Twitter Text Capture and Analysis

Student Name : Pranalee Jadhav
Student ID : 801030690

1. How does the program work?
Ans:
Working of the given python program is as follows:

1. Get access token, access secret, consumer key, consumer secret from twitter api.
2. Get Inputs (keyword, to date, from date, number of tweets) from user.
3. Configure tweepy object using the keys defined in step 1.
4. Using api.search, list of tweets are retrieved containing the keyword and between to and from date
5. Iterate through tweets using tweepy.Cursor.

2. How do you think you can use this code?
Ans:
We can use this code to get a list of tweets. Along with the tweet, the api will also return name of the person responsible for tweet, his followers count, retweet count, tweet time.
http://docs.tweepy.org/en/v3.5.0/api.html

3. Can you think of different scenarios where this code could be used for data collection?
Ans:

- The code can be used to detect active twitter followers and members.
- It can also be used to find fake twitter holders
- It can be used to find inactive users

- It can be useful for the administrator to delete users who are inactive on twitter
- It can be used to send offer or update related emails to users who are often active on twitter

For the given example, run following commands on terminal/ command prompt:

python3 twitter_analysis.py

```
pythonpractice — Python twitter_analysis.py — 80×24
4
Apples-iMac-2:pythonpractice Apple$ python3 twitter_analysis.py

Choose Options:
press 0 for top n users who have tweeted the most for the entire timeline
press 1 for top n users who have tweeted the most for every hour
press 2 for top n users who have the maximum followers
press 3 for top n tweets which have the maximum retweet count
press 4 for exit
```

Twitter_analysis.py is using textfile 'charlotte'.txt

For running the file: twitter.py, type below command

python3 twitter.py

Github Link:
https://github.com/lee0392/ssdi_twitter.git

Output

a. The top n users who have tweeted the most for the entire timeline.

```
pythonpractice — Python twitter_analysis.py — 80×24
press 2 for top n users who have the maximum followers
press 3 for top n tweets which have the maximum retweet count
press 4 for exit

0
Enter n to know top n users who have tweeted the most for the entire timeline :
10

Result:

CharlotteCP : 2
MariePotee : 1
QueenPlz : 1
MLBRaysUp : 1
TerezaSmurf : 1
charlietuna43 : 1
tmj_clt_transp : 1
jackiecdeac : 1
LushLtd : 1
LRNROSE : 1



Choose Options:
```

b. The top n users who have tweeted the most for every hour.

```
pythonpractice — Python twitter_analysis.py — 80×24
press 4 for exit

1
Enter n to know top n users who have tweeted the most for every hour : 10
Top users for time Interval: From 2018-02-17 16:24:56 till 2018-02-17 16:22:57
CharlotteCP : 2
MariePotee : 1
QueenPlz : 1
MLBRaysUp : 1
TerezaSmurf : 1
charlietuna43 : 1
tmj_clt_transp : 1
jackiecdeac : 1
LushLtd : 1
LRNROSE : 1
```

c. The top n users who have the maximum followers.

```
pythonpractice — Python twitter_analysis.py — 80×24
2
Enter n to know top n users who have the maximum followers : 10

Result:

LushLtd : 196620
SocNCharlotte : 38190
gr8musicvenues : 14185
SCSportsReport : 7442
MsCharlotteLace : 6855
LRNROSE : 5478
xaltd : 3718
FipNowPlays : 3101
IncidentsPolice : 2426
Charlotte_Foxxx : 2130
```

d. The top n tweets which have the maximum retweet count.

```
pythonpractice — Python twitter_analysis.py — 82×33
3
Enter n to know top n tweets which have the maximum retweet count : 10

Result:

RT @davidminpdx: Charlotte, NC District Attorney announces that his office will no
 longer require nonviolent, first-time defendants to pay… : 184
RT @VVSupremo: 🍬Try this delicious dessert with your sweetheart.#LoveMyQueso #Vale
ntinesDayRecipe: https://t.co/Sog5hGrrlo https://t.co/W… : 52
RT @OHaraNews: This week's #EverydayHero: @BrittBogues! Her impact has been a slam
 dunk in #Charlotte! Check out what she is doing and nomi… : 4
RT @SCSportsReport: If you need to market to active families and MOMS, please cons
ider the South Charlotte Sports Report.   Your support al… : 2
@lizpeek By staging a Pro Hillary rally here in Charlotte? : 2
WHEN WILL THE TICKETS GO ON SALE FOR WHAT MAKES YOU COUNTRY TOUR AT PNC MUSIC PAVI
LLION IN CHARLOTTE NC . PLZ SEE T… https://t.co/YYCT7MyHp9 : 1
RT @TBTimes_Rays: More early reporters and impressive pitching for #Rays  https://
t.co/tyNr5tFSTk : 1
RT @TheCharlotteSE1: LIVE TONIGHT- Find out more here: https://t.co/FT12qA9Ac4 : 1
@Ccampbellmusic Hi Charlotte, will you be busking tomorrow? 🎸: 0
See our latest #Charlotte, NC #job and click to apply: CDL A Owner Operator, Drop
&amp; Hook - https://t.co/E7qH9q0ZUl… https://t.co/o9YtGQjh53 : 0
```

Source code :

```
import re #importing regular expression library
import operator  #for sorting dictionary
import datetime #importing date time library

#Defining global variables
data = [[]]
count1 = {}
count2 = {}
interval = {}


#Defining functions

#Parse text file using regex and store in an array
def getData():
	global data
	file = open("'charlotte'.txt", 'r')
	matches = re.findall('(\w+) \\[([^\\]]*)\\] "([^"]*)" ([^ ]*) ([^ ]*\n)',file.read())
#regular expression
	for match in matches:
		data.append(match)
	return

#Find users with maximum tweets and followers in entire timeline
def findHighestTweeter1():
	global data
	for i in range(len(data)):
		x = 1
		y = int(data[i][3])
		if data[i][0] not in count1:
			for j in range(i+1, len(data)):
				if data[j][0]==data[i][0]: #comparision
					x = x + 1
```

```
                        count1[data[i][0]] = [x,y]
        return

#Find tweets with maximum retweet count
def findHighestTweets():
        global data
        global count2
        temp = {}
        for i in range(len(data)):
                temp[i] = int(data[i][4].replace('\n','')) #convert string to int
        count2 = sorted(temp.items(), key=operator.itemgetter(1),reverse=True)
        return


#Convert timestring into datetime object
def gettime(str):
        dt = datetime.datetime.strptime(str, "%d/%b/%Y:%H:%M:%S ")
        return dt

#Find time difference
def timediff(diff):
        days = diff.days
        temp = days * 24
        diff2 = (diff.seconds) / 3600
        tot_hrs = temp + diff2
        return tot_hrs

#Find users with maximum tweets in a given range
def findHighestTweeter2(start,last):
        global data
        count3 = {}
        for i in range(start,last):
                x = 1
```

```
        if data[i][0] not in count3:
            for j in range(i+1, last):
                if data[j][0]==data[i][0]:
                    x = x + 1
            count3[data[i][0]] = x
    return count3


#Find the list of users with their number of tweets for every hour of the timeline
def getEveryHourTweets():
    global data
    global interval
    c = 0
    i = 0
    new_s = 0
    while (i + new_s)<len(data):
        i = i + new_s
        parent_time = gettime(data[i][1])
        for j in range(i+1, len(data)):
            cur_time = gettime(data[j][1])
            if timediff(cur_time-parent_time)>1 or j==(len(data)-1):
                temp = findHighestTweeter2(i,j)
                interval[c] = [parent_time,gettime(data[j-1][1]),temp]
                c = c + 1
                new_s = j

#Displaying List

def displayDict(dict,n,index):
    #iterator = iter(dict.items())
    print('\nResult:\n')
    for i in range(n):
        print(dict[i][0]+" : "+str(dict[i][1][index]))
```

```
        print('\n')

def displayTweets(dict,n):
        #iterator = iter(dict.items())
        print('\nResult:\n')
        for i in range(n):
                print(data[dict[i][0]][2]+" : "+str(dict[i][1]))
        print('\n')


def disp_perhr_tweet(n):
        for i in range(len(interval)):
                sorted_dict = sorted(interval[i][2].items(),
key=operator.itemgetter(1),reverse=True)
                print("Top users for time Interval: From " +
interval[i][0].strftime("%Y-%m-%d %H:%M:%S")+" till
"+interval[i][1].strftime("%Y-%m-%d %H:%M:%S"))
                for j in range(n):
                        print(sorted_dict[j][0]+" : "+str(sorted_dict[j][1]))


#Calling functions
getData()
del data[0]
findHighestTweeter1()
findHighestTweets()
getEveryHourTweets()

#Display Menu
while 1:
        try: #exception handling for allowing integer input only
                query = int(input("\nChoose Options:\npress 0 for top n users who
have tweeted the most for the entire timeline\npress 1 for top n users who have
```

```
tweeted the most for every hour\npress 2 for top n users who have the maximum
followers\npress 3 for top n tweets which have the maximum retweet count\npress
4 for exit\n\n"));


            if query==0:
                n1 = int(input("Enter n to know top n users who have tweeted
the most for the entire timeline : "))
                sorted_dict = sorted(count1.items(), key=lambda i: i[1][0],
reverse=True)
                displayDict(sorted_dict,n1,0)

            elif query==1:
                n1 = int(input("Enter n to know top n users who have tweeted
the most for every hour : "))
                disp_perhr_tweet(n1)


            elif query==2:
                n1 = int(input("Enter n to know top n users who have the
maximum followers : "))
                sorted_dict = sorted(count1.items(), key=lambda i: i[1][1],
reverse=True)
                displayDict(sorted_dict,n1,1)

            elif query==3:
                n1 = int(input("Enter n to know top n tweets which have the
maximum retweet count : "))
                displayTweets(count2,n1)
            else:
                break

        except ValueError:
```

```
print('\nYou did not enter a valid integer. Please try again!!')
```