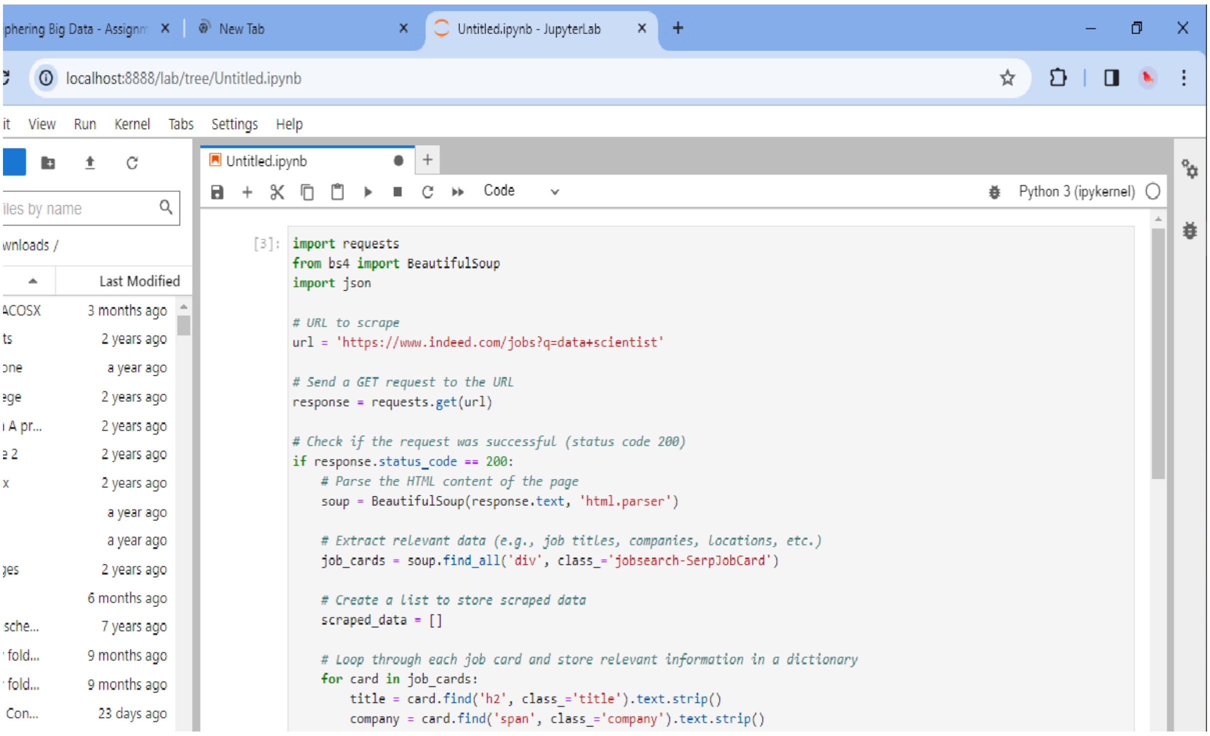
UNIT 2: Web scraping Exercise

The main artefact for Unit 3 was to use the beautifulsoup4 and Request program modules in Python to perform web scraping , the key word being 'Data Scientist' and then parsing the data into either an XML or JSON file. Perform the web scraping with the beautifulsoup4 and Request program modules. Although i did not successfully scrape the data, I learned a lot about data privacy laws as I kept getting a “Failed to retrieve data. Status code: 403” output upon running my code. Status codes indicate information about what happened with a request, for example 403: The resource you’re trying to access is forbidden: you don’t have the right permissions to see it (Grupman, 2020).

Learning Outcomes

- Identify and manage challenges, security issues and risks, limitations, and opportunities in data wrangling.
- Critically analyse data wrangling problems and determine appropriate methodologies, tools, and techniques (involving preparing, cleaning, exploring, creating, optimising and evaluating big data) to solve them.
- Systematically develop and implement the skills required to be effective member of a development team in a virtual professional environment, adopting real life perspectives on team roles and organisation. (University of Essex Online , 2024)



```
import requests
from bs4 import BeautifulSoup
import json

# URL to scrape
url = 'https://www.indeed.com/jobs?q=data+scientist'

# Send a GET request to the URL
response = requests.get(url)

# Check if the request was successful (status code 200)
if response.status_code == 200:
    # Parse the HTML content of the page
    soup = BeautifulSoup(response.text, 'html.parser')

    # Extract relevant data (e.g., job titles, companies, locations, etc.)
    job_cards = soup.find_all('div', class_='jobsearch-SerpJobCard')

    # Create a list to store scraped data
    scraped_data = []

    # Loop through each job card and store relevant information in a dictionary
    for card in job_cards:
        title = card.find('h2', class_='title').text.strip()
        company = card.find('span', class_='company').text.strip()
```

REFERENCES:

Grupman, C. (2020) Python API Tutorial: Getting Started with APIs: https://www.dataquest.io/blog/python-api-tutorial/ [Accessed 18th December 2023].