Questions

1. What do you expect the difference between the Brown bigram and trigram models to look like? Which model will provide you with more coherent text? How will the perplexity of each compare? You should test your predictor and perplexity function using the brown_bigrams and brown_trigrams to confirm your expectations. For perplexity, an average over 2-5 sentences from the Brown corpus should be fine, but make sure you use the same sentences both times. If something you did not expect occurs, explain what happened and why you believe it happened.

Answer: As per the theoratical study and the the concepts which I understood in class, trigrams should generate more relatable and readable sentences given a corpus. The trigram model should give us the more coherent text.
I have tested my predictor fucntion and perplexity function using the brown_bigrams and brown_trigrams (5 random sentences taken from the corpus and kept same for both, trigrams and bigrams model) from the brown text corpus. I had expected that my model should have less perplexity for the trigrams model but the result said otherwise. I could see the bigram perplexity was lower than the trigram, which was unexpected.
In a way I could make sense of it as, the corpus (5 lines) are too low for the trigrams model to given proper perplexity but having three set of words appear in 5 sentences has a low probability in this small samle space. As perplexity is inversely proportional to the Probability of the occuring words, the perplexity increased for trigrams.

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: When testing bigram models trained on the Brown and Webtext corpora on Reuters sentences, the Webtext model performed better, yielding a lower perplexity score on average. This indicates that Webtext's structure and vocabulary align

more closely with Reuters' modern financial news style, whereas the Brown corpus contains older, more formal language that does not generalize as well.
Comparing the performance results:

Webtext contains more modern terms, making it more suited to Reuters' language which in general led to overlapping vocabulary.

Financial and news-related text structures align better with Webtext than Brown.

The Brown corpus includes legal, literary, and outdated news styles, which do not suit well to Reuters' financial news domain.

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 improve next-word prediction by providing better probability estimates and reducing perplexity of the trigram and bigram models both. With more data, the model learns a wider range of word combinations, making predictions more accurate and reducing the chances of encountering unknown sequences. It will also help capture rare words and domain-specific terms, leading to a more balanced and natural output.

References Used for this assignment

1. https://web.stanford.edu/~jurafsky/slp3/3.pdf
2. Lectrues Slides
3. https://training.continuumlabs.ai/data/datasets/what-is-perplexity