**NLP Lab Session**
Week 13, December 2, 2015
Reading Final Project Data for Classification

**Getting Started**

In this lab, we will be running some stand-alone python programs that demonstrate how to work with the three datasets available for the classification option of the Final Projects, which I have downloaded for you from the web. All of the programs and data are zipped together in one zip file on Blackboard:

FinalProjects_ClassificationData_Fall2015.zip

Download this zip file to your NLP class folder in the lab and unzip it there. The folder is organized by the 3 datasets, but there are python programs within each folder that might be useful in other parts.

The additional Final Project options are also available as zip files on Blackboard:

FinalProjects_Annotation_Fall2015.zip
FinalProjects_Programming_Fall2015.zip

**Looking at and Reading the datasets**

Each folder contains a python program that we will run to process the data in the folder.

- First look at the raw data.
- Run the program and observe some of the tokenized data.

**Notes:**

1. The limitations of using the classifier in NLTK is that you may be limited by the number of documents and the number of features that you can use in order for the classifier to run in a reasonable time and memory space.
   a. each program allows you to limit the number of documents (emails, phrases or tweets)
   b. you can limit the number of features in the python program

2. In the Enron email dataset, if you want to run the POS tagger, you must first separate the text into sentences, using NLTK's sentence tokenizer, for example, this gets a list of sentences, each of which can then be tokenized and run in the POS tagger.

```
sentences = nltk.tokenize.sent_tokenize(text)
```

3. Sentiment Lexicons: These lexicons and the program to read subjectivity words (with all three types of positive, negative, and neutral) are found in the kaggle folder. If you want to try using positive and negative sentiment words from the LIWC lexicon, I have given a python program that will read them in

sentiment__read_pos_neg_words.py

**There is no Lab Exercise this week.**