Data Science | 30 Days of Machine Learning | Day - 5

Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

#Data Gathering (Previous Classes)

1. **Working with CSV Files – Day 3**
2. **Working with JSON/SQL – Day 3**
3. **Fetching data from an API – Day 4**

#Framing a Machine Learning Problem – Day 4

---------- Today Topics | Day 5-------

#Data Gathering

4. **Fetching data using web scraping**

----Upcoming Topics----

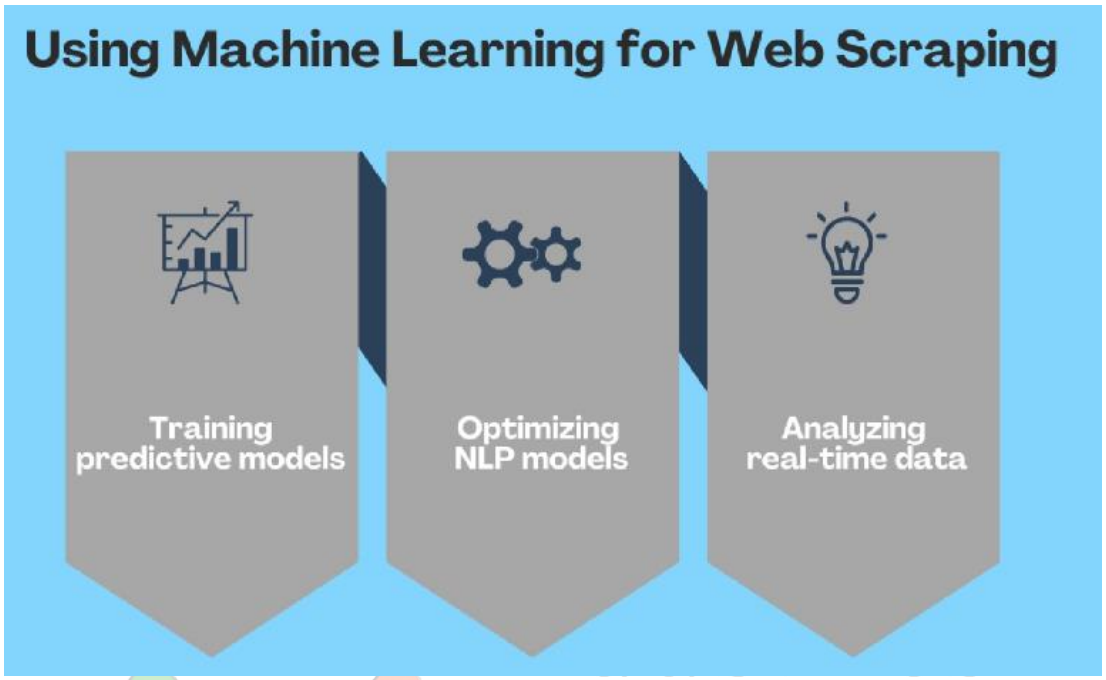
-EDA & Feature Engineering

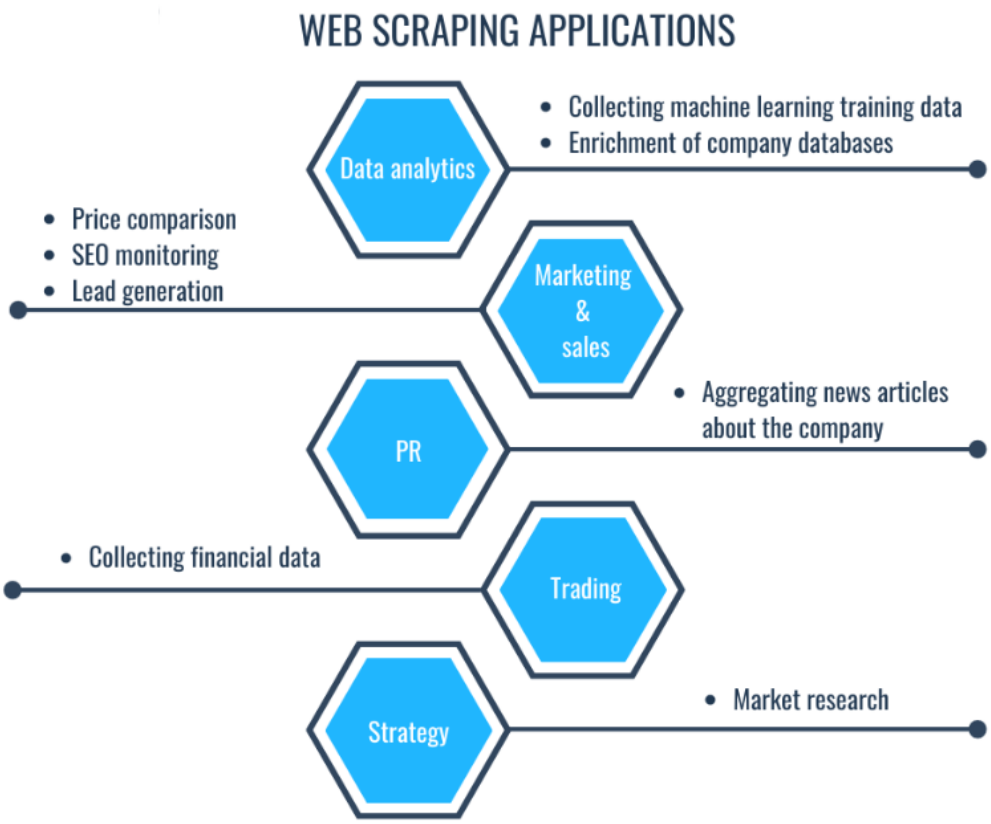
 # How We Understand the Data?

 # What is Web Scraping: It's is a term used for automatically retrieving data from the internet and structuring it in a useful manner. The standard scraping algorithm makes use of static paths to navigate the program through the HTML to the sought-after data.

Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications.

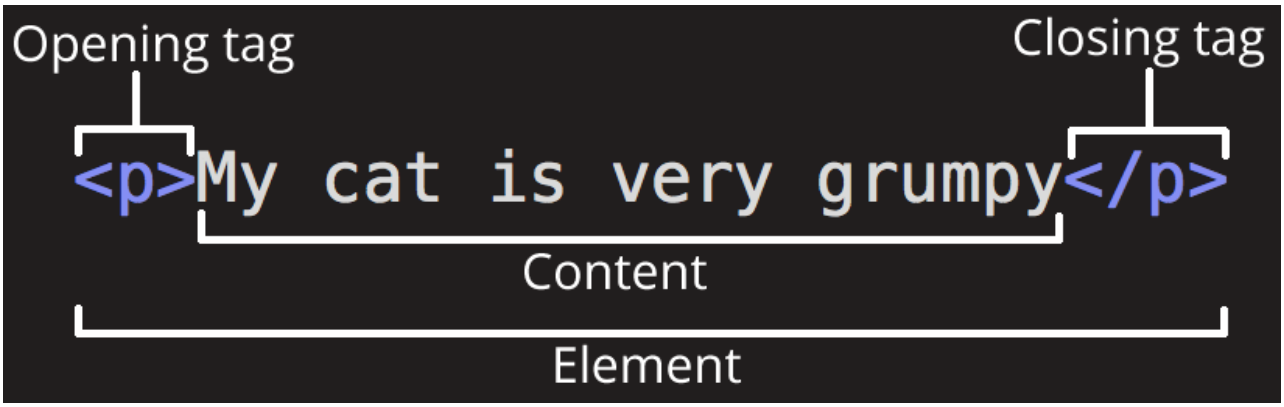
Read Blog: https://research.aimultiple.com/machine-learning-web-scraping/

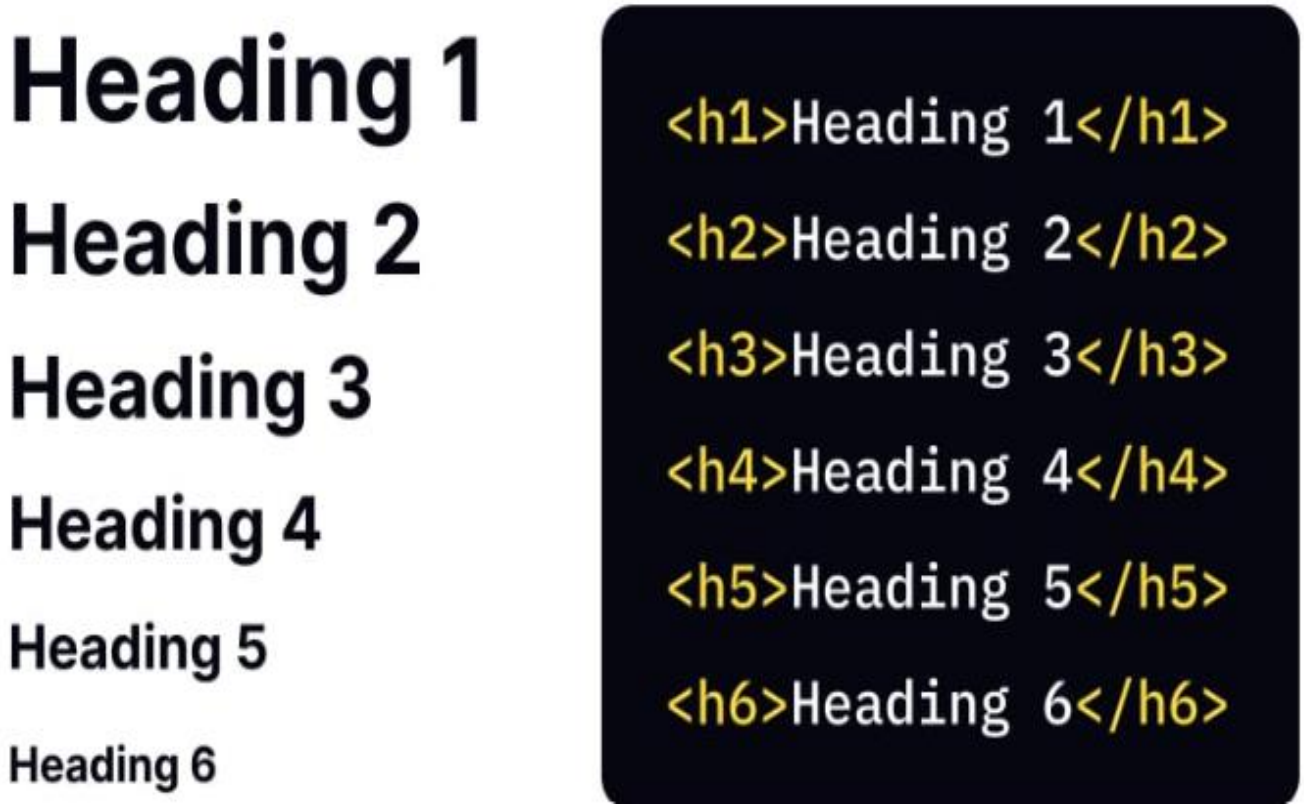




HTML Hierarchy

```
<!DOCTYPE html>   ← Tells the document type
<html>            ← The Root Element
    <head>        ← Contains the header information
        <title>Title of the Page</title>   ← Defines Title of the Page
    </head>
    <body>        ← Holds the Content of the Page
    Tags related to layout and formatting
    </body>
</html>
```


Opening tag
Closing tag
My cat is very grumpy
Content
Element


Heading 1
Heading 2
Heading 3
Heading 4
Heading 5
Heading 6
<h1>Heading 1</h1>
<h2>Heading 2</h2>
<h3>Heading 3</h3>
<h4>Heading 4</h4>
<h5>Heading 5</h5>
<h6>Heading 6</h6>

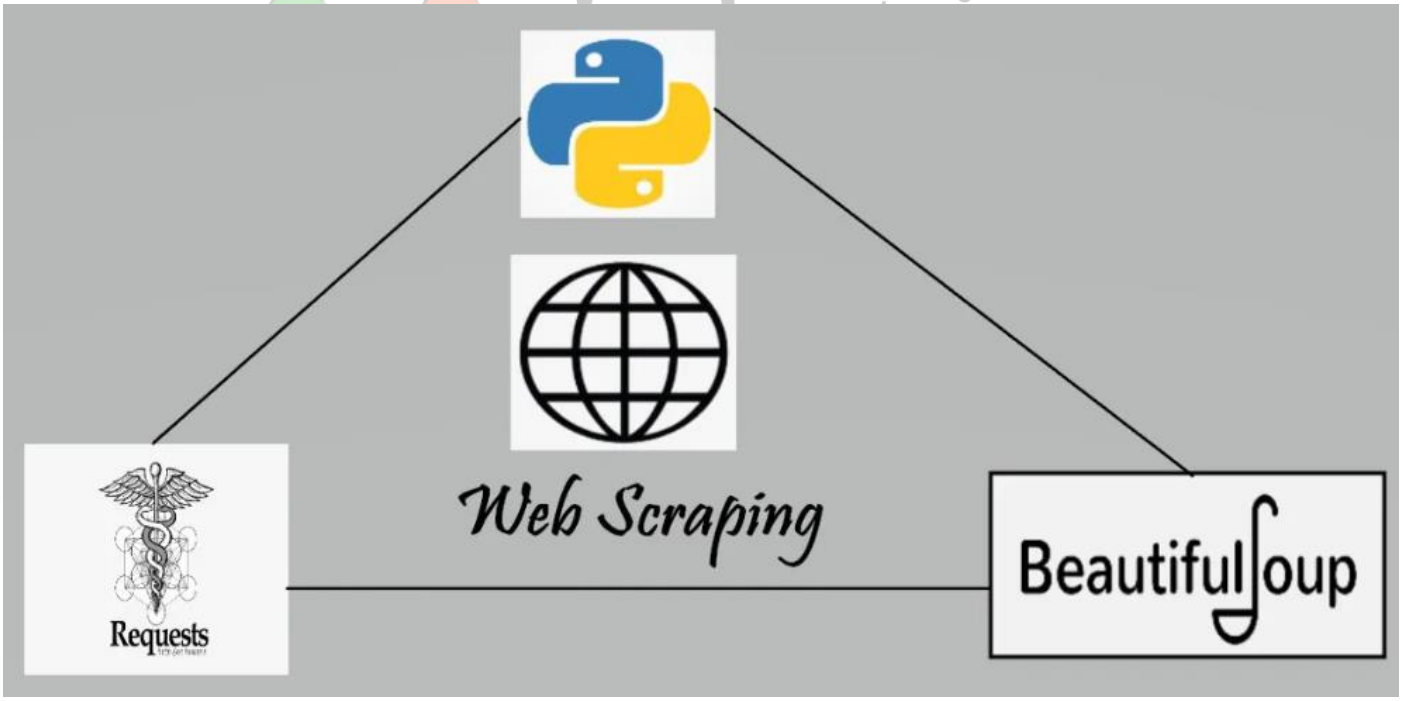
```
<body>

  <div id="content">
    <h1>Heading here</h1>
    <p>Lorem ipsum dolor sit amet.</p>
    <p>Lorem ipsum dolor <em>sit</em> amet.</p>
    <hr>
  </div>

  <div id="nav">
    <ul>
      <li>item 1</li>
      <li>item 2</li>
      <li>item 3</li>
    </ul>
  </div>

</body>
```

```
<div class="container-fluid">
  <h3 class="text-primary text-center">jQuery Playground</h3>
  <div class="row">
    <div class="col-xs-6">
      <div class="well">
        <button class="btn btn-info">Info</button>
        <button class="btn btn-primary">Like</button>
        <button class="btn btn-danger">Delete</button>
      </div>
    </div>
    <div class="col-xs-6">
      <div class="well">
      <div class="btn btn-info">More</div>
      <div class="btn btn-primary">Click</div>
      <div class="btn btn-danger">Report</div>
      </div>
    </div>
  </div>
</div>
```



<Open Jupiter Notebook>

Welcome to 30 Days Machine Learning

Import Libraries

```
import pandas as pd
import requests
from bs4 import BeautifulSoup
```

We use link for web scraping: https://www.ambitionbox.com/list-of-companies?page=1

Import Webpage

```
headers={'User-Agent':'Mozilla/5.0 (Windows NT 6.3; Win 64 ; x64) Apple WeKit
/537.36(KHTML , like Gecko) Chrome/80.0.3987.162 Safari/537.36'}

webpage=requests.get('https://www.ambitionbox.com/list-of-
companies?page=1',headers=headers).text
```

Define “Soup” as BeautifulSoup Library

```
soup=BeautifulSoup(webpage,'lxml')
```

Use Print

```
#print(soup.prettify())
```

Extract the heading line

```
for n in soup.find_all('h1'):
    print(n.text.strip())
```

Extract the Company Name

```
for i in soup.find_all('h2'):
  print(i.text.strip())
```

Extract the rating value

```
for k in soup.find_all('span',class_='companyCardWrapper__companyRatingValue'):
    print(k.text.strip())
```

Extract the all card details

```
for j in soup.find_all('span',class_='companyCardWrapper__interLinking'):
    print(j.text.strip())
```

<End Code>

Upcoming Class- Day 6

- **Exploratory Data Analysis (EDA)**

- **EDA Univariate Analysis**
- **EDA Bivariate Analysis**
- **EDA Multivariate Analysis**

- **Pandas Profiler**
- **Asking Basic Question**

Data Story Telling : Curious Data Minds

ASIAN PAINTS: INDIA'S BIGGEST DATA SCIENCE COMPANY THAT SELLS PAINT

Read Blogs:
https://startuptalky.com/asian-paints-case-study/

https://d3.harvard.edu/platform-digit/submission/asian-paints-indias-biggest-data-science-company-that-sells-paint/

https://yourstory.com/2023/08/asian-paints-data-science-revolution

https://tradebrains.in/asian-paints-case-study/