MULTIMEDIA UNIVERSITY®

Inquire, Inspire and Innovate

FACULTY COMPUTING AND INFORMATIC

TDS 3751
Social Media Computing

Assignment Part 2

Alternate Method To Scrape Twitter Data

GROUP 6

BY

NO.	STUDENT NAME	STUDENT ID	HANDPHONE
1.	Alvin Fong Weng Yew	1191100280	011-20295617
2.	Tan Sin Zhung	1191100281	010-3661060
3.	Leong Yi Hong	1191100292	011-72892995
4.	Chow Yi Ying	1191100981	012-457 2690

TO

LECTURER NAME	MR. SOO WOOI KING
LECTURE GROUP	TC1L

MULTIMEDIA UNIVERSITY

TRIMESTER 2, SESSION 2022/2023

PART 1

The table below compares different methods of crawling Twitter data: Manual Scraping, PhantomMaster, and APIFY.

Manual Scraping: This method refers to the manual process of gathering Twitter data. It involves manually searching, browsing, and copying the desired information from Twitter profiles. Users must visit the Twitter website, search for specific profiles, and manually extract the data by copying and pasting it into a document or spreadsheet. Manual scraping is time-consuming and susceptible to human errors.

PhantomMaster: PhantomMaster is a tool that provides API-based data extraction solutions. It offers pre-built APIs specifically designed for scraping data from various platforms, including Twitter. With PhantomMaster, users can make API requests to extract Twitter data programmatically. This method offers more automation and efficiency compared to manual scraping.

APIFY: APIFY is a platform that offers a range of tools and services for web scraping and automation. It provides a marketplace with pre-built scraping tools, including Twitter scrapers. With APIFY, users can access ready-to-use scrapers designed for Twitter data extraction. These scrapers can be configured with various parameters, such as the Twitter profile page link and the desired data to extract. APIFY provides a user-friendly interface and simplifies the process of scraping Twitter data.

In summary, manual scraping involves manually extracting data from Twitter profiles, while PhantomMaster and APIFY offer API-based solutions that enable automated and efficient data extraction from Twitter. These API-driven methods eliminate the need for manual browsing and copying, saving time and effort. PhantomMaster and APIFY provide more streamlined approaches for retrieving Twitter data by leveraging the power of APIs and automation.

	Manual Scrapping	PhantomBuster **PhantomMaster**	APIFY **APIFY**
Link		https://phantombuster.com/	https://apify.com/
Cost	Free	14 Days Free Trial	Free $5 Monthly Usage
Speed	Slow	Medium	Fast
Ease of use	Takes time to scrape ourself	Need to perform some setup on Chrome before executing the API	Easy to use.

Flexibility	Flexible, can scrape whatever data we want, but manually	Medium flexibility, need to update the profile link for each runtime in the settings.	High flexibility, the tweet can be crawled after putting in the profile links and the number of posts.
Scalability	Not suitable for large-scale data extractions takes a lot of time and effort	Suitable for large-scale data extractions	Suitable for large-scale data extractions
Data Consistency	Not consistent, human error might occur	Consistent	Consistent
Maintenance and Upgrade	Require continuous efforts, especially when the process needs to repeat periodically	Easy to accommodate any changes to datasets	Easy to accommodate any changes to datasets
Data Access	Public access to data	May have some problems, if Twitter policy changes	May have some problems if Twitter policy changes
Error Handling	Manual scrapping will not be having any issues	Network errors, timeout and other issues may occur during scraping	Network errors, timeout and other issues may occur during scraping
Data Format and Structure	Unstructured or semi-structured data, depending on how it is organized during the copy-paste process.	CSV, JSON	EXCEL, JSON, CSV, XML, HTML Tables, RSS, JSONL
Advanced Filtering and Analysis	No	No	Can select which features to filter and download

PART 2

After comparing the methods of scraping data from Twitter, we have decided to extract the data using APIFY. APIFY offers several benefits compared to the other two methods, manual scraping and PhantomMaster.

APIFY provides fast data extraction capabilities, allowing us to retrieve Twitter data efficiently. Its user-friendly interface and simplified setup process make it easy to navigate and utilize the tool effectively. We can customize the extraction parameters to fit our specific needs, including selecting the desired profile links, defining the number of posts to retrieve, and choosing the data format for export.

APIFY is well-suited for large-scale data extractions and ensures consistent and reliable results. It incorporates robust error-handling mechanisms, such as network error management and timeout handling, to minimize potential interruptions. We can rely on APIFY to handle changes in datasets or Twitter's API requirements seamlessly, allowing for easy maintenance and upgrades.

By utilizing APIFY for Twitter data extraction, we can streamline the process, enhance efficiency, and ensure accurate and consistent results. The tool's flexibility in supporting various data formats, such as JSON, CSV, Excel, and XML, enables seamless integration into our downstream analysis and processing workflows.

Overall, APIFY offers speed, ease of use, flexibility, scalability, data consistency, easy maintenance and upgrades, error-handling capabilities, and support for multiple data formats. These advantages make APIFY the ideal choice for extracting Twitter data efficiently and effectively.

Using APIFY for Twitter data extraction offers several advantages, but there are also some talking points regarding potential issues, problems, and requirements:

1. Cost: While APIFY offers a free usage tier, there may be additional costs associated with higher usage or specific features. It's important to understand the pricing structure and evaluate if the cost aligns with your budget and project requirements.

2. Setup and Configuration: APIFY requires some initial setup and configuration before executing the API. This setup typically involves setting up the Chrome browser and configuring the API calls. Users need to familiarize themselves with these setup steps and ensure they are performed correctly for the desired functionality.

3. Rate Limits and Throttling: Twitter has rate limits in place to control API usage. When using APIFY to scrape Twitter data, it's important to monitor and adhere to these rate limits to avoid errors or restrictions. Understanding and managing rate limits is crucial to ensure smooth data extraction.

4. Data Privacy and Compliance: Twitter has specific policies and terms of service that dictate how data can be accessed, used, and stored. Ensure that your usage of APIFY and the extracted Twitter data comply with these policies to avoid any legal or ethical issues. Be aware that changes in Twitter's policies can potentially impact API access or data usage.

5. Network Errors and Timeout Handling: APIFY, like any web scraping tool, can encounter network errors or timeouts during the data extraction process. It's important to implement error handling and retry mechanisms to handle these issues gracefully and ensure data integrity.

6. Data Format and Structure: APIFY offers various data formats for exporting extracted Twitter data, such as JSON, CSV, Excel, XML, and more. Consider the specific format and structure requirements of your downstream data processing and analysis tools to ensure compatibility and seamless integration.

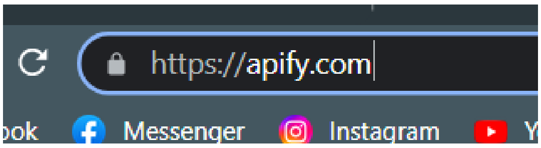
7. Scalability: APIFY is designed to handle large-scale data extractions. However, it's essential to monitor performance and scalability as your data volume increases. Consider factors such as API response times, resource utilization, and potential bottlenecks to ensure efficient and scalable data extraction.

8. Documentation and Support: Familiarize yourself with APIFY's documentation and available support resources. Understanding the API endpoints, parameters, and best practices will help you make the most of the tool. Additionally, access to reliable customer support or community forums can be beneficial when facing issues or seeking guidance.

Below are the steps to extract data from Twitter using APIFY:

Step 1:

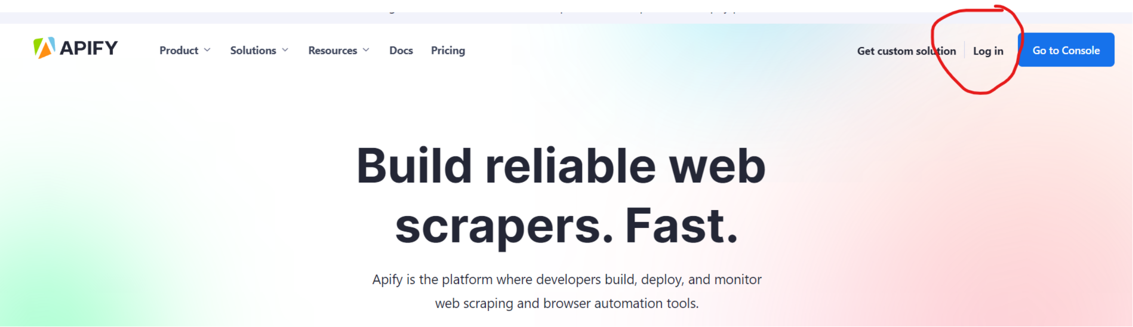
Open the Apify website by visiting https://apify.com/. The Apify platform offers various tools and services for web scraping and automation.



https://apify.com

Step 2:

If you already have an Apify account, log in using your credentials. If you don't have an account, you must register for a new account to proceed.

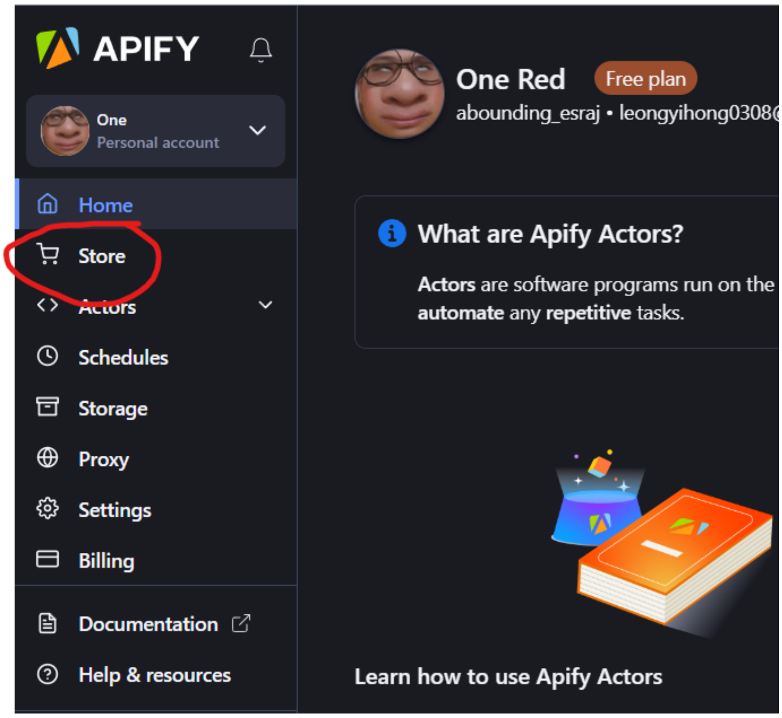


Build reliable web scrapers. Fast.

Apify is the platform where developers build, deploy, and monitor web scraping and browser automation tools.

Step 3:

Once logged in, navigate to the "Store" section of the Apify website. This section provides a marketplace for different pre-built scraping tools and solutions.



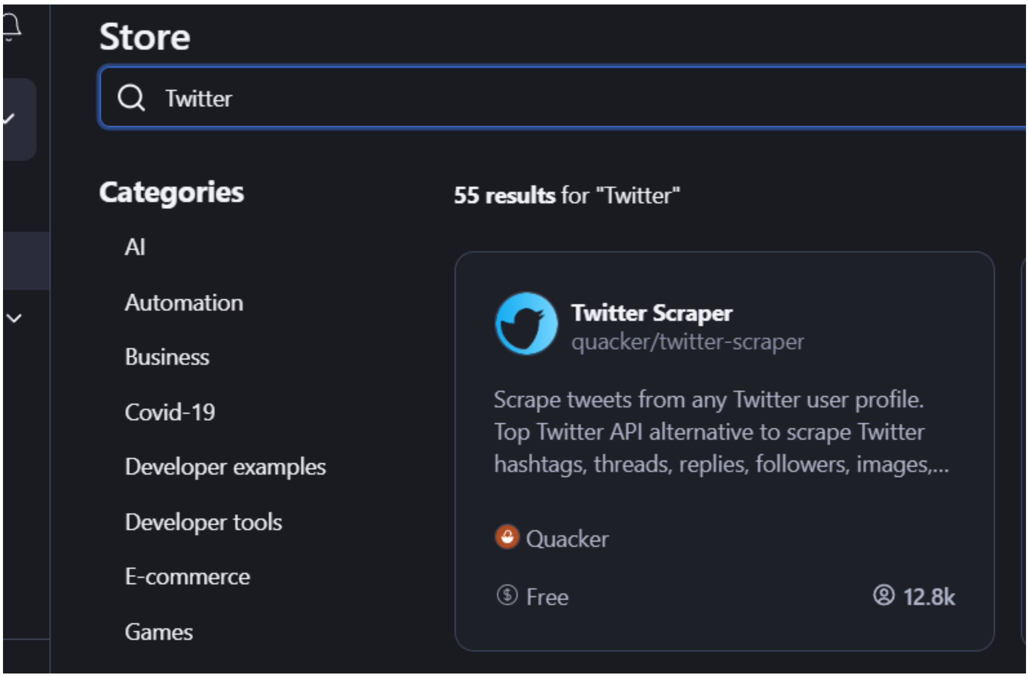
What are Apify Actors?

Actors are software programs run on the
automate any repetitive tasks.

Learn how to use Apify Actors

Step 4:

In the search bar within the Store, type "Twitter Scraper" and hit enter. This will search for available Twitter scraping solutions on Apify. Select the Twitter Scrapper provided by Quaker.



55 results for "Twitter"

Twitter Scraper
quacker/twitter-scraper

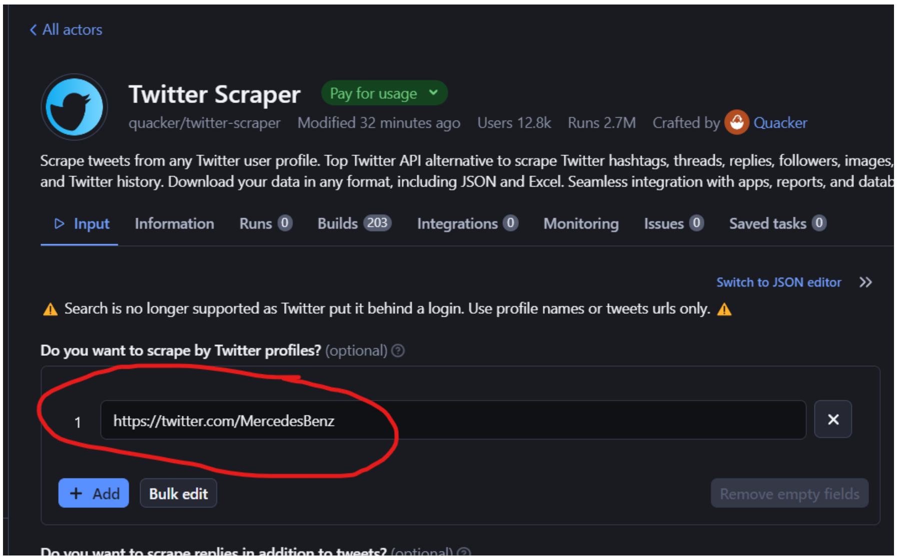
Scrape tweets from any Twitter user profile. Top Twitter API alternative to scrape Twitter hashtags, threads, replies, followers, images,...

Quacker

Free 12.8k

Step 5:

After selecting the Twitter scraper tool, you'll typically need to provide the Twitter profile page link that you want to scrape. This link could be the URL of a specific user's Twitter profile.


< All actors
Twitter Scraper
Pay for usage
quacker/twitter-scraper Modified 32 minutes ago Users 12.8k Runs 2.7M Crafted by Quacker

Scrape tweets from any Twitter user profile. Top Twitter API alternative to scrape Twitter hashtags, threads, replies, followers, images, and Twitter history. Download your data in any format, including JSON and Excel. Seamless integration with apps, reports, and datab

⚠ Search is no longer supported as Twitter put it behind a login. Use profile names or tweets urls only. ⚠

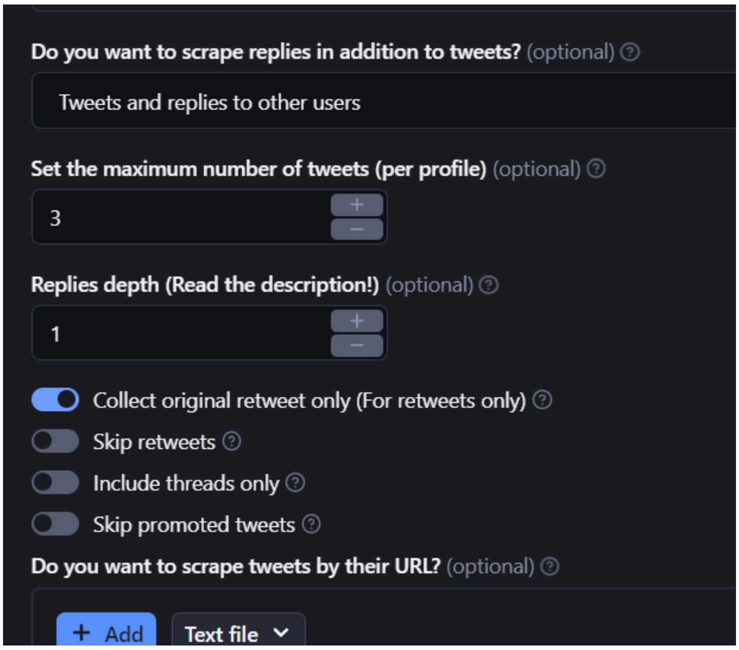
Do you want to scrape by Twitter profiles? (optional)

1 https://twitter.com/MercedesBenz

Do you want to scrape replies in addition to tweets? (optional)

Step 6:

Depending on the specific scraper tool you choose, there may be additional configuration options available. These configurations can include specifying the data you want to extract, such as tweets, followers, or user details. Adjust the configurations as necessary.



Do you want to scrape replies in addition to tweets? (optional)

Tweets and replies to other users

Set the maximum number of tweets (per profile) (optional)

3

Replies depth (Read the description!) (optional)

1

● Collect original retweet only (For retweets only)

○ Skip retweets

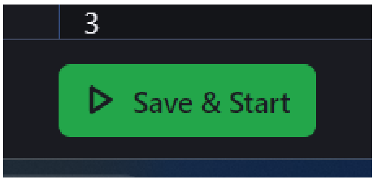
○ Include threads only

○ Skip promoted tweets

Do you want to scrape tweets by their URL? (optional)

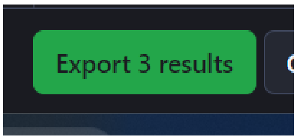
Step 7:

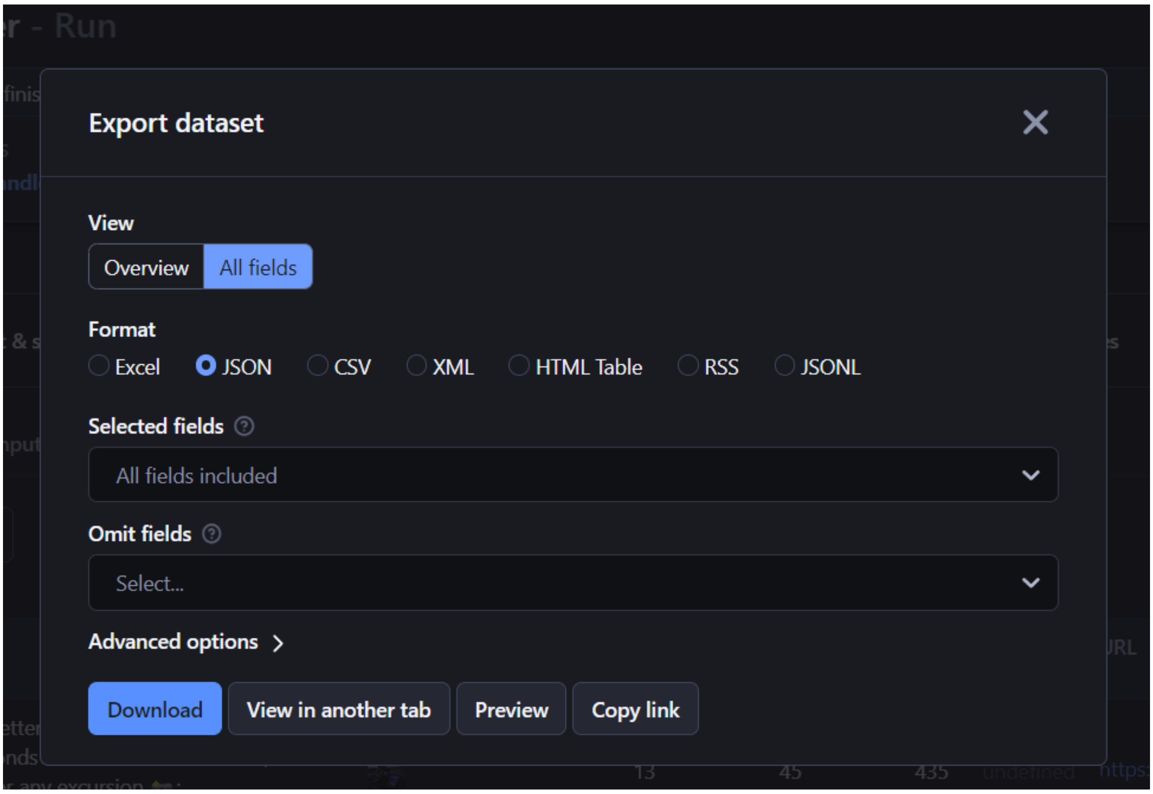
Once you have configured the scraper tool, look for the "Save & Start" button or a similar option to initiate the scraping process. This will start the tool, and it will begin fetching and extracting data from the provided Twitter profile.



Step 8:

After the scraping process is completed, you can typically export the results and download the data files. The specific method of exporting and downloading the data can vary based on the scraper tool or Apify's interface. Look for options to export the data in formats such as CSV, JSON, or Excel.





Export dataset

View

Overview [✓ All fields]

Format

○ Excel ● JSON ○ CSV ○ XML ○ HTML Table ○ RSS ○ JSONL

Selected fields

All fields included

Omit fields

Select...

Advanced options