Text and Sentiment Analysis on Tweets - 2015 Cricket World Cup

The most popular communication tool to share everyday opinions and life events is social media. Twitter, an online social networking service can be used to discover such events in real time. As tweets are short 140-character messages being generated continuously, they are a good source of streaming data for opinion and sentiment analysis.

These tweets reflect public sentiment when aggregated, we can extract the sentiment and look at the general correlation between these sentiments and an event. A large collection of such tweets could be taken to provide a useful reflection of public sentiment towards some special events. We have decided to use the 2015 cricket world cup as our event.

The following abstractly discusses the steps we have decided:

**Collecting Data**

The first step is to build a data set. Twitter provides [REST APIs](https://dev.twitter.com/rest/public) we can use to interact with their service. [Tweepy](http://tweepy.readthedocs.org/) is one very good example.

**Text Pre-processing**

In this step we want to do some data clean up as well as pre-process (understand the anatomy and tokenize the tweets) the data to make it usable for our analysis further.

**Analysis**

We extract the words of interest and remove some stop words, we can use some NLTK components for this approach. We might also wish to add analysis around co-occurrences (sometimes terms that occur together might be of interest).

**Data Visualization**

We might want to do some time series plots to analyze the frequency of tweets on a team by team basis.

**Sentiment Analysis**

We want to apply some natural language processing techniques to figure out the opinion towards a certain team.

By

Vishwas Vijaya Kumar

Anirudh ..

Gautham ..