VELLORE INSTITUTE OF TECHNOLOGY

VIT

உழைப்பே உயர்வு தரும்

CSE 3024 – Web Mining

Week 1

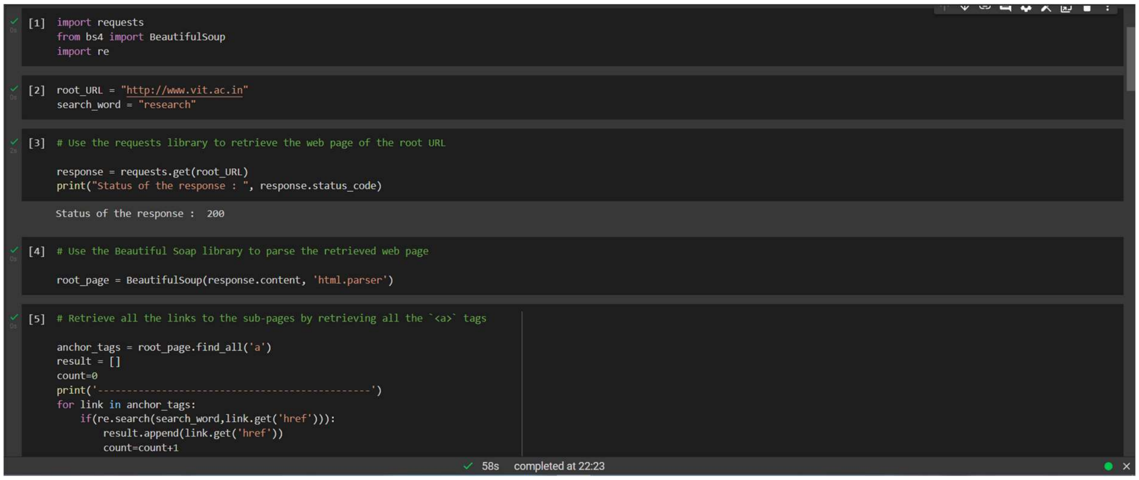
Name-Tanmay Mahajan

Reg. no.-19bce1735

Prof.- Dr. Bhuvaneswari Anbalagan

Exercise 1: Simple Web Crawlers

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). Store the crawled urls into a .txt/ .csv/ .xls file for quick retrieval in future. (CT)



[1]

```python
import requests
from bs4 import BeautifulSoup
import re
```

[2]

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

[3]

```python
# Use the requests library to retrieve the web page of the root URL

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

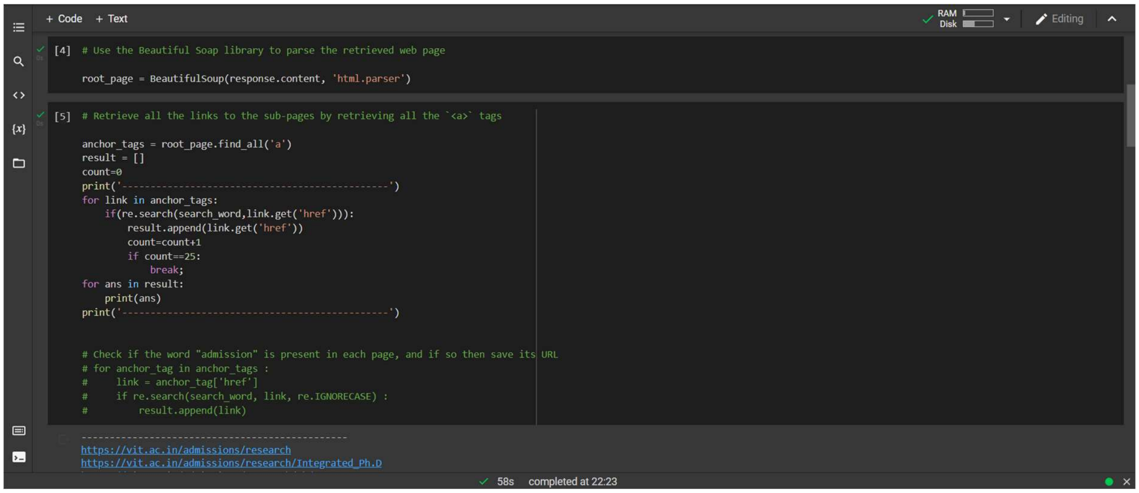
```
Status of the response :  200
```

[4]

```python
# Use the Beautiful Soap library to parse the retrieved web page

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

[5]

```python
# Retrieve all the links to the sub-pages by retrieving all the `<a>` tags

anchor_tags = root_page.find_all('a')
result = []
count=0
print('----------------------------------------------')
for link in anchor_tags:
    if(re.search(search_word,link.get('href'))):
        result.append(link.get('href'))
        count=count+1
```



[4]

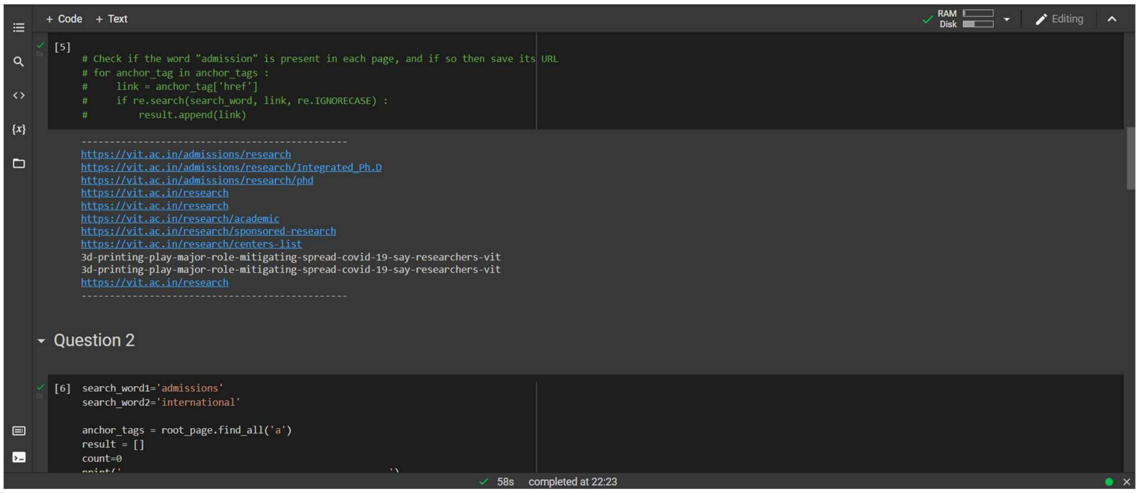
```python
# Use the Beautiful Soap library to parse the retrieved web page

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

[5]

```python
# Retrieve all the links to the sub-pages by retrieving all the `<a>` tags

anchor_tags = root_page.find_all('a')
result = []
count=0
print('----------------------------------------------')
for link in anchor_tags:
    if(re.search(search_word,link.get('href'))):
        result.append(link.get('href'))
        count=count+1
        if count==25:
            break;
for ans in result:
    print(ans)
print('----------------------------------------------')


# Check if the word "admission" is present in each page, and if so then save its URL
# for anchor_tag in anchor_tags :
#     link = anchor_tag['href']
#     if re.search(search_word, link, re.IGNORECASE) :
#         result.append(link)
```

```
----------------------------------------------
https://vit.ac.in/admissions/research
https://vit.ac.in/admissions/research/Integrated_Ph.D
```

Output-



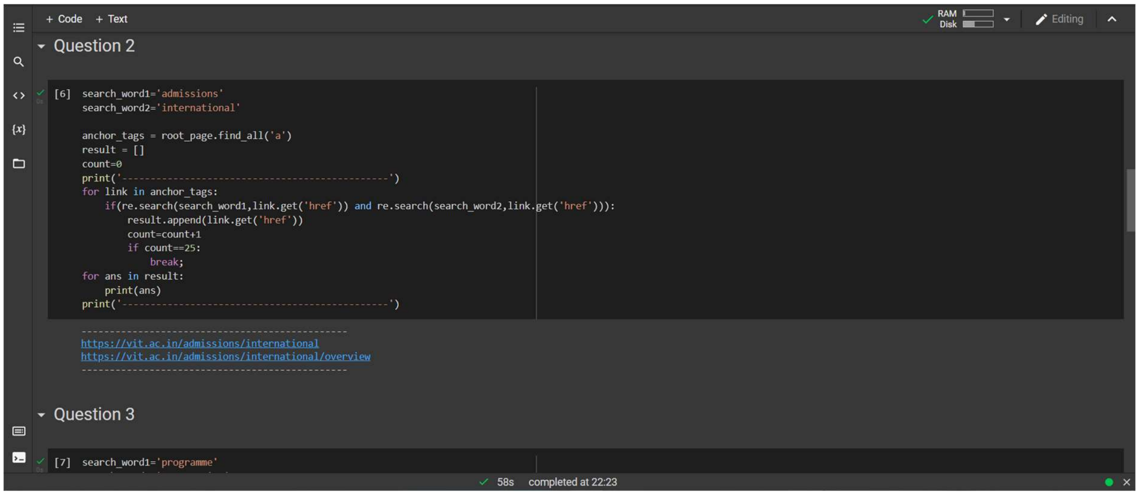
```
[5]
    # Check if the word "admission" is present in each page, and if so then save its URL
    # for anchor_tag in anchor_tags :
    #     link = anchor_tag['href']
    #     if re.search(search_word, link, re.IGNORECASE) :
    #         result.append(link)
```

```
    ----------------------------------------------
    https://vit.ac.in/admissions/research
    https://vit.ac.in/admissions/research/Integrated_Ph.D
    https://vit.ac.in/admissions/research/phd
    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
    3d-printing-play-major-role-mitigating-spread-covid-19-say-researchers-vit
    3d-printing-play-major-role-mitigating-spread-covid-19-say-researchers-vit
    https://vit.ac.in/research
    ----------------------------------------------
```

Question 2

```
[6] search_word1='admissions'
    search_word2='international'

    anchor_tags = root_page.find_all('a')
    result = []
    count=0
```

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



Question 2

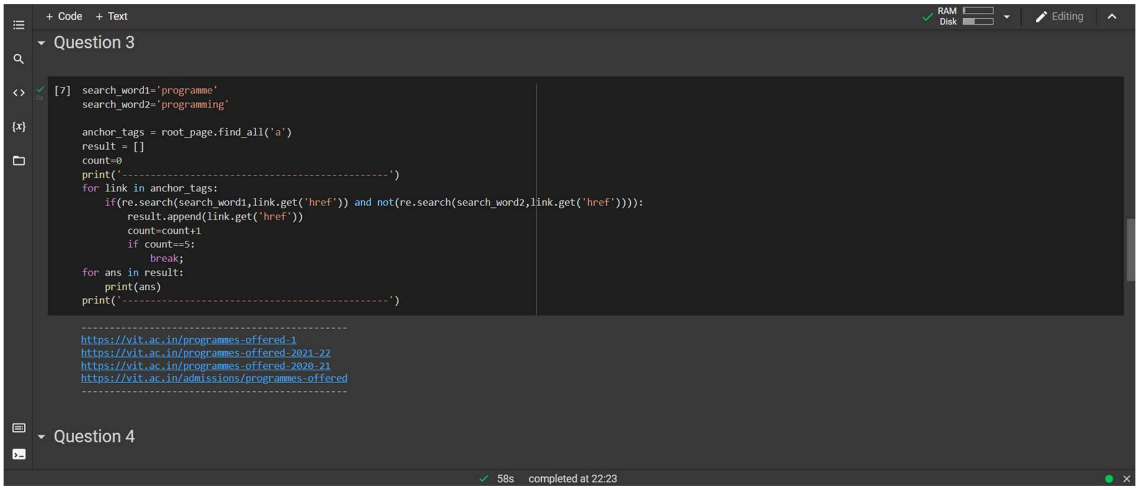
```
[6] search_word1='admissions'
    search_word2='international'

    anchor_tags = root_page.find_all('a')
    result = []
    count=0
    print('----------------------------------------------')
    for link in anchor_tags:
        if(re.search(search_word1,link.get('href')) and re.search(search_word2,link.get('href'))):
            result.append(link.get('href'))
            count=count+1
            if count==25:
                break;
    for ans in result:
        print(ans)
    print('----------------------------------------------')
```

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

Question 3

```
[7] search_word1='programme'
```

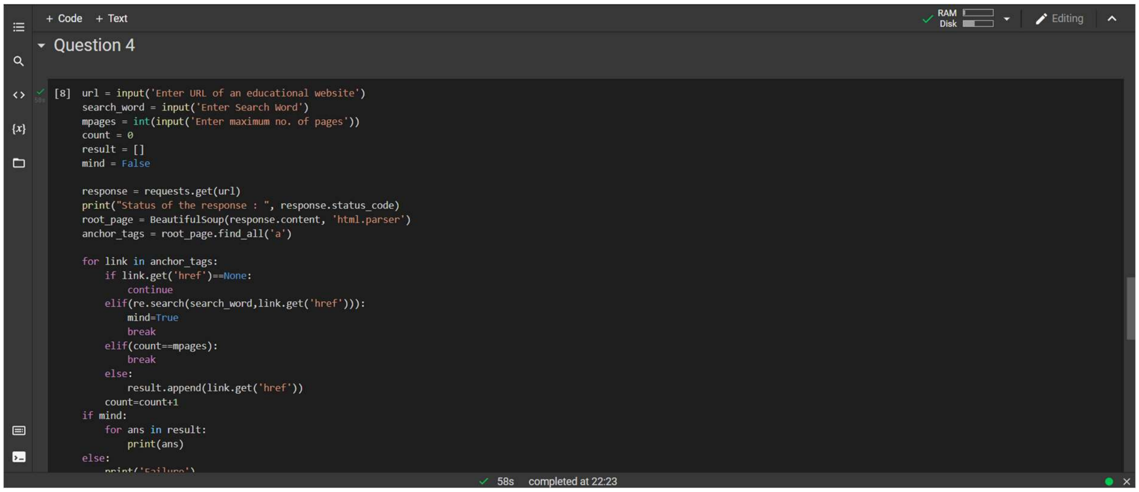
3.Find documents that contain the word “Programme” but not the word “programming” within the URL “Vit.ac.in” using Python. (5 pages)



```
[7] search_word1='programme'
    search_word2='programming'

    anchor_tags = root_page.find_all('a')
    result = []
    count=0
    print('---------------------------------------------')
    for link in anchor_tags:
        if(re.search(search_word1,link.get('href')) and not(re.search(search_word2,link.get('href')))):
            result.append(link.get('href'))
            count=count+1
            if count==5:
                break;
    for ans in result:
        print(ans)
    print('---------------------------------------------')
```

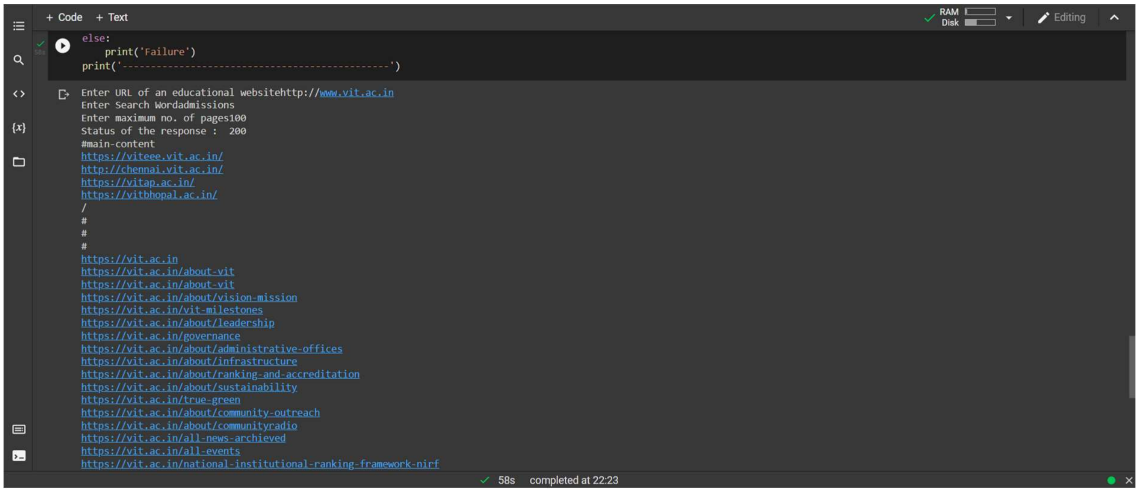
```
---------------------------------------------
https://vit.ac.in/programmes-offered-1
https://vit.ac.in/programmes-offered-2021-22
https://vit.ac.in/programmes-offered-2020-21
https://vit.ac.in/admissions/programmes-offered
---------------------------------------------
```

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.



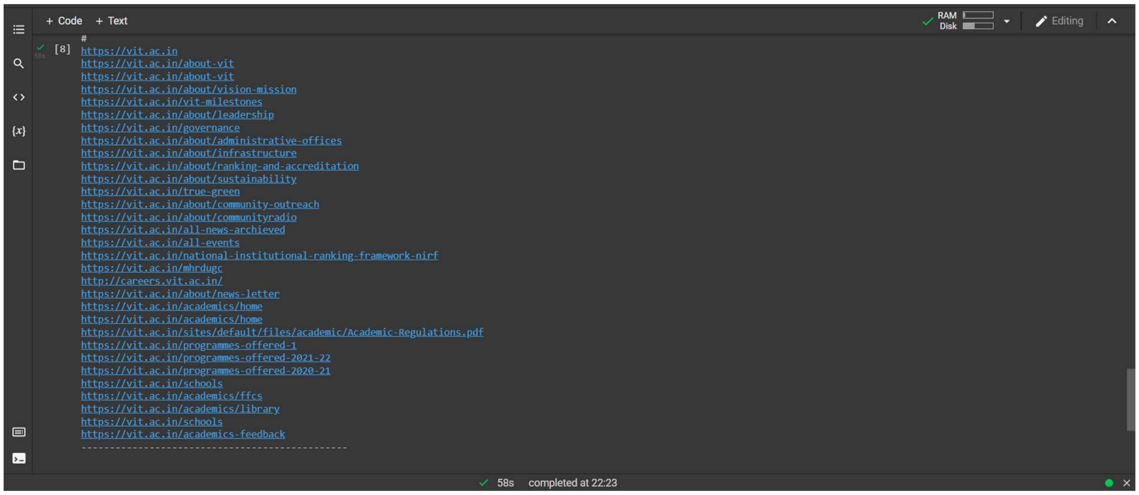
```
[8] url = input('Enter URL of an educational website')
    search_word = input('Enter Search Word')
    mpages = int(input('Enter maximum no. of pages'))
    count = 0
    result = []
    mind = False

    response = requests.get(url)
    print("Status of the response : ", response.status_code)
    root_page = BeautifulSoup(response.content, 'html.parser')
    anchor_tags = root_page.find_all('a')

    for link in anchor_tags:
        if link.get('href')==None:
            continue
        elif(re.search(search_word,link.get('href'))):
            mind=True
            break
        elif(count==mpages):
            break
        else:
            result.append(link.get('href'))
        count=count+1
    if mind:
        for ans in result:
            print(ans)
    else:
        print('Failure')
```

Output-



```
else:
    print('Failure')
print('-----------------------------------------------')
```

```
Enter URL of an educational websitehttp://www.vit.ac.in
Enter Search Wordadmissions
Enter maximum no. of pages100
Status of the response :  200
#main-content
https://viteee.vit.ac.in/
http://chennai.vit.ac.in/
https://vitap.ac.in/
https://vitbhopal.ac.in/
/
#
#
#
https://vit.ac.in
https://vit.ac.in/about-vit
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
https://vit.ac.in/about-vit
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
http://careers.vit.ac.in/
https://vit.ac.in/about/news-letter
https://vit.ac.in/academics/home
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
https://vit.ac.in/schools
https://vit.ac.in/academics/ffcs
https://vit.ac.in/academics/library
https://vit.ac.in/schools
https://vit.ac.in/academics-feedback
-----------------------------------------------
```