Programming Exercise

Meant only to improve your proficiency in this subject.
These programming assignments will not be evaluated/graded

1. Extend the program[1][2] that predicts the word into another program that computes the probability and perplexity of a test sentence

   (a) Build the language model using corpus
   - Big corpus → Long time to build a model
   - Development → Choose a smaller corpus
   - Testing/Production → use a bigger corpus to build your model

   (b) Check the model parameters using the debugger

   (c) Once satisfied with the learned model, test your sentences using the model

   Input:

   (d) A sentence consisting of words in the corpus that you have used for creating the language model

   (e) A sentence with one more words that are OOV

   Output:Probability of the input sentence

   (f) Try this exercise for trigram and 4-gram language models

2. Develop a Naive Bayes model to predict whether an incoming email is a Spam or !Spam. Use the dataset available at https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/ Use 70-80% of the dataset for training and the rest for testing your model

[1] `https://github.com/Ramaseshanr/anlp/blob/master/BigramLM.ipynb`
[2] `https://github.com/Ramaseshanr/anlp/blob/master/TrigramLM.ipynb`