Web Scraping Instagram with python

AUTOMATING IMAGE EXTRACTION

you can download **top instagram photos** for a **hashtag** using this code.

REQUIREMENTS:

- selenium and wget libraries: `pip install selenium wget`
- builtin os, time, and getpass libraries
- driver for browser you use

Notice:

> *I put three types of code for the three most used browsers so you should change the code relative to the browser you are using.*

In [1]:

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.webdriver.support.wait import WebDriverWait
from getpass import getuser, getpass
import time, os, wget
```

Download driver for browser you use

- just search `download driver for [browser name]` and download it.
- extract them to any where you want.
- add their path here

Notice:

> *you can just add driver to PATH and then you wont need add their path here*
> *so you can open a new window using `driver = webdriver.[browser]()` and not specifying `driver_path` into code.*

you need to login into an account

In [2]:

```
username = getuser()
ig_user = input("Enter username: ")
ig_pass = getpass(prompt="Enter password: ")
hashtag = "#" + input("Enter tag: ").replace("#"," ").strip().replace(" ","_").replace("-","_"

# Brave browser
chromedriver_path = f"/home/{username}/Documents/chromedriver_linux64/chromedriver"
brave_path = "/usr/bin/brave-browser"

# Firefox browser
#firefoxdriver_path = f"/home/{username}/Documents/geckodriver-linux64/geckodriver"

# Chrome browser
#chromedriver_path = f"/home/{username}/Documents/chromedriver_linux64/chromedriver"
```

```
Enter username: mr.azaryazdi
Enter password: ········
Enter tag: cat
```

Open browser window

In [3]:

```
# Brave Browser
option = webdriver.ChromeOptions()
option.binary_location = brave_path
option.add_argument("--incognito")
driver = webdriver.Chrome(executable_path=chromedriver_path, options=option)


# Firefox Browser
#driver = webdriver.Firefox(executable_path=firefoxdriver_path)


# Chrome Browser
#driver = webdriver.Firefox(executable_path=chromedriver_path)
```

Login to Instagram account

1. open instagram webpage.

2. click accept cookies button (comment cookies line if not exist)
3. target the username and password input fields.
4. enter username and password.
5. click login button. </br>

screenshot

cookies screenshot </br>

In [4]:

```
driver.get("http://www.instagram.com/")

cookies = WebDriverWait(driver, 15).until(EC.element_to_be_clickable(
    (By.XPATH, '//button[contains(text(), "Accept All")]'))).click()

username = WebDriverWait(driver, 10).until(EC.element_to_be_clickable(
    (By.CSS_SELECTOR, "input[name='username']")))

password = WebDriverWait(driver, 10).until(EC.element_to_be_clickable(
    (By.CSS_SELECTOR, "input[name='password']")))

username.clear()
username.send_keys(ig_user)
password.clear()
password.send_keys(ig_pass)

try:
    button = WebDriverWait(driver, 2).until(EC.element_to_be_clickable(
        (By.CSS_SELECTOR, "button[type='submit']"))).click()
except:
    button2 = WebDriverWait(driver, 15).until(EC.element_to_be_clickable(
        (By.XPATH, '//div[contains(text(), "Log In")]'))).click()
```

Handle alerts

you might only get a single alert, or you might get 2 of them.

- save your login info?
- turn on notification

you should adjust the code below accordingly
</br></br>

save info screenshot

notification screenshot

In [5]:

```
time.sleep(5)
save_info = WebDriverWait(driver, 15).until(EC.element_to_be_clickable(
    (By.XPATH, '//button[contains(text(), "Not Now")]'))).click()

notifications = WebDriverWait(driver, 15).until(EC.element_to_be_clickable(
    (By.XPATH, '//button[contains(text(), "Not Now")]'))).click()
```

Search for a certain hashtag

1. target the searchbox input field and clear it
2. hit enter

Notice:

> *maybe there will be a problem for submiting the hashtag</br> for fixing that enter the "Enter" twice.*

In [6]:

```
searchbox = WebDriverWait(driver, 10).until(EC.element_to_be_clickable(
    (By.XPATH, "//input[@placeholder='Search']")))
searchbox.clear()
searchbox.send_keys(hashtag)

#FIXING THE DOUBLE ENTER
time.sleep(5) # Wait for 5 seconds to let everything load correctly
my_link = WebDriverWait(driver, 10).until(EC.element_to_be_clickable(
    (By.XPATH, "//a[contains(@href, '/" + hashtag[1:] + "/')]"))).click()
```

Scroll Down¶

Increase `n_scrolls` to select more photos (depending on screen resolution)

Example:

- 2 scrolls cover approx. 35 photos

- 3 scrolls cover approx. 45 photos

In [7]:

```
n_scrolls = 2
for j in range(0, n_scrolls):
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
    time.sleep(5)
```

target all links elements on the page

In [8]:

```
anchors = driver.find_elements_by_tag_name('a')
anchors = [a.get_attribute('href') for a in anchors]
anchors = [a for a in anchors if str(a).startswith("https://www.instagram.com/p/")]

print('Found ' + str(len(anchors)) + ' links to images')
```

```
Found 51 links to images
```

In [9]:

```
print(anchors[:5])
```

```
['https://www.instagram.com/p/CSRB5vQqODP/', 'https://www.instagram.com/p/CSRGq6vqbbD/', 'https://www.instagram.com/p/CSRYlMmJxt0/', 'https://www.instagram.com/p/CSRJfTLqktw/', 'https://www.instagram.com/p/CSROMqHqocd/']
```

Convert links of the posts to their direct links of the images

In [10]:

```
images = []

#follow each image link and extract only image at index=1
for a in anchors:
    driver.get(a)
    time.sleep(5)
    img = driver.find_elements_by_tag_name('img')
    img = [i.get_attribute('src') for i in img]
    images.append(img[1])
```

In [11]:

```
print(images[:5])
```

```
['https://scontent-mxp1-1.cdninstagra
m.com/v/t51.2885-15/e35/p1080x1080/23
3664782_106435858338396_6541243200928
988170_n.jpg?_nc_ht=scontent-mxp1-1.c
dninstagram.com&_nc_cat=100&_nc_ohc=4
D0RTAyInIMAX9Zwc44&tn=_n0mBMqImc0PqVK
d&edm=AABBvjUBAAAA&ccb=7-4&oh=84c533e
4e17254e6d829279415d3ca0c&oe=6115E6A8
&_nc_sid=83d603', 'https://scontent-m
xp1-1.cdninstagram.com/v/t51.2885-15/
e35/s1080x1080/233664001_118328223771
911_793040319641020613_n.jpg?_nc_ht=s
content-mxp1-1.cdninstagram.com&_nc_c
at=111&_nc_ohc=5vqwR2xbabAAX_dYutH&ed
m=AABBvjUBAAAA&ccb=7-4&oh=2601e9b9104
d3c4c58fc7315319b45f7&oe=6115890B&_nc
_sid=83d603', 'https://scontent-mxp1-
1.cdninstagram.com/v/t51.2885-19/s150
x150/222791783_3992569747522537_15812
01377043246540_n.jpg?_nc_ht=scontent-
mxp1-1.cdninstagram.com&_nc_ohc=Q2pln
FZfl6kAX-fJ614&edm=AABBvjUBAAAA&ccb=7
-4&oh=31380b5e8240378ccb8a1aa5e05c7a0
0&oe=6116459F&_nc_sid=83d603', 'http
s://scontent-mxp1-1.cdninstagram.com/
v/t51.2885-15/e35/s1080x1080/23403618
6_352228559639175_7545262275618375141
_n.jpg?_nc_ht=scontent-mxp1-1.cdninst
agram.com&_nc_cat=103&_nc_ohc=dVqN-4J
FK0kAX8G_mA4&edm=AABBvjUBAAAA&ccb=7-4
&oh=0094ca133fe68e54dcb2de6e5338b533&
oe=611519B0&_nc_sid=83d603', 'http
s://scontent-mxp1-1.cdninstagram.com/
v/t51.2885-15/e35/p1080x1080/23413934
7_574397313593858_1279168334984311620
_n.jpg?_nc_ht=scontent-mxp1-1.cdninst
agram.com&_nc_cat=102&_nc_ohc=NSOVNOh
IyKwAX-LyaM9&edm=AABBvjUBAAAA&ccb=7-4
&oh=9abbc663c32d2263b85b5339241b11cf&
oe=6115C23F&_nc_sid=83d603']
```

Save images to computer

create a new folder for our images somewhere on our computer. Then, download and save all the images there.

In [12]:

```
path = os.getcwd()
path = os.path.join(path, hashtag[1:])
os.mkdir(path)
```

In [13]:

```
print(path)
```

```
/home/ali/MyJupyterNoteBooks/cat
```

In [15]:

```
#download images
counter = 1
for image in images:
    save_as = os.path.join(path, hashtag[1:] + str(counter) + '.jpg')
    wget.download(image, save_as)
    counter += 1
```

```
100%
[.....................................
 272234 / 272234
```

Done!

By Momento