Pfizer_scrape

February 10, 2021

[1]:

```
credentials = {'CONSUMER_KEY':'',
               'CONSUMER_SECRET':'',
               'bearer':'',
               'ACCESS_TOKEN':'',
               'ACCESS_TOKEN_SECRET':''}
```

[2]:

```
import tweepy
from datetime import datetime
import pandas as pd
import time
import os
```

[3]:

```
# Authenticate to Twitter
# https://www.linkedin.com/pulse/
 ↪tweepy-tutorial-how-scrape-data-from-twitter-using-python-revanth/

auth = tweepy.OAuthHandler(credentials["CONSUMER_KEY"],␣
 ↪credentials["CONSUMER_SECRET"])
auth.set_access_token(credentials["ACCESS_TOKEN"],␣
 ↪credentials["ACCESS_TOKEN_SECRET"])

api = tweepy.API(auth)

try:
    api.verify_credentials()
    print("Authentication OK")

except:
    print("Error during authentication")
```

```
Authentication OK
```

[12]:

```
def scraptweets(search_words, date_since, numTweets, numRuns, since_id):

    # Define a for-loop to generate tweets at regular intervals
    # We cannot make large API call in one go. Hence, let's try T times
    # Define a pandas dataframe to store the date:
    db_tweets = pd.DataFrame(columns = ['username', 'acctdesc', 'location',␣
↪'following',
                                        'followers', 'totaltweets',␣
↪'usercreatedts', 'tweetcreatedts',
                                        'retweetcount', 'favoritecount',␣
↪'text', 'hashtags', 'id']
                                )
    program_start = time.time()
    for i in range(0, numRuns):
        # We will time how long it takes to scrape tweets for each run:
        start_run = time.time()

        # Collect tweets using the Cursor object
        # .Cursor() returns an object that you can iterate or loop over to␣
↪access the data collected.
        # Each item in the iterator has various attributes that you can access␣
↪to get information about each tweet
        tweets = tweepy.Cursor(api.search, q=search_words, lang="en",␣
↪since_id=since_id, since=date_since,␣
↪max_id=since_id+int(10e14),tweet_mode='extended').items(numTweets)
        # Store these tweets into a python list
        tweet_list = [tweet for tweet in tweets]
        # Obtain the following info (methods to call them out):
        # user.screen_name - twitter handle
        # user.description - description of account
        # user.location - where is he tweeting from
        # user.friends_count - no. of other users that user is following␣
↪(following)
        # user.followers_count - no. of other users who are following this user␣
↪(followers)
        # user.statuses_count - total tweets by user
        # user.created_at - when the user account was created
        # created_at - when the tweet was created
        # retweet_count - no. of retweets
        # (deprecated) user.favourites_count - probably total no. of tweets␣
↪that is favourited by user
        # retweeted_status.full_text - full text of the tweet
        # tweet.entities['hashtags'] - hashtags in the tweet
        # Begin scraping the tweets individually:
        noTweets = 0
        for tweet in tweet_list:
            # Pull the values
            username = tweet.user.screen_name
            acctdesc = tweet.user.description
            location = tweet.user.location
            following = tweet.user.friends_count
            followers = tweet.user.followers_count
            totaltweets = tweet.user.statuses_count
            usercreatedts = tweet.user.created_at
            tweetcreatedts = tweet.created_at
            retweetcount = tweet.retweet_count
            favoritecount = tweet.favorite_count
            hashtags = tweet.entities['hashtags']
            idx = tweet.id
            try:
                text = tweet.retweeted_status.full_text
            except AttributeError:  # Not a Retweet
                text = tweet.full_text
            # Add the 11 variables to the empty list - ith_tweet:
            ith_tweet = [username, acctdesc, location, following, followers,␣
↪totaltweets,
                     usercreatedts, tweetcreatedts, retweetcount,␣
↪favoritecount, text, hashtags, idx]
            # Append to dataframe - db_tweets
            db_tweets.loc[len(db_tweets)] = ith_tweet
            # increase counter - noTweets
            noTweets += 1

        # Run ended:
        since_id = db_tweets['id'].max()
        end_run = time.time()
        duration_run = round((end_run-start_run)/60, 2)

        print('no. of tweets scraped for run {} is {}'.format(i + 1, noTweets))
        print('time take for {} run to complete is {} mins'.format(i+1,␣
↪duration_run))

        time.sleep(920) #15 minute sleep time

    # Once all runs have completed, save them to a single csv file:
    # Obtain timestamp in a readable format
    to_csv_timestamp = datetime.today().strftime('%Y%m%d_%H%M%S')
    # Define working path and filename
    path = os.getcwd()
    filename = path +'/vaccine_tweets.csv'
    # Store dataframe in csv with creation date timestamp
    db_tweets.to_csv(filename, index = False)

    program_end = time.time()
    print('Scraping has completed!')
    print('Total time taken to scrap is {} minutes.'.format(round(program_end -␣
↪program_start)/60, 2))
```

```
    return db_tweets
```

[11]:

```
# Initialise these variables:
search_words = "#pfizer OR #biontech OR #pfizerbiontech OR #pfizervaccin␣
 ↪-filter:retweets"
date_since = "2021-01-30"
numTweets = 2500
numRuns = 12
# Call the function scraptweets
df = scraptweets(search_words, date_since, numTweets, numRuns,␣
 ↪1358789204231278596)
df.head()
# 1355592162940997634
# 1358463997998206979
# 1358789204231278596
```

```
no. of tweets scraped for run 1 is 577
time take for 1 run to complete is 0.36 mins
no. of tweets scraped for run 2 is 1106
time take for 2 run to complete is 0.64 mins
no. of tweets scraped for run 3 is 971
time take for 3 run to complete is 0.62 mins
no. of tweets scraped for run 4 is 106
time take for 4 run to complete is 0.09 mins
no. of tweets scraped for run 5 is 3
time take for 5 run to complete is 0.02 mins
no. of tweets scraped for run 6 is 5
time take for 6 run to complete is 0.01 mins
no. of tweets scraped for run 7 is 5
time take for 7 run to complete is 0.01 mins
no. of tweets scraped for run 8 is 5
time take for 8 run to complete is 0.01 mins
no. of tweets scraped for run 9 is 4
time take for 9 run to complete is 0.01 mins
no. of tweets scraped for run 10 is 9
time take for 10 run to complete is 0.01 mins
no. of tweets scraped for run 11 is 5
time take for 11 run to complete is 0.02 mins
no. of tweets scraped for run 12 is 3
time take for 12 run to complete is 0.02 mins
Scraping has completed!
Total time taken to scrap is 185.85 minutes.
```

[11]:

```
          username                                           acctdesc  \
0        BrazilSFE  Brazil SFE®| We are passionate about improving…
1    _Indiaupdates  India Updates is an independent news & Informa…
2        TMReserve                             Join the real conversation
```

```
3  DrFariyaBukhari  Reality bites & so does my Blog. Dare to indul…
4   TheUltraAliens                                     Intuipreneur

            location following followers totaltweets       usercreatedts  \
0  São Paulo, Brasil      1240        94       48337 2015-01-02 14:13:17
1   New Delhi, India       102       232       10937 2019-02-26 16:12:39
2           Malaysia       189      7352       73126 2011-05-05 16:27:46
3           Pakistan       254       597       53084 2014-04-20 14:54:05
4         Via Lactea      3141       722        7497 2014-11-01 08:39:00

       tweetcreatedts retweetcount favoritecount  \
0 2021-02-02 13:15:27            0             1
1 2021-02-02 13:15:00            0             0
2 2021-02-02 13:07:13            2             1
3 2021-02-02 13:05:28            1             1
4 2021-02-02 13:00:35            0             1

                                                text  \
0  Dê Like! https://t.co/wGCPT8qVpc\nGlobal Pharm…
1  Pfizer-BioNTech to produce 2 bn doses of Covid…
2  Pfizer forecasts $15b in Covid-19 vaccine sale…
3  Valid point. Only PCR negative &amp; Non-react…
4  "7 die at Spanish care home after getting #Pfi…

                                           hashtags                   id
0  [{'text': 'Top10', 'indices': [113, 119]}, {'t…  1356592067955339267
1  [{'text': 'Pfizervaccine', 'indices': [97, 111…  1356591952376987649
2  [{'text': 'Pfizer', 'indices': [59, 66]}, {'te…  1356589995620884481
3  [{'text': 'Pfizer', 'indices': [159, 166]}, {'…  1356589555126788096
4  [{'text': 'Pfizer', 'indices': [42, 49]}, {'te…  1356588324098572288
```

[14]:

```
a = df['id'].max()
a
```

[14]:

```
1358789204231278596
```

[ ]:

```
```