Assignment 4, Data Processing & Visualization QMSS G4063
Navid Hassanpour, nh2519@columbia.edu
Due Thu. April 7, 2016

Please submit your assignment on Courseworks and include links to your 1) code and 2) web-based visualizations in the report. If you have static visualizations, embed them in your submission file. There are bonus points for conducting the challenge part (note: it is optional) at the end of the assignment and uploading your Shiny app online.
Do not upload large data files, all coding should be executable on the data files made available to you online, no need for large file uploads.

Detecting major topics in candidates' conversation corpus, visualizing the contrasts, (9 visualizations, 1000 words, optional link to online app)
Using the same link as the one you used for the third assignment, access a daily collection of tweets pertaining to the Primaries. Download the 10 files `tweets.03.16.2016.json` to `tweets.03.25.2016.json` (or the smaller versions, `tweets.03.16.2016.summary.json` to `tweets.03.25.2016.summary.json`, if your computer can not process the larger file). Using the streamR parsing command and what you learned in assignments 1 and 3, change the .json file into a R data frame and divide the file, into

1. `Democrat, Republican`, or
2. `Clinton, Cruz, Sanders, Trump` subcategories.

You will need both (R, D), and (HC, TC, BS, DT) divisions for conducting comparisons among topics used in tweets in each of the subgroups. Now, for each of the 10 following *topics*, think of 5 words, which, in your opinion, can best represent each of the topics:

Topic List: `Economy, Immigration, Health Care, Military, Gun Control, China, Trade, Race, Climate Change, Religion`

Visualizations: Using the concrete techniques you have learned in the class, parse the tweets in each subcategory into a corpus, and compare the levels of the usage of each topic in the following visualizations (for the static visualizations append all the 10 data files you have)

- Two static visualizations comparing the levels of the usage of each of the *topics* for a) Between the two Democratic candidates b) Between the two Republican candidates

- One static visualization comparing the levels of the usage of each of the *topics* between the democrats (as a whole) and the republicans
- Six dynamic visualizations for each of the candidates, and for democrats and republicans as a whole, showing the *daily* change in the level of the usage of each of the topics during the time period you are considering (March 16 to March 25)

For generating dynamic visualizations you build, use transition and transformation techniques you have learned, and the following plot, at this link, as a guide. Although you do not have to adhere to this single design, feel free to explore different methods of comparing the usage of topics among each pair (barcharts, bubble charts, rings, etc), and use either R or D3.

Report part 1: In 600 words describe the contrasts between 1) the two parties 2) the two democrat candidates 3) the two republican candidates. Why do you think such contrasts exist? You can explain the substantive differences among the campaigns if you are aware of their political positions. Do your findings map onto your expectations about the candidates' and parties' positions?

Report part 2: In 400 words describe the transformations of the daily frequency profiles for topic usage, across the ten days you are considering. Do you see any major refractions/transformations in the usage of a topic by a candidate or a party? Why did such transformations (if any) occur? Try to pinpoint the real world events that caused such shifts.

Extra points: Using the first order topic detection you implemented above, explore the tweets containing your keywords for each topic. Per topic, what are the other most frequent words in the corpus? Are they good candidates for augmenting your topic's "bag of words"? Use the most frequent words which could be good additions to your topic-forming list of words (additional 5 for ten words per topic). Run your topic detection algorithm again, and see if any of your results are different from those with 5-word topics. Explain the differences, if any. When do you think you should stop adding words to the topic-defining lists?