Vaibhavi Dharashivkar
CSCI 630 – Foundations of Artificial Intelligence
Lab 2 Write up

A description of your features and how you chose them:

After careful observations, here are some of the features I tried to increase the accuracy of my code:

Contains word a	Boolean. Checks if the line contains a. Usually a most commonly used word in English.
Contains word the	Boolean. Checks if the line contains the. Usually a most commonly used word in English.
Contains word of	Boolean. Checks if the line contains of. Usually a most commonly used word in English.
Contains word and	Boolean. Checks if the line contains and. Usually a most commonly used word in English to connect multiple sentences.
Contains word een	Boolean. Checks if the line contains een. Usually a most commonly used word in Dutch.
Contains substring aa	Boolean. Checks if the line contains substring aa. Usually a most commonly used substring in Dutch.
Contains substring en	Boolean. Checks if the line contains substring en. Usually a most commonly used substring in Dutch.
Contains word het	Boolean. Checks if the line contains het. Usually a most commonly used word in Dutch.
Contains word in	Boolean. Checks if the line contains in. Usually a most commonly used word in English.
Contains word de	Boolean. Checks if the line contains de. Usually a most commonly used word in Dutch.

A description of the decision tree learning, how you came up with the best parameters and your own testing results:

Features/parameters:
There are certain features in the English language which are particular to only itself. I had studied about it in my Cryptography class. By using these unique characteristics, we had tried to decrypt an encoded message in English by brute force. This helped me select the above features.
Decision tree:
Calculate the entropy of the whole list. Then calculate the entropy of each attribute. Then found the remainder and then gain. This helps to find the root of the tree i.e. the attribute best suited and hence should be asked first since it has the highest gain. Apart from the algorithm this is the main part of the program to focus in in order to get accurate results.
Accuracy:
I ran my code for different set of 15 word lines, out of which I am submitting one that has 2000 lines out of which first 1000 are Dutch and remaining are English. I found the accuracy of my code to be approximately 94 percent.

A description of the boosting, how many trees turned out to be useful, and your own testing:

I tested the code for the same files as decision tree and discovered that the accuracy is lower than that of the decision tree. 'Contains_de, Contains_een, Contanins_the' are the most prominent features which turned out to be very useful.
Boosting was tried on all the individual stumps of the decision tree. It heavily relies on error correction.