Group 3

Web Scraping for Online Job Search

MIE 1624

Nikunj Viramgama
Ganesh Vedula
Aakash Iyer
Maharshi Trivedi
Vaibhav Gupta
Mohamad Danish

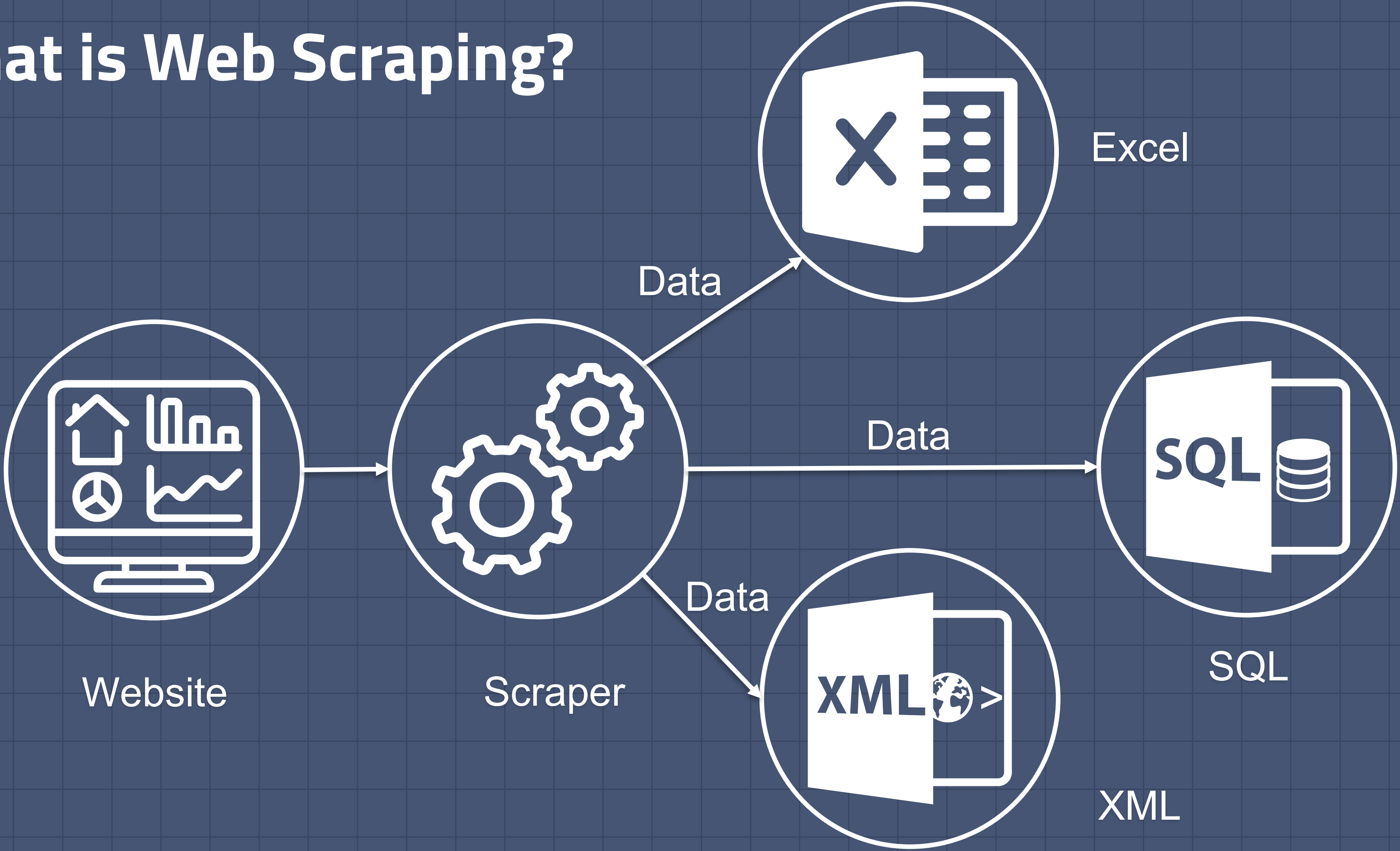
Outline on Web Scraping

Problem with Data Acquisition

44ZB [2]	90%	80% Time [1]	20% Time
Data in 2020	**Unstructured data**	**Data Preparation**	**Data Analysis**

(1) https://www.infoworld.com/article/3228245/data-science/the-80-20-data-science-dilemma.html
(2) Deloitte University Press

What is Web Scraping?

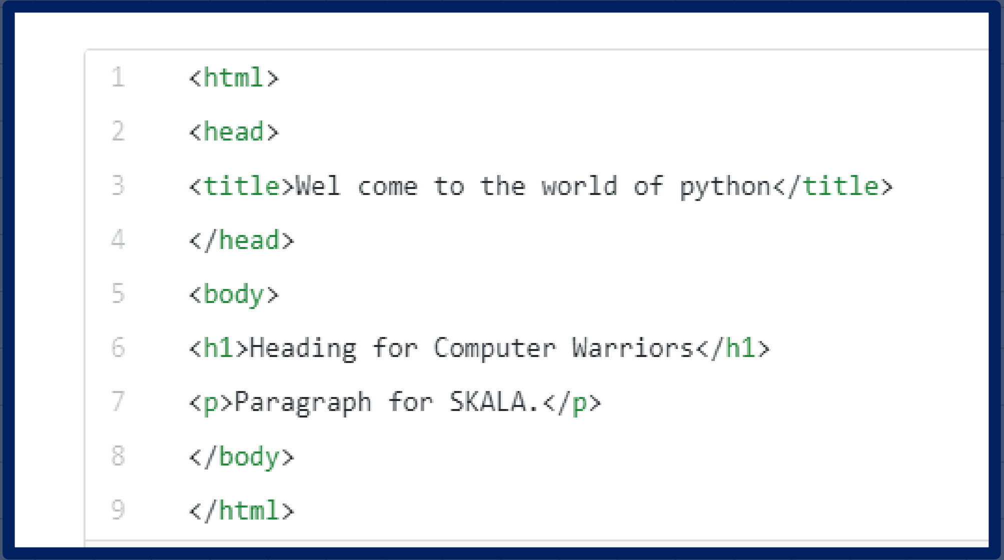


Web Scraping Libraries

Parse Web pages

Fetch data by directly visiting the website

Basic HTML and Steps



```html
<html>
<head>
<title>Wel come to the world of python</title>
</head>
<body>
<h1>Heading for Computer Warriors</h1>
<p>Paragraph for SKALA.</p>
</body>
</html>
```

page = requests.get(url)

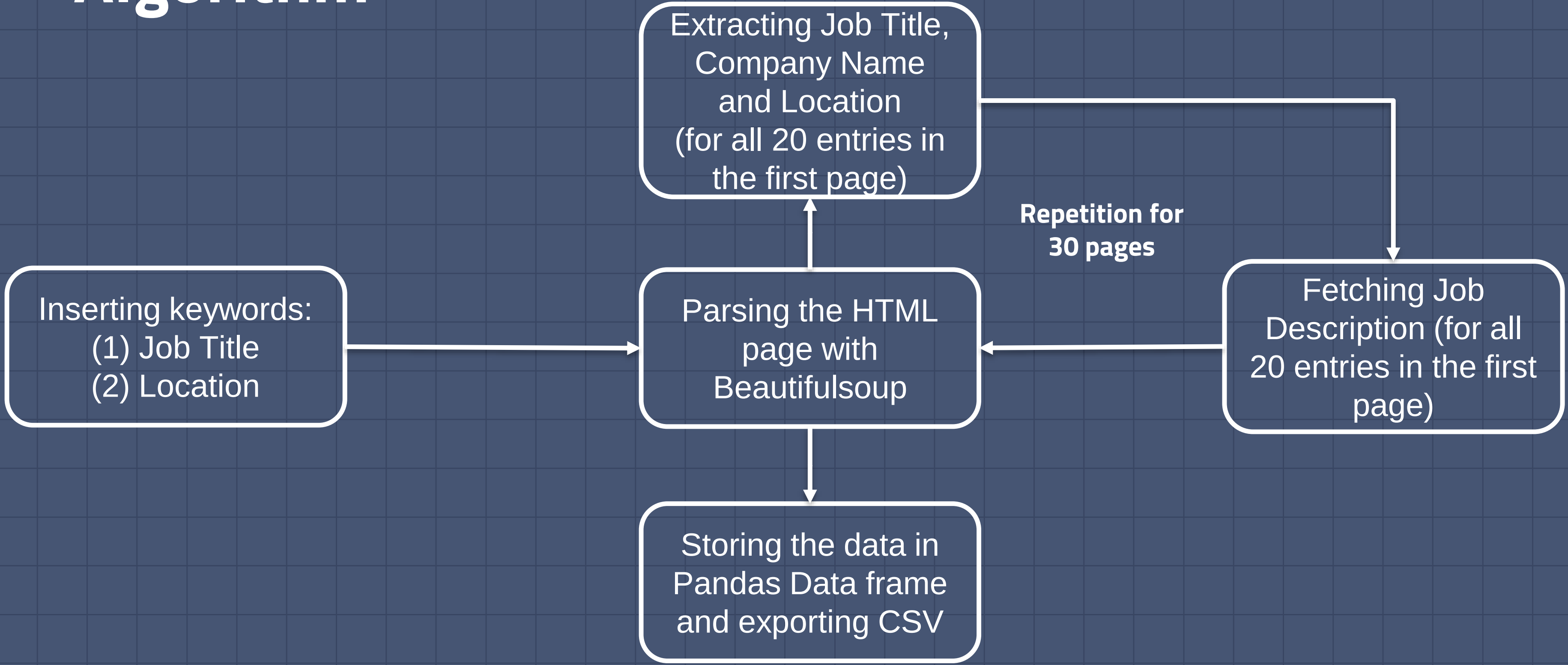
soup = BeautifulSoup(page.content, 'html.parser’)

data = soup.find_all(“tag",”h1"})

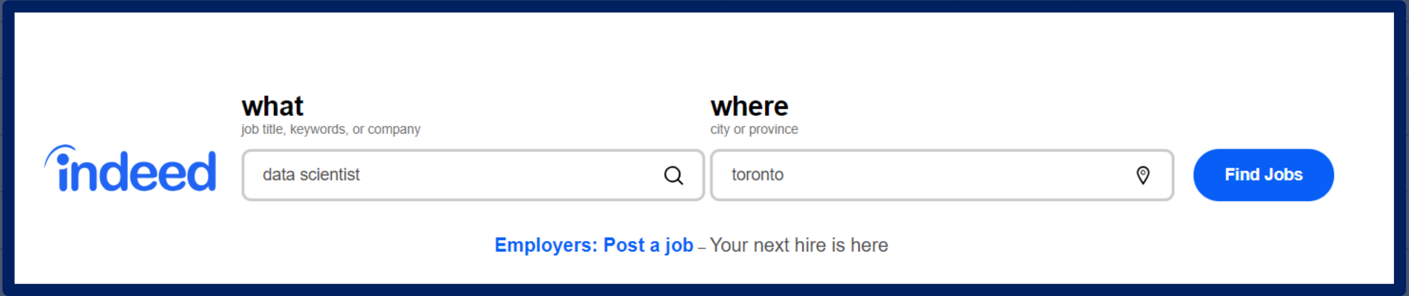
Image Source: GitHub

Implementation with Beautifulsoup

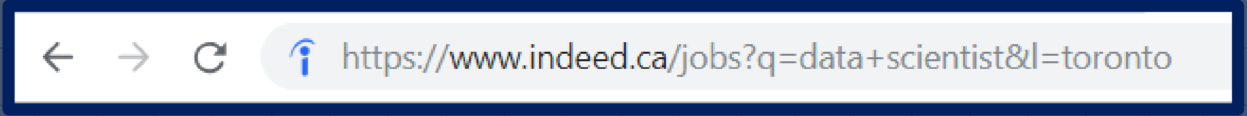
Algorithm



Extracting Job Title, Company Name and Location (for all 20 entries in the first page)
Repetition for 30 pages
Inserting keywords: (1) Job Title (2) Location
Parsing the HTML page with Beautifulsoup
Fetching Job Description (for all 20 entries in the first page)
Storing the data in Pandas Data frame and exporting CSV



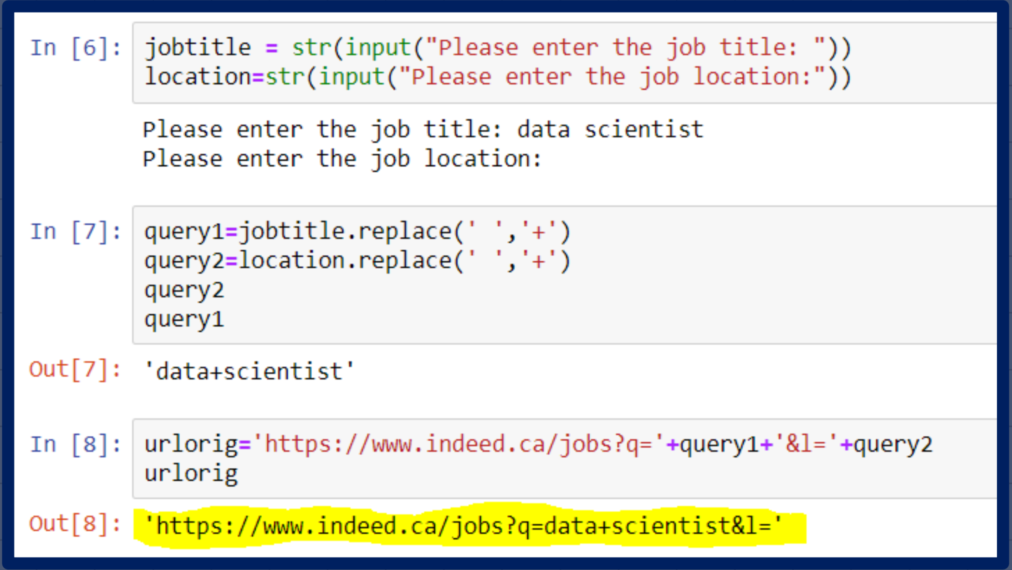
indeed
what
job title, keywords, or company
data scientist
where
city or province
toronto
Find Jobs
Employers: Post a job – Your next hire is here



```
https://www.indeed.ca/jobs?q=data+scientist&l=toronto
```

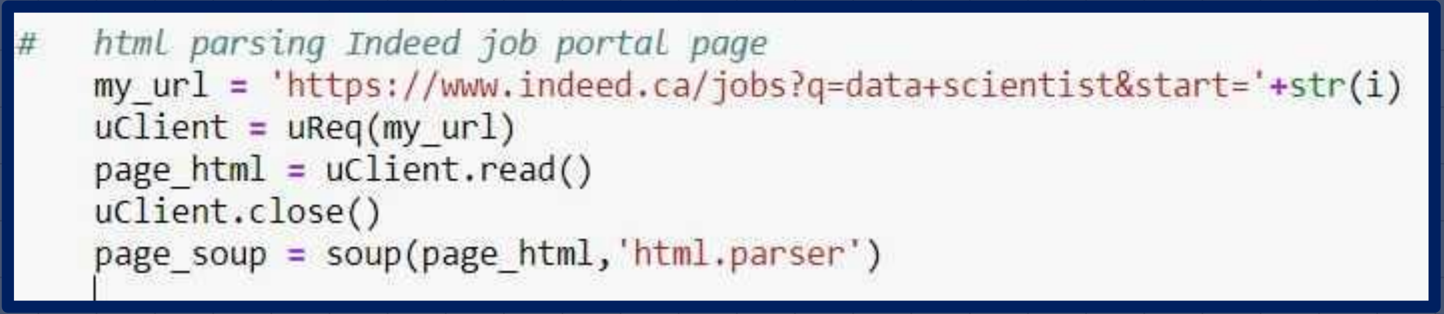
1st Step

Interacting with www.indeed.ca

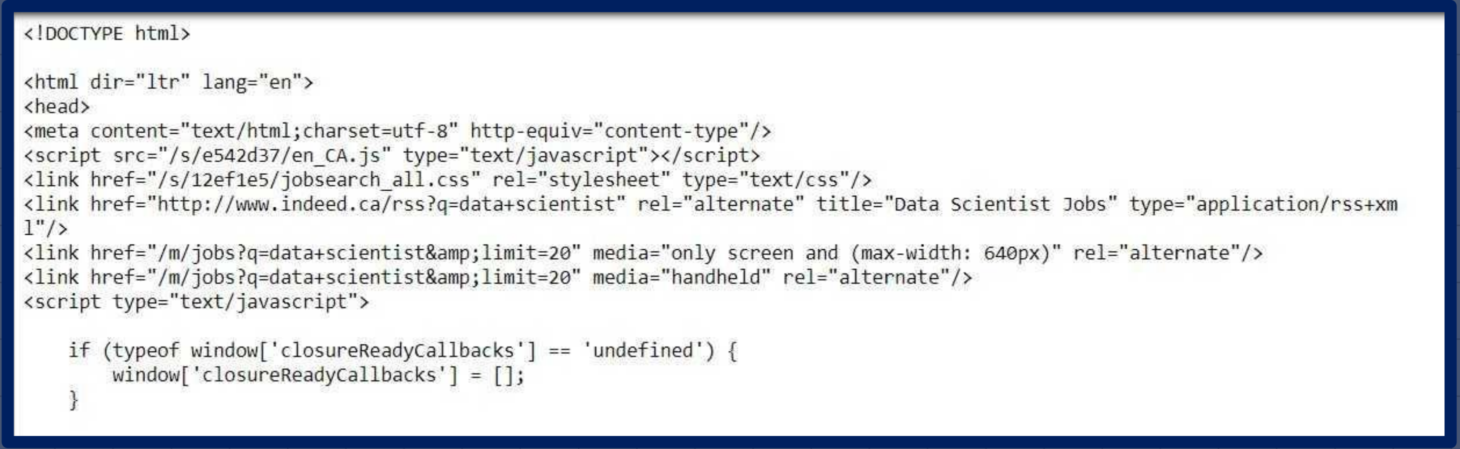


```
In [6]: jobtitle = str(input("Please enter the job title: "))
        location=str(input("Please enter the job location:"))

        Please enter the job title: data scientist
        Please enter the job location:

In [7]: query1=jobtitle.replace(' ','+')
        query2=location.replace(' ','+')
        query2
        query1

Out[7]: 'data+scientist'

In [8]: urlorig='https://www.indeed.ca/jobs?q='+query1+'&l='+query2
        urlorig

Out[8]: 'https://www.indeed.ca/jobs?q=data+scientist&l='
```

2nd Step
Web page Parsing



```
#   html parsing Indeed job portal page
    my_url = 'https://www.indeed.ca/jobs?q=data+scientist&start='+str(i)
    uClient = uReq(my_url)
    page_html = uClient.read()
    uClient.close()
    page_soup = soup(page_html,'html.parser')
```



```
<!DOCTYPE html>

<html dir="ltr" lang="en">
<head>
<meta content="text/html;charset=utf-8" http-equiv="content-type"/>
<script src="/s/e542d37/en_CA.js" type="text/javascript"></script>
<link href="/s/12ef1e5/jobsearch_all.css" rel="stylesheet" type="text/css"/>
<link href="http://www.indeed.ca/rss?q=data+scientist" rel="alternate" title="Data Scientist Jobs" type="application/rss+xm
l"/>
<link href="/m/jobs?q=data+scientist&amp;limit=20" media="only screen and (max-width: 640px)" rel="alternate"/>
<link href="/m/jobs?q=data+scientist&amp;limit=20" media="handheld" rel="alternate"/>
<script type="text/javascript">

    if (typeof window['closureReadyCallbacks'] == 'undefined') {
        window['closureReadyCallbacks'] = [];
    }
```

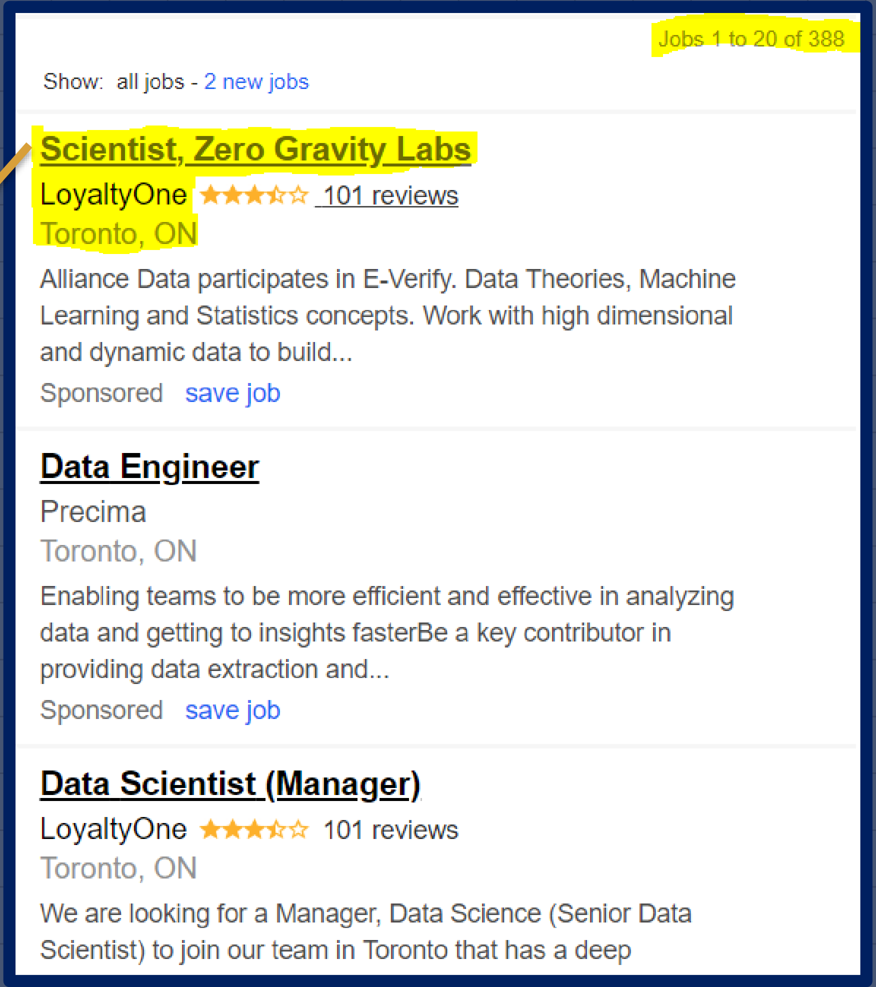
3rd Step

Extracting Job title and Company details

```
#   extracts job_title
    jobs = page_soup.findAll("div", class_="row" )
    for job in jobs:
        titles.append(job.a["title"])
```

```
#   extracts company_name
    companies = page_soup.findAll("span", class_="company" )
    for company in companies:
        company_names.append(company.text.strip())
```

```
#   extracts location
    locations = page_soup.findAll( class_="location")
    for location in locations:
        location_names.append(location.text)
```



Jobs 1 to 20 of 388

Show: all jobs - 2 new jobs

Scientist, Zero Gravity Labs

LoyaltyOne ★★★☆☆ 101 reviews

Toronto, ON

Alliance Data participates in E-Verify. Data Theories, Machine Learning and Statistics concepts. Work with high dimensional and dynamic data to build...

Sponsored save job

Data Engineer

Precima

Toronto, ON

Enabling teams to be more efficient and effective in analyzing data and getting to insights fasterBe a key contributor in providing data extraction and...

Sponsored save job

Data Scientist (Manager)

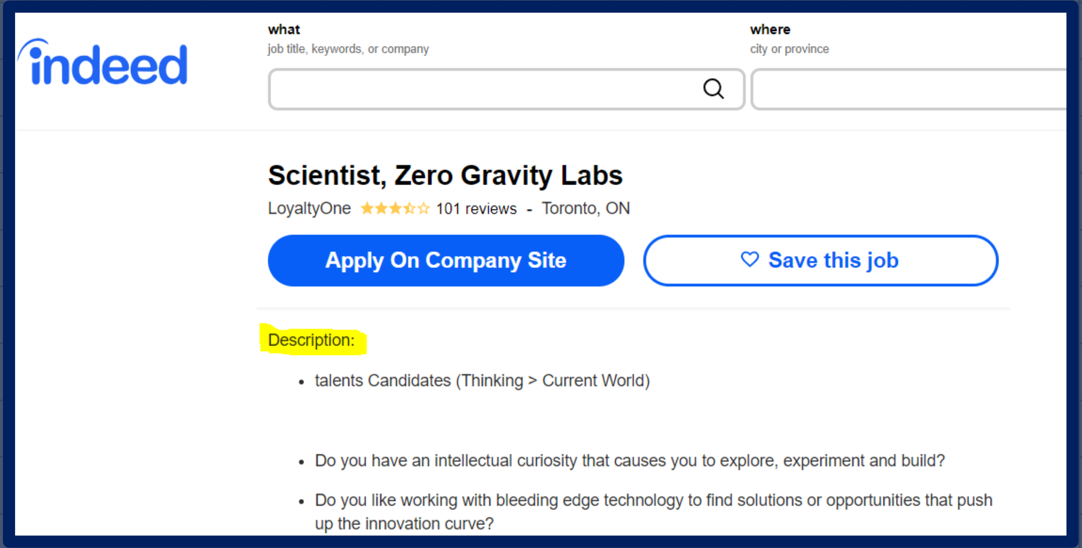
LoyaltyOne ★★★☆☆ 101 reviews

Toronto, ON

We are looking for a Manager, Data Science (Senior Data Scientist) to join our team in Toronto that has a deep

4th Step

Fetching Company Description



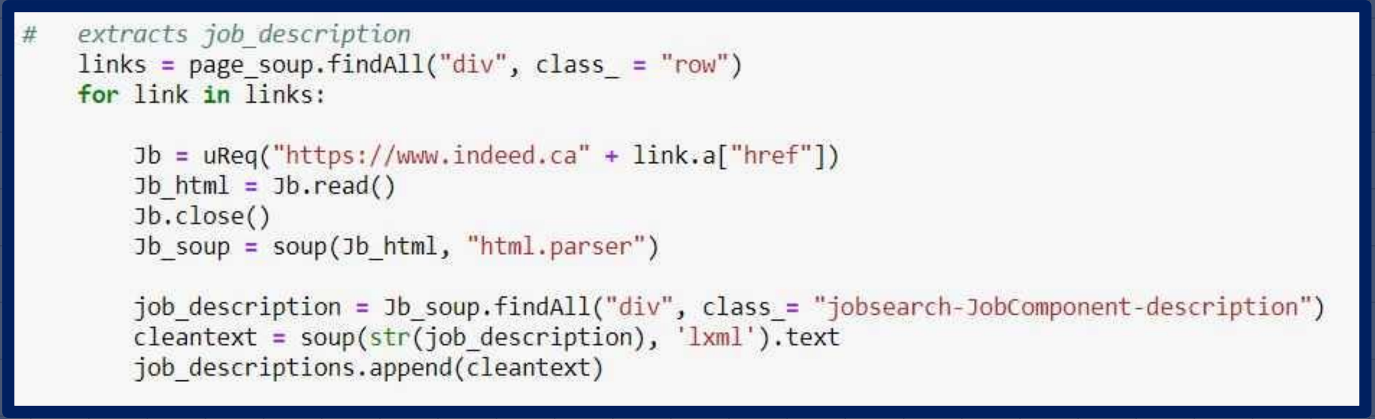
Scientist, Zero Gravity Labs

LoyaltyOne ★★★☆☆ 101 reviews - Toronto, ON

Description:

- talents Candidates (Thinking > Current World)

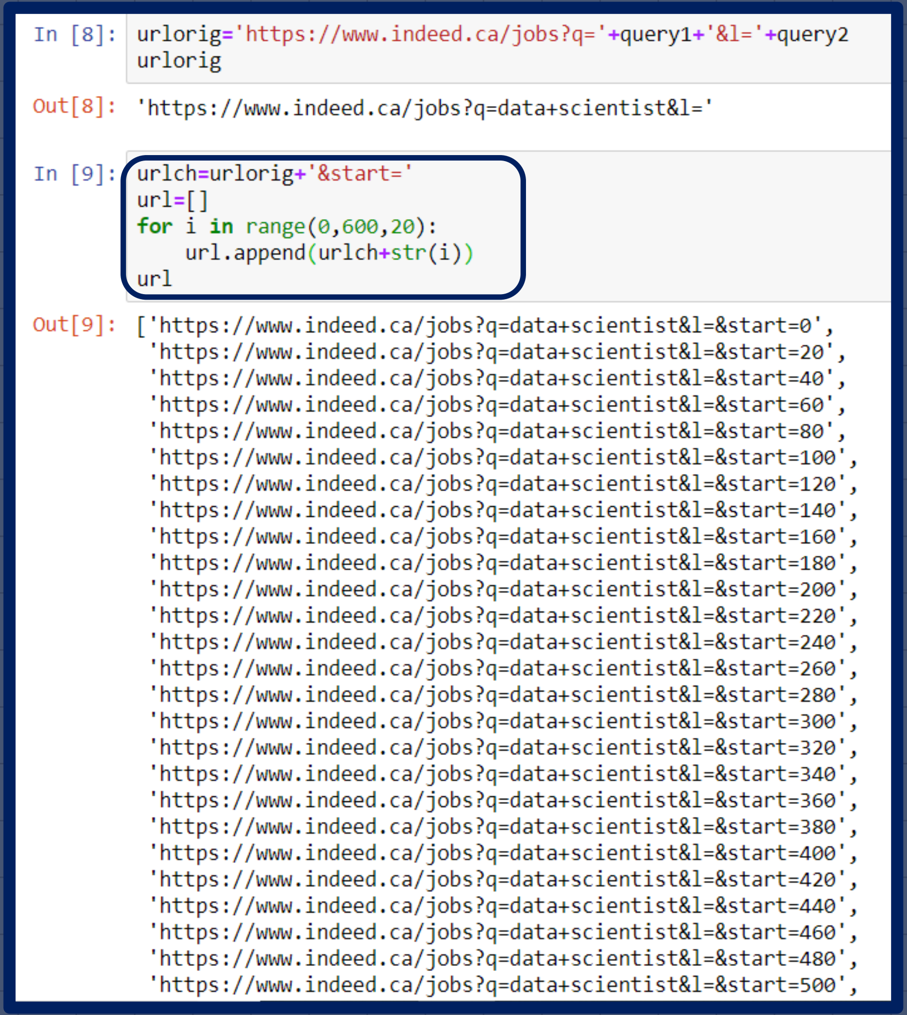
- Do you have an intellectual curiosity that causes you to explore, experiment and build?
- Do you like working with bleeding edge technology to find solutions or opportunities that push up the innovation curve?



```
#   extracts job_description
    links = page_soup.findAll("div", class_ = "row")
    for link in links:

        Jb = uReq("https://www.indeed.ca" + link.a["href"])
        Jb_html = Jb.read()
        Jb.close()
        Jb_soup = soup(Jb_html, "html.parser")

        job_description = Jb_soup.findAll("div", class_= "jobsearch-JobComponent-description")
        cleantext = soup(str(job_description), 'lxml').text
        job_descriptions.append(cleantext)
```

4th Step

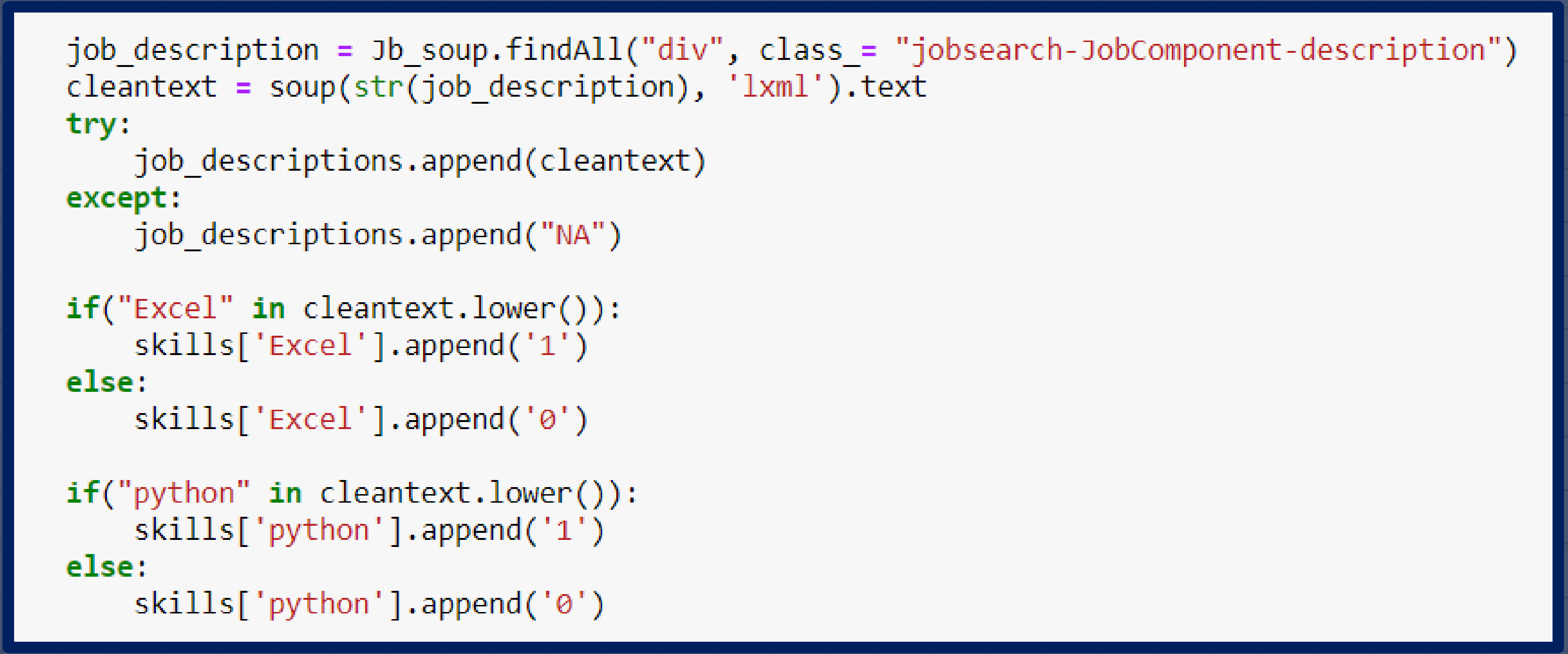
Repeating Step 2nd 3rd and 4th for 30 pages



```
In [8]:  urlorig='https://www.indeed.ca/jobs?q='+query1+'&l='+query2
         urlorig

Out[8]:  'https://www.indeed.ca/jobs?q=data+scientist&l='

In [9]:  urlch=urlorig+'&start='
         url=[]
         for i in range(0,600,20):
             url.append(urlch+str(i))
         url

Out[9]:  ['https://www.indeed.ca/jobs?q=data+scientist&l=&start=0',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=20',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=40',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=60',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=80',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=100',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=120',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=140',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=160',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=180',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=200',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=220',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=240',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=260',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=280',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=300',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=320',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=340',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=360',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=380',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=400',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=420',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=440',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=460',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=480',
          'https://www.indeed.ca/jobs?q=data+scientist&l=&start=500',
```

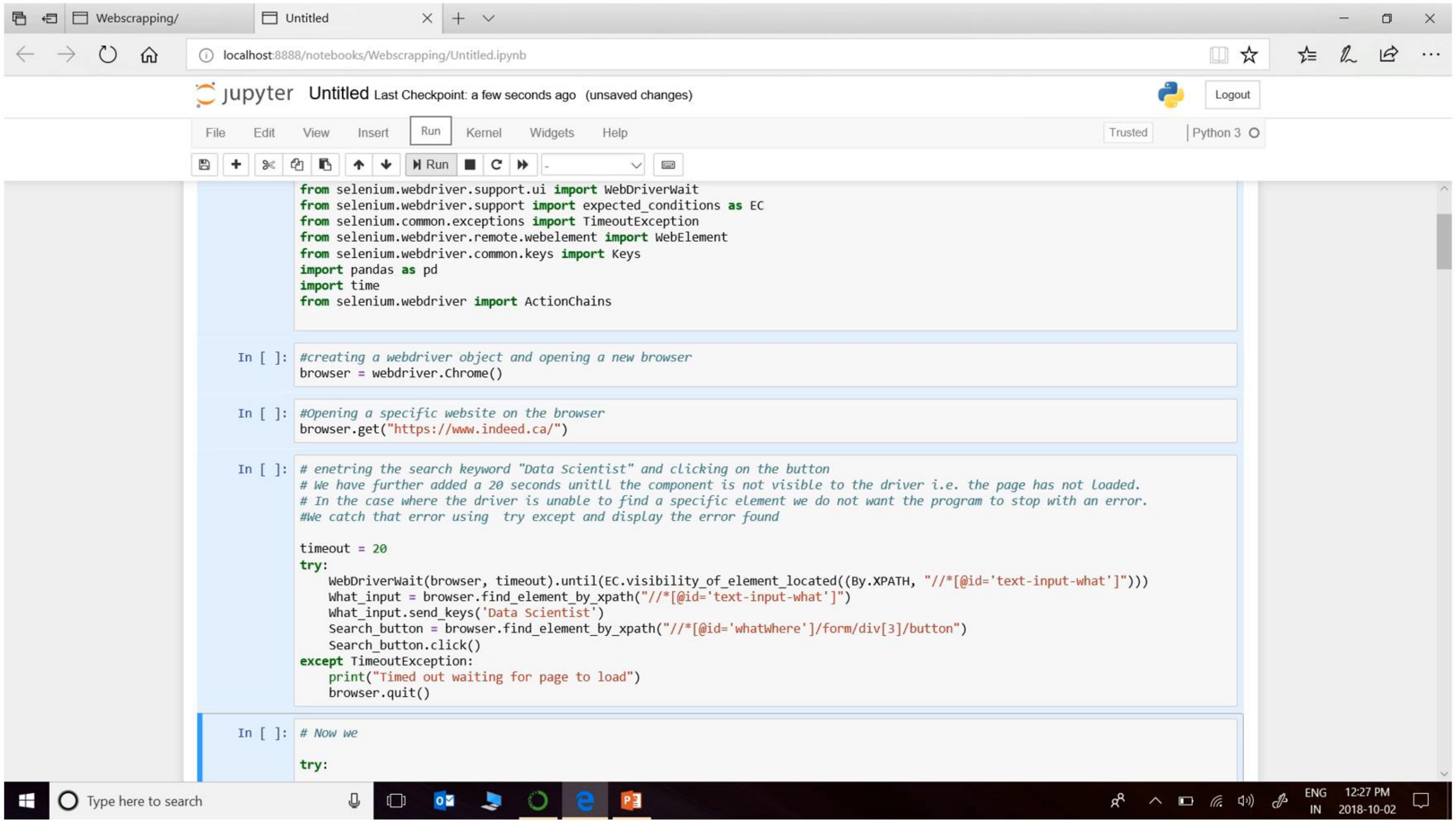
5th Step

Extracting Skills



```python
job_description = Jb_soup.findAll("div", class_= "jobsearch-JobComponent-description")
cleantext = soup(str(job_description), 'lxml').text
try:
    job_descriptions.append(cleantext)
except:
    job_descriptions.append("NA")

if("Excel" in cleantext.lower()):
    skills['Excel'].append('1')
else:
    skills['Excel'].append('0')

if("python" in cleantext.lower()):
    skills['python'].append('1')
else:
    skills['python'].append('0')
```

Implementation with Selenium


Webscrapping/
Untitled
localhost:8888/notebooks/Webscrapping/Untitled.ipynb
jupyter Untitled Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File Edit View Insert Run Kernel Widgets Help
Trusted
Python 3
Run
Type here to search
ENG IN
12:27 PM
2018-10-02

```
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import TimeoutException
from selenium.webdriver.remote.webelement import WebElement
from selenium.webdriver.common.keys import Keys
import pandas as pd
import time
from selenium.webdriver import ActionChains
```

In [ ]:
```
#creating a webdriver object and opening a new browser
browser = webdriver.Chrome()
```

In [ ]:
```
#Opening a specific website on the browser
browser.get("https://www.indeed.ca/")
```

In [ ]:
```
# enetring the search keyword "Data Scientist" and clicking on the button
# We have further added a 20 seconds untill the component is not visible to the driver i.e. the page has not loaded.
# In the case where the driver is unable to find a specific element we do not want the program to stop with an error.
#We catch that error using  try except and display the error found

timeout = 20
try:
    WebDriverWait(browser, timeout).until(EC.visibility_of_element_located((By.XPATH, "//*[@id='text-input-what']")))
    What_input = browser.find_element_by_xpath("//*[@id='text-input-what']")
    What_input.send_keys('Data Scientist')
    Search_button = browser.find_element_by_xpath("//*[@id='whatWhere']/form/div[3]/button")
    Search_button.click()
except TimeoutException:
    print("Timed out waiting for page to load")
    browser.quit()
```

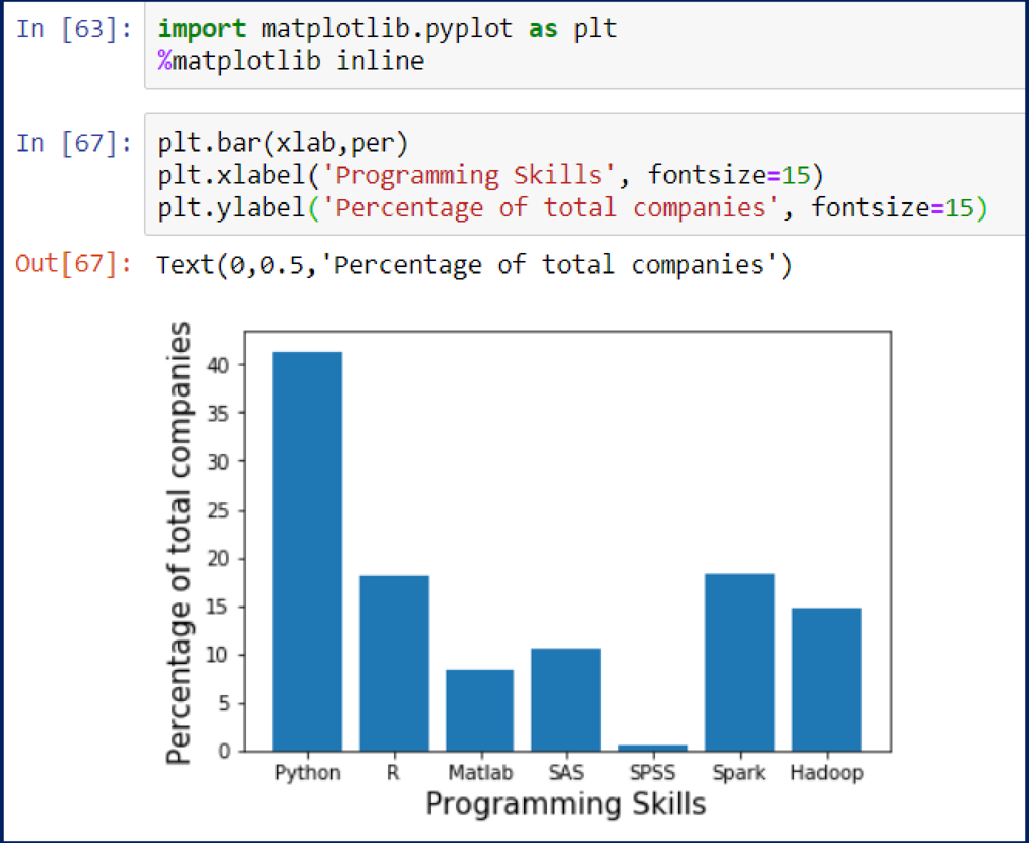
In [ ]:
```
# Now we

try:
```

Comparison

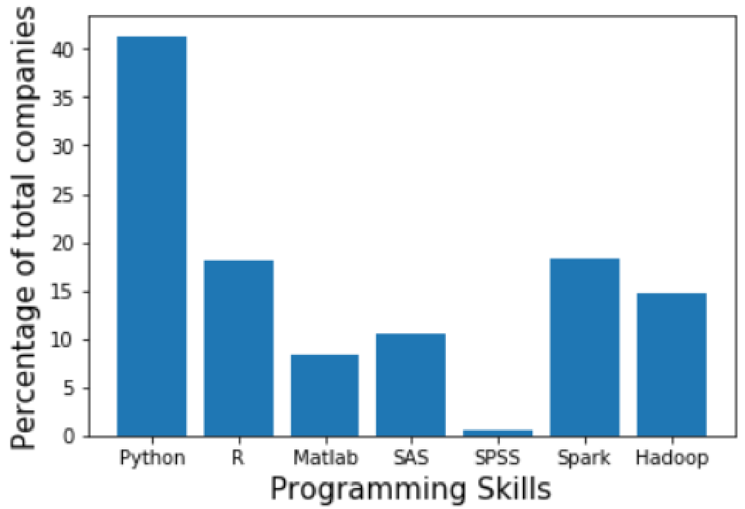
	BeautifulSoup	Se
Web Driver Requirement	✗	✓
Ease of Use	✓	✗

Comparison

	BeautifulSoup	Se
Can handle AJAX requests	✗	✓



```
In [63]: import matplotlib.pyplot as plt
         %matplotlib inline

In [67]: plt.bar(xlab,per)
         plt.xlabel('Programming Skills', fontsize=15)
         plt.ylabel('Percentage of total companies', fontsize=15)

Out[67]: Text(0,0.5,'Percentage of total companies')
```



References

- https://www.incapsula.com/web-application-security/web-scraping-attack.html#What-is-Web-Scraping
- https://www.dataquest.io/blog/web-scraping-beautifulsoup/
- http://webextract.net/
- https://www.dataquest.io/blog/web-scraping-tutorial-python/
- https://www.dataquest.io/blog/web-scraping-beautifulsoup/
- https://www.crummy.com/software/BeautifulSoup/
- https://www.crummy.com/software/BeautifulSoup/
- https://stackoverflow.com/questions/17436014/selenium-versus-beautifulsoup-for-web-scraping

Slide deck from SlidesCarnival

Icons from Flaticon

Thank you!