Ryan Chan 503943165 CS112 HW4

Part II

b) Accuracy = 0.8, M.S.E. = 0.137
c) Accuracy = 0.65, M.S.E. = 0.215

Part III

For the Hasselhoff, when the classifier is run with the same pair of training data files given to the classifier as both the training input and the test input:

Accuracy = 0.875,
M.S.E. = 0.0960

When the classifier is tested with the fresh test data:

Accuracy = 0.829,
M.S.E. = 0.111

Therefore, the classifier has a higher accuracy when tested on the training data. This is because when the classifier is "trained", it is more biased towards data that is similar to the training data. Therefore, the distribution of counts will be higher for alphabets that occur more in the training data, and thus the higher accuracy when the same data is tested against the classifier.

Using the training and test data of "democrat" and "republican", the results:

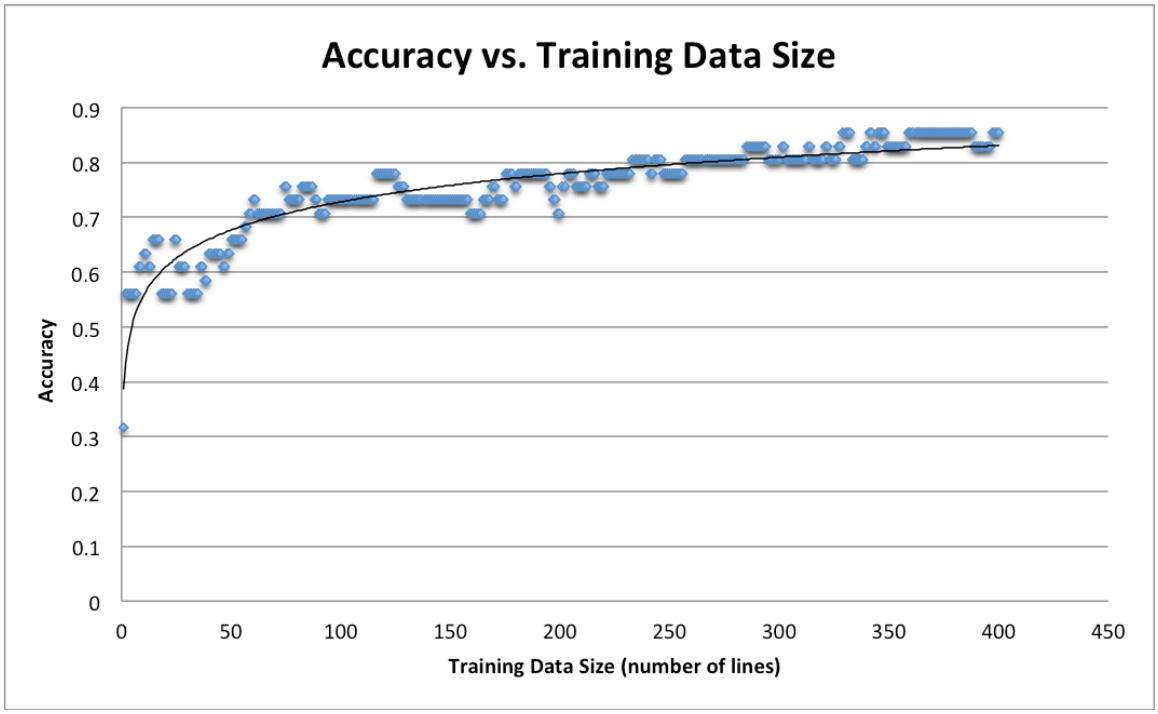
Accuracy = 0.524
M.S.E. = 0.253

Part IV (1)

For this part, I implemented an additional feature to determine if a string has 40% of the letters being vowels. The accuracy increased for the Hasselhoff data tested, accuracy grew from 0.875 to 0.882. However, the accuracy decreased for the cities data (part II b), from 0.65 to 0.64. This is because cities data are usually atypical words that are not English, so adding the vowel feature would make the classifier perform worse.

Part IV (3)

To create the plot between the amounts of data vs. accuracy, I used the hasselhoff_en_test.txt and hasselhoff_en_train.txt:



Part IV (4)

Another type of dataset is to train the classifier with two dictionaries of two languages, then test the languages on books that are written in the respective languages. In my example, I used the English and Spanish dictionaries as training set, and tested them on two books in the respective languages.
Accuracy = 0.814
M.S.E. = 0.182