COMP 551 ASSIGNMENT 3

Section: 2

Name: Shatil Rahman

ID: 260606042

**Q1) Text to Vector format:**

Both the IMDB and yelp datasets were converted to fixed length vector representations, in binary bag of words and frequency bag of words forms. The 10000 words that comprise the feature vector, along with their ID’s are in the vocabulary files, named “IMDB-vocab.txt” and “yelp-vocab.txt”. The training, validation and test sets were then converted to vector representations, and saved in “IMDB-train.txt”, “IMDB-valid.txt” and “IMDB-test.txt” for the IMDB set, and “yelp-train.txt”, “yelp-valid.txt” and “yelp-test.txt” for the yelp set.

**Q2) Classification performance on yelp dataset:**

5 classifiers were run on the yelp dataset. Two of them were for baseline performance only: a random classifier which assigns a sample to a randomly (uniform probability) chosen class, and a majority classifier, which calculates the majority class in the training set, and predicts all examples as that class.