CS410 Project Status Report

Name: Chris Toombs

NetID: ctoombs2@illinois.edu

Current Progress:

For this project, I am working on the twitter sarcasm detector as part of the text classification project. I have successfully created a python project with git integration to my project repository. I have verified that the repository connects successfully to Live Data Lab.

In terms of implementation of the project, I was going to use word2vec, but I am looking at utilizing gensim's doc2vec along with scikit-learn for classification. The reason for this is that doc2vec will allow me to retain semantics, whereas word2vec does not.

I have successfully loaded in both the train and test data sets in my programs and have preprocessed the words using the NLTK package and it's associated stop words. I am not going to perform stemming at this time, but I am adding some custom stop words (such as @USER) to clean up the documents (i.e. tweets)

Remaining Tasks:

I will need to make documentation on how to run the program, as well as fully comment my code. To be completed, I need to instantiate my documents with word2vec and implement the logistic classifier. I also want to clean up my pre-processing step, so I will be working on that in the next couple of days.

Challenges:

I've never used Gensim before, so I will need to read up on that, but there seems to be a lot of content online regarding. I will be getting married this week, so I will not be completing this project (most likely) until next week, so I am going to try and get as much work done as possible from 11/29-12/1. So far, nothing blocking me from finishing this on time.