**Week 8 – Lab Homework**

Let’s try using a stopword list to prune the word features. We’ll start with the NLTK stop word list, but we’ll remove some of the negation words, or parts of words, that our negation filter uses. This list is still pretty large.

(In this question the python parts are preceded by the prompt >>> .)

>>> stopwords = nltk.corpus.stopwords.words('english')

>>> len(stopwords)

127

>>> stopwords

>>> newstopwords = [word for word in stopwords if word not in negationwords]

>>> len(newstopwords)

>>> newstopwords

Now take the new stop words out of the collection of all words, and then take the top 2000 to be the word features.

>>> new\_all\_words\_list = [word for word in all\_words\_list if word not in newstopwords]

Now continue to get new word features of length 2000 after the stopwords are removed:

>>> new\_all\_words = nltk.FreqDist(new\_all\_words\_list)

>>> new\_word\_items = new\_all\_words.most\_common(2000)

>>> new\_word\_features = [word for (word,count) in new\_word\_items]

>>> print(new\_word\_features[:30])

Now choose to re-run one of the classifiers with the word\_features having stop words removed. Noting that the definition of the feature functions uses the word\_features variable, choose to redefine either

featuresets

SL\_featuresets

or NOT\_featuresets

**Choose one option to report results—Option A. or Option B.**

Option A.

Re-run the training and test sets, train the classifier, and report on classifier accuracy. Be sure to report the baseline accuracy that you got for that type of feature set when you first ran it and the new accuracy score with stopwords removed.

Option B.

Another option would be to redefine the SL\_features function to have just one numeric feature that would subtract the number of negative words from positive words. Again you would report the baseline accuracy score for the original SL\_features and a new accuracy score for the new definition of features.

**Homework:**

Submit a report summary that includes the results along with the baseline accuracy achieved and new accuracy score achieved.  Include any observations and lessons learned in your report.  In addition to the report summary, upload a copy of the python code you developed as part of this lab.