Python Development – Tumelo Matobo

Task 1: Python Script for Web Scraping and Automation

CNN News Article Scraper

This Python script is designed to scrape news article headlines from the CNN World News Section (https://edition.cnn.com/world). It utilizes the **requests** library to fetch the webpage content, **BeautifulSoup** from **bs4** to parse the HTML content, and **pandas** to organise and save the data into a CSV file. The script operates by sending a GET request to the CNN World News URL, parsing the returned HTML to extract news headlines contained within specific HTML elements, and organising there elements into a structured format. Each scraped headline is timestamped with the current date and time when the script is run. The primary purpose of this script is to automate the collection of new headlines for data analysis, content aggregation, or monitoring the news cycle. It can be particularly useful for researchers, journalists, and data analysts looking to analyse trends in news coverage or for developers building applications that require real-time new data.

Dependencies

- **Python:** The script is written in Python and requires a Python interpreter to run. If you haven't installed Python on your system, download and install it from the official website. https://www.python.org/downloads/
- **Requests:** These are used for sending HTTP requests to the CNN website.
- **BeautifulSoup:** A library for parsing HTML and XML documents, making it easier to navigate and search the parse tree.
- **Pandas:** An open-source data analysis and manipulation tool, used for organising data into a DataFrame and exporting it to a CSV file.

To install these dependencies, open a terminal or command prompt and run the following commands:

```
pip install requests beautifulsoup4 pandas
```

How to Run the Script

1. Ensure that Python and the required libraries are installed in your system.
2. Download and save the Python script **Web_Scraper_News_Website.py** on your computer in a desired directory.
3. Open a terminal or command prompt.
4. Navigate to the directory where the script is saved.
5. Run the script in a Python environment by using the following command:

```
python Web_Scraper_News_Website.py
```

Script Breakdown

- **Fetching Web Content:** The script starts by sending a GET request to the CNN World News URL using the **requests.get** method. The response from this request contains the HTML content of the page.
- **Parsing HTML:** The **BeautifulSoup** library is then used to parse the HTML content. The **html.parser** navigates the HTML structure to locate content of interest.
- **Scraping Headlines:** The script searches for all elements with a **container__headline-text** class, which contain the news headlines. These headlines are extracted, stripped of leading or trailing whitespace, and stored in a list.
- **Data Organization:** the headlines are then organised into a list of dictionaries, which is converted into a **pandas** DataFrame. This structured format makes the data easier to manipulate and analyse.
- **Timestamp and File Saving:** A timestamp is generated using the current date and time to ensure that each file is uniquely named. The DataFrame is then saved to a CSV file named **cnn_news_article_<time>.csv** where **<time>** is replaced with the actual timestamp at the time of running the script.
- **Output:** The script prints a message indicating the successful scraping and saving of news articles, along with the name of the CSV file created.
- **Automation:** The script can be scheduled to run at specific intervals, such as daily or weekly, using built-in scheduling tools such as Task Scheduler for Windows and Cron for Linux and macOS.

Limitations and Considerations

- The script is tailored to the specific structure of the CNN World News section as of the time it was written. Any changes to the website's HTML structure may require adjustments to the script

Code

cnn2_news_articles_scraper.py

Task 1 > Web Scraping > cnn2_news_articles_scraper.py > ...

```python
import requests
from bs4 import BeautifulSoup
import pandas as pd
from datetime import datetime

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

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

#Use BeautifulSoup to parse the HTML content
soup = BeautifulSoup(response.text, 'html.parser')

#Find all the article headline elements
articles = soup.find_all('span', class_='container__headline-text')
#Initialize a list to store the scraped data
news_article = []

for article in articles:
    #Extract the headline text
    headline = article.text.strip()

    #Append the scraped data to the list
    news_article.append({'Headline': headline})

#Convert the list of dictionaries to a pandas DataFrame
df = pd.DataFrame(news_article)

# Create a timestamp for the CSV file
time = datetime.now().strftime('%Y%m%d_%H%M%S')

file_name = f'cnn_news_article_{time}.csv'
#Save the DtaFrame to a CSV file
df.to_csv(file_name, index = False)

print(f"News articles scraped and saved to {file_name}")

```

Output

The output of the CNN News Article Scraper script is a CSV file containing the headlines of news articles scraped from the CNN World News section. The CSV file is structured with a single column labelled Headline where each row corresponds to a unique news headline. The output file is named using a specific format that includes a timestamp to ensure uniqueness for each run of the script.

	A
1	**Headline**
2	Three UAE soldiers killed in attack on military base in Somalia
3	The latest on the Israel-Hamas war
4	King Charles makes first public outing since cancer diagnosis
5	Russian forces push into Ukraine's Avdiivka, piling pressure on new army chief
6	How Indonesia's future is in the hands of young voters, in 5 charts
7	Unimaginable devastation seen inside Khan Younis, the southern Gaza city once a safe haven for the displaced
8	King Charles thanks public in first message since cancer diagnosis
9	Venezuela builds forces near border with Guyana despite agreement to de-escalate
10	A philosopher's words emerge from charred, ancient scrolls
11	Where is Hind? Calls for answers more than a week after rescuers go missing trying to save trapped 5-year-old
12	How Gaza's hospitals became battlegrounds
13	Enter the Year of the Dragon: A 2024 guide to Lunar New Year
14	How the climate crisis fuels gender inequality
15	Harrowing audio reveals the moment a family was killed in Gaza
16	Nigerian Banking CEO, family among those killed in helicopter crash
17	Study: Longtime ice sheet could melt again; raise sea levels
18	Concerns over the rise of Germany's far-right AfD party
19	Ukraine air defenses under pressure as two Russian missile types again evade interception
20	How much does the public have a right to know about King Charles' cancer diagnosis?
21	King Charles' cancer was 'caught early,' British PM says
22	One killed during attack on main courthouse in Istanbul
23	I envy people who have a grave to visit': Earthquake survivors in Turkey struggle to rebuild their lives one year on
24	King Charles III has cancer and will step back from public duties
25	Five-year-old Palestinian girl found dead after being trapped in car under Israeli fire
26	Why only a trickle of aid is getting into Gaza
27	Netanyahu directs Israeli military to draw up plan to evacuate more than one million people from Rafah as offensive looms
28	Israel's repudiation of a deal with Hamas draws fury from hostages' families
29	Netanyahu says Hamas' demands on hostage and ceasefire deal are 'delusional'

30	A barred gate, a musty chamber and dirty dishes: Inside the underground compound where Israel says hostages were held
31	In pictures: Lunar New Year celebrations
32	The week in 32 photos
33	People we've lost in 2024
34	In pictures: Britain's King Charles III
35	Internet blackout hits Sudan as UN appeals for $4.1 billion to ease 'epic suffering' caused by war
36	West Africa bloc urges Burkina, Niger and Mali not to withdraw
37	Thousands sheltering in hospital as fighting escalates in Democratic Republic of Congo
38	Suspect awaiting extradition over US nurse murder escapes Kenya cell
39	Haiti elections will take place when security improves, PM Henry says as protests grow
40	Brazil's former president Bolsonaro under investigation in probe into attempted coup
41	Rio declares dengue emergency as Brazil gears up for Carnival
42	Nicaragua grants political asylum to former Panamanian President Ricardo Martinelli
43	Pakistan releases official election results, independents affiliated with Khan's PTI secure most seats
44	Myanmar junta enforces compulsory military service law
45	Russia has recruited as many as 15,000 Nepalis to fight its war. Many returned traumatized. Some never came back
46	In shock result, allies of jailed ex-leader Khan win most seats in Pakistan election
47	It's not too late: Where to see April's total solar eclipse
48	Coldplay has a request: They really, really want to play in China
49	Ancient Xi'an was once a key starting point for Silk Road journeys. It also hosts one of China's most stunning lantern shows
50	Chinese zodiac fortune predictions: What's in store for the Year of the Dragon
51	It's a brand new route on the world's most famous train - but it'll cost you $8,500 one way
52	A photographer's fantastical portrait of rural China during Lunar New Year
53	Look of the Week: Zendaya is the latest A-lister to lean into method dressing
54	Joy for maker of tallest matchstick Eiffel Tower as Guinness World Records reverses initial rejection
55	New documentary 'Your Fat Friend' turns unflinching gaze towards anti-fatness
56	British Vogue features 40 'legendary' cover stars for editor Edward Enninful's final issue

57	It's spring in the Midwest. The only problem? It's midwinter
58	Severe thunderstorms could bring heavy rain to the Southeast as a snowmaker takes aim at the Northeast
59	Super El Nino' is here, but La Nina looks likely. What's in store for the coming months
60	Volcanoes Fast Facts
61	Atlantic Ocean circulation nearing devastating tipping point, study finds
62	Taylor Swift performing in Tokyo before flying to Las Vegas for Super Bowl
63	Violent clashes erupt in northern India following the demolition of a mosque and Islamic school
64	Farmers' protests have erupted across Europe. Here's why
65	Allies of Pakistan's jailed ex-leader deliver surprise victory in general elections
66	Biden, German Chancellor push for Ukraine aid
67	Growing discontent among Europe's farmers
68	Protests in Pakistan as election results trickle in
69	Girl's 'miraculous; rescue offers hope for families of dozens buried in Philippines landslide
70	Ukrainian military sees changes after top brass shakeup
71	February 9, 2024 Israel-Hamas war
72	Outmanned and outgunned: Ukraine's new army chief faces big challenges in taking the fight to Russia
73	See inside Sofia Vergara's sumptuous Los Angeles home
74	From Bulgari to Armani, ultra-luxe branded residences are booming in Dubai
75	The quiet American: US reveals pavilion design for World Expo 2025
76	Remains of ancient Roman triumphal arch unearthed in Serbia
77	Step inside Emma Stone's sunlit LA home, up for sale at $4 million
78	Super Bowl: What to know about the Chiefs-49ers matchup
79	Extraordinary quality': The 20-year-old Englishman leading Real Madrid's La Liga title charge
80	Host nation Ivory Coast continues miraculous run to AFCON final to set up a matchup against Nigeria
81	Chinese authorities cancel second Argentina soccer match after Messi backlash in Hong Kong
82	Dr. Kwane Stewart, who cares for the pets of those experiencing homelessness, is CNN's Hero of the Year
83	Jason Momoa, Amanda Seyfried, other stars to take part in tonight's CNN Heroes event
84	Meet the people who are making the world a better place
85	CNN Heroes 2023 Voting Disclosures

This documentation provides a comprehensive overview of the CNN News Article Scraper script, including its purpose, dependencies, operation, and the detailed process it follows to scrape news headlines from CNN World News. By understanding how the script works, users can effectively run it, analyse the collected data, and adapt it to their specific needs or similar projects.