Webscraping with Python BeautifulSoup4

Date updated: 12th May, 2020

Written by: Ruhaila Maskat (PhD)

#Disclaimer: This code serves only as a teaching material for ITS480/ISP610 Business Data Analytics @ FSKM, UiTM Shah Alam on the topic of Webscraping.

Note: Do not copy and paste these codes. Type them. Characters may differ across platforms.

1. Install latest Python.
2. Open Windows command line.
3. To install BeautifulSoup, go to C:\[your Python path]. Type,
   ```
   C:\[your Python path]\Scripts>pip install beautifulsoup4
   ```
   If it is already installed, you will get,
   ```
   Requirement already satisfied: beautifulsoup4 in C:\[your Python
   path]\lib\site-packages
   ```
   else, installation will take place.
4. To know if BeautifulSoup is successfully installed, you must invoke Python first,
   ```
   C:\[your Python path]\Scripts>python
   ```
   You may get,
   ```
   Python 3.8.3rc1 (tags/v3.8.3rc1:802eb67, Apr 29 2020, 21:39:14) [MSC
   v.1924 64 b
   it (AMD64)] on win32
   Type "help", "copyright", "credits" or "license" for more
   information.
   ```
   Then type,
   ```
   >>> import bs4
   ```
   If you get the following empty Python cursor, this means BeautifulSoup4 is available to be used and has been imported into your Python environment.
   ```
   >>>
   ```
5. At the header, to grab a html page, type,
   ```
   >>> from urllib.request import urlopen as uReq
   ```
6. Next, to parse html tags, call BeautifulSoup by typing,
   ```
   >>> from bs4 import BeautifulSoup as soup
   ```
7. Now, we need to define the html page's url, type
   ```
   >>> my_url = 'https://www.lelong.com.my'
   ```
8. To check contents of my_url variable, type
   ```
   >>> my_url
   ```
   You should get,
   ```
   'https://www.lelong.com.my'
   ```
9. To open a connection to the web page and downloading into your machine, type,
   ```
   >>> uClient = uReq(my_url)
   ```
10. To read the scraped contents, type,
    ```
    >>> page_html = uClient.read()
    ```
    Note: Warning, it is not advisable to view the contents at this point in time, because if the web page is huge, the command prompt will crash.
11. Make sure that you close the connection.
    Note: It is unethical to leave a connection open.
    ```
    >>> uClient.close()
    ```

12. To parse the contents, type,

```
>>> page_soup = soup(page_html,"html.parser")
```

13. To view the header of the contents, type,

```
>>> page_soup.h1
```

You should get,

```
<h1>Lelong.my - Largest online shopping marketplace in Malaysia</h1>
```

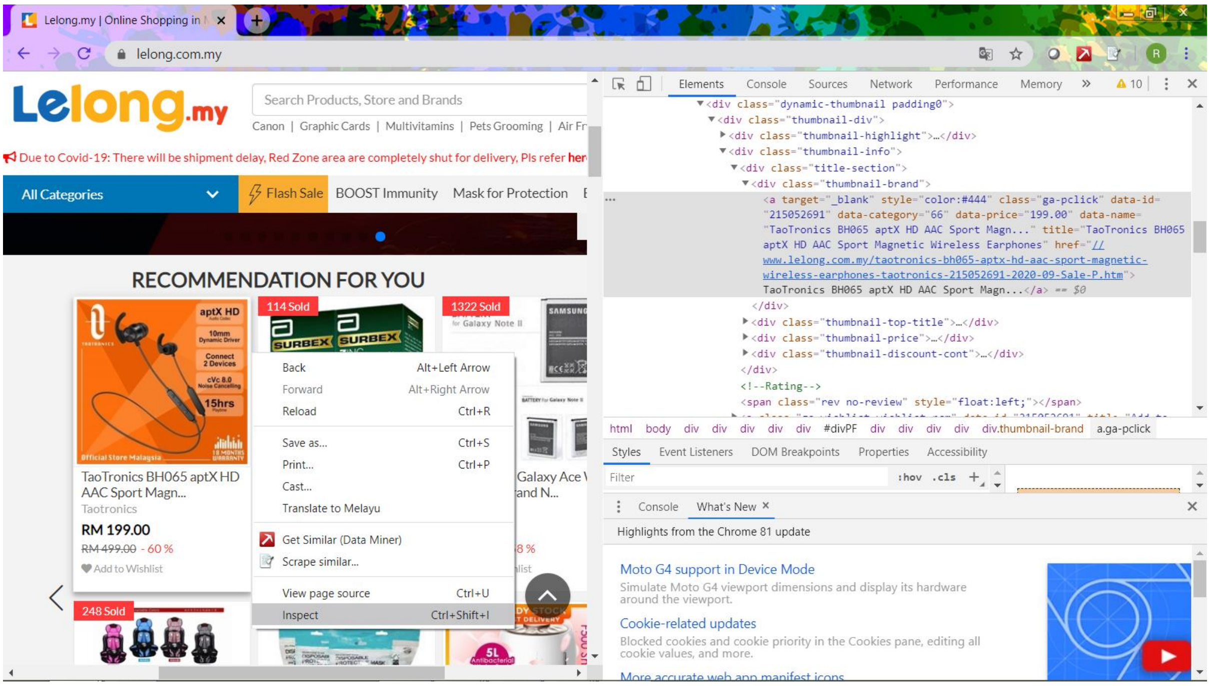
14. To view any paragraph in the contents, type,

```
>>> page_soup.p
```

You should get,

```
<p id="responseresult">
                    Loading ...
                </p>
```

15. Focus cursor on a particular part of the webpage of interest, highlight the text and right mouse click to choose Inspect.



```
▼<div class="dynamic-thumbnail padding0">
  ▼<div class="thumbnail-div">
    ▶<div class="thumbnail-highlight">…</div>
    ▼<div class="thumbnail-info">
      ▼<div class="title-section">
        ▼<div class="thumbnail-brand">
            <a target="_blank" style="color:#444" class="ga-pclick" data-id=
            "215052691" data-category="66" data-price="199.00" data-name=
            "TaoTronics BH065 aptX HD AAC Sport Magn..." title="TaoTronics BH065
            aptX HD AAC Sport Magnetic Wireless Earphones" href="//
            www.lelong.com.my/taotronics-bh065-aptx-hd-aac-sport-magnetic-
            wireless-earphones-taotronics-215052691-2020-09-Sale-P.htm">
            TaoTronics BH065 aptX HD AAC Sport Magn...</a> == $0
          </div>
        ▶<div class="thumbnail-top-title">…</div>
        ▶<div class="thumbnail-price">…</div>
        ▶<div class="thumbnail-discount-cont">…</div>
        </div>
        <!--Rating-->
        <span class="rev no-review" style="float:left;"></span>
```

Lelong.my | Online Shopping in
lelong.com.my
Lelong.my
Search Products, Store and Brands
Canon | Graphic Cards | Multivitamins | Pets Grooming | Air Fr
Due to Covid-19: There will be shipment delay, Red Zone area are completely shut for delivery, Pls refer her
All Categories
Flash Sale
BOOST Immunity
Mask for Protection
RECOMMENDATION FOR YOU
114 Sold
1322 Sold
TaoTronics BH065 aptX HD AAC Sport Magn...
Taotronics
RM 199.00
RM 499.00 - 60 %
Add to Wishlist
248 Sold
Galaxy Ace \ and N...
Back Alt+Left Arrow
Forward Alt+Right Arrow
Reload Ctrl+R
Save as... Ctrl+S
Print... Ctrl+P
Cast...
Translate to Melayu
Get Similar (Data Miner)
Scrape similar...
View page source Ctrl+U
Inspect Ctrl+Shift+I
Elements
Console
Sources
Network
Performance
Memory
10
html body div div div div div #divPF div div div div div.thumbnail-brand a.ga-pclick
Styles
Event Listeners
DOM Breakpoints
Properties
Accessibility
Filter
:hov .cls +
Console
What's New
Highlights from the Chrome 81 update
Moto G4 support in Device Mode
Simulate Moto G4 viewport dimensions and display its hardware around the viewport.
Cookie-related updates
Blocked cookies and cookie priority in the Cookies pane, editing all cookie values, and more.
More accurate web app manifest icons

16. Identify where the data you want is located. In this case, we want the brand name and price. Thus, it is stored in a html class called **ga-pclick** which holds element **data-name** and **data-price**.

17. To grab these data, type

```
>>> brandcontainers = page_soup.findAll("a",{"class":"ga-pclick"})
```

18. To save these data into a .csv file,

```
>>> filename = "lelongproducts_2020.csv"
>>> f = open(filename, "w")
>>> headers = "brand, price\n"
>>> f.write(headers)
```

19. To parse the text data and write into the file, you need a for loop.

Note: Make sure you observe Python's indentation rules.

```
>>> for i in range(0, len(brandcontainers)):
      brand = brandcontainers[i]["title"]
      price = brandcontainers[i]["data-price"]
      f.write(brand.replace(",","|")+","+price.replace(",",".")+"\n")
```

20. Close the file stream,

```
>>> f.close()
```

```
from urllib.request import urlopen as uReq
from bs4 import BeautifulSoup as soup

my_url = 'https://www.lelong.com.my/'

uClient = uReq(my_url)
page_html = uClient.read()
uClient.close()

page_soup = soup(page_html,"html.parser")

filename = "lelongproducts_2020.csv"
f = open(filename, "w")
headers = "brand, price\n"
f.write(headers)

brandcontainers = page_soup.findAll("a",{"class":"ga-pclick"})

for i in range(0, len(brandcontainers)):
      brand = brandcontainers[i]["title"]
      price = brandcontainers[i]["data-price"]
      f.write(brand.replace(",","|")+","+price.replace(",",".")+"\n")

f.close()
```