Assignment 4 - Text Classification

Sharath Chandra Bagur Suryanarayanaprasad
800974802
sbagursu@uncc.edu

Accuracies-

Naive Bayes Model = 63.25%
Bigram model accuracy = 62.75%
Binarised Naive Bayes Model = 55.5%

Discussion-

Given the preprocessed debate transcript data we first parse every line and get the total document count. Then while parsing every line we generate a dictionary containing a bag of words used by the respective speaker(class). We further use the formula to estimate the likelihood of words and calculate the probabilities. We apply Naive Bayes in order to develop a model to predict and classify as to which class a particular text belongs to. We use the following formula to do so.

Difficulties faced-

Parsing and developing data structures to store the parsed data. Iterating over large dataset and storing relevant data was a key hurdle in implementation. Calculating values from given data and plugging it into the respective formula.

Execution Steps-

classify.py is the file that contains the Python script to run and execute the assignment. Use this file to execute the program. The output of the file will display the accuracies of the models.

The visualization is obtained by taking the output of python script and feeding it into Rscript. The path /Rscript/FrequentWordsVisualisation is the R project required to execute the R script. The Visualization.pdf file present in the root directory of the assignment is the output visualization from the R Script.