

Name: Yash Srivastava

PGID: 12010060

Email: Yashy_Srivastava_AMPBA2021S@isb.edu

Subject: **Data Collection**

Q1

Scraping Jobs from Naukri

Goal of the assignment: Identify Key Tech Hires in the Hyderabad area

The screenshot shows the Naukri.com website interface. At the top, there's a navigation bar with the Naukri logo and various links. Below that, a search bar shows the current search for 'hyderabad secunderabad'. The main content area displays a list of job listings. On the left, there are filters for Location, Freshness, Salary, and Experience. The job listings include details like the company name (Virtusa Consulting Services Pvt Ltd), location (Chennai, Pune, Mumbai), and job title. The right sidebar features a section for 'Get Personalised Job Recommendations' and a grid of featured companies including NIIT, Prodapt, IQVIA, kony, Inexwave, HCL, ONPASSIVE, inspirisys, Sierra-Cedar, and Wells Fargo.

We scrap the following from this website.

- Name of the company
- Number of Positions
- Segmentation
- Skills they are looking for

```

In [8]: ##Import all the necessary Libraries. Selenium, BeautifulSoup, Pandas, csv, String etc
import requests
from bs4 import BeautifulSoup
from selenium import webdriver
import urllib
import csv
import os
import pandas as pd

#Create an empty CSV file with column names
with open('Hyderabad_Yash_Srivastava.csv', 'a', encoding='utf-8', newline='') as f_out:
    csv_print = csv.writer(f_out)
    fileIsEmpty = os.stat('Hyderabad_Yash_Srivastava.csv').st_size == 0
    if fileIsEmpty:
        csv_print.writerow(['Job-Title', 'Company Name', 'Segmentation', 'Skills they are looking for', 'Date Posted'])

    max_page_num = 10 #no of pages that we want to scrap
    CompanyName = [] #empty List
    JobName = [] #empty List
    Skills = [] #empty List
    Dates = [] #empty List
    segmentation=[] #empty List
    Segmentation='Information Technology'

#For Loop starts for achieving pagination
for i in range(1, max_page_num + 1):

    #If Condition for first page
    if i == 1:
        url = "https://www.naukri.com/information-technology-jobs?cityType=25.1.4&industryTypeId=25"
    else:
        url = "https://www.naukri.com/information-technology-jobs-" + str(i) + "?cityType=25.1.4&industryTypeId=
25"

#Chrome driver to use Selenium
driver = webdriver.Chrome(executable_path = r'C:\Users\ysrivastava\Downloads\chromedriver_win32\chromedriver.
exe')

#Fetching the URL of Naukri through selenium driver
driver.get(url)
#Contains page source
html = driver.page_source

#Creating beautiful soup object
soup = BeautifulSoup(html, "html.parser")

#Finding particular class and searching its subclasses to find our respective elements
for jobs in soup.find_all(class_='jobTuple'):

    #Adding try and except blocks to ensure code doesnt fail
    try:
        company = jobs.find('a', class_='subTitle').text.strip() #Stripping out Company Name
    except Exception as e:
        company = "Not Found"
    CompanyName.append(company)
    print(company)

    #Adding try and except blocks to ensure code doesnt fail
    try:
        title = jobs.find('a', class_='title').text.strip() #Stripping out Job Title
    except Exception as e:
        title = "Not Found"
    JobName.append(title)
    print(title)

    #Adding try and except blocks to ensure code doesnt fail
    try:
        skill = jobs.find('ul', class_='tags').text.split(sep=" ") #Stripping out Skills
    except Exception as e:
        skill = "Not Found"
    Skills.append(skill)
    print(skill)

    #Adding try and except blocks to ensure code doesnt fail
    try:
        dates = jobs.find('div', class_='type br2 flleft grey').text.strip() #Stripping out duration of the jo
b posted
    except Exception as e:

```

```
        dates = "Not Found"
    Dates.append(dates)
    print(dates)

    print(segmentation)
    print('=====')
    segmentation.append('Information Technology')

    csv_print.writerow([title,company,Segmentation,skill,dates])
    #Writting the content to CSV file.
```

```
In [35]: #Creating a Data Frame to show output
df = pd.DataFrame(list(zip(CompanyName, JobName, Skills,segmentation,Dates)),
                   columns=['Company', 'Title', 'Skills', 'Category', 'Date Posted'])
```

```
In [36]: df
```

Out[36]:

	Company	Title	Skills	Category	Date Posted
0	Nisarga Information Technology Solutions Pvt Ltd	Business Development Manager	[Communication, SkillsSalesBusiness, Developme...	Information Technology	19 Days Ago
1	TERRA TECHNOLOGY	Information Technology	[UnixService, managementglobal, operationsLinu...	Information Technology	30+ Days Ago
2	TERRA TECHNOLOGY	Information Technology	[UnixService, managementglobal, operationsData...	Information Technology	30+ Days Ago
3	Stellar Information Technology Private Limited	Relationship Manager IT Product Sales Exp. ...	[CRM, SalesIt, Product, SalesB2B, SalesHardwar...	Information Technology	25 Days Ago
4	International Institute of Information Technology	Technology Management- Lecturer	[CounselorMentorTrainerAdvisorEducatorTeaching]	Information Technology	30+ Days Ago
...
195	CDK GLOBAL, LLC.	Consultant - .Net Developer	[GITRDBMSJavascriptsoftware, qualityDigital, m...	Information Technology	10 Days Ago
196	CDK GLOBAL, LLC.	Senior UI Developer (React.Js/Node.Js)	[IT, SkillsHTMLCSSJavascriptJenkinsAzureWeb, s...	Information Technology	10 Days Ago
197	CDK GLOBAL, LLC.	Senior Java Developer	[IT, SkillsJavaJavascriptTrainingExcelDigital,...	Information Technology	10 Days Ago
198	Han Digital Solution (P)	AVP , Threat Detection	[UnixSocial, mediaNetwork, analysisForensicWin...	Information Technology	30+ Days Ago
199	Infronics Systems Ltd	Support Engineer - Linux	[AdministrationLinuxNetworkingRHELAdministrati...	Information Technology	30+ Days Ago

200 rows × 5 columns

```
In [37]: #Counting the number of jobs posted by a company  
df.groupby(['Company']).count()
```

Out[37]:

	Title	Skills	Category	Date Posted
Company				
AGI INDUSTRIES PRIVATE LIMITED	2	2	2	2
ARY Technologies Private Limited	1	1	1	1
Accenture Solutions Pvt Ltd	2	2	2	2
Acel Solutions Pvt. Ltd.	1	1	1	1
Anantha Cyber Tech Pvt Limited	12	12	12	12
BM Global Solutions	1	1	1	1
Biarca	1	1	1	1
Big Idea Technology LLC	1	1	1	1
CDK GLOBAL (INDIA) PRIVATE LIMITED	9	9	9	9
CDK GLOBAL, LLC.	13	13	13	13
CGI Information Systems and Management Consultants	1	1	1	1
CRAW Security	1	1	1	1
CUETHREE Technologies	2	2	2	2
Cee Chat LLC	1	1	1	1
Comint Systems and Solutions Pvt Ltd	1	1	1	1
Creative Hands HR Consultancy	1	1	1	1
Dhruvsoft Services Pvt Ltd	1	1	1	1
Diverse Lynx India Pvt. Ltd	1	1	1	1
Edification Technologies	1	1	1	1
FactSet Systems India Pvt Ltd	1	1	1	1
Future Focus Infotech Pvt. Ltd.	1	1	1	1
Han Digital Solution (P)	1	1	1	1
Infronics Systems Ltd	1	1	1	1
Integra Micro Software Services	1	1	1	1
International Institute of Information Technology	77	77	77	77
Kensium Solutions Private Limited	1	1	1	1
Lincoln Tech	1	1	1	1
LogonB2B Soft Solutions Pvt. Ltd.	1	1	1	1
Nisarga Information Technology Solutions Pvt Ltd	1	1	1	1
Omics Group	5	5	5	5
OpsMx	1	1	1	1
OpsRamp	3	3	3	3
Orcapod Consulting Services Private Limited	1	1	1	1
Overseas Information Technology	6	6	6	6
Pamten Software Solutions Pvt. Ltd.	1	1	1	1
Pegasystems Worldwide India Pvt. Ltd.	1	1	1	1
Prolifics	2	2	2	2
Ram IT Solutions	2	2	2	2
S2TECH.COM INDIA PVT LTD	1	1	1	1
Sahasya Global Solutions	4	4	4	4
Secureworks	1	1	1	1
ServiceNow	1	1	1	1
Shuchita Technologies (P) Ltd.	1	1	1	1
Solugenix India Private Limited	7	7	7	7
Stellar Information Technology Private Limited	1	1	1	1
Superme enterprises	1	1	1	1
TERRA TECHNOLOGY	5	5	5	5
TTEC	5	5	5	5
Thinksure Information Technology Private Limited	1	1	1	1
Times World Information Technology	1	1	1	1

	Title	Skills	Category	Date Posted
Company				
True Troops Consultants Private Limited	2	2	2	2
Varutra Consulting Pvt. Ltd	1	1	1	1
Virtusa Consulting Services Pvt Ltd	2	2	2	2
Young India Information Technology Services (P) Lt	3	3	3	3
Zyme India	1	1	1	1
iSpatial Techno Solutions Pvt Ltd.	1	1	1	1

In []:

Name: Yash Srivastava

PGID: 12010060

Email ID: yash_srivastava_ampba2021s@isb.edu

Q2 : Extracting data from API

In [2]: `pip install newsapi-python`

```
Collecting newsapi-python
  Downloading newsapi_python-0.2.6-py2.py3-none-any.whl (7.9 kB)
Requirement already satisfied: requests<3.0.0 in c:\users\ysrivastava\anaconda3\lib\site-packages (from newsapi-python) (2.22.0)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in c:\users\ysrivastava\anaconda3\lib\site-packages (from requests<3.0.0->newsapi-python) (3.0.4)
Requirement already satisfied: idna<2.9,>=2.5 in c:\users\ysrivastava\anaconda3\lib\site-packages (from requests<3.0.0->newsapi-python) (2.8)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\ysrivastava\anaconda3\lib\site-packages (from requests<3.0.0->newsapi-python) (2019.11.28)
Requirement already satisfied: urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1 in c:\users\ysrivastava\anaconda3\lib\site-packages (from requests<3.0.0->newsapi-python) (1.25.8)
Installing collected packages: newsapi-python
Successfully installed newsapi-python-0.2.6
Note: you may need to restart the kernel to use updated packages.
```

In [6]: *#Importing News api Library*

```
import pandas as pd
import datetime as dt
from newsapi import NewsApiClient

#Through NewsApi we receive a key that we are embedding.
apikey= NewsApiClient(api_key='48882599569f4067a8ab0ba599adc373')

#We have used get_everything() and have passed certain parameters
get_data = apikey.get_everything(q='Coronavirus',language='en',page_size=100,from_param='2020-07-16')
get_data.keys()
get_data['articles'][0]
articles= get_data['articles']
#Articles has the data formatted for use. Let us test its output
articles[0]
```

Out[6]: `{'source': {'id': None, 'name': 'Lifehacker.com'},
 'author': 'Beth Skwarecki on Vitals, shared by Beth Skwarecki to Lifehacker',
 'title': 'How to Keep Track of All the Potential Coronavirus Treatments',
 'description': 'There's still no cure for the coronavirus, but dozens of drugs and treatments are being tested against it. And you're not alone if you've gotten confused about which ones are mere possibilities and which are widely understood to be useful. The science changes...',
 'url': 'https://vitals.lifehacker.com/how-to-keep-track-of-all-the-potential-coronavirus-treatments-1844419504',
 'urlToImage': 'https://i.kinja-img.com/gawker-media/image/upload/c_fill,f_auto,f1_progressive,g_center,h_675,pg_1,q_80,w_1200/wcjl3ipr32kmjr75go9.jpg',
 'publishedAt': '2020-07-17T17:15:00Z',
 'content': 'Theres still no cure for the coronavirus, but dozens of drugs and treatments are being tested against it. And youre not alone if youve gotten confused about which ones are mere possibilities and which... [+2342 chars]}'`

```
In [7]: #This will be in list form hence we convert it into a DataFrame
Newsdata = pd.DataFrame(articles,columns=['author','title','content','source','description',])
Newsdata
```

Out[7]:

	author	title	content	source	description
0	Beth Skwarecki on Vitals, shared by Beth Skwar...	How to Keep Track of All the Potential Coronav...	Theres still no cure for the coronavirus, but ...	{'id': None, 'name': 'Lifehacker.com'}	There's still no cure for the coronavirus, but...
1	Kate Dore on Two Cents, shared by Kate Dore to...	Don't Skip These Estate Planning Moves During ...	As the coronavirus pandemic sweeps the nation,...	{'id': None, 'name': 'Lifehacker.com'}	As the coronavirus pandemic sweeps the nation,...
2	Mariella Moon	Google will ban coronavirus conspiracy ads to ...	Google is amping up its fight against coronavi...	{'id': 'engadget', 'name': 'Engadget'}	Google is amping up its fight against coronavi...
3	Karissa Bell	Facebook has removed 7 million posts for coron...	The company removes posts that spread false cl...	{'id': 'engadget', 'name': 'Engadget'}	If it seems like there's a lot of misinformati...
4	https://www.facebook.com/bbcnews	Coronavirus: Trump defends hydroxychloroquine ...	Image copyrightEPA\r\nUS President Donald Trum...	{'id': 'bbc-news', 'name': 'BBC News'}	The president said the drug is only rejected a...
...
95	https://www.facebook.com/bbcnews	Coronavirus: Los Angeles to shut off water and...	Image copyrightGetty ImagesImage caption\r\n A...	{'id': 'bbc-news', 'name': 'BBC News'}	Houses have become nightclubs despite coronavi...
96	Christina Maxouris, CNN	US returns to 1,000 coronavirus deaths in a day	(CNN)At least 1,000 American deaths linked to ...	{'id': 'cnn', 'name': 'CNN'}	At least 1,000 American deaths linked to coron...
97	None	Boris Johnson: Signs of a second coronavirus w...	Chat with us in Facebook Messenger. Find out w...	{'id': 'cnn', 'name': 'CNN'}	UK Prime Minister Boris Johnson has warned of ...
98	None	10-year-old creates video journal while fighti...	Chat with us in Facebook Messenger. Find out w...	{'id': 'cnn', 'name': 'CNN'}	Kendall Hanson, a 10-year-old girl in Orange C...
99	Naomi Thomas, CNN	Coronavirus stresses Americans more than other...	(CNN)The coronavirus pandemic has turned life ...	{'id': 'cnn', 'name': 'CNN'}	The coronavirus pandemic has turned life upsid...

100 rows × 5 columns

```
In [8]: #We can see the source column has a dict so Lets extract it from the DataFrame
Newsdata1 = Newsdata['source']
Newsdata1
```

```
Out[8]: 0    {'id': None, 'name': 'Lifehacker.com'}
1    {'id': None, 'name': 'Lifehacker.com'}
2    {'id': 'engadget', 'name': 'Engadget'}
3    {'id': 'engadget', 'name': 'Engadget'}
4    {'id': 'bbc-news', 'name': 'BBC News'}
...
95   {'id': 'bbc-news', 'name': 'BBC News'}
96   {'id': 'cnn', 'name': 'CNN'}
97   {'id': 'cnn', 'name': 'CNN'}
98   {'id': 'cnn', 'name': 'CNN'}
99   {'id': 'cnn', 'name': 'CNN'}
Name: source, Length: 100, dtype: object
```



```
In [9]: #Let us split the ID and Name column and add it to our existing Data Frame.
News_sources = pd.DataFrame(Newsdata1.tolist())
Newsdata['source_id']=News_sources['id']
Newsdata['source_name']=News_sources['name']

#Creating a new DataFrame with formatted columns
Newsdata_df = pd.DataFrame(Newsdata,columns=['source_id','source_name','author','title','description','content'])
Newsdata_df
```

Out[9]:

	source_id	source_name	author	title	description	content
0	None	Lifehacker.com	Beth Skwarecki on Vitals, shared by Beth Skwar...	How to Keep Track of All the Potential Coronav...	There's still no cure for the coronavirus, but...	Theres still no cure for the coronavirus, but ...
1	None	Lifehacker.com	Kate Dore on Two Cents, shared by Kate Dore to...	Don't Skip These Estate Planning Moves During ...	As the coronavirus pandemic sweeps the nation,...	As the coronavirus pandemic sweeps the nation,...
2	engadget	Engadget	Mariella Moon	Google will ban coronavirus conspiracy ads to ...	Google is amping up its fight against coronavi...	Google is amping up its fight against coronavi...
3	engadget	Engadget	Karissa Bell	Facebook has removed 7 million posts for coron...	If it seems like there's a lot of misinformati...	The company removes posts that spread false cl...
4	bbc-news	BBC News	https://www.facebook.com/bbcnews	Coronavirus: Trump defends hydroxychloroquine ...	The president said the drug is only rejected a...	Image copyrightEPA\r\nUS President Donald Trum...
...
95	bbc-news	BBC News	https://www.facebook.com/bbcnews	Coronavirus: Los Angeles to shut off water and...	Houses have become nightclubs despite coronavi...	Image copyrightGetty ImagesImage caption\r\nA...
96	cnn	CNN	Christina Maxouris, CNN	US returns to 1,000 coronavirus deaths in a day	At least 1,000 American deaths linked to coron...	(CNN)At least 1,000 American deaths linked to ...
97	cnn	CNN	None	Boris Johnson: Signs of a second coronavirus w...	UK Prime Minister Boris Johnson has warned of ...	Chat with us in Facebook Messenger. Find out w...
98	cnn	CNN	None	10-year-old creates video journal while fighti...	Kendall Hanson, a 10-year-old girl in Orange C...	Chat with us in Facebook Messenger. Find out w...
99	cnn	CNN	Naomi Thomas, CNN	Coronavirus stresses Americans more than other...	The coronavirus pandemic has turned life upsid...	(CNN)The coronavirus pandemic has turned life ...

100 rows × 6 columns

```
In [10]: #Let us export the resulting dataset into a csv file.
Newsdata_df.to_csv("News_api_Yash_Srivastava.csv")
```

In []: