Assignment for Stir Tech Internship

Introduction

This document describes how I approached and completed the task of web scraping Twitter's trending topics with Selenium and ProxyMesh, storing the data in MongoDB, and displaying the results on a simple web page.

Task Overview

1. Write a Selenium script to read the Twitter home page and fetch the top 5 trending topics under the "What's Happening" section.
2. Use a valid Twitter account to log in, since it is required to see the trending topics.
3. Use ProxyMesh to ensure each new request originates from a different IP address.
4. Create a unique ID for each Selenium script run, and store the results in a MongoDB database (unique ID, trend1–5, date/time, IP address).
5. Provide a simple HTML page with a button that triggers the script, and then shows the results along with a JSON extract from MongoDB.

Tools and Technologies

- **Selenium** for automating the Twitter login and retrieving trending topics.
- **ProxyMesh** to change the IP address for each request.
- **MongoDB** for storing the fetched trend data.
- **Node.js and Express** for building a simple server to run and display results.
- **Axios** for fetching the external IP address.
- **EJS** (Embedded JavaScript) templates for rendering the data on a webpage.

Implementation Steps

1. **Set Up Environment and Dependencies**
    - Installed Node.js, MongoDB, and Selenium packages (including any required WebDriver like ChromeDriver).
    - Created a .env file for Twitter credentials and ProxyMesh settings.
2. **Selenium Script**
    - Logged into Twitter using credentials from the .env file.
    - Navigated to the homepage and scraped the top 5 trending topics under the "What's Happening" section.
    - Ensured that each new request used a different IP address by integrating ProxyMesh credentials into the Selenium session.
3. **Unique Trends and IP Address**
    - Filtered out duplicate trends so only 5 unique items remained.

- Used Axios to fetch the external IP address from an API (e.g., https://api.ipify.org).
- Stored this IP address, along with the trends and a unique ID (using the uuid library), in the MongoDB database.

4. **Express Server and Routes**
    - Set up an Express server with two main routes:
        - **GET "/"**: Renders a simple HTML page (index.ejs).
        - **POST "/run-script"**: Executes the Selenium script, fetches the trending topics, saves them to MongoDB, and responds with a JSON containing the unique ID, trends, IP address, and date/time.
5. **Displaying Results**
    - On the HTML page, a button issues a POST request to /run-script.
    - When the data is returned, it displays the trending topics, the IP address, and a JSON extract of the MongoDB record.

MongoDB Schema

```
const trendSchema = new mongoose.Schema({
  uniqueId: String,
  trends: [String],
  dateTime: Date,
  ipAddress: String
});
```

The schema stores a unique ID, the list of trends, the scrape's date/time, and the IP address from ProxyMesh or the external IP service.

Workflow Summary

1. **Click Button** on the webpage →
2. **POST "/run-script"** →
3. **Selenium** logs into Twitter→
4. **Scrapes** top 5 trending topics →
5. **ProxyMesh** changes IP →
6. **Axios** fetches external IP →
7. **Data** stored in MongoDB →
8. **JSON** response returned to client.

Conclusion

This setup efficiently handles web scraping from Twitter behind a login, rotates IP addresses using ProxyMesh, stores the trends in MongoDB, and serves the results on a simple webpage. It ensures each run has a unique ID, captures the date and time, and logs the actual IP address in use.

Final Output

Below is an example screenshot or snippet that you can include showing the data returned by the /run-script route and stored in MongoDB.

Twitter Trending Topics

Fetch Trends

These are the most happening topics as on 26/12/2024, 2:06:47 am

- #AtalBihariVajpayee
- #मनुस्मृति_अमरग्रंथ_है
- #AtalJanmShatabdi
- #Retro
- #मनुस्मृति_भारत_का_कलंक_है

The IP address used for this query was 124.123.182.58.

Here's a JSON extract of this record from the MongoDB:


```
{
  "uniqueId": "481f1ae5-d7a5-4379-9668-192ff772f7eb",
  "trends": [
    "#AtalBihariVajpayee",
    "#मनुस्मृति_अमरग्रंथ_है",
    "#AtalJanmShatabdi",
    "#Retro",
    "#मनुस्मृति_भारत_का_कलंक_है"
  ],
  "dateTime": "2024-12-25T20:36:47.259Z",
  "ipAddress": "124.123.182.58",
  "_id": "676c6cdf1fa41af2b3cb22ac",
  "__v": 0
}
```


Click here to run the query again