Balaji Paiyur Mohan
1001576836

Structure of Code:

- The code is designed using Python and contains a function for data handling
- First, the input folders are retrieved as a list from the parent directory.
- Then the model starts with training of first 500 files in each category where the words and its frequency are stored in a dictionary for easy access to test it in the future
- At the end of training model, we will have dictionary of trained files under each category, dictionary of words under each category and a master dictionary of total words.
- Now, using the trained model we test the remaining 500 files in each category and classify the words by computing log probabilities(because it improves accuracy of model) over the testing data using the formula,

$$P(i \mid j) = \frac{word_{ij} + \propto}{word_j + |N| + 1}$$

Where N is the total number of words in vocabulary and ∝ = 0.0001(Laplace smoothing)

$$P(j) = \log \pi_j + \sum_{n=1}^{N} \log(1 + f_i)\log(P(i|j))$$

Which is the optimal model.

- The number of collisions of maximum probabilities of a particular word in a category is computed and is divided by the total population in order to calculate accuracy of the model.
- Accuracy of this model comes out be **65.00 %**

Screenshot of output: