Task : 7 Author:- Sagar Bhoi

In [1]:
```
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd

urls = ["http://www.studyguideindia.com/Colleges/Engineering/default.asp?State=DL",
        "http://www.studyguideindia.com/Colleges/Engineering/default.asp?State=MH&ct=159",
        "http://www.studyguideindia.com/Colleges/Engineering/default.asp?State=WB&ct=221",
        "http://www.studyguideindia.com/Colleges/Engineering/default.asp?State=TN&ct=1"]

state_city_url= "http://www.studyguideindia.com/Courses/Engineering-Courses.asp"



def metrocity_all_clgs():
    all_clgs = []
    for url in urls:
        content = requests.get(url)
        html_content = content.content
        soup = bs(html_content, "html.parser")
        clglists = soup.find_all("table", {"class": "clg-listing"})
        anchors = clglists[0].find_all('a')
        for i in range(len(anchors)):
            all_clgs.append(anchors[i]['href'])
    content.close()
    return all_clgs


clg_name = []
clg_address = []
clg_url = []
clg_email = []
clg_phone = []


def appending(info_list):
    if "College Name" in info_list:
        index = info_list.index('College Name')
        clg_name.append(info_list[index + 1])
    else:
        clg_name.append("NULL")
    if "Address" in info_list:
```

```
        index = info_list.index("Address")
        clg_address.append(info_list[index + 1])
    else:
        clg_address.append("NULL")
    if "Website" in info_list:
        index = info_list.index("Website")
        clg_url.append(info_list[index + 1])
    else:
        clg_url.append("NULL")
    if "E-Mail" in info_list:
        index = info_list.index("E-Mail")
        clg_email.append(info_list[index + 1])
    else:
        clg_email.append("NULL")
    if "Phone" in info_list:
        index = info_list.index("Phone")
        clg_phone.append(info_list[index + 1])
    else:
        clg_phone.append("NULL")


def scrap_data(all_clgs):
    for link in all_clgs:
        r = requests.get(link)
        html = r.content
        soup = bs(html, 'html.parser')
        clg_data = soup.find_all("table", {"class": "altcolor1"})
        if len(clg_data) == 0:
            continue
        clg_info = clg_data[0].find_all("td")
        info_list = []
        for i in range(len(clg_info)):
            info_list.append(clg_info[i].text.strip())
        appending(info_list)
    r.close()
    dataframe= pd.DataFrame({"College Name":clg_name,"Address":clg_address,"Url Address":clg_url,"E-Mail":clg_email,"Phone":clg_phone})
    return dataframe

def save_data(dataframe,filename):
    dataframe.to_csv(filename,index=False)

number= int(input("Enter 1 for colleges in Metrocity \nEnter 2 for colleges  Cities \nEnter 3 for state colleges  :"))

if number== 1:
    all_clgs = metrocity_all_clgs()
    data= scrap_data(all_clgs)
```

```
    save_data(data,filename="metrocity_collage_data.csv")

elif number == 2:
    r = requests.get(state_city_url)
    http = r.content
    soup = bs(http, "html.parser")
    box = soup.find_all("div", {"class": "tab_inner_full"})
    required_div_of_cities = box[2]
    anchores = required_div_of_cities.find_all("a")
    list_of_link_of_city = []
    for a in anchores:
        list_of_link_of_city.append(a['href'])
    all_clg = []
    for url in list_of_link_of_city:
        content = requests.get(url)
        html_content = content.content

        soup = bs(html_content, "html.parser")
        clglists = soup.find_all("table", {"class": "clg-listing"})

        anchors = clglists[0].find_all('a')

        for i in range(len(anchors)):
            all_clg.append(anchors[i]['href'])
    content.close()
    data = scrap_data(all_clg)
    save_data(data,filename='city_college_list.csv')

elif number == 3:
    r = requests.get(state_city_url)
    http = r.content
    soup = bs(http, "html.parser")
    box = soup.find_all("div", {"class": "tab_inner_full_2col"})
    required_div_of_state = box[2]
    anchores = required_div_of_state.find_all("a")
    list_of_link_of_state = []
    for a in anchores:
        list_of_link_of_state.append(a['href'])
    all_clg = []
    for url in list_of_link_of_state:
        content = requests.get(url)
        html_content = content.content

        soup = bs(html_content, "html.parser")
        clglists = soup.find_all("table", {"class": "clg-listing"})
```

```
        anchors = clglists[0].find_all('a')

        for i in range(len(anchors)):
            all_clg.append(anchors[i]['href'])
    content.close()
    data = scrap_data(all_clg)
    save_data(data,filename="state_college_list.csv")

else:
    print("enter valid input , given input is wrong")
```

```
Enter 1 for colleges in Metrocity
Enter 2 for colleges  Cities
Enter 3 for state colleges  :1
```

In [ ]: