INTERN CAREER

"WE SPEAK DATA"

Python Development Intern's

Task 1

Introduction

The project is to work on web scraping and automation we write a python script to extract data from a website and automate a specific task. The main goal of this project is to make use of web scraping techniques to obtain useful information from a selected website, making it easier to create and maintain an updated dataset. The chosen website, eBay, is notable for its extensive and freely available material, which makes it a great place to find pertinent information.Using Python scripts libraries like Beautiful Soup and Requests, our goal is to methodically gather particular data points.The requirement for fast and accurate information retrieval, which serves as a basis for data analysis, market insights, and other applications.The project's automation component, which is accomplished by having a script run on an established schedule daily, weekly, etc. guarantees that the dataset is up to date and consistent with the website's dynamic content.

Website selection

- For this project ,I have choosen BBC news channel website which is recognized globally with a wealth of publicly accessible data.I have choosed this bbc website for webscraping and to extract the headline news because of several reasons,

➢ Structured HTML parser

➢ Availability of content of daily bases

➢ Consistent live updates

➢ Global reputed organization

➢ Trusted by people which is very familiar

Python code

```
import pandas as pd
import matplotlib.pyplot as plt
from apscheduler.schedulers.blocking import BlockingScheduler

def scrape_bbc_headlines():
    url = 'https://www.bbc.com/news' # URL of the bbc News website
    response = requests.get(url)
    if response.status_code == 200: # Check if the request was successful
        soup = BeautifulSoup(response.text, 'html.parser')
        # Find elements that contain headlines (specific to the bbc structure)
        headlines = soup.find_all('h3', class_='gs-c-promo-heading__title')
        # Extract text from the headline elements
        headline_texts = [headline.get_text(strip=True) for headline in headlines]
        return headline_texts
    else:
        print("Failed to retrieve data")
```

Headline Extraction

```
# Test the function
headlines = scrape_bbc_headlines()
if headlines:
    for idx, headline in enumerate(headlines, start=1):
        print(f"Headline {idx}: {headline}")
    else:
        print("No headlines found")

# Create a scheduler
scheduler = BlockingScheduler()

# Schedule the scraping function to run every hour (you can change the interval as needed)
scheduler.add_job(scrape_bbc_headlines, 'interval', hours=1)
try:
    scheduler.start()
except KeyboardInterrupt:
    # If you want to stop the scheduler manually (Ctrl+C)
    passscheduler = BlockingScheduler()
```

Output:

- Headline 1: Iraq warns of disastrous consequences for region after US strikes
- Headline 2: Why did US wait to retaliate for drone attack on its troops?
- Headline 3: Cancer doctor takes gamble to treat his brain tumour
- Headline 4: Misinformation spreads in China on ‘civil war’ in Texas
- Headline 5: 'My mother doesn't know if she's married or a widow'
- Headline 6: Brilliant Bumrah blows England away in second Test
- Headline 7: Malaysia halves ex-PM's jail term for corruption
- Headline 8: Imran Khan and wife jailed for illegal marriage
- Headline 9: Inert nuclear missile found in US man's garage
- Headline 10: US approves $4bn sale of armed drones to India
- Headline 11: Inert nuclear missile found in US man's garage
- Headline 12: US approves $4bn sale of armed drones to India
- Headline 13: China executes couple who killed two toddlers
- Headline 14: Three wounded in Paris train station knife attack
- Headline 15: Manhunt for London attack suspect continues
- Headline 16: Pakistan's king of comebacks looks set to win again
- Headline 17: Who is Bushra Bibi, the faith healer wife of Imran Khan?
- No headlines found

Dependencies

For the successful execution of the web scraping and automation script in this project, several external libraries and modules are utilized. Below is a list of dependencies along with their roles in this project are:

1. Requests: - Purpose: Used for making HTTP requests to the bbc news channel website, enabling the retrieval of HTML content for subsequent parsing. - **Installation:** `pip install requests`

2. Beautiful Soup: - Purpose: Employs HTML parsing to extract specific data elements from the retrieved HTML content. - Installation: `pip install beautifulsoup4`

3. Pandas: - Purpose: Facilitates the organization, manipulation, and storage of the extracted data in a structured format (DataFrame). - Installation: `pip install pandas`

4. APScheduler:- Purpose: Provides scheduling functionality for automating the execution of the web scraping script at predefined intervals. - Installation: `pip install apscheduler`

5. CSV (Built-in Python Module): - Purpose: Used to read from and write to CSV files, enabling the storage of the processed data. - No additional installation required as it is part of the Python standard library.

6. Schedule: - Purpose: An alternative to `APScheduler`, used for scheduling tasks in a straightforward manner. - Installation: `pip install schedule`