**Assignment 4**

In this assignment you are to create a sentiment analysis classifier using the Natural Language Toolkit classifier module.  As training and test data, use the Movie Reviews Corpus provided by NLTK. Write a short report describing what you did and why. See the sections below for more details on what to include in the report.

**Training**

The training data consists of 2,000 documents (movie reviews) that have been tagged as having negative or positive sentiment.

**Classifier Type**

You will decide which type of classifier to use. Sticking with Naïve Bayes is fine, as is using the code from Ch.6 as a starting point. Experimenting with different classifiers will earn you 5-10 points of extra credit. If you do this, include the code from all the different classifiers you tried, but indicate which system is your final one. Describe what you did in your report and include the results from using each classifier.

**Features**

You will also decide which features to use, but you should experiment with **at least** **two** types of features not described in Ch. 6, Sect. 1.3. Your grade will depend in part on the thoughtfulness of the features you try, whether they improve your accuracy or not. You may use supplementary files for features, such as a sentiment vocabulary (but it is not necessary). Be sure to upload these to D2L.

In your report, describe the features you tried and why you thought they would be useful. Explain which features you kept in your final system and why you chose them over any you may have eliminated.

**Evaluation**

Use 10-fold cross validation on the Movie Reviews Corpus, outputting the average accuracy to the screen and including the results in your report. You can earn an extra 3 points of extra credit if, in addition to accuracy, you report precision and recall and an F1-score. Your output should report accuracy as a decimal to 3 digits, as follows:

Accuracy: .000

If you do the extra credit, add an additional line to the output as follows:

Precision: .000; Recall: .000; F1: .000

**What to hand in and when**

By 11:30 p.m., Friday, April 28, turn in your code to the Dropbox, along with a pdf of your report.  Your code should output an average accuracy for your 10-fold cross validation directly to the screen, as describe above. Name your program **Lastname\_PA4\_final.py**. You may upload to the Dropbox separate code that did not get used in the final version (such as code using different classifiers), but name them something different. Have your program run as follows:

python Lastname\_PA4\_final.py training\_file.txt

If your program uses supplementary files, have your program read them in from the same directory as the "\_final.py" file. So, if you have a file like "sentiment.txt", code something like

with open("sentiment.txt",'r') as f:

      <do stuff>

Do not use any subfolders.