Web Scraping

What is our GOAL for this MODULE?

The goal of this module is to learn scraping to get the useful data from different sites.

What did we ACHIEVE in the class TODAY?

- We scraped the stars data from Nasa's site

Which CONCEPTS/CODING BLOCKS did we cover today?

- Usage of selenium
- Usage of Beautiful soup
- Getting data from Html of a page

How did we DO the activities?

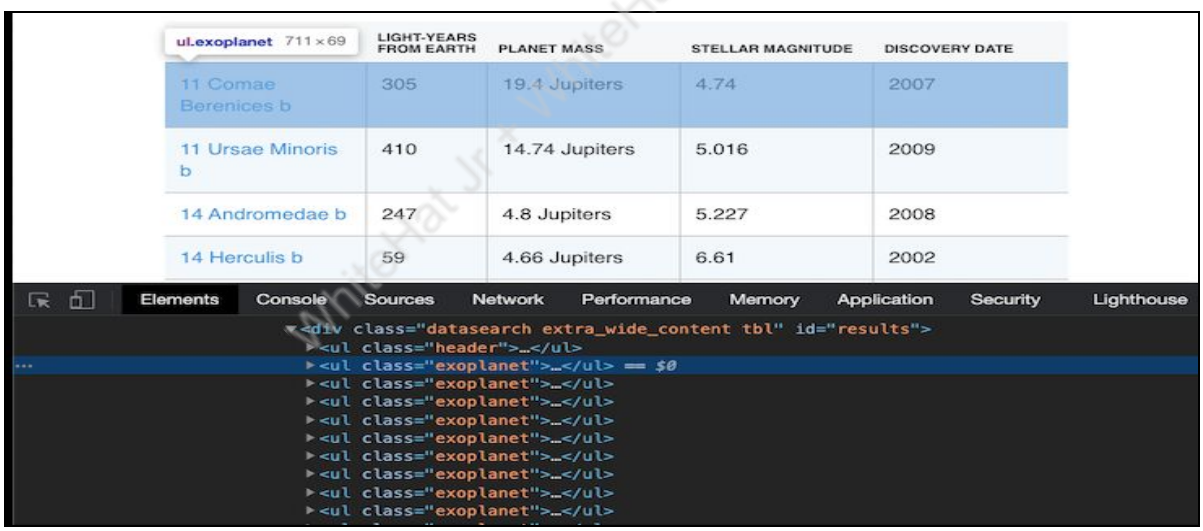
1. We created a virtual environment.
2. We installed the necessary libraries.
3. We imported the selenium and Beautiful soup in our code.

```
from selenium import webdriver
from bs4 import BeautifulSoup
import time
import csv
```

4. We also installed the chrome driver.
5. Then we opened the link we want to scrape using selenium.

```
START_URL = "https://exoplanets.nasa.gov/exoplanet-catalog/"
browser = webdriver.Chrome("/path/to/chromedriver")
browser.get(START_URL)
time.sleep(10)
```

6. We got all the tags from the table to get the data from it.



7. We found all the ul tags to get the data.

```
soup = BeautifulSoup(browser.page_source, "html.parser")
    for ul_tag in soup.find_all("ul", attrs={"class", "exoplanet"}):
```

8. Then we found all the li tags.

```
li_tags = ul_tag.find_all("li")
```

9. We found that all the data is inside the anchor tags.

```
temp_list = []
        for index, li_tag in enumerate(li_tags):
            if index == 0:
                temp_list.append(li_tag.find_all("a")[0].contents[0])
            else:
                try:
                    temp_list.append(li_tag.contents[0])
                except:
                    temp_list.append("")

planet_data.append(temp_list)
```

10. We code to loop the code 428 times.

```python
scrapper_2.py
from selenium import webdriver
from bs4 import BeautifulSoup
import time
import csv

START_URL = "https://exoplanets.nasa.gov/exoplanet-catalog/"
browser = webdriver.Chrome("/Users/apoorvelous/Downloads/chromedriver")
browser.get(START_URL)
time.sleep(10)

def scrape():
    headers = ["name", "light_years_from_earth", "planet_mass", "stellar_magnitude", "discovery_date"]
    planet_data = []
    for i in range(0, 428):
        soup = BeautifulSoup(browser.page_source, "html.parser")
        for ul_tag in soup.find_all("ul", attrs={"class", "exoplanet"}):
            li_tags = ul_tag.find_all("li")
            temp_list = []
            for index, li_tag in enumerate(li_tags):
                if index == 0:
                    temp_list.append(li_tag.find_all("a")[0].contents[0])
                else:
                    try:
                        temp_list.append(li_tag.contents[0])
                    except:
                        temp_list.append("")
            planet_data.append(temp_list)
        browser.find_element_by_xpath('//*[@id="primary_column"]/footer/div/div/div/nav/span[2]/a').click()
    with open("scrapper_2.csv", "w") as f:
        csvwriter = csv.writer(f)
        csvwriter.writerow(headers)
        csvwriter.writerows(planet_data)

scrape()
```

What's NEXT?

In the next class, we will explore more of scraping and data cleaning.

EXTEND YOUR KNOWLEDGE:

You can read the following blog on scraping with selenium to understand more:
https://medium.com/ymedialabs-innovation/web-scraping-using-beautiful-soup-and-selenium-for-dynamic-page-2f8ad15efe25