Analyzing ‘Missouri’ Tweets

Terry Ballou-Crawford

Terry Ballou-Crawford
5/7/16

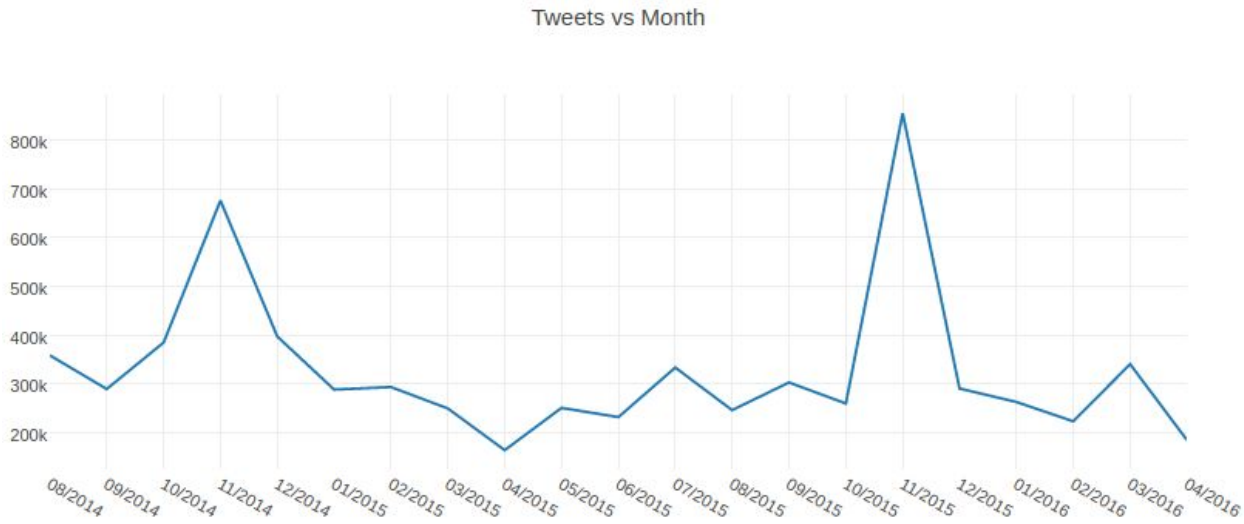
Descriptive Statistics

1. How many tweets are in the collection?
    a. In the tw4_db collection, there are 963,611,753 rows. According to the docs this is only a rough estimate using the command:
        i. SELECT SUM(TABLE_ROWS) FROM INFORMATION_SCHEMA.TABLES WHERE TABLE_SCHEMA = 'tw4_db';
    b. In the Tweet table alone there are:
        i. 215,968,516 rows
    c. In my Missouri subset, there are 6,888,314 rows
2. When do they start?
    a. From the Tweet and Missouri table, the first entry is:
        i. 2014-08-05 12:25:26
3. When do they end?
    a. From the Tweet and Missouri table, the latest entry is:
        i. 2016-04-21 21:35:57
4. What is the trend for tweet volume?
    a. The results for the Tweet table are below, grouped by month. The earliest and latest months of 2015 experienced the most Twitter traffic for data collected, with a few popular early 2016 months thrown in.

DATE_FORMAT(created_at,'%m/%Y')	count(*)
01/2015	18909295
12/2015	16666281
11/2015	15768330
01/2016	15458980
02/2016	14956411
10/2015	13836101
02/2015	12900174
09/2015	11995710
11/2014	10517843
03/2016	8949121
12/2014	8808602
07/2015	8685637
06/2015	8597748
04/2016	8087416

05/2015	7981375
08/2015	7835104
03/2015	7289049
10/2014	6514247
04/2015	5050672
09/2014	3654252
08/2014	3506168

5. If you look at the most common words over the lifetime of the search, do you notice any particular trends associated with those words?
    a. November was by far the most popular month for tweets containing 'Missouri'. In each November month contained within the Missouri dataset, there was twice as many tweets in November than the neighboring December and October months. The October and December counts of tweets were relatively similar to tweet counts the rest of the year.



6. What external events might correspond with the differences in the trends of most common words?
    a. For the popularity of Missouri tweets in November 2015, the most likely external event is the Missouri Tigers football team boycotting practice to convince the UM President and Chancellor to step down. This was an issue that was nationally watched and thus garnered much more attention than the fly-over state typically receives. For November 2014, the social issues in

Ferguson were most likely the reason of Tweets about Missouri. This was also a national issue that received much media coverage.

7. What hashtags show up as most prominent in each month of the lifecycle?
    a. The hashtag table does not have a date field. Joined it with the tweet table and grouped by text of hashtag, year of tweet, and month of tweet. The top 20 results are returned. This query is not running in a reasonable amount of time. Check out the SQL in the python1.py script.
8. Which twitter users are the most mentioned?
    a. This query is not running in a reasonable amount of time. Check out the SQL in the python1.py script.
9. How frequently is each user mentioned during each month of the lifecycle?
    a. Because the most mentioned query is not running in a reasonable amount of time, this one cannot be answered. I wrote the code that iterates over the top 20 most mentioned users and groups their mentions by month and prints them out the counts.
10. What is the relationship between the volume of tweets you selected and the volume of tweets for other collections in the data set?
    a. My Missouri dataset has 6,888,314 rows while the Tweet table has 215,968,516 rows. This means Missouri accounts for approximately 3% of the collected data.

Identifying Research Questions

1. Select one research question to pursue.
    a. What are the most popular hashtags for Tweets containing ‘Missouri’?

Answering Research Questions

1. Prepare the data set for analysis
2. Create a GitHub Repository for your final project
3. Describe the specific data set you are using for the final project (which job_id’s).
    1. I created my own dataset from the Tweet table from all tweets containing ‘Missouri’ (case insensitive).
4. Build one directory to answer the research question. Include all code and data set references (SQL) there.

5. REPORT : For the research question:
    1. Describe the data carpentry work and software carpentry work you did to obtain both descriptive statistics and answer the questions
    i. The data carpentry that I did for this project involved creating a new SQLite database from the MariaDB database using a subset of the tweet data. All the types matched up except for geolocation which were cast to strings when they were inserted into the SQLite table. All data to answer questions about the Missouri dataset is therefore pulled from the SQLite database.
    2. Provide one, short 3-5 paragraph explanation of your results for your question
    i. My findings were that of tweets containing 'Missouri', the hashtag #Missouri was by far the most popular. Of the 2,316,055 tweets with a hashtag among the Missouri dataset (about ⅛ of all tweets mentioning Missouri), 458,977 were of #Missouri. This was nearly 20% of all the Missouri tweets. This equated to being .21% of the entire tweet table. Considering that the Missouri dataset spiked significantly in the Novembers contained within the tweet table, much of this traffic was most likely not just from users in Missouri. Other popular hashtags included #Ferguson, #ConcernedStudent1950, #MichaelBrown, #MikeBrown, part of the November issues that blipped Missouri's popularity on the national radar, albeit not in a good way. Many other hashtags include states such as #Mississippi, #Florida, #Illinois, #Ohio, etc. Much of this traffic is most likely due to sporting and political events. Politics also show up including #Trump2016, #TrumpTrain, #MakeAmericaGreatAgain, #FeelTheBern, #and #VoteTrump. Much of the states being hashtagged are most likely due to events there, including political debates and elections.
6. Submit your code to GitHub
7. Submit your REPORT to Canvas. INCLUDE a LINK to the GITHUB repo on the cover page of your report.