

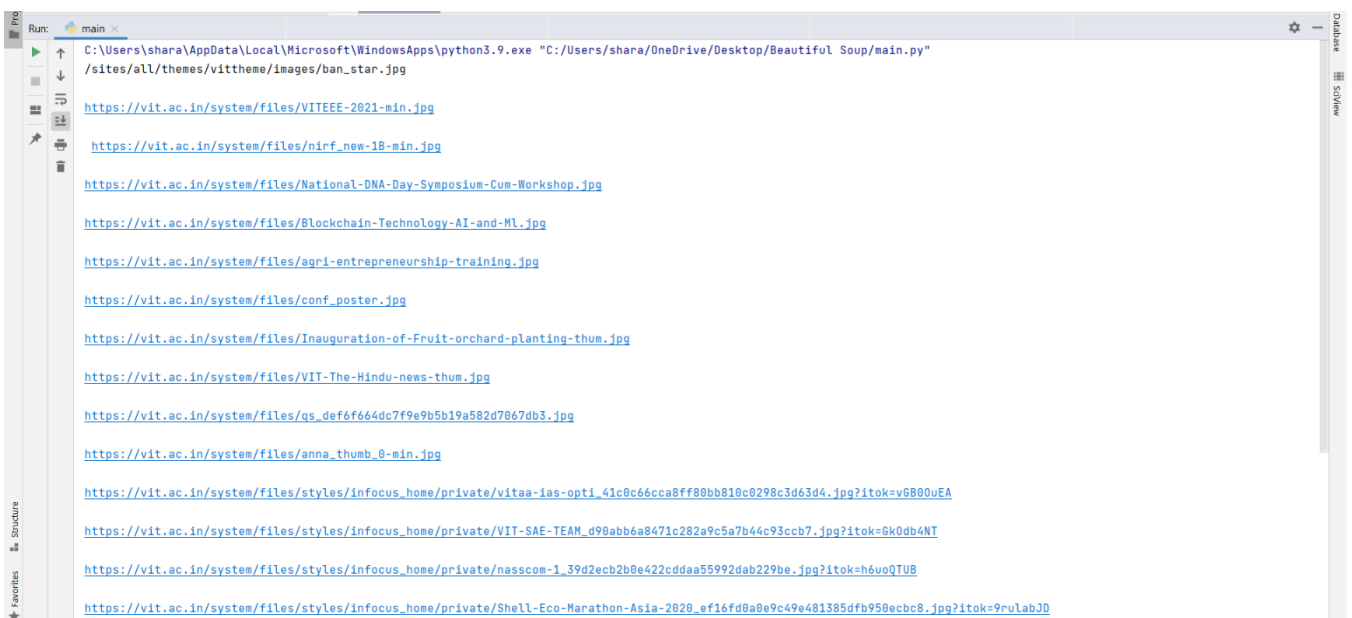
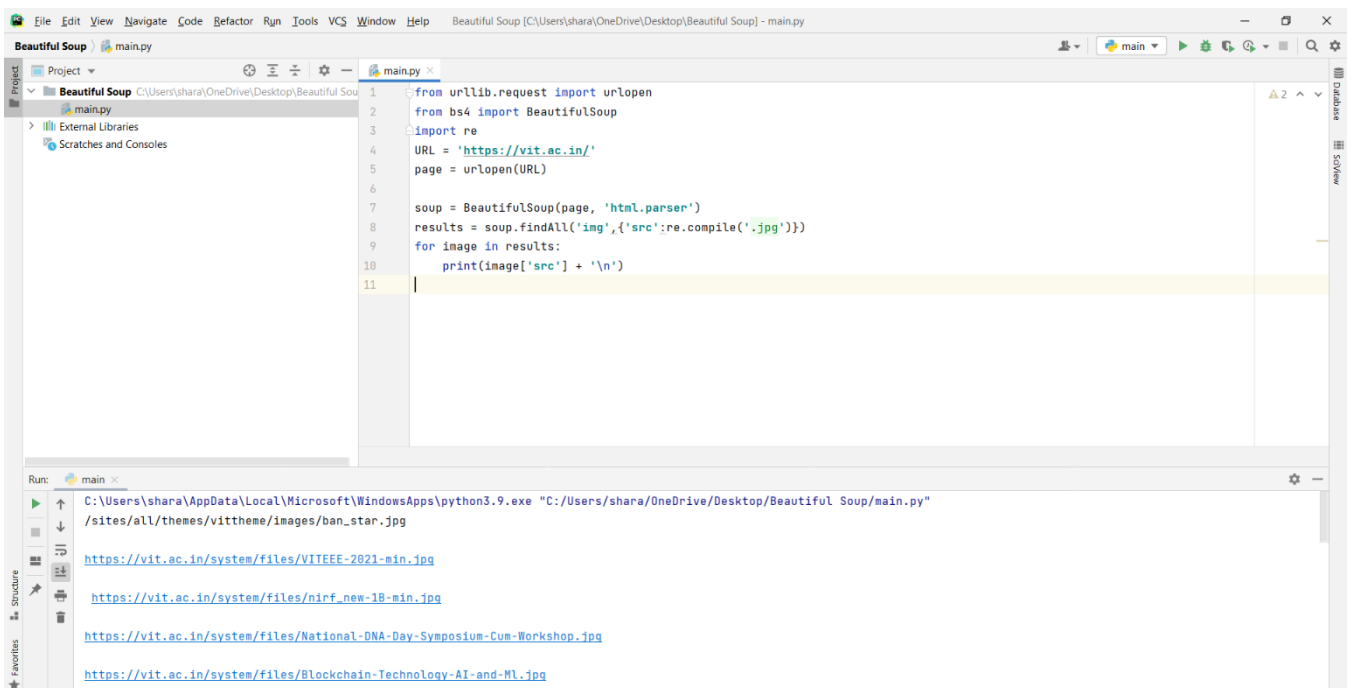
WEB MINING

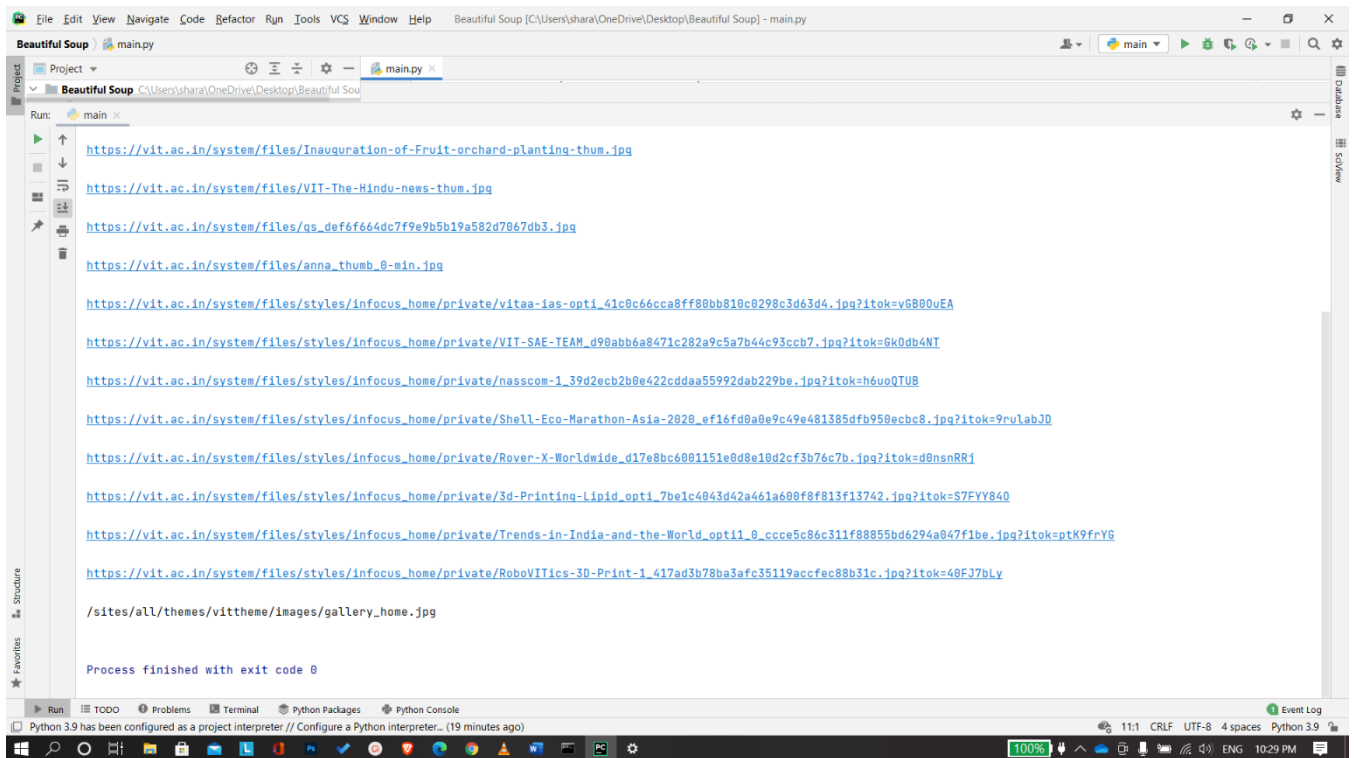
by Sharadindu Adhikari, 19BCE2105

BEAUTIFUL SOUP**• Extracting Image URLs (IMG TAG) from vit.ac.in**

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import re
URL = 'https://vit.ac.in/'
page = urlopen(URL)

soup = BeautifulSoup(page, 'html.parser')
results = soup.findAll('img', {'src': re.compile('.jpg')})
for image in results:
    print(image['src'] + '\n')
```

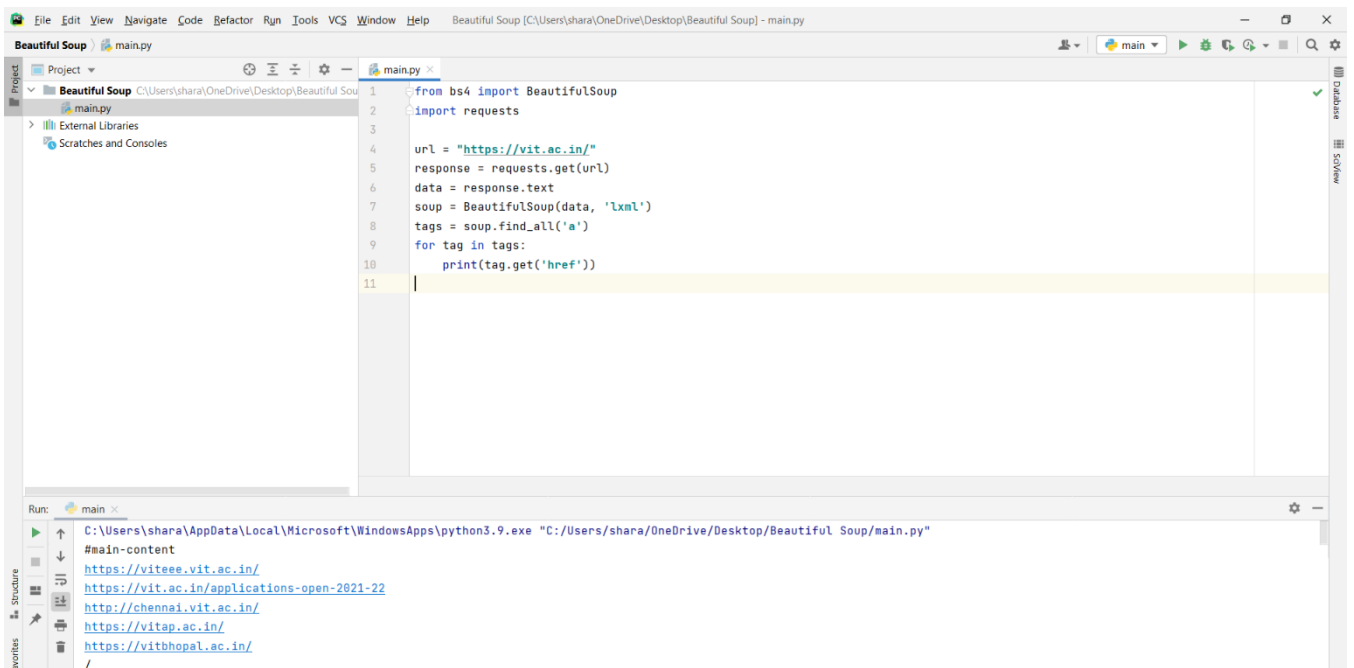




- **Extracting URLs to other websites (HREF TAG) from vit.ac.in**

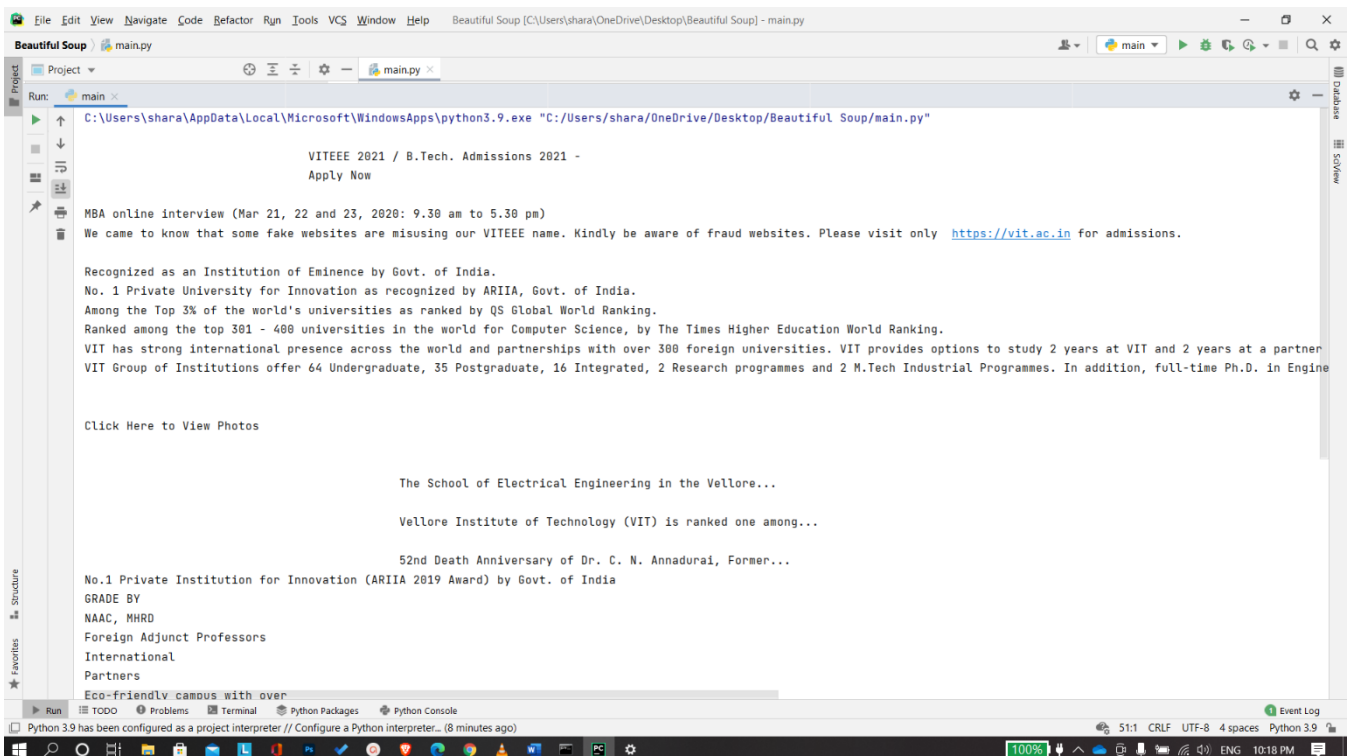
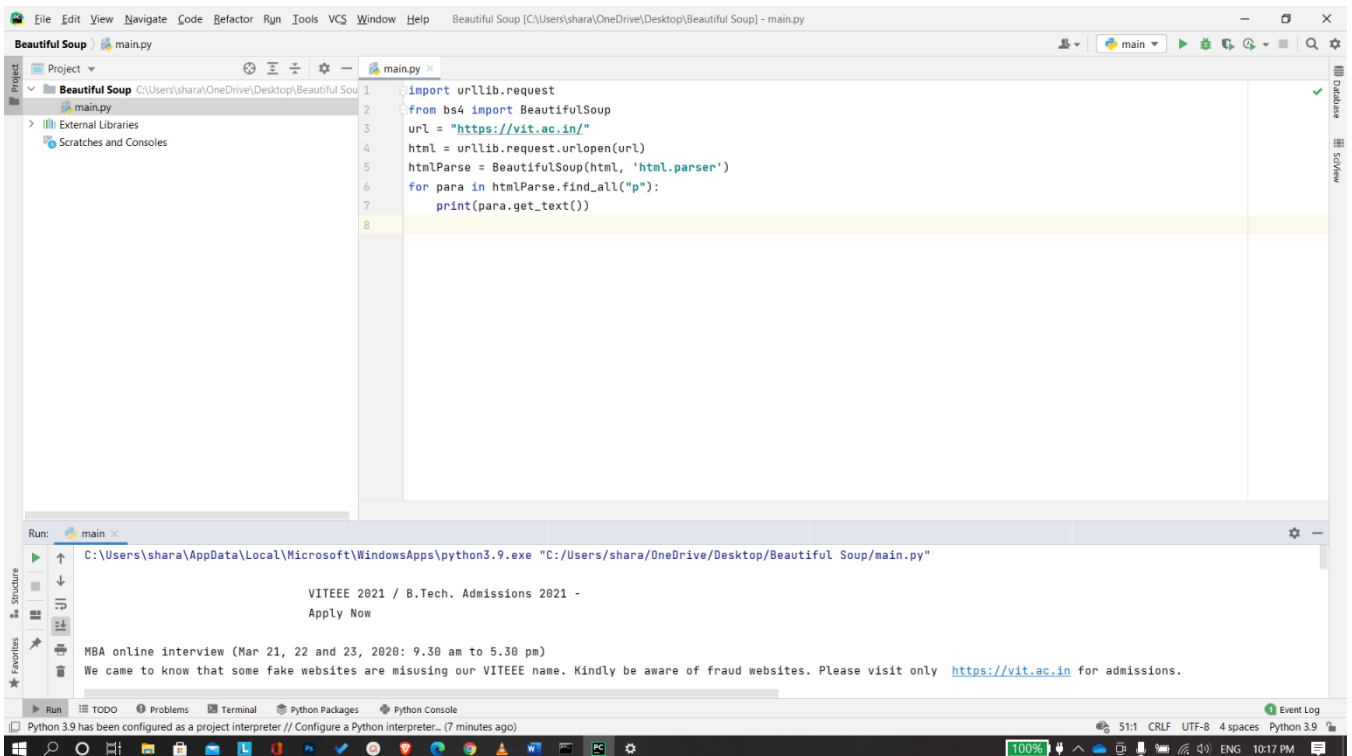
```
from bs4 import BeautifulSoup
import requests

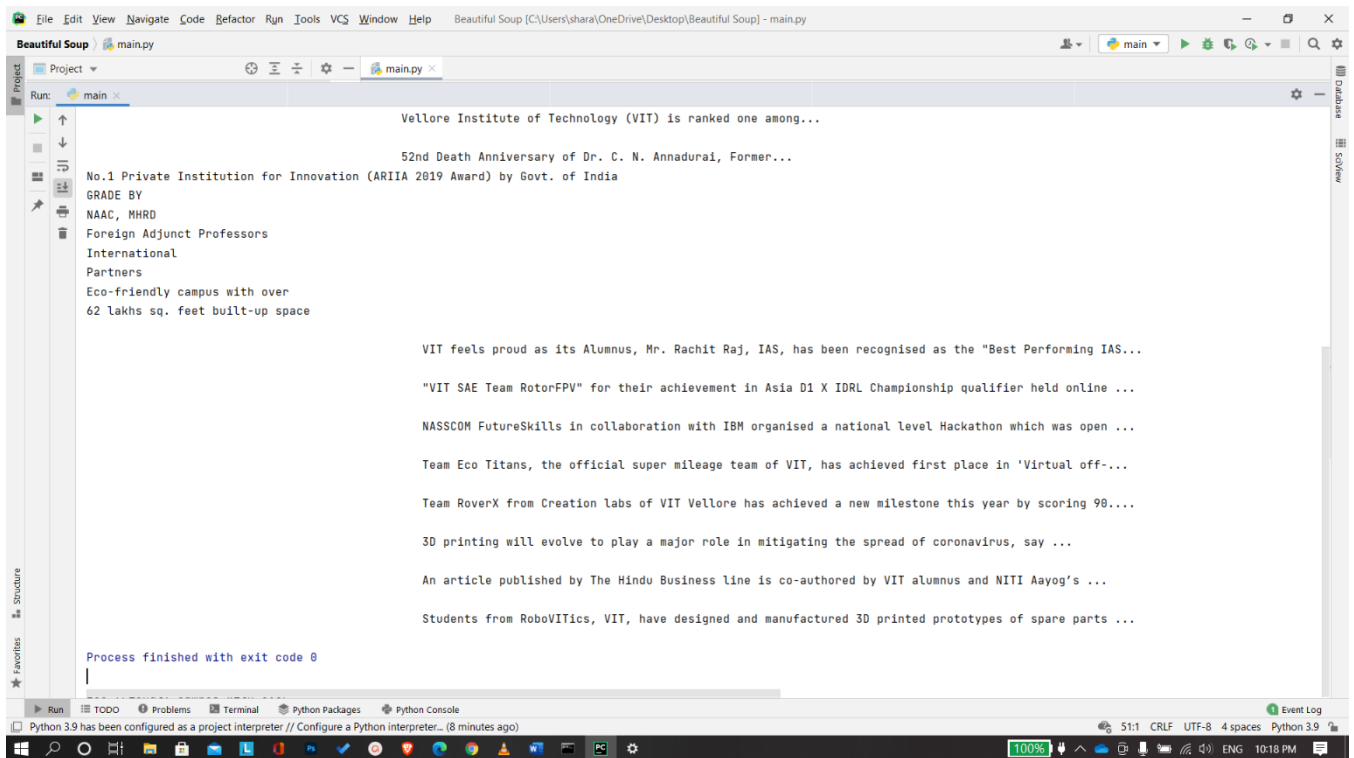
url = "https://vit.ac.in/"
response = requests.get(url)
data = response.text
soup = BeautifulSoup(data, 'lxml')
tags = soup.find_all('a')
for tag in tags:
    print(tag.get('href'))
```



- Extracting paragraphs (P tag) from vit.ac.in

```
import urllib.request
from bs4 import BeautifulSoup
url = "https://vit.ac.in/"
html = urllib.request.urlopen(url)
htmlParse = BeautifulSoup(html, 'html.parser')
for para in htmlParse.find_all("p"):
    print(para.get_text())
```



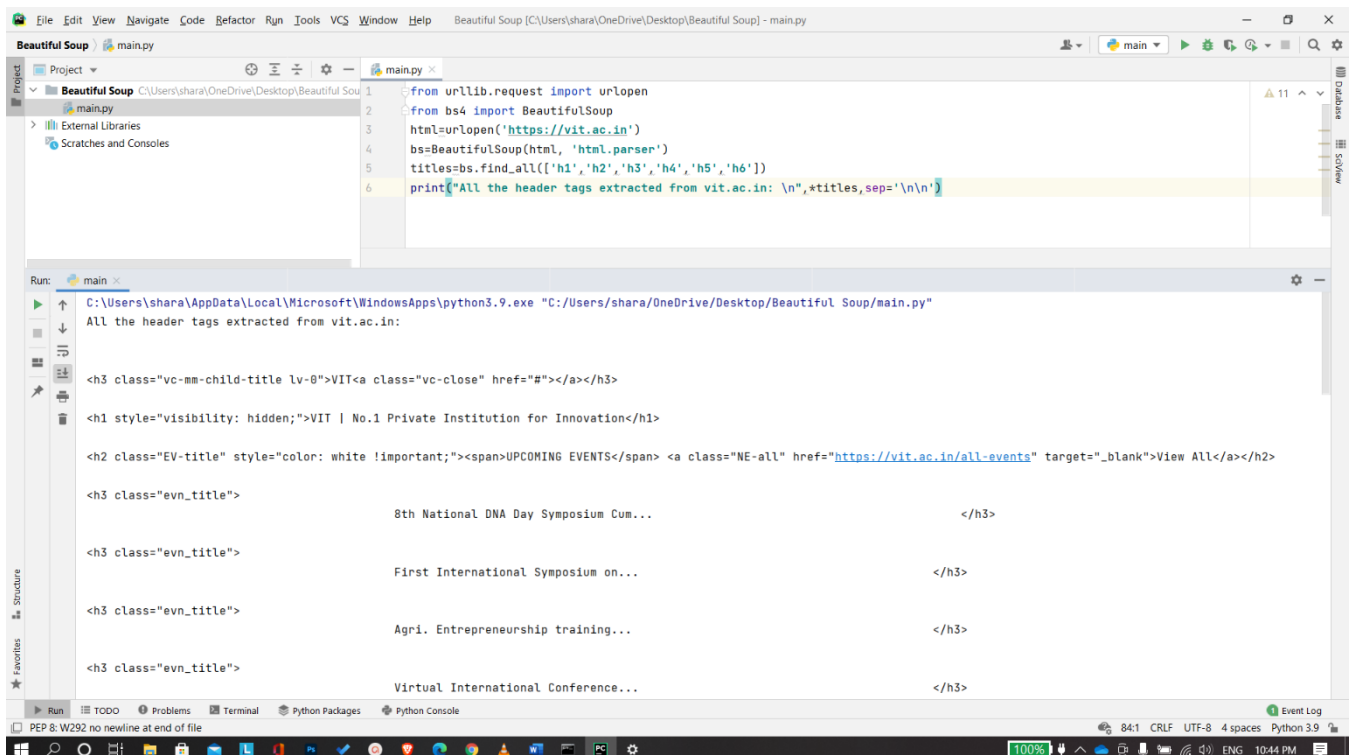


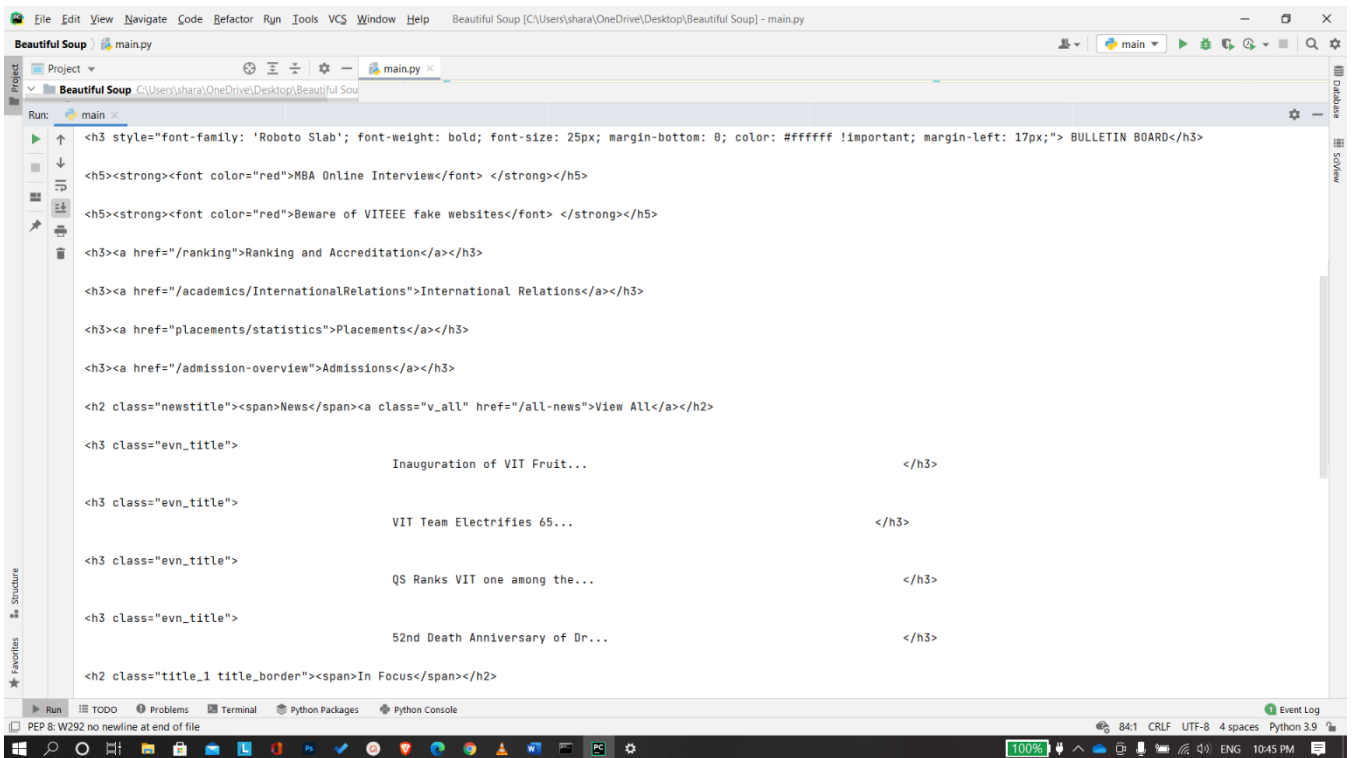
- **Extracting Header tags from vit.ac.in**

```

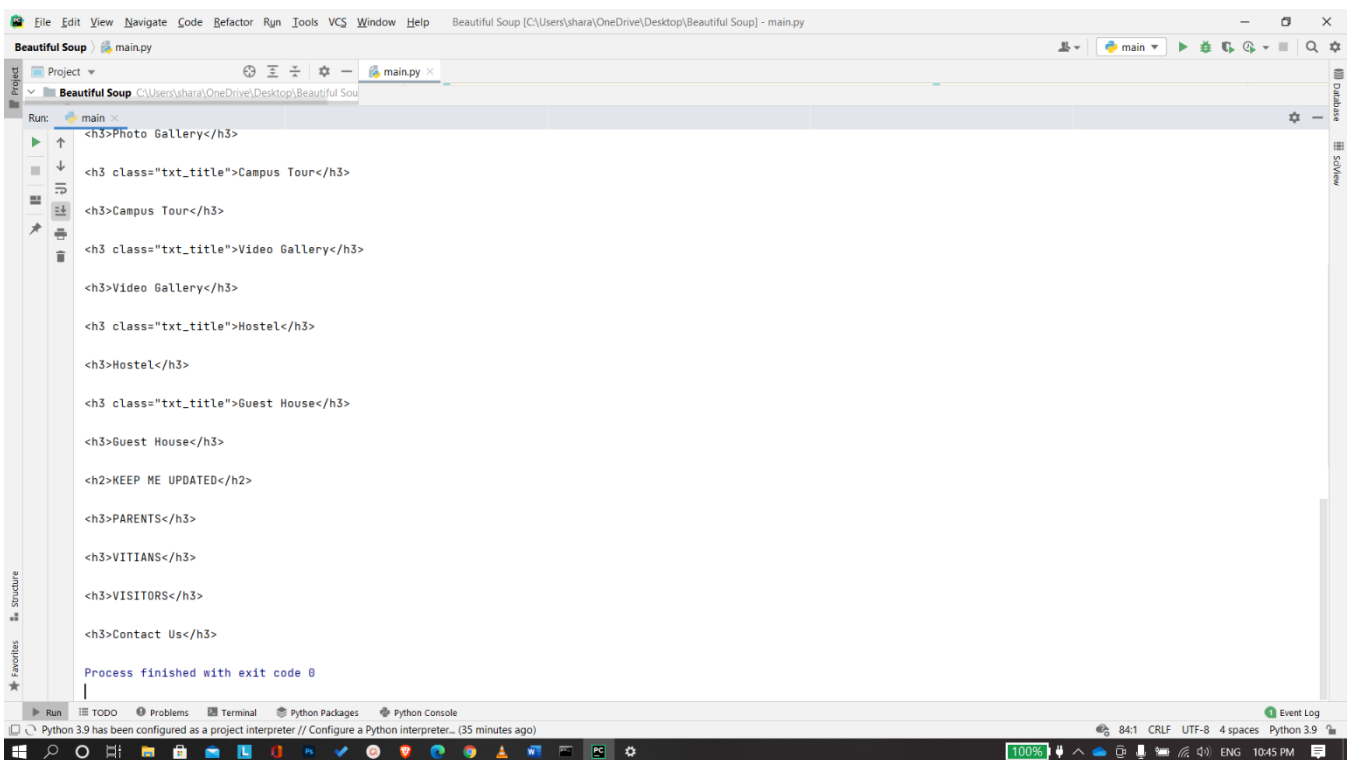
from urllib.request import urlopen
from bs4 import BeautifulSoup
html=urlopen('https://vit.ac.in')
bs=BeautifulSoup(html, 'html.parser')
titles=bs.find_all(['h1', 'h2', 'h3', 'h4', 'h5', 'h6'])
print("All the header tags extracted from vit.ac.in: \n",*titles,sep='\n\n')

```





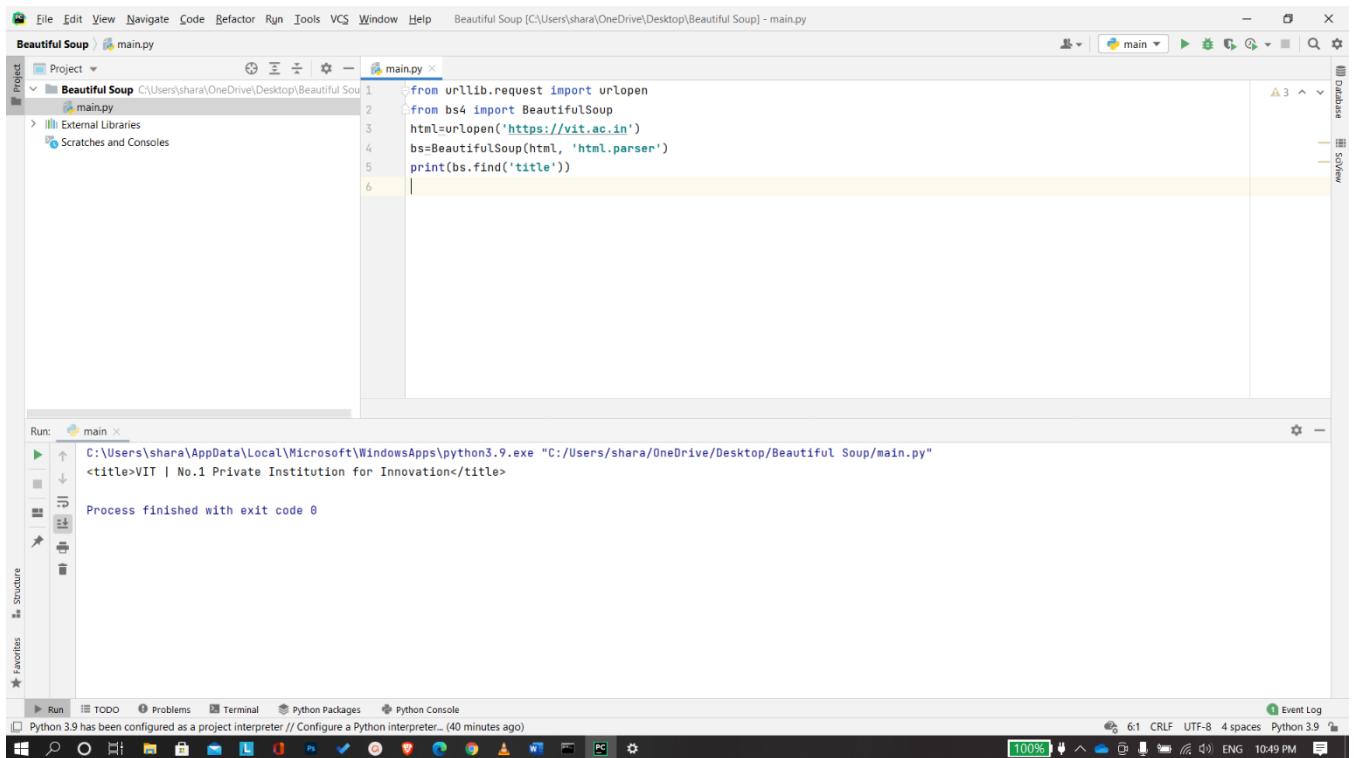
```
File Edit View Navigate Code Refactor Run Tools VCS Window Help Beautiful Soup [C:\Users\shara\OneDrive\Desktop\Beautiful Soup] - main.py
Beautiful Soup main.py
Project Beautiful Soup C:\Users\shara\OneDrive\Desktop\Beautiful Sou
Run main
<h3 style="font-family: 'Roboto Slab'; font-weight: bold; font-size: 25px; margin-bottom: 0; color: #ffffff !important; margin-left: 17px;"> BULLETIN BOARD</h3>
<h5><strong><font color="red">MBA OnLine Interview</font> </strong></h5>
<h5><strong><font color="red">Beware of VITEEE fake websites</font> </strong></h5>
<h3><a href="/ranking">Ranking and Accreditation</a></h3>
<h3><a href="/academics/InternationalRelations">International Relations</a></h3>
<h3><a href="placements/statistics">Placements</a></h3>
<h3><a href="/admission-overview">Admissions</a></h3>
<h2 class="newstitle"><span>News</span><a class="v_all" href="/all-news">View All</a></h2>
<h3 class="evn_title">
Inauguration of VIT Fruit... </h3>
<h3 class="evn_title">
VIT Team Electrifies 65... </h3>
<h3 class="evn_title">
QS Ranks VIT one among the... </h3>
<h3 class="evn_title">
52nd Death Anniversary of Dr... </h3>
<h2 class="title_1 title_border"><span>In Focus</span></h2>
Run TODO Problems Terminal Python Packages Python Console
PEP 8: W292 no newline at end of file
84:1 CRLF UTF-8 4 spaces Python 3.9 100% ENG 10:45 PM
```



```
File Edit View Navigate Code Refactor Run Tools VCS Window Help Beautiful Soup [C:\Users\shara\OneDrive\Desktop\Beautiful Soup] - main.py
Beautiful Soup main.py
Project Beautiful Soup C:\Users\shara\OneDrive\Desktop\Beautiful Sou
Run main
<h3>Photo Gallery</h3>
<h3 class="txt_title">Campus Tour</h3>
<h3>Campus Tour</h3>
<h3 class="txt_title">Video Gallery</h3>
<h3>Video Gallery</h3>
<h3 class="txt_title">Hostel</h3>
<h3>Hostel</h3>
<h3 class="txt_title">Guest House</h3>
<h3>Guest House</h3>
<h2>KEEP ME UPDATED</h2>
<h3>PARENTS</h3>
<h3>VITIANS</h3>
<h3>VISITORS</h3>
<h3>Contact Us</h3>
Process finished with exit code 0
Run TODO Problems Terminal Python Packages Python Console
Python 3.9 has been configured as a project interpreter // Configure a Python interpreter... (35 minutes ago)
84:1 CRLF UTF-8 4 spaces Python 3.9 100% ENG 10:45 PM
```

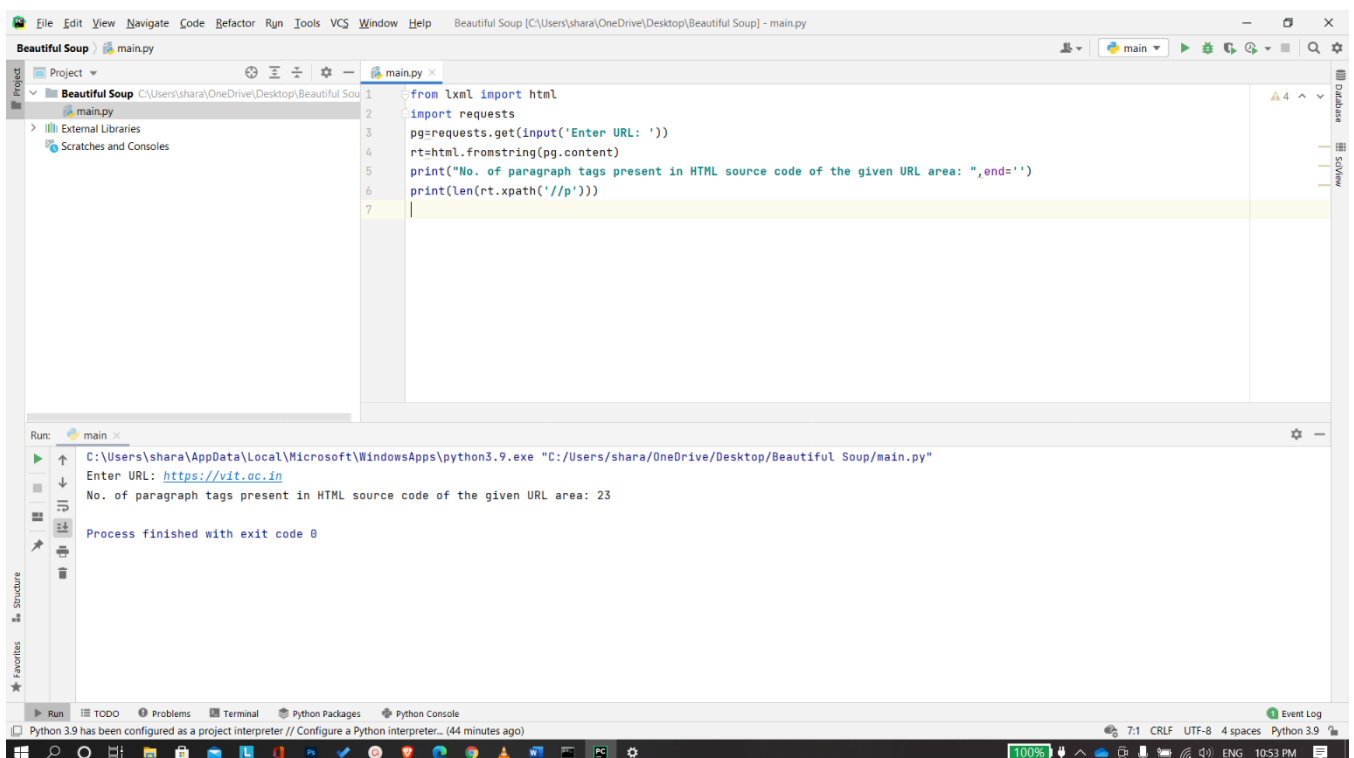
- **Extracting all title tags from vit.ac.in**

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html=urlopen('https://vit.ac.in')
bs=BeautifulSoup(html, 'html.parser')
print(bs.find('title'))
```



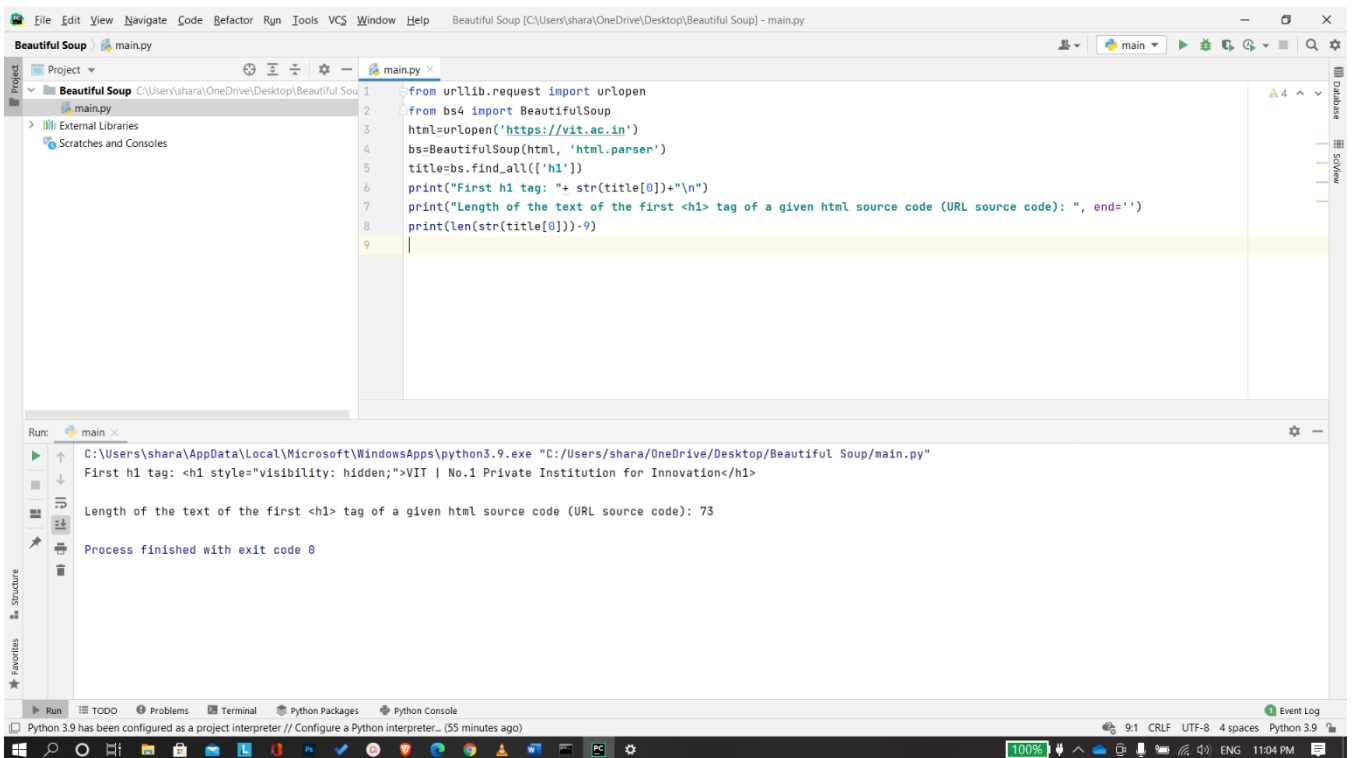
- Count the number of paragraph tags of an HTML document

```
from lxml import html
import requests
pg=requests.get(input('Enter URL: '))
rt=html.fromstring(pg.content)
print("No. of paragraph tags present in HTML source code of the given URL area: ",end='')
print(len(rt.xpath('//p')))
```



- Find the length of the text of the first <h1> tag of a given HTML document

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html=urlopen('https://vit.ac.in')
bs=BeautifulSoup(html, 'html.parser')
title=bs.find_all(['h1'])
print("First h1 tag: "+ str(title[0])+"\n")
print("Length of the text of the first <h1> tag of a given html source code (URL source code): ", end='')
print(len(str(title[0]))-9)
```



The screenshot shows a Python IDE window titled 'Beautiful Soup' with a file named 'main.py'. The code in the editor is as follows:

```
1 from urllib.request import urlopen
2 from bs4 import BeautifulSoup
3 html=urlopen('https://vit.ac.in')
4 bs=BeautifulSoup(html, 'html.parser')
5 title=bs.find_all(['h1'])
6 print("First h1 tag: "+ str(title[0])+"\n")
7 print("Length of the text of the first <h1> tag of a given html source code (URL source code): ", end='')
8 print(len(str(title[0]))-9)
```

The output window shows the following results:

```
First h1 tag: <h1 style="visibility: hidden;">VIT | No.1 Private Institution for Innovation</h1>
Length of the text of the first <h1> tag of a given html source code (URL source code): 73
Process finished with exit code 0
```

The status bar at the bottom indicates the Python 3.9 interpreter is configured and the file encoding is UTF-8.