CSE 3024

Web Mining

LAB ASSESSMENT - 2

NAME: Vibhu Kumar Singh
REG. NO: 19BCE0215
TEACHER: Mr. Hiteshwar Kumar Azad

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

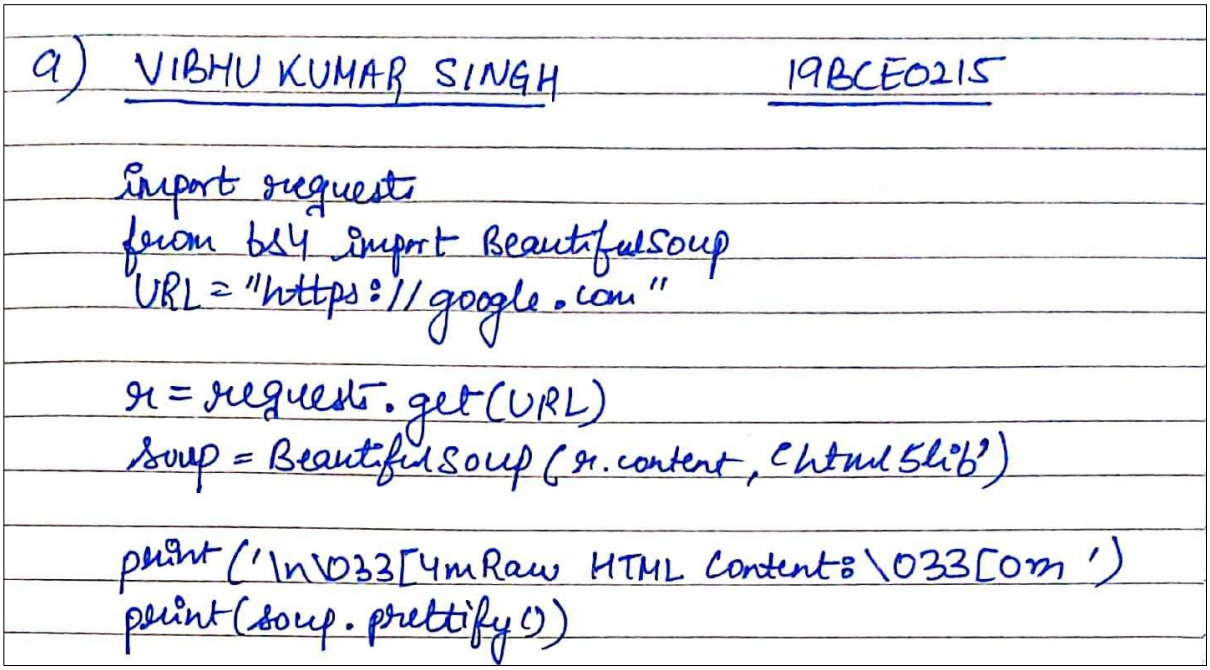
a) **raw HTML content**
b) **tags (title, p, a, div)**
c) **all textual content.**

Note: *Consider the input to be any website of your choice.*

Ans 1.

a) Raw HTML content

HANDWRITTEN CODE:



CODE:

```
import requests
from bs4 import BeautifulSoup
URL = "https://google.com"

r=requests.get(URL)
soup=BeautifulSoup(r.content, 'html5lib')

print('\n\033[4mRaw HTML Content:\033[0m ')
print(soup.prettify())
```

CODE SCREENSHOT:

C:\Users\Vibhu\OneDrive - vit.ac.in\Desktop\Winter Semester 21-22\Web Mining\ELA\LAB-2\rawHTML.py

rawHTML.py

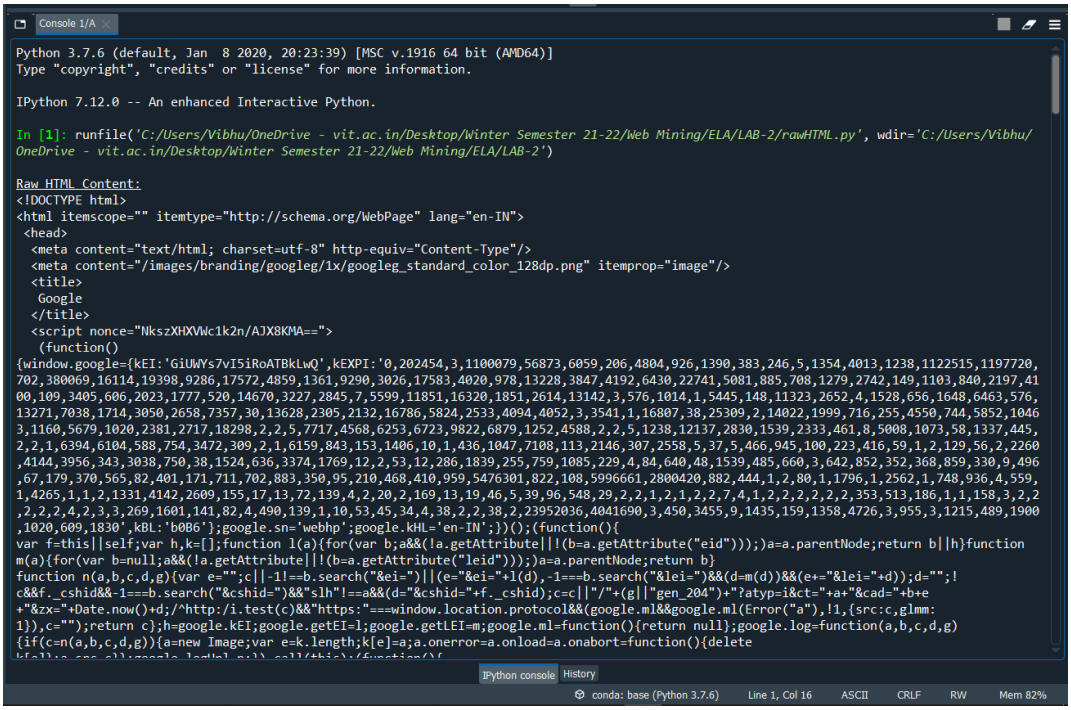
```
1  import requests
2  from bs4 import BeautifulSoup
3  URL = "https://google.com"
4
5  r=requests.get(URL)
6  soup=BeautifulSoup(r.content, 'html5lib')
7
8  print('\n\033[4mRaw HTML Content:\033[0m ')
9  print(soup.prettify())
```

OUTPUT:

Raw HTML Content:

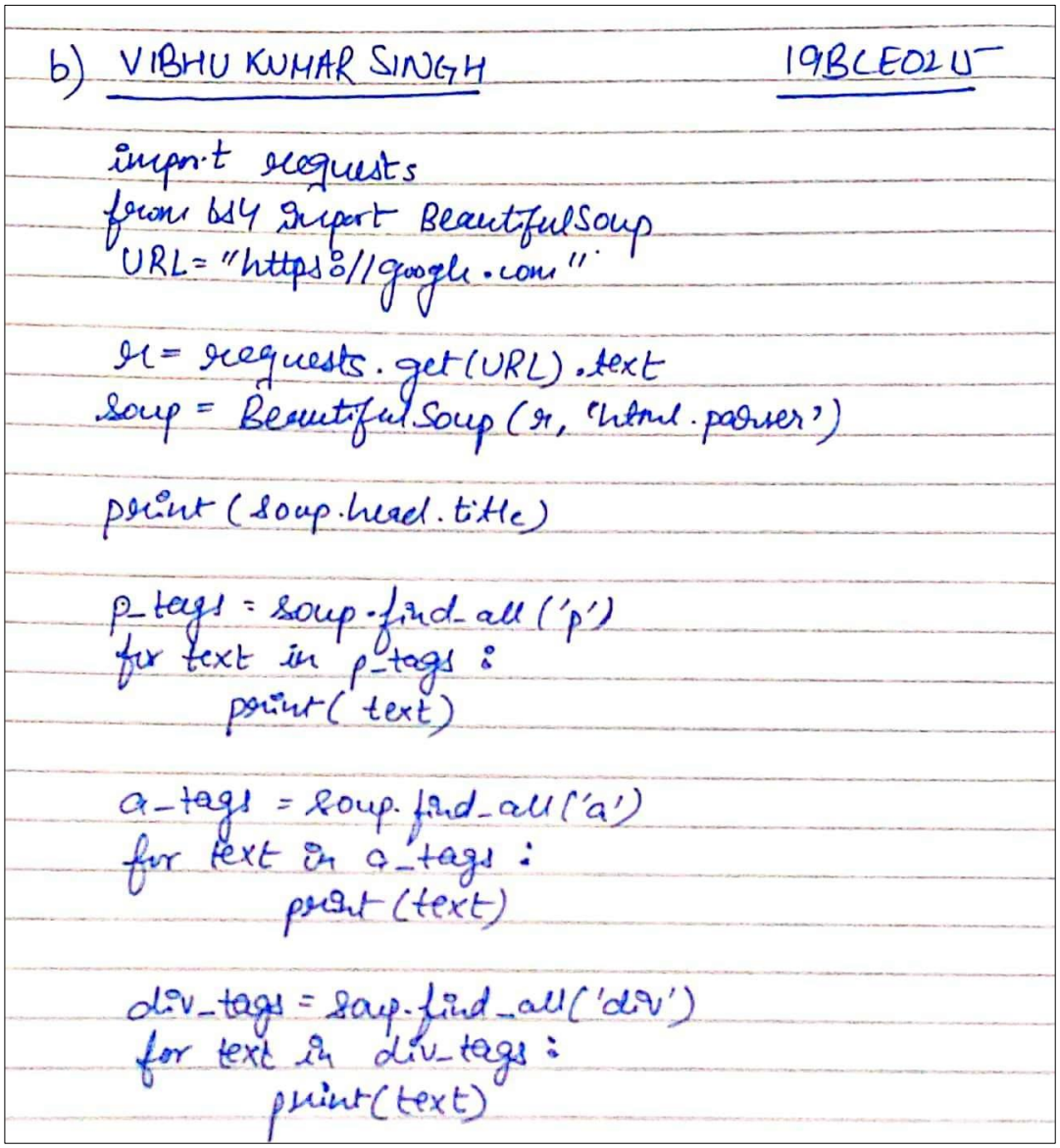
```
<!DOCTYPE html>
<html itemscope="" itemtype="http://schema.org/WebPage" lang="en-IN">
 <head>
  <meta content="text/html; charset=utf-8" http-equiv="Content-Type"/>
  <meta content="/images/branding/googleg/1x/googleg_standard_color_128dp.png" itemprop="image"/>
  <title>
   Google
  </title>
  <script nonce="NkszXHXVWc1k2n/AJX8KMA==">
   (function(){window.google={k…
```

OUTPUT SCREENSHOT:



b) Tags (title, p, a, div)

HANDWRITTEN CODE:


b) VIBHU KUMAR SINGH 19BCE0214

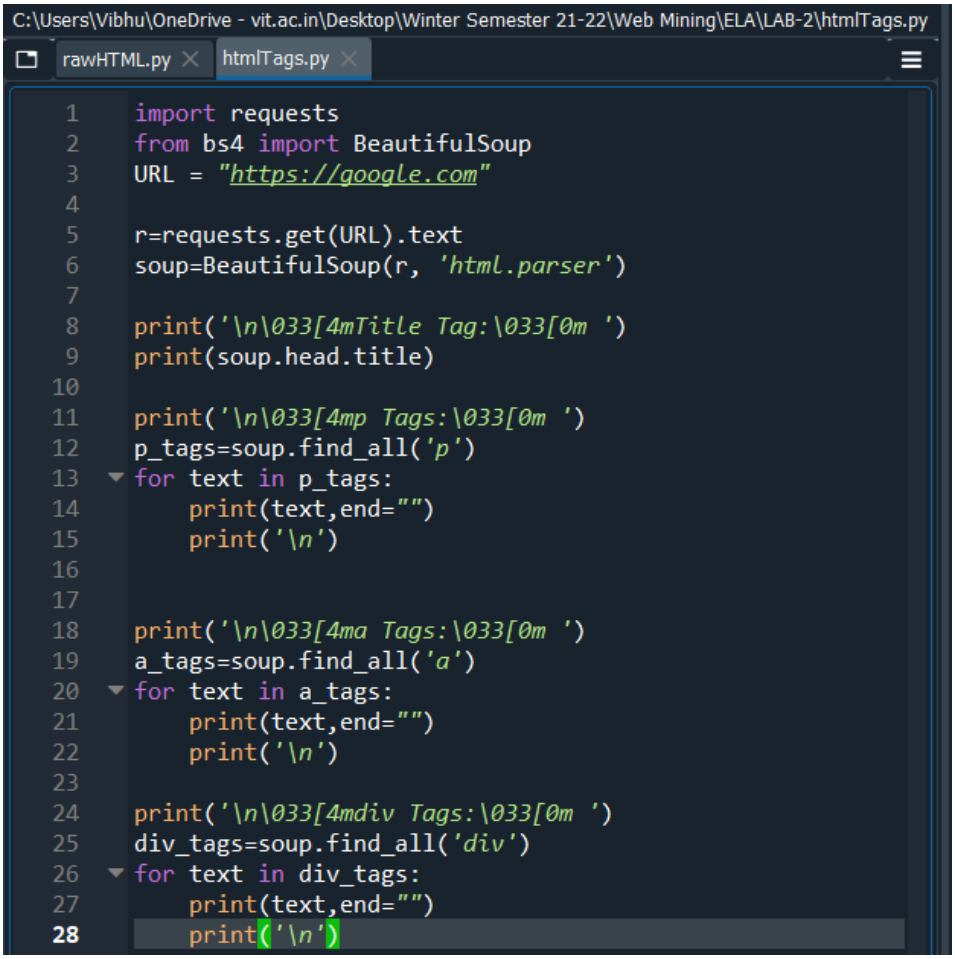
```
import requests
from bs4 import BeautifulSoup
URL = "https://google.com"

r = requests.get(URL).text
soup = BeautifulSoup(r, 'html.parser')

print(soup.head.title)

p_tags = soup.find_all('p')
for text in p_tags:
    print(text)

a_tags = soup.find_all('a')
for text in a_tags:
    print(text)

div_tags = soup.find_all('div')
for text in div_tags:
    print(text)
```

CODE:

```
import requests
from bs4 import BeautifulSoup
URL = "https://google.com"

r=requests.get(URL).text
soup=BeautifulSoup(r, 'html.parser')

print('\n\033[4mTitle Tag:\033[0m ')
print(soup.head.title)

print('\n\033[4mp Tags:\033[0m ')
p_tags=soup.find_all('p')
for text in p_tags:
    print(text,end="")
```

```
    print('\n')

print('\n\033[4ma Tags:\033[0m ')
a_tags=soup.find_all('a')
for text in a_tags:
    print(text,end="")
    print('\n')

print('\n\033[4mdiv Tags:\033[0m ')
div_tags=soup.find_all('div')
for text in div_tags:
    print(text,end="")
    print('\n')


print('\n\033[4mTitle Tag:\033[0m ')
print(soup.head.title)

print('\n\033[4mp Tags:\033[0m ')
p_tags=soup.find_all('p')
for text in p_tags:
    print(text,end="")
    print('\n')

print('\n\033[4ma Tags:\033[0m ')
a_tags=soup.find_all('a')
for text in a_tags:
    print(text,end="")
    print('\n')

print('\n\033[4mdiv Tags:\033[0m ')
div_tags=soup.find_all('div')
for text in div_tags:
    print(text,end="")
    print('\n')
```

CODE SCREENSHOT:



OUTPUT:

Title Tag:
<title>Google</title>

p Tags:
© 2022 - Privacy - Terms

a Tags:
Images
Maps
Play
YouTube
Sign in
Advertising Programs
Business Solutions
About Google
…

div Tags:

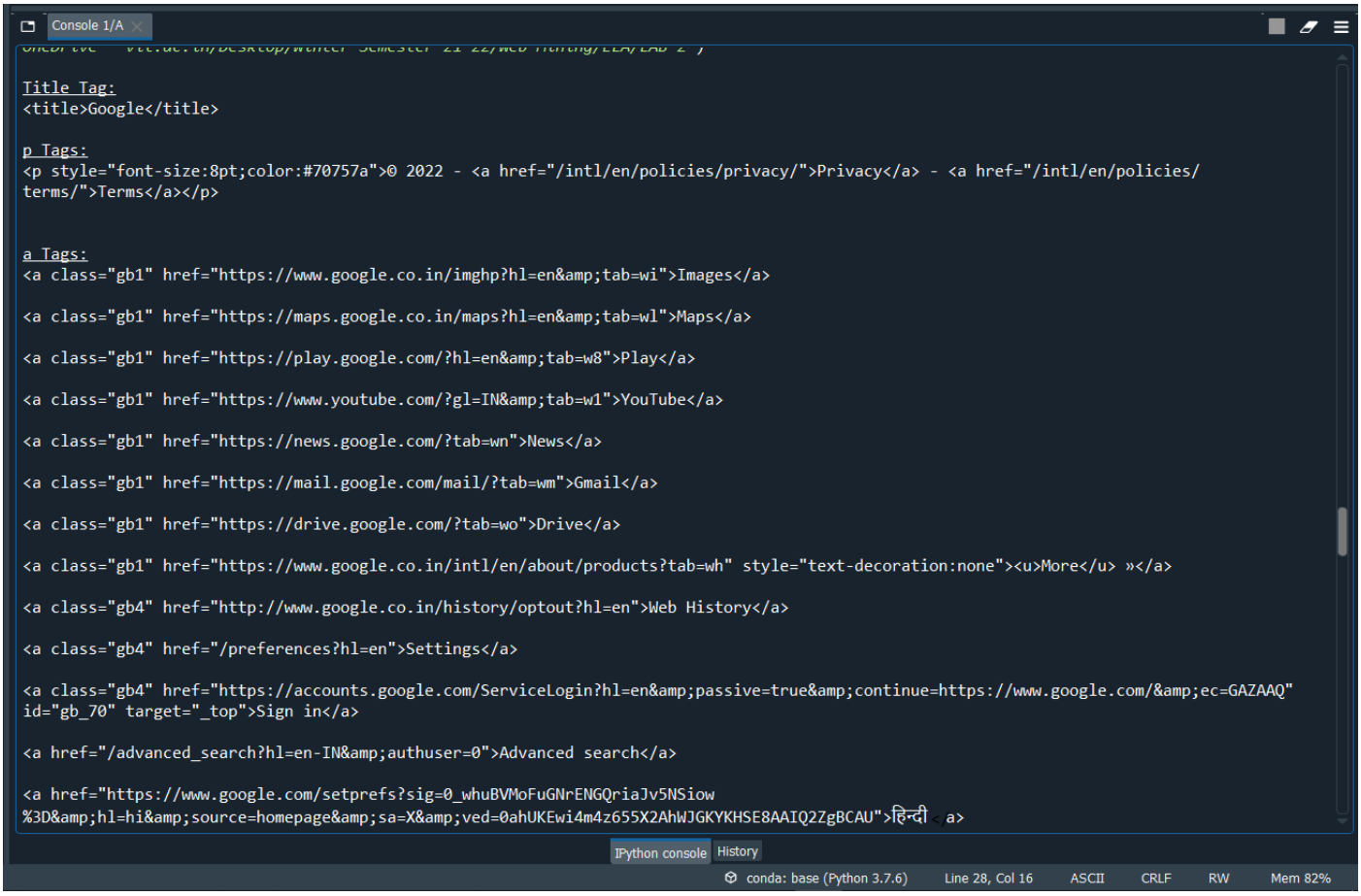
<nobr>Search Images Maps Play YouTube News Gmail Drive <u>More</u> »</nobr> <nobr>Web History | Settings | Sign in</nobr>

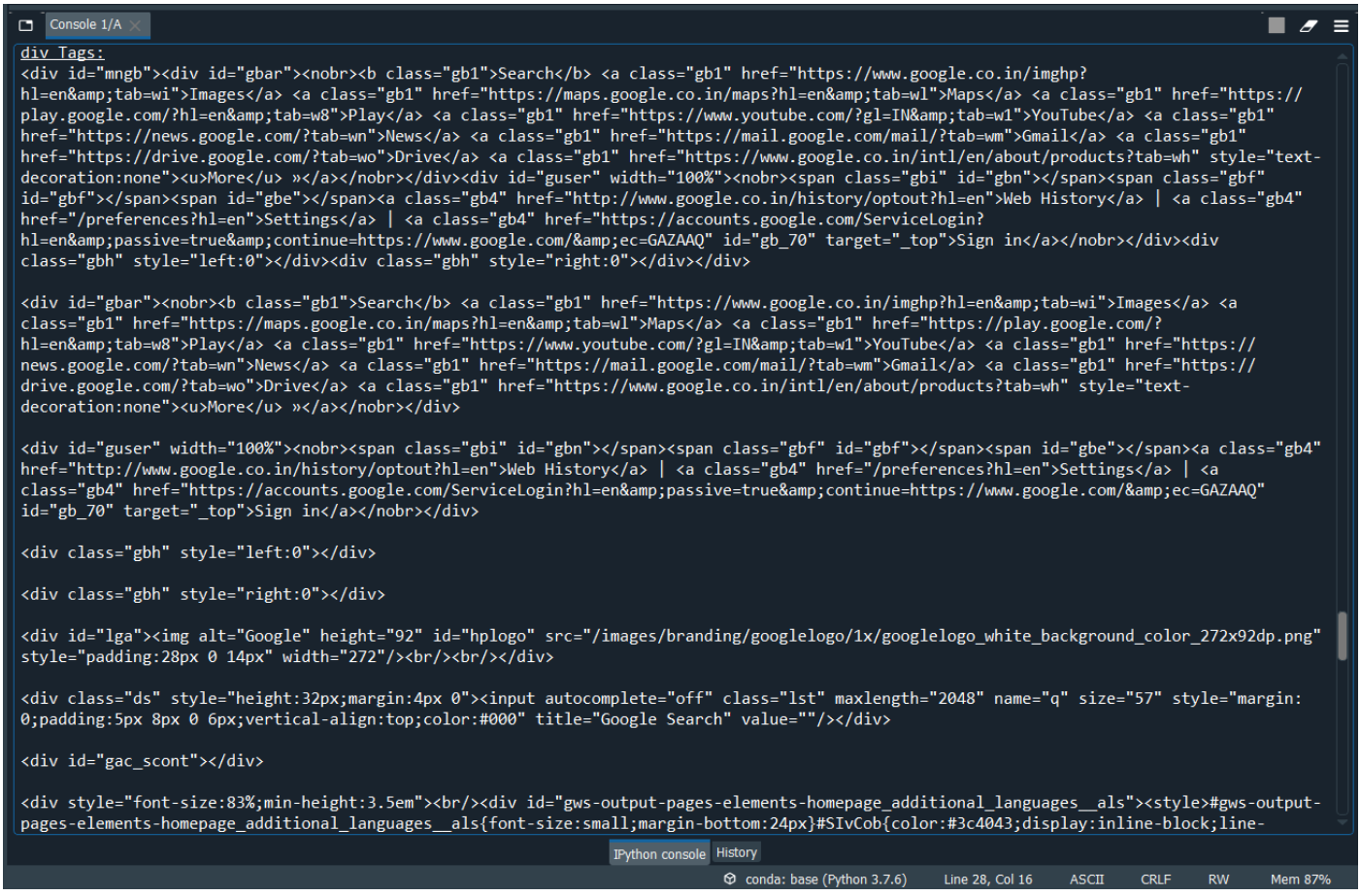
<nobr>Search Images Maps Play YouTube News Gmail Drive <u>More</u> »</nobr>

<nobr>Web History | Settings | Sign in</nobr>

…

OUTPUT SCREENSHOT:


Console 1/A
Title Tag:
<title>Google</title>
p Tags:
© 2022 - Privacy - Terms
a Tags:
Images
Maps
Play
YouTube
News
Gmail
Drive
<u>More</u> »
Web History
Settings
Sign in
Advanced search
हिन्दी
IPython console
History
conda: base (Python 3.7.6)
Line 28, Col 16
ASCII
CRLF
RW
Mem 82%

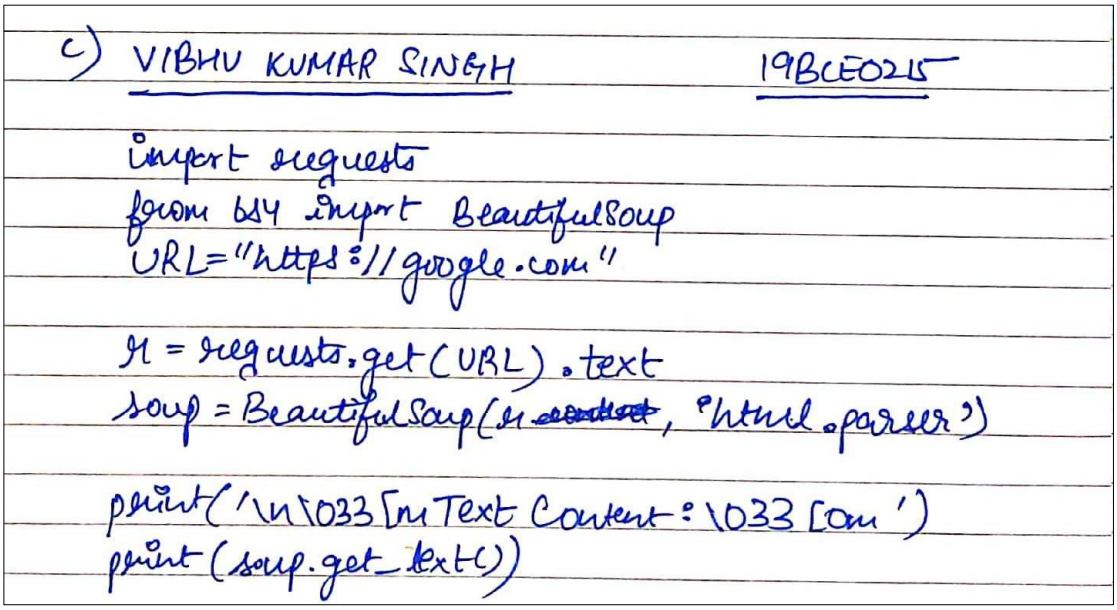

Console 1/A
div Tags:
<nobr>Search Images Maps Play YouTube News Gmail Drive <u>More</u> »</nobr> <nobr>Web History | Settings | Sign in</nobr>
<nobr>Search Images Maps Play YouTube News Gmail Drive <u>More</u> »</nobr>
<nobr>Web History | Settings | Sign in</nobr>

<input autocomplete="off" class="lst" maxlength="2048" name="q" size="57" style="margin:0;padding:5px 8px 0 6px;vertical-align:top;color:#000" title="Google Search" value=""/>

 <style>#gws-output-pages-elements-homepage_additional_languages__als{font-size:small;margin-bottom:24px}#SIvCob{color:#3c4043;display:inline-block;line-
IPython console
History
conda: base (Python 3.7.6)
Line 28, Col 16
ASCII
CRLF
RW
Mem 87%

c) All Textual Content

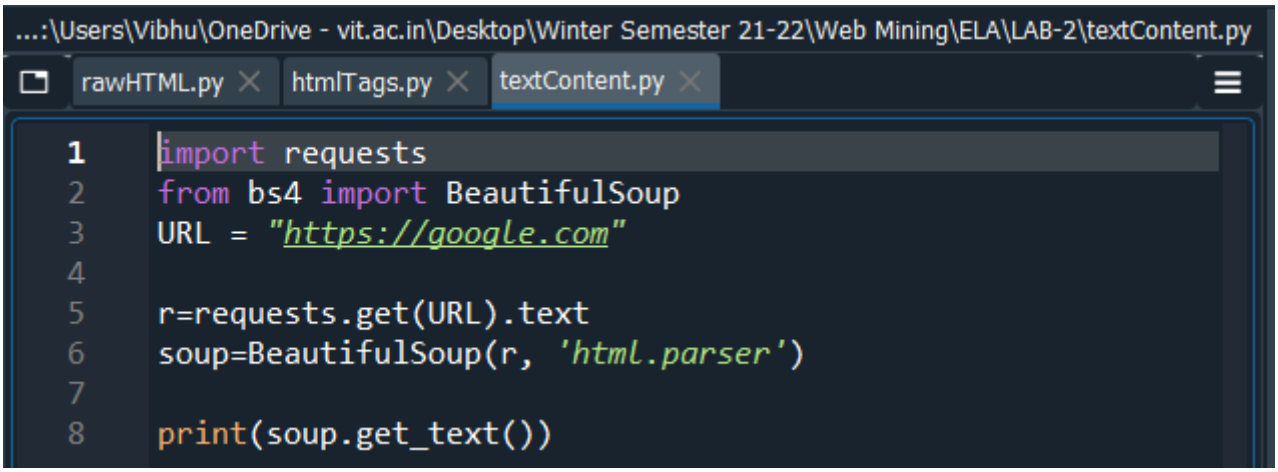
HANDWRITTEN CODE:



CODE:

```
import requests
from bs4 import BeautifulSoup
URL = "https://google.com"

r=requests.get(URL).text
soup=BeautifulSoup(r, 'html.parser')

print(soup.get_text())
```

CODE SCREENSHOT:



OUTPUT:

Text Content:
Google Search Images Maps Play YouTube News Gmail Drive More »Web History | Settings | Sign in Advanced search Google offered in: हिन्दी বাংলা తెలుగు मराठी தமிழ் ગુજરાતી ಕನ್ನಡ മലയാളം ਪੰਜਾਬੀ Advertising Programs Business Solutions About GoogleGoogle.co.in© 2022 - Privacy – Terms

OUTPUT SCREENSHOT:

```
PS C:\Users\Vibhu\OneDrive - vit.ac.in\Desktop\Winter Semester 21
-22\Web Mining\ELA\LAB-2> python -u "c:\Users\Vibhu\OneDrive - vi
t.ac.in\Desktop\Winter Semester 21-22\Web Mining\ELA\LAB-2\textCo
ntent.py"

Text Content:
GoogleSearch Images Maps Play YouTube News Gmail Drive More »Web
History | Settings | Sign in Advanced searchGoogle offered in:  ह
िन्दी বাংলা తెలుగు मराठी தமிழ் ગુજરાતી ಕನ್ನಡ മലയാളം ਪੰਜਾਬੀ Adverte
ising ProgramsBusiness SolutionsAbout GoogleGoogle.co.in© 2022 -
Privacy - Terms
```