Assignment 1: Written Report

Shanmukha Vamshi Kuruba

Question 1

1. **What do you expect the difference between the Brown bigram and trigram models to look like?**

   Answer: The Bigram model, relying on just one preceding word for context, tends to generate more abrupt and disjointed sentences. In contrast, the Trigram model, which considers two previous words has more context, allowing it to produce sentences with more flow and coherency, compared to the Bigram model.

2. **Which model will provide you with more coherent text?**

   Answer: The Trigram model generates more coherent text compared to the Bigram model because it utilizes larger context. Trigram model considers two previous words, allowing for smoother sentence construction.

3. **How will the perplexity of each compare?**

   Answer: Logically, the Trigram model should yield a lower perplexity score than the Bigram model. Since it has more context, it should be **less surprised** when encountering an unseen sentence, leading to more accurate word predictions.

4. **Observations after testing brown bigram and brown trigram.**

   Sentence Predictor: Trigrams outperforms Bigram with respect to Coherency just as expected.
   Reason: By considering a larger context, trigrams generally provide more accurate predictions compared to bigrams.

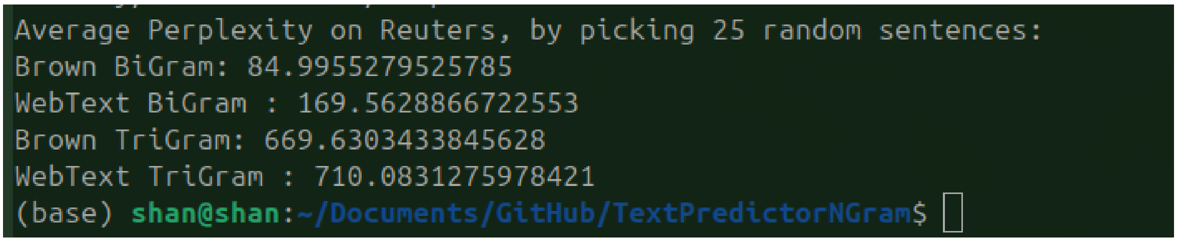
   Perplexity: The perplexity score of Bigram is lower that Trigram which is not what was expected.
   Reason: Since the number of unique three-word combinations is lower than two-word combinations, the conditional probabilities for trigrams tend to be lower. Which lead to higher Perplexity Score.

Question 2

When testing our bigram models on the Reuters data, do you think a model trained on Brown or Webtext will perform best? Pick any 25 sentences from the Reuters corpus and calculate the average perplexity using each of your bigram datasets.

Compare the results of each and provide explanation as to why you believe that one performed better than the other.

Answer:



```
Average Perplexity on Reuters, by picking 25 random sentences:
Brown BiGram: 84.9955279525785
WebText BiGram : 169.5628866722553
Brown TriGram: 669.6303433845628
WebText TriGram : 710.0831275978421
(base) shan@shan:~/Documents/GitHub/TextPredictorNGram$
```

Figure 1: Perplexity scores for all models

The above images shows the Perplexity scores I got on each of the model.

Observations:

1. **Comparing corpus:** Model trained on Brown performed than Webtext with a lower perplexity, the reason for this is that
   (a) The Brown corpus consists of well-structured, formal writing, making it more similar to Reuters.
   (b) The Webtext corpus contains informal content, such as internet conversations and blogs, which makes it less suited for predicting the structured language of Reuters.
2. **Comparing Models:** Despite expectations, Trigram models showed a higher perplexity score. This happened because the training dataset contained only 5,000 sentences. As a result, the number of unique three-word combinations was significantly lower than the number of unique two-word combinations. This led to lower probability estimates, ultimately increasing perplexity.

Question 3

When predicting the next word in a sentence, what do you believe would happen if we increased the number of sentences in our training data?

Answer: Increasing the number of sentences in the training data would help capture more unique bigrams and trigrams, improving probability predictions. This would ultimately enhance the model's ability to generate more coherent sentences and reduce perplexity.