

## CSE3024 - WEB MINING (L41 + L42)

**NAME – ANMOL**  
**REG. NO. - 19BCE0891**

## DIGITAL ASSIGNMENT – 2

**URL used :** <https://anmol1804.github.io/MyPortfolio/>

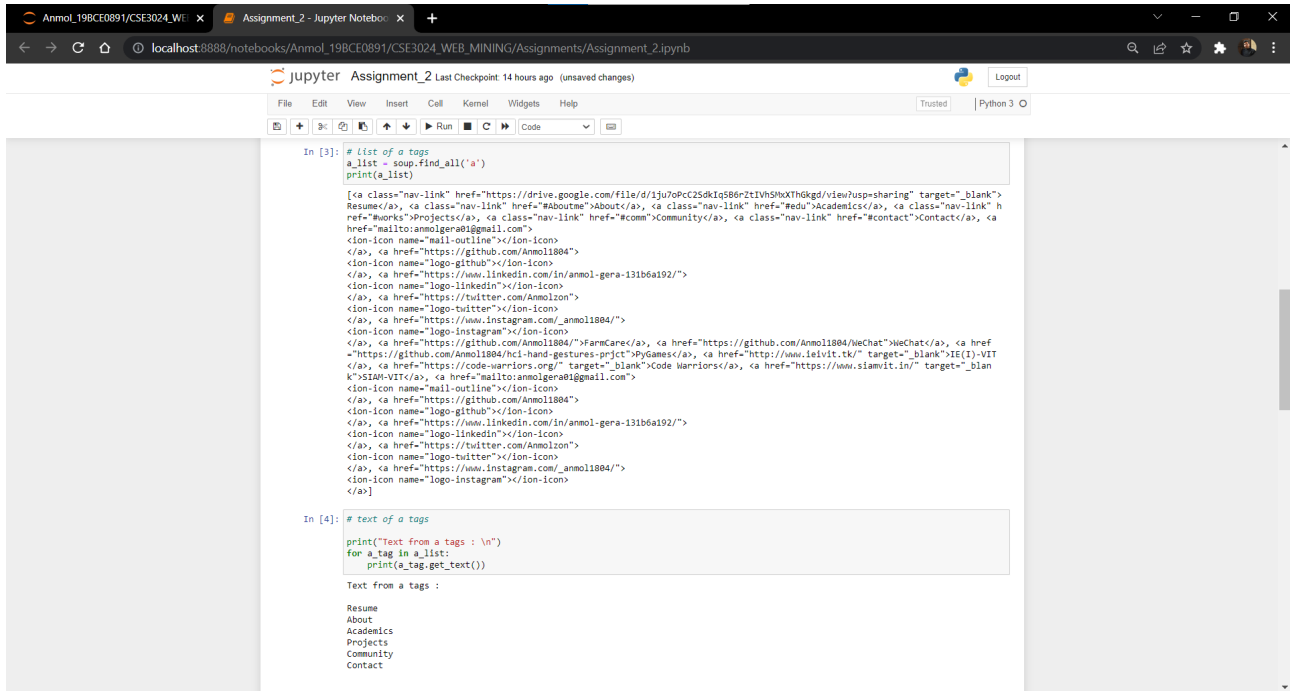
**1. Write a python program to scrape website to extract the following:**

**a) raw HTML content**

[illegible][illegible]

**b) tags (title, p, a, div)**

## CSE3024 - WEB MINING (L41 + L42)



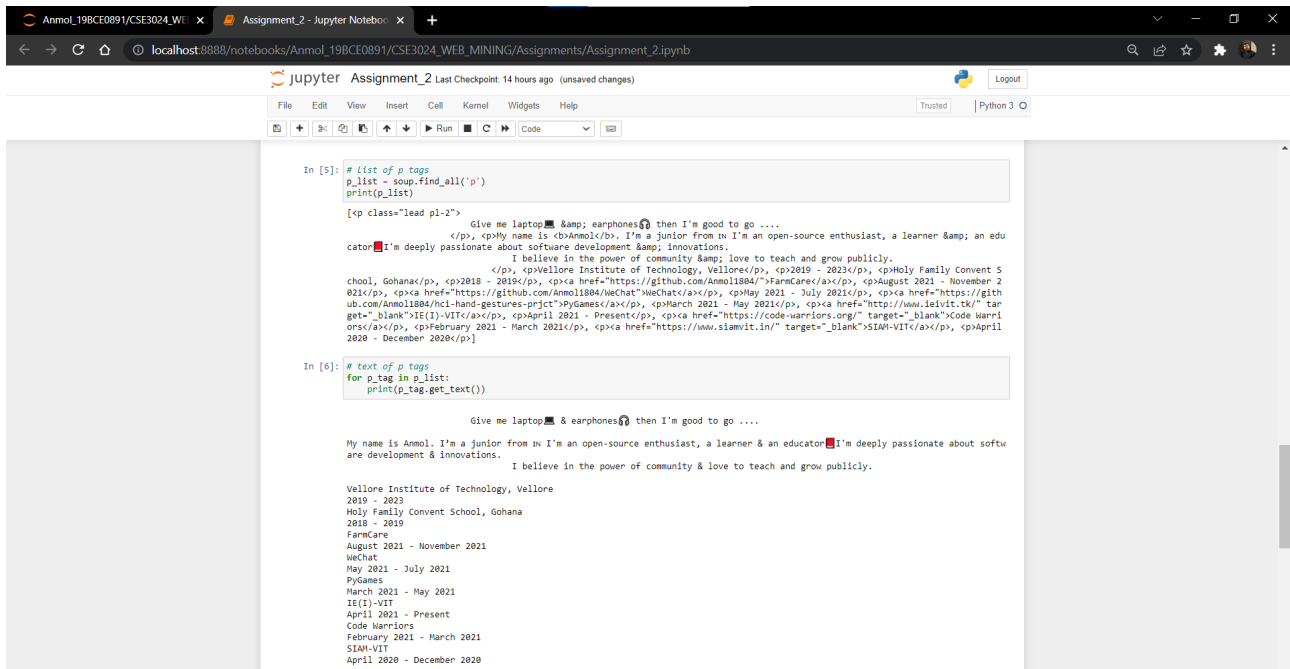
```
In [3]: # list of a tags
a_list = soup.find_all('a')
print(a_list)

[<a class="nav-link" href="https://drive.google.com/file/d/1ju7oPcC25dKiQ586rZtIVh5hXtHdkgd/view?usp=sharing" target="_blank">Resume</a>, <a class="nav-link" href="#aboutme">About</a>, <a class="nav-link" href="#edu">Academics</a>, <a class="nav-link" href="#works">Projects</a>, <a class="nav-link" href="#com">Community</a>, <a class="nav-link" href="#contact">Contact</a>, <a href="mailto:anmolgera91@gmail.com"></a>, <ion-icon name="mail-outline"></ion-icon>
</a>, <a href="https://github.com/Anmol1804"></ion-icon>
<ion-icon name="logo-github"></ion-icon>
</a>, <a href="https://www.linkedin.com/in/anmol-gera-131b6a192/"></ion-icon>
<ion-icon name="logo-linkedin"></ion-icon>
</a>, <a href="https://twitter.com/Anmol1804"></ion-icon>
<ion-icon name="logo-twitter"></ion-icon>
</a>, <a href="https://www.instagram.com/_anmol1804/"></ion-icon>
<ion-icon name="logo-instagram"></ion-icon>
</a>, <a href="https://github.com/Anmol1804"/>FamCare</a>, <a href="https://github.com/Anmol1804/WeChat">WeChat</a>, <a href="https://github.com/Anmol1804/hci-hand-gestures-prjct">PyGames</a>, <a href="http://www.ielvit.tk/" target="_blank">IE(I)-VIT</a>, <a href="https://code-warriors.org/" target="_blank">Code Warriors</a>, <a href="https://www.siamvit.in/" target="_blank">SIAM-VIT</a>, <a href="mailto:anmolgera91@gmail.com"></a>, <ion-icon name="mail-outline"></ion-icon>
</a>, <a href="https://github.com/Anmol1804"></ion-icon>
<ion-icon name="logo-github"></ion-icon>
</a>, <a href="https://www.linkedin.com/in/anmol-gera-131b6a192/"></ion-icon>
<ion-icon name="logo-linkedin"></ion-icon>
</a>, <a href="https://twitter.com/Anmol1804"></ion-icon>
<ion-icon name="logo-twitter"></ion-icon>
</a>, <a href="https://www.instagram.com/_anmol1804/"></ion-icon>
<ion-icon name="logo-instagram"></ion-icon>
</a>]

In [4]: # text of a tags
print("Text from a tags : \n")
for a_tag in a_list:
    print(a_tag.get_text())

Text from a tags :

Resume
About
Academics
Projects
Community
Contact
```



```
In [5]: # list of p tags
p_list = soup.find_all('p')
print(p_list)

[<p class="lead pl-2">
    Give me laptop & earphones then I'm good to go ....
    </p>, <p>My name is <b>Anmol</b>. I'm a Junior from I'm an open-source enthusiast, a learner & an educator. I'm deeply passionate about software development & innovations.
    I believe in the power of community & love to teach and grow publicly.
    </p>, <p>Vellore Institute of Technology, Vellore</p>, <p>2019 - 2023</p>, <p>Holy Family Convent School, Gohana</p>, <p>2018 - 2019</p>, <p>FamCare</p>, <p>August 2021 - November 2021</p>, <p>WeChat</p>, <p>May 2021 - July 2021</p>, <p>PyGames</p>, <p>March 2021 - May 2021</p>, <p>IE(I)-VIT</p>, <p>April 2021 - Present</p>, <p>Code Warriors</p>, <p>February 2021 - March 2021</p>, <p>SIAM-VIT</p>, <p>April 2020 - December 2020</p>]

In [6]: # text of p tags
for p_tag in p_list:
    print(p_tag.get_text())

    Give me laptop & earphones then I'm good to go ....

My name is Anmol. I'm a Junior from I'm an open-source enthusiast, a learner & an educator. I'm deeply passionate about software development & innovations.
    I believe in the power of community & love to teach and grow publicly.

Vellore Institute of Technology, Vellore
2019 - 2023
Holy Family Convent School, Gohana
2018 - 2019
FamCare
August 2021 - November 2021
WeChat
May 2021 - July 2021
PyGames
March 2021 - May 2021
IE(I)-VIT
April 2021 - Present
Code Warriors
February 2021 - March 2021
SIAM-VIT
April 2020 - December 2020
```

## CSE3024 - WEB MINING (L41 + L42)

[illegible]

The screenshot displays a Jupyter Notebook environment within a web browser. The browser's address bar indicates the notebook is located at `localhost:8888/notebooks/Anmol_19BCE0891/CSE3024_WEB_MINING/Assignments/Assignment_2.ipynb`. The Jupyter interface features a top menu bar with options like File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu is a toolbar with icons for file management and execution. The main area contains a code cell with the following Python code:

```
In [9]: title = soup.find('title')
        print(title.get_text())
```

The output area below the code cell is currently empty.

**c) all textual content.**

Assignment\_2 Last Checkpoint: 14 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```
In [10]: # All text
text = soup.get_text()
print(text)
```

August 2021 - November 2021

Created a Seller account where a seller can add his farm and sell multiple crops,vegetables etc.  
Created a Customer account where customer can buy any number of farm products from any number of farms.

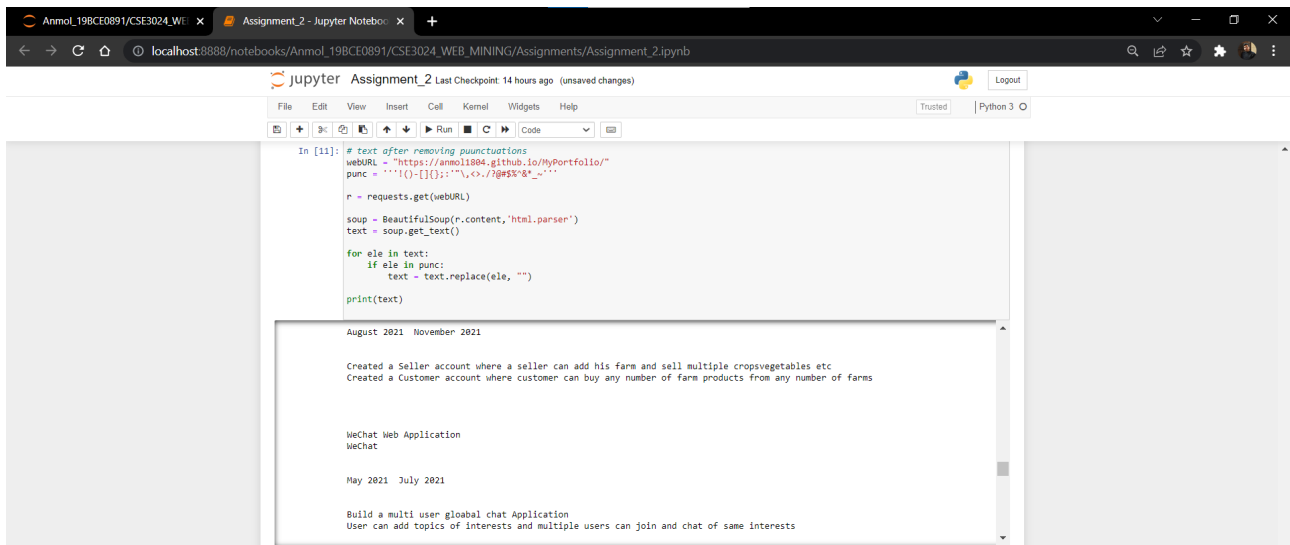
WeChat Web Application  
WeChat

May 2021 - July 2021

Build a multi user global chat Application  
User can add topics of interests and multiple users can join and chat of same interests

(textual content after removing punctuation)

## CSE3024 - WEB MINING (L41 + L42)



### CODE -

```
# raw HTML content
import requests
webURL = "https://anmol1804.github.io/MyPortfolio/"
r = requests.get(webURL)
print(r.content)
```

```
# raw HTML content using BeautifulSoup
from bs4 import BeautifulSoup
soup = BeautifulSoup(r.content, 'html.parser')
print(soup.prettify())
```

```
# list of a tags
a_list = soup.find_all('a')
print(a_list)
```

```
# text of a tags
print("Text from a tags : \n")
for a_tag in a_list:
    print(a_tag.get_text())
```

```
# list of p tags
p_list = soup.find_all('p')
print(p_list)
```

```
# text of p tags
for p_tag in p_list:
    print(p_tag.get_text())
```

```
# list of div tags
div_list = soup.find_all('div')
print(div_list)
```

## CSE3024 - WEB MINING (L41 + L42)

```
# text of div tags
for div_tag in div_list:
    print(div_tag.get_text())

title = soup.find('title')
print(title.get_text())

# All text
text = soup.get_text()
print(text)

# text after removing puunctuations
webURL = "https://anmol1804.github.io/MyPortfolio/"
punc = "!()-[]{};:'\".,<>./?@#$$%^&* _~"
r = requests.get(webURL)
soup = BeautifulSoup(r.content,'html.parser')
text = soup.get_text()
for ele in text:
    if ele in punc:
        text = text.replace(ele, "")

print(text)
```