DSBDA: C1- WEB SCRAPING

1 ) !pip install requests beautifulsoup4 ------already satisfied written

2) import requests -----------nothing works

from bs4 import BeautifulSoup

3) url = "https://books.toscrape.com/"

headers = {

    "User-Agent": "Mozilla/5.0"

}

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

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

4) books = soup.find\_all("article", class\_="product\_pod")

print("Total books found:", len(books))

for book in books[:5]: # Print first 5 books

title = book.h3.a["title"]

price = book.find("p", class\_="price\_color").text

rating = book.p["class"][1] # e.g., "One", "Two", "Three"

print("Title:", title)

print("Price:", price)

print("Rating:", rating)

print("-" \* 50)

5) soup.select('article h3 a')[0]['title']

6) import pandas as pd

import time

base\_url = 'https://books.toscrape.com/catalogue/page-{}.html'

headers = {'User-Agent': 'Mozilla/5.0'}

titles = []

prices = []

ratings = []

availability = []

# Loop through the first 5 pages (change range as needed)

for page in range(1, 6):

print(f"Scraping page {page}...")

response = requests.get(base\_url.format(page), headers=headers)

soup = BeautifulSoup(response.content, 'html.parser')

books = soup.find\_all('article', class\_='product\_pod')

for book in books:

title = book.h3.a['title']

price = book.find('p', class\_='price\_color').text

availability\_status = book.find('p', class\_='instock availability').text.strip()

rating\_class = book.find('p', class\_='star-rating')['class'][1]

titles.append(title)

prices.append(price)

ratings.append(rating\_class)

availability.append(availability\_status)

time.sleep(1) # Respectful scraping

# Create DataFrame

df = pd.DataFrame({

'Title': titles,

'Price': prices,

'Rating': ratings,

'Availability': availability

})

df.head()

7) soup.select('article p')[1].string

8) soup.select('article p')[0]['class']