BBM497: Introduction to Natural Language Processing Lab. **(Due: 30/03/2020)**

Assignment #1

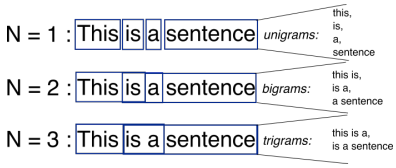
Instructor: Necva Bölücü *Name:* MELTEM TOKGÖZ, *Netid:* 21527381

1. INTRODUCTION AND BUILD LANGUAGE MODEL

I build unigram, bigram and trigram language models to predict the probability and perplexity that generated sentence. The program works as follows:

- First, the rows in the data set are read and returned as a list from the data set function.
- Then, language models are created by giving parameters 1, 2, 3 to n in the main function. In order not to read the data again each time, I read the data once in main function and sent it to the n-gram function as a parameter.
- After creating the language models, I created sentences in different language models (uni-bi-tri gram) according to the given length and count, and in the meantime, I took the words according to their weight in the "next function".
- I found the probability of MLE of each of the sentences I created. If the probability of MLE turns out to be 0, I found the possibility of making Smoothing by calling the sprob.
- Then I calculated perplexity from the possibilities I found.

While creating the language model, I added tokens to the beginning and end of the sentences. Punctuation is also included in the language model.



2. GENERATED SENTENCE

- In this task, using (unigram, bigram, trigram), I automatically created the for each language model and as many sentences as desired.While doing this, I arranged the words according to their weight and selected randomly. While selecting, I used the pairs that contained the word given in the bigram and the trigram.

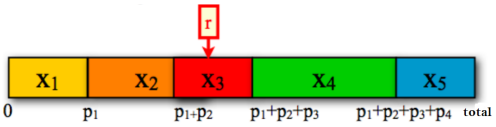


Figure 1: Weighted Random Choice

- Error analysis of your generated sentences

  - The sentences are not very correct. I interpret this as follows. Firstly, we make a weighted choice. So the word with more frequency is more likely to come. This increases the accuracy of the sentences, but since it is a random choice, it does not provide absolute accuracy.

  - Another reason may be that I use punctuation marks. This affects the accuracy of the sentence. I have observed that sometimes nonsense punctuation marks appear in meaningless parts of the sentence.

  -The language model we choose is the factor that changes the accuracy of sentences a lot. I will talk about this in the last part.

3. CALCULATION OF PERPLEXITY

Perplexity is the inverse probability of the set, normalized by the number of words. To calculate perplexity, I had to first find the possibilities of the sentences. So I found the possibility according to the formulas below.

- MLE estimate:

$$p(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

- Add-1 estimate:

$$p(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)+1}{C(w_{i-1})+V}$$

For example, a sentence created in the unigram has a probability of 0 in a bigram and a trigram, so here I have smoothing using the sprob function. After calculating the probabilities, I calculated perplexity according to the formula below.

$$\begin{aligned} PP(W) &= P(w_1 w_2 \cdots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_1 w_2 \cdots w_N)}} \end{aligned}$$

4. RESULT OF SENTENCE GENERATION

When I compared the results, I analyzed the following.

- Whatever language model we form the sentence, the perplexity calculated with that model is lower. The low Perplexity indicates that the model is better. So it is more convenient. For example, the trigram perplexity of the sentence created with the tri-gram language model is lower than the unigram and bigram perplexity. The same goes for other language models.
- Probability decreases as the length of the sentence increases. So Perplexity increases. For example, in the diagram below, we can observe this difference between the sentence of 8 in length and 20 in length. Sentence of the length of 8 have lower perplexity.
- The sentences are not very correct. I interpret this as follows. Because we select completely random words in a chart, the most meaningless sentences come out of the unigram structure. More meaningful sentences emerge from the bigram. Because the new word is chosen according to the word that precedes it. The most meaningful trigram sentences. That is, as n increases numerically, the accuracy of the sentence increases.

UNIGRAM SENTENCES		**Probability**
Sentence : (length=8)	the , Smith availed too four little sharp	
Unigram Perplexity :	4.491461389525396	1.7869216186594197e-27
Bigram Perplexity :	9.571856947825554	6.01389475608878e-41
Trigram Perplexity :	10.758698401555552	4.987132907437123e-43
Sentence : (length=8)	the of the know pretty . have has	
Unigram Perplexity :	3.5323422530424744	8.20216275725767e-19
Bigram Perplexity :	15.195550920113504	1.0077600625702914e-39
Trigram Perplexity :	19.188618664956017	4.566226991933434e-43
Sentence : (length=16)	from With remained terrible stooped Bowser them and do otter touch on dealt Rilla	
Unigram Perplexity :	4.379065846045087	6.937341986815344e-56
Bigram Perplexity :	9.987275568267584	1.1157198408905453e-86
Trigram Perplexity :	13.153499070884083	5.786931382203608e-97

BIGRAM SENTENCES		**Probability**
Sentence : (length=8)	However , who was very clever and learned	
Unigram Perplexity :	3.5019205410909553	4.824387741497877e-23
Bigram Perplexity :	1.9004043947515856	3.692435614094328e-12
Trigram Perplexity :	1.7090070658657184	2.867489321945085e-10
Sentence : (length=20)	This vexed the books she was very clever and learned , who was very clever and learned , who was	
Unigram Perplexity :	3.6697459124045237	6.2365504647847354e-55
Bigram Perplexity :	2.5784533356784674	3.2322839078288085e-40
Trigram Perplexity :	8.061225103969953	9.671622259322864e-88
Sentence : (length=15)	However , who was very clever and learned , she was very clever and learned	
Unigram Perplexity :	3.421511376083837	1.456136530229371e-43s
Bigram Perplexity :	2.1175308720567356	6.888401100254302e-30
Trigram Perplexity :	7.283493962486419	1.5376497516088256e-64

TRIGRAM SENTENCES		Probability
Sentence : (length=8)	`` You have asked all the books she	
Unigram Perplexity :	3.958939423707856	1.21545817489666e-21
Bigram Perplexity :	2.991479876721852	2.2079386344655302e-17
Trigram Perplexity :	1.9185456594366146	1.2473120561903373e-10
Sentence : (length=20)	replied the XXXXX ; for the christening party , and did not believe in fairies : she said that they	
Unigram Perplexity :	3.363569853649943	7.026763065992641e-53
Bigram Perplexity :	2.2262748741419336	3.887767619915021e-35
Trigram Perplexity :	1.7073799959055498	9.984154838858806e-24
Sentence : (length=15)	On the arm of the Royal Family was full of chapters about nothing else .	
Unigram Perplexity :	4.5543847800657185	3.917961806311576e-48
Bigram Perplexity :	2.420651291462955	2.2744654370099e-28
Trigram Perplexity :	1.4995900767381263	2.1378431154359654e-13