15Z332 Ex2 - WebScrapping

October 17, 2018

1 Exercise 2

1.1 Perform web scrapping on PSG IM website to display names of faculty with Ph.D

1.1.1 Step 1: Get web page response

Using requests package in python, get the response for the URL of PSG IM Faculty web page

```
In [5]: import requests
        r  = requests.get("http://psgim.ac.in/2017/01/full-time-faculty/")
        print(r)
```

```
<Response [200]>
```

1.1.2 Step 2: Retrieve all faculty names

The names of the faculty are wrapped within a 'div' tag with class 'vc_column-inner' and has heading 4 as text formatting. We retrieve such data by inspecting the webpage in the web browser. To pull such data from a HTML file, we use 'BeautifulSoup' package

```
In [10]: #Import BeautifulSoup for extracting data from a HTML file
         from bs4 import BeautifulSoup

         #Extract HTML content from the HTML response
         soup = BeautifulSoup(r.content, 'html.parser')

         #Filter div tags with class 'vc_column_inner'
         faculty = soup.findAll('div',{"class":"vc_column-inner"})

         #Retrieve all faculty names (written in 'h4') and store in a list
         facultyName = []
         for i in range(len(faculty)):
             if faculty[i].find('h4'):
                 facultyName.append(faculty[i].find('h4').text)

         #This person does not have any qualification or designation, hence it was removed
         facultyName.remove('Ms Muthu Janaki')
         print('Number of faculties',len(facultyName))
         for i in range(len(facultyName)):
             print(str(i+1)+" "+facultyName[i])
```

```
Number of faculties 31
1 Dr Archana Krishnan
2 DR ARUL RAJAN K
3 DR Balasudarsun N L
4 DR DEEPA R
5 Mr Harish V
6 DR JAGAJEEVAN R
7 DR JOSHUA SELVAKUMAR J
8 Mr Karthikeyan M S
9 Mr KARTHIKEYAN L
10 DR KAVITHA D
11 Dr KRISHNAVENI R
12 DR KRISHNAVENI MUTHIAH R
13 Dr MANJU P GEORGE
14 MR Manikandan S S
15 DR MANSURALI A
16 Ms Rajeswari R
17 DR RAMAN H
18 Dr RAMKUMAR N
19 DR SATHISH M
20 DR SATHYANARAYANAN R S
21 DR SEKKIZHAR J
22  Dr Srigayathri Devi K
23 Dr SRIVIDYA V
24 DR SUJATHA R
25 DR SUDHARANI RAVINDRAN D
26 DR SWAMYNATHAN R
27 DR THILAGAM V
28 DR UMA MAHESWARI B
29 DR UMESH CHANDRASEKHAR
30 Dr VIJAYA T G
31 DR VIVEK N
```

1.1.3 Step 3: Filter names of faculty with Ph.D

Scrape through every div tag with text 'Qualification', and identify those with 'Ph.D'. If 'Ph.D' is present in the faculty's qualification, print the faculty name.

```
In [16]: #index variable to keep track of faculty names
         index = -1
         count = 0
```

```
#Finds all 'p' tag that has the class 'wpb_text_column'
faculty = soup.findAll('div',{"class":"wpb_text_column"})

print("Faculty with Ph.D: ")
for i in range(len(faculty)):
    if faculty[i].find('p'):
        #if the text of 'p' tag has 'Qualification'
        if faculty[i].find('p').text.find('Qualification')!=-1:
            index=index+1
            #if the qualification has 'Ph.D' or 'PhD'
            if faculty[i].find('p').text.find('Ph.D')!=-1 or faculty[i].find('p').tex
                #print the faculty name from the list created above
                count=count+1
                print(count,facultyName[index])
```

```
Faculty with Ph.D:
1 Dr Archana Krishnan
2 DR ARUL RAJAN K
3 DR Balasudarsun N L
4 DR DEEPA R
5 Mr Harish V
6 DR JAGAJEEVAN R
7 DR JOSHUA SELVAKUMAR J
8 DR KAVITHA D
9 Dr KRISHNAVENI R
10 DR KRISHNAVENI MUTHIAH R
11 Dr MANJU P GEORGE
12 DR MANSURALI A
13 DR RAMAN H
14 Dr RAMKUMAR N
15 DR SATHISH M
16 DR SATHYANARAYANAN R S
17 DR SEKKIZHAR J
18  Dr Srigayathri Devi K
19 Dr SRIVIDYA V
20 DR SUJATHA R
21 DR SUDHARANI RAVINDRAN D
22 DR SWAMYNATHAN R
23 DR THILAGAM V
24 DR UMA MAHESWARI B
25 DR UMESH CHANDRASEKHAR
26 Dr VIJAYA T G
27 DR VIVEK N
Total number of faculties with Ph.D =  27
```

```
In [ ]:
```