05/02/2025

Web Scraping

- **What is Web Scraping?**

Web scraping is the process of extracting data from websites using automated scripts. It allows us to collect and structure data that is publicly available on web pages.

- **Why Do We Do Web Scraping?**

Web scraping is useful for:
✅ **Data collection** – Gathering information from websites for analysis.
✅ **Automation** – Extracting data without manual copy-pasting.
✅ **Price monitoring** – Tracking product prices on e-commerce sites.
✅ **Research** – Collecting data for academic or business insights.

- **Requirements for Web Scraping**

To perform web scraping, you need:

1. **Python Installed** – The programming language to run the script.
2. **Libraries** – Required tools for scraping, such as:
   - selenium – Automates web browsers.
   - pandas – Helps store and process extracted data.
3. **Web Browser & Driver** – Example: Google Chrome + ChromeDriver.
4. **Understanding of HTML & XPath** – To locate and extract specific elements from web pages.

Extracts **table data** from a webpage and saves it as a CSV file

1) Importing the Libraries

```
from selenium import webdriver

from selenium.webdriver.chrome.service import Service

from selenium.webdriver.common.by import By

import pandas as pd

import time
```

□ selenium: Automates web browser actions.

□ pandas: Manages and stores scraped data.

□ time: Adds delays to avoid loading issues.

2) **Set Up Chrome WebDriver**

```
options = webdriver.ChromeOptions()

options.add_argument("--headless")  # Runs Chrome in headless mode (no UI)

options.add_argument("--window-size=1920,1080")  # Sets the window size

driver = webdriver.Chrome()  # Initializes the Chrome browser
```

Why?

- We use Chrome WebDriver to **automate** browsing.
- --headless mode means the browser runs **in the background** (without opening a window).

3) **Open the Web Page**

```
driver.get("https://kb.corel.com/en/125936")  # Open the Corel webpage
time.sleep(1)  # Wait 1 second to ensure page loads completely
```

Why?

- driver.get(url): Opens the given URL.
- time.sleep(1): Waits for the page to load properly.

4) Locate the Table on the Webpage

```
tables = driver.find_element(By.XPATH, "//table")  # Find the table element
all_table_rows = tables.find_elements(By.XPATH, ".//tr")  # Find all rows in the table
```

Why?

- find_element(By.XPATH, "//table"): Finds the **first** table on the page.
- find_elements(By.XPATH, ".//tr"): Finds **all rows (<tr>)** inside the table.

5) Extract Data from the Table

```
list_of_rows = []  # Create an empty list to store extracted data

for each_row in all_table_rows:  # Loop through each row
    list_of_data = []  # Create an empty list for row data
    all_data = each_row.find_elements(By.XPATH, ".//td")  # Find all columns in the row
    for data in all_data:  # Loop through each column
        list_of_data.append(data.text)  # Extract and store text
        print(list_of_data)  # Print extracted data (optional)
    list_of_rows.append(list_of_data)  # Add row data to the main list
```

6) Convert Extracted Data to a CSV File

```
df = pd.DataFrame(list_of_rows[1:], columns=list_of_rows[0])  # Create a Pandas DataFrame
df.to_csv("korel.csv", index=False)  # Save it as "korel.csv"
print(df)  # Print the extracted data
```