Title: Web Scraping Project- Hacker News

Name : Muskan Ejaz

ID: su91-bietm-f23-024

Introduction:

This lab report documents the web scraping project on Hacker News, a social news website focusing on technology, computer science, and entrepreneurship. The goal was to extract valuable data from the website using Python scripts and BeautifulSoup library.

Methods:

1. Import necessary libraries:

The project started by importing the required libraries, BeautifulSoup and requests. BeautifulSoup is used for parsing HTML content, while requests is used for sending HTTP requests.

2. Specify the user-agent header:

A user-agent header was specified to mimic a browser request. This is done to avoid anti-scraping measures that may be in place on the website.3. Define the URL of the webpage to scrape:

The URL of the webpage to scrape was defined as '(link unavailable)'.

4. Send a GET request to the URL with headers:

A GET request was sent to the URL with the specified headers using the requests library.

5. Check if the request was successful:

The status code of the response was checked to ensure that the request was successful. A status code of 200 indicates a successful request.

6. Parse the content using BeautifulSoup:

The HTML content of the webpage was parsed using BeautifulSoup.

7. Open a file for writing the table rows:

A file was opened in write mode to save the scraped data.

8. Find and write all 'tr' elements (table rows):

All 'tr' elements (table rows) were found and written to the file using a loop.

Results:

The project successfully scraped the table rows from Hacker News and saved them to an HTML file named 'hn_table_rows.html'.

Discussion:

The project demonstrates the ability to scrape data from a website using Python and BeautifulSoup. The specified user-agent header helped to mimic a browser request and avoid anti-scraping measures.

Conclusion:

This project showcases the power of web scraping in extracting valuable data from websites. The skills learned in this project can be applied to various applications, such as data analysis, market research, and social media monitoring.

Python Code:

```
from bs4 import BeautifulSoup

import requests


headers = {

'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_2) AppleWebKit/601.3.9 (KHTML, like Gecko) Version/9.0.2 Safari/601.3.9'

}


url = 'https://news.ycombinator.com/'


response = requests.get(url, headers=headers)


output_file = 'hn_table_rows.html'


if response.status_code == 200:

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

with open(output_file, 'w', encoding='utf-8') as file:

for item in soup.find_all('tr'):

try:
```

file.write(str(item) + '\n')

except Exception as e:

print(f"Error: {e}")

else:

print(f"Failed to retrieve page: {response.status_code}")

print(f"Scraped table rows saved to {output_file}")

Hacker News new | past | comments | ask | show | jobs | submit

Hacker News new | past | comments | ask | show | jobs | submit login

1. A reawakening of systems programming meetups (eatonphil.com)
   32 points by paulgb 1 hour ago | hide | 5 comments
2. Execute JavaScript in a WebAssembly QuickJS Sandbox (github.com/sebastianwessel)
   57 points by sebastianwessel 3 hours ago | hide | 16 comments
3. Servant Leadership Theory (pon.harvard.edu)
   13 points by Bluestein 39 minutes ago | hide | 7 comments
4. Malloc broke Serenity's JPGLoader, or: how to win the lottery (2021) (sin-ack.github.io)
   141 points by fanf2 6 hours ago | hide | 101 comments
5. The staggering science and art behind Wimbledon's legendary grass courts (go.com)
   6 points by hbcondo714 1 hour ago | hide | 5 comments
6. Numeronymize (leancrew.com)
   7 points by surprisetalk 1 hour ago | hide | 1 comment
7. How the 18th-Century French Media Stoked a Werewolf Panic (openculture.com)
   26 points by PaulHoule 4 hours ago | hide | 2 comments
8. Show HN: BeaconDB – An Alternative to Mozilla Location Services (beacondb.net)
   142 points by joelkoen 8 hours ago | hide | 41 comments