Task - Develop a solution to extract the following information from websites

Social Media Links

Tech Stack

Meta Title

Meta Description

Payment Gateways

Website Language

Category Of Website

Programming Language Used- Python

Approach of the problem -

BeautifulSoup is a python library used for parsing HTML and XML documents.It creates a parse tree from page source code that can be used to extract data easily

Working of Scrapping data from websites

When I created a BeautifulSoup object, the library first takes the input HTML or XML document as a string. It also allows us to specify the parser We want to use

Choosing a Parser :

html.parser

lxml

html5lib

Parsing the Document :

The chosen parser reads the HTML or XML document and creates a parse tree. This tree represents the structure of the document , with nodes corresponding to tags,attributes,text,and comments.

Creating the Parse Tree:

Tag Objects: Each HTML Or XML tag becomes a Tag object. These objects contain:

The tag name(div,p)

A dictionary of attributes(class, id)

Alist of child nodes, which can be other tags,strings, or comments

Navigating and Searching the Parse Tree :

BeautifulSoup provides various methods to navigate and search the parse tree. These methods and search the parse tree. These methods allow for powerful and flexible data extraction.

Navigating :

tag.parent : Access the parent of a tag

tag.contents : Access the direct children of a tag.

tag.next_sibling and tag.previous_sibling: Navigate between sibling tags

Searching :

soup.find() : Finds the first tag that matches the criteria

soup.find_all() : Finds all tags that match the criteria

soup.select() : Uses CSS selectors to find tags

Handling Bad Markup :

One of BeautifulSoup’s strengths is its ability to handle poorly-formed or broken HTML.