

Web mining Assignment -3

Note: You need to use “Beautiful Soup” package and “Requests” module. Programs need to be written in Python language 3.8 version. Use of Regular expressions allowed [re module]

NAME: GEROSH GEORGE

REG NO: 19BCE1403

1. Web page: <https://www.vit.ac.in>
 - Print the “title” of the page
 - Print out all the anchor tags with the class = “nav-link”

CODE:

```
import requests
from bs4 import BeautifulSoup

url=input('Enter the url: ')

resp = requests.get(url)

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

print(f'Title of page: {soup.title.text}')

a_tags = soup.findAll(name='a', attrs={'class':'nav-link'})

for a in a_tags:
    print(a)
```

OUTPUT:

```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Mining\Programs> python exercise1.py
Enter the url: https://www.vit.ac.in

Title of page: VIT | No.1 Private Institution for Innovation

<a class="nav-link vc-mm-mobile-toggle-btn" href="#"> Menu<i class="fa fa-bars"></i> </a>
<a class="nav-link vc-mm-mobile-toggle-btn" href="#"><i class="fa fa-bars"></i></a>
<a class="nav-link" href="https://vit.ac.in">Home</a>
<a class="nav-link" href="https://vit.ac.in/about-vit" title="About VIT">About Us</a>
<a class="nav-link" href="https://vit.ac.in/about-vit" title="Overview">Overview</a>
<a class="nav-link" href="https://vit.ac.in/about/vision-mission" title="Vision & Mission">Vision & Mission</a>
<a class="nav-link" href="https://vit.ac.in/vit-milestones" title="VIT Milestones">VIT Milestones</a>
<a class="nav-link" href="https://vit.ac.in/about/leadership" title="Leadership">Leadership</a>
<a class="nav-link" href="https://vit.ac.in/governance" title="Governance">Governance</a>
<a class="nav-link" href="https://vit.ac.in/about/administrative-offices" title="Administrative Offices">Administrative Offices</a>
<a class="nav-link" href="https://vit.ac.in/about/infrastructure" title="Infrastructure">Infrastructure</a>
<a class="nav-link" href="https://vit.ac.in/about/ranking-and-accreditation" title="Ranking & Accreditation">Ranking & Accreditation</a>
<a class="nav-link" href="https://vit.ac.in/about/sustainability" title="Sustainability">Sustainability</a>
<a class="nav-link" href="https://vit.ac.in/true-green" title="True Green project">True Green project</a>
<a class="nav-link" href="https://vit.ac.in/about/community-outreach" title="Community Outreach">Community Outreach</a>
<a class="nav-link" href="https://vit.ac.in/about/communityradio" title="Community Radio">Community Radio</a>
<a class="nav-link" href="https://vit.ac.in/all-news-archieved" title="Archieved News">Archieved News</a>
<a class="nav-link" href="https://vit.ac.in/all-events" title="Events">Events</a>
<a class="nav-link" href="https://vit.ac.in/national-institutional-ranking-framework-nirf" title="NIRF">NIRF</a>
<a class="nav-link" href="https://vit.ac.in/mhrdugc" title="MHRD/UGC">MHRD/UGC</a>
<a class="nav-link" href="https://careers.vit.ac.in/" title="Careers@VIT">Careers@VIT</a>
<a class="nav-link" href="https://vit.ac.in/about/news-letter" title="Newsletter">Newsletter</a>
<a class="nav-link" href="https://vit.ac.in/academics/home" title="Academics">Academics</a>
<a class="nav-link" href="https://vit.ac.in/academics/home" title="Overview">Overview</a>
<a class="nav-link" href="https://vit.ac.in/sites/default/files/academic/Academic-Regulations.pdf" title="Academic Regulations">Academic Regulations</a>
<a class="nav-link" href="https://vit.ac.in/programmes-offered-1" title="Programmes Offered">Programmes Offered</a>
<a class="nav-link" href="https://vit.ac.in/programmes-offered-2021-22" title="AY 2021-22">AY 2021-22</a>
<a class="nav-link" href="https://vit.ac.in/programmes-offered-2020-21" title="AY 2020-21">AY 2020-21</a>
<a class="nav-link" href="https://vit.ac.in/schools" title="Curriculum">Curriculum</a>

<a class="nav-link" href="https://vit.ac.in/programmes-offered-1" title="Programmes Offered">Programmes Offered</a>
<a class="nav-link" href="https://vit.ac.in/programmes-offered-2021-22" title="AY 2021-22">AY 2021-22</a>
<a class="nav-link" href="https://vit.ac.in/programmes-offered-2020-21" title="AY 2020-21">AY 2020-21</a>
<a class="nav-link" href="https://vit.ac.in/schools" title="Curriculum">Curriculum</a>
<a class="nav-link" href="https://vit.ac.in/academics/ffcs" title="FFCS">FFCS</a>
<a class="nav-link" href="https://vit.ac.in/academics/library" title="Library">Library</a>
<a class="nav-link" href="https://vit.ac.in/schools" title="Schools">Schools</a>
<a class="nav-link" href="https://vit.ac.in/academics-feedback" title="Feedback">Feedback</a>
<a class="nav-link" href="https://vit.ac.in/admissions/overview" title="Admissions">Admissions</a>
<a class="nav-link" href="https://vit.ac.in/admissions/overview" title="Overview">Overview</a>
<a class="nav-link" href="https://vit.ac.in/admissions/programmes-offered" title="Programmes Offered">Programmes Offered</a>
<a class="nav-link" href="https://vit.ac.in/all-courses/ug" title="Undergraduate">Undergraduate</a>
<a class="nav-link" href="https://vit.ac.in/all-courses/pg" title="Postgraduate">Postgraduate</a>
<a class="nav-link" href="https://vit.ac.in/admissions/research" title="Research">Research</a>
<a class="nav-link" href="https://vit.ac.in/admissions/research/Integrated_Ph.D" title="Integrated Ph.D">Integrated Ph.D</a>
<a class="nav-link" href="https://vit.ac.in/admissions/research/phd" title="PHD">PHD</a>
<a class="nav-link" href="https://vit.ac.in/admissions/international" title="International">International</a>
<a class="nav-link" href="https://vit.ac.in/stars-support-advancement-rural-students-0" title="STARS">STARS</a>
<a class="nav-link" href="https://vit.ac.in/placements/overview" title="Career Development Centre">Career Development Centre</a>
<a class="nav-link" href="https://vit.ac.in/career-development-centre" title="Overview">Overview</a>
<a class="nav-link" href="https://vit.ac.in/placements/superdreamoffers" title="Super Dream offer">Super Dream offer</a>
<a class="nav-link" href="https://vit.ac.in/placements/dreamoffers" title="Dream Offer">Dream Offer</a>
<a class="nav-link" href="https://vit.ac.in/placements/internship" title="Internships">Internships</a>
<a class="nav-link" href="https://vit.ac.in/placements/statistics" title="Statistics">Statistics</a>
<a class="nav-link" href="https://vit.ac.in/placements/pat-Office" title="CDC Office">CDC Office</a>
<a class="nav-link" href="https://vit.ac.in/placements/contact-us" title="Contact Us">Contact Us</a>
<a class="nav-link" href="https://vit.ac.in/InternationalRelations" title="International Relations">International Relations</a>
<a class="nav-link" href="https://vit.ac.in/InternationalRelations" title="Overview">Overview</a>
<a class="nav-link" href="https://vit.ac.in/internationalrelations/itp" title="International Transfer Programs (ITP)">International Transfer Programs (ITP)</a>
<a class="nav-link" href="https://vit.ac.in/internationalrelations/partneruniversities" title="Partner Universities">Partner Universities</a>
```

```
<a class="nav-link" href="https://vit.ac.in/internationalrelations/sap" title="SAP">SAP</a>
<a class="nav-link" href="https://vit.ac.in/admissions/international/overview" title="International Admissions">International Admissions</a>
<a class="nav-link" href="https://vit.ac.in/academics-more/Contact us" title="Contact us">Contact us</a>
<a class="nav-link" href="https://vit.ac.in/research" title="Research">Research</a>
<a class="nav-link" href="https://vit.ac.in/research" title="Overview">Overview</a>
<a class="nav-link" href="https://vit.ac.in/research/academic" title="Academic Research">Academic Research</a>
<a class="nav-link" href="https://vit.ac.in/research/sponsored-research" title="Sponsored Research">Sponsored Research</a>
<a class="nav-link" href="https://vit.ac.in/iprcell" title="IPR Cell">IPR Cell</a>
<a class="nav-link" href="https://vit.ac.in/research/centers-list" title="Research Centers">Research Centers</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/overview" title="Campus Life">Campus Life</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/overview" title="Overview">Overview</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/fests" title="Fests">Fests</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/studentswelfare" title="Students' Welfare">Students' Welfare</a>
<a class="nav-link" href="https://vit.ac.in/campus-category/Clubs" title="Student Clubs">Student Clubs</a>
<a class="nav-link" href="https://vit.ac.in/campus-category/Chapters" title="Student Chapters">Student Chapters</a>
<a class="nav-link" href="https://vit.ac.in/campus-category/Campus-Events" title="Campus Events">Campus Events</a>
<a class="nav-link" href="https://vit.ac.in/campus-category/Counselling-Division" title="Counselling Division">Counselling Division</a>
<a class="nav-link" href="https://vit.ac.in/campus-category/grievancecell" title="General Grievance Redressal Committee">General Grievance Redressal Committee</a>
<a class="nav-link" href="https://vit.ac.in/campus-category/Student-Council" title="Student Council">Student Council</a>
<a class="nav-link" href="https://vit.ac.in/academics/library" title="Library">Library</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/sports" title="Sports">Sports</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/hostels" title="Hostels">Hostels</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/healthservices" title="Health Services">Health Services</a>
<a class="nav-link" href="https://vit.ac.in/campuslife/otheramenities" title="Other Amenities">Other Amenities</a>
<a class="nav-link" href="https://vit.ac.in/detailview/green-vit" title="Green Campus">Green Campus</a>
<a class="nav-link" href="https://vit.ac.in/vitol/" target="_blank">Online Institute (VITOL)</a>
<a class="nav-link" href="javascript:void(0)">Examinations</a>
<a class="nav-link" href="https://vit.ac.in/academics/coe" target="_blank">Controller of Examinations</a>
<a class="nav-link" href="https://vit.ac.in/transcripts-alumni" target="_blank">Transcripts for Alumni</a>
<a class="nav-link" href="https://vit.applydirect.org/student" target="_blank">WES attestation</a>
<a class="nav-link" href="https://applications.wes.org/required-documents/Search/Result?cId=2&eId=38&iNm=Vellore%20Institute%20of%20Technology%20%28VIT%20&cNm=Bachelor%20of%20Applied%20Arts&iSec=University%20Education&iSub=Bachelor%27s%20Degree%20Programs&p;comp=1&isEca=False" target="_blank">WES Application Process</a>
<a class="nav-link" href="https://www.wes.org/advisor-blog/send-electronic-academic-records/" target="_blank">WES Advisory Blog</a>
```

```
<a class="nav-link" href="https://vit.ac.in/files/e-SANAD-Notification.pdf" target="_blank">e-Sanad</a>
<a class="nav-link" href="https://web.vit.ac.in/CertificateVerification/login" target="_blank">Certificate Verification for Registered Agencies</a>
<a class="nav-link" href="https://vit.ac.in/centers/asc" target="_blank">Academic Staff College</a>
<a class="nav-link" href="http://www.vittbi.com/#/" target="_blank">VIT TBI</a>
<a class="nav-link" href="javascript:void(0)">Parents</a>
<a class="nav-link" href="https://vtop.vit.ac.in/" target="_blank">Parent Login</a>
<a class="nav-link" href="https://vit.ac.in/campus-category/Counselling-Division" target="_blank">Counselling Division</a>
<a class="nav-link" href="https://vit.ac.in/guest-house" target="_blank">Guest House</a>
<a class="nav-link" href="https://vit.ac.in/redressal" target="_blank">Grievance Redressal</a>
<a class="nav-link" href="https://vit.ac.in/hotels-in-vellore" target="_blank">Hotels in Vellore</a>
<a class="nav-link" href="javascript:void(0)">Students</a>
<a class="nav-link" href="https://mail.google.com/" target="_blank">VIT Gmail</a>
<a class="nav-link" href="https://vtop.vit.ac.in/" target="_blank">Student Login</a>
<a class="nav-link" href="http://intranet.vit.ac.in" target="_blank">VIT Intranet</a>
<a class="nav-link" href="https://vit.ac.in/sites/default/files/SBST_Freshers_App.rar" target="_blank">SBST Freshers App</a>
<a class="nav-link" href="https://vit.ac.in/anti-ragging-committee" target="_blank">Anti Ragging Committee</a>
<a class="nav-link" href="https://vit.ac.in/capability-enhancement-scheme" target="_blank">Capability Enhancement Schemes</a>
<a class="nav-link" href="https://vit.ac.in/sites/default/files/FormatGuidelines.doc" target="_blank">Industrial Visit</a>
<a class="nav-link" href="https://vit.ac.in/internal-complaints-committee" target="_blank">IICC</a>
<a class="nav-link" href="javascript:void(0)">Alumni</a>
<a class="nav-link" href="https://vit.ac.in/alumni_progression" target="_blank">Alumni Progression</a>
<a class="nav-link" href="https://vit.ac.in/academics/transcripts" target="_blank">Transcripts</a>
<a class="nav-link" href="https://vit.ac.in/instruction" target="_blank">Medium of Instruction / Migration Certificate</a>
<a class="nav-link" href="https://www.campusinteraction.com/" target="_blank">Lateral Hiring</a>
<a class="nav-link" href="https://vtop.vit.ac.in/vtop/initialProcess" target="_blank">Other services to Alumni</a>
<a class="nav-link" href="https://vit.ac.in/alumni-events" target="_blank">Alumni - Events</a>
<a class="nav-link" href="https://vit.ac.in/detailview/alumni-photo-gallery" target="_blank">Alumni - Photo Gallery</a>
<a class="nav-link" href="http://www.vitaa.org/" target="_blank">VITAA Website</a>
<a class="nav-link" href="https://vit.ac.in/alumni-office-contact" target="_blank">Alumni contact</a>
<a class="nav-link" href="https://campustour.vit.ac.in/" target="_blank">Campus Tour</a>
```

2. Web page : <https://vit.ac.in/school/allfaculty/site/computer-applications>
- Print out all the faculty names using the class id "title2" and their research area by devising appropriate algorithm – you need to use methods of BeautifulSoup
 - Find the facebook link, twitter link, instagram link and linkedin link of VIT from the page content and identify the class names given for each.
 - List out the DOM hierarchy of the page [Find all the children and the relationship between the children

CODE:

```
import requests
from bs4 import BeautifulSoup

url=input('Enter the url: ')

resp = requests.get(url)

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

names = []
research=[]
print("Faculty names along with their research area: ")
for h3 in soup.findAll('h3',{'class':'title2'}):
    parent = h3.parent
    p_tags = parent.find_all('p')
    if len(p_tags) == 3:
        if len(p_tags[2].text)==1:
            research.append('None')
        else:
            research.append(p_tags[2].text)
    else:
        research.append('None')
    names.append(h3.text)

for faculty,research in zip(names,research):
    print(f'{faculty}: {research}')

print("\nSocial Links: ")

links_span = soup.find('span',{'class':'soclia_links'})
links = links_span.find_all('a')
for link in links:
    print(f"{link['title']} : {link['href']}")
    print(f"class name: {link['class']}\n")
```

OUTPUT:

a)

```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Minning\Programs> python exercise2.py
Enter the url: https://vit.ac.in/school/allfaculty/site/computer-applications
Faculty names along with their research area:
Dr.Ramkumar T: Data Mining & Big Data Analytics
Dr.Ephzibah E.P: Data Mining and Artificial Intelligence
Dr. Karthikeyan P: Cloud computing, Web Services
Dr.Manivannan S.S: Network and Information Security, IoT and Machine Learning
Dr. Meenatchi S: Computer Science Hardware and Architecture
Dr.Senthil Murugan B: None
Dr.Shynu P.G: Cloud computing, Information Security, Data Science
Dr. Uma Maheswari G: Computer Science Information Systems
Dr.Venketesh P: None
Dr.Deepa N: Predictive Analytics
Dr.Jayalakshmi P: Networks
Dr. Senthil Kumar N: Semantic Web; Information Retrieval
Ms. Deepa P: None
Ms. Manisha R. Patil: None
Ms. Manjupriya R: None
Mr. Ravindran U: None
Mr.Sreeraag G: None
```

b)

```
Social Links:
facebook : https://www.facebook.com/VITUniversity/
class name: ['face_book_icon', 'f_icon_be']

twitter : https://twitter.com/vit_univ
class name: ['twitter_icon', 'f_icon_be']

LinkedIn : https://www.linkedin.com/school/vellore-institute-of-technology/
class name: ['linkedin_icon', 'f_icon_be']

youtube : https://www.youtube.com/c/VITUniversityVellore
class name: ['youbtube_icon', 'f_icon_be']

instagram : https://www.instagram.com/vellore_vit/
class name: ['insta_icon', 'f_icon_be']
```

c)

```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Mining\Programs> python find_dom.py
Enter the url: https://vit.ac.in/school/allfaculty/site/computer-applications
-> head
|
| -> meta
| -> meta
| -> meta
| -> meta
| -> link
| -> meta
|   -> link
|   -> meta
|     -> meta
|     -> meta
|     -> meta
|     -> meta
|     -> meta
|     -> meta
|     -> meta
|     -> title
|     -> style
|     -> meta
|     -> meta
|     -> meta
|     -> meta
|     -> meta
|   -> meta
|   -> meta
|   -> meta
|   -> meta
|   -> meta
| -> body
|   -> div
|     -> div
|       -> a
|     -> div
|       -> script
|       -> script
|       -> script
```

```
-> section
| -> div
| | -> div
| | | -> div
| | | | -> div
| | | | | -> div
| | | | | -> div
| | | | | -> a
| | | | | | -> img
| | | -> div
| | | | -> div
| | | | -> ul
| | | | | -> li
| | | | -> ul
| | | | | -> li
| | | | | -> li
| | | | | -> li
| | | | | -> li
| | | | | -> li
| | | | | -> li
| | | | | -> li
| | | | | -> li
| | | | | -> li
| | | | -> span
| | | | -> div
| | | | | -> div
| -> header
| | -> div
| | | -> div
| | | | -> img
| | | -> div
| | | -> div
| | | | -> div
| | | | -> div
| | | | | -> div
```

```
| | | | | | | | | | -> div
| | | | | | | | | | -> div
| | | | | | | | | | -> div
| | | | | | | | | | -> div
| | | | | | | | | | -> div
| | | | | | | | | | -> span
| | | | | | | | | | -> ul
| | | | | | | | | | -> li
-> div
-> div
| -> div
| | -> div
| | | -> div
| | | | -> div
| | | | -> div
| | | | -> ul
| | | | | -> li
| | | | | -> a
| | | | | -> li
| | | | | -> a
| | | | | -> li
| | | | | -> a
| | | | | -> li
-> div
| -> div
| | -> div
| | | -> div
| | | | -> div
| | | | -> div
| | | | -> div
| | | | -> div
-> script
-> footer
| -> div
| | -> div
| | | -> div
```



```
-> div
| -> div
| | -> div
| | | -> div
| | | | -> h3
| | | | -> ul
| | | -> div
| | | | -> h3
| | | | -> ul
| | | -> div
| | | | -> h3
| | | | -> ul
| | | -> div
| | | | -> h3
| | | | -> ul
| | | -> div
| | | | -> h3
| | | | -> ul
| | | -> div
| | | | -> h3
| | | | -> ul
| | | -> br
| -> div
| -> div
| | -> div
| | | -> ul
| | | | -> li
| | | -> span
| | | | -> a
| | | | -> a
| | | | -> a
| | | | -> a
| | | | -> a
| -> hr
| -> div
| | -> div
| | | -> div
| | | | -> span
```


3.

Web page : <https://sermitsiaq.ag/english>

- a. Find all the items of class="menu" and print out the items of the menu with class names With class names "first leaf", "leaf" and "last leaf".
- b. Find all items with ids containing string "menu" in them
- c. Find all items with tag "article".
- d. List out the DOM hierarchy of the page

CODE:

```
import requests
from bs4 import BeautifulSoup as bs
import re

url=input('Enter the url: ')

resp = requests.get(url,headers={'User-Agent': 'Mozilla/5.0'})

page = resp.text
soup= bs(page,'html.parser')

items = soup.findAll('ul',{'class':'menu'})
print(f"Number of items of class menu: {len(items)}")

print("\nItems inside class Menu with classes 'first leaf', 'leaf' and 'last leaf'")
for item in items:
    elements = item.findAll('li',{'class':['leaf','last leaf','first leaf']})
    for elem in elements:
        print(elem)
    print("\n")

items = soup.findAll(id=re.compile("menu"))
print("\nAll elements with id containing 'menu':\n",*items,sep="\n")

articles = soup.find_all("article")
print("\nAll article tags:\n",*articles,sep="\n")
```

OUTPUT:

a)

```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Mining\Programs> python exercise3.py
Enter the url: https://sermitsiaq.ag/english
Number of items of class menu: 2

Items inside class Menu with classes 'first leaf', 'leaf' and 'last leaf'
<li class="first leaf"><a href="/node/195195">Bestil foto</a></li>
<li class="leaf"><a href="http://aviisi.sermitsiaq.ag/abonnemeter/" target="_blank">Abonnement</a></li>
<li class="leaf"><a href="/annoncer">Annoncer</a></li>
<li class="leaf"><a href="/node/155045">Kontakt</a></li>
<li class="leaf"><a href="http://aviisi.sermitsiaq.ag" target="_blank">E-aviser</a></li>
<li class="leaf job-link"><a href="http://job.sermitsiaq.ag" target="_blank">JOB</a></li>
<li class="last leaf"><a href="/pilivik">Pilivik</a></li>

<li class="first leaf"><a href="/">Forsiden</a></li>
<li class="leaf"><a class="menu-indland" href="/indland">Indland</a></li>
<li class="leaf"><a href="/nuuk">Nuuk</a></li>
<li class="leaf"><a class="menu-politik" href="/politik">Politik</a></li>
<li class="leaf"><a href="/taxonomy/term/6/">Erhverv</a></li>
<li class="leaf"><a href="/politi">Politi</a></li>
<li class="leaf"><a class="menu-udland" href="/udland">Udland</a></li>
<li class="leaf"><a href="/kultur">Kultur</a></li>
<li class="leaf"><a href="/sport">Sport</a></li>
<li class="leaf"><a href="/nyhedsoversigt">Nyhedsoversigt</a></li>
<li class="last leaf job-link"><a class="menu-job" href="http://job.sermitsiaq.ag" style="font-weight: bold" target="_blank">Job</a></li>
```

b)

```
All elements with id containing 'menu':

<button id="menuToggle"></button>
<div class="panel-pane pane-block pane-menu-menu-secondary-menu" id="secondarymenu">
<ul class="menu"><li class="first leaf"><a href="/node/195195">Bestil foto</a></li>
<li class="leaf"><a href="http://aviisi.sermitsiaq.ag/abonnemeter/" target="_blank">Abonnement</a></li>
<li class="leaf"><a href="/annoncer">Annoncer</a></li>
<li class="leaf"><a href="/node/155045">Kontakt</a></li>
<li class="leaf"><a href="http://aviisi.sermitsiaq.ag" target="_blank">E-aviser</a></li>
<li class="leaf job-link"><a href="http://job.sermitsiaq.ag" target="_blank">JOB</a></li>
<li class="last leaf"><a href="/pilivik">Pilivik</a></li>
</ul> </div>
<div id="mainmenu">
<nav class="mainmenu"><div class="panel-pane pane-block pane-system-main-menu">
<ul class="menu"><li class="first leaf"><a href="/">Forsiden</a></li>
<li class="leaf"><a class="menu-indland" href="/indland">Indland</a></li>
<li class="leaf"><a href="/nuuk">Nuuk</a></li>
<li class="leaf"><a class="menu-politik" href="/politik">Politik</a></li>
<li class="leaf"><a href="/taxonomy/term/6/">Erhverv</a></li>
<li class="leaf"><a href="/politi">Politi</a></li>
<li class="leaf"><a class="menu-udland" href="/udland">Udland</a></li>
<li class="leaf"><a href="/kultur">Kultur</a></li>
<li class="leaf"><a href="/sport">Sport</a></li>
<li class="leaf"><a href="/nyhedsoversigt">Nyhedsoversigt</a></li>
<li class="last leaf job-link"><a class="menu-job" href="http://job.sermitsiaq.ag" style="font-weight: bold" target="_blank">Job</a></li>
</ul> </div>
</nav>
</div>
<div id="thememenu">
<nav class="thememenu"><div class="panel-pane pane-views-panes pane-theme-panel-pane-1">
```

```

<div class="theme-title"><span class="theme">TEMAER</span><span class="corner"><svg enable-background="new 0 0 16 32" height="32px" id="Layer_1" version="1.1" viewBox="0 0 16 32" width="16px" x="0px" xml:space="preserve" xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink" y="0px">
  <g>
    <path d="M13.184,16.153" fill="#A8130D"></path>
    <g>
      <polyon fill="#FFFFFF" points="14.497,13.914 5.188,0 0,0 9.309,13.914 10.812,16.161 9.289,18.393 0.007,31.996 5.196,31.996 14.477,18.393 16,16.161" />
      <polyon fill="#A8130D" points="10.812,16.161 9.309,13.914 0,0 0.007,31.996 9.289,18.393" />
    </g>
  </g>
</svg></span></div>
<ul class="theme">
  <li>
    <div class="views-field views-field-name"> <span class="field-content"><a href="/emne/coronavirus">Coronavirus</a></span> </div> </li>
    <li>
      <div class="views-field views-field-name"> <span class="field-content"><a href="/emne/nye-lufthavne">nye lufthavne</a></span> </div> </li>
    <li>
      <div class="views-field views-field-name"> <span class="field-content"><a href="/emne/vaccinesagen">Vaccinesagen</a></span> </div> </li>
    <li>
      <div class="views-field views-field-name"> <span class="field-content"><a href="/emne/kuannersuitkvanevfjeld">Kuannersuit/Kvanevfjeld</a></span> </div> </li>
    <li>
      <div class="views-field views-field-name"> <span class="field-content"><a href="/emne/sexchikane">Sexchikane</a></span> </div> </li>
  </ul>
<div class="clearfix"></div> </div>
</nav>
</div>
<div id="omenu"></div>

```

c)

```

All article tags:

<article class="node node--article node--w615 node--article--w615" role="article">
  <h1>Arctic people concerned for the future </h1>
</article>
<article class="node node--article node--w300 node--article--w300" role="article">
  <h1>Press release: Common Arctic search and rescue service agreed</h1>
</article>
<article class="node node--article node--w300 node--article--w300" role="article">
  <h1>Address Kuupik Kleist Arctic Council 7 Meeting 2011</h1>
</article>
<article class="node node--article node--w300 node--article--w300" role="article">
  <h1>Open letter sent to the Foreign Ministers of Canada, U.S., Norway, Denmark, Greenland and Russia</h1>
</article>
<article class="node node--article node--w300 node--article--w300" role="article">
  <h1>Resource Development Principles in Inuit Nunaat</h1>
</article>
<article class="node node--article node--w300 node--article--w300" role="article">
  <h1>Arctic Council Nuuk Ministerial Agenda</h1>
</article>
<article class="node node--article node--w300 node--article--w300" role="article">
  <h1>Editorial: Our friends in Norway</h1>
</article>
<article class="node node--article node--term-list node--article--term-list" role="article">
  <h1>Climate change - it's about the people</h1>
  <h2>
    Living conditions and economic development in the face of climate change are the challenges the Arctic Council needs to deal with, Greenland's premier says</h2>
</article>
<article class="node node--article node--term-list node--article--term-list" role="article">
  <h1>Tight security during Arctic summit</h1>
  <h2>
    Terrorism threat levels remain "serious" as Arctic foreign ministers gather in Nuuk</h2>
</article>
<article class="node node--article node--term-list node--article--term-list" role="article">
  <h1>New Arctic strategy to be presented in Nuuk</h1>

```

```

  As climate scientists paint an increasingly dire picture of global warming, it has begun to sink in with lawmakers that slowing the development requires immediate action</h2>
</article>
<article class="node node--article node--term-list node--article--term-list" role="article">
  <h1>Arctic Council Nuuk Ministerial Agenda</h1>
  <h2>
    </h2>
</article>
<article class="node node--article node--term-list node--article--term-list" role="article">
  <h1>Participants in Arctic Council Nuuk Ministerial</h1>
  <h2>
    The smallest delegations have as few as six members, the largest up to 25</h2>
</article>

```

d)

Enter the url: <https://sermitsiaq.ag/english>

-> head

-> link

-> link

-> link

-> script

-> link

-> link

-> meta

-> script

-> link

-> meta

-> link

-> link

-> link

-> link

-> meta

-> meta

-> link

-> link

-> meta

-> meta

-> meta

-> meta

-> meta

-> meta

-> meta

-> meta

-> title

-> link

-> link

-> link

-> style

-> script

-> script

-> script

```
| | | | -> script
| | | | -> script
| | | | -> script
| | | | -> script
| | | | -> meta
| | | | -> script
| | | | -> script
| | | | -> link
| | | | -> link
| | | | -> link
| | | | -> script
| | | | -> script
| | | | -> script
| | | | -> script
| | | | -> script
| -> body
| | -> noscript
| | | -> iframe
| | -> div
| | -> script
| | -> script
| | -> a
| | -> header
| | | -> div
| | | | -> div
| | | | | -> div
| | | | | | -> ul
| | | | | | | -> li
| | | | | | | | -> a
| | | | | | | | | -> img
| | | | | | | -> li
| | | | | | | -> a
| | | | | | | | -> img
| | | | -> div
| | | | | -> button
| | | | | -> button
```

```

| -> div
| | -> ul
| | | -> li
| | | | -> a
| | | -> li
| | | | -> a
| | | -> li
| | | | -> a
| | | -> li
| | | | -> a
| | | -> li
| | | | -> a
| | | -> li
| | | | -> a
| | | -> li
| | | | -> a
| -> div
| -> div
| | -> form
| | | -> div
| | | | -> input
| | | | -> input
| | | | -> div
| | | | | -> label
| | | | | -> input
| | | | -> input
| | | | -> input
| | | | -> input
| | | -> div
| | | | -> gcse:search
| | | -> noscript
| -> div
-> div
| -> div
| | -> div
| | | -> a

```



```
| | | | | | -> p
| | | | | | | -> a
| | | | -> div
| | | | -> div
| | | | -> h2
| | | | | -> a
| | | | -> h2
| | | | | -> a
| | | | -> h2
| | | | | -> a
| | | | -> h2
| | | | | -> a
| | | | -> h2
| | | | | -> a
| | | | -> h2
| | | | | -> a
| | | | -> h2
| | | | | -> a
| | | | -> p
| | | | -> h2
| | | | -> div
| | | -> div
| | | | -> p
| | | | -> strong
| | -> script
| | -> script
| | -> script
| | -> script
| | -> div
| | -> script
| | -> script
| | -> script
| | -> script
| | -> script
```

4. Take website :<https://www.batimes.com.ar>
 - a. List items of class “nav-item text-uppercase px-0”
 - b. Search for string “Matías Lammens” and list out the HTML item of which the string occurs
 - c. List all the images in the page
 - d. List out the DOM hierarchy of the page.

CODE:

```
import requests
from bs4 import BeautifulSoup as bs
import re

url=input('Enter the url: ')

resp = requests.get(url)

page = resp.text
soup= bs(page,'html.parser')

items = soup.find_all(class_="nav-item text-uppercase px-0")
print('\nList items of class “nav-item text-uppercase px-0”: ',*items,sep='\n')

elements = soup.find_all(string=re.compile('Matías Kulfas'))
print("\nElements containing string 'Matías Kulfas'")
for elem in elements:
    print(elem.parent)

images = soup.find_all("img")
print("\nAll Images :",*images,sep="\n\n")
```

OUTPUT:

a)

```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Mining\Programs> python exercise4.py
Enter the url: https://www.batimes.com.ar

List items of class "nav-item text-uppercase px-0":
<li class="nav-item text-uppercase px-0">
<a class="px-0 nav-link" href="/last-news/" target="_self" title="Topics of the day">
Topics
</a>
</li>
<li class="nav-item text-uppercase px-0">
<a class="px-0 nav-link" href="/news/argentina/president-fernandez-under-fire-for-olivos-birthday-party-during-lockdown.phtml" target="_self" title="President Fer
nández under fire for Olivos birthday party during lockdown">
Olivos party
</a>
</li>
<li class="nav-item text-uppercase px-0">
<a class="px-0 nav-link" href="/news/sports/messis-hunger-for-more-means-star-is-sure-to-shine-in-city-of-lights.phtml" target="_self" title="Messi's hunger for m
ore means star is sure to shine in City of Lights">
Messi moves to Paris
</a>
</li>
<li class="nav-item text-uppercase px-0">
<a class="px-0 nav-link" href="/news/argentina/matias-kulfas-to-lower-inflation-you-need-multi-sector-agreements-and-clear-objectives.phtml" target="_self" title=
"Matias Kulfas: 'To lower inflation you need multi-sector agreements and clear objectives'">
Matias Kulfas interview
</a>
</li>
<li class="nav-item text-uppercase px-0">
<a class="px-0 nav-link" href="/news/opinion-and-analysis/what-does-the-new-ipcc-report-mean-for-latin-america.phtml" target="_self" title="What does the new IPCC
report mean for Latin America?">
IPCC report
</a>
</li>
```

b)

```
Elements containing string 'Matias Kulfas'
<a class="px-0 nav-link text-secondary" href="/news/argentina/matias-kulfas-to-lower-inflation-you-need-multi-sector-agreements-and-clear-objectives.phtml" target
="_self" title="Matias Kulfas: 'To lower inflation you need multi-sector agreements and clear objectives'">
Matias Kulfas interview
</a>
<a class="px-0 nav-link" href="/news/argentina/matias-kulfas-to-lower-inflation-you-need-multi-sector-agreements-and-clear-objectives.phtml" target="_self" title=
"Matias Kulfas: 'To lower inflation you need multi-sector agreements and clear objectives'">
Matias Kulfas interview
</a>
<h2>Matias Kulfas: 'To lower inflation you need multi-sector agreements and clear objectives'</h2>
<h4><span>13-08-2021 22:36</span> Productive Development Minister Matias Kulfas on trying to lower inflation, the IMF agreement and the role of the state.
</h4>
```

c)

```
All Images :





























/270/152/passengers-ezeiza-airport-tourism-coronavirus-covid-1221313.jpg 270w,https://fotos.perfil.com/2021/08/24/trim/540/304/passengers-ezeiza-airport-tourism-coronavirus-covid-1221313.jpg 540w,https://fotos.perfil.com/2021/08/24/trim/720/405/passengers-ezeiza-airport-tourism-coronavirus-covid-1221313.jpg 720w,https://fotos.perfil.com/2021/08/24/trim/720/355/passengers-ezeiza-airport-tourism-coronavirus-covid-1221313.jpg 720w," src="https://via.placeholder.com/720x405?text=BATIMES"/>























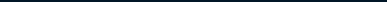












Scorecard Research

```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Mining\Programs>
```

```
(venv) PS C:\Users\Gerosh Shibu George\Desktop\VIT College\Third Year\Web Mining\Programs> python find_dom.py
Enter the url: https://www.batimes.com.ar
```

[illegible]

```
| | | | | -> link
| | | | | -> link
| | | | | -> script
| | | | | -> script
| | | | | -> script
| | | | | -> script
| | | | | -> script
| | | | | -> script
| -> body
| | -> script
| | -> script
| | -> noscript
| | | -> iframe
| | -> div
| | | -> header
| | | | -> div
| | | | | -> div
| | | | | -> div
| | | | | -> div
| | | | | -> div
| | | | | -> div
| | | | | -> a
| | | | | | -> img
| | | | -> div
| | | | -> div
| | | | | -> nav
| | | | | -> button
| | | | | | -> span
| | | | | -> a
| | | | | -> img
| | | | -> div
| | | | | -> ul
| | | | | | -> li
| | | | | | | -> a
| | | | | | -> li
| | | | | | | -> a
| | | | | | -> li
| | | | | | | -> a
```

```
-> div
 -> ul
 -> li
 -> a
 -> li
 -> a
 -> li
 -> a
 -> li
 -> a
 -> li
 -> a
 -> div
 -> a
 -> i
 -> a
 -> i
 -> a
 -> i
 -> a
 -> i
 -> main
 -> section
 -> section
 -> div
 -> div
 -> div
 -> div
 -> div
 -> div
 -> div
 -> div
 -> article
 -> h1
 -> a
```



```
 | -> a
 | | -> figure
 | | -> div
 | -> div
 | | -> div
 | | -> article
 | | -> article
 | -> div
 | | -> article
 | | -> a
 | | -> article
 | | -> a
 | -> div
 | | -> div
 | -> div
 | | -> div
 | -> div
 | -> div
 -> section
 | -> div
 | | -> div
 | | | -> article
 | | | | -> a
 | | | | -> figure
 | | | | -> img
 | | | | -> div
 | | | | -> h3
 | | | | -> h2
 | | | | -> h4
 | | | | -> h5
 | | -> div
 | | | -> div
 | | | -> article
 | | | | -> a
 | | | | -> figure
 | | | -> div
```

```
| | | | -> h2
-> div
| -> div
-> article
| -> span
| -> a
| -> figure
| -> img
| -> h4
| -> div
| -> h3
| -> h2
-> article
| -> span
| -> a
| -> figure
| -> img
| -> h4
| -> div
| -> h3
| -> h2
-> div
| -> div
-> article
| -> span
| -> a
| -> figure
| -> img
| -> h4
| -> div
| -> h3
| -> h2
| -> h5
-> article
| -> span
| -> a
```





CODE FOR FINDING THE DOM STRUCTURE:

```
import requests
from bs4 import BeautifulSoup as bs

url=input('Enter the url: ')

resp = requests.get(url,headers={'User-Agent': 'Mozilla/5.0'})

page = resp.text
soup= bs(page,'html.parser')

def graph(value,size):
 if size>=12:
 return
 else:
 for i in value.findChildren(recursive=False):
 for j in range(-1,size,1):
 print("|",end=" ")
 print("->",i.name)
 graph(i,size+1)

graph(soup.html,0)
```