lab9-ui22cs03-anywebs

October 28, 2023

Let's Make a Webscraping tool using python that extract phone number and email from any webpage

[10]:
```
pip install requests
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (2.31.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests) (3.3.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests) (2023.7.22)
```

[11]:
```
pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-
packages (4.11.2)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-
packages (from beautifulsoup4) (2.5)
```

[14]:
```
#Importing the requried library
#Request library in Python is used to make HTTP requests, and it serves a wide␣
 ↪range of purposes

import requests
#Beautiful Soup is a popular Python library used in web scraping and parsing␣
 ↪HTML and XML documents
from bs4 import BeautifulSoup
# re module in Python, short for "regular expressions," is a powerful tool for␣
 ↪working with text and pattern matching.
import re
import pandas as pd
```

[17]:
```
#Let's take input of website to extract contact details
url = input("Enter the Webpage URL: ")
```

```python
urlrequest=requests.get(url)
```

```
Enter the Webpage URL: http://www.iiitsurat.ac.in
```

[18]:

```python
#Soup is used to parse and navigate the HTML content of a webpage
soup = BeautifulSoup(urlrequest.text, 'html.parser')
```

[19]:

```python
# Below is used to extract plain text content from the BeautifulSoup object soup
text = soup.get_text()
```

[20]:

```python
#It will print all html webpage texts
print(text)
```

```




IIIT Surat :: Home
```

Home

About IIIT Surat

About IIIT Surat
Vision & Mission
Director's Desk
Board of Governors
Finance Committee
Senate

Administration

Faculty

MOU

Academics

B.Tech 2018-2022
B.Tech 2022-23 onwards
Academic Calendar
Holiday List
B.Tech Academic Rules
Ph.D Academic Rules
Ph.D Fee Structure

Admission

Admission 2023

Results
RTI
T & P
Career

Links for Students

IIIT-Surat Moodle
IIIT-Surat Coursera
IIIT-Surat Github

IIIT-MATLAB
NEWSLETTERS

Events

2nd Convocation Ceremony 19-08-2023
International Yoga Day 21st June 2023
World Environment Day 2023

Notification

Useful Links

National Overseas Scholarship Scheme for the 2022-23
Loksabha Research Fellowships
National Scholarship Portal
Uttar Pradesh Government - Scholarship System

Read More

MoE Initiatives

Follow us on Facebook

Follow us on Twitter

Gallery

Follow us on Youtube

About Us

About IIIT Surat
Vision & Mission
Director's Desk
Board of Governors
Administration

Contact Us

Where we are?

Contact

Indian Institute of Information Technology, Surat
Kholvad Campus
Kamrej, Surat - 394190
Gujarat
Phone : 02621-298060
Email : office@iiitsurat.ac.in

Our Partners

GNFC

Gujarat Informatics Limited

Gujarat Gas Limited

SBI Collect

[24]:

```
#Let's find all the occurance of String "Phone" in webpage text
phone= soup.find_all(text=re.compile(r'Phone', re.I))
df_phone=pd.DataFrame(phone)
print(df_phone)
```

```
                      0
0  Phone : 02621-298060
```

```
<ipython-input-24-e8906c7e3018>:2: DeprecationWarning: The 'text' argument to
find()-type methods is deprecated. Use 'string' instead.
  phone= soup.find_all(text=re.compile(r'Phone', re.I))
```

[22]:

```
#Here we will using regular expression to find Email which having "@" in text␣
  ↪and display it
email= re.findall(r'\S+@\S+', text)
print("Email:",email)
```

```
Email: ['office@iiitsurat.ac.in']
```

[23]:

```
#CODE BY UI22CS03
#Note : The Phone no. and Email ID will only get extracted if the input url␣
  ↪having that on webpage so further
#I will more features like it can even natigate to another webpages of website␣
  ↪so it can find more informations of a website
```