

SINDH MADERSATUL ISLAM UNIVERSITY

DAPARTEMENT OF SOFTWARE ENGINERING

PROJECT NAME : WEB SCRAPING

COURSE COORDINATOR : AMEEN CHAHJRO

COURSE NAME : DATA STRUCTURE & ALGORITHM

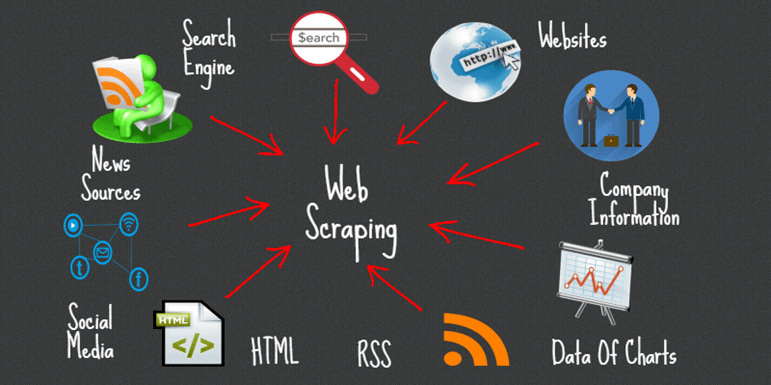
GROUP MEMBER:

* USAMA ABDUL JABBAR (BSE-20F-120)
* JAWAD IRSHAD(BSE-20F-045)
* ALEEM ARIF(BSE-20F-017)

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| S.NO | TOPIC | PAGE NO |
| 01 | INTRO. TO WEB SCRAPING | 03 |
| 03 | WHAT IS BEAUTIFUL SOUP | 04 |
| 04 | HOW TO DO WEB SCARPING | 05 |
| 05 | CONCLUSION | 09 |

INTRODUCTION TO WEBSCRAPING

Web scraping is a term used to describe the use of a program or algorithm to extract and process large amounts of data from the web. Whether you are a data scientist, engineer, or anybody who analyzes large amounts of datasets, the ability to scrape data from the web is a useful skill to have. Let's say you find data from the web, and there is no direct way to download it, web scraping using Python is a skill you can use to extract the data into a useful form that can be imported. 

WHAT IS BEAUTIFUL SOUP

**Beautiful Soup** is a Python library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser and provides Pythonic idioms for iterating, searching, and modifying the parse tree.



WEB SCRAPING USING BEAUTIFUL SOUP

Using Jupyter Notebook, you should start by importing the necessary modules (pandas, numpy, requests). If you don't have Jupyter Notebook installed, I recommend installing it using the Anaconda Python distribution which is available on the internet. Graphical user interface, text, application, Word

Description automatically generated

After importing necessary modules, you should specify the URL

A picture containing text

Description automatically generated

Getting the html of the page is just the first step. Next step is to create a Beautiful Soup object from the html. This is done by passing the html to the BeautifulSoup() function. The Beautiful Soup package is used to parse the html, that is, take the raw html text and break it into Python objects.

Chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

You can view the html of the webpage by right-clicking anywhere on the webpage and selecting "Inspect." This is what the result looks like.Letter

Description automatically generated with medium confidence

You can use the find\_all() method of soup to extract useful html tags within a webpage. Examples of useful tags include < a > for hyperlinks, < table > for tables, < tr > for table rows, < th > for table headers, and < td > for table cells.

Graphical user interface, text, application

Description automatically generated

As you can see from the output above, html tags sometimes come with attributes such as class, src, etc. These attributes provide additional information about html elements. You can use a for loop and the get('"href") method to extract and print out only hyperlinks

Text

Description automatically generated

The next step is to convert the list into a dataframe and covert into csv file by using pandas.

Graphical user interface, text

Description automatically generated

Now in the last we read the csv file im pythom by using pandas.

Graphical user interface, text, application, email

Description automatically generated

Conclusion :

In this documentation, we performed web scraping using Python. we used the Beautiful Soup library to parse html data and convert it into a form that can be used for analysis. Web scraping is a technique that includes many positive and beneficial aspects for those who use it. So, the following are some of the main but substantial advantages that have made this method so popular among various individuals and industries:

A picture containing text

Description automatically generated

The end.