Extra Credit NLP Assignment

Natural Language Processing (NLP) is a branch of data science/AI that deals with analyzing text and language data. The goal of this assignment will be to build a simple language model and gain insight into the application of probability to text data.

First, read the attached section taken from “Speech and Language Processing " by Jurafsky which introduces counting words, calculating joint probability of sentences, and N-grams. In class we covered counting, conditional probabilities, and discrete distributions. This section also introduces a new concept of Markov approximations.

The first part of this assignment will require you to load and tokenize the text dataset. We will use the same dataset as in the chapter- the Berkeley Restaurant dataset. You can use python packages to achieve this (e.g. nltk).

The main steps of the assignments:

1. Pre-process and tokenize the dataset provided. Only keep actual words and remove disfluencies such as “uh”, “uhm”. Add special tokens to indicate beginning of each sentence (e.g. </s>)
2. Count the words and report the size of your vocabulary. Also report the number of sentences.
3. Read the book chapter on N-grams and re-generate figures 4.1 and 4.2 for bigram counts. It doesn’t have to be exact.
4. Calculate the joint probability for at least 5 sentences (with vocabulary in the dataset) using bigrams.
5. Repeat step 2 using trigrams. Did the estimates change?
6. Submit the code or pdf of your program and output. If we can’t run the program, no credit will be assigned.