**REPORT**

**PROBLEM 1:**

To calculate the accuracy, the following procedure is applied:-

* Once the train and test data is read into the environment, punctuation marks like ‘,’ and ‘.’ are removed.
* Initially the total number of classes (Speakers in this case) is determined and the number of documents under each class is found out.
* Once it is done, the prior is calculated for each of the classes using the principle: documents under that class divided by the total number of classes.
* Then the total vocabulary count is determined, which is the count of the total number of unique words.
* The next step would be counting the total number of words under each class by placing the class name and the corresponding count in a dictionary.
* Later, count of each word is determined for every speaker, which consists of a class name as a key and the corresponding set of a word and its count as value.
* For calculating accuracy, probability that a line of test dataset belongs to a particular class is determined and the class with the maximum probability is chosen as the predicted class. If the original class is same as the predicted class we count it as a positive or correct case. Probability of every word is calculated using Naïve Bayes formula – (N+1)/(C+V)
* Finally, the accuracy is determined by dividing the count of the positive or correct cases by the total number of test classes.

In this problem, two main principles are applied:

1) Smoothing – To avoid the problem with null values, smoothing is applied. In the first case, Laplacian or Add-one smoothing is applied, and the accuracy in that case turned out to be 52.0. In order to further improve the performance of the model, the alpha value or smoothing coefficient in the formula is changed to 0.1. Due to this change, the overall accuracy of the model turned out to 64.0

2) Log transformation – To avoid underflow of the values obtained, log transformation is applied to the Naïve Bayes formula.

**Accuracy of the model using Add-one or Laplacian smoothing: 52.0**

**Accuracy of the model using an alpha of 0.1: 64.0**

**PROBLEM 2:**

In this problem, in addition to the Naïve Bayes bag of words model, two additional kinds of features are implemented as described below:

**a) Using Bigrams and Trigrams:**

In the first case, instead of using unigram count in the Naïve Bayes formula, bigram model is built and the count of bigrams is used in calculation of the probability.

The priors and vocabulary remains the same, but the count of the total number of bigrams under each class is to be determined.

And the number of appearances of each bigram in every class is determined using the same principle.

Finally, the probability of each document in test data is determined using the Naïve Bayes formula and if the predicted class turns out to be same as the actual class, it is considered as a positive case.

**For the bigram model, the accuracy is determined as: 47.25**

Similarly, the same process is followed for calculating with the trigram as well and the output is as follows:

**For the trigram model, the accuracy is determined as: 5.0**

This implies that using Bigram and Trigram models has reduced the accuracy of the model to some extent. Because of the presence of very few such combinations, the accuracy has reduced.

**b) Using Binomial Naïve Bayes:**

In the second case, binomial Naïve Bayes approach is followed where instead of taking all the words in the train data, only the unique words are considered while calculating the count of the words or the occurrences of each word.

This process makes sure that there are no repetitions to occur in the list of words. So while calculating the count of words under each class all the repeated words are eliminated using set().

**The final accuracy of the model using binomial method is determined as: 64.5**

**PROBLEM 3:**

For problem 3, the top 20 key-value pairs of words and the corresponding count for the available speakers (17 in total) are exported as CSV files. Later the CSV files are imported into R environment to plot the words.

Bar graphs are plotted with words on the X-axis and their corresponding frequencies on the Y-axis.