Data Analyst Bootcamp

Web Scraping in Python

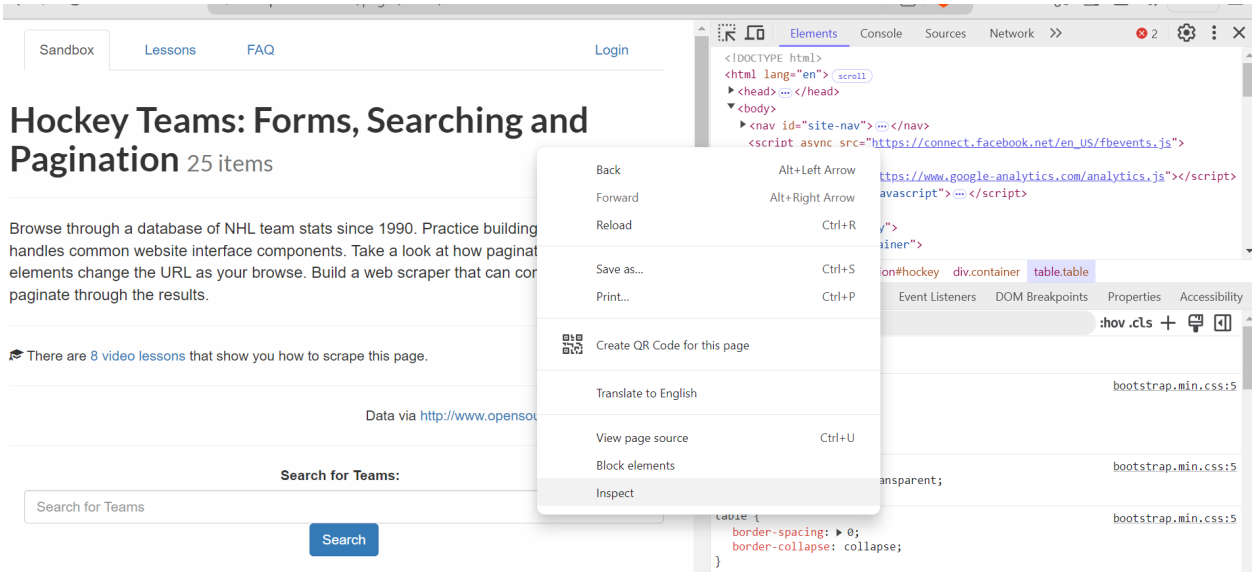
Inspecting Web Pages with HTML

```
<html>

<head>
    <title>My First Web Page</title>
</head>

<body>
    <h1>My First Web Page</h1>
    <p><b>Hello World Wide Web!</b></p>
    <p><i>Hello World Wide Web!</i></p>
    <p><u>Hello World Wide Web!</u></p>
    <p>This is my first web page.</p>
    <p>HTML tags can give <b><i>various</i></b>
    <u>looks and format</u> to the content of this web page.</p>
</body>

</html>
```



inspect and then click on arrow(left of elements in inspect) and then inspect whatever u like

BeautifulSoup + Requests

```
from bs4 import BeautifulSoup
import requests

url ='https://www.scrapethissite.com/pages/forms/'

page=requests.get(url)

#204 400 401 404 bad req 204 no content 404 error
soup = BeautifulSoup(page.text,'html')

print(soup)

print(soup.prettify)
```

find and find all

```
soup.find('div') #finds only first response

soup.find_all('div') #finds every div response

soup.find_all('p', class_ = 'lead')

soup.find_all('p', class_ = 'lead').text #error cause find all

soup.find('p', class_ = 'lead').text.strip()

soup.find_all('th')

soup.find('th').text.strip()
```

Scraping data from real website + Pandas

```
from bs4 import BeautifulSoup
import requests
```

```
url='https://en.wikipedia.org/wiki/List_of_largest_companies_in_
page=requests.get(url)
soup=BeautifulSoup(page.text,'html')

print(soup)

soup.find_all('table')[0]

#soup.find('table',class_='wikitable sortable') we can use this

table=soup.find_all('table')[0]

print(table)

world_titles=table.find_all('th')

world_titles

world_table_titles=[title.text.strip() for title in world_titles
print(world_table_titles)

import pandas as pd

df=pd.DataFrame(columns=world_table_titles)
df

column_data=table.find_all('tr')

for row in column_data[1:]:
    row_data=row.find_all('td')
    individual_row_data=[data.text.strip() for data in row_data]
    #print(individual_row_data)
    length=len(df)
    df.loc[length]=individual_row_data

df
```

```
df.to_csv(r'C:\Users\User\Documents\FileSorter\companies.csv',in
```