NAME: Meith Navlakha
SAP: 60004190068
BDI Practical

EXPERIMENT 11

AIM : Case study to perform Sentiment analysis using Spark Streaming.

THEORY:

Data is being generated at an unprecedented rate, and by analyzing it correctly and providing valuable and meaningful insights at the right time, it can result in valuable solutions for an array of domains involved with data. Real-time streaming data is widely used across a range of industries, from health care and banking to media and retail. Netflix, for example, provides real-time recommendations tailored to individual preferences. Similarly to every business that streams large amounts of data and relies on various analytics, Amazon tracks its users' interactions with its products and makes prompt recommendations of related items.

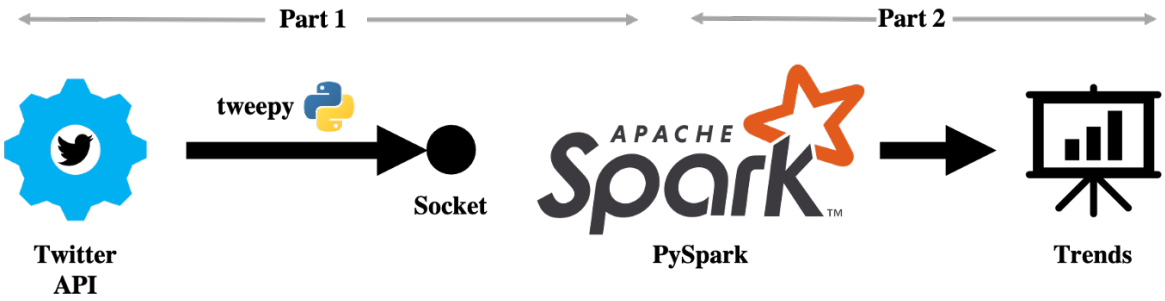
Apache Spark is an efficient framework for analyzing large amounts of data in real-time and performing various analyses on the data. Many resources discuss Spark and its popularity in the big data space, but it is worthwhile to highlight that its core features include real-time big data processing using Resilient Distributed Datasets (RDDs), streaming, and machine learning. In this tutorial, we will demonstrate how Spark Streaming components can be used in conjunction with PySpark to resolve a business problem.

Motivation

In today's society, the importance of social media cannot be denied. Most businesses collect feedback from their Twitter followers in order to gain insight and to gain a better understanding of their customers. Social media feedback changes rapidly, and the ability to analyze that feedback in real time is essential for the success of any business. There are several ways to discover how people react to a new product, brand, or event. For example, the sentiment expressed through tweets about a particular topic, product, brand or event can provide an indication of the level of willingness or trust in a product. Hence, we use this premise in our tutorial and extract trending #tags related to our desired topic every few minutes since hashtagged tweets are more engaging.

We will use Tweepy to access Twitter's streaming API and the Spark streaming component with TCP socket to receive tweets. We will layer tweets on RDD and then retrieve the most popular hashtags. After that, we use Spark SQL to save the top hashtags to a temporary database. Finally, we visualize the results using Python's visualization tools.

The following image illustrates the overall architecture of our program.



NAME: Meith Navlakha
SAP: 60004190068
BDI Practical

CODE:

Receive_tweets:

```
from tweepy.auth import OAuthHandler
from tweepy import Stream
from tweepy.streaming import StreamListener
import socket
import json

consumer_key    = '<consumer_key>'
consumer_secret = '<consumer_secret>'
access_token    = '<access_token>'
access_secret   = '<access_secret>'

class TweetsListener(StreamListener):
    def __init__(self, csocket):
        self.client_socket = csocket

    def on_data(self, data):
        try:
            message = json.loads( data )
            print( message['text'].encode('utf-8') )
            self.client_socket.send( message['text'].encode('utf-8') )
            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):
    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'])

new_skt = socket.socket()
host = "127.0.0.1"
port = 5555
new_skt.bind((host, port))

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

new_skt.listen(5)
```

```
c, addr = new_skt.accept()
print("Received request from: " + str(addr))
send_tweets(c)
```

read_tweets.py

```
import findspark
findspark.init()
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from pyspark.sql import SQLContext
from pyspark.sql.functions import desc

sc = SparkContext()
ssc = StreamingContext(sc, 10)
sqlContext = SQLContext(sc)
socket_stream = ssc.socketTextStream("127.0.0.1", 5555)
lines = socket_stream.window(60)

from collections import namedtuple
fields = ("hashtag", "count" )
Tweet = namedtuple( 'Tweet', fields )
( lines.flatMap( lambda text: text.split( " " ) )
  .filter( lambda word: word.lower().startswith("#") )
  .map( lambda word: ( word.lower(), 1 ) )
  .reduceByKey( lambda a, b: a + b )
  .map( lambda rec: Tweet( rec[0], rec[1] ) )
  .foreachRDD( lambda rdd: rdd.toDF().sort( desc("count") )
  .limit(10).registerTempTable("tweets") ) )

ssc.start()

import time
from IPython import display
import matplotlib.pyplot as plt
import seaborn as sns
get_ipython().run_line_magic('matplotlib', 'inline')
count = 0
while count < 5:
    time.sleep(5)
    top_10_tags = sqlContext.sql( 'Select hashtag, count from tweets' )
    top_10_df = top_10_tags.toPandas()
    display.clear_output(wait=True)
    plt.figure( figsize = ( 10, 8 ) )
    sns.barplot( x="count", y="hashtag", data=top_10_df)
    plt.show()
```

```
    count = count + 1
    print(count)
ssc.stop()
```

Output:

receive_tweet.py:

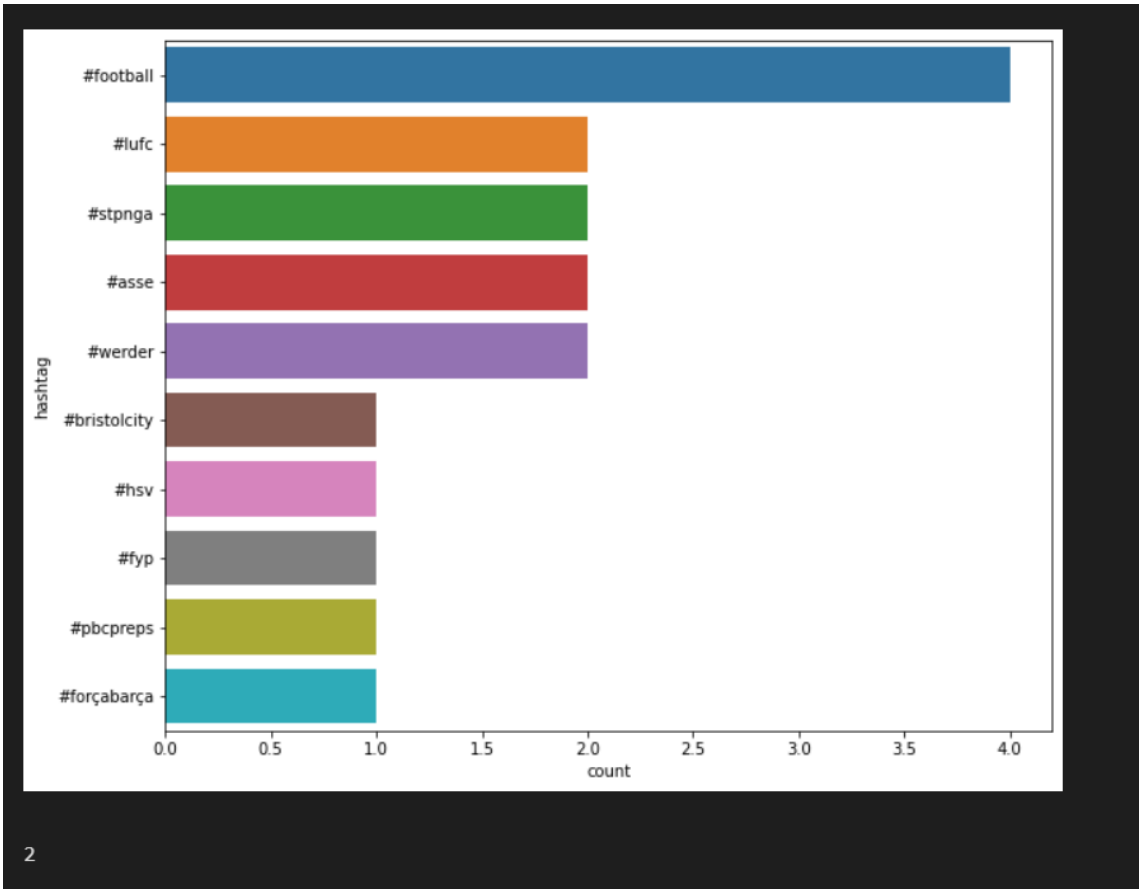
The tweets related to football are obtained as shown below.

```
Output exceeds the size limit. Open the full output data in a text editor
Now listening on port: 5555
Received request from: ('127.0.0.1', 50600)
b'@RetourConfrerie \xc3\xa9conomique c un peu culott\xc3\xa9 de vouloir se d\xc3\xa9tacher d pays du Maghreb en sachant ke vous devez plus\xe2\x80\xa6
https://t.co/g9nkLw1AQl'
b'@NoeIFlantier @JoelPostbad Le football il a chang\xc3\xa9'
b'Chris Sutton leaves Robbie Savage red-faced with embarrassing Zoom video of chair gaffe\nhttps://t.co/9w5QGnTNXK https://t.co/p0yxBl4vF9'
b'It is never that serious\xe2\x80\xa6 fr.'
b'RT @LFCTransferRoom: \xe2\x9d\x97\xef\xb8\x8fThe five substitutions rule is now permanent in professional football.'
b'RT @FranceRMCF: Thibaut Courtois : "Merci Marcelo, pour tout ce temps pass\xc3\xa9 ensemble. Je te souhaite le meilleur ! T\xe2\x80\x99es une l\xc3\xa9gende du
foot\xe2\x80\xa6'
b'Send him to the championship'
b'RT @JustGiving: CW: Baby loss\n\nThis Month is @SandsUK Awareness Month \xf0\x9f\xa4\x8d\n\nA few weeks ago, the football team at @Upp_Meadhurst hosted a
char\xe2\x80\xa6'
b'@indykaila A sidline kick as in Gaelic Football'
b'RT @christofidellis: \xce\x9f\xce\xbb\xcf\x85\xce\xbc\xcf\x80\xce\xb9\xce\xb1\xce\xba\xcf\x8c\xcf\x82:
\xce\xa4\xce\xb5\xce\xbb\xce\xb5\xce\xb9\xcf\x8e\xce\xbd\xce\xb5\xce\xb9 \xce\xbc\xce\xb5 \xcf\x84\xce\xb9\xcf\x82
\xce\xb1\xce\xbd\xce\xb1\xce\xbd\xce\xb5\xcf\x8e\xcf\x83\xce\xb5\xce\xb9\xcf\x82 \xcf\x84\xce\xbf\xcf\x85 \xce\x91\xcf\x81\xce\xb1\xce\xbc\xcf\x80\xce\xaf
\xce\xba\xce\xb1\xce\xb9 \xcf\x84\xce\xbf\xcf\x85 \xce\x92\xce\xb1\xce\xbb\xce\xbc\xcf\x80\xce\xbf\xcf\x85\xce\xb5\xce\xbd\xce\xac #Olympiacos #Olympiakos
@sport24 @elarabiyoussef\xe2\x80\xa6'
b'we call it fucking good business, and it\xe2\x80\x99s a very well run football club'
b'RT @RyanBurnsMN: I connected with @Crimsonfootball head coach @lombocrimson to learn more about #Gophers tight end commit @sampetersss, and\xe2\x80\xa6'
b"RT @Sportsbetio: A spot in Qatar awaits \xe2\x8f\xb3\n\n#DidYouKnow \xf0\x9f\x91\x89\xf0\x9f\x8f\xbb It's been 21 years since both teams met in a competitive
game \xf0\x9f\xa4\xaf\n\nPredict here: ht\xe2\x80\xa6"
b"i'm sure toni will delete his tweet now that a guy larping as a brazilian football scout has informed him that his\xe2\x80\xa6 https://t.co/pH6HfqaNON"
b'Haaa wee le football il a chang\xc3\xa9, comme sais-tu que c\xe2\x80\x99est un acteur Porno \xf0\x9f\xab\xb3\xf0\x9f\x8f\xbe humm'
b'RT @coachtf: BadAss Coaches understand that 7 on 7 in the summertime can develop the worst quarterback habits in football. Make your QB tak\xe2\x80\xa6'
b'RT @ARanganathan72: Hindus can form a football team but they won\xe2\x80\x99t be able to give it a name as they are not united.'
b"RT @ManUtdMEN: A report claims Erik ten Hag is 'definitely' going to sign Antony and Frenkie de Jong #mufc https://t.co/Lbba9ni1Gb https://\xe2\x80\xa6"
b'RT @JBreezyII: Swear to God I\xe2\x80\x99d retire from football if this happened to me. https://t.co/Z1kopsX4pK'
```

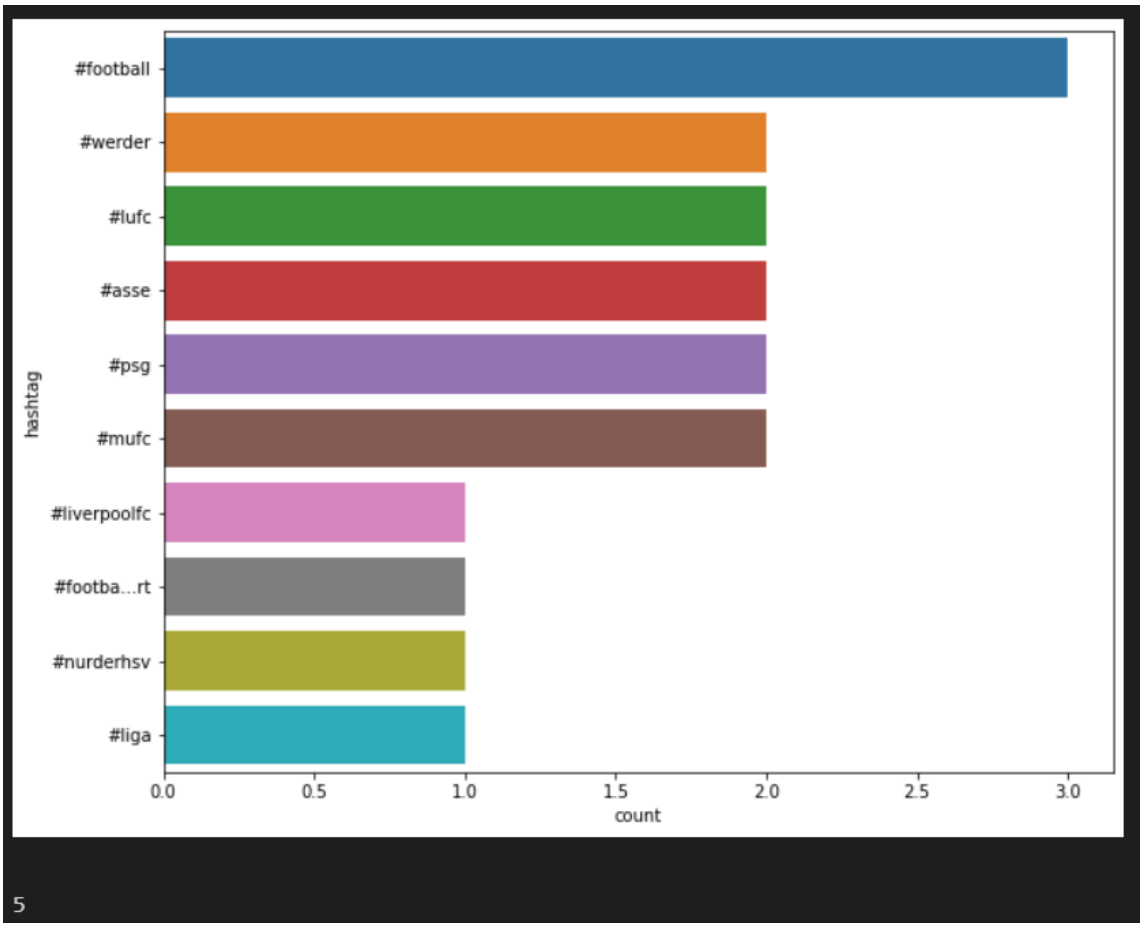
```
b'@niiteGodfrey @JoanOzil @Arsenal We dont care...\nMost important..he comes from Brazil..The Home of football'
b'RT @MercatoPlein: DECLA \xf0\x9f\x9a\xa8 Mathieu #Valbuena : \xc2\xab\xc2\xa0Bien s\xc3\xbbr que je suis marseillais, mais, pour \xc3\xaatre honn\xc3\xaate, je
trouve que l\xe2\x80\x99ambiance du V\xc3\xa9lo\xe2\x80\xa6'
...
b'RT @SkySportsPL: \xf0\x9f\x97\xa3\xef\xb8\x8f "Manchester United\'s offer was bigger than Liverpool\'s but the project of Liverpool playing in the Champions
League was\xe2\x80\xa6'
b'RT @MirrorFootball: Ryan Gravenberch explains why he snubbed last-gasp Man Utd transfer attempt by Erik ten Hag\nhttps://t.co/T3aIVymAWr htt\xe2\x80\xa6'
b'RT @MadridXtra: \xf0\x9f\x9a\xa8| The five substitutions rule in professional football is now PERMANENT.'
b'@ElFloco_ : Juste un passionn\xc3\xa9 de football qui nous en parle sur son compte et depuis peu via une cha\xc3\xaene Youtube do\xe2\x80\xa6
https://t.co/IykkTN8e08'
```

The next images contain the 2nd and 5th iteration of the analysis using Spark Streaming respectively; and we can see the hashtags that are being used the most with the word football.

NAME: Meith Navlakha
SAP: 60004190068
BDI Practical



2



5

NAME: Meith Navlakha
SAP: 60004190068

CONCLUSION :

In this experiment, I explored and learnt about Spark Streaming for real-time streaming data processing. For the streaming data we used tweetpy library and performed Sentimental Analysis on the same. Initially, #football, #ufc, #stpnga were the top hashtags, later on, #football, #werder , #lufc become the top hashtags -- clearly displaying how we processed the trend with time. The output for the experiment were observed and attached. Thus in this experiment, we implemented Sentimental Analysis on twitter streaming data.