464e+107	1.0007152731 626001e+86	3.59433506 4566249e+68	3.841955762 701439e+49	1.016146664 4599868e+32	1998'in sonunda Netscape'in pazar payı liderliğini Internet Explorer'a kaptırmasıyla sona eren bu süreci başlatan bu etkiler, daha sonra bir Microsoft yöneticisi tarafından "Netscape'in hava desteğinin kesilmesi" olarak değerlendirildi
5.124874101 180784e+18	29071717240 493.94	528396040 57.868416	15897298.76 3425615	89506.0388 313062	Netscape kısa süre sonra AOL tarafından 4
103599288 7725.4878	103824657 5.3234429	9219684.4 84885193	89541.7240 1617849	3374.44459 32490153	2 milyar dolara satın alındı
3.66679495 7973366e+60	2.12056068 07724642e+49	1.01763599 10371956e+40	1.342671639 655486e+29	2.367958170 3683207e+19	Pazarın yeni önderi olan Internet Explorer, 2002 yılında, Netscape'in en iyi zamanlarında bile yakalayamadığı %98'lik pazar payına ulaştı
5.02193666 5058856e+95	1.8648799 765154925e+77	2.89776657 8133435e+60	4.201681633 691643e+41	2.712116758 3853684e+27	Microsoft'un bu sırada ortaya koyduğu davranışlar, tekel yasalarının hiçe sayılması ve işletim sistemi pazarından tekel olma durumunun kötüye kullanıldığı yönündeki eleştirilere destek sağlar nitelikteydi
1.03994962 89715157e+35	1.37535428 56023257e+26	2.67000048 0627648e+21	1.33454343 3813731e+16	89229251 5918.3645	Bu kadar büyük oranda kullanılan Internet Explorer bu savaşı kazanmış oldu
3.344197482 128507e+54	4.668618431 8697746e+42	1.852218885 130392e+33	1.562668864 5362043e+25	1.901431531 929407e+17	Ancak bu noktadan sonra Microsoft'un IE'nin geliştirilmesine yatırım yapmamasıyla, ağ tarayıcılarındaki gelişim durdu
1.164966293 0064272e+47	1.4582758030 110866e+37	1.009231941 1377499e+30	6.8817454824 13085e+21	8031790930 328.503	Sonuçta iki tür "kötü durum" güçlendi: Ağ standartları, tek tarayıcı egemenliği ile geri plana atıldı
454798635 4.848445	1195716.78 51835755	8385.13991 6173854	518.5338124 395215	60.05509917 7131176	Internet Explorer 6
1.35168976696 61269e+45	1.7787093932 257532e+36	5.219831051 637177e+25	1.3377561430 994858e+17	75392225 22.127008	Geçişli Biçem Sayfası, PNG görüntü biçimi ve XHTML gibi biçimleri düzgün olarak gösteremiyor
281822338278 7371.5	751025199.4 059864	49836.8157 2068452	216.096017 99299878	2.38239635 007118	Glutensiz ve kazeinsiz diyet http://www
2.78486636674 39213e+17	157981906 5767.7695	5726388.1 15751906	357.418582 4295706	8.06883954 0813077	Yaşlılık da pnömoniye yatkınlaştırır
1953224551587 557.2	402852259.7 0778936	161202.89 29263195	997036024 64.76222	449.294388 8247688	ilaçlar için kanıtlar yetersizdir

According to the table, large n grams worked better than the small ones.

Random Sentences for Each Gram

1-Gram	2-Gram	3-Gram	4-Gram	5-Gram
“la ledade “	“ ara ara”	“ ve arasında”	“ olarak kabul “	“ olarak kabul e”
“la ledade “	“siysk ara”	“işler ve ara”	“ye fransız devrimi”	“de dönüşmüştür ö”