Web Mining Lab - 4

Aadhitya Swarnesh

[redacted]

- 3 March 2021

Question

Write a Python program to read the given website and extract the phone numbers and emails and contact addresses.

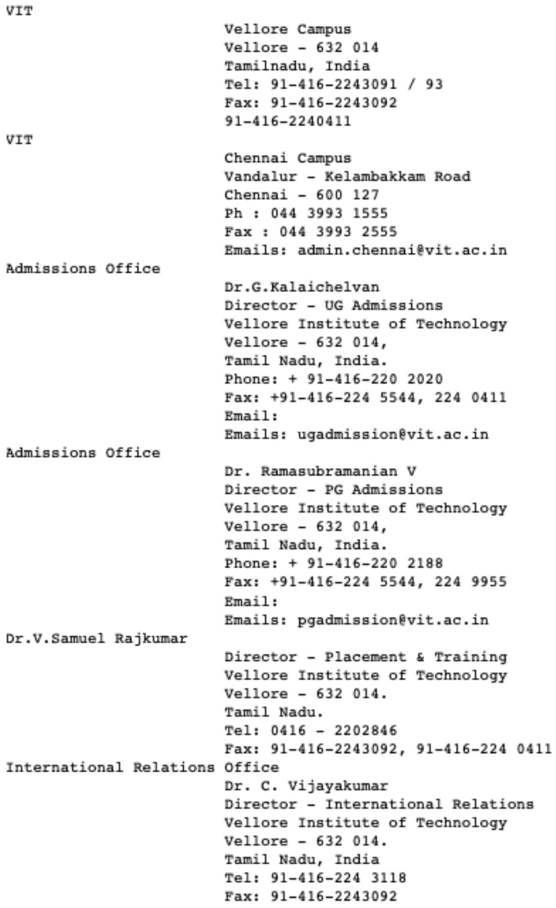
We will be using the Selenium library and the Python Programming language to accomplish this task.

We have first imported the necessary packages and then created a safari browser supported web driver using the selenium library.

We then use this library to open the VIT University home page. From here we locate the "contact us" link. We then use the driver to click on this link and redirect us to the "Contact Us" web page.

In this page, we can notice that there are three tables containing the contact details. The first table has the details of the university officials along with their phone numbers and addresses. The next two tables share common format stating the names, designations and their email addresses.

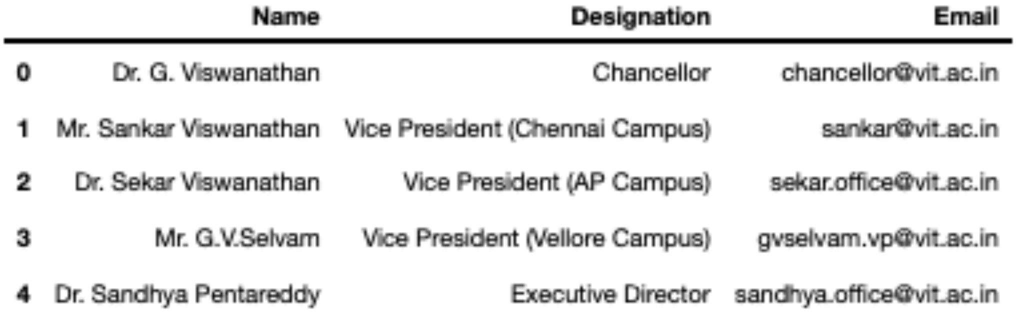
For the first table, we process differently than the other two. We get the text inside each of the cells. There are anchor tags and para tags etc placed randomly inside these cells and so we cannot take the text from these cells directly. So we carefully navigate through these DOM structures only if they are present, and gather the details carefully from the child nodes and the text directly placed. We run the email addresses through a REGEX filter to ensure its authenticity. We then store all these details in a text file. Each paragraph of this text file is a new contact information, and contains details like their phone and physical as well as their email addresses and phone numbers.



```
VIT
                        Vellore Campus
                        Vellore - 632 014
                        Tamilnadu, India
                        Tel: 91-416-2243091 / 93
                        Fax: 91-416-2243092
                        91-416-2240411
VIT
                        Chennai Campus
                        Vandalur - Kelambakkam Road
                        Chennai - 600 127
                        Ph : 044 3993 1555
                        Fax : 044 3993 2555
                        Emails: admin.chennai@vit.ac.in
Admissions Office
                        Dr.G.Kalaichelvan
                        Director - UG Admissions
                        Vellore Institute of Technology
                        Vellore - 632 014,
                        Tamil Nadu, India.
                        Phone: + 91-416-220 2020
                        Fax: +91-416-224 5544, 224 0411
                        Email:
                        Emails: ugadmission@vit.ac.in
Admissions Office
                        Dr. Ramasubramanian V
                        Director - PG Admissions
                        Vellore Institute of Technology
                        Vellore - 632 014,
                        Tamil Nadu, India.
                        Phone: + 91-416-220 2188
                        Fax: +91-416-224 5544, 224 9955
                        Email:
                        Emails: pgadmission@vit.ac.in
Dr.V.Samuel Rajkumar
                        Director - Placement & Training
                        Vellore Institute of Technology
                        Vellore - 632 014.
                        Tamil Nadu.
                        Tel: 0416 - 2202846
                        Fax: 91-416-2243092, 91-416-224 0411
International Relations Office
                        Dr. C. Vijayakumar
                        Director - International Relations
                        Vellore Institute of Technology
                        Vellore - 632 014.
                        Tamil Nadu, India
                        Tel: 91-416-224 3118
                        Fax: 91-416-2243092
```

For the remaining tables, we follow the same strategy. We take these details and store them in a CSV file. We go through each row, take the first column data, store it in name, the next in designation, and the last column has an anchor tag, so we collect its data and then use a REGEX validator to verify the authenticity of the email address, if valid we store it in the file.

The below image just shows the first 5 contact information retrieved from the web page.



	Name	Designation	Email
0	Dr. G. Viswanathan	Chancellor	chancellor@vit.ac.in
1	Mr. Sankar Viswanathan	Vice President (Chennai Campus)	sankar@vit.ac.in
2	Dr. Sekar Viswanathan	Vice President (AP Campus)	sekar.office@vit.ac.in
3	Mr. G.V.Selvam	Vice President (Vellore Campus)	gvselvam.vp@vit.ac.in
4	Dr. Sandhya Pentareddy	Executive Director	sandhya.office@vit.ac.in

We will now go through the code :

Web Mining Lab - 4 Selenium

Write a Python program to read the given website and extract the phone numbers and emails and contact addresses.

In [1]:

```
from selenium import webdriver
import re
import pandas as pd
import copy
```

In [2]:

```
# This opens the safari automated window.
# We need to allow automation in safari settings before running this.

driver = webdriver.Safari()
```

In [3]:

```
# This fetches the web page in the new window

driver.get("https://vit.ac.in")
```

In [4]:

```
# Open the Contact Us page

driver.find_element_by_xpath("//a[@title='Contact Us']").click()
```

In [79]:

```
# In the contact us page, all the contact details are in tables with class nam
e as "table al_left table-bordered table-striped custom-style"

# Get all the tables with that class name
tables = driver.find_elements_by_css_selector('.table.al_left.table-bordered.t
able-striped.custom-style')

tables
```

Out[79]:

```
[<selenium.webdriver.remote.webelement.WebElement (session="1BB043
90-3F06-4125-8B84-1B2692FE857E", element="node-E98624DE-BCE7-4EA2-
8B28-FA87524A889F")>,
 <selenium.webdriver.remote.webelement.WebElement (session="1BB043
90-3F06-4125-8B84-1B2692FE857E", element="node-6B5F557B-F6BF-45DB-
9AAE-A9448F2A78EE")>,
 <selenium.webdriver.remote.webelement.WebElement (session="1BB043
90-3F06-4125-8B84-1B2692FE857E", element="node-5D34AA5F-AF7F-48A4-
9164-4C722578E5E6")>]
```

In [99]:

```
# The first table is different

item = tables[0]
# Get the rows
tds = item.find_elements_by_xpath("./tbody/tr/td")

# Store them in an array of strings
contact_temp_arr = []

for td in tds :
    # Get all the emails from the anchor tags there
    cur_emails = []
    try :
        anchors = td.find_elements_by_xpath("./a")
        for anchor in anchors :
            cur_emails.append(anchor.text)
    except :
        pass

    # Get the text except the ones inside the anchor tags
    OWN_TEXT_SCRIPT = "if (arguments[0].hasChildNodes()) { \
                            var res = ''; \
                            var children = arguments[0].childNodes; \
                            for (var n = 0; n < children.length; n++) { \
                                if (children[n].nodeType == Node.TEXT_NODE) {
\
                                    res += ' ' + children[n].nodeValue; \
                                } \
                            } \
                            return res.trim() \
                        } \
                        else { \
                            return arguments[0].innerText \
                        }"
    # Some td's have p-tags and font-tags, so we go cross it and then use the js above
    it = td
    temp = None
    try :
        temp = it.find_element_by_xpath("./p")
    except :
        pass
    if temp is not None :
        it = temp
    try :
        temp = it.find_element_by_xpath("./font")
    except :
        pass
    if temp is not None :
        it = temp

    # Execute the above js script
    text = driver.execute_script(OWN_TEXT_SCRIPT, it)
    if len(cur_emails) > 0 :
        text += "\n\t\t\tEmails: " + ",".join(cur_emails)


    print(text)
    contact_temp_arr.append(text)
```

VIT
Vellore Campus
Vellore - 632 014
Tamilnadu, India
Tel: 91-416-2243091 / 93
Fax: 91-416-2243092
91-416-2240411

VIT
Chennai Campus
Vandalur - Kelambakkam Road
Chennai - 600 127
Ph : 044 3993 1555
Fax : 044 3993 2555
Emails: admin.chennai@vit.ac.in

Admissions Office
Dr.G.Kalaichelvan
Director - UG Admissions
Vellore Institute of Technology
Vellore - 632 014,
Tamil Nadu, India.
Phone: + 91-416-220 2020
Fax: +91-416-224 5544, 224 0411
Email:
Emails: ugadmission@vit.ac.in

Admissions Office
Dr. Ramasubramanian V
Director - PG Admissions
Vellore Institute of Technology
Vellore - 632 014,
Tamil Nadu, India.
Phone: + 91-416-220 2188
Fax: +91-416-224 5544, 224 9955
Email:
Emails: pgadmission@vit.ac.in

Dr.V.Samuel Rajkumar
Director - Placement & Training
Vellore Institute of Technology
Vellore - 632 014.
Tamil Nadu.
Tel: 0416 - 2202846
Fax: 91-416-2243092, 91-416-224 0411

International Relations Office
Dr. C. Vijayakumar
Director - International Relations
Vellore Institute of Technology
Vellore - 632 014.
Tamil Nadu, India
Tel: 91-416-224 3118
Fax: 91-416-2243092

In [100]:

```
# As these details are in organizational level, we save this seperately in a f
ile

with open('Institutuion_Contact.txt', 'w') as file :
    for row in contact_temp_arr :
        file.write(row)
        file.write("\n\n")
```

In [49]:

```
# Create arrays to store the names, designations, and email address :

names = []
designations = []
emails = []
```

In [50]:

```
# For the remaining tables, the format is similar, so we will store them in a
CSV file in the end

for i in range(1, len(tables)) :
    table = tables[i]
    trs = table.find_elements_by_xpath("./tbody/tr")

    for i in range(1, len(trs)) :
        tds = trs[i].find_elements_by_xpath("./td")

        # First column is designation
        designations.append(tds[0].text)

        # Second column is Name
        names.append(tds[1].text)

        # Third is email which is inside an anchor tag
        try :
            cur_email = tds[2].find_element_by_xpath("./a").text
        except :
            cur_email = '-'
        regex_pattern = '^[a-z0-9]+[\._]?[a-z0-9]+[@]\w+[.]\w'
        if (cur_email != '-') and (not re.search(regex_pattern, cur_email)) :
            print("Email Pattern does not match", cur_email)
            break
        emails.append(cur_email)
```

In [57]:

```
# Convert these into a CSV file

officials_df = pd.DataFrame(list(zip(names, designations, emails)), columns=['Name', 'Designation', 'Email'])
officials_df.head()
```

Out[57]:

```
                        Name                        Designation                     Email
0         Dr. G. Viswanathan                         Chancellor     chancellor@vit.ac.in
1   Mr. Sankar Viswanathan   Vice President (Chennai Campus)          sankar@vit.ac.in
2     Dr. Sekar Viswanathan        Vice President (AP Campus)     sekar.office@vit.ac.in
3              Mr. G.V.Selvam   Vice President (Vellore Campus)   gvselvam.vp@vit.ac.in
4   Dr. Sandhya Pentareddy                Executive Director  sandhya.office@vit.ac.in
```

In [59]:

```
# Save this CSV file

officials_df.to_csv('Officials_Details.csv', index=None)
```

In [101]:

```
# This Closes the connection and closes the window

driver.close()
```

In [ ]: