John Fields

Dr. Gregory Block

IST652 - Homework #2

August 21,2019

Introduction

With 321 million active monthly users, the social media platform Twitter has had a global impact on the way we consume news and communicate.[i,ii] Twitter started by allowing users to only send short messages of 140 characters and in 2017 they doubled the length to 280 characters. Twitter has also made an Application Programming Interface (API) available for developers to download these "tweets" in real-time for storage in databases where additional analysis can be performed.

The goal for this project is to download tweets from Twitter with hashtags related to Los Angeles traffic such as #LAtraffic and use this information in the final project for this class (Los Angeles Traffic Collisions). Hashtags in Twitter are preceded with the # symbol and are used to group together people who have similar interests so they can "follow" others or search based on these hashtags. Through the use of API's this process can be automated to collect real-time tweets using Python software and a Mongo DB database on a local computer.

Data and Source

The data set for this project consists of the tweets that were collected over several days using the following hashtags:

- #LATraffic
- #latraffic
- #LATraffic24
- #latraffic24
- #lacrash
- #laaccident

The hashtags above did not return the desired goal of approximately 300 tweets, so the following hashtags were added:

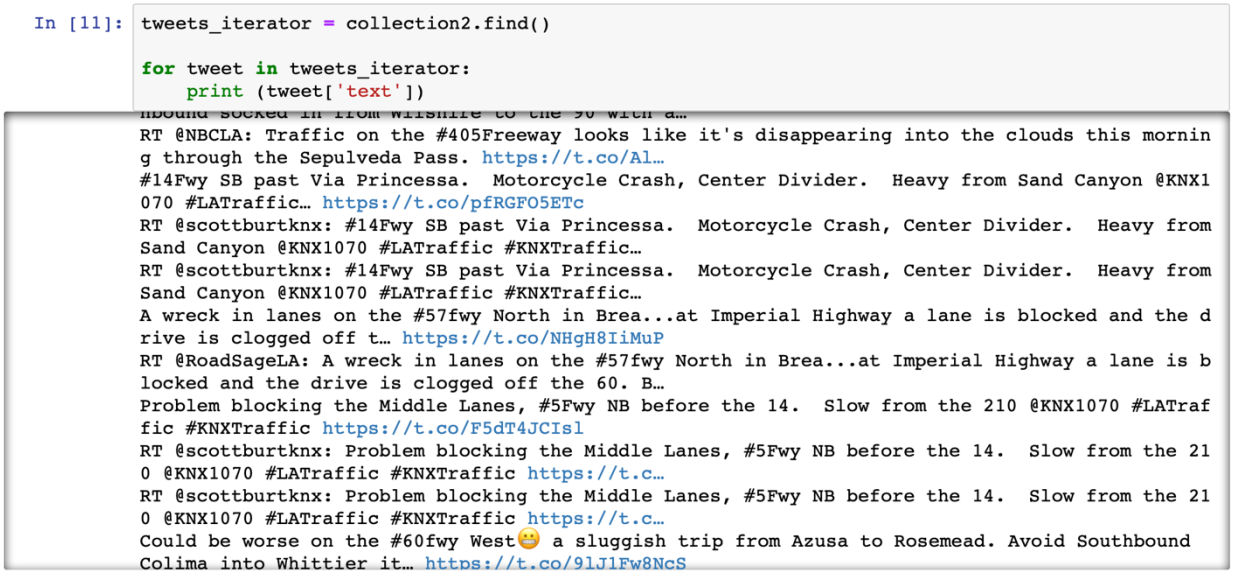
- #crash
- #traffic

A total of 339 tweets related to traffic were collected during this time that will be used in this analysis. It should also be noted that during testing, the hashtag #trump was used since there are several tweets per second using this hashtag. However, the results from these tweets will not be included in this paper due to the offensive nature of some of the tweets.

Data Exploration and Cleaning

The requirements for this assignment provided the option to pull files from a file for content from Facebook or Twitter. Although much more challenging, the decision to sign up for a Twitter API and spend over 20 hours debugging the code was beneficial since the knowledge gained will be very useful in the future.

Once the software coding was complete to download the tweets, the next step was to understand how to query the Mongo DB for the desired information. The initial exploration included a review of the full text of each tweet without having any details on the time the tweet was sent and whether it was a "retweet" where someone re-posts another person's tweet. See Figure 1 below for an example of the output of tweets from MongoDB.



```
In [11]: tweets_iterator = collection2.find()

         for tweet in tweets_iterator:
             print (tweet['text'])
```

```
nbound locked in from Wilshire to the 90 with a…
RT @NBCLA: Traffic on the #405Freeway looks like it's disappearing into the clouds this mornin
g through the Sepulveda Pass. https://t.co/Al…
#14Fwy SB past Via Princessa.  Motorcycle Crash, Center Divider.  Heavy from Sand Canyon @KNX1
070 #LATraffic… https://t.co/pfRGFO5ETc
RT @scottburtknx: #14Fwy SB past Via Princessa.  Motorcycle Crash, Center Divider.  Heavy from
Sand Canyon @KNX1070 #LATraffic #KNXTraffic…
RT @scottburtknx: #14Fwy SB past Via Princessa.  Motorcycle Crash, Center Divider.  Heavy from
Sand Canyon @KNX1070 #LATraffic #KNXTraffic…
A wreck in lanes on the #57fwy North in Brea...at Imperial Highway a lane is blocked and the d
rive is clogged off t… https://t.co/NHgH8IiMuP
RT @RoadSageLA: A wreck in lanes on the #57fwy North in Brea...at Imperial Highway a lane is b
locked and the drive is clogged off the 60. B…
Problem blocking the Middle Lanes, #5Fwy NB before the 14.  Slow from the 210 @KNX1070 #LATraf
fic #KNXTraffic https://t.co/F5dT4JCIsl
RT @scottburtknx: Problem blocking the Middle Lanes, #5Fwy NB before the 14.  Slow from the 21
0 @KNX1070 #LATraffic #KNXTraffic https://t.c…
RT @scottburtknx: Problem blocking the Middle Lanes, #5Fwy NB before the 14.  Slow from the 21
0 @KNX1070 #LATraffic #KNXTraffic https://t.c…
Could be worse on the #60fwy West😬 a sluggish trip from Azusa to Rosemead. Avoid Southbound
Colima into Whittier it… https://t.co/9lJlFw8NcS
```

Figure 1 - Sample of collected tweets from MongoDB

The NoSQL Booster is a Graphical User Interface (GUI) which was very helpful to understand the structure that Twitter uses for the API to then develop more detailed queries.

Structure of Mongo DB:

Database = twitterdb2

Collection = twitter_search

Field = text

```
RT @scottburtknx: #210Fwy EB before Osborne Street.  3
    Right Lanes Blocked, Injury Crash.  JAMMED from the
    118 - #118Fwy EB heavy from the…
```

Figure 2 - Example of tweet viewed in NoSQL Booster

Although the number of tweets with geolocation data was limited, the following code in Python was used to locate the precise location where this tweet originated (blue dot in Figure 2).

text	#indoosmoke #latraffic @ Puffpuffpass420 https://t.co/YTpW9l9gyd

coordinates	Array[2]
0	34.084
1	-118.041

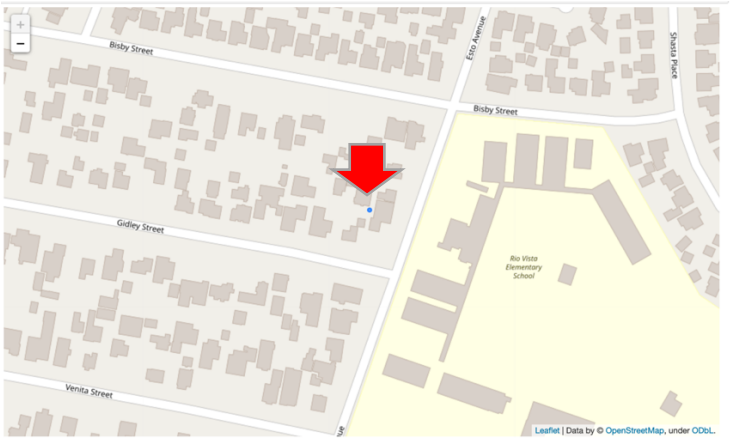


Figure 3 - Location of tweet from 34.084, -118.041

In addition to understanding the location of the tweet, the next steps was to query the Mongo DB to find how many tweets have the text "latraffic". NoSQL Booster was used to construct the code that provided the result shown in Figure 3 and summarized the number of tweets (12).

```
#LATraffic is the worst yooo.. and it’s like this everyday smh 😞
https://t.co/RaLFioxZy5
#14Fwy SB past Via Princessa.  Motorcycle Crash, Center Divider.  Heavy fr
om Sand Canyon @KNX1070 #LATraffic… https://t.co/pfRGFO5ETc
RT @scottburtknx: #14Fwy SB past Via Princessa.  Motorcycle Crash, Center
Divider.  Heavy from Sand Canyon @KNX1070 #LATraffic #KNXTraffic…
Problem blocking the Middle Lanes, #5Fwy NB before the 14.  Slow from the
210 @KNX1070 #LATraffic #KNXTraffic https://t.co/F5dT4JCIsl
RT @scottburtknx: Problem blocking the Middle Lanes, #5Fwy NB before the 1
4.  Slow from the 210 @KNX1070 #LATraffic #KNXTraffic https://t.c…
Relax In Los Angeles City Daily News is out! https://t.co/OAO5nAEQRq #405f
reeway #latraffic
#indoosmoke #latraffic @ Puffpuffpass420 https://t.co/YTpW9l9gyd
The latest Breaking LA News! https://t.co/JBei8RFa9r #california #latraffi
c
RT @scottburtknx: Stalled Car just cleared, #134Fwy WB at Lankershim Blvd.
  Slow from Hollywood Way @KNX1070 #LATraffic #KNXTraffic https:/…
12
```

Figure 4 - Output of MongoDB query for "latraffic"

The final step in the processing of the tweets was to import the MongoDB data to Python using a pandas dataframe.

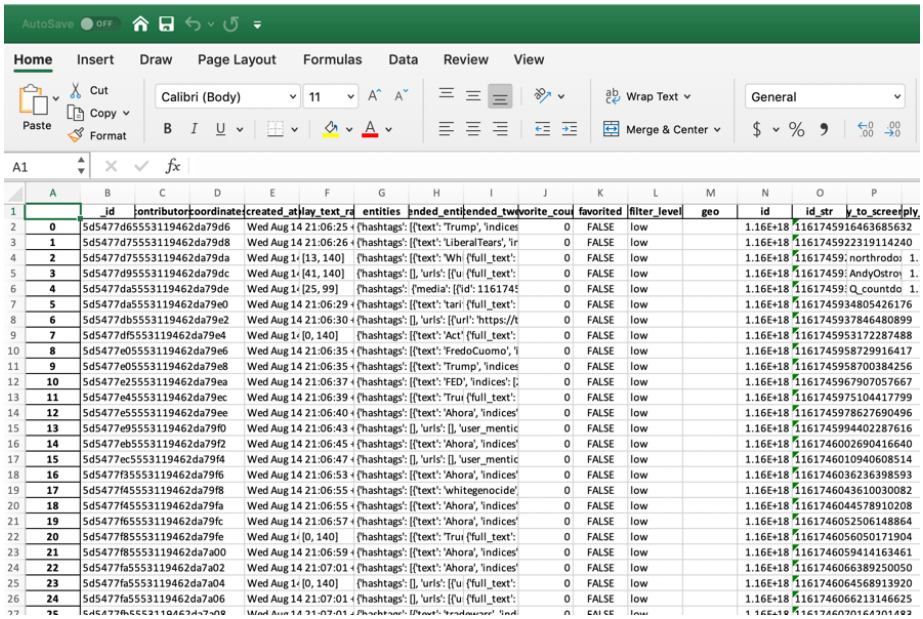


Figure 5 - Sample output of dataframe export to Excel

Description of the Program

The Python program to import the dataset, format the data and perform the analysis was developed in Jupyter Notebook with MongoDB. The project code was split into the following sections:

1. Twitter API to MongoDB

This was the most challenging part of the project due to the limited information online for setting up the Twitter API and then writing the collected data to MongoDB. Although difficult and frustrating at times, the effort was worth the struggle to now understand how to create a real-time feed from Twitter.

2. MongoDB Queries

There were similar challenges to query the information that was stored in MongoDB. The vast amount of detailed Twitter information is overwhelming, and the queries are much more complex. However, NoSQL Booster has a visual query tool that was extremely helpful in writing the queries to explore the information contained in the 339 tweets that were collected.

Conclusion

The dramatic increase in the use of social media platforms such as Twitter has changed the way we communicate and interact around the world. Having the skills and knowledge to "mine" this data is useful skill for anyone working in the computer science field. The information and skills gained through this assignment will be invaluable in the future.

---

[i] Shaban, Hamza. "Twitter Reveals Its Daily Active User Numbers for the First Time." *The Washington Post*, WP Company, 7 Feb. 2019, www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/?noredirect=on.

[ii] Cresci, Elena. "12 Ways Twitter Changed Our Lives." *The Guardian*, Guardian News and Media, 21 Mar. 2016, www.theguardian.com/technology/2016/mar/21/12-ways-twitter-changed-our-lives-10th-birthday.