PROBLEM STATEMENT : WEB SCRAPING CHALLENGE

PRESENTED BY
Team React

Agenda

1. Problem Statement
2. Methodology
3. Code Structure & Website Structure
4. **Data Extracted**
5. **Challenges Faced**
6. **Conclusion**

Problem Statement

Extracting valuable data from web pages.

Develop a program capable of extracting specific information from web pages.

Back to Agenda

Proposed Solutions

Back to Agenda

Solution # 1

Employ Selenium, a browser automation tool, to mimic human interactions and extract data.

Solution # 2

Utilize Python libraries like Beautiful Soup and Requests to parse HTML and extract desired information

We will be implementing **Solution #1**

Code Structure & Website Structure

1. **Importing Modules**

```
In [71]: from selenium import webdriver
         import time
         import re
         import pandas as pd
         from selenium.webdriver.common.by import By
```

2. **Website Structure**

- **Interacts with "blog-items" containing "blog-item" elements.**
- **Each "blog-item" comprises "img" and "content" divisions.**
- **Extraction: Image link, blog title, publication date, and likes count.**

```
In [ ]: """
        -----STRUCTURE OF CONTENT IN WEBSITE-----

        driver-> blog-items(div) -> list of (blog-item[div])
        blog-item -> img(div),content(div)
        img -> a(tag) -> data-bg(attribute) -----> IMG_LINK--------(1)
        content -> h6(tag) -> a(tag) ------ BLOG TITLE(text)---------(2)
        content -> blog-detail(div) -> bd-item[0](div) -> span(tag) -----> BLOG DATE(text)-------(3)
        content -> zilla-likes(tag) -> span(tag) -----> Likes Count-------(4)

        """
```

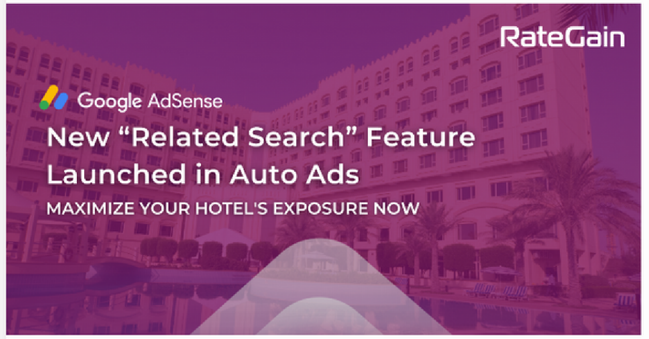
Code Structure

1. **Scraping Logic:**
    - Defined **scrape_blog_data()** function to execute scraping.
    - Utilized Chrome WebDriver to access the target URL **(https://rategain.com/blog/)**.
    - Employed a **while** loop to navigate through multiple pages.
    - Extracted information using **find_elements** and **find_element** methods of WebDriver.
2. **Data Processing:**
    - Retrieved data elements like **image link, title, date, and likes count** from specific HTML elements.
    - Organized extracted data into a **structured format (dictionary)** for storage.
3. **Data Export:**
    - Compiled scraped data into a **Pandas DataFrame**.
    - Exported the data to an Excel file named **"Scraped data.xlsx"**.
4. **Automation and Error Handling:**
    - Incorporated error handling for elements not found or scrolling issues.
    - Automated page navigation using the **"next"** button in pagination.

Data Extracted

Here is some data extracted form the website (on 21-11-23)

1	Maximize Your Hotel’s Exposure with Google AdSense’s New "Related Search" Feature in Auto Ads	October 20, 2023	https://rategaincom.wpengine	19
2	Beyond Reach & Frequency: Hotels' New Era with Facebook's 'Reservation' Buying Type	October 12, 2023	https://rategaincom.wpengine	5
3	Managing Overbookings and Cancellations with Hotel Booking Engines	October 5, 2023	https://rategaincom.wpengine	3
4	Global Distribution System (GDS) vs. Channel Manager: Which is Right for Your Hotel	October 1, 2023	https://rategaincom.wpengine	4
5	Jingle All the Way: Europe Christmas Travel Trends	September 28, 2023	https://rategaincom.wpengine	2
6	Why Bing Hotel Ads Should Be in Your Marketing Mix	September 28, 2023	https://rategaincom.wpengine	3



Maximize Your Hotel’s Exposure with Google AdSense’s New "Related Search" Feature in Auto Ads

OCTOBER 20, 2023 BLOG

Attention, hotel marketing professionals! Google AdSense has just rolled out a game-changing feature that could redefine the way you engage with potential guests: the "Related Search" within Auto Ads. What's New Google has introduced a new feature within AdSense Auto ads called "Related Search for Auto Ads." This feature allows publishers to display search terms related to the content on the pages their users are viewing. When a user chooses one of these suggested se...

Read More

19

book-reservation-buying-type/

Beyond Reach & Frequency: Hotels' New Era with Facebook's 'Reservation' Buying Type

OCTOBER 12, 2023 BLOG

Hey there, hoteliers and hospitality heroes! Gone are the days when Facebook's "Reach and Frequency" was the go-to for your ad campaigns. Say a big hello to "Reservation" - the new kid on the block, and oh boy, does it have perks for your hotel business! Facebook Ad Buying Types Facebook offers businesses ways to purchase ad space, known as "buying types." Diving deeper into the bustling world of Facebook ad strategies, it's crucial to understand the landscape before...

Challanges

- Observing and Adapting to Complex HTML Structures.
- Scraping paginated content or pages with infinite scrolling requires additional logic to navigate through multiple pages.
- Some websites use dynamic loading techniques (e.g., AJAX) to load content after the initial page load.
- Websites frequently update their layout, classes, or IDs, causing previously working scraping scripts to break.

Thank You!