# 30 Days Of Python: Day 22 - Web Scraping


## Python Web Scraping

### What is Web Scrapping

The internet is full of huge amount of data which can be used for different purposes. To collect this data we need to know how to scrape data from a website.

Web scraping is the process of extracting and collecting data from websites and storing it on a local machine or in a database.

In this section, we will use beautifulsoup and requests package to scrape data. The package version we are using is beautifulsoup 4.

To start scraping websites you need *requests*, *beautifoulSoup4* and a *website*.

```
pip install requests
pip install beautifulsoup4
```

To scrape data from websites, basic understanding of HTML tags and CSS selectors is needed. We target content from a website using HTML tags, classes or/and ids. Let us import the requests and BeautifulSoup module

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

Let us declare url variable for the website which we are going to scrape.

```
import requests
from bs4 import BeautifulSoup
url = 'https://archive.ics.uci.edu/ml/datasets.php'

# Lets use the requests get method to fetch the data from url

response = requests.get(url)
# lets check the status
status = response.status_code
print(status) # 200 means the fetching was successful
```

```
200
```

Using beautifulSoup to parse content from the page

```
import requests
from bs4 import BeautifulSoup
url = 'https://archive.ics.uci.edu/ml/datasets.php'

response = requests.get(url)
content = response.content # we get all the content from the website
soup = BeautifulSoup(content, 'html.parser') # beautiful soup will give a
chance to parse
print(soup.title) # <title>UCI Machine Learning Repository: Data
Sets</title>
print(soup.title.get_text()) # UCI Machine Learning Repository: Data Sets
print(soup.body) # gives the whole page on the website
print(response.status_code)

tables = soup.find_all('table', {'cellpadding':'3'})
# We are targeting the table with cellpadding attribute with the value of 3
# We can select using id, class or HTML tag , for more information check
the beautifulsoup doc
table = tables[0] # the result is a list, we are taking out data from it
for td in table.find('tr').find_all('td'):
    print(td.text)
```

If you run this code, you can see that the extraction is half done. You can continue doing it because it is part of exercise 1. For reference check the [beautifulsoup documentation]

🍪 You are so special, you are progressing everyday. You are left with only eight days to your way to greatness. Now do some exercises for your brain and muscles.

# 💻 Exercises: Day 22

1. Scrape the following website and store the data as json file(url = 'http://www.bu.edu/president/boston-university-facts-stats/').
2. Extract the table in this url (https://archive.ics.uci.edu/ml/datasets.php) and change it to a json file

3. Scrape the presidents table and store the data as json(https://en.wikipedia.org/wiki/List_of_presidents_of_the_United_States). The table is not very structured and the scrapping may take very long time.

🎉 CONGRATULATIONS ! 🎉