

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Programme | : | **B.Tech** | Semester | : | **WIN Sem 21-22** |
| Course | : | **Web Mining Lab** | Code | : | **CSE3024** |
| Faculty | : | **Dr.Bhuvaneswari A** | Slot | : | **L7+L8** |
| Date | : | **04-1-2022** | Marks | : | **10 Points** |

**Exercise 1: Simple Web Crawlers**

**GOOGLE COLAB LINK:**

[**https://colab.research.google.com/drive/1ffKBrGY\_FJIjbZ7phKTcorPshZaZyjBF?usp=sharing**](https://colab.research.google.com/drive/1ffKBrGY_FJIjbZ7phKTcorPshZaZyjBF?usp=sharing)

**PROCEDURE:**

1. Import the required libraries i.e requests, beautiful Soup, and re

2. Set the root url and word used for web crawling accordingly

3. Check for the status of the response (200 for ok)

4. Parse the page and retrieve all the anchor tags

5. Store the href attribute in temporary variable ‘link’ and append the link in the result array

6. Print the array

**Question 1:** Given a seed/root URL, e.g., "Vit.ac.in", Design a simple crawler to return allpages (URLs) that contains a keyword "research" from this site. (25 pages)

**Code:**

import requests

from bs4 import BeautifulSoup

import re

root\_URL="http://www.vit.ac.in"

search\_word = "admissions"

#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

#use the beautiful soup library to parse the retrieved web page

root\_page = BeautifulSoup(response.content,'html.parser')

#retrieve all the links to the sub pages by retrieving all the <a> tags

anchor\_tags = root\_page.find\_all('a')

result = []

#check if the word "admission" is present in each page

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 are: ")

for url in result :

  print("\t", url)

The links in the root URL page are:

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

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

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

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

<https://vit.ac.in/admissions/research/Integrated_Ph.D>

<https://vit.ac.in/admissions/research/phd>

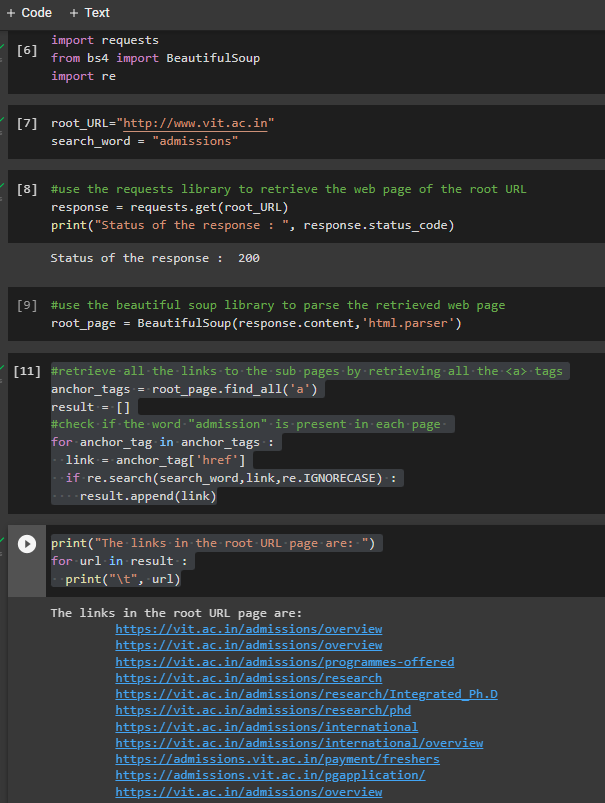
<https://vit.ac.in/admissions/international>

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

<https://admissions.vit.ac.in/payment/freshers>

<https://admissions.vit.ac.in/pgapplication/>

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



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

**Code:**

root\_URL="http://www.vit.ac.in"

search\_word1 = "admissions"

search\_word2 = "international"

#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#use the beautiful soup library to parse the retrieved web page

root\_page = BeautifulSoup(response.content,'html.parser')

#retrieve all the links to the sub pages by retrieving all the <a> tags

anchor\_tags = root\_page.find\_all('a')

result = []

#check if the word "admission" is present in each page

for anchor\_tag in anchor\_tags :

  link = anchor\_tag['href']

  if re.search(search\_word1,link,re.IGNORECASE) and re.search(search\_word2,link,re.IGNORECASE):

    result.append(link)

print("The links in the root URL page are: ")

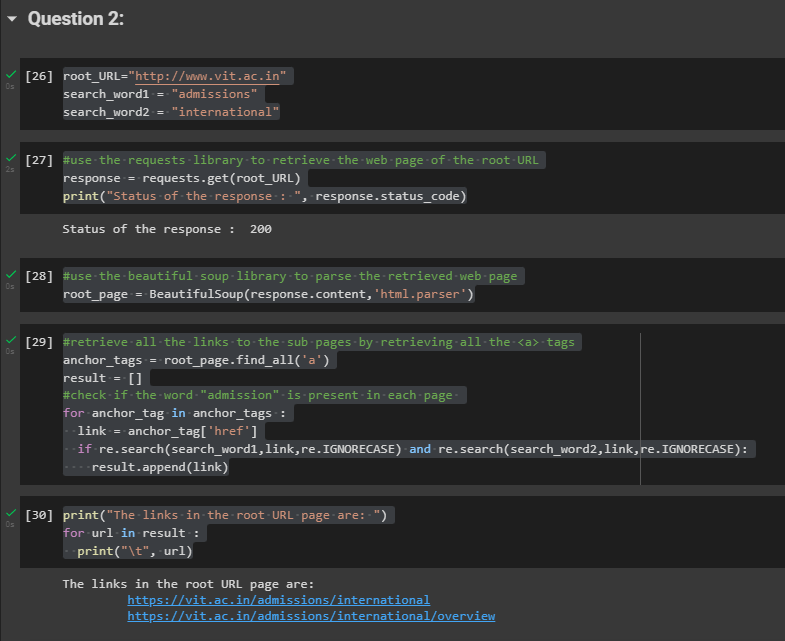
for url in result :

  print("\t", url)

The links in the root URL page are:

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

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



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

**Code:**

root\_URL="http://www.vit.ac.in"

search\_word1 = "programme"

search\_word2 = "programming"

#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#use the beautiful soup library to parse the retrieved web page

root\_page = BeautifulSoup(response.content,'html.parser')

#retrieve all the links to the sub pages by retrieving all the <a> tags

anchor\_tags = root\_page.find\_all('a')

result = []

#check if the word "admission" is present in each page

for anchor\_tag in anchor\_tags :

  link = anchor\_tag['href']

  if re.search(search\_word1,link,re.IGNORECASE) and not re.search(search\_word2,link,re.IGNORECASE):

    result.append(link)

print("The links in the root URL page are: ")

for url in result :

  print("\t", url)

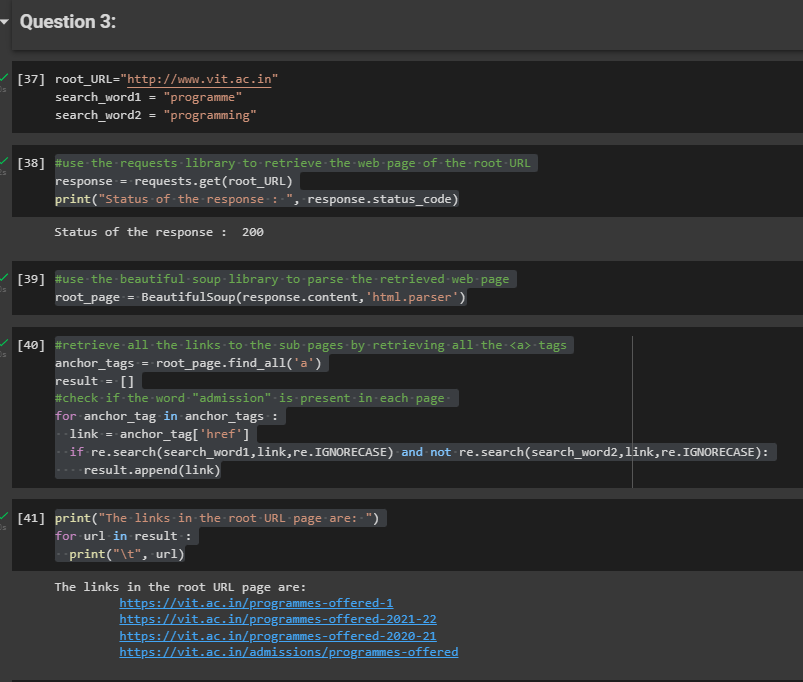
The links in the root URL page are:

<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>

****

**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.

**Code:**

root\_URL=input('enter the name of the website you want to crawl:')

search\_word=input('enter the name of the keyword you want: ')

n=int(input('entet the number of pages to be checked: '))

enter the name of the website you want to crawl:<https://www.iitg.ac.in/>

enter the name of the keyword you want: academics

entet the number of pages to be checked: 25

#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#use the beautiful soup library to parse the retrieved web page

root\_page = BeautifulSoup(response.content,'html.parser')

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

anchor\_tags = root\_page.find\_all('a')

result = []

counter = n

# Check if the word "admission" is present in each page, and if so then save its URL

for anchor\_tag in anchor\_tags:

  if counter !=1 :

    counter = counter -1

    link = anchor\_tag['href']

    print("\t", link)

    if re.search(search\_word, link, re.IGNORECASE) and result.count == 0 :

        result.append(link)

        break

# javascript:void(0)

javascript:void(0)

javascript:void(0)

/?lang=hindi&lang=hindi

<https://login.microsoftonline.com/>

iitg\_page\_details.php?page=16/screen-reader

<https://iitg.ac.in/old>

<https://www.facebook.com/iitgwt/>

<https://twitter.com/IITGuwahati>

<https://www.linkedin.com/school/iitg/>

<https://www.instagram.com/iitgwt/>

<https://www.youtube.com/IITGuwahatiOfficial>

<https://www.iitg.ac.in/acad/contactus.php>

iitg\_page\_details?page=25/frequently-asked-questions

iitg\_videos

iitg\_tenders\_all

pdf/c05cfcc8b4cd031b6377462e59575e1c-12-51-170.pdf

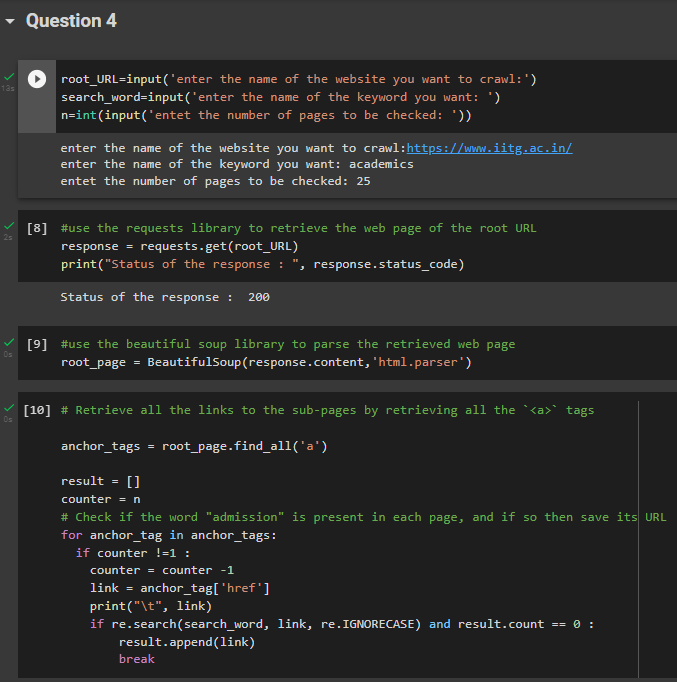
iitg\_page\_details.php?page=5/visitor-s-information

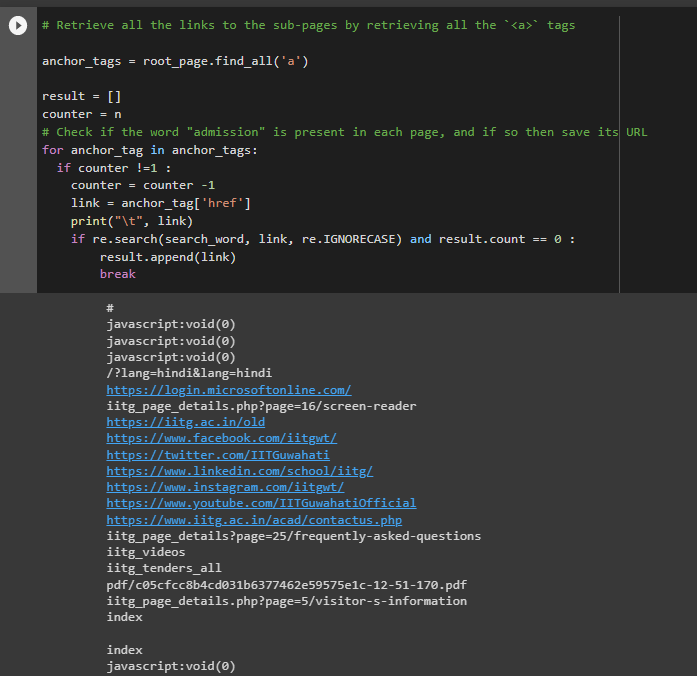
index

index

javascript:void(0)

iitg\_page\_details.php?page=1/about-us

****

****