Summary Report Web Scraping with BeautifulSoup and Requests

Mahir Hashimov
12695120
Unit 3
Deciphering Big Data
University of Essex Online
16 June 2024

Contents

Web Scraping

Objective

The objective of this task was to write a web scraping script in Python to extract job listings for the keyword "Data Scientist" from Indeed.com and parse this data into an XML file. The process involved using the beautifulsoup4 and requests Python libraries.

Steps Taken

1. *Setting Up the Environment*

- Installed necessary libraries using pip:

```
pip install requests
```

```
pip install beautifulsoup4
```

2. *Identifying the Webpage and Data for Scraping*

- Target URL: https://www.indeed.com/jobs?q=Data+Scientist&l=
- Identified that the required data includes job titles, company names, locations, and job links.

3. *Writing the Initial Code*

- Sent an HTTP GET request to the URL.
- Parsed the HTML response using BeautifulSoup.
- Located job postings using HTML element classes and tags.

4. *Handling Errors and Challenges*

- HTTP 403 Error (Forbidden): Initial requests to the website returned a 403 error, indicating forbidden access. This was resolved by adding headers to mimic a browser request:

```
headers = {
```

```
 'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36'
```

```
}
```

```
response = requests.get(url, headers=headers)
```

5. *Parsing HTML and Extracting Data*

- Used BeautifulSoup to locate the job elements on the page:

```
job_elements = soup.find_all('div', class_='slider_container')
```

- Extracted job titles, company names, locations, and job links from the elements.

6. *Storing Data in XML Format*

- Used the xml.etree.ElementTree library to create an XML structure and save the data:

```
import xml.etree.ElementTree as ET
jobs = ET.Element('jobs')
for job_element in job_elements:
  title_element = job_element.find('h2', class_='jobTitle')
  company_element = job_element.find('span', class_='companyName')
  location_element = job_element.find('div', class_='companyLocation')
  link_element = job_element.find('a', class_='jcs-JobTitle')['href']
  job = ET.SubElement(jobs, 'job')
  title = ET.SubElement(job, 'title')
  title.text = title_element.text.strip()
  location = ET.SubElement(job, 'location')
  location.text = company_element.text.strip() +
location_element.text.strip()
  link = ET.SubElement(job, 'link')
  link.text = 'https://www.indeed.com' + link_element
tree = ET.ElementTree(jobs)
tree.write('jobs.xml', encoding='utf-8', xml_declaration=True)
```

7. *Saving and Reviewing the Output*

- Saved the XML file and reviewed it to ensure the correct data was extracted and formatted.

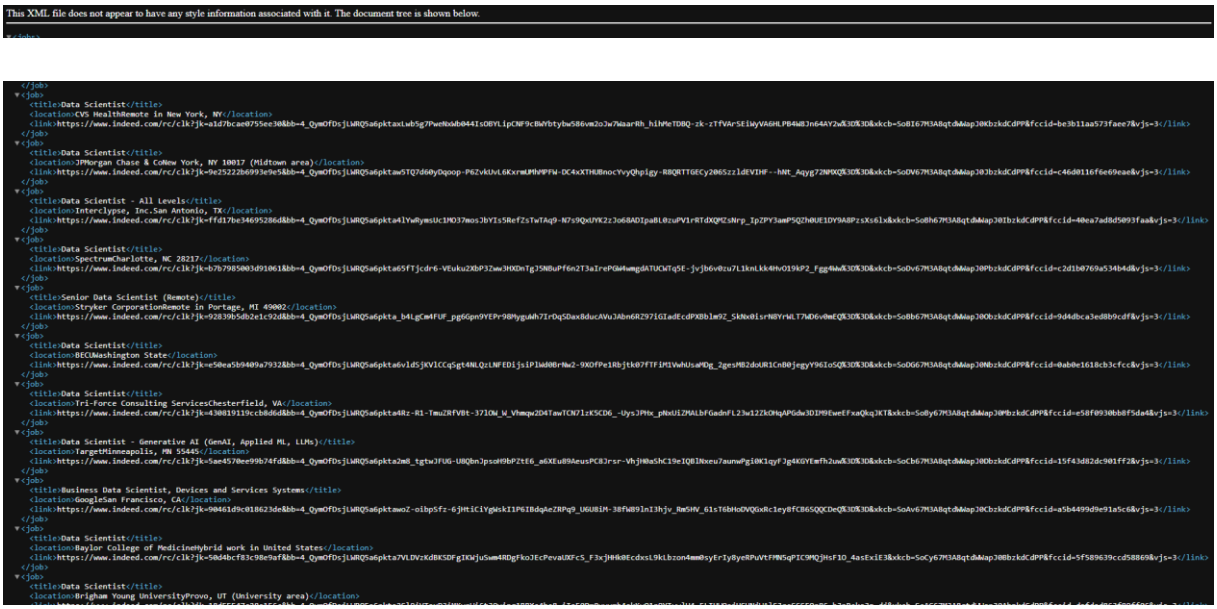
8. *Challenges and Resolutions*

- 403 Forbidden Error: Initially faced a 403 error which was resolved by adding appropriate headers to the request to mimic a browser.
- Locating HTML Elements: Had to inspect the webpage to find the correct HTML tags and classes to extract the required data.
- XML Formatting: Ensured the extracted data was correctly formatted in the XML structure.

Conclusion

The project involved a combination of web scraping techniques, handling HTTP errors, parsing HTML content, and structuring data in XML format. Despite facing initial challenges with access restrictions and HTML parsing, these were overcome through methodical troubleshooting and adjusting the code to handle different scenarios. The final script successfully extracts job listings from Indeed.com and saves them in an XML file, which was then added to a GitHub repository for easy access and sharing.

Output


This XML file does not appear to have any style information associated with it. The document tree is shown below.
</job>
</job>
<title>Data Scientist</title>
<location>CVS HealthRemote in New York, NY</location>
<link>https://www.indeed.com/rc/clk?jk=a1d7bcae0755ee30&bb=4_QymOfDsjLWRQ5a6pktaxlwb5g7PweNoAb044IsOBYLipCNF9cBWYbtybw586vm2oJw7WaarRh_hihWeTDBQ-zk-zTfVArSEiWyVA6HLPB4WBJn64AY2w%3D%3D&xkcb=SoBI67M3A8qtdWWap30KbzkdCdPP&fccid=be3b11aa573faee7&vjs=3</link>
</job>
<job>
<title>Data Scientist</title>
<location>JPMorgan Chase & CoNew York, NY 10017 (Midtown area)</location>
<link>https://www.indeed.com/rc/clk?jk=9e25222b6993e9e5&bb=4_QymOfDsjLWRQ5a6pktaw5TQ7d60yDqoop-P6ZvkUvL6KxrmUMNMPFW-DC4xXTHU8nocYvyQhpigy-R8QRTTGECy2065zzldEVIHF--hNt_Aqyg72NMXQ%3D%3D&xkcb=SoDV67M3A8qtdWWap30JbzkdCdPP&fccid=c46d0116f6e69eae&vjs=3</link>
</job>
<job>
<title>Data Scientist - All Levels</title>
<location>Interclypse, Inc.San Antonio, TX</location>
<link>https://www.indeed.com/rc/clk?jk=ffd17be34695286d&bb=4_QymOfDsjLWRQ5a6pkta4lYwRymsUc1MO37mosJbYIs5RefZsTwTAq9-N7s9QxUYK2zJo6ADIpaBL0zuPV1rRTdXQMZsNrp_IpZPY3amPSQZhOUE1DY9A8PzsXc61xkxkcb=So8h67M3A8qtdWWap30IbzkdCdPP&fccid=40ea7ad8d5093faa&vjs=3</link>
</job>
<job>
<title>Data Scientist</title>
<location>SpectrumCharlotte, NC 28217</location>
<link>https://www.indeed.com/rc/clk?jk=b7b7985003d91061&bb=4_QymOfDsjLWRQ5a6pkta65fTjcdr6-VEuku2XbP32zw3HXDnTgj5NBuPf6n2T3aIrePGWwmgdATUCWTq5E-jvjb6v0zu7L1knLkk4HvO19kP2_Fgg4WW%3D%3D&xkcb=SoDv67M3A8qtdWWap30PbzkdCdPP&fccid=c2d1b0769a534b4d&vjs=3</link>
</job>
<job>
<title>Senior Data Scientist (Remote)</title>
<location>Stryker CorporationRemote in Portage, MI 49002</location>
<link>https://www.indeed.com/rc/clk?jk=92839b5db2e1c92d&bb=4_QymOfDsjLWRQ5a6pkta_b4LgCm4fUF_pg6Gpn9YEPr98MygaWh7IrDqSDax8ducAVuJAbn6RZ97IGIadEcdPXBblm9Z_SkNx0isrNBYrWLT7WD6v0mEQ%3D%3D&xkcb=SoBb67M3A8qtdWWap30ObzkdCdPP&fccid=9d4dbca3ed8b9cdf&vjs=3</link>
</job>
<job>
<title>Data Scientist</title>
<location>BECUWashington State</location>
<link>https://www.indeed.com/rc/clk?jk=e50ea5b9409a7932&bb=4_QymOfDsjLWRQ5a6pkta6vldSjKVlCqSgt4NLQzLNFEDijsiPlw80BrNw2-9XOfPe1Rbjtk07fIfiM1VwhUsaMDg_2gesMB2doUR1CnB0jegyY96IoSQ%3D%3D&xkcb=SoDG67M3A8qtdWWap30NbzkdCdPP&fccid=0ab8e1618cb3cfcc&vjs=3</link>
</job>
<job>
<title>Data Scientist</title>
<location>Tri-Force Consulting ServicesChesterfield, VA</location>
<link>https://www.indeed.com/rc/clk?jk=430819119ccb8d6d&bb=4_QymOfDsjLWRQ5a6pkta4Rz-R1-TmuZRfVBt-371OW_W_Vhmqv2D4TawTCN7lzKSCD6_-UysJPHx_pNxUi2MALbfGadnFL23w12ZkOHqAPGdw3DIM9EweEFxaQkqJKT&xkcb=SoBy67M3A8qtdWWap30MbzkdCdPP&fccid=e58f0930bb8f5da4&vjs=3</link>
</job>
<job>
<title>Data Scientist - Generative AI (GenAI, Applied ML, LLMs)</title>
<location>TargetMinneapolis, MN 55445</location>
<link>https://www.indeed.com/rc/clk?jk=5ae457@ee99b74fd&bb=4_QymOfDsjLWRQ5a6pkta2m8_tgtwJFUG-UBQbnJpsoH9bPZtE6_a6XEu89AeusPC8Jrsr-VhjH0a5hC19eIQBlNxeu7aunwPgi0K1qyFJg4KGYEmfh2uw%3D%3D&xkcb=SoCb67M3A8qtdWWap30ObzkdCdPP&fccid=15f43d82dc901ff2&vjs=3</link>
</job>
<job>
<title>Business Data Scientist, Devices and Services Systems</title>
<location>GoogleSan Francisco, CA</location>
<link>https://www.indeed.com/rc/clk?jk=90461d9c018623de&bb=4_QymOfDsjLWRQ5a6pktawoZ-oibpSfz-6jHtiCiYgWsk11P6I8dqAeZRPq9_U6U8iM-38fW891mI3hjv_Rm5HV_61sT6bHoDVQGxRc1eyBfCB6SQQCDeQ%3D%3D&xkcb=SoAv67M3A8qtdWWap30CbzkdCdPP&fccid=a5b4499d9e91a5c6&vjs=3</link>
</job>
<job>
<title>Data Scientist</title>
<location>Baylor College of MedicineHybrid work in United States</location>
<link>https://www.indeed.com/rc/clk?jk=58d4bcf83c98e9af&bb=4_QymOfDsjLWRQ5a6pkta7VLDVzKdBKSDFgIKWjuSwm4RDgFkoJEcPevaUXFcS_F3xJHHk0EcdxsL9kibzon4mm0syErIyByeRPuVtFMN5qPIC9MQjHsF1O_4asExiE3%3D%3D&xkcb=SoCy67M3A8qtdWWap30BbzkdCdPP&fccid=5f589639ccd58869&vjs=3</link>
</job>
<job>
<title>Data Scientist</title>
<location>Brigham Young UniversityProvo, UT (University area)</location>
</job>
<job>
<title>Data Scientist</title>
<location>Brigham Young UniversityProvo, UT (University area)</location>
<link>https://www.indeed.com/rc/clk?jk=a1d7bcae0755ee30&bb=4_QymOfDsjLWRQ5a6pktaxlwb5g7PweNoAb044IsOBYLipCNF9cBWYbtybw586vm2oJw7WaarRh_hihWeTDBQ-zk-zTfVArSEiWyVA6HLPB4WBJn64AY2w%3D%3D&xkcb=SoBI67M3A8qtdWWap30KbzkdCdPP&fccid=be3b11aa573faee7&vjs=3</link>