**Muhammad Kaleem**

**56614**

**Bs – Data Science**

**3rd semester**

***Assignment no 4***

***Web Scraping :***

Web scraping is the process of extracting data from websites. It involves using automated tools, such as scripts or software, to gather and parse the content of web pages. The extracted data can then be stored, analyzed, or used for various purposes, such as building datasets, tracking prices, or gathering information for research.

### Key Steps in Web Scraping:

1. **Fetching the Webpage**: The scraper sends an HTTP request (often using libraries like requests) to a website’s URL to get the HTML content of the page.
2. **Parsing the HTML**: After fetching the webpage, a parser (such as BeautifulSoup in Python) is used to extract specific data from the HTML structure by targeting tags, classes, or IDs.
3. **Data Extraction**: The desired information (like product prices, titles, reviews, etc.) is then extracted by identifying and navigating through the webpage's HTML structure.
4. **Data Storage**: After extracting the data, it can be stored in various formats such as CSV files, databases, or even directly used in further applications like data analysis.

#### Libraries of python for web scraping:

* requests: For making HTTP requests to access the HTML content of the webpage.
* BeautifulSoup or lxml: For parsing and navigating the HTML content to extract specific elements.
* Selenium: For scraping websites that require JavaScript interaction.

***Code:***

import requests

from bs4 import BeautifulSoup

def scrape\_news\_headlines(url):

"""

Scrapes news headlines from the given URL.

Args:

url (str): The URL of the news website.

Returns:

list: A list of news headlines.

"""

response = requests.get(url)

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

# Find all headline elements (on CNN, they may be within <h3> or <span> tags)

headlines = soup.find\_all('h3', class\_='cd\_\_headline')

# Extract and clean the text for each headline

headline\_list = [headline.text.strip() for headline in headlines]

return headline\_list

if \_\_name\_\_ == '\_\_main\_\_':

url = 'https://edition.cnn.com/world'

headlines = scrape\_news\_headlines(url)

# Print each headline

for idx, headline in enumerate(headlines, 1):

print(f"{idx}. {headline}")

***Summary of the code :***

### How This Code Works:

1. **Requests** is used to fetch the HTML content of the webpage (in this case, a news section from CNN).
2. **BeautifulSoup** is used to parse the HTML and search for headlines. News headlines are typically within certain tags (like <h3> or <span>), so we target those using find\_all.
3. **Extract Text**: For each headline found, the text is extracted, cleaned (by removing extra spaces), and stored in a list.
4. **Display Results**: Finally, the script prints out the scraped headlines.