**QUESTIONS**

1. **Describe text processing pipeline you have selected.**

The text processing pipeline used was both scikit-learn and NLTK (natural language toolkit) for the logistic regression and naïve bayes respectively in which a Count Vectoriser was used for the vectorization. This is basically because scikit learn helps in removing words which is called tokenization in which the words are converted to integers or floating point values so as to be used as an input for a machine learning algorithm. And scikit learn is easy to use, while NLTK helps in tokenization, parsing, classification and stemming.

1. **Why you have selected these two classification methods?**

First is the Naïve Bayes algorithm, and since our model is a classification model then one of the best algorithm to use is a Naïve Bayes algorithm which is because it is a popular algorithm for classifying text. Naïve Bayes is also highly scalable.

Also logistic regression is a popular classification model algorithm which is used when the dependent variable is binary like our dataset which contain both Positive and Negative sentiments.

1. **Compare selected classification methods. Which one is better? Why?**

For logistic regression, I had to check out the performance with different regularization values to get the best value before finally agreeing to select the one with the highest accuracy score and I had to divide the dataset into two different segment containing train and test datasets.

The train dataset had an equal number of positive reviews and negative reviews and also for the test dataset, it contained equal numbers of positive and negative reviews.

Using Naïve Bayes, I was able to call each positive and negative review differently and created a split function that divides each dataset into test and train in which I later trained using the naïve bayes classifier and also using the NLTK for tokenization and stemming. Naïve bayes get the best classifier and renders the best with its accuracy.

Naïve bayes seems to be the better classifier because it assumes all the features are conditionally independent and it worked well with the training data since the estimate are based on a joint density function.