

University of Missouri – Kansas City

CS5990: Big Data Programming

Summer 2020

PROJECT 2

By: Uyen Dang (16227610)

07/29/2020

1. PROJECT DESCRIPTION

--Individual Contribution:

- Part 1: Hadoop Map Reduce Spark - Finding Facebook Common Friends
 - Gabriella Willis:
- Part 2: Spark Data Frame
 - Elizabeth Nastoff (part a+b):
 - Jun Yang(part c+d): <https://github.com/jun0405/CS490/wiki/Lab-2>
- Part 3: Spark Streaming Task
 - Uyen Dang: <https://github.com/uyendang/CS5590---Big-Data-Programming/wiki/Project-2>
- Part 4: Spark GraphX Task
 - Brett Recker: <https://github.com/brettrecker/CS5590-BDP/wiki/Project-based-Exam-%232>

--Source Code:

- https://github.com/uyendang/CS5590---Big-Data-Programming/tree/master/Project_2

--Technology Used:

- GitHub
- Twitter Developer
- PyCharm

-- The Idea:

- In this project, I created a Twitter Developer account and get credentials then I extracted tweets through socket. I filtered data in the trending topic of "corona virus COVID19 SARSCOV2" hashtag. Then, I used PyCharm to group and count the number of appearances of the hashtags found.

-- Project in Today's World:

- This type of project could be used to predict what is the trend in Twitter and how many people tweet at what rate and what is the frequency of that trend.

-- The Portion I Worked:

- I completed part 3 - Spark Streaming Task.

-- Challenges:

- The most difficult part was the algorithm and Twitter Developer account need several days to get approved.

-- The Milestones and Teammates:

The teammates were all down-to-earth and awesome, we helped each other whenever possible and they are all organized and always on-time.

We had 9 days to complete this project.

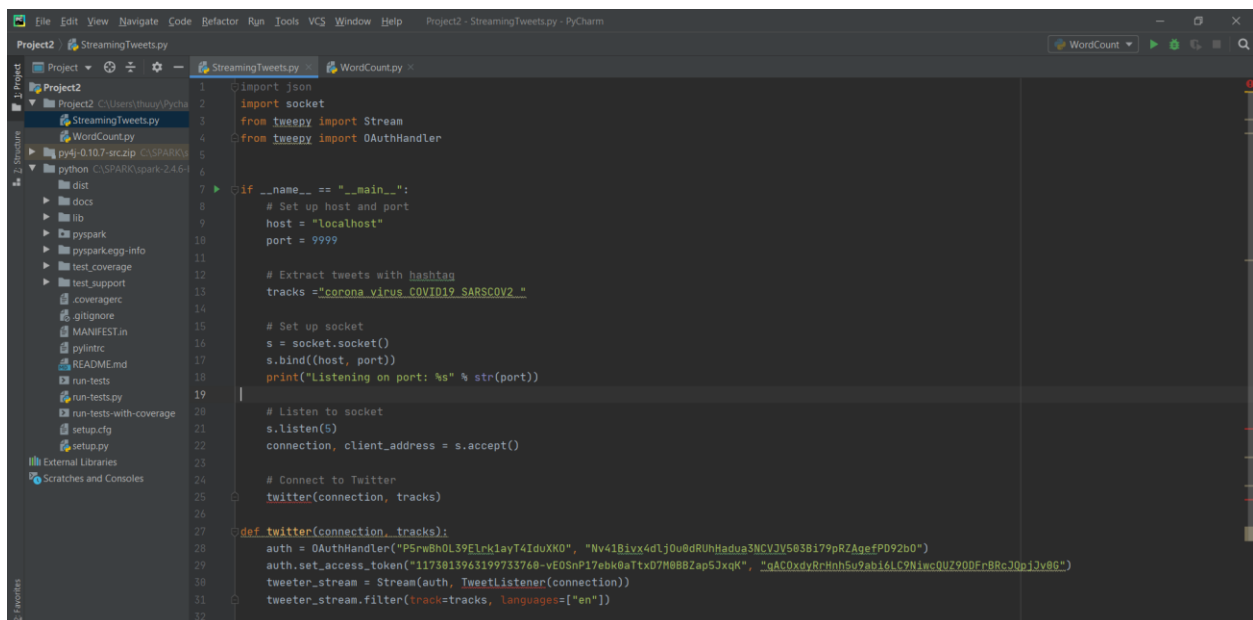
- On 7/22/2020 We divided parts among all team members
 - On 7/29/2020 We set the goal for each individual to complete their parts on this date
 - On 7/29/2020 We planned to do a final check and record videos for our project.
-

2. Twitter Spark Streaming

---Step 1: Create a Twitter account and gather credential information.

- api_key = "P5rwBhOL39Elrk1ayT4IduXKO"
- api_secret = "Nv41Bivx4dljOu0dRUhHadia3NCVJV503Bi79pRZAgefPD92bO"
- access_token = "1173013963199733760-vEOSnP17ebk0aTtxD7M0BBZap5JxqK"
- access_token_secret = "qAC0xdyRrHnh5u9abi6LC9NiwcQUZ9ODFrBRcJQpjJv0G"

---Step 2: StreamingTweets.py



```
1 import json
2 import socket
3 from tweepy import Stream
4 from tweepy import OAuthHandler
5
6
7 if __name__ == "__main__":
8     # Set up host and port
9     host = "localhost"
10    port = 9999
11
12    # Extract tweets with hashtag
13    tracks = "corona virus COVID19 SARS-CoV2"
14
15    # Set up socket
16    s = socket.socket()
17    s.bind((host, port))
18    print("Listening on port: %s" % str(port))
19
20
21    # Listen to socket
22    s.listen(5)
23    connection, client_address = s.accept()
24
25    # Connect to Twitter
26    twitter(connection, tracks)
27
28    def twitter(connection, tracks):
29        auth = OAuthHandler("P5rwBhOL39Elrk1ayT4IduXKO", "Nv41Bivx4dljOu0dRUhHadia3NCVJV503Bi79pRZAgefPD92bO")
30        auth.set_access_token("1173013963199733760-vEOSnP17ebk0aTtxD7M0BBZap5JxqK", "qAC0xdyRrHnh5u9abi6LC9NiwcQUZ9ODFrBRcJQpjJv0G")
31        tweeter_stream = Stream(auth, TweetListener(connection))
32        tweeter_stream.filter(track=tracks, languages=["en"])
```

---Step 3: WordCount.py

---Step 4: Tweets that are being streamed and extracted

---Step 5: Work Count is performed on Twitter Streaming Data

Batch: 1

window	tags	count
[2020-07-29 14:39:30, 2020-07-29 14:40:20]	#	20
[2020-07-29 14:39:00, 2020-07-29 14:39:50]	#	20
[2020-07-29 14:38:00, 2020-07-29 14:38:50]	#MTVHottest	9
[2020-07-29 14:39:00, 2020-07-29 14:39:50]	#1's	9
[2020-07-29 14:39:30, 2020-07-29 14:40:20]	#1's	8
[2020-07-29 14:39:30, 2020-07-29 14:40:20]	#MTVHottest	8
[2020-07-29 14:38:00, 2020-07-29 14:38:50]	#1	8
[2020-07-29 14:37:30, 2020-07-29 14:38:20]	#1	7
[2020-07-29 14:38:30, 2020-07-29 14:39:20]	#1	6
[2020-07-29 14:37:30, 2020-07-29 14:38:20]	#MTVHottest	6
[2020-07-29 14:38:30, 2020-07-29 14:39:20]	#MTVHottest	5
[2020-07-29 14:38:30, 2020-07-29 14:39:20]	#10YearsOf1D	5
[2020-07-29 14:40:00, 2020-07-29 14:40:50]	#MTVHottest	5
[2020-07-29 14:39:00, 2020-07-29 14:39:50]	#1	5
[2020-07-29 14:38:00, 2020-07-29 14:38:50]	#BTS	5
[2020-07-29 14:37:30, 2020-07-29 14:38:20]	#BTS	5
[2020-07-29 14:40:00, 2020-07-29 14:40:50]	#1's	5
[2020-07-29 14:39:00, 2020-07-29 14:39:50]	#BTS	5
[2020-07-29 14:38:30, 2020-07-29 14:39:20]	#1's	4
[2020-07-29 14:38:30, 2020-07-29 14:39:20]	#BTS	4

only showing top 20 rows

Batch: 2

+-----+-----+-----+		
window	tags	count
+-----+-----+-----+		
[2020-07-29 14:42:00, 2020-07-29 14:42:50] #		43
[2020-07-29 14:42:30, 2020-07-29 14:43:20] #		23
[2020-07-29 14:40:30, 2020-07-29 14:41:20] #		21
[2020-07-29 14:39:30, 2020-07-29 14:40:20] #		20
[2020-07-29 14:39:00, 2020-07-29 14:39:50] #		20
[2020-07-29 14:40:00, 2020-07-29 14:40:50] #		20
[2020-07-29 14:41:30, 2020-07-29 14:42:20] #1's		12
[2020-07-29 14:41:00, 2020-07-29 14:41:50] #1's		11
[2020-07-29 14:42:00, 2020-07-29 14:42:50] #1's		9
[2020-07-29 14:38:00, 2020-07-29 14:38:50] #MTVHottest		9
[2020-07-29 14:39:00, 2020-07-29 14:39:50] #1's		9
[2020-07-29 14:39:30, 2020-07-29 14:40:20] #1's		8
[2020-07-29 14:39:30, 2020-07-29 14:40:20] #MTVHottest		8
[2020-07-29 14:38:00, 2020-07-29 14:38:50] #1		8
[2020-07-29 14:40:30, 2020-07-29 14:41:20] #1's		7
[2020-07-29 14:40:00, 2020-07-29 14:40:50] #1's		7
[2020-07-29 14:41:00, 2020-07-29 14:41:50] #10YearsOf1D		7
[2020-07-29 14:37:30, 2020-07-29 14:38:20] #1		7
[2020-07-29 14:41:30, 2020-07-29 14:42:20] #10YearsOf1D		7
[2020-07-29 14:40:30, 2020-07-29 14:41:20] #1		6
+-----+-----+-----+		

only showing top 20 rows

Batch: 3

window	tags	count
[2020-07-29 14:42:00, 2020-07-29 14:42:50] #		43
[2020-07-29 14:42:30, 2020-07-29 14:43:20] #		23
[2020-07-29 14:40:30, 2020-07-29 14:41:20] #		21
[2020-07-29 14:39:30, 2020-07-29 14:40:20] #		20
[2020-07-29 14:39:00, 2020-07-29 14:39:50] #		20
[2020-07-29 14:40:00, 2020-07-29 14:40:50] #		20
[2020-07-29 14:44:00, 2020-07-29 14:44:50] #1's		14
[2020-07-29 14:44:30, 2020-07-29 14:45:20] #1's		13
[2020-07-29 14:43:30, 2020-07-29 14:44:20] #1's		13
[2020-07-29 14:44:00, 2020-07-29 14:44:50] #1		13
[2020-07-29 14:44:30, 2020-07-29 14:45:20] #1		12
[2020-07-29 14:41:30, 2020-07-29 14:42:20] #1's		12
[2020-07-29 14:41:00, 2020-07-29 14:41:50] #1's		11
[2020-07-29 14:43:30, 2020-07-29 14:44:20] #1		10
[2020-07-29 14:43:00, 2020-07-29 14:43:50] #1's		9
[2020-07-29 14:42:00, 2020-07-29 14:42:50] #1's		9
[2020-07-29 14:38:00, 2020-07-29 14:38:50] #MTVHottest		9
[2020-07-29 14:39:00, 2020-07-29 14:39:50] #1's		9
[2020-07-29 14:39:30, 2020-07-29 14:40:20] #1's		8
[2020-07-29 14:39:30, 2020-07-29 14:40:20] #MTVHottest		8

only showing top 20 rows

3. REFERENCES:

- <https://www.linkedin.com/pulse/apache-spark-streaming-twitter-python-laurent-weichberger/>
- <https://github.com/stefanobaghino/spark-twitter-stream-example>