Name : Aryan Vigyat

Registration Number : 20BCE1452

Question 1:

Given a seed/root URL, e.g., "Vit.ac.in", Design a simple crawler to return all pages (URLs) that contains a keyword "research" from this site. (25 pages)

In [ ]:
```
import re
from bs4 import BeautifulSoup as soup
import requests
seed_url="http://www.vit.ac.in"
searchWord="research"
maxPages=25
response=requests.get(seed_url)
print("Response Status Code :",response.status_code)
rootPage=soup(response.content,'html.parser')
atags=rootPage.find_all('a')
result=list()
pages=0
for atag in atags:
    if(pages==maxPages):
        break
    link=atag['href']
    if re.search(searchWord,link,re.IGNORECASE):
        result.append(link)
        pages+=1

print("The Links in the SEED URL Page which consists of the word research are as fo
for res in result:
    print("\t",res)
```

```
Response Status Code : 200
The Links in the SEED URL Page which consists of the word research are as follows
         https://vit.ac.in/admissions/research
         https://vit.ac.in/research
         https://vit.ac.in/research
         https://vit.ac.in/research/academic
         https://vit.ac.in/research/sponsored-research
         https://vit.ac.in/research/centers-list
         https://vit.ac.in/schools-centres-list-research-guides-2022
         3d-printing-play-major-role-mitigating-spread-covid-19-say-researchers-vi
t
         3d-printing-play-major-role-mitigating-spread-covid-19-say-researchers-vi
t
         https://vit.ac.in/research
```

Question 2:

Find documents that contain the word “admissions” and the word “international” within the URL “Vit.ac.in” using Python. (25 pages)

In [ ]:
```
import requests
from bs4 import BeautifulSoup as soup
import re
seed_URL = "http://www.vit.ac.in"
searchWords = ['admissions', 'international']
maxPages=25
response=requests.get(seed_URL)
print("Response Status Code :",response.status_code)
rootPage=soup(response.content,'html.parser')
atags=rootPage.find_all('a')
validLinks=list()
for atag in atags:
    link=atag['href']
    if link.startswith("http") :
        if link not in validLinks :
            validLinks.append(link)
print("Total Number of Documents is {}".format(len(validLinks)))
final=list()
foundPages=0
failed=list()
pages=0
for link in validLinks :
    if(pages==maxPages):
        break
    try :
        page = requests.get(link).text
    except requests.ConnectionError :
        try :
            page = requests.get(link, verify=False).text
        except :
            failed.append(link)
        continue
    if (re.search(searchWords[0], page, re.IGNORECASE)) and (re.search(searchWords
        final.append(link)
        foundPages+=1
    pages+=1
print("The Documents that Contain the Words Admission and International is ")
for i in final:
    print("\t",i)
```

```
Response Status Code : 200
Total Number of Documents is 166
```

```
C:\Users\ayuar\AppData\Local\Programs\Python\Python310\lib\site-packages\urllib3\c
onnectionpool.py:1045: InsecureRequestWarning: Unverified HTTPS request is being m
ade to host 'chennai.vit.ac.in'. Adding certificate verification is strongly advis
ed. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
```

```
The Documents that Contain the Words Admission and International is
        https://vitap.ac.in/
        https://vitbhopal.ac.in/
        https://vit.ac.in
        https://vit.ac.in/about-vit
        https://vit.ac.in/about/vision-mission
        https://vit.ac.in/vit-milestones
        https://vit.ac.in/about/leadership
        https://vit.ac.in/governance
        https://vit.ac.in/about/administrative-offices
        https://vit.ac.in/about/infrastructure
        https://vit.ac.in/about/sustainability
        https://vit.ac.in/true-green
        https://vit.ac.in/about/community-outreach
        https://vit.ac.in/about/communityradio
        https://vit.ac.in/all-news-archieved
        https://vit.ac.in/all-events
        https://vit.ac.in/national-institutional-ranking-framework-nirf
        https://vit.ac.in/mhrdugcaicte
        https://vit.ac.in/about/news-letter
        https://vit.ac.in/academics/home
        https://vit.ac.in/programmes-offered-1
        https://vit.ac.in/programmes-offered-2021-22
        https://vit.ac.in/programmes-offered-2020-21
        https://vit.ac.in/schools
```

Question 3:

Find documents that contain the word “Programme” but not the word “programming” within the URL “Vit.ac.in” using Python. (5 pages)

In [ ]:

```python
import re
from bs4 import BeautifulSoup as soup
import requests
seedURL="http://vit.ac.in"
searchWords=['Programme','Programming']
response=requests.get(seedURL)
maxPages=5
print("Response Status Code :",response.status_code)
rootPage=soup(response.content,'html.parser')
atags=rootPage.find_all('a')
validLinks=list()
for atag in atags:
    link=atag['href']
    if link.startswith("http") :
        if link not in validLinks :
            validLinks.append(link)
print("Total Number of Documents Linked to the Root URL is {}".format(len(validLink
final=list()
foundPages=0
failed=list()
pages=0
for link in validLinks :
    if(pages==maxPages):
        break
```

```
    try :
        page = requests.get(link).text
    except requests.ConnectionError :
        try :
            page = requests.get(link, verify=False).text
        except :
            failed.append(link)
        continue
    if (re.search(searchWords[0], page, re.IGNORECASE)) and (not re.search(searchWo
        final.append(link)
        foundPages+=1
    pages+=1
print("The Documents that Contain the Word Programme and not Programming is ")
for i in final:
    print("\t",i)
```

```
Response Status Code : 200
Total Number of Documents Linked to the Root URL is 166
C:\Users\ayuar\AppData\Local\Programs\Python\Python310\lib\site-packages\urllib3\c
onnectionpool.py:1045: InsecureRequestWarning: Unverified HTTPS request is being m
ade to host 'chennai.vit.ac.in'. Adding certificate verification is strongly advis
ed. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
The Documents that Contain the Word Programme and not Programming is
         https://vitap.ac.in/
         https://vitbhopal.ac.in/
         https://vit.ac.in
         https://vit.ac.in/about-vit
         https://vit.ac.in/about/vision-mission
```

Question 4:

Write a web crawler program which takes as input a url (Educational website) and a search key word and maximum number of pages (15-20 Pages) to be searched and returns as output all the web pages it searched till it found the search word on a web page or return failure.

In [ ]:
```
import re
from bs4 import BeautifulSoup as soup
import requests
seedURL4 = input("Enter the Input URL:")
searchWord = input("Enter the Search Word: ")
maxPages = int(input("Enter the Max Pages:"))
response = requests.get(seedURL4)
print("Status of the response : ", response.status_code)
rootPage=soup(response.content,'html.parser')
atags=rootPage.find_all('a')
validLinks=list()
for atag in atags:
    try:
        link=atag['href']
        if link.startswith("http") :
            if link not in validLinks :
                validLinks.append(link)
```

```
    except:
        pass
print("Total Number of Documents is {}".format(len(validLinks)))
final=list()
foundPages=0
failed=list()
pages=0
for link in validLinks :
    if(pages==maxPages):
        break
    try :
        page = requests.get(link).text
    except requests.ConnectionError :
        try :
            page = requests.get(link, verify=False).text
        except :
            failed.append(link)
        continue
    if (re.search(searchWord, page, re.IGNORECASE)):
        final.append(link)
        foundPages+=1
    pages+=1
if(foundPages==0):
    print("Failure")
else:
    print("The Documents that Contain the Word {} is ".format(searchWord))
    for i in final:
        print("\t",i)
```

```
Enter the Input URL:https://pes.edu/
Enter the Search Word: campus
Enter the Max Pages:19
Status of the response :  200
Total Number of Documents is 66
The Documents that Contain the Word campus is
         https://pessat.com/
         https://pessat.com
         https://pes.edu/btech/
         https://pes.edu/programs/
         https://pes.edu/scholarships/
         https://pes.edu/placements/
         https://news.pes.edu/
         https://news.pes.edu/11033
         https://news.pes.edu/11031
         https://news.pes.edu/11032
         https://events.pes.edu/
         https://events.pes.edu/10657#new_tab
         https://www.youtube.com/watch?v=5PllbH06p50
         https://pes.edu/bca
         https://pes.edu/tag/new/
         https://pes.edu/tag/undergraduate/
         https://pes.edu/bscmedical
         https://pes.edu/bsceconomics
         https://pes.edu/law/
```

Question 5:

Write a Python program to read the given website and extract the phone numbers and emails and contact addresses from Chennai, Amaravathi, Bhopal vit website. (5 Marks)

In [ ]:
```
import re
from bs4 import BeautifulSoup as soup
import requests
seedURL5=["https://vitap.ac.in/","https://vitbhopal.ac.in/","https://chennai.vit.ac
f = open("myfile.txt", "w")
for url in seedURL5:
    response = requests.get(url,verify=False)
    print("Status of the response : ", response.status_code)
    rootPage=soup(response.content,'html.parser')
    phpattern = re.compile(r'[7-9][0-9]{9}')
    emailpattern=re.compile(r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b')
    phone_numbers=re.findall(phpattern,str(rootPage))
    email_ids=re.findall(emailpattern,str(rootPage))
    f.write("-------{} Email IDS and Phone Number-------\n".format(url))
    f.write((' '.join(email_ids)))
    f.write((' '.join(phone_numbers)))
    f.write("\n")
f.close()
f2 = open("myfile.txt", "r")
print(f2.read())
```

```
C:\Users\ayuar\AppData\Local\Programs\Python\Python310\lib\site-packages\urllib3\c
onnectionpool.py:1045: InsecureRequestWarning: Unverified HTTPS request is being m
ade to host 'vitap.ac.in'. Adding certificate verification is strongly advised. Se
e: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
Status of the response :  200
C:\Users\ayuar\AppData\Local\Programs\Python\Python310\lib\site-packages\urllib3\c
onnectionpool.py:1045: InsecureRequestWarning: Unverified HTTPS request is being m
ade to host 'vitbhopal.ac.in'. Adding certificate verification is strongly advise
d. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
Status of the response :  200
C:\Users\ayuar\AppData\Local\Programs\Python\Python310\lib\site-packages\urllib3\c
onnectionpool.py:1045: InsecureRequestWarning: Unverified HTTPS request is being m
ade to host 'chennai.vit.ac.in'. Adding certificate verification is strongly advis
ed. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
```

Status of the response : 200
-------https://vitap.ac.in/ Email IDS and Phone Number-------
7868934148 7868934148 7868934148 7797590012 9520281387
-------https://vitbhopal.ac.in/ Email IDS and Phone Number-------
info@vitbhopal.ac.in placement@vitbhopal.ac.in admissions@vitbhopal.ac.in wardenmh
@vitbhopal.ac.in wardenlh@vitbhopal.ac.in9662422233 8194455437 8262477098 88766599
32 7795006127 8612877226 7709999738 7986318372 8845910171 8876348337 8945379300 89
45479108 8612870136 8845918969 8845910863 8845911379 8845911923 8194455437 8262477
098 8876348337 8945379300 8945479108 8876659932 7795006127 8612877226 8612870136 7
709999738 8061303498 8061303498 7035880534 8864788979 7035880534 8864788979 703588
0534 8864788979 7035880534 8864788979 7035880534 8864788979 7035880534 8864788979
8061303498 8264704521 8264704521 8551857321 7465352127 7174252163 7465352127 71742
52163 7465352127 7174252163 7465352127 7174252163 7465352127 7174252163 7465352127
7174252163 7979272149 7979272149 8728174513 9344318998 9344318998 9344318998 93443
18998 9344318998 9344318998 7906040746 7906040746 9568028126 8286208450 7317899495
8286208450 7317899495 8286208450 7317899495 8286208450 7317899495 8286208450 73178
99495 8286208450 7317899495 8150108223 8150108223 8797830589 8318240920 9602035035
8318240920 9602035035 8318240920 9602035035 8318240920 9602035035 8318240920 96020
35035 8318240920 9602035035 7977097410 7977097410 7662683301 9741738558 8399911470
9741738558 8399911470 9741738558 8399911470 9741738558 8399911470 9741738558 83999
11470 9741738558 8399911470 8324916849 8324916849 9672133224 8482824318 9672133224
8482824318 9672133224 8482824318 9672133224 8482824318 9672133224 8482824318 96721
33224 8482824318 8324916849 8113170000 7729692478 7397009434 9618935664 9260750302
7217597803 7075121594 9065360540 9275123713 8756951133 8955924605 9158682827 96552
73119 8113170000 7729692478 7729692478 7729692478 7729692478 7729692478 7729692478
7939801444 8213782599 7254431471 8503397689 7710044413 7859852193 7528554828 82460
52324 7107344378 9287253478 9906576570 8377194160 8351581336 7089927560 8180286903
8443864678 8797978927 7939801444 8213782599 7254431471 8213782599 7254431471 82137
82599 7254431471 8213782599 7254431471 8213782599 7254431471 8213782599 7254431471
7979757306 7979757306 8110484234 7831648607 8110484234 7831648607 8110484234 78316
48607 8110484234 7831648607 8110484234 7831648607 8110484234 7831648607 7979757306
7997123712 7997123712 8061489449 8187953360 8061489449 8187953360 8061489449 81879
53360 8061489449 8187953360 8061489449 8187953360 8061489449 8187953360 7997123712
7953510457 7953510457 8358086840 7958327780 8358086840 7958327780 8358086840 79583
27780 8358086840 7958327780 8358086840 7958327780 8358086840 7958327780 7953510457
7926112081 7926112081 8941056685 7563129879 8941056685 7563129879 8941056685 75631
29879 8941056685 7563129879 8941056685 7563129879 8941056685 7563129879 7926112081
7893933214 7893933214 7998754603 9141007060 7998754603 9141007060 7998754603 91410
07060 7998754603 9141007060 7998754603 9141007060 7998754603 9141007060 7893933214
7888279366 7888279366 7078686666 9281519859 7078686666 9281519859 7078686666 92815
19859 7078686666 9281519859 7078686666 9281519859 7078686666 9281519859 7888279366
7892372833 7892372833 7097665995 9446229641 7097665995 9446229641 7097665995 94462
29641 7097665995 9446229641 7097665995 9446229641 7097665995 9446229641 7892372833
7968996840 7968996840 7496227456 8514161077 7496227456 8514161077 7496227456 85141
61077 7496227456 8514161077 7496227456 8514161077 7496227456 8514161077 7968996840
7974585219 7974585219 9986160257 7917539790 9986160257 7917539790 9986160257 79175
39790 9986160257 7917539790 9986160257 7917539790 9986160257 7917539790 7974585219
8312719179 8851818198 9034194288 9216018669 9344396816 8998725771 7553521655 84354
24529 8213626783 8312719179 8851818198 9034194288 8851818198 9034194288 8851818198
9034194288 8851818198 9034194288 8851818198 9034194288 8851818198 9034194288 79703
12688 7970312688 7884614597 7449419967 7884614597 7449419967 7884614597 7449419967
7884614597 7449419967 7884614597 7449419967 7884614597 7449419967 7970312688 79703
12688 7884614597 7449419967 8312719179 8851818198 9034194288 7968996840 7496227456
8514161077 7974585219 9986160257 7917539790 7892372833 7097665995 9446229641 78882
79366 7078686666 9281519859 7926112081 8941056685 7563129879 7893933214 7998754603
9141007060 7953510457 8358086840 7958327780 7979757306 8110484234 7831648607 79971
23712 8061489449 8187953360 8324916849 9672133224 8482824318 8113170000 7729692478
7939801444 8213782599 7254431471 8264704521 7465352127 7174252163 7979272149 93443
18998 7906040746 8286208450 7317899495 8150108223 8318240920 9602035035 7977097410
9741738558 8399911470 8061303498 7035880534 8864788979 7986318372 8845910171 88459
18969 8845910863 8845911379 8845911923 9212651252

-------https://chennai.vit.ac.in/ Email IDS and Phone Number-------
cw.cc@vit.ac.in wlh.cc@vit.ac.in transport.cc@vit.ac.in cw.cc@vit.ac.in wlh.cc@vi
t.ac.in transport.cc@vit.ac.in admin.chennai@vit.ac.in8272122876 8272215041 827223
9968 8272270359 8272298574 8659071608 8588684987 8679772278 8634261762 8634270077
9038225429 7121194580 8634261762 8588708437 8634261762 8588708437 8201114257 82011
14257 8588708437 8588708437 8659971350 9315459946 8679772278 7358782569 8272122876
8634261762 9038225429 8634270077 7121194580 7358782569 8634261762 8272215041 85887
08437 8272239968 8634261762 8272270359 8588708437 8272298574 8201114257 8201114257
8588708437 8659071608 8588708437 8659971350 9315459946 8588684987