Big Data Project Update - November 5, 2015

Clouded Minds

by Deepak Gautam, Sandra Lee Gibson, Michael S Nichols, William Wheeler Carter

We have worked with the tweets we have collected from several games of English Premier Soccer League. We ran into several speed bumps in the process. We discovered a problem with quotations in the tweets producing invalid JSON files. This issue has been addressed and should not cause any further issues. Also, some of the data previously collected was corrupted. This affected several games data, which has resulted in less data to test and analyze.

We are working with sentiment140 for performing sentiment analysis on the data. Sentiment140 will only support English and Spanish. This has caused us to have to discard some of the data already collected. We now plan to capture 5 to 6 games for full analysis. In order for the sentiment analysis to work as intended, it is necessary to separate the tweets by team. We are looking at possible approaches to properly separate the tweets.

We have been able to calculate slopes of the spikes to get match events. We were also able to analyze the tweets in the detected spikes. The major problem we are facing is making single term while vectorizing for similar words like ‘Goal’ or ‘Goooal’ so that such words’ real count within the spike can be calculated accurately. We were able to detect scorers of the soccer game with good accuracy values, but for detection of team which scored, we are planning to have sentiment analysis on tweets within

considered spike. The accuracy values for red cards in soccer game isn’t that bad either, but it has been difficult to get other minor events such as yellow cards or substitutions. The result has is looking better for tweets tweeted in english language, but unfortunately the tweets collected earlier has no language parameter that twitter gives, so we need to collect more tweets with language parameter.