Streaming: Twitter Sentiment Analysis

Here, I would use Spark Streaming and Twitter API and implement my Sentiment Analysis system with running for 3 hours.

We need to write the HTTP Client File as below:

```python
import socket
import sys
import requests
import requests_oauthlib
import json

ACCESS_TOKEN = '817228248304103424-PbFDe6KA1XYHrrskObkHpLgDFIJkth6'
ACCESS_SECRET = 'RQBCJ2fdZUmCOVCEQFmKkCx6vl3Zrb1lm9kChUhgFo1XT'
CONSUMER_KEY = 'Q7dBHVt2Y3OEaqoXHIHJs9kpk'
CONSUMER_SECRET = 'x5I72aPKMNvBWZN3bZTzNwXycDTUVZnPZTLsWARcsqAeNRVJZX'
my_auth = requests_oauthlib.OAuth1(CONSUMER_KEY, CONSUMER_SECRET,ACCESS_TOKEN,
ACCESS_SECRET)

def get_tweets():
    url = 'https://stream.twitter.com/1.1/statuses/filter.json'
    query_data = [('language', 'en'), ('locations', '-130,-20,100,50'),('track','#')]
    query_url = url + '?' + '&'.join([str(t[0]) + '=' + str(t[1]) for t in query_data])
    response = requests.get(query_url, auth=my_auth, stream=True)
    print(query_url, response)
    return response

def send_tweets_to_spark(http_resp, tcp_connection):
    for line in http_resp.iter_lines():
        try:
            full_tweet = json.loads(line)
            tweet_text = full_tweet['text']
            print("Tweet Text: " + tweet_text)
            print ("----------------------------------------")
            tcp_connection.send(tweet_text + '\n')
        except:
            e = sys.exc_info()[0]

TCP_IP = "localhost"
TCP_PORT = 9009
conn = None
s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
s.bind((TCP_IP, TCP_PORT))
```

```
s.listen(1)
print("Waiting for TCP connection...")
conn, addr = s.accept()
print("Connected... Starting getting tweets.")
resp = get_tweets()
# send_tweets_to_spark(resp, conn)

http_resp = resp
tcp_connection = conn
for line in http_resp.iter_lines():
    try:
        full_tweet = json.loads(line)
        tweet_text = full_tweet['text']
        #print("Tweet Text: " + tweet_text)
        #print ("------------------------------------------")
        tweet_text += '\n'
        tcp_connection.send(tweet_text.encode())
    except:
        e = sys.exc_info()
```

```
In [*]: TCP_IP = "localhost"
        TCP_PORT = 9009
        conn = None
        s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
        s.bind((TCP_IP, TCP_PORT))
```

```
In [*]: s.listen(1)
        print("Waiting for TCP connection...")
        conn, addr = s.accept()
        print("Connected... Starting getting tweets.")
        resp = get_tweets()
        # send_tweets_to_spark(resp, conn)

        http_resp = resp
        tcp_connection = conn
        for line in http_resp.iter_lines():
            try:
                full_tweet = json.loads(line)
                tweet_text = full_tweet['text']
                #print("Tweet Text: " + tweet_text)
                #print ("------------------------------------------")
                tweet_text += '\n'
                tcp_connection.send(tweet_text.encode())
            except:
                e = sys.exc_info()

        Waiting for TCP connection...
        Connected... Starting getting tweets.
        https://stream.twitter.com/1.1/statuses/filter.json?language=en&locations=-130,-20,100,50&track=# <Response [200]>
```

Here, in order to avoid seeing hundreds of pages of tweet_text being printed, we don't print every line of tweet_text. And we could see the valid results in Spark_streaming file.

We need to write Spark Streaming File as below:

```
import findspark
findspark.init()
```

```python
from pyspark import SparkConf,SparkContext
from pyspark.streaming import StreamingContext
from pyspark.sql import Row,SQLContext
import sys
import requests

conf = SparkConf()
conf.setMaster('local[2]')
conf.setAppName("TwitterStreamApp")
sc = SparkContext(conf=conf)
sc.setLogLevel("ERROR")
ssc = StreamingContext(sc, 100)
ssc.checkpoint("checkpoint_TwitterApp")

def aggregate_tags_count(new_values, total_sum):
    return sum(new_values) + (total_sum or 0)

def get_sql_context_instance(spark_context):
    if ('sqlContextSingletonInstance' not in globals()):
        globals()['sqlContextSingletonInstance'] = SQLContext(spark_context)
    return globals()['sqlContextSingletonInstance']

def process_rdd(time, rdd):
    print("----------- %s -----------" % str(time))
    try:
        sql_context = get_sql_context_instance(rdd.context)
        row_rdd = rdd.map(lambda w: Row(hashtag=w[0], hashtag_count=w[1]))
        hashtags_df = sql_context.createDataFrame(row_rdd)
        hashtags_df.registerTempTable("hashtags")
        hashtag_counts_df = sql_context.sql("select hashtag, hashtag_count from
hashtags order by hashtag_count desc limit 10")
    except:
        e = sys.exc_info()[0]
    finally:
        try:
            hashtag_counts_df.show()
        except:
            pass

dataStream = ssc.socketTextStream("localhost",9009)
dataStream.pprint()

words = dataStream.flatMap(lambda line: line.split(" "))
hashtags = words.filter(lambda w: '#' in w).map(lambda x: (x, 1))

tags_totals = hashtags.updateStateByKey(aggregate_tags_count)
```

```
tags_totals.foreachRDD(process_rdd)

ssc.start()
ssc.awaitTermination()
```

This is my screen shot at the start of running code:

```
-------------------------------------------
Time: 2018-10-21 17:45:00
-------------------------------------------
@enesessen Playback 🤔
Floppppped
Running up the middle with Smallwood just isn't going to work Doug
How does he get there before @OpTicH3CZ  lol
Rae Carruth is about to get out of prison. He's been in there for 18 years!! That's not long enough imo. Also, he g…
https://t.co/LhrmKZLvtH
at 6pm i'll be dressed and ready to go sooo LMK!!
So, this @BreweryOmmegang is out of Brooklyn ... and, the #Jets are on the tube.

Gee, you'd think I was in NYC. 😉 https://t.co/84buIRKMDC
My dad's classic move:
...

----------- 2018-10-21 17:45:00 -----------
+--------------------+-------------+
|             hashtag|hashtag_count|
+--------------------+-------------+
|           #CareerArc|            6|
|              #Hiring|            5|
|                 #job|            5|
|      #Pharmaceutical|            2|
|              #travel|            1|
|           #TMBTMEDIA|            1|
|                  #me|            1|
|             #october|            1|
|           #thesmiths|            1|
|#bigmouthstrikesa...|            1|
+--------------------+-------------+
```

And after 3h running:

```
-------------------------------------------
Time: 2018-10-21 20:46:40
-------------------------------------------
@JumpOttawa keep on breath'in 💐 @ArianaGrande
My dumb ass lost my card bro
@stoolpresidente Keep sucking dick Dave!
BUCS WIN BABY!!!!!!!! https://t.co/dXEuKrcYun
@BellaTwins Nikki you has lost your mind you are not going beat ronda. Ronda is going to beat you Nikki and Ronda i…
https://t.co/npU388DJ0T
The announcers were just talking about the kicker getting cut too lol
Todd haylies offense had plenty of chances &amp; that's what HAPPENS when u call shit plays in critical moments. Grea
t… https://t.co/anFwsPUGBe
@SShawFL @CoryBooker "New Jersey Senator, Cory Booker, accused of sexual assault by a man" by @RAMRANTS

https://t.co/K3R138MayP
...

----------- 2018-10-21 20:46:40 -----------
+----------+-------------+
|   hashtag|hashtag_count|
+----------+-------------+
|#CareerArc|         8780|
|      #job|         7265|
|   #Hiring|         6842|
|      #Job|         2386|
|  #hiring!|         2385|
|    #Jobs|         2385|
|   #hiring|         1409|
|     #job:|         1196|
|     #job?|         1169|
|   #Browns|          751|
+----------+-------------+
```