Web Mining project report

Digital Assignment 2-3

Social media scrapper

(Twitter and Instagram)

Aswin S – 17BCE1201

J Meghana – 17BCE1113

**Summary –**

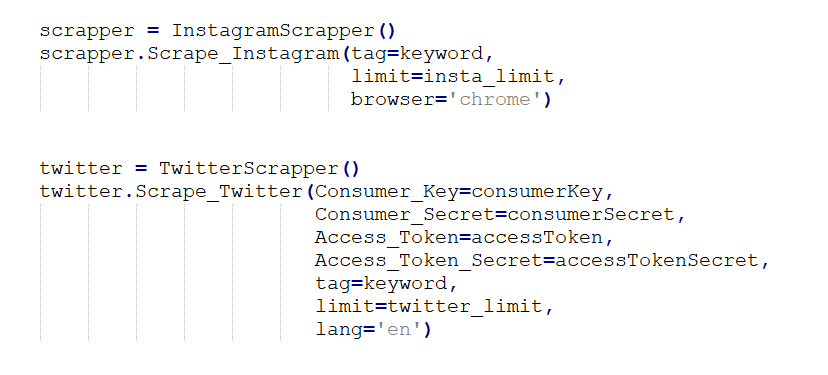
Most social media sites are a collection of user-generated data in various formats in an unstructured manner accessible through the web. This is a strong scrapper/crawler that gets images, tweets, descriptions, hashtags from social media sites – Instagram and Twitter in a well-organized form. Data scrapped from social media is the largest and most dynamic dataset that can be used to analyze various trends and gather useful information.

**Scrapper.py -**

The main snippet of the code, this includes importing the two scrappers and takes in the user input for

* The keyword to scrap for
* No of posts from Instagram
* No of posts from Twitter

Information is scrapped using the keyword, limit and browser. Scraping Twitter requires consumer and accesstoken key and secret key.



**InstagramScrapper.py**

Opens browser and stores images, links and captions. Creates a data directory where the images are downloaded. Next, it opens excel and dumps all the data in the file where hashtags and captions are sorted and stored based on their frequencies.

**TwitterScrapper.py**

Tweepy supports oauth authentication. It appends to the existing images scrapped from Instagram. And a workbook in excel for getting sorted url’s, hashtags further sorted by importance or frequency.

**Software used –**

Python

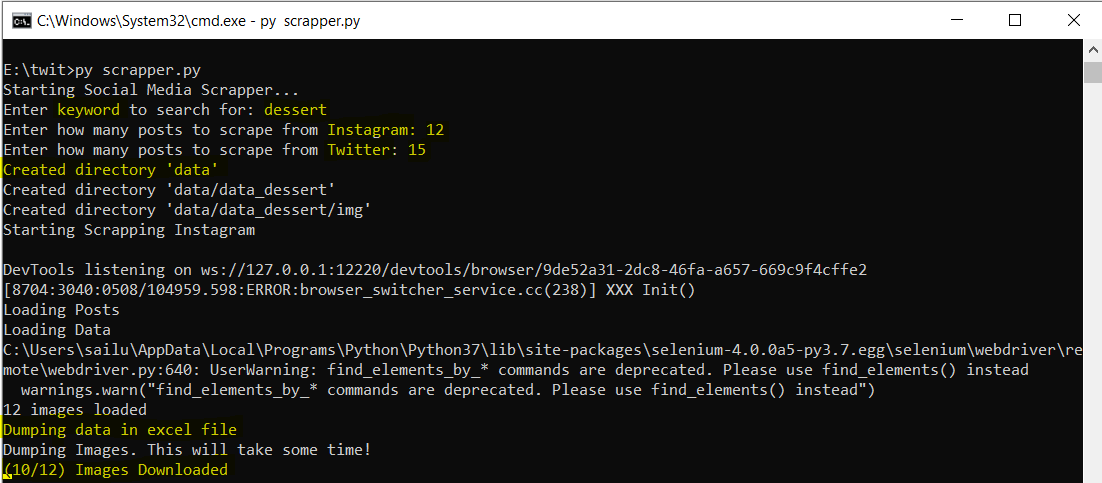
Tweepy – ( To access Twitter API)

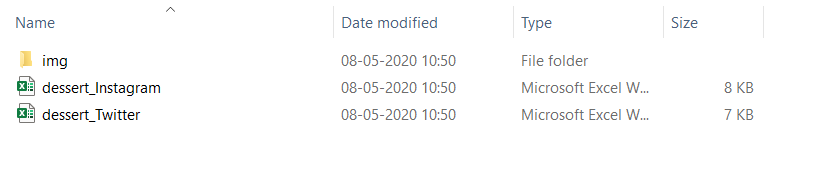
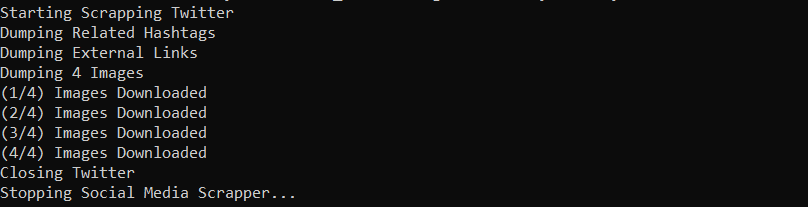
Selenium – (To automate testing across browser)

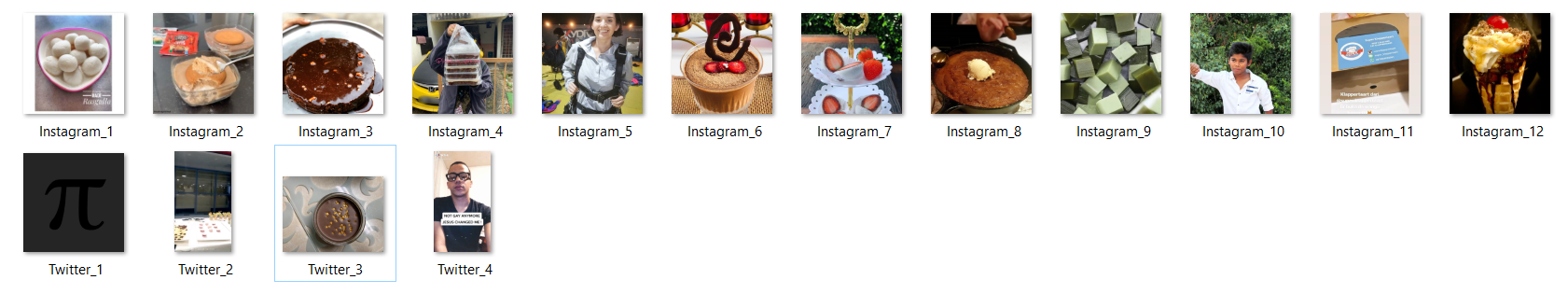
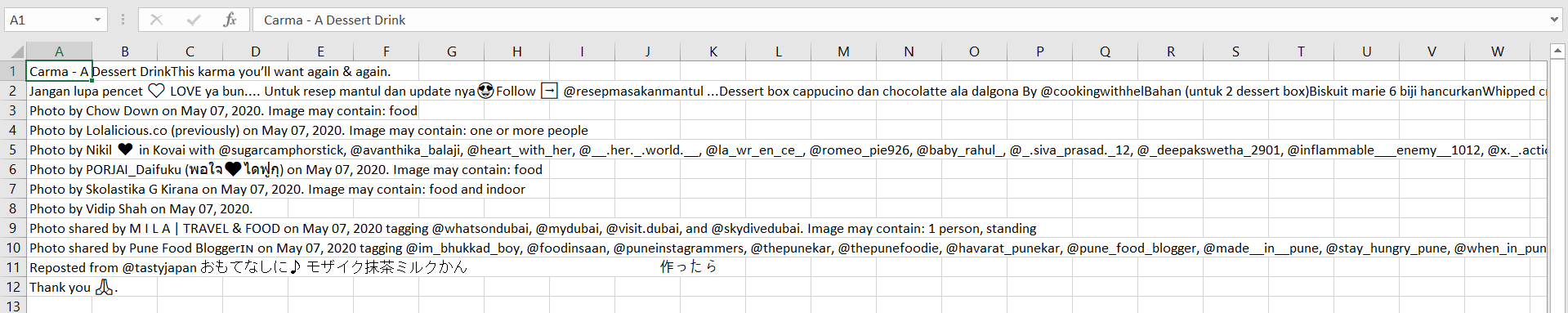
URLLIB – (URL handling module)

Openpyxl – (Used to read/write excel type files)

**Example run –**

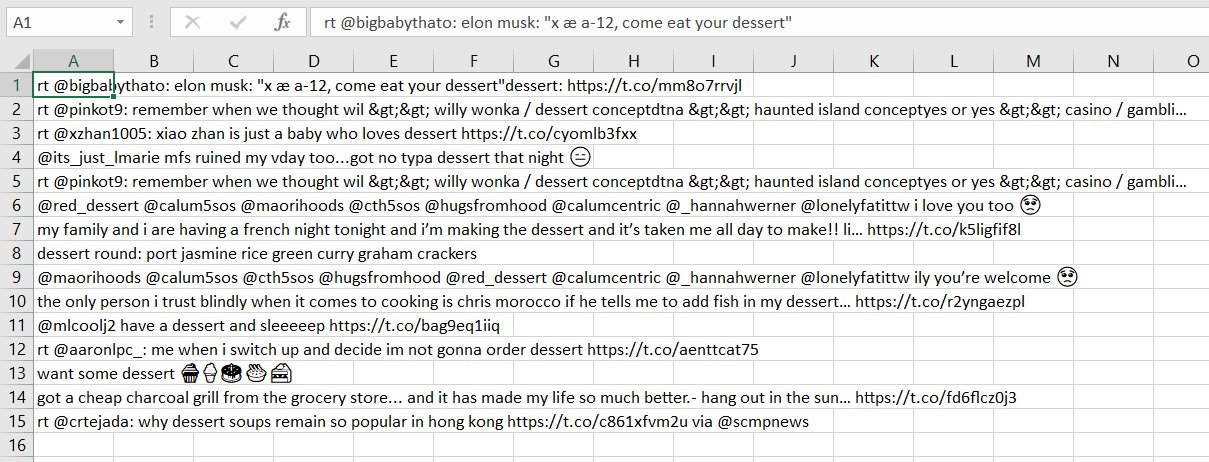


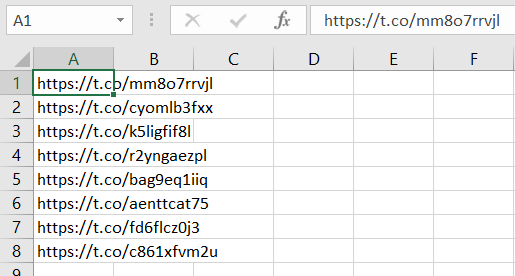
**Instagram images, descriptions**



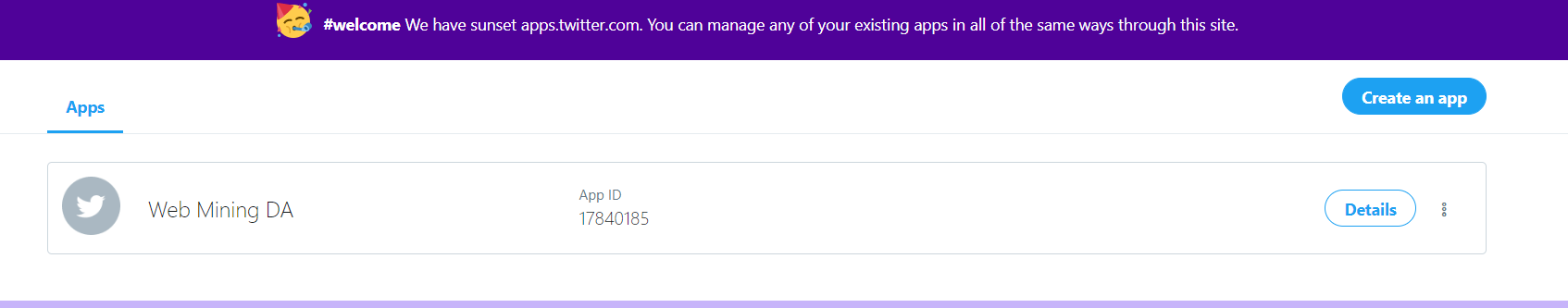
**Twitter descriptions, tweets and links**



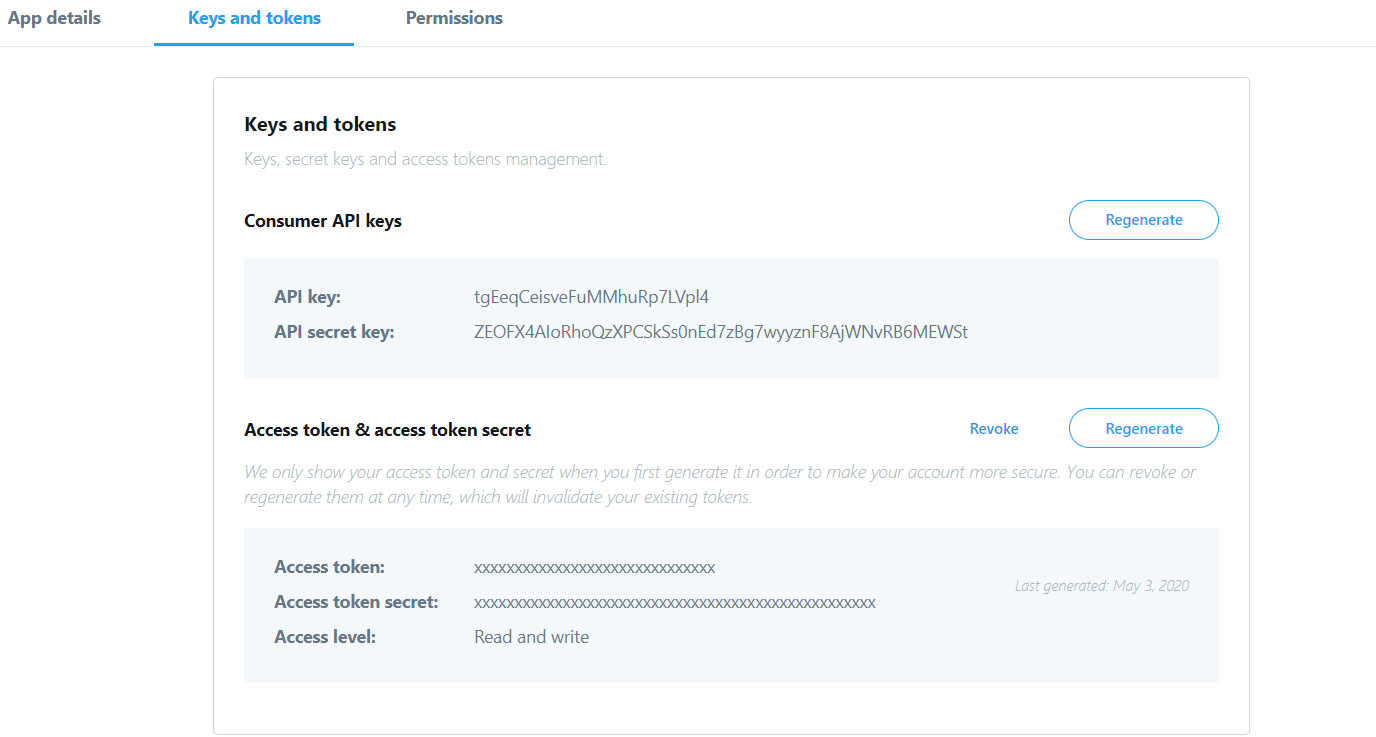


Creating a developer twitter account:

Create an app:



Get the keys and tokens



Add the credentials to Scrapper.py