INSY 5378: Group Project 1

Post-Election Social Media Analytics

**DUE DATE**: Nov 18th, Friday by 11:59 p.m.

Instructions:

1. This is the first group project. The total is 100 points.
2. Late submissions will not be accepted.
3. Submit (1) your project report and (2) Python code in Blackboard. DO NOT email your project.

The purpose of this project is to analyze how people’s sentiment and topics change after the US presidential election. Please follow the following steps:

1. **[20 points]** Using Twitter Streaming API, collect 10K tweets for each of three important people: Hillary, Trump, and Obama. The total number of tweets should exceed 30K. You may use Twython or any other Python module. As shown in the lecture, you can use track=KEYWORD to get keyword-filtered tweets. (Note: In case you loose the API connection with Twitter, you may need to run your code multiple times to collect sufficient number of tweets.)
2. **[20 points]** Calculate the sentiment score for each person. You can use TextBlob (or other sentiment analysis modules) to get the polarity and subjectivity scores for each tweet. Then aggregate the average scores over all the tweets for a given day and the given person. Who has the most positive comments? Who has the most negative ones?
3. **[20 points]** Create the word clouds for each of the three people. Please remove stop words and do stemming before feeding into the word cloud module.
4. **[20 points]** Repeat Steps A (collection), B (sentiment analysis), C (word cloud) by adding the locations filter to the Streaming API. You can use locations=lat1,lon1,lat2,lon2 to collect tweets from a specific region.
   1. Collect tweets from the following 7 cities: Dallas TX, New York NY, San Francisco CA, Los Angeles CA, Chicago IL, Washington DC, Atlanta GA.
   2. Who has the most positive/negative comments?
   3. Create word clouds for each city and person. (See Streaming documentation for the details: <https://dev.twitter.com/streaming/overview/request-parameters#track>)
5. **[20 points]** Using the 30K text corpus from Step A, run topic analysis with 20 and 50 topics using NNF and LDA from GENSIM.