1. Introduction to Web Scraping

Definition & Purpose

- Web Scraping is the process of extracting data from websites using automated scripts.
- Commonly used for tasks like price tracking, job listings, and data analysis.

Key Points

- Web scraping automates the retrieval of online data, saving time and effort.
- Always review the website's Terms of Service (ToS) before scraping.

---

2. Understanding Website Structure

HTML Basics

- Websites are built using HTML elements (e.g., `<div>`, `<span>`, `<li>`).
- These elements can have classes or IDs to help identify them.

Inspecting the Website

- Use Developer Tools in your browser (right-click → "Inspect") to locate HTML elements.
- Look for repeated patterns or tags that contain the information you want to scrape.

---

3. Using the requests Module

What Is the `requests` Library?

- A Python library for sending HTTP requests to websites.
- Makes it simple to fetch web pages (HTML, JSON, etc.).

Key Steps

1. **Install & Import**
   - In Colab, `!pip install requests` (usually it's already installed).
   - Then, `import requests`.
2. **Send a GET Request**
   - Use `requests.get("your_website_url")` to retrieve the page content.
   - Check `response.status_code` to ensure the request was successful (`200` means OK).

```
import requests
from bs4 import BeautifulSoup

Key = "Data Scientist" # User input

url = f"https://www.timesjobs.com/candidate/job-search.html?searchType=personaliz

response = requests.get(url)

print("Status Code:",response.status_code)
```

```
Status Code: 200
```

```
response.text
```

```
{"type":"string"}
```

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(response.text,"html.parser")
```

```
jobs = soup.find_all('li',class_="clearfix job-bx wht-shd-bx")
```

```
print(f"Number of jobs on a page:{len(jobs)}")
```

```
Number of jobs on a page:25
```

Extract Company Name

```
jobs[1].find("h3",class_="joblist-comp-name").text.strip()
```

```
{"type":"string"}
```

```
company_names = []

for job in jobs:
  company_tag = job.find("h3",class_="joblist-comp-name")

  if company_tag:
    company_name = company_tag.text.strip()
  else:
    company_name = "Company name not found"

  company_names.append(company_name)

company_names
```

```
['Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data',
 'Hevo Data']
```

Extract Skills

```
jobs[0]
```

```
jobs[0].find("div",class_="more-skills-sections")
```

```
<div class="more-skills-sections">
<span>
                     data integration solutions
                 </span>
<span>
                     cloud based architectures
                 </span>
<span>
                     distributed systems knowledge
                 </span>
<span>
                     sql and nosql expertise
                 </span>
<span>
                     programming languages proficiency
                 </span>
<span>
                     python
                 </span>
<span>
                     c
                 </span>
<span>
                     cache
                 </span>
<span>
                     database
                 </span>
<span>
                     java
                 </span>
<span>
                     software engineering
                 </span>
</div>
```

```
skills_list=[]

for skill in jobs:
  skill_tag = skill.find("div",class_="more-skills-sections")

  if skill_tag:
    skill = skill_tag.text.strip().replace("\n","").replace("\t","")
  else:
    skill = "N/A"

  skills_list.append(skill)
```

```
skills_list
```

```
['data integration solutions\r   \r   cloud based architectures\r   \r   distributed
 'javascript programming\r   \r   front end development\r   \r   web performance opt
 'object oriented design   multithreaded applications   performance optimization   p
 'forecasting models\r   \r   hubspot administration\r   \r   data analytics\r   \r
 'backend development\r   \r   distributed systems\r   \r   cloud native architectur
 'operational management\r   \r   stakeholder engagement\r   \r   strategic planning
 'outbound sales strategies\r   \r   team leadership experience\r   \r   performance
 'object oriented design\r   \r   data integration solutions\r   \r   multithreaded
 'partner sales strategy\r   \r   outbound sales execution\r   \r   stakeholder enga
 'data integration solutions\r   \r   scalable architectures\r   \r   cloud native s
 'front end development\r   \r   ui  /  ux design\r   \r   javascript expertise\r
 'distributed systems\r   \r   cloud infrastructure\r   \r   programming languages\r
 'outbound sales strategies\r   \r   business case development\r   \r   technical ac
 'python programming\r   \r   distributed systems\r   \r   cloud infrastructure\r
 'outbound sales strategies\r   \r   team management skills\r   \r   performance tra
 'object oriented design\r   \r   multithreaded applications\r   \r   system perform
 'data pipeline development\r   \r   elt workflow design\r   \r   sql query proficie
 'sales cycle management\r   \r   outbound prospecting\r   \r   business case develo
 'excellent communication\r   \r   deep research skills\r   \r   content optimizatio
 'product ownership\r   \r   data integration\r   \r   stakeholder collaboration\r
 'content creation\r   \r   seo optimization\r   \r   research proficiency\r   \r
 'outbound sales strategies\r   \r   technical solution selling\r   \r   stakeholder
 'outbound sales strategies\r   \r   pipeline management\r   \r   technical acumen\r
 'sql proficiency\r   \r   data reporting\r   \r   dashboard creation\r   \r   stake
 'account management\r   \r   revenue growth\r   \r   customer retention\r   \r   st
```

Extract location

```
jobs[0]
```

```
<li class="clearfix job-bx wht-shd-bx">
<header class="clearfix">
<!--
-->
<!-- -->
<div class="d-flex d-flex-l-r job-title__logo">
<div class="d-flex d-flex-l-r">
<span class="logo-container">
<i class="default-company-logo"></i>
</span>
<div>
<h2 class="heading-trun" title="Staff Engineer">
<a href="https://www.timesjobs.com/job-detail/staff-engineer-hevo-data-bengaluru-ban
      Staff Engineer</a> </h2>
<div class="d-flex d-flex-align-item">
<h3 class="joblist-comp-name">
```

```
            Hevo Data

            </h3>
<span class="sim-posted">
<span>1 day ago</span>
</span>
</div>
</div>
</div>
</div>
</header>
<ul class="list-job-dtl clearfix">
<li class="job-description__">
            About Hevo DataUnleash the power of data at Hevo! Join our team of brill
         </li>
<li>
<div class="srp-skills">
<!-- The fixed Length to compare against -->
<div class="more-skills-sections">
<span>
                     data integration solutions
                 </span>
<span>
                     cloud based architectures
                 </span>
<span>
                     distributed systems knowledge
                 </span>
<span>
                     sql and nosql expertise
                 </span>
<span>
                     programming languages proficiency
                 </span>
<span>
                     python
                 </span>
<span>
                     c
                 </span>
<span>
                     cache
                 </span>
<span>
                     database
                 </span>
<span>
                     java
                 </span>
<span>
                     software engineering
                 </span>
</div>
</div>
</li>
</ul>
<div class="d-flex d-flex-l-r d-flex-align-item">
<ul class="top-jd-dtl mt-16 clearfix">
<li class="srp-zindex location-tru" title="Bengaluru / Bangalore">
<i class="srp-icons location"></i>

              Bengaluru / Bangalore



              </li>
<li><i class="srp-icons experience"></i>8 - 11 Years</li>
<li><i class="srp-icons salary"></i>Not disclosed</li>
</ul>
<div class="list-job-bt clearfix">
<div class="list-action">
<div class="applied-dtl clearfix" id="showPostApplyData_71577169">
<a class="apply-btn" href="javascript:callExtJobApply('71577169','adId=yfv__PLUS__LX
</div>
</div>
</div>
</div>
<!--
            <li>
              <i class="material-icons">location_on</i>
              Bengaluru / Bangalore
              </li>
-->
<a class="posoverlay_srp" href="https://www.timesjobs.com/job-detail/staff-engineer-
</li>
```

```python
jobs[9].find("li",class_="srp-zindex location-tru").text.strip()
```

```
{"type":"string"}
```

```python
locations_list=[]

for location in jobs:
  location_tag=location.find("li",class_="srp-zindex location-tru")

  if location_tag:
    location = location_tag.text.strip()
  else:
    location = "N/A"

  locations_list.append(location)

locations_list
```

```
['Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore',
 'Pune',
 'Bengaluru / Bangalore',
 'San Francisco,  Chicago,  Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore,  Pune',
 'San Francisco,  Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore,  Pune',
 'San Francisco,  Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore',
 'San Francisco,  Bengaluru / Bangalore,  Pune',
 'San Francisco,  Bengaluru / Bangalore']
```

Extract experience

```
jobs[0]
```

```
<li class="clearfix job-bx wht-shd-bx">
<header class="clearfix">
<!--
-->
<!-- -->
<div class="d-flex d-flex-l-r job-title__logo">
<div class="d-flex d-flex-l-r">
<span class="logo-container">
<i class="default-company-logo"></i>
</span>
<div>
<h2 class="heading-trun" title="Staff Engineer">
<a href="https://www.timesjobs.com/job-detail/staff-engineer-hevo-data-bengaluru-ban
      Staff Engineer</a> </h2>
<div class="d-flex d-flex-align-item">
<h3 class="joblist-comp-name">
             Hevo Data

             </h3>
<span class="sim-posted">
<span>1 day ago</span>
</span>
</div>
```

```
</div>
</div>
</div>
</header>
<ul class="list-job-dtl clearfix">
<li class="job-description__">
            About Hevo DataUnleash the power of data at Hevo! Join our team of brill
         </li>
<li>
<div class="srp-skills">
<!-- The fixed Length to compare against -->
<div class="more-skills-sections">
<span>
                    data integration solutions
                </span>
<span>
                    cloud based architectures
                </span>
<span>
                    distributed systems knowledge
                </span>
<span>
                    sql and nosql expertise
                </span>
<span>
                    programming languages proficiency
                </span>
<span>
                    python
                </span>
<span>
                    c
                </span>
<span>
                    cache
                </span>
<span>
                    database
                </span>
<span>
                    java
                </span>
<span>
                    software engineering
                </span>
</div>
</div>
</li>
</ul>
<div class="d-flex d-flex-l-r d-flex-align-item">
<ul class="top-jd-dtl mt-16 clearfix">
```

```
<li class="srp-zindex location-tru" title="Bengaluru / Bangalore">
<i class="srp-icons location"></i>

              Bengaluru / Bangalore



              </li>
<li><i class="srp-icons experience"></i>8 - 11 Years</li>
<li><i class="srp-icons salary"></i>Not disclosed</li>
</ul>
<div class="list-job-bt clearfix">
<div class="list-action">
<div class="applied-dtl clearfix" id="showPostApplyData_71577169">
<a class="apply-btn" href="javascript:callExtJobApply('71577169','adId=yfv__PLUS__LX
</div>
</div>
</div>
</div>
<!--
            <li>
              <i class="material-icons">location_on</i>
              Bengaluru / Bangalore
              </li>
-->
<a class="posoverlay_srp" href="https://www.timesjobs.com/job-detail/staff-engineer-
</li>
```

```python
experiences = []

for job in jobs:
    # Find all <li> elements in the job post
    li_tags = job.find_all('li')

    exp_text = "N/A"  # Default if experience is not found

    for li in li_tags:
        # Check if this <li> contains the experience <i> tag
        exp_text = li.find('i', class_='srp-icons experience')

        if exp_text:
          exp = li.text.strip()
          break

    experiences.append(exp)
experiences
```

```
['8 - 11 Years',
 '2 - 5 Years',
 '3 - 6 Years',
 '5 - 8 Years',
 '10 - 13 Years',
 '3 - 5 Years',
 '7 - 13 Years',
 '5 - 8 Years',
 '6 - 10 Years',
 '8 - 11 Years',
 '3 - 5 Years',
 '5 - 8 Years',
 '5 - 8 Years',
 '5 - 8 Years',
 '8 - 11 Years',
 '3 - 6 Years',
 '3 - 5 Years',
 '5 - 8 Years',
 '1 - 3 Years',
 '5 - 9 Years',
 '1 - 5 Years',
 '6 - 9 Years',
 '6 - 9 Years',
 '3 - 6 Years',
 '10 - 13 Years']
```

Extract Salary

```
jobs[10]
```

```
<li class="clearfix job-bx wht-shd-bx">
<header class="clearfix">
<!--
-->
<!-- -->
<div class="d-flex d-flex-l-r job-title__logo">
<div class="d-flex d-flex-l-r">
<span class="logo-container">
<i class="default-company-logo"></i>
</span>
<div>
<h2 class="heading-trun" title="Engineering Manager">
<a href="https://www.timesjobs.com/job-detail/engineering-manager-hevo-data-bengalur
      Engineering Manager</a> </h2>
<div class="d-flex d-flex-align-item">
<h3 class="joblist-comp-name">
            Hevo Data

            </h3>
<span class="sim-posted">
<span>few days ago</span>
</span>
</div>
</div>
</div>
</div>
</header>
<ul class="list-job-dtl clearfix">
<li class="job-description__">
```

```
            About HevoAt Hevo ,  we are changing the way companies leverage data to
         </li>
<li>
<div class="srp-skills">
<!-- The fixed Length to compare against -->
<div class="more-skills-sections">
<span>
                     backend development
                 </span>
<span>
                     distributed systems
                 </span>
<span>
                     cloud native architecture
                 </span>
<span>
                     team leadership
                 </span>
<span>
                     system design
                 </span>
<span>
                     mentor
                 </span>
<span>
                     technical leader
                 </span>
<span>
                     technical leadership
                 </span>
<span>
                     database
                 </span>
<span>
                     security
                 </span>
<span>
                     software development
                 </span>
<span>
                     mentoring
                 </span>
<span>
                     saas
                 </span>
<span>
                     infrastructure
                 </span>
<span>
                     performance optimization
                 </span>
```

```
</div>
</div>
</li>
</ul>
<div class="d-flex d-flex-l-r d-flex-align-item">
<ul class="top-jd-dtl mt-16 clearfix">
<li class="srp-zindex location-tru" title="Bengaluru / Bangalore">
<i class="srp-icons location"></i>

              Bengaluru / Bangalore



              </li>
<li><i class="srp-icons experience"></i>10 - 13 Years</li>
<li><i class="srp-icons salary"></i>Not disclosed</li>
</ul>
<div class="list-job-bt clearfix">
<div class="list-action">
<div class="applied-dtl clearfix" id="showPostApplyData_71639022">
<a class="apply-btn" href="javascript:callExtJobApply('71639022','adId=bNF__PLUS__4V
</div>
</div>
</div>
</div>
<!--
            <li>
              <i class="material-icons">location_on</i>
              Bengaluru / Bangalore
              </li>
-->
<a class="posoverlay_srp" href="https://www.timesjobs.com/job-detail/engineering-man
</li>
```

```python
salary_list=[]


for salary in jobs:
  li_tag = salary.find_all("li")

  for li in li_tags:
    salary = li.find("i",class_="srp-icons salary")

    if salary:
      salary = li.text.strip().replace("\n","").replace("\t","")
      break

  salary_list.append(salary)

salary_list
```

```
['Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed',
 'Not disclosed']
```

Putting Things Together

```
import requests
from bs4 import BeautifulSoup

def job_finder():

  Key = "Data Scientist"


  url = f"https://www.timesjobs.com/candidate/job-search.html?searchType=personalize

  response = requests.get(url)
  soup = BeautifulSoup(response.text,"html.parser")

  jobs = soup.find_all('li',class_="clearfix job-bx wht-shd-bx")

  company_names = []
  locations_list=[]
  experiences = []
  salary_list=[]

  if not jobs:
    print("No job has found,let inspect html:")
    print(soup.prettify()[:2000])

  #COMPANY_NAME
  for job in jobs:
    company_tag = job.find("h3",class_="joblist-comp-name")
    if company_tag:
      company_name = company_tag.text.strip()
    else:
      company_name = "Company name not found"

    company_names.append(company_name)

  #LOCATION_NAME

    location_tag=job.find("li",class_="srp-zindex location-tru")

    if location_tag:
      location = location_tag.text.strip()
    else:
      location = "N/A"

    locations_list.append(location)

  #EXPERIENCE

    # Find all <li> elements in the job post
    li_tags = job.find_all('li')

    exp_text = "N/A"  # Default if experience is not found

    for li in li_tags:
        # Check if this <li> contains the experience <i> tag
        exp_text = li.find('i', class_='srp-icons experience')

        if exp_text:
          exp = li.text.strip()
          break

    experiences.append(exp)

  #SALARY

    li_tag = job.find_all("li")

    for li in li_tags:
      salary = li.find("i",class_="srp-icons salary")

      if salary:
        salary = li.text.strip().replace("\n","").replace("\t","")
        break

    salary_list.append(salary)



  result = {"Company_name":company_names,
            "Location_name":locations_list,
            "Experiences":experiences,
            "Salary":salary_list}

  print(f"Extracted {len(jobs)} job postings:")
  return result


data =  job_finder()

import pandas as pd

# Convert dictionary of lists into a Pandas DataFrame
df = pd.DataFrame(data)

# Save to CSV
df.to_csv("scraped_jobs.csv", index=False)
print("Data saved to scraped_jobs.csv")

# Display the first few rows in tabular form
display(df)
```

```
Extracted 25 job postings:
Data saved to scraped_jobs.csv
```

{"summary":"{\n \"name\": \"df\",\n \"rows\": 25,\n \"fields\": [\n   {\n   \

```
from google.colab import files

# Download the CSV file
files.download('scraped_jobs.csv')
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.Javascript object>
```