# find-top-hashtags-twitter

Do real-time analysis on the tweets, find the top #Hashtags from twitter data

Tools & Technologies:

Tweepy (And your own pair of API Keys from Twitter);

Pyspark (Python 3.7, Spark 2.4);

Jupyter Notebook;

Spark Streaming;

Spark SQL;

pymongo;

MongoDB;

Set your credentials on config.json file to get Twitter API access.

{ "asecret": "XXX...XXX",

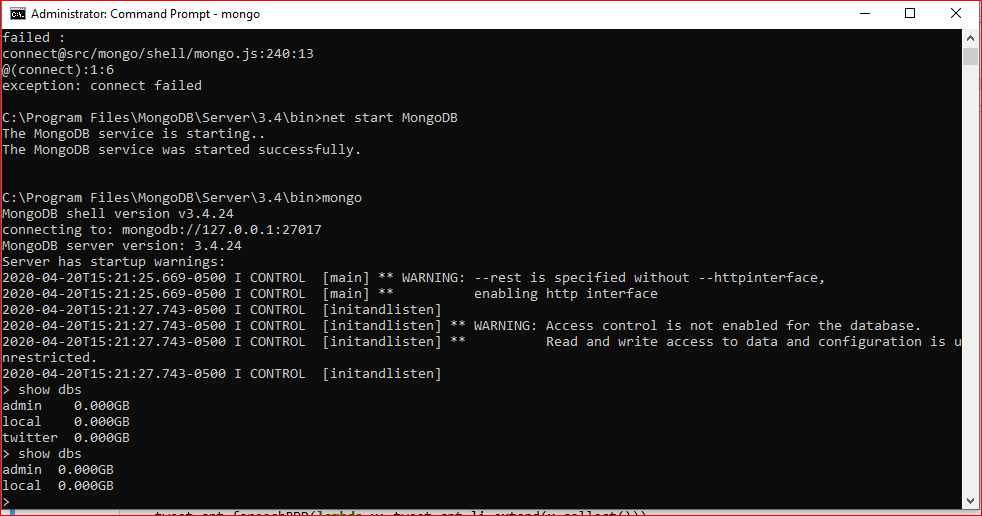
"atoken": "XXX...XXX",

"csecret": "XXX...XXX",

"ckey": "XXX...XXX" }

Modify the parameters.json file to set your own parameters

Start MongoDB Database process to store the data into MongoDB



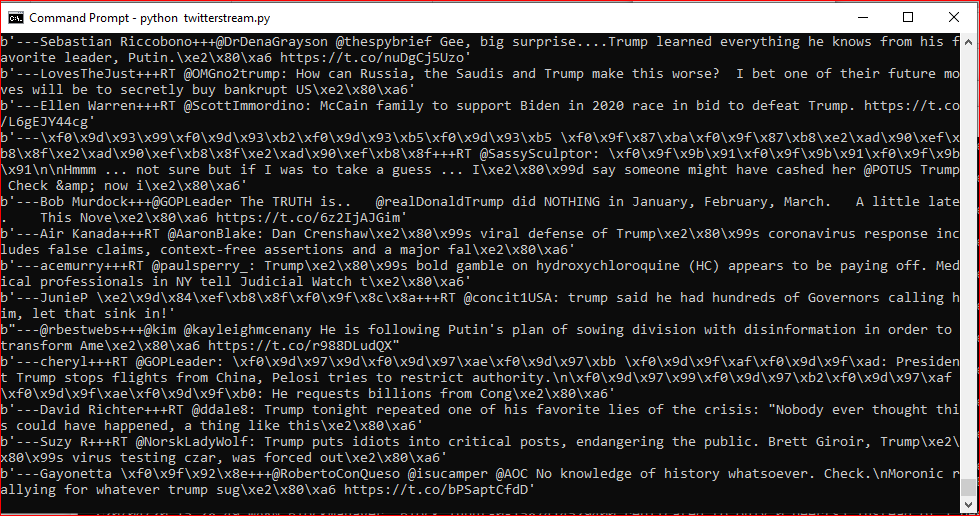
Implementation:

Twitter API

Use the tweepy Twitter API to stream tweets

Filter out the tweets containing the specific keywords / hashtags we want to track.

To give the tweets to the spark job, using TCP / IP socket

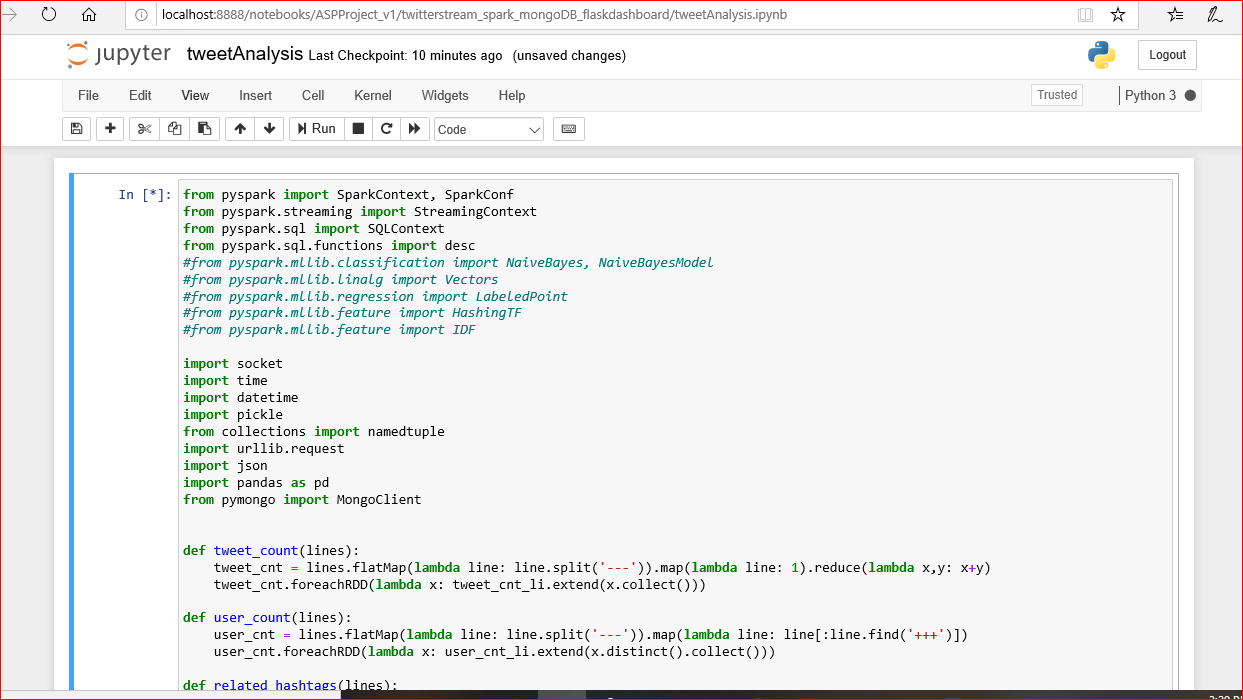


Real-time Analysis

Using Spark Streaming to real-time tweet review

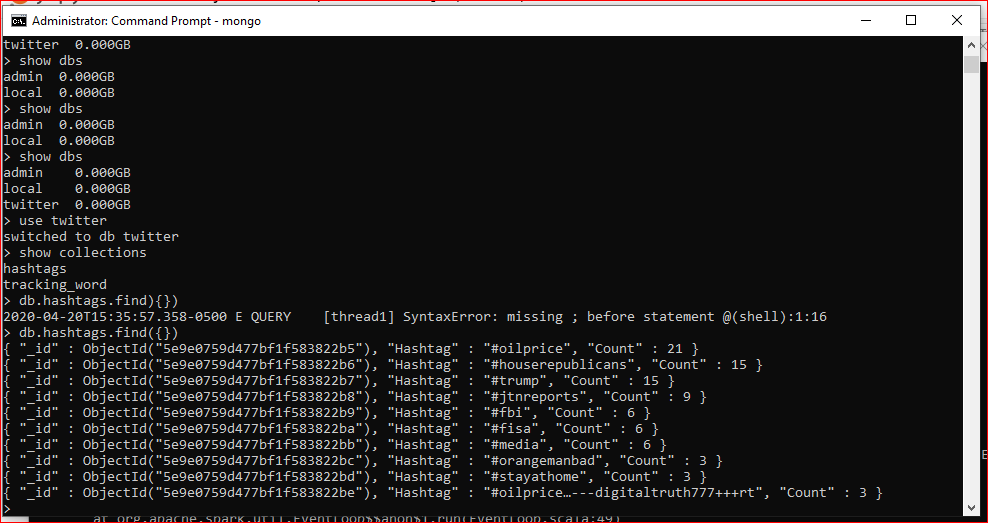
Clean the tweets

Locate the most important hashtags



Database

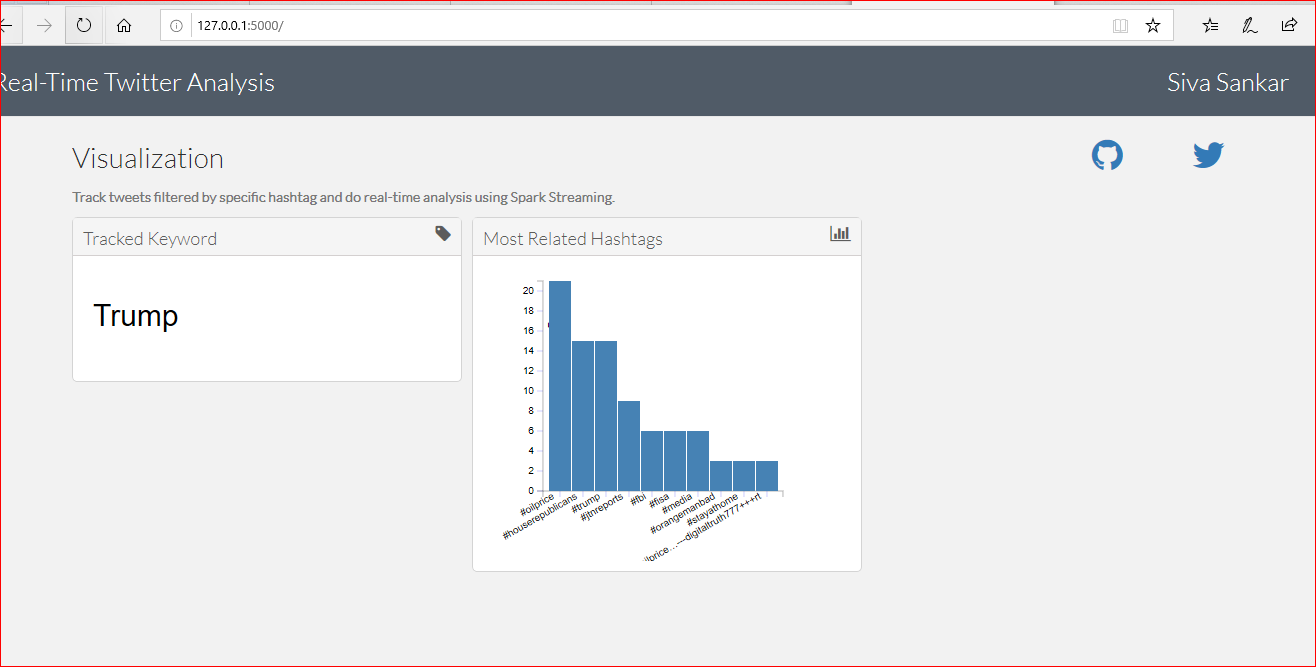
Using MongoDB to store the results of an study



Visualization (Output)

Time line of related most related hashtags, search keyword

Now can browse localhost:5000 to see the output



HOW TO RUN:

1. run twitterstream.py file from command prompt

python twitterstream.py

2. Then run tweetAnalysis from jupyter notebook (open anaconda prompt and type "jupyter notebook", this opens jupyter notebook

3. Then run dashboard/dashboard.py file from the command prompt

py dashboard/dashboard.py

4. Now can browse localhost:5000 to see the output