Project 5

ReadMe

Mike Roylance
roylance@uw.edu

Language coded	Java
Approach	I created 3 classes to help me complete this project

App.java	This was in charge of doing the File IO to bring in the words and their count and then the sentences. This also prints out the result of each word
BayesianReporter.java	This class was in charge of counting up the probabilities for each sentence. It takes in all the languages (in a nested HashMap) and introduces Laplace smoothing for calculating the probability of a word belonging to a particular language.
Tests.java	This class tests the logic in Bayesian Report. I only needed a few tests to verify my functionality

Smoothing	I decided to use LaPlace Smoothing for this problem because I find that method effective and easy to implement. Here is the algorithm: $\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$ On line 20 of BayesianReporter, I am adding in the denominator portion (1 for every word). On lines 54 and 62, I'm adding in the numerator addition (+1). These ensure that I will never have a probability of 0.
Extra Credit	I created a function CalculateMaximumUnknownThreshold that sets a threshold of "all but 3" of the words being unknown. In other words, if only three words are known, it's too little and the sentence is classified as "Unknown". This is calculated by the LaPlace smoothing of 1 / TotalCount where TotalCount is n + k (n is the total domain count where k is the unique count). This is stored in the extra-credit.txt file.
Results	Please look at output.txt created from condor.cmd for more

	information