# Web Scraping-2


## What is our GOAL for this MODULE?

The goal of this module is to scrape more data from the exoplanets link.

## 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?

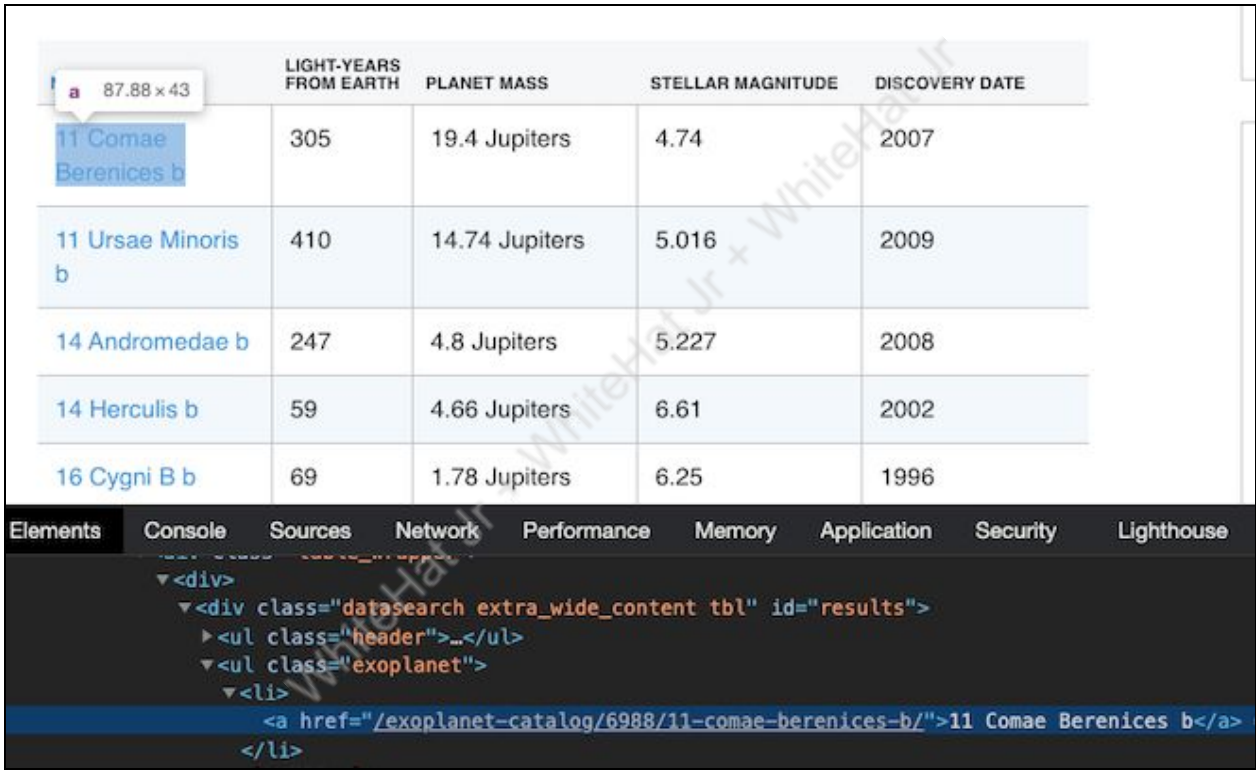
- Usage of selenium
- Usage of Beautiful soup
- Getting data from Html of a page

## 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.
4. We added a new hyperlink to the header.

```
headers = ["name", "light_years_from_earth", "planet_mass", "stellar_magnitude", "discovery_date", "hyperlink"]
```

5. We checked the hyperlinks.



6. We saw that the https://exoplanets.nasa.gov was missing so we coded to add it.

scrapper_2.py

```python
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", "hyperlink"]
    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("")
            hyperlink_li_tag = li_tags[0]
            temp_list.append("https://exoplanets.nasa.gov"+hyperlink_li_tag.find_all("a", href=True)[0]["href"])
            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()
```

7. We coded to get the data from the hyperlink.

scrapper_2.py

```python
from selenium import webdriver
from bs4 import BeautifulSoup
import requests
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)
headers = ["name", "light_years_from_earth", "planet_mass", "stellar_magnitude", "discovery_date", "hyperlink", "planet_type", "planet_r
planet_data = []
new_planet_data = []

def scrape():
    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("")
            hyperlink_li_tag = li_tags[0]
            temp_list.append("https://exoplanets.nasa.gov"+hyperlink_li_tag.find_all("a", href=True)[0]["href"])
            planet_data.append(temp_list)
        browser.find_element_by_xpath('//*[@id="primary_column"]/footer/div/div/div/nav/span[2]/a').click()
```

```python
def scrape_more_data(hyperlink):
    page = requests.get(hyperlink)
    soup = BeautifulSoup(page.content, "html.parser")
    for tr_tag in soup.find_all("tr", attrs={"class": "fact_row"}):
        td_tags = tr_tag.find_all("td")
        temp_list = []
        for td_tag in td_tags:
            try:
                temp_list.append(td_tag.find_all("div", attrs={"class": "value"})[0].contents[0])
            except:
                temp_list.append("")
        new_planet_data.append(temp_list)

scrape()
for data in planet_data:
    scrape_more_data(data[5])

final_planet_data = []

for index, data in enumerate(planet_data):
    final_planet_data.append(data + final_planet_data[index])

with open("final.csv", "w") as f:
        csvwriter = csv.writer(f)
        csvwriter.writerow(headers)
        csvwriter.writerows(final_planet_data)
```

## 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