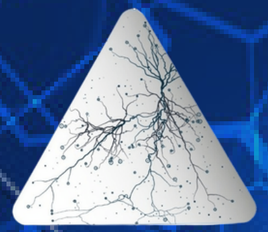
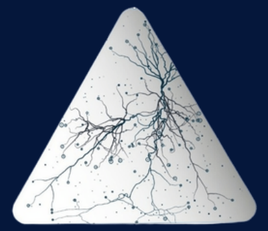


ensias_ai_club

DATA CELL

DATA COLLECTION

WEB SCRAPING



ensias_ai_club

PLAN

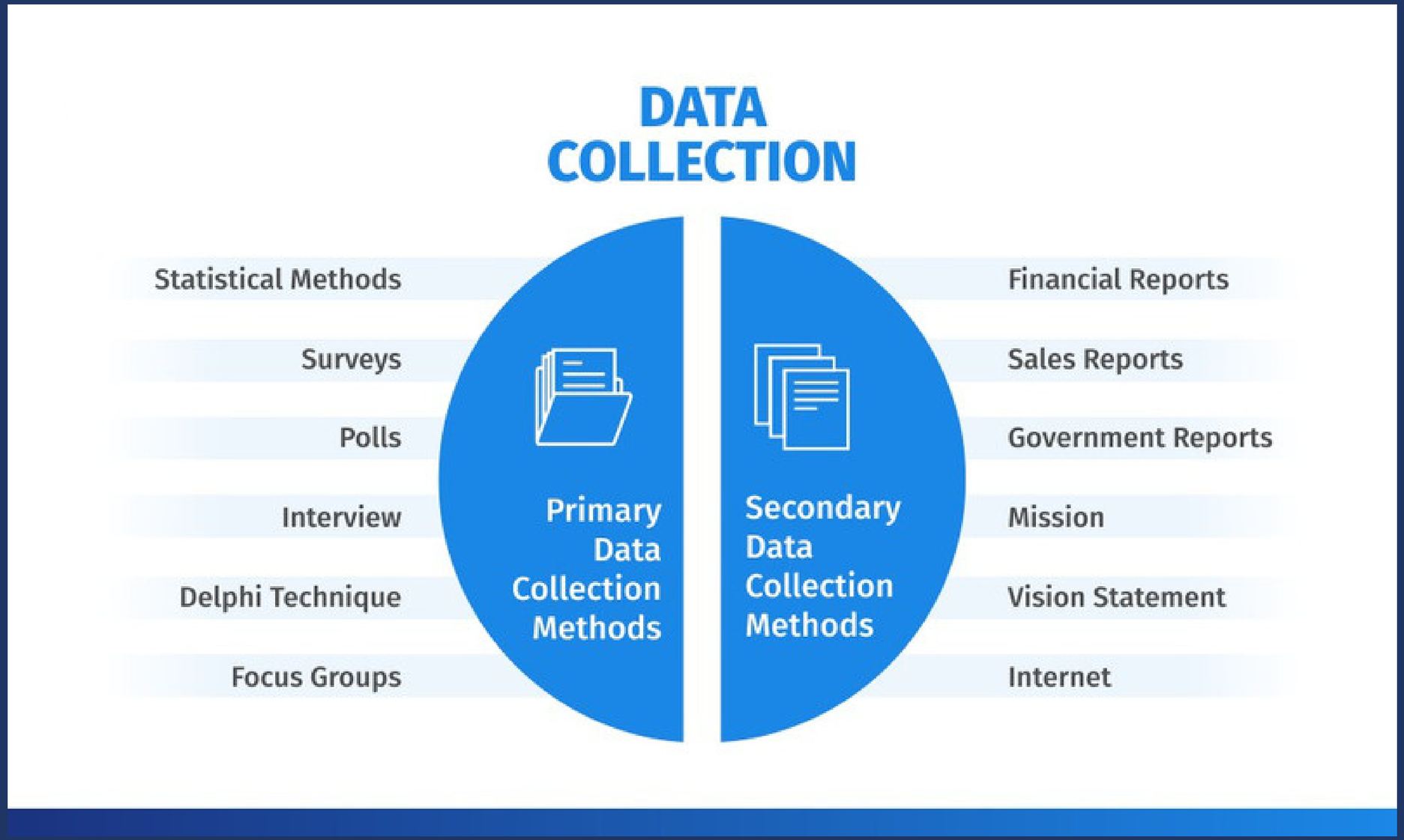
WHAT IS DATA COLLECTION ?

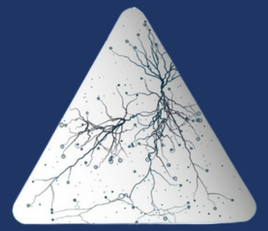
DATA SOURCES

WEB SCRAPING

ensias_ai_club

DATA COLLECTION





ensias_ai_club

DATA SOURCES

Government and political Data

- Data.gov
- The Census Bureau

Social Data

- Google Trends

Economic Data

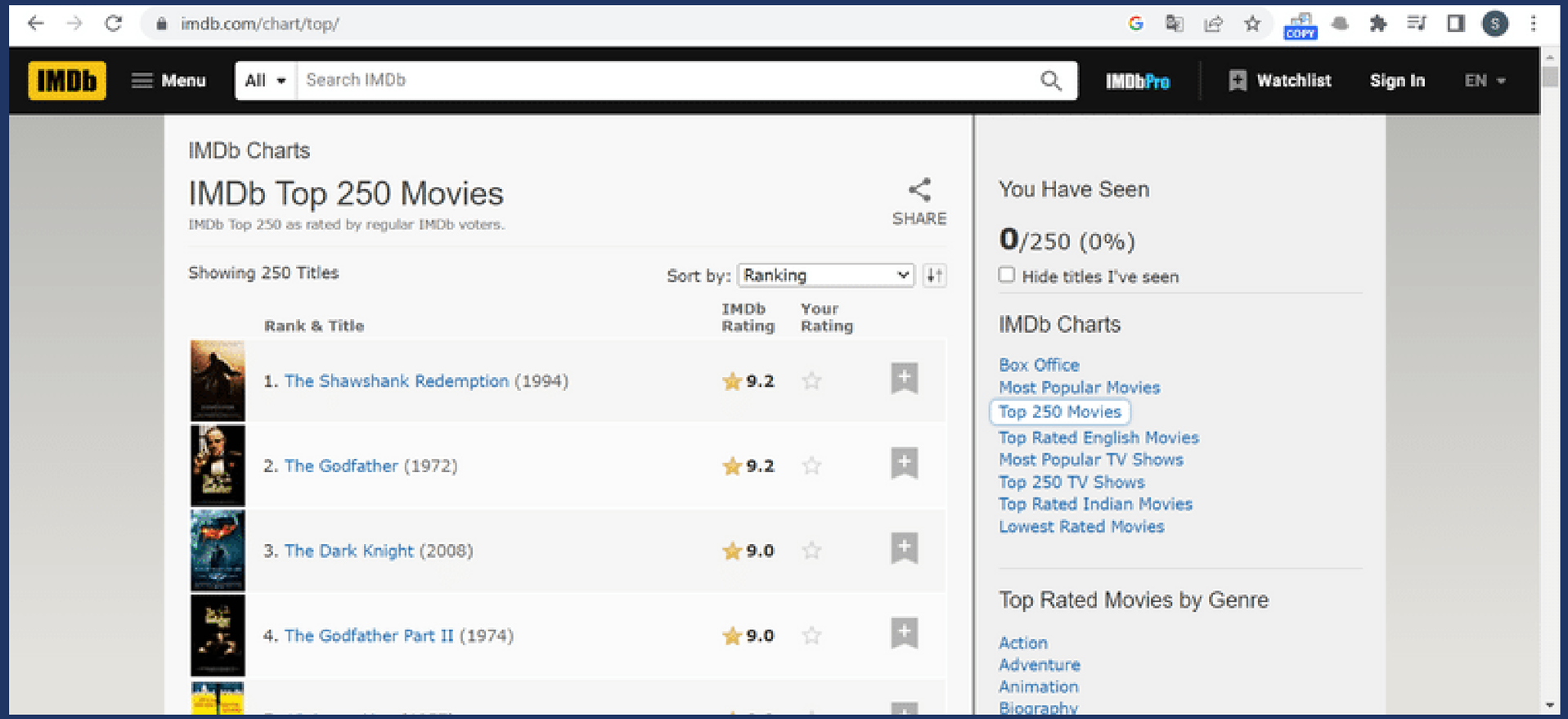
- World Bank Open Data

kaggle

ensias_ai_club

WEB SCRAPING

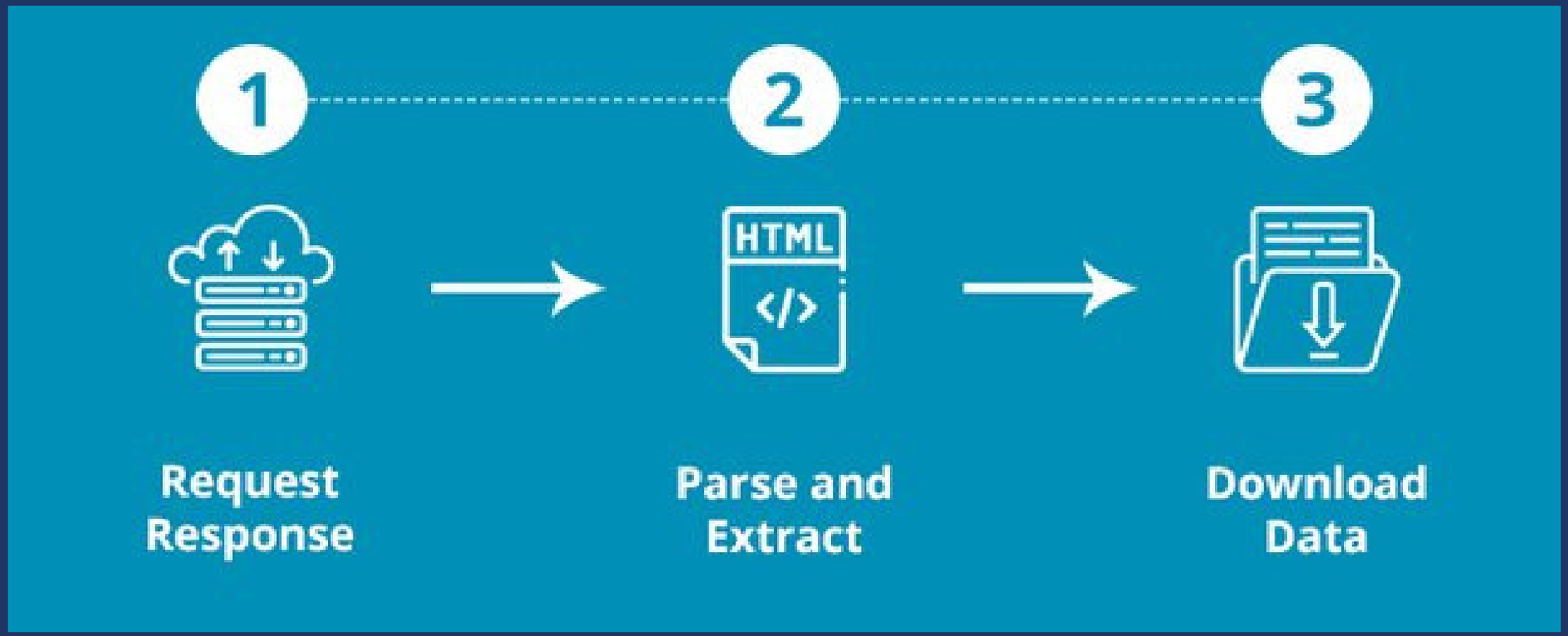
What is Web Scraping in Python?



ensias_ai_club

WEB SCRAPING

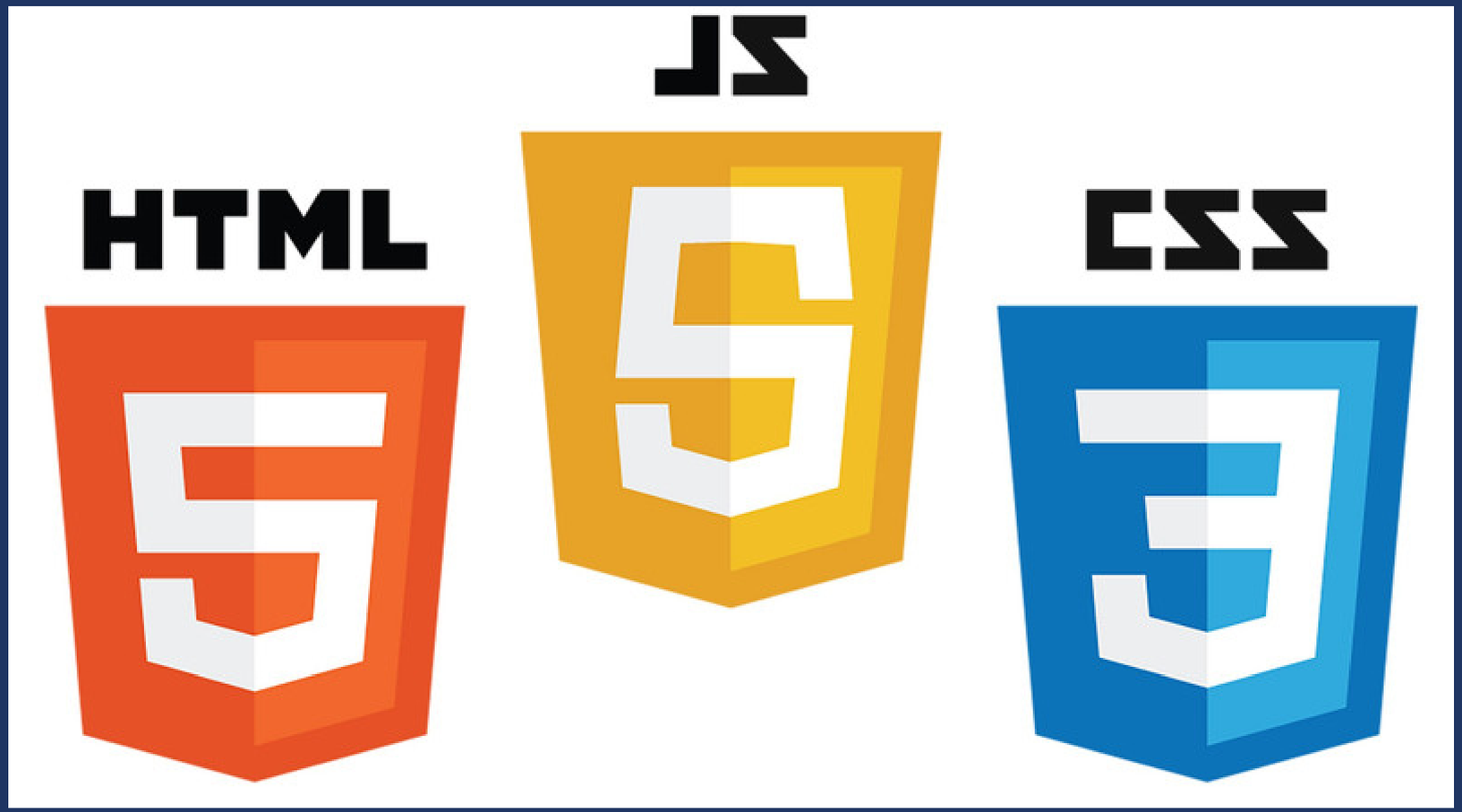
How Does Web Scraping Work?

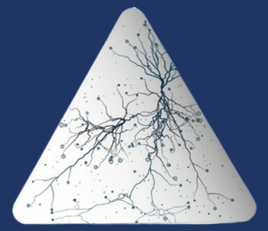


ensias_ai_club

WEB SCRAPING

The Components of a Web Page





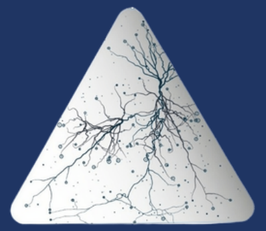
ensias_ai_club

WEB SCRAPING

HTML (HyperText Markup Language)


```
<html>
<head>
</head>
<body>
<p>
Paragraphe 1
</p>
<p>
Paragraphe 2
</p>
</body>
</html>
```

```
<html>
<head>
</head>
<body>
<p>
Here's a paragraph of text!
<a href="http://ensias.um5.ac.ma/">Ensias</a>
</p>
<p>
Here's a second paragraph of text!
<a href="https://www.python.org">Python</a> </p>
</body></html>
```



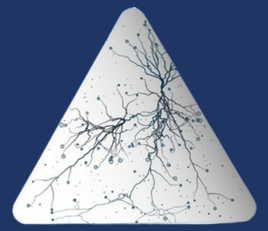
ensias_ai_club

WEB SCRAPING

HTML (HyperText Markup Language)

- div — indicates a division, or area, of the page.
- b — bolds any text inside.
- i — italicizes any text inside.
- table — creates a table.
- form — creates an input form.

```
<html>
<head>
</head>
<body>
<p class="bold-paragraph">
Here's a paragraph of text!
<a href="http://ensias.um5.ac.ma/" id="link">ENSIAS</a>
</p>
<p class="bold-paragraph extra-large">
Here's a second paragraph of text!
<a href="https://www.python.org" class="extra-large">Python</a>
</p>
</body>
</html>
```



ensias_ai_club

WEB SCRAPING

The requests library

```
import requests
page = requests.get("https://dataquestio.github.io/web-scraping-pages/simple.html")
page
```

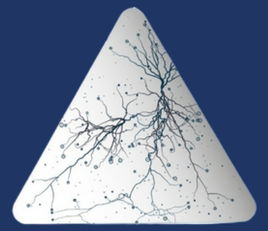
```
page.status_code
```

```
page.content
```

ensias_ai_club

WEB SCRAPING

Parsing a page with BeautifulSoup



ensias_ai_club

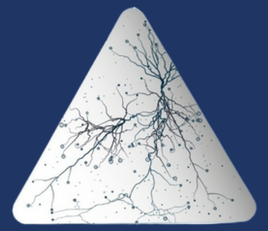
WEB SCRAPING

Parsing a page with BeautifulSoup

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(page.content, 'html.parser')
```

```
list(soup.children)
```

```
[type(item) for item in list(soup.children)]
```



ensias_ai_club

WEB SCRAPING

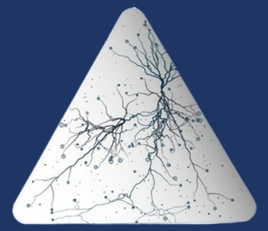
Parsing a page with BeautifulSoup

```
html = list(soup.children)[2]
```

```
list(html.children)
```

```
body = list(html.children)[3]
list(body.children)
```

```
p = list(body.children)[1]
p.get_text()
```



ensias_ai_club

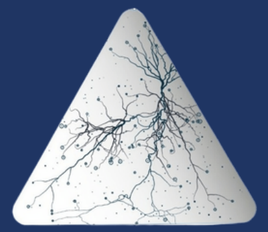
WEB SCRAPING

Finding all instances of a tag at once

```
soup = BeautifulSoup(page.content, 'html.parser')
soup.find_all('p')
```

```
soup.find_all('p')[0].get_text()
```

```
soup.find('p')
```



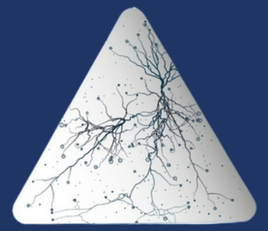
ensias_ai_club

WEB SCRAPING

Searching for tags by class and id

```
<html>
  <head>
    <title>A simple example page</title>
  </head>
  <body>
    <div>
      <p class="inner-text first-item" id="first">
        First paragraph.
      </p>
      <p class="inner-text">
        Second paragraph.
      </p>
    </div>
      <p class="outer-text first-item" id="second">
```

```
<b>
First outer paragraph.
</b>
</p>
<p class="outer-text">
<b>
Second outer paragraph.
</b>
</p>
</body>
</html>
```



ensias_ai_club

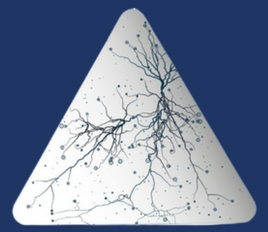
WEB SCRAPING

Searching for tags by class and id

```
page = requests.get("https://dataquestio.github.io/web-scraping-pages/ids_and_classes.html")
soup = BeautifulSoup(page.content, 'html.parser')
soup
```

```
soup.find_all('p', class_='outer-text')
```

```
soup.find_all(id="first")
```

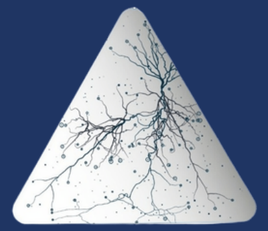


ensias_ai_club

WEB SCRAPING

Downloading IMdb data

Beautiful Soup

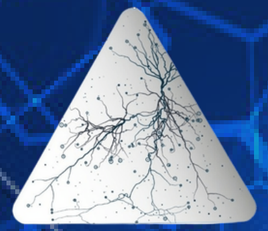


ensias_ai_club

WEB SCRAPING

Downloading IMdb data

```
excel=openpyxl.Workbook()
sheet=excel.active
sheet.title=""
sheet.append([ ])
excel.save(" ")
```



ensias_ai_club

THANK YOU