Data and Artificial Intelligence

Cyber Shujaa Program

Week 1 Assignment
Web Scraping and Data Handling in Python

Student Name: Deborah Kwamboka Omae

Student ID: CS-DA02 -25075

Introduction

First week's assignment was about web scrapping which simply means automatically extracting data from a website. To achieve this, I used python programming language together with helpful libraries such as requests, Beautifulsoup and pandas. Requests is used to download the HTML content from the website, Beautifulsoup to parse through the HTML content and pandas to organize the data in a structured format. The website had structured data about Hockey team scores and these scrapping was done on Google Colab which is a free cloud-based platform provided by google that lets you write python code in your web browser, in my case Google Chrome. After extracting the data, it was saved as a csv file.

The objectives of the assignment were:

1. Practical Python coding on Jupiter Notebooks hosted on Google Colab
2. Use requests and BeautifulSoup to extract data from a web page.
3. Parse and clean the extracted data.
4. Store structured data into a Pandas DataFrame.
5. Export the final dataset to a .csv file.

Tasks Completed

Step 1: Importing of python libraries used for web scrapping.

I imported the libraries: Requests, BeautifulSoup and pandas to be able to collect data from the website and organize it in a structured format.

The code:

from bs4 import BeautifulSoup

import requests

import pandas as pd

These lines of code import the given libraries through the keyword ‘import’ in python

Step2: Fetching the HTML webpage content

The Code:

url = 'https://www.scrapethissite.com/pages/forms/'

page = requests.get(url)

I set the URL of the website we are scrapping in a variable called url. To fetch the html content, the get method of requests is used and the object returned is stored in the variable page. That response object contains the HTML of the page (in .text), the HTTP status code (in. status code) and headers.

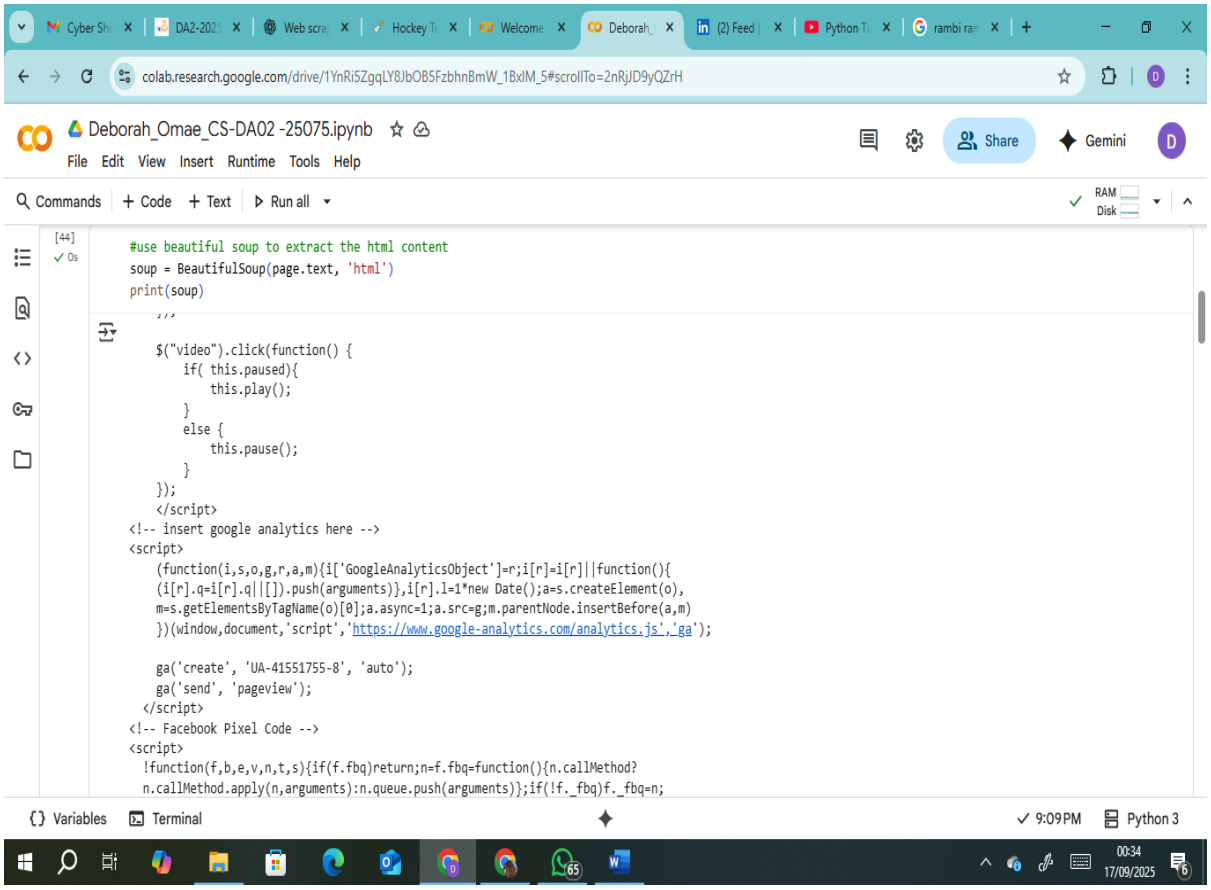
Step 3: parsing the html content into a navigable structure

The code:

#Use BeautifulSoup to extract the HTML content

soup = BeautifulSoup(page.text, 'html')

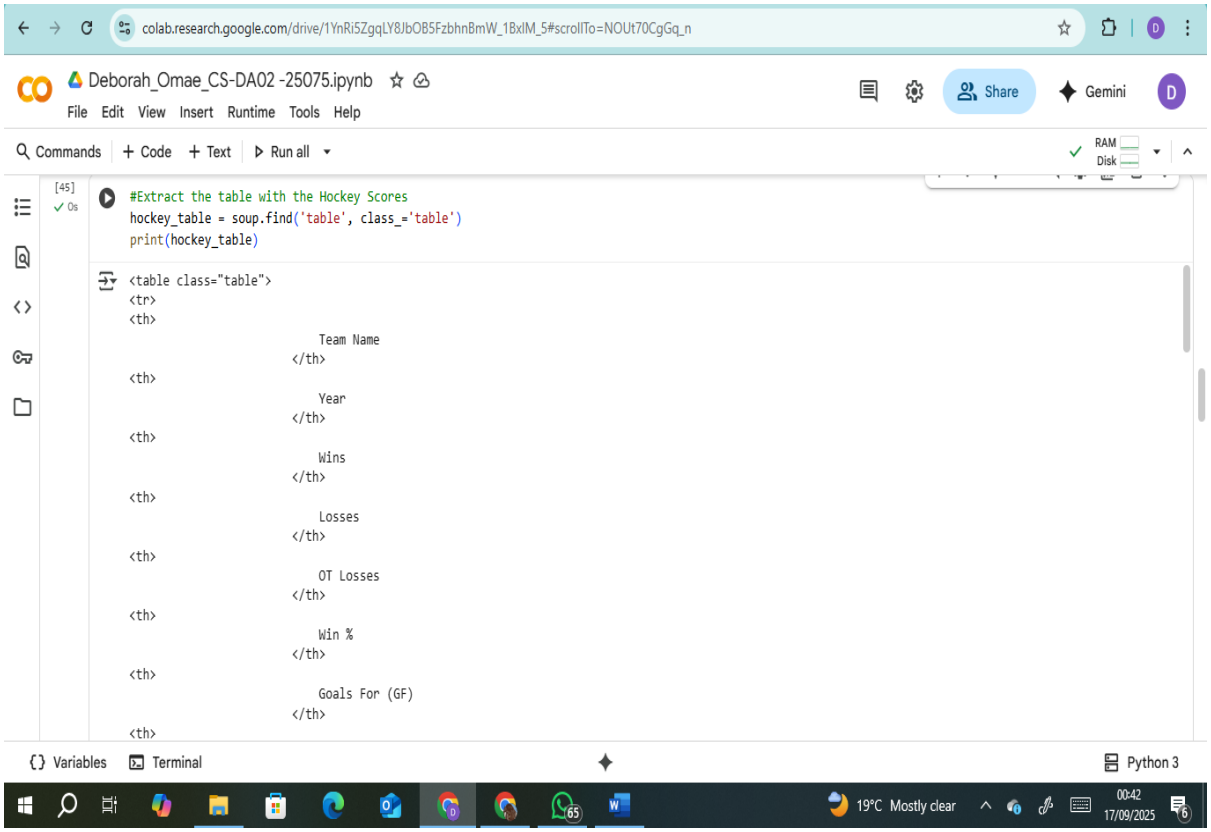
print(soup)



#Extract the table with the Hockey Scores

hockey_table = soup.find('table', class_='table')

print(hockey_table)



#Extract the column headings

table_titles = hockey_table.find_all('th')

hockey_table_title = [title.text.strip() for title in table_titles]

print(hockey_table_title)

Step 4: storing the rows into a table like structure using pandas' data frame

The code:

#Save the column headings onto a Pandas DataFrame

df = pd.DataFrame(columns=hockey_table_title)

df

#Extract the data row by row. First get all rows, then loop through each while stripping and saving data into the DataFrame

table_data = hockey_table.find_all('tr')

for row in table_data[1:]:

raw_data = row.find_all('td')

each_raw_data = [data.text.strip() for data in raw_data]

print(each_raw_data)

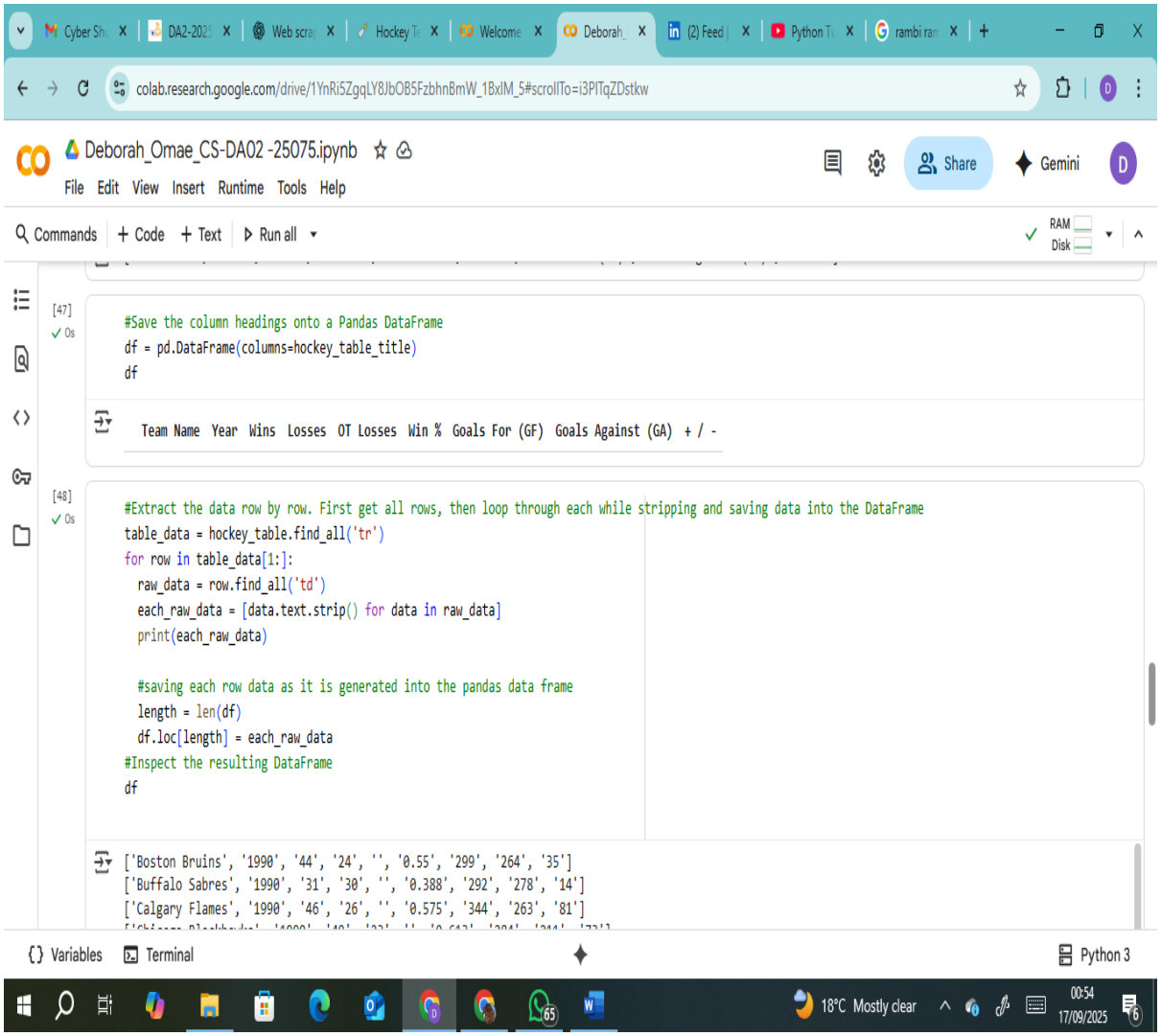
#saving each row data as it is generated into the pandas data frame

length = len(df)

df.loc[length] = each_raw_data

#Inspect the resulting DataFrame

df



I extracted the data row by row then looped through each while stripping and saving the data into a data frame. The .loc in pandas is used to access or insert rows and columns by their labels, and in the above code it's adding each new scraped row into the DataFrame at the next index. The data is now in a structured format.

Step 5: storing the data frame to disk (.csv file)

The code:

df.to_csv(r'./Hockey.csv')

Link to Code:

https://colab.research.google.com/drive/1YnRi5ZgqLY8JbOB5FzbhnBmW_1BxlM_5?usp=sharing

Conclusion

This first week I have had an insightful introduction to the history of AI and how data plays an important role in the emergence and development of AI. Job roles in the field of data and AI were highlighted. Starting my first project on web scrapping is an eye opener of the data science methodology in particular, the fetching of data and organizing it into a well-structured format. I have posted my writeup on my blog and I look forward to building a portfolio that I can showcase on my CV as I look for jobs in Data and AI.