Programme	:	**B.Tech** Semester : **Win Sem 21-22**
Course	:	**Web Mining Lab** Code : **CSE3024**
Faculty	:	**Dr.Bhuvaneswari A** Slot : **L7 + L8**
Date	:	**07-01-2022** Marks : **10 Points**

Vaibhav Agarwal

19BCE1413

Exercise 1: Simple Web Crawlers

1. Given a root URL, e.g., "Vit.ac.in", Design a simple crawler to return all pages that contains a string "research" from this site.

Code with execution:

```
import requests
from bs4 import BeautifulSoup
import re


root_URL = "http://www.vit.ac.in"
search_word = "research"

response = requests.get(root_URL)
print("Status of the response : ", response.status_code)


 root_page = BeautifulSoup(response.content, 'html.parser')

anchor_tags = root_page.find_all('a')
result = []


for anchor_tag in anchor_tags :
  link = anchor_tag['href']
```

```
  if re.search(search_word, link, re.IGNORECASE) :
    result.append(link)


print("The links in the root URL page which contains the word
'research' are:")
for url in result :
  print("\t", url)
```

OUTPUT:

The links in the root URL page which contains the word 'research' are:

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

http://info.vit.ac.in/Faculty-Research-Awards/default.htm

3d-printing-play-major-role-mitigating-spread-covid-19-say-researchers-vit

3d-printing-play-major-role-mitigating-spread-covid-19-say-researchers-vit

2. Find documents that contain the word “admissions” and the word “international” within the URL “Vit.ac.in” using Python.

CODE:

```
import requests

from bs4 import BeautifulSoup

import re

root_URL = "http://www.vit.ac.in"

search_words = ['admissions', 'international']
```

```
response = requests.get(root_URL)

print("Status of the response : ", response.status_code)

root_page = BeautifulSoup(response.content,

'html.parser')

anchor_tags = root_page.find_all('a')

valid_links = []

for anchor_tag in anchor_tags :

    link = anchor_tag['href']

    if link.startswith("http") :

        if link not in valid_links :

            valid_links.append(link)




print("The number of documents/pages linked to the current root

page is : ", len (valid_links))

result = []

failed = []

for link in valid_links :
```

```
    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(search_words[0], page, re.IGNORECASE)) and
 (re.search(search_words[1], page, re.IGNORECASE)) :
        result.append(link)
print("The links in the root URL page which contains the word
 'admissions', and 'international' are :")
for url in result :
    print("\t", url)
```

```
The links in the root URL page which contains the word 'admissions', and
'international' are :
```

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/ranking-and-accreditation

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/mhrdugc

https://vit.ac.in/about/news-letter

https://vit.ac.in/academics/home

https://vit.ac.in/programmes-offered-2021-22

https://vit.ac.in/programmes-offered-2020-21

https://vit.ac.in/schools

https://vit.ac.in/academics/ffcs

https://vit.ac.in/academics/library

https://vit.ac.in/academics-feedback

https://vit.ac.in/admissions/overview

https://vit.ac.in/admissions/programmes-offered

https://vit.ac.in/all-courses/ug

https://vit.ac.in/all-courses/pg

https://vit.ac.in/admissions/research

https://vit.ac.in/admissions/international

https://vit.ac.in/stars-support-advancement-rural-students-0

https://vit.ac.in/placements/overview

https://vit.ac.in/career-development-centre

https://vit.ac.in/placements/superdreamoffers

https://vit.ac.in/placements/dreamoffers

https://vit.ac.in/placements/internship

https://vit.ac.in/placements/statistics
https://vit.ac.in/placements/pat-Office

https://vit.ac.in/career-development-centrecdc-contact

https://vit.ac.in/InternationalRelations

https://vit.ac.in/internationalrelations/itp

https://vit.ac.in/internationalrelations/partneruniversities

https://vit.ac.in/internationalrelations/sap

https://vit.ac.in/admissions/international/overview

https://vit.ac.in/academics-more/Contact us

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/campuslife/overview

https://vit.ac.in/campuslife/fests

https://vit.ac.in/campuslife/studentswelfare

https://vit.ac.in/campuslife/sports

https://vit.ac.in/campuslife/hostels

https://vit.ac.in/campuslife/startups
https://vit.ac.in/campuslife/healthservices

https://vit.ac.in/campuslife/otheramenities

https://vit.ac.in/detailview/green-vit

https://vit.ac.in/academics/coe

https://vit.ac.in/transcripts-alumni

https://vit.ac.in/centers/asc

https://vit.ac.in/campus-category/Counselling-Division

https://vit.ac.in/guest-house

https://vit.ac.in/redressal

https://vit.ac.in/hotels-in-vellore

https://vit.ac.in/anti-ragging-committee

https://vit.ac.in/capability-enhancement-scheme

https://vit.ac.in/internal-complaints-committee

https://vit.ac.in/academics/transcripts

https://vit.ac.in/instruction

https://vit.ac.in/alumni-events

https://vit.ac.in/detailview/alumni-photo-gallery

https://vit.ac.in/alumni-office-contact

https://www.youtube.com/c/VITUniversityVellore

https://vit.ac.in/school-mechanical-engineeringsmec/virtual-international-conference-product design-development-and

https://vit.ac.in/school-civil-engineering-sce/2nd-international-conference-recent-trends-cons

truction-materials-and

https://vit.ac.in/school-computer-science-and-engineering-scope/international-conference-computational-methods-and

https://vit.ac.in/school-electrical-engineering-select/innovations-power-and-advanced-compu t ing-technologies-i

https://vit.ac.in/applications-open-2021-22

https://vit.ac.in/detailview/35th-annual-convocation

https://vit.ac.in/detailview/vit-wishes-warm-%E2%80%98happy-birthday%E2%80%99-our-honourable-chancellor

https://vit.ac.in/vit-institution-eminence-ioe

https://vit.ac.in/ariia-award

https://vit.ac.in/qs-ranks-vit-one-among-top-12-institutions-india-engineering-and-technology

https://vit.ac.in/world-university-rankings-2020

https://vit.ac.in/vit-university-sets-record-limca-book-records

https://vit.ac.in/vellore-institute-technology-vit-triumphs-tata-steel-materialnext-20-0

https://vit.ac.in/vit-donates-%E2%82%B9150-cr-cm%E2%80%99s-fund

https://vit.ac.in/international-yoga-day-2021-0

https://vit.ac.in/inauguration-vit-fruit-orchard-planting

https://vit.ac.in/galleries

https://vit.ac.in/video

https://vit.ac.in/campus-hostel/hostels

https://vit.ac.in/academics/iqac

https://vit.ac.in/iprcell

https://vit.ac.in/campus-category/grievancecell

https://vit.ac.in/contactus

```
print("The links that we failed to open are : ")

for url in failed :

    print("\t", url)
```

The links that we failed to open are :

http://intranet.vit.ac.in

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

```
import requests

from bs4 import BeautifulSoup

import re

root_URL = "http://www.vit.ac.in"

search_word_1 = "Programme"

search_word_2 = "Programming"
```

```
response = requests.get(root_URL)

print("Status of the response : ", response.status_code)

root_page = BeautifulSoup(response.content, 'html.parser')

anchor_tags = root_page.find_all('a')

valid_links = []

for anchor_tag in anchor_tags :

    link = anchor_tag['href']

    if link.startswith("http") :

        if link not in valid_links :
            valid_links.append(link)

print("The number of documents/pages linked to the current root
page is : ", len (valid_links))

result = []

failed = []

for link in valid_links :

    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(search_word_1, page, re.IGNORECASE)) and (not
re.search(search_word_2, page, re.IGNORECASE)) :
    result.append(link)
print("The links in the root URL page which contains the word
 'Programme' but not the word 'programming' are :")
for url in result :
    print("\t", url)
```

The links in the root URL page which contains the word 'Programme' but not the word 'programming' are :

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/ranking-and-accreditation

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/national-institutional-ranking-framework-nirf

https://vit.ac.in/mhrdugc

https://vit.ac.in/about/news-letter

https://vit.ac.in/academics/home

https://vit.ac.in/sites/default/files/academic/Academic-Regulations.pdf

https://vit.ac.in/programmes-offered-1

https://vit.ac.in/programmes-offered-2021-22
https://vit.ac.in/programmes-offered-2020-21

```
print("The links that we failed to open are : ")

for url in failed :

    print("\t", url)
```

```
The links that we failed to open are :

        http://intranet.vit.ac.in
```

http://intranet.vit.ac.in/

4. Write a web crawler program which takes as input a url(Educational Website), a search 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.

CODE:

```
import requests

from bs4 import BeautifulSoup

import re

root_URL = input("Input URL:")

search_word = input("Search Word: ")

Max_pages = int(input("Max Pages"))

response = requests.get(root_URL)

print("Status of the response : ", response.status_code)

root_page = BeautifulSoup(response.content, 'html.parser')
```

```
anchor_tags = root_page.find_all('a')

result = []

found = False
for anchor_tag in anchor_tags :

    link = anchor_tag['href']

    while (found == False and Max_pages > 0):

       Max_pages -=1

    if re.search(search_word, link, re.IGNORECASE):

        result.append(link)

        found = True

        break

    else:

        result.append(link)

    if(found == False):

        result = ["failure"]

print("The links in the root URL are given below")

for url in result :

    print("\t", url)
```

```
print("The links that we failed to open are : ")
for url in failed :
    print("\t", url)
```

INPUT:

https://www.annauniv.edu/

Hostel

16

OUTPUT:

The links in the root URL are given below

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

http://www.annauniv.edu/index.php

failure