Social Media Scrapper

Prof Vaishali Jabade, Mandar Kulkarni

Department of Electronics and Telecommunication Engineering

***Abstract* — Social media is an ocean of data and information. There are thousands of images shared daily. But all the data on internet is scattered. This paper aims at collecting that data from that ocean through an efficient and automatic ways. The technique for Scrapping of social media websites like twitter and Instagram to create images or tweets dataset is discussed in detail. A GUI is also made so as to automate the process and make it more user friendly.**

Keywords *— social media, scrapping, GUI*

I. Introduction

Social media scrapping refers to collecting data from websites like Twitter, Instagram etc. We have done this using Python programming language and various libraries like NumPy, Pandas, selenium, wget etc. For GUI development tkinter library is used. This paper discusses scrapping of two websites, Instagram and Twitter. Instagram is a website where people share short videos and pictures. Thousands of pictures with various hashtags associated with picture are posted every second on Instagram. Similarly, Twitter is a website where people tweet their opinions about any topic. Hot discussions are trended due to mass use of the hashtags used in the tweet. The sites are active from so many years that results of all almost hashtags are available. The proposed GUI asks the user about his/her website choice and then directs the user to login page. After logging in the user can search the hashtag and the images/tweets from the website with that hashtag will be downloaded with the help of selenium web browser and wget library The process flow is discussed in detail in this paper

II. Methodology/Experimental

A. Libraries

- Tkinter

Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit. Tkinter provides various controls, such as buttons, labels and text boxes used in a GUI application. These controls are commonly called widgets.

- Selenium

Selenium Python bindings provides a simple API to write functional/acceptance tests using Selenium WebDriver. Through Selenium Python API you can access all functionalities of Selenium WebDriver in an intuitive way. Selenium Python bindings provide a convenient API to access Selenium Web Drivers like Firefox, Ie, Chrome, Remote etc. The current supported Python versions are 3.5 and above.

- Wget

The wget command is a non-interactive utility to download remote files from the internet which is built-in with Unix based operating systems. It supports HTTP, HTTPS, and FTP protocols, as well as retrieval through HTTP proxies.

B. Process Flow

The major steps are:

1. **A simple GUI**

GUI was made using Tkinter library. It mainly involves using the proper widgets like Buttons, text Label so as to create different pages. In this project we have made login pages for Instagram and Twitter so user needs to login to scrap data. A GUI which asks the user for the “search term” is also made where the user needs to enter the hashtag according to his/her choice**.**

2. **Creating Web driver instance**

Any webdriver like chrome, Firefox can be used. Selenium library is used to create the webdriver instance. Using that variable we can open any desired webpage and we can access the data on that page using the inspect element, we can even access any particular section on the webpage by copying the xpath which makes it easier to grab image link .

3. **Scrapping the data**

Images can be easily downloaded ones we can get their links from the Xpath. Wget library is used for it. It takes link and location as an input and then downloads the link and stores it into desired location. The tweet information is first stored into dataframes which are then converted into csv file.

III. RESULTS/DISCUSSIONS

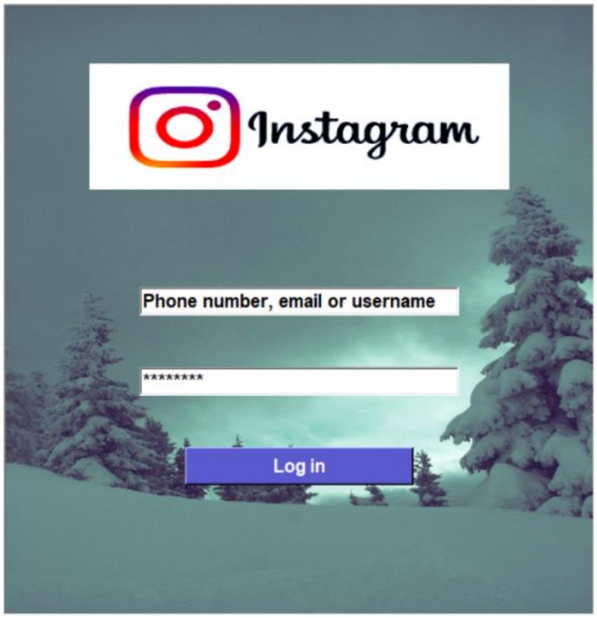
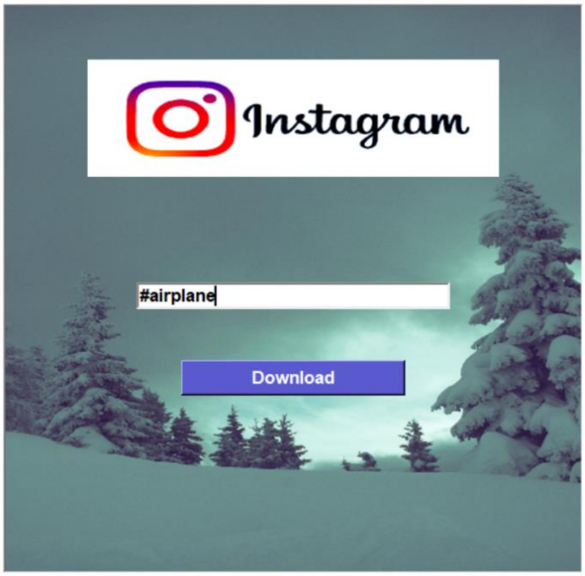
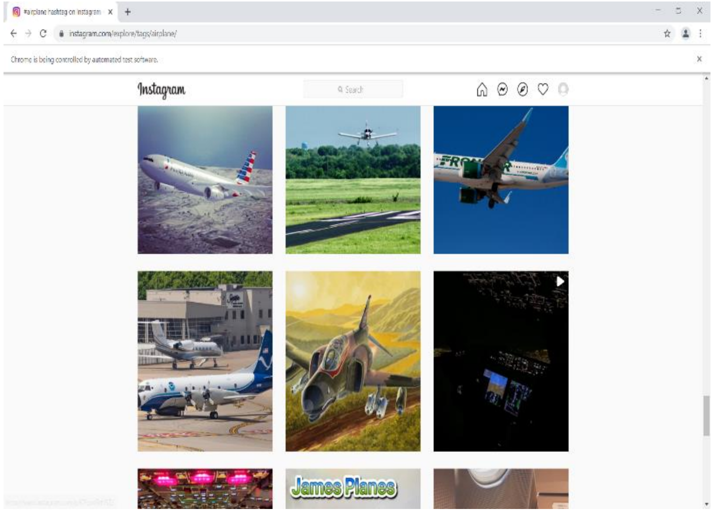


Fig: Instagram Login UI



#airplane

Fig: Searching for desired hashtag


#airplane hashtag on Instagram
instagram.com/explore/tags/airplane/
Instagram
Search
James Planes

Chrome is being controlled by automated test software.

Fig: #airplane searched automatically on Instagram and all images are being downloaded

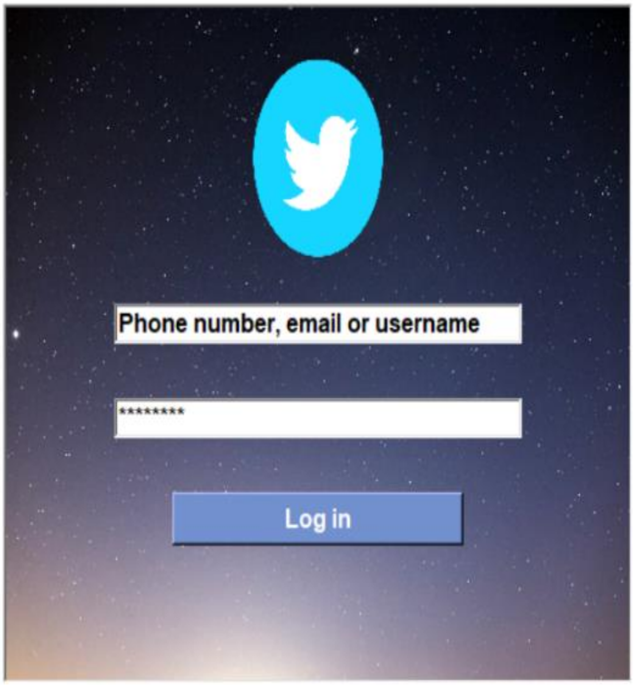
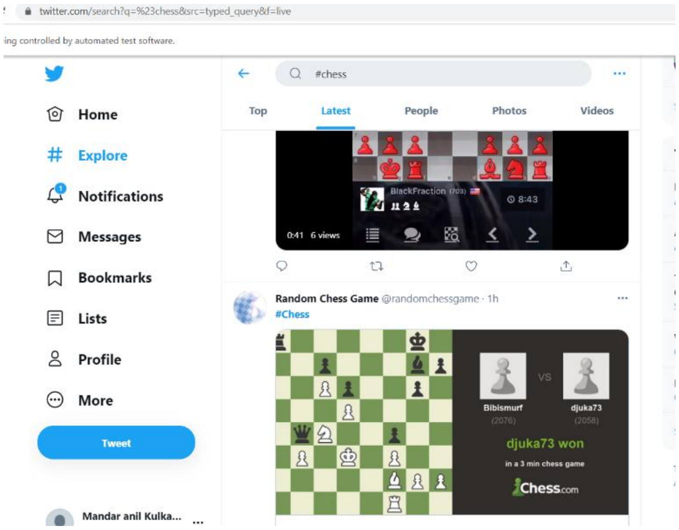
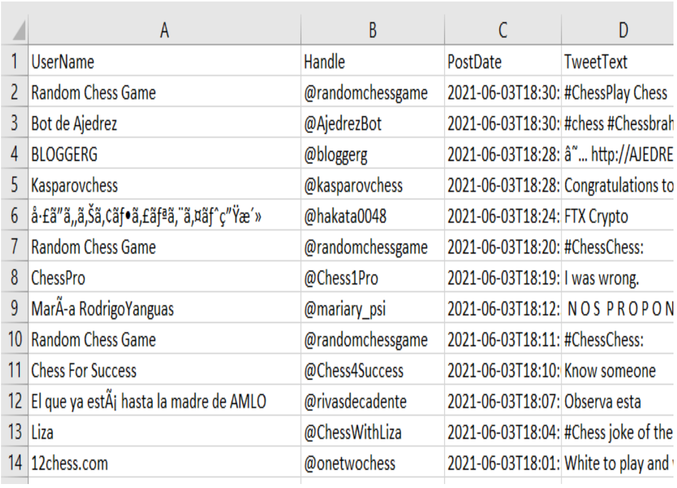


Fig: Twitter Login UI



Random Chess Game @randomchessgame · 1h
#Chess

Fig: #chess searched on twitter and all data regarding the tweets with hashtag chess is being converted to a csv file



	A	B	C	D
1	UserName	Handle	PostDate	TweetText
2	Random Chess Game	@randomchessgame	2021-06-03T18:30:	#ChessPlay Chess
3	Bot de Ajedrez	@AjedrezBot	2021-06-03T18:30:	#chess #Chessbrah
4	BLOGGERG	@bloggerg	2021-06-03T18:28:	â˝… http://AJEDRE
5	Kasparovchess	@kasparovchess	2021-06-03T18:28:	Congratulations to
6	å·£ã"ã„ã‚Šã‚¢ãƒ•ã‚£ãƒªã‚¨ã‚¤ãƒˆç"Ÿæ´»	@hakata0048	2021-06-03T18:24:	FTX Crypto
7	Random Chess Game	@randomchessgame	2021-06-03T18:20:	#ChessChess:
8	ChessPro	@Chess1Pro	2021-06-03T18:19:	I was wrong.
9	MarÃa RodrigoYanguas	@mariary_psi	2021-06-03T18:12:	NOS PROPON
10	Random Chess Game	@randomchessgame	2021-06-03T18:11:	#ChessChess:
11	Chess For Success	@Chess4Success	2021-06-03T18:10:	Know someone
12	El que ya estÃ¡ hasta la madre de AMLO	@rivasdecadente	2021-06-03T18:07:	Observa esta
13	Liza	@ChessWithLiza	2021-06-03T18:04:	#Chess joke of the
14	12chess.com	@onetwochess	2021-06-03T18:01:	White to play and

Fig: Tweet Information stored in a CSV file

IV. LIMITATIONS

1] Data is boon if used for good purposes but it is a curse if it falls in wrong hands so there are chances that the scrapping may lead to privacy breach

2] Social media sites especially Instagram are very concerned about privacy rules hence there are certain limitations and barrier for bot scrapper. Scrapping data for anything other than personal use is strictly prohibited.

3] If wrong tags are used to describe an image, then it might happen that some false image will be added to the dataset so manual sorting is required once the whole process is finished.

4] Certain tweets contains images which are ignored in current prototype hence the tweet might get misinterpreted which can give some false results.

V. FUTURE SCOPE

1] Image/Video links posted in a tweet can also be downloaded so as to get the complete meaning of the tweet.

2]Downloading multiple images in a post is not possible in wget library hence in any only the first image is downloaded.

3] Sometimes only a single hashtag is insufficient to describe the type of dataset we want, so data sorted by multiple hashtags is possible

4] The proposed systems can also be used to scrap other websites like LinkedIn to organise internships / job as per the title which will be beneficial to lot of aspirants.

VI. CONCLUSION

Social media is a great way of connecting people. People share photos, videos with one another. Social media is also a great medium to share views /opinions about any topic. This data can be used for various purpose/projects like sentiment analysis or for business use if organized properly. With the algorithm and technique discussed in this paper image and tweets can be properly organized using tags. A simple GUI is also made which can make this process more user friendly. Overall, we conclude that the organized data by scrapping can be a boon if used in legal way.

ACKNOWLEDGMENT

Our sincere appreciation goes to staff members and out guide of the Department of Electronics and Telecommunication, Vishwakarma Institute of Technology, Pune for the successful completion of this project.

REFERENCES

1. https://selenium python.readthedocs.io/installation.html#introduction
2. https://ieeexplore.ieee.org/document/8822022
3. https://towardsdatascience.com/web-scrape-twitter-by-python-selenium-part-1-b3e2db29051
4. https://towardsdatascience.com/web-scrape-twitter-by-python-selenium-part-2-c22ae3e78e03
5. https://www.geeksforgeeks.org/python-gui-tkinter/#:~:text=Python%20offers%20multiple%20options%20for,to%20create%20the%20GUI%20applications.