Web Mining Lab - 1

Aadhitya Swarnesh

[redacted]

- 8 February 2021

Question - 1

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

We use the BeautifulSoup library of Python in order to parse through the response.

We use the Requests library of Python in order to make the web page requests.

We use the RE library of Python for pattern matching irrespective of the case or capitalisation of the content in the website.

The code for the program is as follows :

Question 1

Given a root URL, e.g., `vit.ac.in` , Design a simple crawler using Python to return all pages that contains a string `admissions` from this site.

In [3]:

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

In [4]:

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

In [5]:

```
# 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
```

In [6]:

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

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

In [7]:

```
# 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, 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)
```

In [11]:

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

```
The links in the root URL page which contains the word 'admissions'
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/international
        https://vit.ac.in/admissions/international/overview
        https://admissions.vit.ac.in/pgapplication/
        https://admissions.vit.ac.in/irapplicationug/
        https://admissions.vit.ac.in/irapplicationug/
        https://admissions.vit.ac.in/pgapplication/
        https://admissions.vit.ac.in/pgirapplication/
        https://vit.ac.in/files/MBA_online_interview/MBA-2020-Admis
sions-online-interview-candidates-date-and-time-schedule.pdf
        https://vit.ac.in/admissions/programmes-offered
```

In [ ]:

```
```

Question - 2

Find documents that contain the word “Data” and the word “analytics” within the URL “Vit.ac.in” using Python.

We use the BeautifulSoup library of Python in order to parse through the response.

We use the Requests library of Python in order to make the web page requests.

We use the RE library of Python for pattern matching irrespective of the case or capitalisation of the content in the website.

After we get the root URL’s web page, we take all the anchor tags from the parsed structure, and then retrieve the anchor tag’s `href` property which holds the link for the other pages.

Using these links, we perform a get request to these pages and save only the links which have the words “data" and “analytics” somewhere in its document.

We use regular expression to find if a page contains the required words or not.

During the process of making such requests, there is a chance of failure like due to SSL certificate authentication and verification, or just a bad time to connect to the server in some cases.

The code for the program is as follows :

Question 2

Find documents that contain the word `Data` and the word `analytics` within the URL `vit.ac.in` using Python.

In [1]:

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

In [2]:

```
root_URL = "http://www.vit.ac.in"
search_words = ['data', 'analytics']
```

In [3]:

```
# 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
```

In [4]:

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

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

In [22]:

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

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

In [23]:

```
# Accumulate all the unique links from the anchor tags with valid syntax (starts
with http, and not just inter page reference)

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)
```

In [25]:

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

```
The number of documents/pages linked to the current root page is :
133
```

In [26]:

```
# Arrays to store the links

result = []
failed = []
```

In [27]:

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

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_w
ords[1], page, re.IGNORECASE)) :
            result.append(link)
```

```
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
```

In [28]:

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

The links in the root URL page which contains the word 'data', and 'analytics' are :

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/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/programmes-offered-2019-20
https://vit.ac.in/schools
https://vit.ac.in/academics/ffcs
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/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/placement-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/centers-list
https://vit.ac.in/research/sponsored-research
https://vit.ac.in/campuslife/overview
https://vit.ac.in/campuslife/fests
https://vit.ac.in/campuslife/studentswelfare
https://vit.ac.in/academics/library
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
http://www.vittbi.com/#/
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
http://www.vitaa.org/
https://campustour.vit.ac.in/
https://vit.ac.in/vit-among-top-9-institutions-india-shanghai-world-universities-ranking-2020
https://www.facebook.com/VITuniversity/
https://vit.ac.in/school-electronics-engineering-sense/2nd-international-conference-microelectronic-devices-circuits
https://vit.ac.in/school-bio-sciences-technology-sbst/agricultural-biotechnology-quality-assurance-and-testing-tissue
https://vit.ac.in/school-advanced-sciences-sas/6th-biennial-international-group-theory-conference-2021
https://vit.ac.in/general/world-nano-congress-advanced-science-and-technology-2021
https://vit.ac.in/research-interests-eligible-guides-dec-2020-phd-session
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://www.vidyalakshmi.co.in/Students/
http://info.vit.ac.in/CDAC/html/index3.html
https://vit.ac.in/52nd-death-anniversary-dr-c-n-annadurai-former-chief-minister-tamil-nadu-0
https://vit.ac.in/72nd-republic-day-celebration-0
https://vit.ac.in/great-placements-vit-students-inspite-pandemic
https://vit.ac.in/smt-rajeswari-viswanathan-memorial-inter-school-tournament-and-honouring-staff-vellore-corporation
https://vit.ac.in/galleries
http://campustour.vit.ac.in/
https://vit.ac.in/video
https://vit.ac.in/campus-hostel/hostels
https://vit.ac.in/vit-institution-eminence-ioe
https://vit.ac.in/iprcell
https://vit.ac.in/campus-category/grievancecell
https://vit.ac.in/academics/iqac
https://vitap.ac.in/careers/
http://www.vitaa.org
https://vit.ac.in/contactus

In [29]:

```
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
```

In [ ]:

```

```

Question - 3

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

We use the BeautifulSoup library of Python in order to parse through the response.

We use the Requests library of Python in order to make the web page requests.

We use the RE library of Python for pattern matching irrespective of the case or capitalisation of the content in the website.

After we get the root URL’s web page, we take all the anchor tags from the parsed structure, and then retrieve the anchor tag’s `href` property which holds the link for the other pages.

Using these links, we perform a get request to these pages and save only the links which have the words “data" and “analytics” somewhere in its document.

We use regular expression to find if a page contains the word “programme” or not. We similarly use it to ensure that the chosen pages do not contain the word “programming” in any form.

During the process of making such requests, there is a chance of failure like due to SSL certificate authentication and verification, or just a bad time to connect to the server in some cases. We also list these URL’s along with the output to demonstrate the efficiency of our algorithm and code.

The code for the program is as follows :

Question 3

Find documents that contain the word `Programme` and the word `Programming` within the URL `vit.ac.in` using Python.

In [1]:

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

In [2]:

```
root_URL = "http://www.vit.ac.in"
search_word_1 = "Programme"
search_word_2 = "Programming"
```

In [3]:

```
# 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
```

In [4]:

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

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

In [5]:

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

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

In [6]:

```
# Accumulate all the unique links from the anchor tags with valid syntax (starts
with http, and not just inter page reference)

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)
```

In [7]:

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

```
The number of documents/pages linked to the current root page is :
133
```

In [8]:

```
# Arrays to store the links

result = []
failed = []
```

In [9]:

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

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)
```

```
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
/usr/local/lib/python3.6/dist-packages/urllib3/connectionpool.py:84
7: InsecureRequestWarning: Unverified HTTPS request is being made. A
dding certificate verification is strongly advised. See: https://url
lib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
```

In [12]:

```
print("The links in the root URL page which contains the word 'Programme' but no
t the word 'programming' are :")
for url in result :
    print("\t", url)
```

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

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/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-2019-20
https://vit.ac.in/schools
https://vit.ac.in/academics/ffcs
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/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/placement-contact
https://vit.ac.in/InternationalRelations
https://vit.ac.in/internationalrelations/itp
https://vit.ac.in/internationalrelations/partneruniversitie
s
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/centers-list
https://vit.ac.in/research/sponsored-research
https://vit.ac.in/campuslife/overview
https://vit.ac.in/campuslife/fests
https://vit.ac.in/campuslife/studentswelfare
https://vit.ac.in/academics/library
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
http://www.vittbi.com/#/
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/sites/default/files/FormatGuidelines.doc
https://vit.ac.in/internal-complaints-committee
https://vit.ac.in/alumni_progression
https://vit.ac.in/academics/transcripts
https://vit.ac.in/instruction
https://vit.ac.in/vit-among-top-9-institutions-india-shanghai-world-universities-ranking-2020
https://www.facebook.com/VITuniversity/
https://www.youtube.com/c/VITUniversityVellore
https://vit.ac.in/school-electronics-engineering-sense/2nd-international-conference-microelectronic-devices-circuits
https://vit.ac.in/school-bio-sciences-technology-sbst/agricultural-biotechnology-quality-assurance-and-testing-tissue
https://vit.ac.in/school-advanced-sciences-sas/6th-biennial-international-group-theory-conference-2021
https://vit.ac.in/general/world-nano-congress-advanced-science-and-technology-2021
https://vit.ac.in/research-interests-eligible-guides-dec-2020-phd-session
https://youtu.be/2JK5Q8iVHBI
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/52nd-death-anniversary-dr-c-n-annadurai-former-chief-minister-tamil-nadu-0
https://vit.ac.in/72nd-republic-day-celebration-0
https://vit.ac.in/great-placements-vit-students-inspite-pandemic
https://vit.ac.in/smt-rajeswari-viswanathan-memorial-inter-school-tournament-and-honouring-staff-vellore-corporation
https://vit.ac.in/galleries
https://vit.ac.in/video
https://vit.ac.in/campus-hostel/hostels
https://vit.ac.in/vit-institution-eminence-ioe
https://vit.ac.in/iprcell
https://vit.ac.in/campus-category/grievancecell
https://vit.ac.in/academics/iqac
http://www.mhrdnats.gov.in/
https://vitap.ac.in/careers/
https://vit.ac.in/contactus

In [11]:

```
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
```

In [11]:

```

```