CSE 472: Social Media Mining

Project I - Social Media Data Analysis

Prof. Huan Liu
TA: Zhen Tan
ztan36@asu.edu
Due at 2022, September 22nd, 11:59PM

This is an ***individual*** project assignment. Please refer to the Academic Integrity information at the end of the document before you start this project. You will work on the steps described in the project guideline to create the dataset, codes, and report (See the "Submission" section for details). We highly recommend you finish reading the whole document before starting your implementation. To ensure you can submit on time, we encourage you to start working ***ASAP***!

Project Objectives

In this project, you will learn how to crawl social media data as well as process and perform exploratory analysis on your extracted data. For this project, you are required to crawl data from Twitter about a specific topic: COVID-19 Vaccine.

Project Guideline

Step 1. Choose between the Non-API method and API method to crawl Twitter data:
(1) Non-API. If you choose the Non-API method, there is a tool: snscrape, which is a scraping tool for social networking services (SNS). It can scrape tweets and user profiles without using Twitter's API. snscrape is not just for scraping tweets but also across various other social networking sites like Facebook, Instagram, Reddit, VKontakte, and Weibo. For this project, you need to learn from a simple tutorial to crawl data on Twitter.
(2) API. If you choose API method, you first need to obtain the credentials for scraping the data (i.e., API-key). Detailed instructions for obtaining the credentials for some of the platforms are described in the appendix. We recommend requesting credentials as fast as you can. After you get the credentials, you need to learn how to use API by yourself. You can refer to the official documentation or utilize existing libraries such as Tweepy.

Also you can choose any other Python-based library/tool you like. But whichever you choose, you are required to crawl Twitter data about the specific topic: COVID-19 Vaccine. Please save the collected data in the JSON format and write a brief description in your report on how you crawled your data.
Step 2. Now that you need to collect your data, construct and visualize your data as networks. There are multiple packages and software available for network analysis such as, networkx (recommended), snap, NodeXL, etc.. Choose one and read the instruction on how to build your graph. Each package requires a certain graph data format, such as edge list, adjacency matrix, or adjacency list.

In this step, you are asked to build two social networks with each at least 200 nodes. (To achieve that, you can either collect two folds of data in two JSON files [recommended] or crawl all data in one run and then split it into two JSON files). We offer two options for collecting data and constructing these two networks and you must choose one of them.
(1) Crawl Twitter data about two attitudes towards the COVID-19 vaccine: **anti-vaccine** and **pro-vaccine**. To achieve that, you need to crawl data using specific *searching keywords* or *hashtags*. For example, *"#GetVaccinatedOrGetCovid" and "vaccination work"* for pro-vaccine tweets, *"#NoVaccineForMe" and "#StopVaccination"* for anti-vaccine tweets. You can find more keywords/hashtags from Pro-vaccine and Anti-vaccine to get enough data. For each attitude with corresponding crawled data, you need to build a network independently, and for analysis, you need to contrast the characteristics of these two networks.
(2) Crawl COVID-19 vaccine-related tweets from **two time periods**. To decide the time periods, you need to look at data sources provided by JHU Covid Data Repository, such as the CDC Data Tracker, where you can find various Spatio-temporal COVID-19 statistics (e.g. waves of daily confirmed cases, death, and vaccinations). Choose two time periods that you think are meaningful and comparable (e.g. peak and bottom of new cases), and crawl two folds of Twitter data with the same searching keywords (any keyword related to the COVID-19 vaccine). Then you need to construct two networks independently based on the two folds, and contrast the characteristics of these two networks.

For either option, you need to provide a description of the data collection process in your report. Some representative network types are described as follows, but you are free to define your own network type as long as you can justify it. Please note that besides the twitter content crawled, different types of network may request collecting different features and additional information of tweets and users. In the report, write what features you used for constructing the graphs and take a snapshot of your returned graphs and add it to your report.

1. Friendship Network. A user's friendship network can be represented as a graph where the nodes are the users, and the edges show whether there is a friendship relationship between them. Example: Users and their follower and followee relationships as a directed graph.
2. Diffusion Network. A node represents a user, which can publish, receive, and propagate information. A directed edge between nodes represents the direction of information propagation. Example: tweet propagation where the nodes are users and the edges are retweets/comments/mentions.
3. Word Co-occurrence Network. A network in which the nodes are the words and two words occurring together are linked by an edge.

Step 3. You will learn different network measures in class (Degree Distribution, Clustering Coefficient, Pagerank, Diameter, Closeness, Betweenness, etc.). For each built network, use your chosen package or software from step 2 to obtain "Degree Distribution" and plot it as a *histogram*. Besides this measure, choose two other measures from what you have learned and plot them in case they are returned as distribution, or a number, otherwise. Attach your results to your report.
Step 4. Optional Question (No credit)
Can you spot any significant difference in the characteristics (which do not have to be the network measures, any feature is OK) between the two networks? Can you briefly explain why there is (or is not) such a difference in one or two sentences?

Submission

We will run your code to see if it works for all the steps. The final submission should include data, source codes, and a report (preferred to be PDF format). Submit everything on "Gradescope". You can decide whether to upload your submission as a zip file or not because it will be automatically uncompressed once you upload the zip file. Most importantly, you are only allowed to use **Python** as the programming language. Please make sure your source code has the **".py"** extension. i.e., If you used iPython, then you have to convert it into ".py" extension before you submit your codes.

In summary, your final submission should contain the following items:

1. The project report that satisfies the requirements suggested in each step.
2. The dataset generated by your code (The dataset should be in the format of two ".json" files).
3. Source codes with ".py" extension and related files (e.g., config file, environmental file, etc).

Grading Criteria

pts	Description
1	Select a scraping tool
3	Data Collection
3	Network Construction and Visualization
3	Network Measures Calculation
10	

Table 1: Grading Rubric

Academic Integrity

- To prevent any potential plagiarism, we will randomly select students for each phase of the project and ask them to talk with the TA.
- Your codes in the submission will be automatically checked by the similarity detection tools .
- For Step 2, you have to develop your own code for data scraping. It is NOT permissible to use publicly available datasets.
- For all the steps, you can only *refer* to others' code and use libraries, software, and packages but it is NOT permissible to copy any existing code from others.
- Use a "Reference" section and cite all the tutorials, packages, software, and libraries you used in your data.

APPENDIX: Instructions to obtain API-keys

Twitter:

1. Visit https://developer.twitter.com
2. Log in to your Twitter account or Sign up for a new one.
3. In the top right hand corner click "Apply". Then click "Apply for a developer account".
4. For your primary reason for using Twitter developer tools choose "Student". Then click "Next".
5. You must add a valid phone number to your account to use the Twitter API. Then add your country as "United States" and pick some name for the developer account (you may use your email as a username).
6. Complete the form on how you intend to use your Twitter Developer Account. We recommend rewording the following answers:
   (a) *In your words (how you plan to use Twitter data and/or APIs...)* "I am taking the CSE 472 Social Media Mining course at Arizona State University under Dr. Huan Liu. For my class project, I will use the Twitter API to access user network information and analyze the data using standard metrics."
   (b) *Are you planning to analyze Twitter data?* **Yes** "I will be calculating network measures such as Degree Distribution, Clustering Coefficient, Pagerank, Diameter, Closeness, Betweeness, etc. for my project."
   (c) *Will your app use Tweet, Retweet, like, follow, or Direct Message functionality?* **No**
   (d) *Do you plan to display Tweets or aggregate data about twitter content outside of Twitter?* **Yes** "I will be displaying users as nodes in a graph and use follow / friend relationships as edges. The output will only be displayed in class during the project presentation."
   (e) *Will your product, service, or analysis make Twitter content or derived information available to a government entity?* **No**
7. Review your previous answers and accept the terms and conditions. Then wait for your developer account to be approved. This could take up to a couple days.
8. Once your account has been approved, log back into https://developer.twitter.com and in the top right hand corner click there will be a dropdown menu just to the left of your profile photo. Choose "Apps".
9. Click "Create an app".
10. Fill out the four required fields (App name, Application Description, Website URL, and Tell us how this app will be used) then click "Create".
11. Once the app has been created, you should be able to click "Details" then "Permissions". Configure your application to be Read-only.
12. Next go to the "Keys and tokens" tab. You will see both **Consumer API keys** and **Access token & access token secret**. Use these to invoke Twitter API calls.