HTML Webscrapping: CODE Harry

Bringing the website html content to python program using beautiful soup module.

We can bring the data and dump in any file

How website works:

User----🡪 send requests like www.google.com to server -🡪 web server=🡺 gives html code to browser and the brower will beautify the content and shows in user screen=🡺user

In this we will learn how to get the html from web server and instead of giving it to browser we will dump in python=> This is called webscrapping.

Installation:

Step0: Setup environment in python

a) pip install requests

b) pip install html5lib

b) pip install bs4 (beautifulsoup)

Step1: Get the HTML

Step2: Parse the html (using angular )

Step3: Html tree traverse\*\*\*\* imp step(parse🡪tree🡪manipulating like queries

Step : use Visual studio in the folder to write code (use internet, stack flow)

Notes:

Request is a module that brings the content from a website,

Html5lib it does not parse that content, it will create tree like structure using bs4 (div,paragraph)

Bs4 will parse the content for making our search easy (creating trees)

###We can get data and can manipulate the data

## Beautifulsoup:

We can scape the website using 2 ways:

a)Using API \*\*Genuine …if they are providing API

b) Using HTML web Scrapping using some tools like bs4

## VS code:

1.Create a file main.py in Webscrapping folder through VSCode

2. Open command prompt and go to the webscrapping folder and create a python env

conda create –n webscrap (different from flask\_env)

OR Go to anaconda and activate webscrap

In VSCODE : conda activate webscrap

2. Open a new terminal and install all requirements pip install request, bs4,html5lib

pip install requests

pip install html5lib

pip install bs4 (beautifulsoup)

3.Then import requests

from bs4 import BeautifulSoup

Beautifulsoup is a class

##Now we will scrap the codewithhary website. So give it in url

url=” <https://www.codewithharry.com>”

##Open this website in browser and Rightclick to open Inspect or view source

We can see a html code.

#STEP1: Get the HTMLCode

Content

##request is a module in python that helps in get or post request

To get plane url we will use get method.

#Step2: Parse the HTML

#step: HTML  Tree traversal

iN tree traversing we can list of ul , lists,…

Mostly used functions:

Find\_all text

Tags

get\_text