**Implementation Assignment – 2**

**Submitted by :**

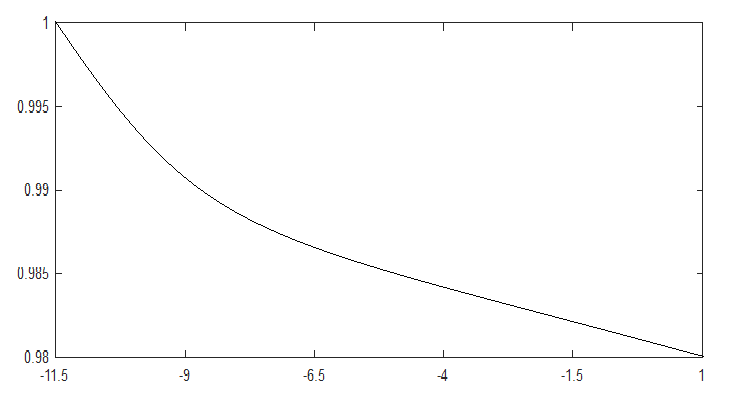
Alex Way Curtis – 932191830 (40%- completed the binomial model)

Purbasha Chatterjee – 931707158 (30%- completed the multinomial model and priors & overfitting)

Meghamala Sinha – 932944466 (30%- completed the heuristic design)

2) **Priors and overﬁtting:**

When we decrease the value of alpha that is from 1 to 0.00001, then the accuracy rate increases. Initially when alpha was 1, then the accuracy rate was 0.9804 and when it gradually decreased to value 0.00001, then the accuracy rate was 0.99806.



log alpha

3) **Identifying important features: For this part, design and test a heuristic to reduce the vocabulary size and improve the classiﬁcation performance. This is intended to be open-ended exploration. Please describe clearly what is your strategy for reducing the vocabulary size and the results of your exploration. A basic pointer to seed your exploration is that we would like to remove words of no discriminative power. How can we measure the discriminative power of a word?**

For this part, we have decided to remove the words from the dictionary which appeared with similar frequency in both Trump and Hillary tweets.

We have calculated the probability of appearance of each word for both in Donald Trump and Hillary Clinton’s tweets. After comparing the probabilities of the common words, we check for the ones having less than 5% difference among each other which could be removed from the dictionary. Since the probability of appearance of these words are the same in both Hillary and Trumps speech, they do not seem to have any discriminative power for judging whether its Donald’s or Hillary’s tweet.

Implementation:

for each word in the dictionary:

if trump\_word\_prob < Clinton\_word\_prob

Set a = trump\_word\_prob / Clinton\_word\_prob

else

Set a = Clinton\_word\_prob / Trump\_word\_prob

if a < 0.05

Remove word from the dictionary

The total number of such words found are 1065, which can be removed to reduce the dictionary size. Since they appear with similar probability in both tweets, they are less likely to have any discriminative power for identifying tweets.

The number of words found is 3556, when we check for words having 10% difference in probability of appearance in Trump and Hillary tweets.