Web mining LAB Assignment -2 [Prof.Ranganthan Sridhar]

Try various options available using Beautiful soup package and using website www.vit.ac.in

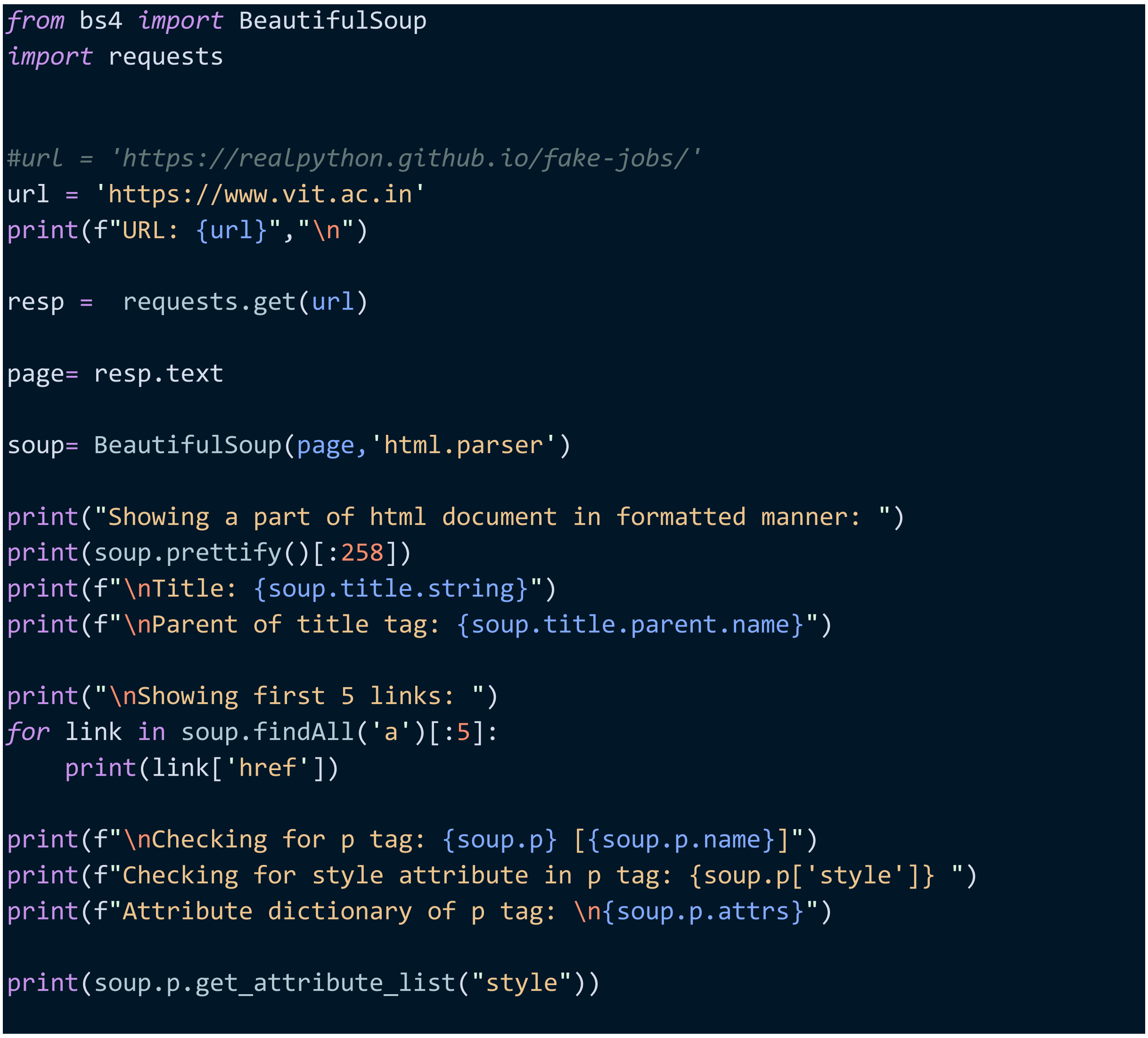
Please capture the output you have got into a word document, convert to a pdf and upload.

CAUTION : DO NOT TRY TO UPDATE OR DELETE ANYTHING IN OUR WEBSITE. IF ANY VIOLATION IS DONE, YOU WILL BE PENALISED.

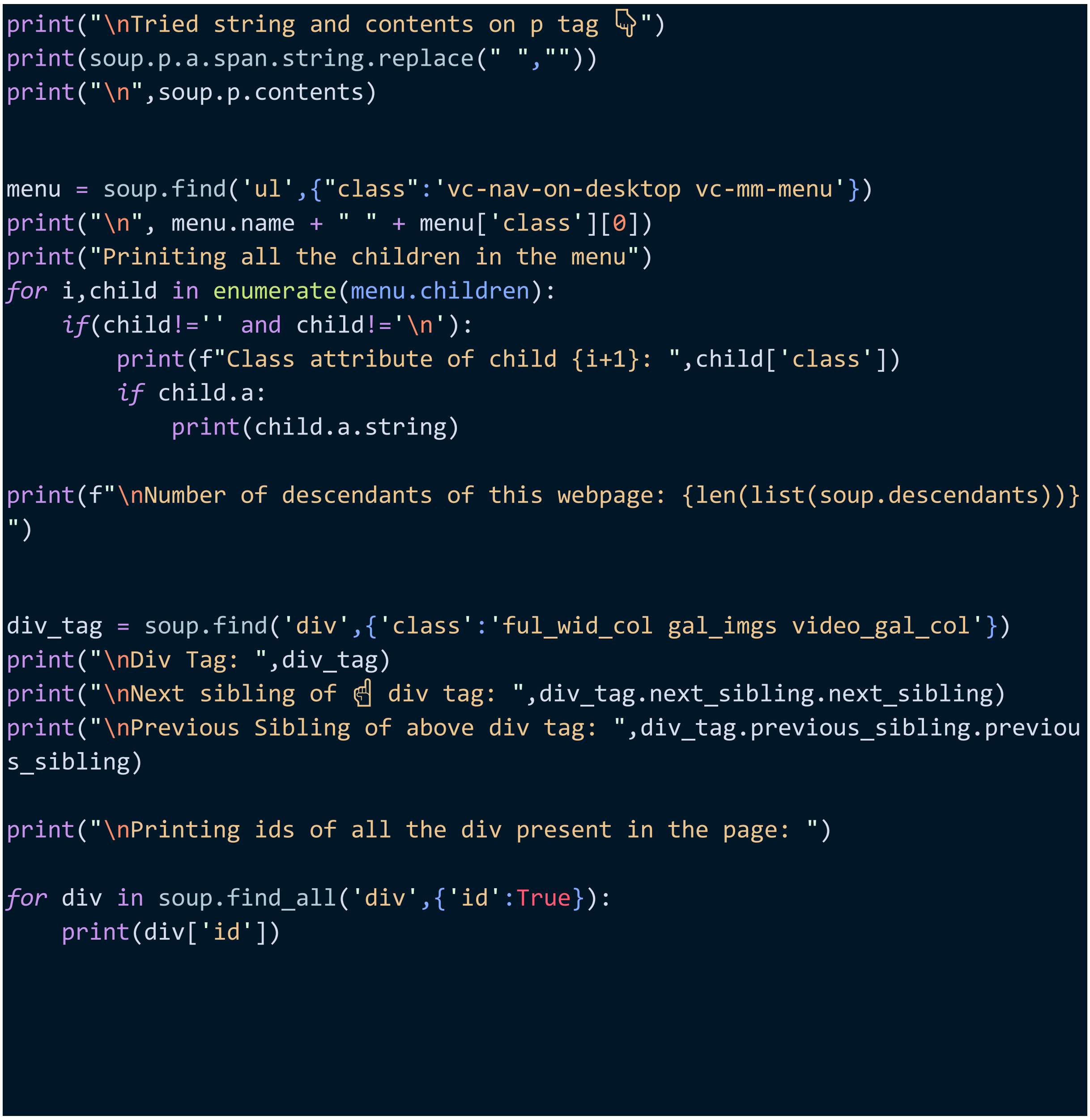
Name: Gerosh George

Reg Number: 19BCE1403

CODE:

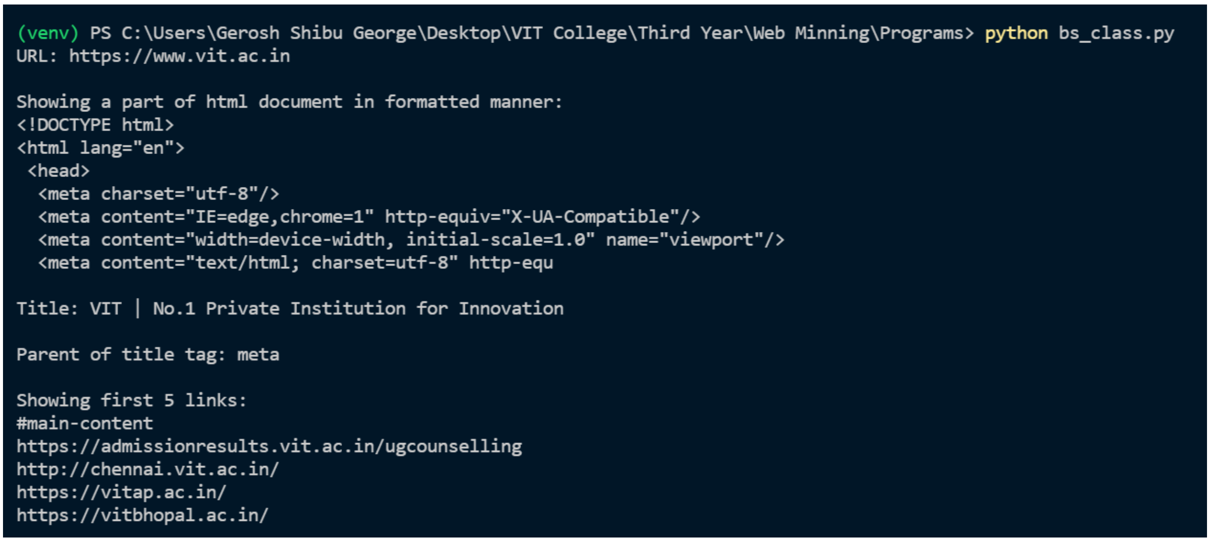


```python
from bs4 import BeautifulSoup
import requests


#url = 'https://realpython.github.io/fake-jobs/'
url = 'https://www.vit.ac.in'
print(f"URL: {url}","\n")

resp =  requests.get(url)

page= resp.text

soup= BeautifulSoup(page,'html.parser')

print("Showing a part of html document in formatted manner: ")
print(soup.prettify()[:258])
print(f"\nTitle: {soup.title.string}")
print(f"\nParent of title tag: {soup.title.parent.name}")

print("\nShowing first 5 links: ")
for link in soup.findAll('a')[:5]:
    print(link['href'])

print(f"\nChecking for p tag: {soup.p} [{soup.p.name}]")
print(f"Checking for style attribute in p tag: {soup.p['style']} ")
print(f"Attribute dictionary of p tag: \n{soup.p.attrs}")

print(soup.p.get_attribute_list("style"))
```

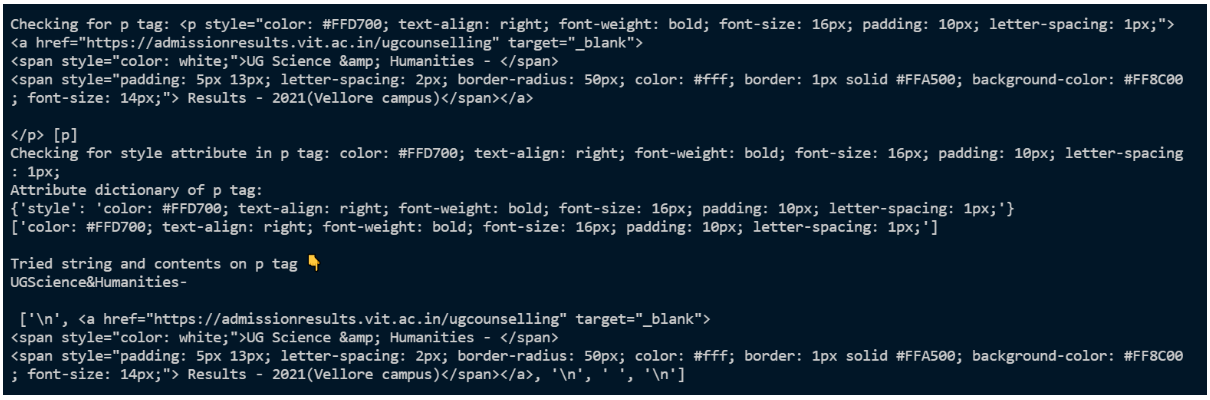


```
print("\nTried string and contents on p tag 👇")
print(soup.p.a.span.string.replace(" ",""))
print("\n",soup.p.contents)


menu = soup.find('ul',{"class":'vc-nav-on-desktop vc-mm-menu'})
print("\n", menu.name + " " + menu['class'][0])
print("Priniting all the children in the menu")
for i,child in enumerate(menu.children):
    if(child!='' and child!='\n'):
        print(f"Class attribute of child {i+1}: ",child['class'])
        if child.a:
            print(child.a.string)

print(f"\nNumber of descendants of this webpage: {len(list(soup.descendants))}")


div_tag = soup.find('div',{'class':'ful_wid_col gal_imgs video_gal_col'})
print("\nDiv Tag: ",div_tag)
print("\nNext sibling of ☝ div tag: ",div_tag.next_sibling.next_sibling)
print("\nPrevious Sibling of above div tag: ",div_tag.previous_sibling.previous_sibling)

print("\nPrinting ids of all the div present in the page: ")

for div in soup.find_all('div',{'id':True}):
    print(div['id'])
```

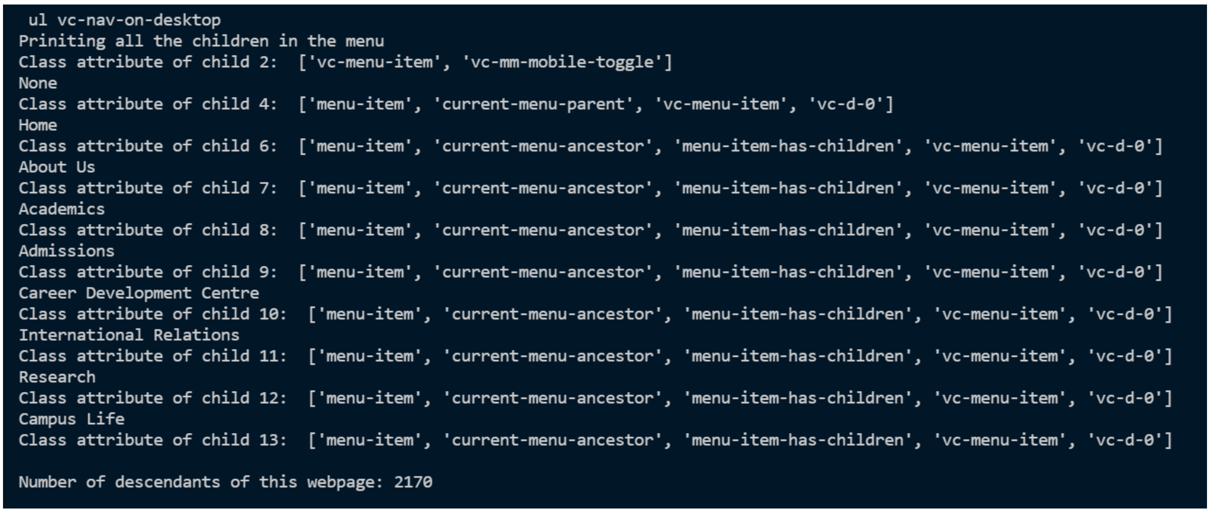
OUTPUT:



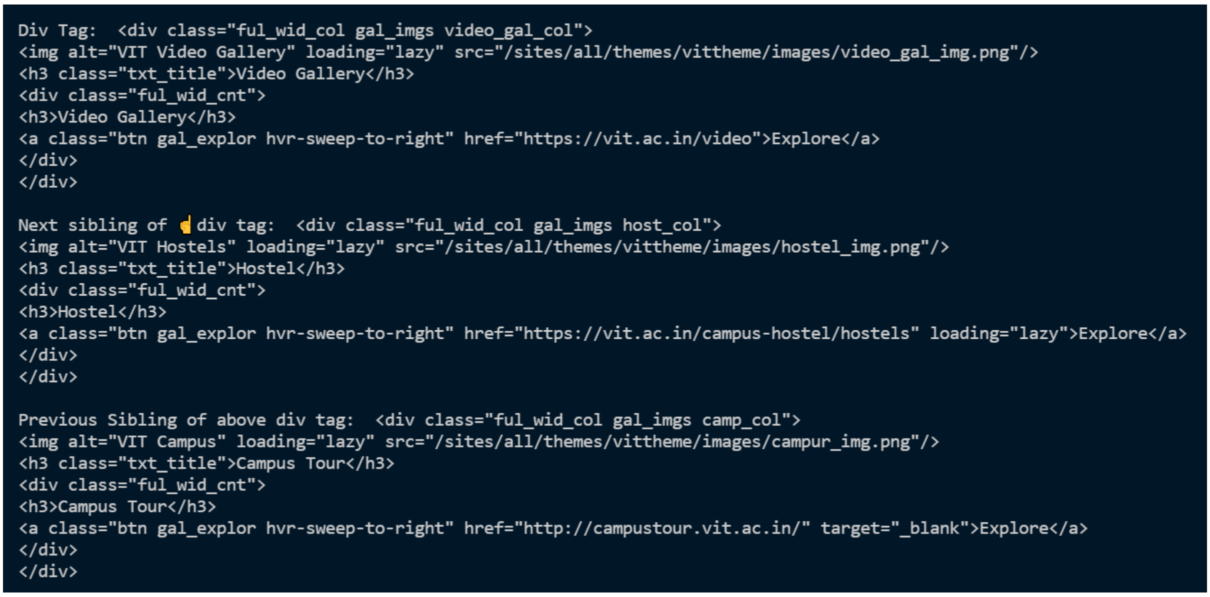
```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Minning\Programs> python bs_class.py
URL: https://www.vit.ac.in

Showing a part of html document in formatted manner:
<!DOCTYPE html>
<html lang="en">
 <head>
  <meta charset="utf-8"/>
  <meta content="IE=edge,chrome=1" http-equiv="X-UA-Compatible"/>
  <meta content="width=device-width, initial-scale=1.0" name="viewport"/>
  <meta content="text/html; charset=utf-8" http-equ

Title: VIT | No.1 Private Institution for Innovation

Parent of title tag: meta

Showing first 5 links:
#main-content
https://admissionresults.vit.ac.in/ugcounselling
http://chennai.vit.ac.in/
https://vitap.ac.in/
https://vitbhopal.ac.in/
```



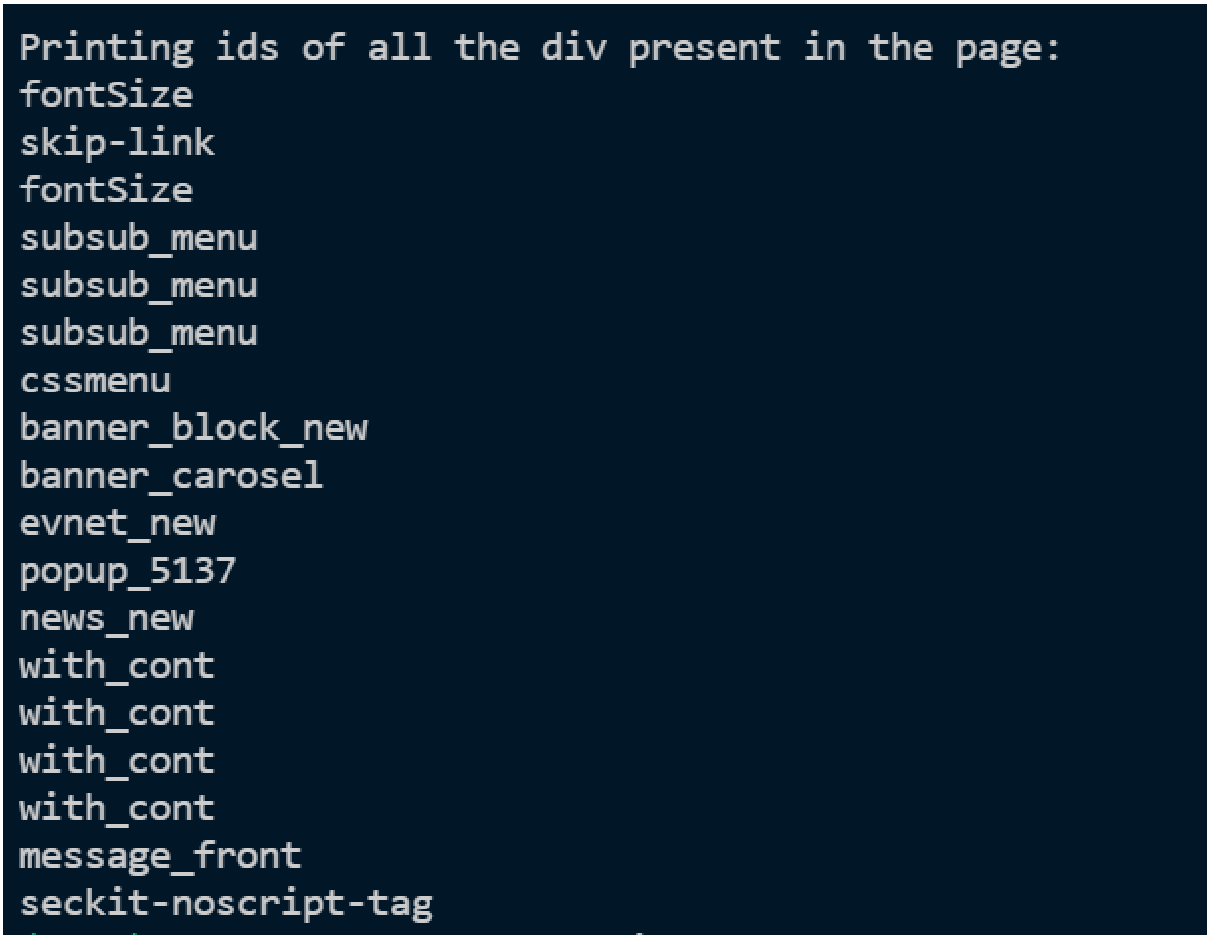
```
Checking for p tag: <p style="color: #FFD700; text-align: right; font-weight: bold; font-size: 16px; padding: 10px; letter-spacing: 1px;">
<a href="https://admissionresults.vit.ac.in/ugcounselling" target="_blank">
<span style="color: white;">UG Science &amp; Humanities - </span>
<span style="padding: 5px 13px; letter-spacing: 2px; border-radius: 50px; color: #fff; border: 1px solid #FFA500; background-color: #FF8C00
; font-size: 14px;"> Results - 2021(Vellore campus)</span></a>

</p> [p]
Checking for style attribute in p tag: color: #FFD700; text-align: right; font-weight: bold; font-size: 16px; padding: 10px; letter-spacing
: 1px;
Attribute dictionary of p tag:
{'style': 'color: #FFD700; text-align: right; font-weight: bold; font-size: 16px; padding: 10px; letter-spacing: 1px;'}
['color: #FFD700; text-align: right; font-weight: bold; font-size: 16px; padding: 10px; letter-spacing: 1px;']

Tried string and contents on p tag 👇
UGScience&Humanities-

 ['\n', <a href="https://admissionresults.vit.ac.in/ugcounselling" target="_blank">
<span style="color: white;">UG Science &amp; Humanities - </span>
<span style="padding: 5px 13px; letter-spacing: 2px; border-radius: 50px; color: #fff; border: 1px solid #FFA500; background-color: #FF8C00
; font-size: 14px;"> Results - 2021(Vellore campus)</span></a>, '\n', ' ', '\n']
```



```
 ul vc-nav-on-desktop
Priniting all the children in the menu
Class attribute of child 2:  ['vc-menu-item', 'vc-mm-mobile-toggle']
None
Class attribute of child 4:  ['menu-item', 'current-menu-parent', 'vc-menu-item', 'vc-d-0']
Home
Class attribute of child 6:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']
About Us
Class attribute of child 7:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']
Academics
Class attribute of child 8:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']
Admissions
Class attribute of child 9:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']
Career Development Centre
Class attribute of child 10:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']
International Relations
Class attribute of child 11:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']
Research
Class attribute of child 12:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']
Campus Life
Class attribute of child 13:  ['menu-item', 'current-menu-ancestor', 'menu-item-has-children', 'vc-menu-item', 'vc-d-0']

Number of descendants of this webpage: 2170
```



```
Div Tag:  <div class="ful_wid_col gal_imgs video_gal_col">
<img alt="VIT Video Gallery" loading="lazy" src="/sites/all/themes/vittheme/images/video_gal_img.png"/>
<h3 class="txt_title">Video Gallery</h3>
<div class="ful_wid_cnt">
<h3>Video Gallery</h3>
<a class="btn gal_explor hvr-sweep-to-right" href="https://vit.ac.in/video">Explore</a>
</div>
</div>

Next sibling of ☝div tag:  <div class="ful_wid_col gal_imgs host_col">
<img alt="VIT Hostels" loading="lazy" src="/sites/all/themes/vittheme/images/hostel_img.png"/>
<h3 class="txt_title">Hostel</h3>
<div class="ful_wid_cnt">
<h3>Hostel</h3>
<a class="btn gal_explor hvr-sweep-to-right" href="https://vit.ac.in/campus-hostel/hostels" loading="lazy">Explore</a>
</div>
</div>

Previous Sibling of above div tag:  <div class="ful_wid_col gal_imgs camp_col">
<img alt="VIT Campus" loading="lazy" src="/sites/all/themes/vittheme/images/campur_img.png"/>
<h3 class="txt_title">Campus Tour</h3>
<div class="ful_wid_cnt">
<h3>Campus Tour</h3>
<a class="btn gal_explor hvr-sweep-to-right" href="http://campustour.vit.ac.in/" target="_blank">Explore</a>
</div>
</div>
```



```
Printing ids of all the div present in the page:
fontSize
skip-link
fontSize
subsub_menu
subsub_menu
subsub_menu
cssmenu
banner_block_new
banner_carosel
evnet_new
popup_5137
news_new
with_cont
with_cont
with_cont
with_cont
message_front
seckit-noscript-tag
```