**Twitter Data and MongoDB**

MongoDB’s flexible, document-based approach to data storage lends itself well to the data of Twitter. Despite my limited exposure to MongoDB I am already a fan. Python is also extremely flexible and my favorite programming language of all. Combining a final favorite of mine, data science, I have decided to download the tweets of some prominent data scientist for the last few months: Hilary Mason (@hmason), John Myles White (@johnmyleswhite), Kaggle (@kaggle, I know this is a website/organization and not a data scientist), Pete Skomoroch (@peteskomoroch), Ryan Rosario (@DataJunkie), and Dj Patil (@dpatil). I got the suggestions from an article1 that I glanced at on kdnuggets.com.

All of the files related to this work (with the exception of the data) is hosted on my github repo: <https://github.com/murrietta/mongo_tweets>.

**Getting Tweets**

Initially I downloaded the tweets, saved them to a json file, and uploaded them to my database. I know I could just use python to write them directly after download but I wanted to keep them in a text file just in case. After this initial try I then tried a few queries and realized that some of the tweets were not stored as dates so I modified my download code to convert the ‘created\_at’ field of all tweets to a unix timestamp times 1000 which is at least one of the formats that is required when uploading data from json to mongoDB. A benefit of doing this is seen at the end of the console output in figure 1, the date range that the tweets in my database span is 2016-11-05 18:25:21 to 2017-04-02 22:12:25 over 12,522 tweets. In the final version of this code I have commented out the line that prints the line “Current tweet created\_at…” as this was for debugging purposes.

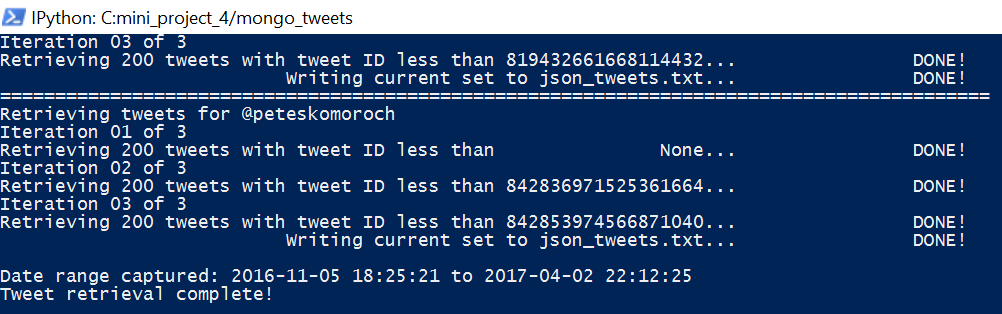


Figure Data Download Final Line of Output

**Loading to MongoDB**

Since my data is in json format within a text file I was able to make a simple powershell script that invoked the mongoimport.exe. Figure 2 shows the output of running mongoimport.exe from the console. The script itself actually changes directory to C:\Program Fiels\MongoDB\Server\3.4\bin, invokes mongod.exe, then invokes mongoimport.exe.

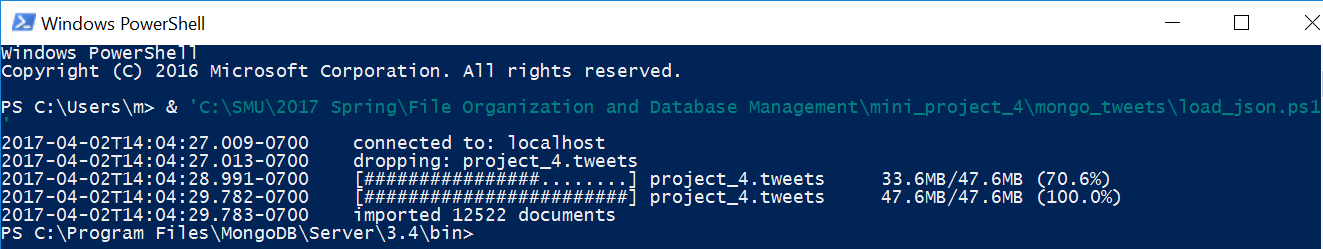
**Analyzing**

Figure Loading tweets to MongoDB via mongoimport.exe

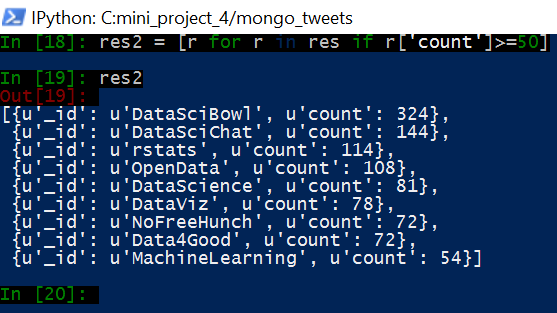
Hashtags were a brilliant idea with twitter since it essentially gets twitter users to self-identify their tweets. My first step was to figure out the hashtags used based on frequency. Figure 2 shows the 9 that were used more than 50 times. I looked up “data science bowl” and found out that the main purpose is to improve lung cancer detection. This seems to be related, at least in concept, to the “Data4Good” hashtag so I decided to see if these shared similar trends over the date range captured in this dataset. Between python and mongoDB’s query language there is a plethora of ways to do this, I chose to focus on mongoDB’s query language and decided to group by the substring of the ‘created\_at’ field that is the first 7 characters, ‘2017-01’ for example. Both hashtags were used only during January, February, and March. #DataSciBowl peaked in January, #Data4Good peaked in February, and #DataSciBowl always had more counts that #Data4Good (a plot is shown in figure 4). The only apparent relation from this view of the data is that both of these have only shown up in January, February, and March of 2017 so far.

Figure Hashtags used more than 50 times

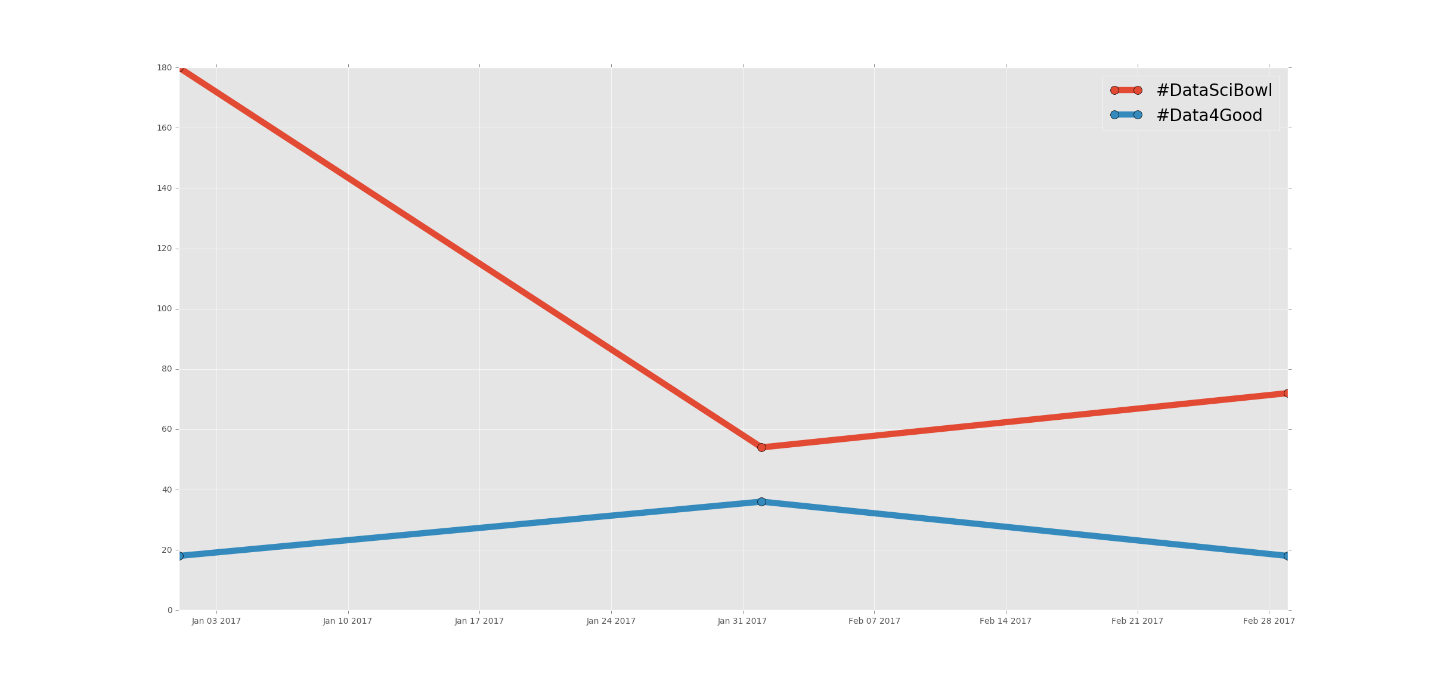
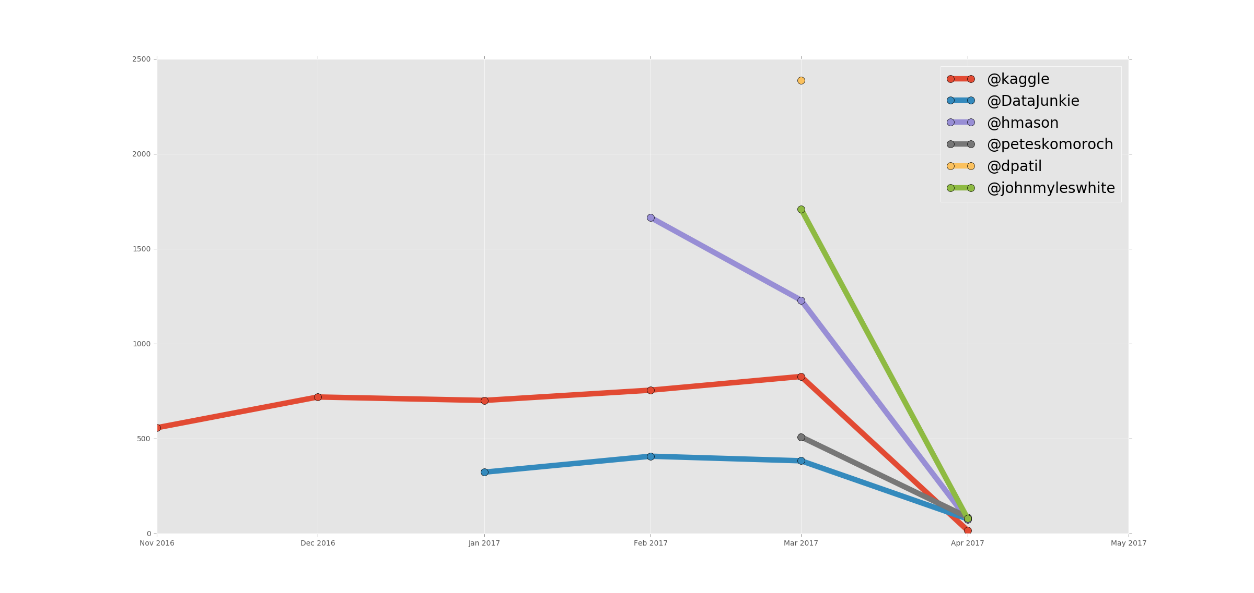
There is a plethora of ways to look at this data but I decided to do one more look, this time at counts of tweets by user over the months shown in the data. This followed a similar method as for the two hashtags but I extended the x-axis a bit to see a little bit more of how the April counts by user looked, the result is best summarized by the chart in figure 5. This trend for @kaggle kind of makes sense, since it is a website/organization it is likely that the team that manages the Kaggle.com twitter account is off for the weekend and so there is almost 0 activity in April (since today is Sunday April 2nd). It was curious that @dpatil only had an excessive amount of tweets in March 2017 but 0 for outside of that month. I printed just the ‘created\_at’ field to screen for all of Dj Patil’s tweets in this dataset just to see if the ‘created\_at’ field got messed up but it was fine.

Figure Counts by User by Month

Figure Count trends for 2017-01, 2017-02, and 2017-03

**Conclusion**

I did not find any methods that allow for downloading a time range of tweets independent of ‘user\_id’ or ‘screen\_name’ so that is why I chose to focus on some users that may likely tweet about things that I find interesting. This was a nice exposure to the twitter API and I will certainly use this periodically to perform investigations into the culture that exists within twitter. MongoDB continues to impress me, especially with the quickness at which it can query documents in a collection. As a result of my work with MongoDB in this course I am going to construct a case for why our organization should store care notes in MongoDB, one of the main reasons being the ease at which we would be able to create programs that can do chart review that is currently done manually.

**References**

1. Lotan, Gilad. *Most Influential Data Scientist on Twitter and Quora*. <http://www.kdnuggets.com/2012/12/most-influential-data-scientists-on-twitter.html>. Retrieved 2017-03-29 at 16:25.