Homework 1

1. Baseline

The accuracy for my unigram model (unigram.py) when trained on english/train and then predicting letters of the development set is 15.33%. This accuracy metric is found by running the program accuracy.py using the unigram model in unigram.py.

2. English

Without any special modifications, my 5-gram model (fivegram.py) trained on english/train and then predicting values on english/dev had an accuracy rate of 55.33%.

To improve the accuracy, I used absolute discounting on the 5-gram model on each subsequent lower model using constant d values calculated by taking a ratio of the one count and two count words for each n-gram model. I also increased the largest n-gram to a 7-gram, storing all n-gram levels from 2 to 7 in a dict object. The set_constants and set_state_constants functions were added to reduce some of the computational time, as each prev_word being checked would have the same values for the constants c(u*) and n1+(u*). The value of c(u,w) had to be calculated for each predicted w value. Utilizing absolute discounting and a 7-gram raised the accuracy of predicting values of the dev set to 62.05%.

Against the test set, the improved model predicted at an accuracy of 61.20%.

3. Chinese

Originally, my predictor running against the development set had an accuracy of 89-90%. This design generated a bigram language model that counted the number of times a Chinese character appeared after another word in the language, and the candidates function returned a list of a space, the token read, and associated Chinese characters to the token.

A flaw I noticed was that if no potential bigrams were found in the training data, then the character selected would be arbitrarily selected out of the list of all candidates without taking into account the probability of each character appearing by themselves. After adding a fix that factored in this probability by taking the sum of all the times that a character/letter appeared regardless of the previous character (as in a unigram model), the accuracy increased to a consistent 91.96%.

Running against the test set yields a consistent accuracy of 87.65%.