**IS620 Project 4**

**Student: Neil Acampa**

**6.10.4**

Using movie reviews: Display and analyze the top 30 features.

The program selects the top 2000 most frequent words from NLTK’s movie review corpus.

Then stop words are removed.

Features are trained and tested twice. The first time trained on the 100Th – 2000th record and tested on the first 100 records. The second time trained on the second 90% of data and tested on the first 10%.

The Naïve Bayes classifier is run on the training set and accuracy is measured on the test set.

The accuracy displays for both runs as well as the top most significant features

6 words are selected and the features are analyzed in terms of whether the result (Positive/Negative) is surprising.