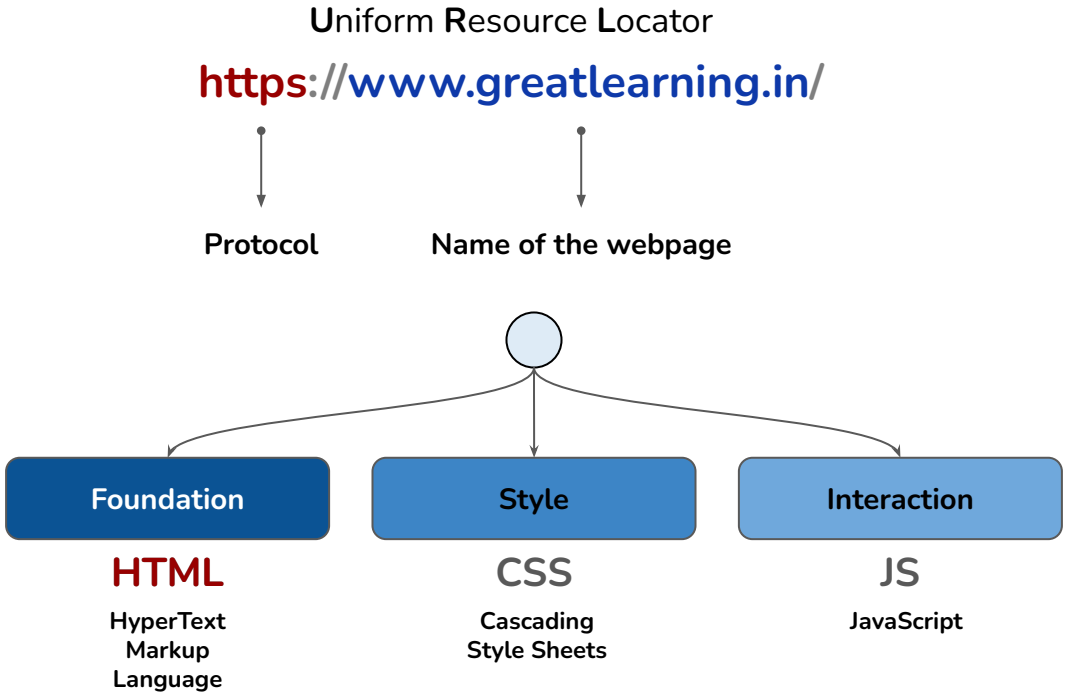
Data Wrangling with Python

Agenda

- Basics of HTML
- Regular Expressions
- Web Scraping with Python
- Exception Handling

Basics of HTML

Look inside a Webpage!



Look inside a Webpage!

```
<html>
<body>
  <center>
    <div class="container" id="intro">
      <h2>Introduction to HTML</h2>
      <p style="font-size:40px" class="title"> Hello World!</p>
      <p class="description">This is a basic HTML file</p>
    </div>
    <div>
      <img src="hello.png" width="200px">
    </div>
  </center>
</body>
</html>
```

Open Tag

Close Tag

Tags

<html> <body>

 <p>

<h2>

Attributes

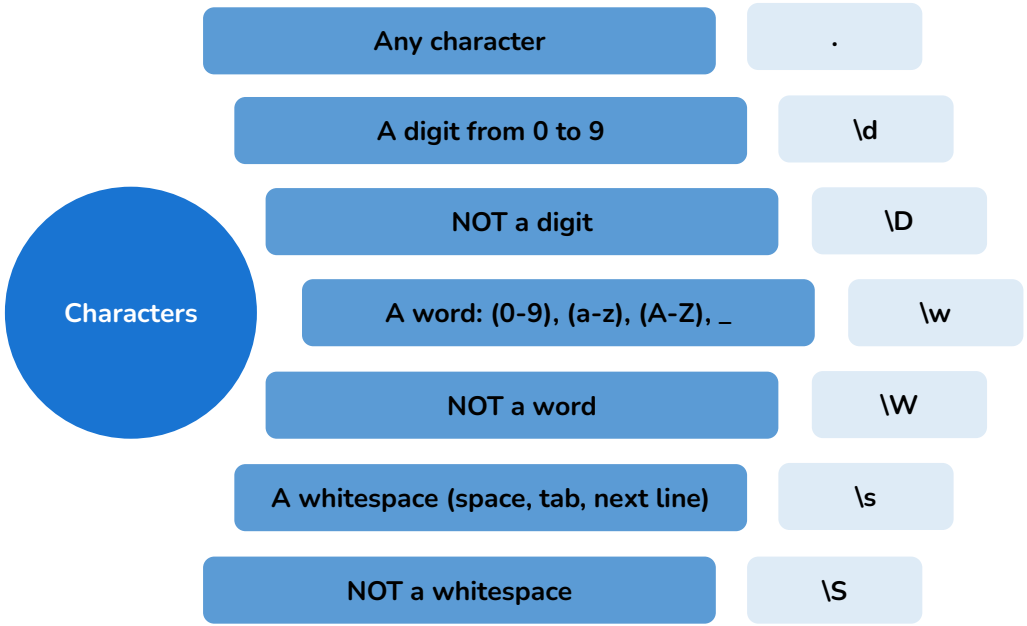
class="container" id="intro"

style="font-size:40px" class="title"

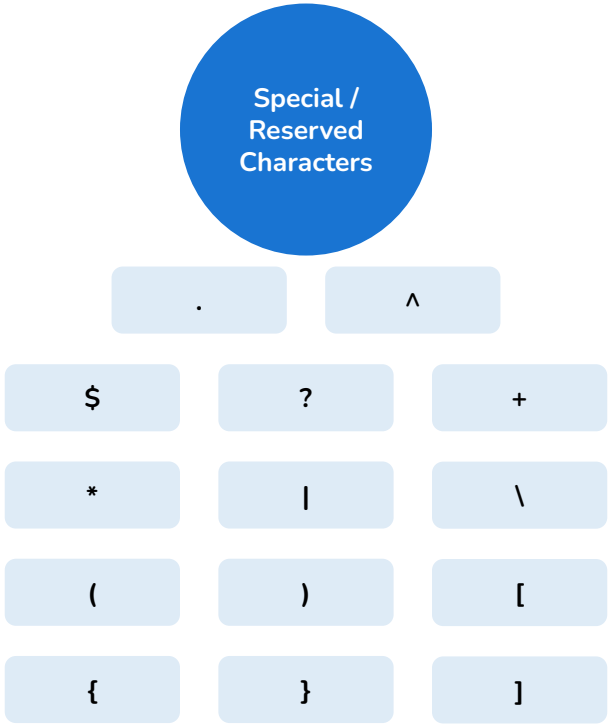
Explore a webpage using **Inspect**

Regular Expressions

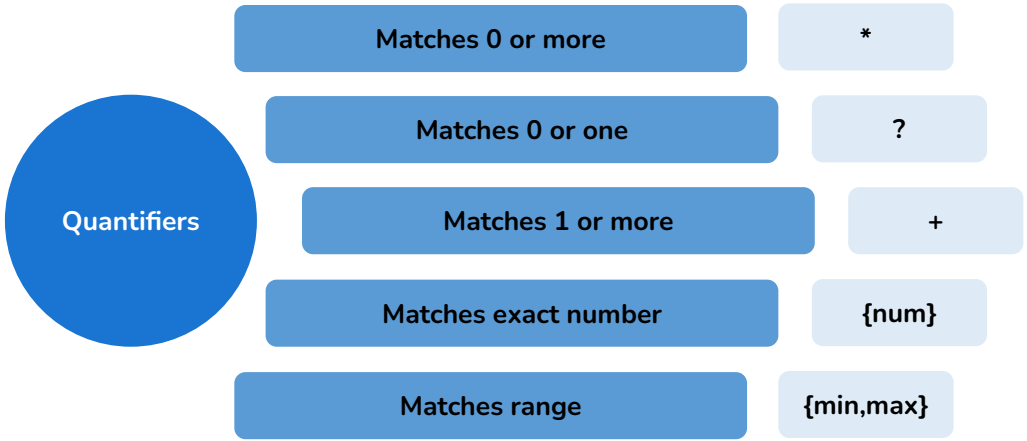
Finding characters in a text



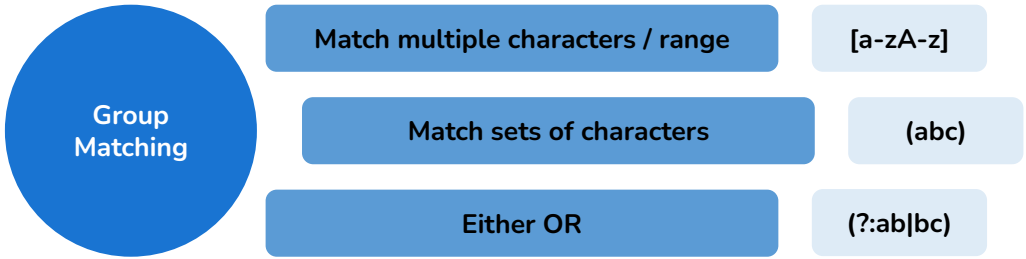
Finding characters in a text



Using quantifiers to match patterns



Matching groups of characters



Web Scraping with Python

Basics of **Web Scraping**

Let’s test our understanding by solving a problem!

Say you are interested to buy a laptop that fits your budget and needs. Assume you found a webpage called cosmic-pc-store.html.

But the webpage does not have any option to play around with the prices and specifications to select based on your needs.

So let’s try to scrape the data from this webpage, clean it using regex and prepare the dataset for analysis.

Post preparing the data, let’s find the different laptops that fit our need by querying the data we created.

Solution Approach

1. View the webpage and understand how the HTML is structured
2. Connect / read the HTML content of the webpage using BeautifulSoup
3. Find the data you need by matching the tags and attributes, and save the data in a pandas dataframe
4. Clean the text columns using RegEx and prepare the data for analysis
5. Perform EDA on the data to select a desired laptop

Exception Handling

Don’t let errors stop your code!

TypeError **AttributeError**

NameError **SyntaxError**

ZeroDivisionError

Do these look familiar?! These are the most common errors you encounter in Python

When Python encounters an error, it STOPs the code execution. You don’t want this to happen. You want your code to try and handle these errors and move on with the execution as much as possible.

Otherwise, everytime something goes wrong, you have to manually sit and debug your code. Instead, you can have python sort it out by itself!

Let’s take a look at how Exception Handling works in Python with a few examples...

What if your web page structure changes?

Great Learning

Happy Learning !