COMP 551 – APPLIED MACHINE LEARNING

Assignment 3 - Sentiment Classification

Name: Rashik Habib

Date submitted: 26th February 2018

Question 1

To generate the datasets, a feature set was constructed using only the training vocabulary; the HTML tags and punctuation marks were removed from the entire text before using the *CountVectorizer* class to extract the 10000 highest frequency words, ignoring common English articles (like 'the') since the presence/frequency of such words was considered to be unrelated to the sentiment behind the review. Afterwards, each review within each dataset was examined, and the corresponding vocabulary lists used to replace the word with its appropriate index. If the feature was absent, it was ignored. The datasets were outputted as "*<dataset>-<type>-submit.txt" (*for example "*yelp-train-submit.txt").*

Question 2