Group Members: Azka Javaid and Caleb Ki (Group3)
Title: WikiLeaks Disclosure Analysis

Purpose: We are interested in performing a time series analysis of the WikiLeaks disclosures and comparing the bias in traditional forms of WikiLeaks reporting through media/radio broadcasting with more recent news developments through social media platforms like Facebook and Twitter. In particular, we want to compare biases around the DNC disclosure.

Analytic Ideas/End Products: We are interested in primarily using the Event Exporter service from GDELT. We would gather the data by specifying a time range for a chosen WikiLeaks event (like DNC, for example). GDELT data has four numeric variables of interest (NumMentions, NumSources, NumArticles and AvgTone). In addition to using these variables to gauge differences in reporting, attitudes and sentiments surrounding WikiLeaks, we would use the SourceURL predictor to perform text analytics on the article linking to the WikiLeaks mention. We plan to scrape Twitter and Facebook data from the APIs. Additionally, we can analyze Google Searches to gauge the average individual's response to such disclosures in addition to the media. Text analytics and Natural Language Processing will be performed by IBM's AlchemyAPI, which parses through the URL and the underlying article to pull key insights, keywords, tone, and entities. Additionally this feature is dynamic, which means that insights pulled on one day from a generic news site like Yahoo will be different from insights pulled the following day since the content of the news typically varies daily. We then plan to perform unsupervised learning techniques like clustering on insights to better understand if WikiLeaks disclosures and the sentiments/tone around the news coverage can be distinctly categorized. Lastly, we plan to show changes in news biases globally through a choropleth mapping (http://blog.gdeltproject.org/new-one-minute-maps-bigquery-udf-cartodb/) and plan to build a ShinyApp to show changes in entities, keywords and sentiments covered by the various news media around WikiLeaks disclosures. If we have time we plan to write a script to pull the text from the actual articles from the sourceURL. We would do a text analysis to complement the clustering as by looking at the actual text of each article we could possibly gain insight into how and why the clustering algorithm grouped the articles.

Data: Besides the Event Exporter service (that provides a raw account of events around the world as well as the actors involved), GDELT also contains an Event Geographic Network, Heatmapper, Timeline, Network map (spreadsheet of most importance influencers in an event), Word Cloud and Tone Timeline, all freely accessible. These datasets will be informative for visualization contexts.

Variables:
Event Action Attribute:
GlobalEventID (integer): uniquely identifies each event record
Day (integer)/MonthYear(integer)/Year/FractionDate: Identify time of event
NumMentions (integer): total number of mentions of this minute across all source documents during the 15 minute update in which it was first seen.

NumSources (integer): total number of information sources containing one or more mentions of this event during the 15 minute update in which it was first seen.
AvgTone (numeric): average tone of all documents containing one or more mentions of this event during the 15 minute update in which it was first seen. Ranges from -100 (extremely negative) to +100 (extremely positive).