



SENIOR ADVISORY PROGRAM
(SAP)

ON

DATA GATHERING

BY

KARTIK TICHKULE

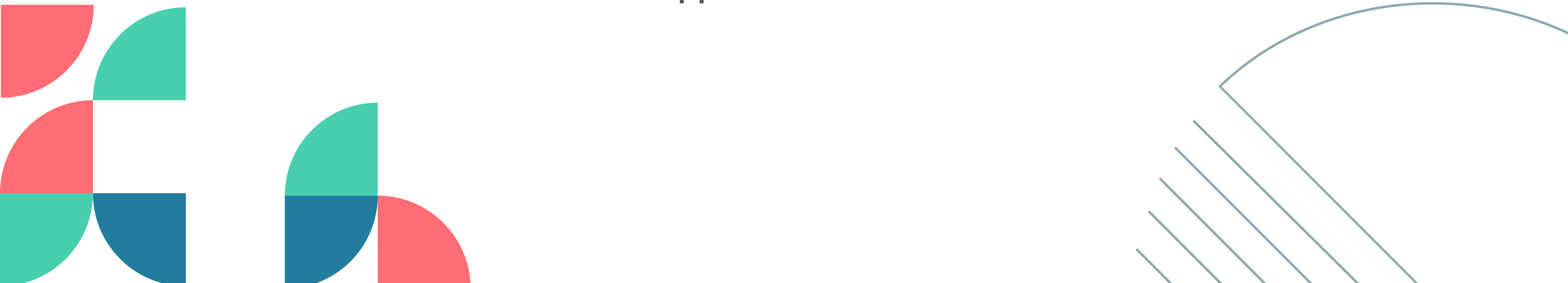


DATA GATHERING USING WEB-SCRAPING



WHAT IS WEB-SCRAPING?

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.



BASIC STEPS IN WEB SCRAPING

01 - PARSING THE WEB

PARSING CAN BE DONE BY **BeautifulSoup** LIBRARY or BY CREATING A LOCAL COPY IN PC

02 - ANALYZING HTML

FOR THIS THE STRUCTURE OF HTML DOC i.e **DOM** must be taken care of

03 - STORING DATA

THE DATA CAN BE STORED IN **DataFrame** using **Pandas** FOR FURTHER ANALYSIS or IN THE CSV FORMAT



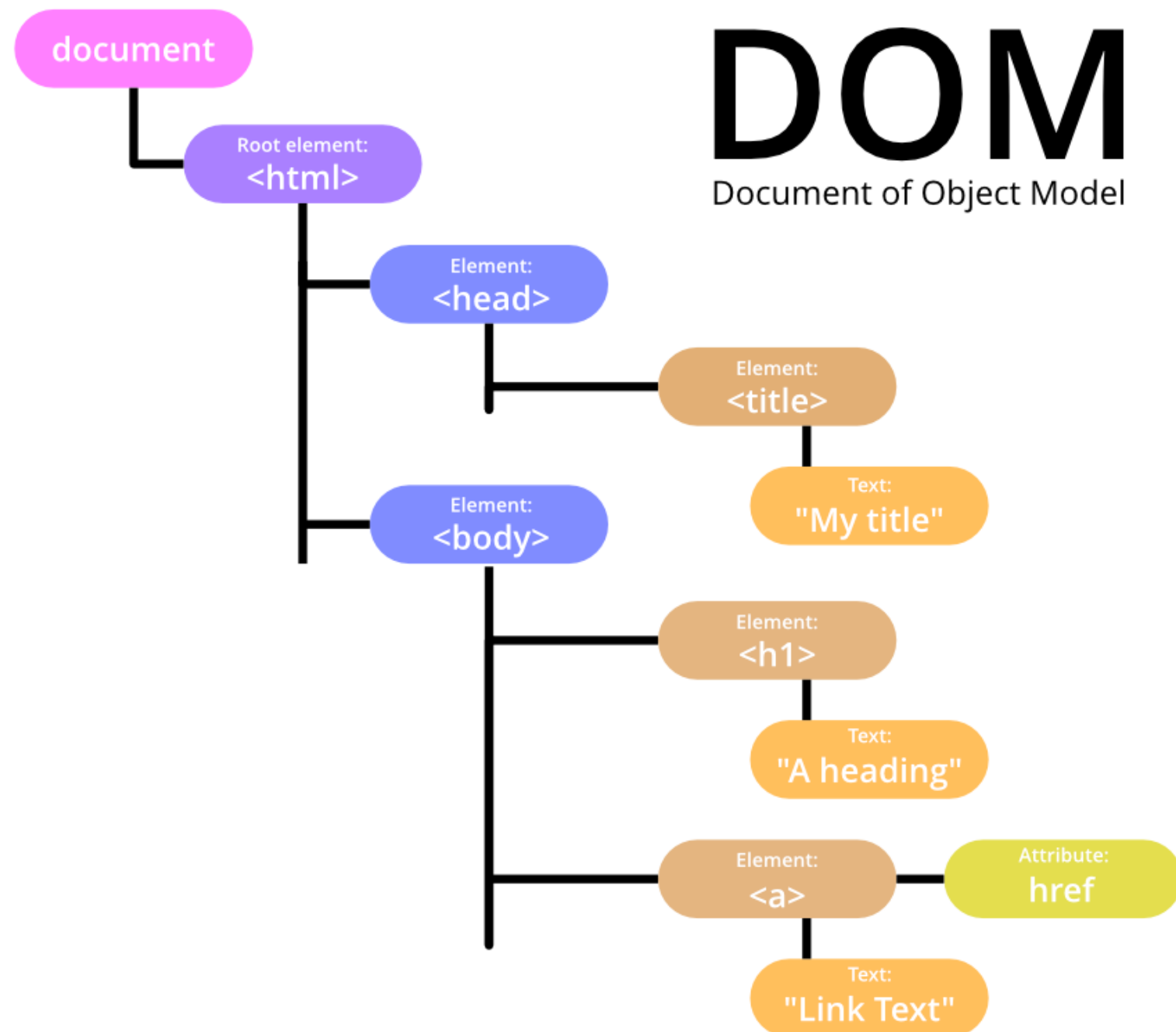
BEAUTIFUL SOUP



- Beautiful Soup is a python library which is named after a Lewis Carroll poem of the same name in “Alice’s Adventures in the Wonderland”.
- Beautiful Soup is a python package and as the name suggests, parses the unwanted data and helps to organize and format the messy web data by fixing bad HTML and present to us in an easily-traversable XML structures.
- In short, Beautiful Soup is a python package which allows us to pull data out of HTML and XML documents.



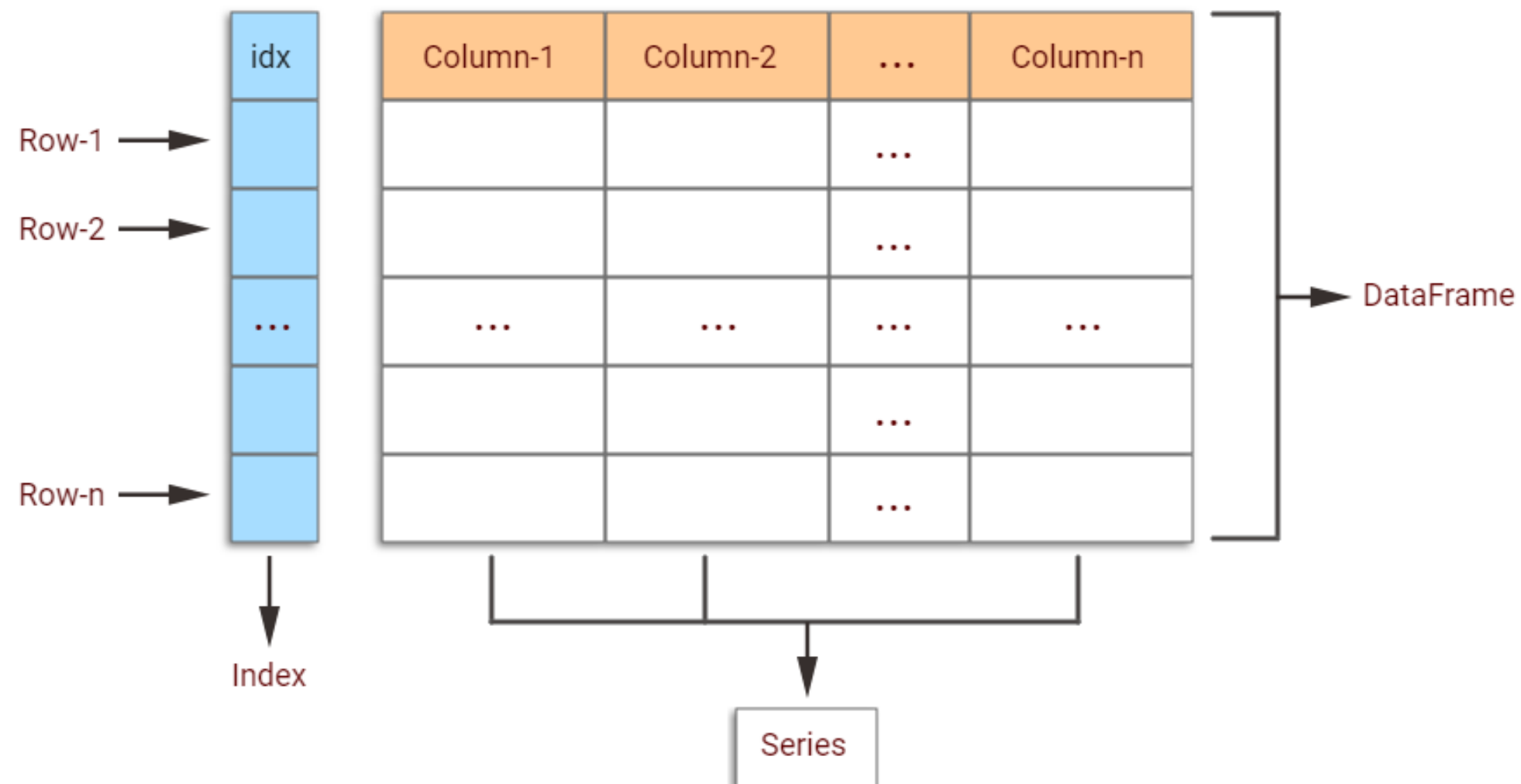
DOCUMENT OF OBJECT MODEL (DOM)



- The Document Object Model (DOM) is a programming interface for web documents. It represents the page so that programs can change the document structure, style, and content. The DOM represents the document as nodes and objects; that way, programming languages can interact with the page.
- A web page is a document that can be either displayed in the browser window or as the HTML source. In both cases, it is the same document but the Document Object Model (DOM) representation allows it to be manipulated. As an object-oriented representation of the web page, it can be modified with a scripting language such as JavaScript.

DATAFRAME IN PANDAS

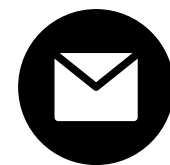
Pandas Data structure



- DataFrame can be thought of as a two-dimensional labeled data structure with columns of potentially different types.
- It resembles a spreadsheet or SQL table, where data is organized into rows and columns, each with its own label. This tabular structure enables efficient manipulation, cleaning, transformation, and analysis of data.



THANK YOU



mailkartik497@gmail.com