Basic guideline for twitter information crawling

# 1. Download

Download “Project\_Twitter\_Doc.ipynb” from github (which you have done!)

# 2. Program requirement

Import “Project\_Twitter\_Doc.ipynb” into your python editor (e.g. spyder3, visual studio code)

(# For better applying the crawler, we suggest you to install the below version of the application for connecting Postgresql/ sqlite3 as the database system engine to python:

* Best python version as above v3.8.0
* Postgresql as above v13.0
* qlite3 as above version 3.38.5

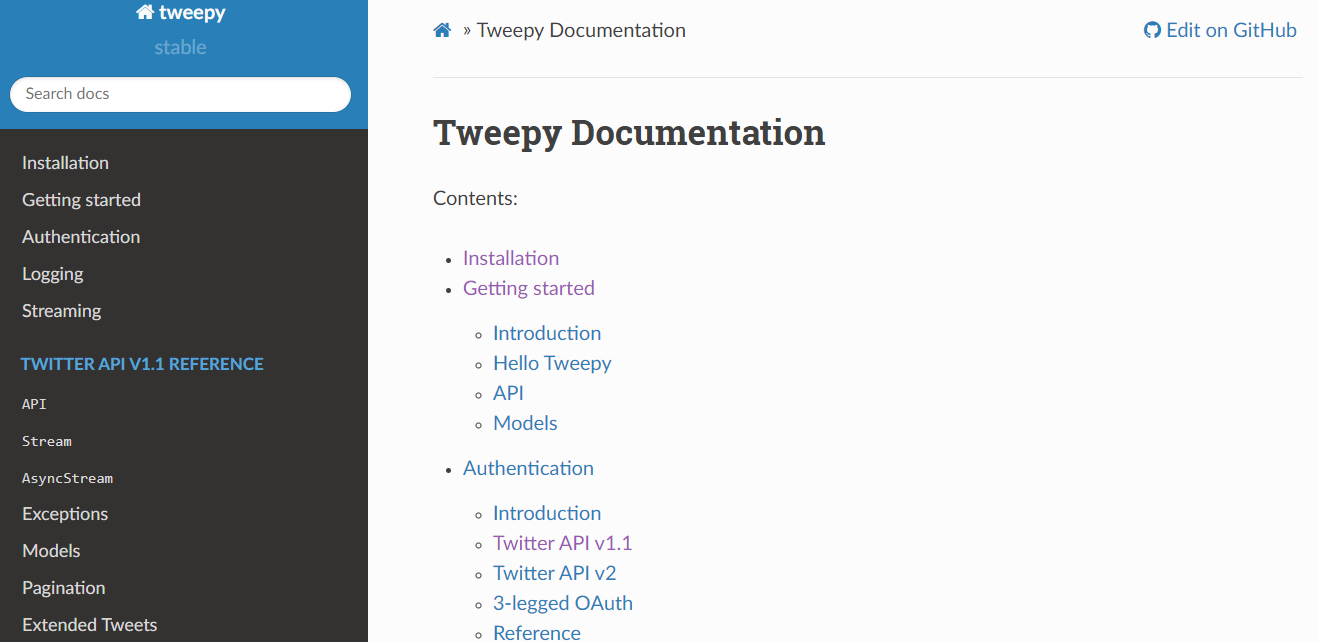
Among the python, we also suggest you to ensure the following tools be installed:

* SQLAlchemy v1.4.36
* tweepy v4.9.0
* psycopg2 v2.8.6
* pandas v1.4.2
* configparser v5.0.2
* json v2020-12

# 3. Tweepy information

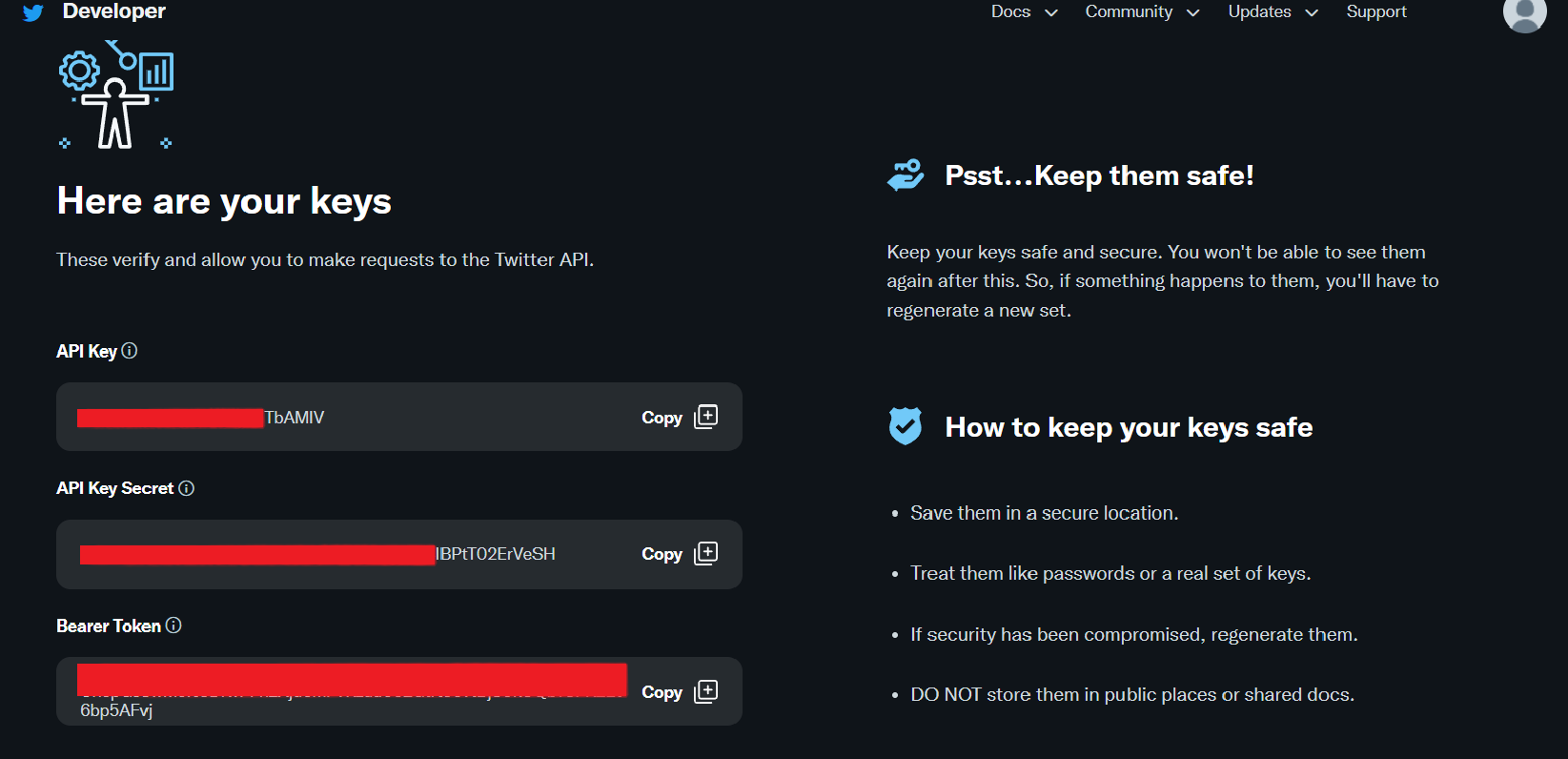
Consumer Keys & Authentication Tokens were required for python library “tweepy”, information for tweepy can be found with the link below:

<https://docs.tweepy.org/en/stable/>



Token should be independent with users, please apply your own account for developer token in twitter official platform:

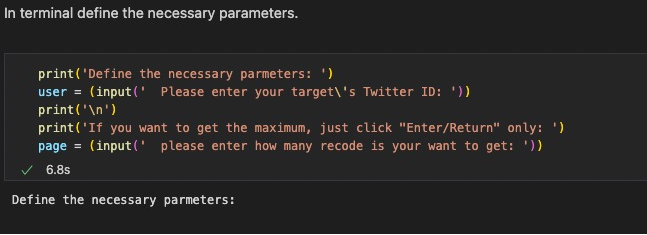
<https://developer.twitter.com/en/portal/dashboard>



# 4. Input target twitter username

Please input “JoeBiden” (president of the US) in the python kernel.

(# In this input section, you can change to other target twitter’s user on your own for analysis.)



# 5. Data filtering

“df2= df[df['tweet'].str.contains('covid|virus|vaccin',case=False)]”

Filtering keywords were default set to be “covid”, “virus” and “vaccin”, which allow users search for the tweets that contain the phrase formed by “covid”, “virus” and “vaccin” in both capital letter or small letter.

These keywords were free to change to other keywords in twitter if you are looking for specific tweets.

Remark:

1. Reasons for keywords chosen:

“covid” goes for “covid-19” , “COVID19”, etc.

“virus” goes for “coronavirus”

“vaccin” goes for “vaccine”, “vaccination”, “unvaccinated”, etc.

1. Keyword “virus” was chosen as the short form of “coronavirus”.

However, searching for “virus” may also contain the tweets that contain keyword only “virus” but not “coronavirus”.

User should beware for the tweets filtered.



# 6. Beware for null data

In the transformation part, the outcome data may be “null”. These data should be dropped.

Only the information about the tweets: the content of tweet, the create time, tweet id, retweet count, and the count of “favorite” would be considered as useful information.

# 7. Connection (pgadmin/sqlite3)

Command “conn = psycopg2.connect(database = 'test',user = 'postgres', password = '1234', host '127.0.0.1',port='5432')”

was used for the connection between python and your sql database.

Which database name ‘test’ , user name ‘postgres’ and password should be followed with your personal setting, host ‘127.0.0.1’ can also be set as ‘localhost’.

# 8. Data Loading

After transformation from python, useful information and data would be loaded to the user's database.

In sqlite3, the database can be stored and output as a database file (.DB).

Or you can also store it in postgresql.

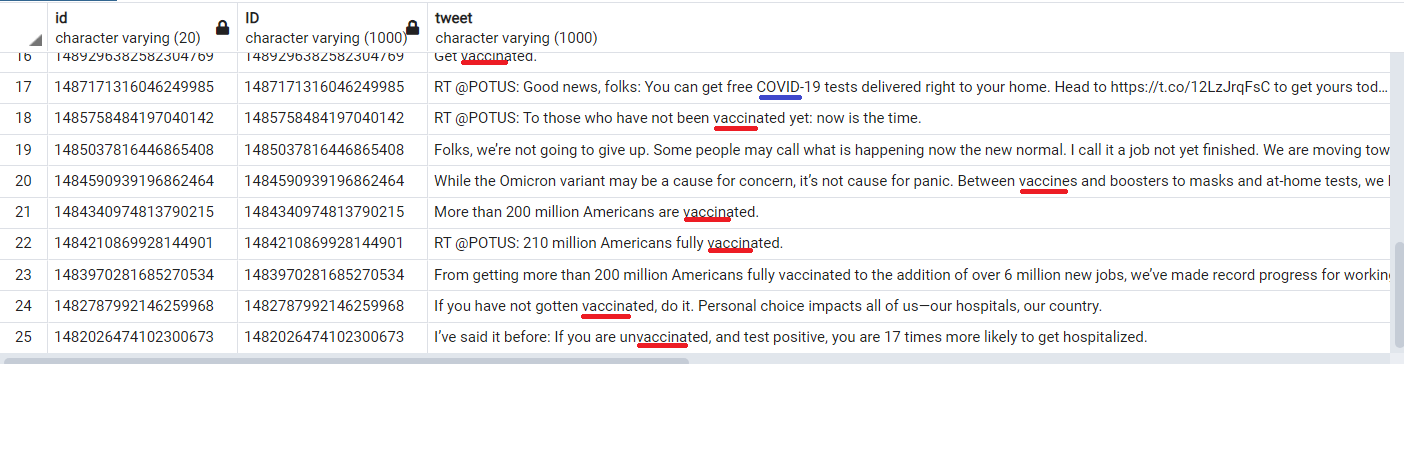


Fig. data successfully loaded and displayed on pgadmin4 after filtration

# 9. ER diagram

