Allyson Mackay
Project Proposal
The Guardian Reporting, the BitCoin Market Value, and Tweeting

1. I want to look at Bitcoin market value data over the last 6 months, the number of article published in The Guardian retrieved by searching for 'bitcoin' and their sentient skew (positive/negative), along with data retrieved from the Twitter API that queries for tweets about Bitcoin. I will perform sentient analysis of the article using the datum sandbox api to determine if the article is negative or positive. Then I will look at negative and positive articles published and the market value of Bitcoin. Ultimately, I am looking to see if articles published in The Guardian about Bitcoin seem to have any relationship with the actual market value and if the slant of an article seems to have any relationship to the market value. I also plan to include the Twitter API. Twitter Get Search will relevant tweets for the query string "bitcoin". I plan to limit to English language tweets in order to do a twitter sentient analysis using the Datum Sandbox Api for sentient analysis. I will compare the tweets about bitcoin to the bitcoin market values and transactional volumes. I will also compare tweets, guardian articles, and bitcoin to look for interesting relationships between the two.

2.The bitcoin data is a csv file with a timestamp (daily), an opening value, a daily high, a daily low, closing value, volume information, and weighted price. The file is relatively small with one daily entry for the last 6 months (August 8, 2013 – February 3, 2013). Querying the Guardian for bitcoin related articles returns

Twitter Get Search returns JSON formatted data. My Get Search request will specify the following parameters: language, since_id, until, result_type. The result_type parameter retrieves only the most popular tweets. I plan to do two different data pulls from the Twitter api. One specifying result_type to get the most popular posts and one without. The relevant fields are the time given by the created_at key and the tweet content provided by the description:key. The file size will vary because Twitter does not index or store all tweets.

The Guardian API has a content search feature which takes a keyword and date parameters and outputs article date, title, and url in either JSON or XML. There are 149 articles for this 6 month time period. The Datum API will take both Twitter and Guardian urls and perform sentient analysis.

3. I will need to loan the csv bitcoin file and read the rows and extract only the values of interest. I am thinking to either extract market close value and/or perhaps calculate the daily difference between market open and close. Once extracted, this will be put into a dictionary. The JSON/XML files for Twitter and the Guardian will be retrieved and read into Python. I will iterate through and extract the date and url and create separate lists/dictionaries. I will also remove erroneous data from the bitcoin file and standardize date format across all there files. In order to process the sentient analysis of each, I will need to send the url through the datum api and take the value it returns and append it to the list/dictionary for each source. Essentially, I will be adding a sentient value to each tweet or article. The resulting output (CSV or plotted directly) will look at media saturation vs bitcoin value and an either positive/negative skew.

4. An interesting visualization would be graphs that chart the rise and fall of bitcoin against the number of articles published and tweets about bitcoin and if these are negative or positive. I am also thinking about doing a tag cloud of the words from articles/tweets that are published about bitcoin if their date of publication corresponds to a considerable peak or fall in the bitcoin value to see what sort of language might be associated with these.