DIG210

Professor Owen Mundy

Data-Based Project Proposal

Brandon Liang

Twitter Sentiment Analysis and Network On Election Night

The past week has been dominated by the topic of the unprecedented presidential election and apparently, people use Twitter to express their opinions and shock. Under this background, I propose a data-based project to scrape through twitter feeds from the election night to analyze how the public reacts to the election night.

I used the "Tweepy", a Twitter API provided in Python, to scrape a sample of tweets on the election night based on different hash-tag searches. The three hash-tag keywords I used are "Trump", "Hillary" and "Election"; thus, I have generated 3 datasets (attached here in the email as well) in csv format. Each dataset contains over 4000 tweets on each hash-tag keyword with features like user, time, location, tweet, retweet count, etc. I then plan to use "NLTK", Natural Language Processing Toolkit written in Python, and sentiment analysis (http://text-processing.com/demo/sentiment/)to process the tweets.

Limitations: due to the limited requests per 15-minute window of Twitter API, I was only able to scrape over 4000 tweets for each hash-tag keyword, much less than I intended.

Assumptions: the data are only representative of the population online, or even more, the sample population of the data. Given the fact that not everyone uses Twitter, we need to understand social media has its limitations; given the actual amount of active Twitter users, over 4000 tweets can hardly reflect the entire Twitter community. Also note the fact that all the tweets collected were tweeted right before midnight of November 8th, when the election was at its most crucial stage before the official result was revealed.

My goal is to discover the trends from tweets for each hash-tag keyword using Natural Language Processing (NLTK) and sentiment analysis (http://text-processing.com/demo/sentiment/). I expect a more neutral and objective tone from the tweets under #Election than that under the other two hash-tag keywords. Moreover, after an initial glance of the datasets, it shows that most of the tweets were retweeted or linked to external articles via URL (since tweets have a character count limit). I would like to use Tweepy to explore the original users of those retweeted tweets and Web Scraping to scrape the titles for all external articles; with the number of tweets and users I have from those datasets, I am expecting to produce a massive, highly connected network with some degree of connection within its components. This would show the connectivity of the digital world we live in and further solidify the theory of "Six Degree of Separation".