Filippos Lymperopoulos   Dimitar Dimitrov
03/03/2014   Software Design 2014

Project Overview

Following the Oscars, we were really interested in mining and further analyzing data from Twitter feeds to determine who and what the general public were discussing during the event. The general approach was to get access to such data, and deduce how and to what extent people talk during, and at the aftermath of the event. Instead of generally looking into the Oscars, we zoomed into the trends of individuals that were nominated for Best Actor, etc.

Implementation

The first part of our code records tweets from the Oscars. We used the Twitter module from pattern.web library. The Twitter API only allows for consecutive tweets to be downloaded – and only 100 at a time – which actually lent itself well to our purpose. We tested some tweet id's by hand to determine the approximate tweet id's at the start and end of the Oscars, and gathered approximately evenly spaced "clusters" between them by re-running the search function 100 times at evenly spaced starting values. We recorded the tweets from each cluster in a list, and recorded all the cluster lists in one main "tweets" list.

We then tested the texts of the tweets against our "buzz"words – a conglomeration of nominees, movie titles and award presenters we thought would be interesting to analyze. Each buzz word in the list is actually a sublist of words associate with it. For each buzz word, each tweet with a match adds a counter to its cluster's item in a master frequency list. Furthermore, each matching tweet also adds the result of a sentiment analysis (from the pattern.en library) to a different sentiment list.

Next, we re-created the time of each cluster by looking at the .date of the first tweet in each. We then proceeded to smooth the data by creating new "average" lists, where each data point was replaced by the average of itself and its four (six for sentiments) closest neighbors. We graphed and formated the results using the matplotlib.pyplot library.

Results

The results that our project yielded were pretty interesting. Since our main goal was to monitor tweets regarding certain people, we saw how observers of the event get influenced by show. The frequency results closely matched awards given. We also created graphs based on sentiment analysis from tweets and observed the reaction of people to the Oscar's results. Though the sentiments averaged positive for the most part, it was clear that they increased when their subjects won an award, and decreased when they lost.

Reflections

In general, the project run well enough and we were able to deduce information that seemed rational. We believe that our results and graphs would be more accurate, through the use of more data. However, the maximum limit of data mining implemented by Twitter (150 querries per 15 minutes maximum and, more importantly, 100 results per query) led to the limitation of data from our source.

We worked very well as a team dividing the work and making sure we both understood what was happening. As a result, we both contributed sufficiently and created a pretty cool project. In future iterations, we would perhaps like to involve other sources, such as Facebook etc. and use computational linguistics to create trees of the progressions of talks around the issue of the Oscars in the cyberspace.