DS 501 Case Study Assignment 1 (120 Points)

Due: 05:00 p.m. on 02/23/2022

Case Study Objective: Twitter Data Collection for a Specific Case Study that Your Group is Interested.

Case Study Deliverables: Please compress all the below files into a zipped file and submit the zip file (**team#_case_study_1.zip**) to Canvas, where **#** is your team number.

1. **Notebook File**: Save this Jupyter Notebook (see the provided template) for Problem 1 to Problem 4 and find the notebook file in your folder (for example, "filename.ipynb"). This is the file that you need to submit. Please make sure that all the answers and plotted tables and figures are in the notebook.

2. **PDF Report**: Please (1) prepare a report (**At Least 5 pages and Not More Than 10 pages**) to report what you found in the data and (2) include figures and tables in the report, but no source code.
   - Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain.
   - Data Sources: What data did you collected? Please also describe your data characteristics in detail.
   - Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail.
   - Results: What did you find in the data? That it, what insights can you obtain from the data?
   - Conclusions: What conclusions can you make from the results?

3. **PPT Slides**: Please prepare PPT slides (for 30 minutes' talk) to present about the case study.
   - 25 minutes for each presentation
   - 5 minutes for questions

Required Readings: Please read the below chapters of the book, **Mining the Social Web (3rd Edition, Jan 2019)**, and the corresponding codes for your case study as a reference. You can read this online book free on https://learning.oreilly.com/library/view/mining-the-social/9781491973547/.

- Chapter 1 (Chapter 1 - Mining Twitter.ipynb)
- Chapter 9 (Chapter 9 - Twitter Cookbook.ipynb)

Note:

(1) The above online book is the latest version that I could find for your reference.

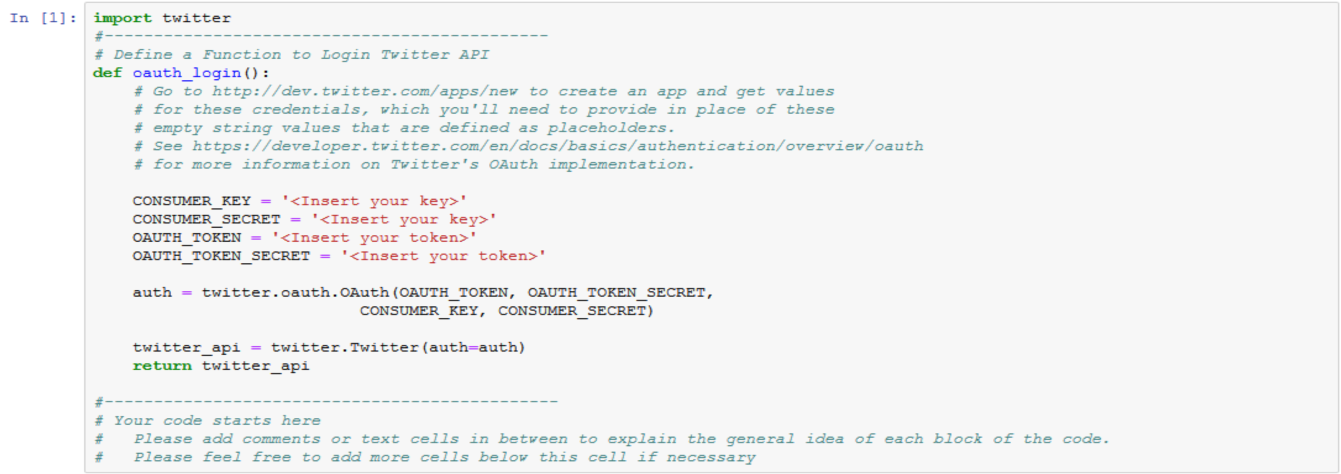
*(2) Please save the given notebook (**.ipynb**) frequently when working in JupyterLab; otherwise, the changes that you made will be lost.*

(3) Note that the setup and configuration for Twitter described in this online book and the given notebook MAY NOT be exactly the same as the latest version of Twitter API. You may like to visit the official website, https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api, to get the up-to-date information if there is any coding issue.

(4) Assignments are accepted in their assigned Canvas drop box without penalty if they are received by 05:00PM EST on the due date, or 10% of the graded score is deducted for the late submission per day. Work submitted after one week of its original due date will not be accepted.

Problem 1: Sampling Twitter Data with Streaming API about a Certain Topic

(1) Select and include a case topic that you are interested in, for example, "WPI" or "Lady Gaga" or "COVID-19".
(2) Use Twitter Streaming API to sample a collection of tweets about this topic in real time. It would be recommended that the number of tweets should be **larger than 1000**, but **smaller than 1 million**.
(3) Store the tweets you downloaded into a local file (txt file or json file).

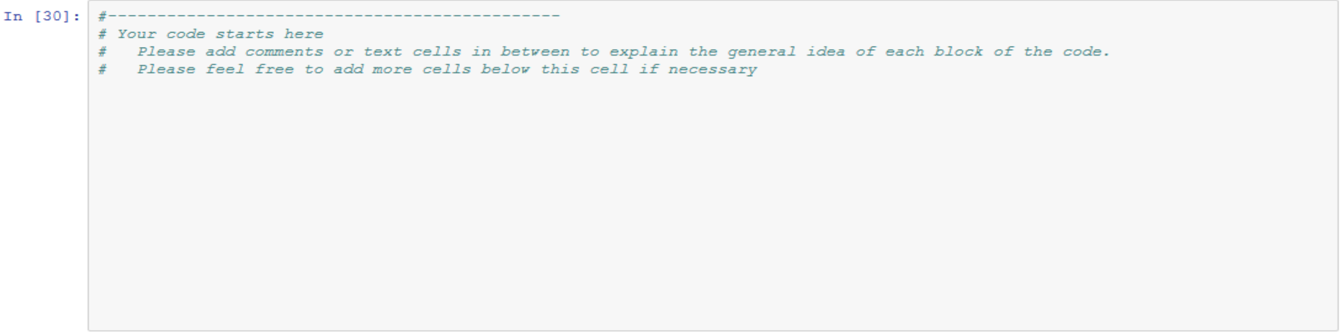


```
In [1]: import twitter
        #----------------------------------------------
        # Define a Function to Login Twitter API
        def oauth_login():
            # Go to http://dev.twitter.com/apps/new to create an app and get values
            # for these credentials, which you'll need to provide in place of these
            # empty string values that are defined as placeholders.
            # See https://developer.twitter.com/en/docs/basics/authentication/overview/oauth
            # for more information on Twitter's OAuth implementation.

            CONSUMER_KEY = '<Insert your key>'
            CONSUMER_SECRET = '<Insert your key>'
            OAUTH_TOKEN = '<Insert your token>'
            OAUTH_TOKEN_SECRET = '<Insert your token>'

            auth = twitter.oauth.OAuth(OAUTH_TOKEN, OAUTH_TOKEN_SECRET,
                                       CONSUMER_KEY, CONSUMER_SECRET)

            twitter_api = twitter.Twitter(auth=auth)
            return twitter_api

        #----------------------------------------------
        # Your code starts here
        #   Please add comments or text cells in between to explain the general idea of each block of the code.
        #   Please feel free to add more cells below this cell if necessary
```

Problem 2: Analyzing Tweets and Tweet Entities with Frequency Analysis

(1) Word Count:

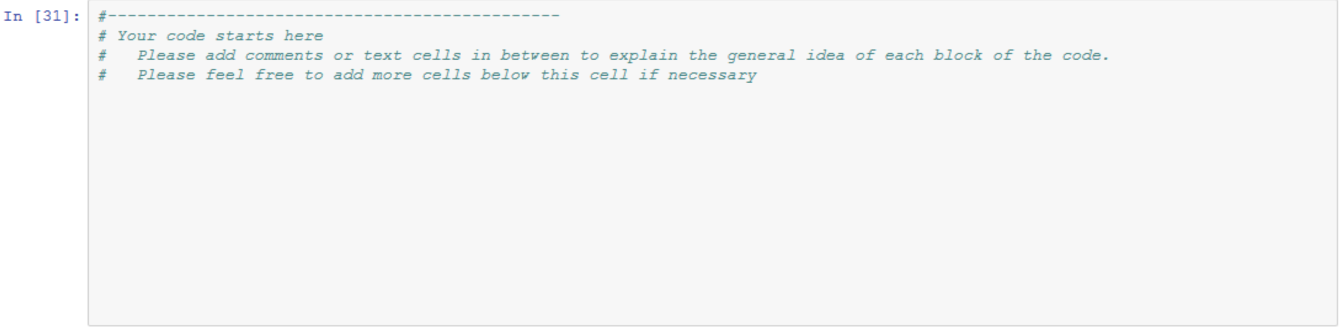
- Use the tweets you collected in Problem 1 and compute the frequencies of the words being used in these tweets.
- Plot a table of the top 30 words with their counts.



```
In [30]: #----------------------------------------------
         # Your code starts here
         #   Please add comments or text cells in between to explain the general idea of each block of the code.
         #   Please feel free to add more cells below this cell if necessary
```

(2) Find the most popular tweets in your collection of tweets.

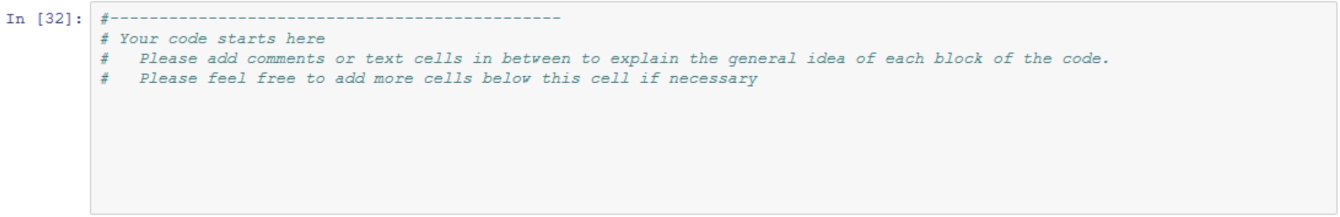
- Please plot a table of the top 10 tweets that are the most popular among your collection, i.e., the tweets with the largest number of retweet counts.



```
In [31]: #----------------------------------------------
         # Your code starts here
         #   Please add comments or text cells in between to explain the general idea of each block of the code.
         #   Please feel free to add more cells below this cell if necessary
```

(3) Find the most popular Tweet Entities in your collection of tweets.

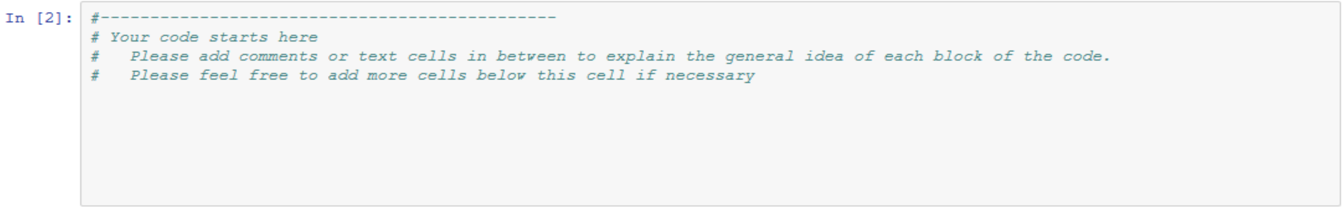
- Please plot a table of the top 10 hashtags and top 10 user mentions that are the most popular in your collection of tweets.



```
In [32]: #----------------------------------------------
         # Your code starts here
         #   Please add comments or text cells in between to explain the general idea of each block of the code.
         #   Please feel free to add more cells below this cell if necessary
```

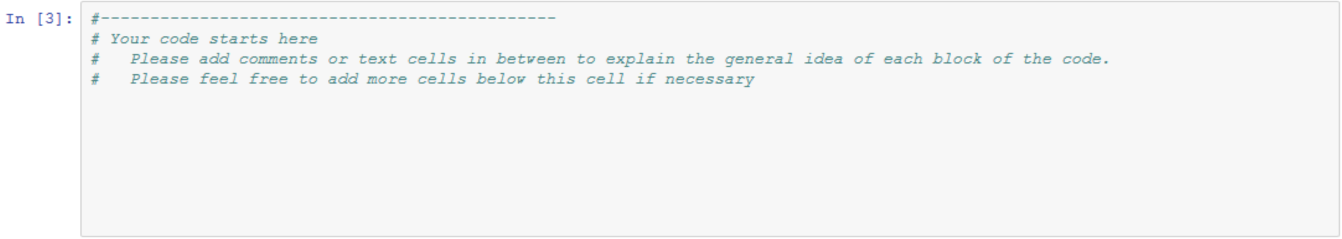
Problem 3: Getting "All" Friends and "All" Followers of a Popular User in Twitter

(1) Choose a popular twitter user who has many followers in the collected tweets of your case topic.
(2) Get the list of all friends and all followers of that twitter user.
(3) Plot 20 out of the followers, plot their ID numbers, and their screen names in a table.
(4) Plot 20 out of the friends (if the user has more than 20 friends), plot their ID numbers, and their screen names in a table.



```
In [2]: #----------------------------------------------
        # Your code starts here
        #   Please add comments or text cells in between to explain the general idea of each block of the code.
        #   Please feel free to add more cells below this cell if necessary
```

(5) Compute the mutual friends within the two groups, i.e., the users who are in both friend list and follower list, plot their ID numbers and their screen names in a table.

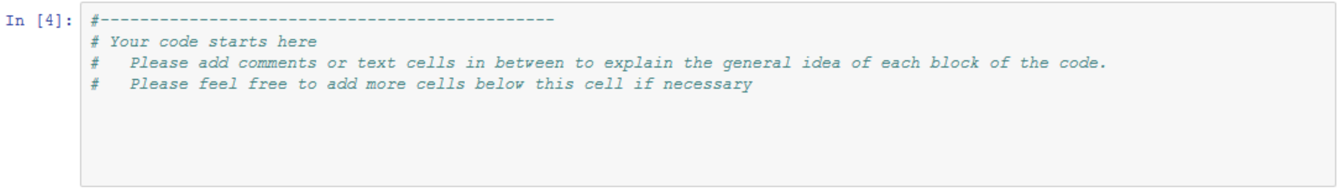


```
In [3]: #----------------------------------------------
        # Your code starts here
        #   Please add comments or text cells in between to explain the general idea of each block of the code.
        #   Please feel free to add more cells below this cell if necessary
```

Problem 4: Domain Question

Run some additional experiments with your collected tweets to gain familiarity with the twitter data and twitter API.

(1) Come up with a question, which Twitter data could help answer from your collected tweets, based upon your case topic of your chosen public organization, private company, social community, etc., in a domain.
(2) How could Twitter data help that company/organization/community spend its resources to answer the above question?



```
In [4]: #----------------------------------------------
        # Your code starts here
        #   Please add comments or text cells in between to explain the general idea of each block of the code.
        #   Please feel free to add more cells below this cell if necessary
```

Grading Criteria:

I. Notebook: Points = 80

Question 1: Points: 20

(1). Select and include a case topic that you are interested in, for example, "WPI" or "Lady Gaga" or "COVID-19". **6 Points**

(2). Use Twitter Streaming API to sample a collection of tweets about this topic in real time. It would be recommended that the number of tweets should be **larger than 1000**, but **smaller than 1 million**. **10 Points**

(3). Store the tweets you downloaded into a local file (txt file or json file). **4 Points**

Question 2: Points: 20

A. <u>Word Count</u>

(1) Use the tweets you collected in Problem 1, and compute the frequencies of the words being used in these tweets. **Points: 4**

(2) Plot a table of the top 30 words with their counts. **Points: 4**

B. <u>Find the most popular tweets in your collection of tweets</u>

Please plot a table of the top 10 tweets that are the most popular among your collection, i.e., the tweets with the largest number of retweet counts. **Points: 4**

C. <u>Find the most popular Tweet Entities in your collection of tweets</u>

(1) Plot a table of the top 10 hashtags. **Points: 4**

(2) Plot the top 10 user mentions that are the most popular in your collection of tweets. **Points: 4**

Question 3: Points: 20

(1) Choose a popular twitter user who has many followers in the collected tweets of your case topic. **Points: 4**

(2) Get the list of all friends and all followers of the twitter user. **Points: 4**

(3) Plot 20 out of the followers, plot their ID numbers and their screen names in a table. **Points: 4**

(4) Plot 20 out of the friends (if the user has more than 20 friends), plot their ID numbers and their screen names in a table. **Points: 4**

(5) Compute the mutual friends within the two groups, i.e., the users who are in both friend list and follower list, plot their ID numbers and screen names in a table. **Points: 4**

Question 4: Points: 20

Run some additional experiments with your data to gain familiarity with the twitter data and twitter API.

(1) Come up with a question, which Twitter data could help answer from your collected tweets, based upon your case topic of your chosen public organization, private company, social community, etc., in a domain.

(2) How could Twitter data help that company/organization/community spend its resources to answer the above question?

Novelty: 10
Interestingness: 10

II. Report (At Least 5 Pages and Not More Than 10 Pages): Points = 20

(1) Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain. **Points: 5**

(2) Data Sources: What data did you collected? Please also describe your data characteristics in detail. **Points: 2**

(3) Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail. **Points: 5**

(4) Results: What did you find in the data? That it, what insights can you obtain from the data? **Points: 5**

(5) Conclusions: What conclusions can you make from the results? **Points: 3**

III. Slides (20 Min of Presentation; 5 Min of Q&A; 5 Min of Changing Teams): Points = 20

(1) Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain. **Points: 2**

(2) Data Sources: What data did you collected? Please also describe your data characteristics in detail. **Points: 2**

(3) Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail. **Points: 2**

(4) Results: What did you find in the data? That it, what insights can you obtain from the data? **Points: 2**

(5) Conclusions: What conclusions can you make from the results? **Points: 2**

(6) Overall Quality of Your Presentation: **Points: 10**

- Storytelling, i.e., how all the above items fit together as a story?
- Description of topic motivation/background, data collection, methodology, results, and conclusions
- Explanation of Question of Interest from the audience