Information Extraction

Team Members:

1. Nishant Agarwal – u1010232
2. Murali Krishna teja Kilari – u1006392

Task: Sarcasm detection using Twitter dataset

Dataset: Since, the tweets have to be manually tagged and of sufficiently high quality and also since Dr. Riloff has previously worked on sarcasm detection, we would like to request the professor to grant us the dataset and we are ready to sign the required documents so as to prevent the distribution of dataset.

Planned Approach: During the course of the semester, we will try different techniques and algorithms to find sarcasm in tweets. We will use Unigrams, Bigrams and Trigrams with Tf-Idf weighting, Word2vec to capture the semantic similarity between words and phrases, POS Tagging of the words in a tweet within the window size 2 (2 because tweets had an 140 character limit and hence will mostly be shorter than a regular sentence), Sentiment analysis and also we will try to extract the phrase in each tweet that caused it to be labelled as sarcastic.

Baseline Approach: The baseline approach will be the extraction of Unigrams, Bigrams and Trigrams with their Tf-Idf weighting and give them to a logistic regression classifier with L2-reguralizer to try and classify tweets as sarcastic or not. Also we will try to extract the bi-grams and tri-grams that caused the tweet to be labelled as sarcastic.