BIG DATA PROGRAMMING

PROJECT_BASED EXAM-2

Spark Streaming Task

OBJECTIVE:

Perform Word-Count on Twitter Streaming Data.

INTRODUCTION:

- Spark streaming helps to access and process Real-time data with the help of different algorithms like map, Reduce, join etc.
- In this project we use MapReduce algorithm to perform word count on streamed **data.**

QUESTION 3: (1)

IDEA OF THE PROJECT:

Idea of the project is to apply different concepts learnt in Big-Data-Programming so far. Here we use MapReduce method to calculate word frequency.

QUESTION 3: (2)

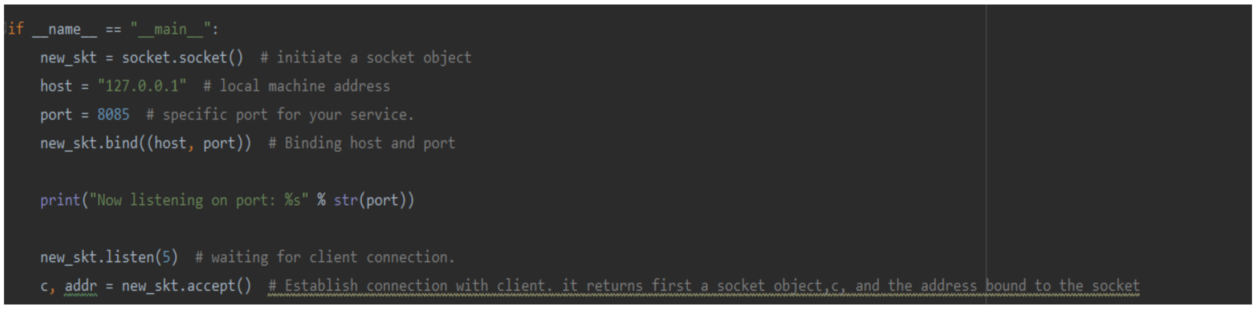
USAGE OF PROJECT IN REAL-TIME:

Spark Streaming Context is used for processing the real-time data streams. In real-time, this idea of streaming helps in prediction, analyzing and data processing workloads etc**.**

QUESTION 3: (3)

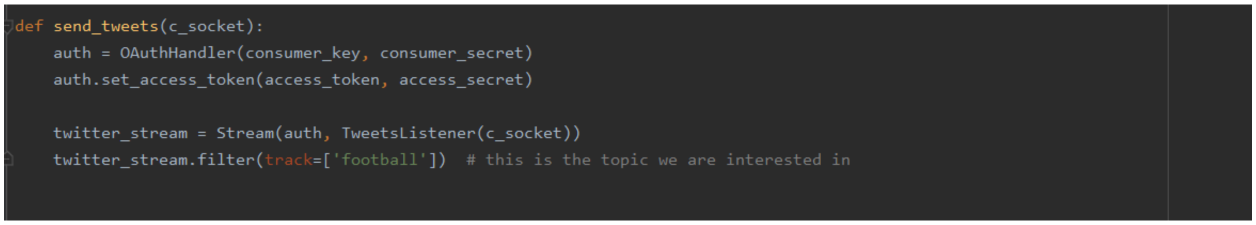
IMPLEMENTATION:

- Initiate a socket object with local machine's IP address and a service specific port number.
- Bind the host and port.
- Make client connection.



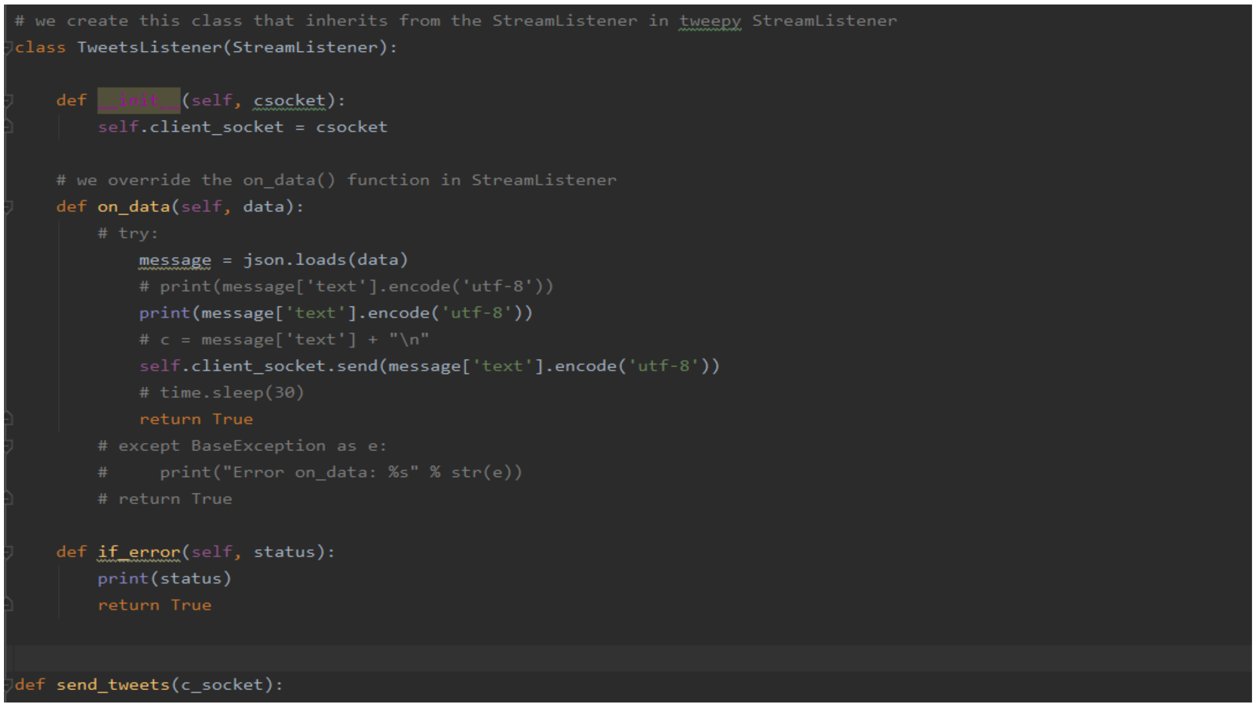
```python
if __name__ == "__main__":
    new_skt = socket.socket()  # initiate a socket object
    host = "127.0.0.1"  # local machine address
    port = 8085  # specific port for your service.
    new_skt.bind((host, port))  # Binding host and port

    print("Now listening on port: %s" % str(port))

    new_skt.listen(5)  # waiting for client connection.
    c, addr = new_skt.accept()  # Establish connection with client. it returns first a socket object,c, and the address bound to the socket
```

- Create a twitter developer account to get access.
- Get authorization and collect tweets under desired topic.
- Here I have extracted tweets under topic football.



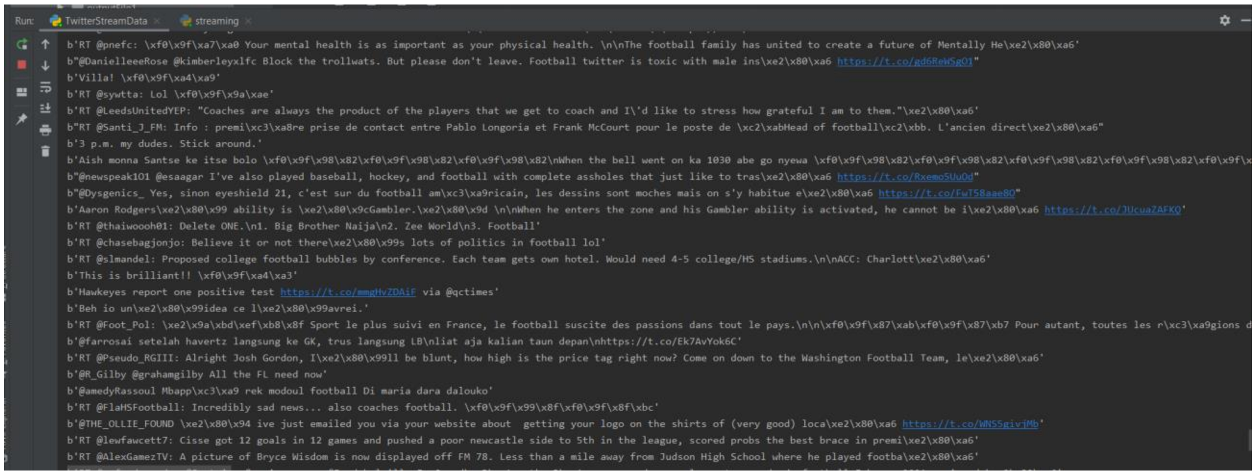
```python
def send_tweets(c_socket):
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_secret)

    twitter_stream = Stream(auth, TweetsListener(c_socket))
    twitter_stream.filter(track=['football'])  # this is the topic we are interested in
```

- Create a class and a model where it extracts only text content from the entire data collected.
- Collect the data and read it to self.
- Run the program and the status gets displayed.



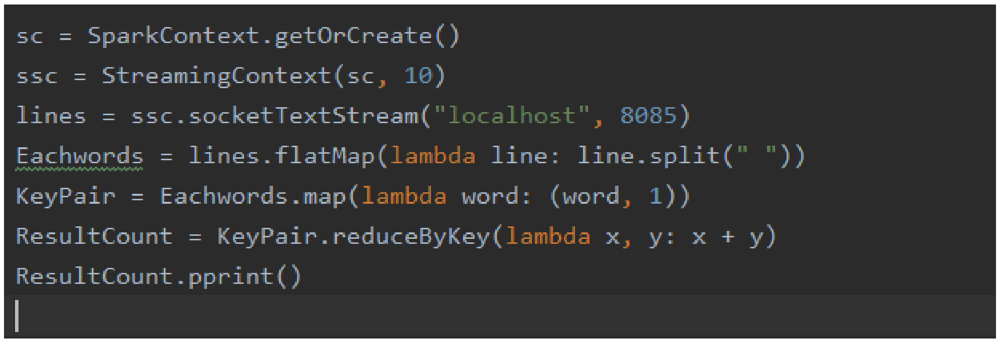
```python
# we create this class that inherits from the StreamListener in tweepy StreamListener
class TweetsListener(StreamListener):

    def __init__(self, csocket):
        self.client_socket = csocket

    # we override the on_data() function in StreamListener
    def on_data(self, data):
        # try:
            message = json.loads(data)
            # print(message['text'].encode('utf-8'))
            print(message['text'].encode('utf-8'))
            # c = message['text'] + "\n"
            self.client_socket.send(message['text'].encode('utf-8'))
            # time.sleep(30)
            return True
        # except BaseException as e:
        #     print("Error on_data: %s" % str(e))
        # return True

    def if_error(self, status):
        print(status)
        return True


def send_tweets(c_socket):
```

OUTPUT:



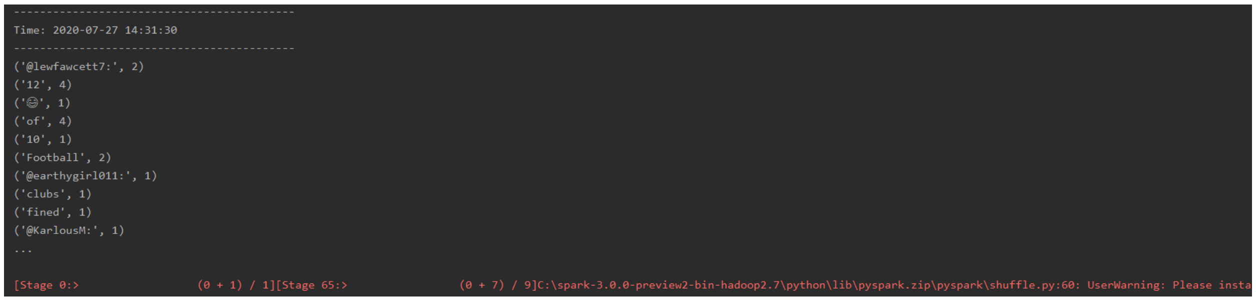
```
b'RT @pnefc: \xf0\x9f\xa7\xa0 Your mental health is as important as your physical health. \n\nThe football family has united to create a future of Mentally He\xe2\x80\xa6'
b"@DanielleeeRose @kimberleyxlfc Block the trollwats. But please don't leave. Football twitter is toxic with male ins\xe2\x80\xa6 https://t.co/gd6ReWSgO1"
b'Villa! \xf0\x9f\xa4\xa9'
b'RT @sywtta: Lol \xf0\x9f\x9a\xae'
b'RT @LeedsUnitedYEP: "Coaches are always the product of the players that we get to coach and I\'d like to stress how grateful I am to them."\xe2\x80\xa6'
b"RT @Santi_J_FM: Info : premi\xc3\xa8re prise de contact entre Pablo Longoria et Frank McCourt pour le poste de \xc2\xabHead of football\xc2\xbb. L'ancien direct\xe2\x80\xa6"
b'3 p.m. my dudes. Stick around.'
b'Aish monna Santse ke itse bolo \xf0\x9f\x98\x82\xf0\x9f\x98\x82\xf0\x9f\x98\x82\nWhen the bell went on ka 1030 abe go nyewa \xf0\x9f\x98\x82\xf0\x9f\x98\x82\xf0\x9f\x98\x82\xf0\x9f\x98\x82\xf0\x9f\x
b"@newspeak101 @esaagar I've also played baseball, hockey, and football with complete assholes that just like to tras\xe2\x80\xa6 https://t.co/RxemoSUuOd"
b"@Dysgenics_ Yes, sinon eyeshield 21, c'est sur du football am\xc3\xa9ricain, les dessins sont moches mais on s'y habitue e\xe2\x80\xa6 https://t.co/FwT58aae8O"
b'Aaron Rodgers\xe2\x80\x99 ability is \xe2\x80\x9cGambler.\xe2\x80\x9d \n\nWhen he enters the zone and his Gambler ability is activated, he cannot be i\xe2\x80\xa6 https://t.co/JUcuaZAFKQ'
b'RT @thaiwoooh01: Delete ONE.\n1. Big Brother Naija\n2. Zee World\n3. Football'
b'RT @chasebagjonjo: Believe it or not there\xe2\x80\x99s lots of politics in football lol'
b'RT @slmandel: Proposed college football bubbles by conference. Each team gets own hotel. Would need 4-5 college/HS stadiums.\n\nACC: Charlott\xe2\x80\xa6'
b'This is brilliant!! \xf0\x9f\xa4\xa3'
b'Hawkeyes report one positive test https://t.co/mmgHvZDAiF via @qctimes'
b'Beh io un\xe2\x80\x99idea ce l\xe2\x80\x99avrei.'
b'RT @Foot_Pol: \xe2\x9a\xbd\xef\xb8\x8f Sport le plus suivi en France, le football suscite des passions dans tout le pays.\n\n\xf0\x9f\x87\xab\xf0\x9f\x87\xb7 Pour autant, toutes les r\xc3\xa9gions d
b'@farrosai setelah havertz langsung ke GK, trus langsung LB\nliat aja kalian taun depan\nhttps://t.co/Ek7AvYok6C'
b'RT @Pseudo_RGIII: Alright Josh Gordon, I\xe2\x80\x99ll be blunt, how high is the price tag right now? Come on down to the Washington Football Team, le\xe2\x80\xa6'
b'@R_Gilby @grahamgilby All the FL need now'
b'@amedyRassoul Mbapp\xc3\xa9 rek modoul football Di maria dara dalouko'
b'RT @FlaHSFootball: Incredibly sad news... also coaches football. \xf0\x9f\x99\x8f\xf0\x9f\x8f\xbc'
b'@THE_OLLIE_FOUND \xe2\x80\x94 ive just emailed you via your website about  getting your logo on the shirts of (very good) loca\xe2\x80\xa6 https://t.co/WNSSgivjMb'
b'RT @lewfawcett7: Cisse got 12 goals in 12 games and pushed a poor newcastle side to 5th in the league, scored probs the best brace in premi\xe2\x80\xa6'
b'RT @AlexGamezTV: A picture of Bryce Wisdom is now displayed off FM 78. Less than a mile away from Judson High School where he played footba\xe2\x80\xa6'
```

- Later initialize the streaming part.
- Give the same port number as given earlier with which client receives the data from server.
- Use flatMap, where it helps in splitting up a string or a sentence Separated and terminated by a delimiter space as shown in the code below.
- Using Map and reduce methods, split up the words and count their repetition.
- Store the resultant word frequency and print them. Thus, output is obtained.

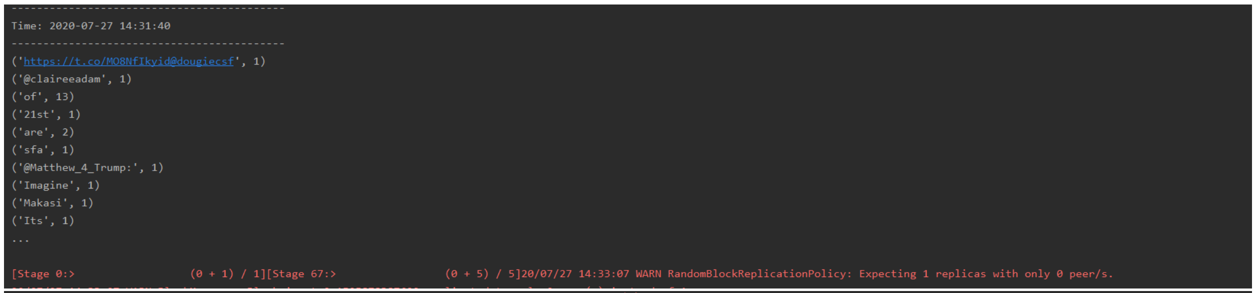


```
sc = SparkContext.getOrCreate()
ssc = StreamingContext(sc, 10)
lines = ssc.socketTextStream("localhost", 8085)
Eachwords = lines.flatMap(lambda line: line.split(" "))
KeyPair = Eachwords.map(lambda word: (word, 1))
ResultCount = KeyPair.reduceByKey(lambda x, y: x + y)
ResultCount.pprint()
```

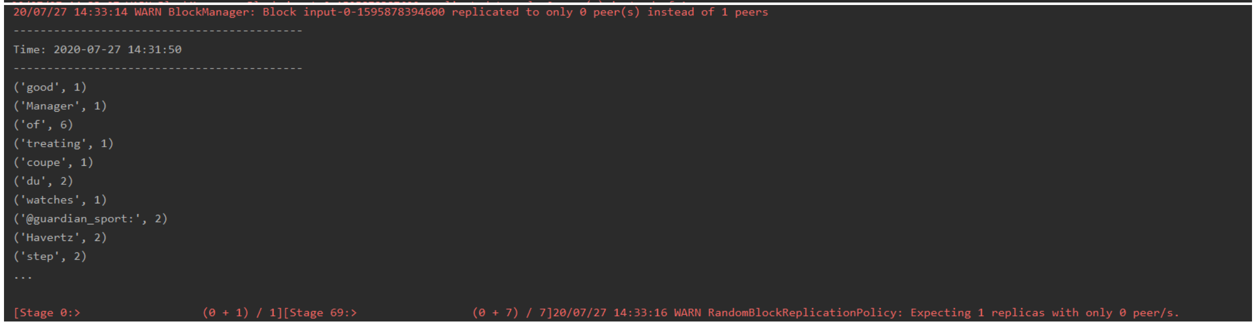
OUTPUT:



```
-------------------------------------------
Time: 2020-07-27 14:31:30
-------------------------------------------
('@lewfawcett7:', 2)
('12', 4)
('😂', 1)
('of', 4)
('10', 1)
('Football', 2)
('@earthygirl011:', 1)
('clubs', 1)
('fined', 1)
('@KarlousM:', 1)
...

[Stage 0:>                    (0 + 1) / 1][Stage 65:>                   (0 + 7) / 9]C:\spark-3.0.0-preview2-bin-hadoop2.7\python\lib\pyspark.zip\pyspark\shuffle.py:60: UserWarning: Please insta
```



```
-------------------------------------------
Time: 2020-07-27 14:31:40
-------------------------------------------
('https://t.co/MO8NfIkyid@dougiecsf', 1)
('@claireeadam', 1)
('of', 13)
('21st', 1)
('are', 2)
('sfa', 1)
('@Matthew_4_Trump:', 1)
('Imagine', 1)
('Makasi', 1)
('Its', 1)
...

[Stage 0:>                 (0 + 1) / 1][Stage 67:>                (0 + 5) / 5]20/07/27 14:33:07 WARN RandomBlockReplicationPolicy: Expecting 1 replicas with only 0 peer/s.
```



```
20/07/27 14:33:14 WARN BlockManager: Block input-0-1595878394600 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2020-07-27 14:31:50
-------------------------------------------
('good', 1)
('Manager', 1)
('of', 6)
('treating', 1)
('coupe', 1)
('du', 2)
('watches', 1)
('@guardian_sport:', 2)
('Havertz', 2)
('step', 2)
...

[Stage 0:>                 (0 + 1) / 1][Stage 69:>                (0 + 7) / 7]20/07/27 14:33:16 WARN RandomBlockReplicationPolicy: Expecting 1 replicas with only 0 peer/s.
```

QUESTION 3: (4)

CHALLENGES FACED:

Had minor issues while writing tweets into socket as it consumes little more time comparatively. And also getting access from twitter also takes time. Irrespective of these two, entire execution procedure went very effectively without any obstacles.

QUESTION 3: (5)

MILESTONES AND INTEGRATION OF THE PROJECT:

Work went very effective and smooth without any obstacles. Team split-up made it easy to perform tasks. As there is no dependency for one task to other, individuals performed one task each and accomplished the tasks given.

TEAM MEMBERS AND CONTRIBUTION:

Roshini varada -- Hadoop MapReduce Algorithm

Link -- https://github.com/RoshiniVarada/BDP_Project2/wiki/CASE1

Sarika Reddy Kota -- Spark Data Frames

Link -- https://github.com/RoshiniVarada/BDP_Project2/wiki/CASE2

Pallavi Arikatla -- Spark streaming

Link -- https://github.com/RoshiniVarada/BDP_Project2/wiki/CASE3

Zakari, Abdulmuhaymin -- Spark Graphx

Link -- https://github.com/RoshiniVarada/BDP_Project2/wiki/CASE4

Video Link:

https://youtu.be/UtsiVZaijyg

PALLAVI ARIKATLA

16301244