Data Scrapping

In [1]:

```
import pandas as pd
from bs4 import BeautifulSoup
import requests
from csv import writer
import requests
import re
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
browser = webdriver.Chrome()
browser.get('https://gim.ac.in/people/faculty')

url= "https://gim.ac.in/people/faculty"
page = requests.get(url)
print(page)
soup = BeautifulSoup(page.content, 'html.parser') # to get the souce code of that page

# lets tell BS to fetch all the links

links = []
for a in soup.find_all('a'):
    links.append(a.get("href"))

#print(links)
#for l in links:
    #if ('https://gim.ac.in/faculty/') in l:
        #print(list(l))


#lists = soup.find_all('a',class_="href") #find all the class section with this class_r

```

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

In [ ]:

```
#links
```

In [2]:

```
# only getting the faculty links
link_list = []
for l in links:
    if ('https://gim.ac.in/faculty/') in l:
        link_list.append(l)
```

In [3]:

```
name_col=[]
email_col=[]
linkedin_col=[]
phone_col=[]
facebook_col=[]
instagram_col=[]

for ele in link_list:
    #print(ele)
    browser.get(ele)
    soup1 = BeautifulSoup(browser.page_source)
    #title = list.find('a',class_="listing-search-item__link listing-search-item__link-

    #Name
    try:
        name = soup1.find('h1',class_="node__title").text.replace('\n','')
        name_col.append(name)
    except:
        name_col.append("NA")

    #email
    try:
        email = soup1.find('div',class_="field__item").text.replace('\n','')
        email_col.append(email)
    except:
        email_col.append("NA")

    #phone
    try:
        phone = soup1.find('div',class_="field field-node--field-email field-formatter-
        phone_col.append(phone)
    except:
        phone_col.append("NA")


    #LinkedIn
    try:
        linkedin = soup1.find('div',class_="link linkedin")
        linkedin_col.append(linkedin.text.replace('\n',''))
    except:
        linkedin_col.append("NA")

    #facebook
    try:
        facebook = soup1.find('div',class_="link facebook")
        #linkedin = linkedin.find('a',target_="_blank")
        facebook_col.append(facebook.text.replace('\n',''))
    except:
        facebook_col.append("NA")

    #instagram
    try:
        instagram = soup1.find('div',class_="link instagram")
        #linkedin = linkedin.find('a',target_="_blank")
        instagram_col.append(instagram.text.replace('\n',''))
    except:
        instagram_col.append("NA")


```

```
df = pd.DataFrame({'Name':name_col,'phone':phone_col, 'email':email_col,'linkedin_link'
df
```

Out[3]:

	Name	phone	email	linkedin_link	
0	Ajit Parulekar	ajitp@gim.ac.in	Director	https://www.linkedin.com/in/ajit-parulekar-600...	
1	Abhishek Ranga	abhishek@gim.ac.in	Associate Professor	https://www.linkedin.com/in/dr-abhishek-ranga-...	https://v
2	Ajay Vamadevan	ajay.vamadevan@gim.ac.in	Professor	NA	
3	Akshay Bhat	akshay@gim.ac.in	Assistant Professor	NA	
4	Alekh Gour	alekh@gim.ac.in	Associate Professor	https://www.linkedin.com/in/dr-alekh-gour-2721...	ht
...	...	...	...	...	
69	Vilasini Devi Nair	devinair@gim.ac.in	Assistant Professor	NA	
70	Vinit Ghosh	vinit@gim.ac.in	Assistant Professor	https://in.linkedin.com/in/vinitghosh	
71	Vishwesh Singbal	singbal@gim.ac.in	Assistant Professor	https://www.linkedin.com/in/vishwesh-s-b2447922/	
72	Vithal S. Sukhathankar	visukh@gim.ac.in	Associate Professor	NA	
73	Yukti Sharma	yukti@gim.ac.in	Assistant Professor	NA	

74 rows × 6 columns

In [ ]: