CC PROJECT REPORT

Web Scraping: Weather Report & News Tracker

Names, Registration Numbers:

Hamna Jamila, 18

Hmael Asif, 52

Introduction

This project explores the use of web scraping techniques in Python within the context of compiler construction. Web scraping involves extracting data from websites, which often requires understanding and interpreting HTML structures. The project demonstrates two Python scripts: one for weather data extraction using regex patterns and another for tracking news headlines from a BBC RSS feed. The aim is to showcase the flexibility of Python in handling real-world applications like gathering live data efficiently.

Code and Output

Code 1: Weather Scraper

```python
import re
import requests

def scrape_weather_with_regex(city_name):
    try:
        # Prepare the URL
        complete_url = f"https://wttr.in/{city_name}"

        # Fetch the raw HTML
        response = requests.get(complete_url)
        if response.status_code != 200:
            print(f"Could not fetch data. Status:
{response.status_code}")
            return

        # Get the raw HTML content
        html_content = response.text
        print(html_content)  # Debugging: View the raw HTML content

        # Update regex patterns based on actual HTML structure
        weather_pattern = r"Weather:.*?(\w+)"  # Adjust this based on
HTML
        temperature_pattern = r"(-?\d+°C)"  # Matches temperature like
25°C
        wind_speed_pattern = r"(\d+\s?km/h)"  # Matches wind speed like
15 km/h
        humidity_pattern = r"(\d+%)"  # Matches humidity like 80%

        # Match patterns in the HTML
        weather_summary = re.search(weather_pattern, html_content)
```

```python
        temperature = re.search(temperature_pattern, html_content)
        wind_speed = re.search(wind_speed_pattern, html_content)
        humidity_level = re.search(humidity_pattern, html_content)

        # Display results
        print("Weather details:")
        print(f"City: {city_name.capitalize()}")
        print(f"Summary: {weather_summary.group(1) if weather_summary
else 'N/A'}")
        print(f"Temperature: {temperature.group(1) if temperature else
'N/A'}")
        print(f"Wind Speed: {wind_speed.group(1) if wind_speed else
'N/A'}")
        print(f"Humidity: {humidity_level.group(1) if humidity_level
else 'N/A'}")

    except Exception as error:
        print(f"Something went wrong: {error}")

# Sample usage
if __name__ == "__main__":
    city_name = "Islamabad"
    scrape_weather_with_regex(city_name)
```

Output :


```
Weather report: Islamabad

      \   /     Clear
       .-.      +10(9) °C
    - (   ) -   ↘ 6 km/h
       `-'      10 km
      /   \     0.0 mm

                             ┌─────────────┐
┌──────────────────────────┤  Sun 26 Jan  ├──────────────────────────┐
│          Morning          └──────┬──────┘         Evening           │
├──────────────────────────┬───────────────────┬─────────────────────┤
│      \   /     Sunny      │    \   /     Sunny │     \   /     Clear │
│       .-.      13 °C      │     .-.      20 °C │      .-.      12 °C │
│    - (   ) -   ↘ 6-9 km/h │  - (   ) -   → 10-12 km/h │ - (   ) -   → 8-17 km/h │
│       `-'      10 km      │     `-'      10 km │      `-'      10 km │
│      /   \     0.0 mm | 0%│    /   \     0.0 mm | 0%│    /   \     0.0 mm | 0%│
```

```
Location: Islamabad, ICT, 44000, پاکستان [33.6945756,73.0643744]

Follow @igor_chubin for wttr.in updates

Weather details:
City: Islamabad
Summary: N/A
Temperature: N/A
Wind Speed: N/A
Humidity: 0%
```

Code 2: News Tracker

```python
import re
import requests

def get_news_with_cleaned_output():
    try:
        # URL of BBC News RSS feed
        url = "http://feeds.bbci.co.uk/news/rss.xml"

        # Fetch the RSS feed content
        response = requests.get(url)
        if response.status_code != 200:
            print(f"Could not fetch data. Status:
{response.status_code}")
            return
```

```python
        # Get the raw XML content
        rss_content = response.text

        # Regular expression to match news headlines
        headline_pattern = r"<title>(<!\[CDATA\[.*?\]\]>|.*?)</title>"

        # Find all matches
        headlines = re.findall(headline_pattern, rss_content)

        # Clean the headlines to remove CDATA tags
        cleaned_headlines = [re.sub(r"<!\[CDATA\[(.*?)\]\]>", r"\1",
headline) for headline in headlines]

        # Display the extracted and cleaned news headlines
        print("Latest News Headlines:")
        # Skip the first headline (it's usually the feed title)
        for headline in cleaned_headlines[1:]:
            print(f"- {headline.strip()}")

    except Exception as error:
        print(f"Something went wrong: {error}")

# Sample usage
if __name__ == "__main__":
    get_news_with_cleaned_output()
```

Output:

```
Latest News Headlines:
- BBC News
- Chancellor hints at support for third runway at Heathrow
- Reeves urges platforms to remove violent content viewed by Southport killer
- Confident, organised, still freewheeling: Trump 2.0 has learned from past
- 'I spent 30 years searching for secret to happiness - the answer isn't what I thought'
- UK hit with high winds as Storm Herminia rolls in
- Trump says he wants Egypt and Jordan to take in Palestinians from Gaza
- Lebanon says 15 killed by Israeli forces after withdrawal deadline missed
- How texting a random number 'hello' led a man to meet his future wife
- Wynne Evans apologises for 'inappropriate' comment at Strictly tour
- The 2010s lost classics that became sleeper hits a decade on
- Holocaust survivors fear Europe is forgetting the lessons of Auschwitz
- 'My story on Tinder predator went live, then my phone rang - it was him'
- Nick Cave says work repelled him after death of sons
- 'I don't want to buy a £4 coffee just so I can use the loo'
- Mexican workers set up tent city to house deportees from US
- The Traitors' best moments: A breakfast bust-up, a steamy shower, a secret priest
- Meta is ditching fact checkers for X-style community notes. Will they work?
- Trump says Keir Starmer doing 'very good job'
- Huge crowds await return to northern Gaza after delays
- Man charged with murdering lecturer
- S Korean president charged with insurrection over martial law attempt
- Traitors' Leanne on wedding plans and baby hopes
- CIA says lab leak most likely source of Covid outbreak
- Holly Willoughby speaks of 'tough' year after kidnap plot
- BBC News app
- Is this a Rachel Reeves reset?
- Bonus Q&A: Trump returns to the White House
- Merciless Sinner wins back-to-back Melbourne titles
- SAS Rogue Heroes: Gentleman Jim was 'quiet but deadly'
- BBC reporter spends a year undercover to expose far-right group
- Police contacted over school rugby tour
- Scotland's dying art of traditional nicknames
```