# COL341 Assignment 5 Report

## Part a

### Preprocessing of data

For all the reviews

- Replace all the punctuations in the review string with spaces.
- Convert the whole string to lowercase.
- Split the string by whitespaces into a list of words.
- Convert the list of words into a set of words so that duplicates get removed.

### Enhancements made in the algorithm

- Used laplace smoothing to get rid of zero probabilities
- Instead of multiplying the probabilities, took log of them and added, so that underflow and precision issues can be avoided.

Accuracy obtained — 84.1975%.

## Part b

- Removed all the stop-words using nltk's list of stop-words.
- Used PorterStemmer to step the remaining words. PorterStemmer was giving a bit better performance than SnowballStemmer.

Accuracy obtained — 85.0975%.

## Part c

- Added all bigrams to the data. Accuracy obtained — 84.7975%.

As we can see, It did not improve the model. Keeping only the frequent bigrams might have helped, but I didn't get to test that.