Assignment 1
Generating Sentences With Use N-gram Models

Doğukan Berat KARATAŞ

March 29, 2020

1 Introduction

In this assignment, I will create sentences using 3 different n-gram models that I created based on the Facebook Children Stories dataset and compare the sentences with the necessary calculations.

2 Dataset

Our dataset was a file called "cbtest-CN-train" consisting of approximately 2.5 million lines. Because of both my ineffective coding and poor performance of my computer, I used a second file called "cbt-train" which is in the same dataset but with fewer lines.

3 Task 1: Building Language Models

In this section, I will be starting to parsing, tokenizing, remove punctuation (exclude ".", "-", "_" punctuations) procedures for each sentence, I read from the dataset, and then creating 3 different models (unigram, bigram, and trigram). Next, I will be using these models both in the next () function to produce sentences and the prob () and sprob () functions to calculate probability.

4 Task 2: Generating Sentences

In this section, I will try to produce sentences using the models I created. While doing this, I will create a Gantt Chart, select random words from it and apply different operations according to each model. In this way, I will have created the desired sentences. In addition, if the words in my sentence reach 20 or if a randomly selected word coincides with the finish token (¡/s¿), the function will stop producing sentences.

4.1 Generating For The Unigram

As is known, there is no contextual relationship between words in the unigram model. Therefore, all I did while creating the sentence was to place the probabilities of all the words in my unigram model on a graph with a range of 0-1 and select random words. Example Unigram Sentences:

- hooty is horse and $< /s >$
- among . rods said in grin story d of clap but people revenge of till folks the king $< s >$ and

4.2 Generating For The Bigram

There is a contextual relationship between the words in the bigram structure, so I calculated the probabilities of the words that might come after the last word and created a graph of 1 probability. Then I picked a word from this chart with a random number that I chose between 0 and 1. I sent the word I selected back to the same function and built a recursive structure. In this way, I have created my sentences.

- the baron you the maiden i am satisfied at present in the smiling into the two years and the country
- bowser the time should be a little knitters . $< /s >$

4.3 Generating For The Trigram

There is a contextual relationship between the words in the trigram structure, so I had to act according to the possibilities of the words that could come after the last two words. Since I normally only have one start token at the beginning of the sentences in my model, I operate according to the bigram in order to be able to select a word come after ¡s¿ when the function is first called. In other steps, I gathered the probabilities of the words that came after the last 2 words, put them on a graph and created a table with 1 probability. Then I create a random number between 0 and 1 and select a random word. And I continue this process with the recursive structure I established.

- $< s >$ i ve gone and i should like to be asleep and had forgotten to wind yarn wash the cups
- $< s >$ like all kindred spirits just as well as everyone else despaired i never did . $< /s >$

5 Task 3: Evaluation of Perplexity

Perplexity of Unigram Sentences			
	Unigram	**Bigram**	**Trigram**
hooty is horse and </s>	5333.108	3784.812	10.195
among . rods said in grin story d of clap but people revenge of till folks the king <s> and	23283.055	7112.811	69.960
story look carriage and enough loud best sea-horses <s> he big hid out he nothing watched . . . everybody	46394.365	12352.500	59.588
a . was she dead sickroom be wakened dark pacifist it ready fair happened young leaked of your you <s>	36207.386	80815.032	75.333
did he . <s> we englishman again it been tears filial had minute mr. baby i as there jonas moved	23094.322	3587.837	604.089

Perplexity of Bigram Sentences			
	Unigram	**Bigram**	**Trigram**
i do you to consider of mind he will be thirty-eight st. gingolf to his eyes shone </s>	24357.150	567.303	668.664
-lrb- suiting the only whine of it s march with tall lady rosalind . </s>	29039.931	478.498	2.563
during the awkward rather distinguished himself . </s>	24130.423	218.397	2.718
the baron you the maiden i am satisfied at present in the smiling into the two years and the country	21376.295	789.383	818.766
bowser the time should be a little knitters . </s>	8416.534	143.234	33.847

Perplexity of Trigram Sentences			
	Unigram	**Bigram**	**Trigram**
<s> i ve gone and i should like to be asleep and had forgotten to wind yarn wash the cups	18826.55	560.872	60.585
<s> like all kindred spirits just as well as everyone else despaired i never did . </s>	15037.418	600.083	15.915
<s> not those which hallow your bridal-night </s>	66055.035	2477.235	13.400
<s> i profess not to be an end so why should we wish to see eden . </s>	10041.831	479.221	56.480
<s> into bowser s little kingdom of iolchos he found himself in a voice from the wall and so tired that	8137.531	913.992	35.427

6 Task 4: Error Analysis

In some of the sentences I have created, there are situations that are contrary to the normal sentence structure:

- There are 3 consecutive points in a sentence I created in Unigram.
- Again, some sentences I created in the Unigram have a dot coming after the first 1-2 words.