CE 314/887 Natural Language Engineering Assignment 1

Designed by: Yunfei Long

Please written your code using jupyter and submit in ipynb file

1 Build an n-gram language model using NLTK's Brown corpus. Provide the code. (You can build a language model in a few lines of code using the NLTK package. You may use one of the bigram, trigram, or higher-order n-grams). (15 pts)

Hint: you might reference this link: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/

2 After completing question 1, make simple predictions using the language model you built in question 1. Start with the words “I am,” and let your n-gram model predict the next word. Show both the code and the model-generated results. (15 pts)

3 Based on the work from question 1 and question 2, generate 10 different sentences that start with “You are.” (15 pts)

4 Using the sentences generated in step 3, apply relevant packages to calculate the probability of these 10 sentences generated by your n-gram model in question 1. (20 pts)

5 Based on the sentence probabilities, calculate the perplexity of the n-gram language model you used. (15 pts)

6 Using the 10 sentences you generated, ask the Google-Gemini-1.5-flash model API we practiced in Lab 2 to generate a story (within 500 words). Ensure the story contains no DEROGATORY, TOXICITY, VIOLENCE, HARASSMENT, HATE_SPEECH, SEXUAL, etc.

content. Provide the code and your generated story (the other parts of your story will not be evaluated, as long as it did not contain problematic content mentioned in red color). (20 pts)