**Assignment 6**

**Cynthia Khan**

**11/30/2018**

The code contains 3 parts:

a) remove\_punctuations.py – which cleans the dataset of all punctuations and wild characters and concatenates the positive and negative training datasets.

b) data\_processor.py – which tokenizes the cleaned dataset and stores each word in the appropriate arrays to be utilized by the algorithm

c) classifier.py – which takes the tokenized input from the data\_processor.py and performs the naïve bayes classification, given the number of classes, which in our case is 2.

The positive and negative training datasets along with the clean data and test data are included.

The accuracy from training 80 reviews and testing on 20 reviews is 65%. While all the positive reviews are correctly identified, only three of the negative reviews are classified correctly.

The accuracy can be greatly improved by increasing the training dataset and selecting more features which carry greater weight or are better indicators for the classification.