**ANLY520 – Sentiment Analytics**

**Term Project**

The term project for this class, due the last week of the semester, consists of a presentation demonstrating the use of the Natural Language Processing techniques you've learned in the course. You will choose a data set, and use the tools of your choice to analyze the data. In your presentation, you will describe the data and present the results of your analysis in the categories shown below. You may use screen shots, diagrams, tables, and text in your slides. You will also need to record audio of yourself presenting each slide of your report. This project accounts for approximately one third of your grade in this class, so be sure to do the best work you can do.

**Data:**

You may choose any data set you would like to work on, as long as it contains at least 1000 distinct unstructured texts. This can be a collection of Twitter data, blog posts, e-mails, news reports, or similar data. The following links contain some suggested data sets:

Enron Email dataset: <https://www.cs.cmu.edu/~./enron/>

Stanford Sentiment Analysis dataset (Twitter): <http://nlp.stanford.edu/sentiment/>

FBS opinion mining data sets: [http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets](http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html" \l "datasets)

Cornell movie review data sets: <https://www.cs.cornell.edu/people/pabo/movie-review-data/>

**Processing:**

Once you've chosen a dataset, you will perform the following tasks:

* Clean the data for processing
* Create a corpus appropriate for your tool
* Tokenize the data
* Perform basic analysis of the data
* Train a sentiment analytic
* Test the sentiment analytic

**Reporting:**

Your report should consist of the following items, supported by appropriate text, diagrams, and screen shots:

* A description of the dataset, including where it was acquired from, the format of the data, and whether or not the data included any tagging.
* Basic statistics about the data
  + Number of documents
  + Average length of document (characters and words)
  + Frequency distribution/lexical dispersion
* Explanation of your sentiment analytic tool and method, including any source code you produce
* Explanation of your training and test set creation
* Results of testing your analytic
  + Findings of the analytic
  + Precision
  + Recall
  + F-score
  + Performance (average processing time per document)
* Your conclusions, including anything notable you learned in the process of completing the project.

Your final report should consist of at least 15-20 slides of content, each of which should have an audio recording of you presenting the material. Remember to properly document your sources.