**Python Course-3**

**Using Python to access web data**

Week- 4>12.4- Retrieving web pages ->

Here we are using urllib library to retrieve the webpages in the form of file, and we don’t have to encode and decode the web data in python, as we were doing in the socket programming.

Steps-> We have to import the urllib library for this-

Basic Code syntax-

Import urllib.request, urllib parse, urllib error

fhand= urllib.request.urlopen(“http://data.pr4e.org/romeo.txt”) // we can use .html extension also, which will give html headers along with content.

for line in fhand:

print(line.decode().rstrip())

Week -4> 12.5- Parsing Web pages->

Here we learn about web crawler using Beautiful soup library. In this we don’t have to use find, split or regular expressions to find the links in the webpages.

To download the Beautiful Soup library-

Commands-

To run this, you can install beautiful soup-

<https://pypi.python.org/pypi/beautifulsoup4>

To Download the file-

<http://www.py4e.com/code3/bs4.zip>

and unzip it in the same directory as this file.

Sample code-

Import urllib.request, urllib.parse, urllib.error

from bs4 import BeautifulSoup

url=input(“Enter\_”)

html=urllib.request.urlopen.url.read()

soup=BeautifulSoup(html,’html.parser’) // soup will have a tree structure of of all the links in the webpage.

#Retrieve all of the anchor tags-

Tags=soup(‘a’)

for tag in tags:

Print(tags.get(‘href’,None))