Name: Manas Tiwari

Section: A

Roll No.: 2023UG1031

Topic: Live News Scrapper Web-Application

Institute: Indian Institute of Information Technology Ranchi

Submitted To: Dr. Shivang Tripathi

Live News Scrapper Web Application

About the Project

The **News Scraper Web Application** is a project designed to scrape and present the latest news headlines along with their associated images and metadata. Users can specify a category (e.g., international news, business, innovation, earth and culture) to fetch relevant news.

The application employs a Python-based backend for web scraping, JavaScript to send request, get the response (data) and handle possible errors and a minimalistic frontend using HTML and CSS for data representation in an appealing way.

This project bridges web scraping and interactive web development, allowing users to retrieve dynamic information from a third-party news source (here www.bbc.com) and display it in a visually appealing way.

Key Features

- Scrapes news headlines, associated images, and additional metadata from a news website (e.g., BBC News).
- Provides the ability to filter news by category such as international news, business, innovation, earth and culture.
- Presents the scraped data through an interactive web interface.

Technologies Used

Backend

- **Python**: Python is a high-level, versatile programming language known for its simplicity and readability. Due to its extensive standard library and support for third-party libraries it is best for web scraping and data processing.
- **Flask**: Flask is a lightweight and flexible web framework for Python. It is designed to build web applications and APIs with minimal boilerplate. It also supports URL routing and request handling.
- **BeautifulSoup**: It is a Python library for web scraping, specifically for parsing HTML and XML documents and extracting relevant data from web pages.
- **Requests**: It is a popular Python library for making HTTP requests to fetch data from a webpage in a simple and human-readable way.

- **Flask-CORS**: A browser security feature that restricts web pages from making requests to a domain different from the one that served the web page, Enables cross-origin requests, allowing the frontend and backend to communicate seamlessly.

  For example, if a frontend hosted on http://example.com wants to fetch data from a backend API at http://api.example.com, CORS needs to be explicitly configured to allow such requests.

Frontend

- **HTML**: It is the standard markup language used to define the structure of a web page. It acts as the skeleton of a webpage by organizing and structuring content.
- **CSS**: Stylesheet language used to control the presentation, for styling the web interface and layout of HTML elements, such as colors, fonts, padding, and spacing.
- **JavaScript**: Versatile programming language used to add interactivity and dynamic behavior to web pages. Also used to interact with the Flask backend and dynamically display news data.

Workflow

1. **User Input**: The user specifies a news category through the select option in the frontend.
2. **Backend Processing**: As the user clicks the “Get News” button, script.js (by Axios) send the request. The Flask backend receives the request and uses the ‘Requests’ and ‘BeautifulSoup’ libraries to fetch the website and scrape relevant news from the specified website an selected topic.
3. **Data Filtering and Saving**: After the data is received, it is filtered to remove invalid or placeholder images. If no data is received or an error occurs, it is also handled and displayed.
4. **Frontend Display**: The data is sent to the frontend and displayed in a clean, user-friendly interface.

How It Works

Backend

- **Flask Endpoint**: The ‘/scrape’ endpoint receives a POST request with a query parameter specifying the news category. Using the parameter, the URL is updated, fetched and scraping is done.
- **Web Scraping**: Using ‘BeautifulSoup, the backend extracts headlines, images, and paragraphs from the targeted news website.
- **Response**: The backend responds with a JSON object containing the filtered news data. The filtered data is converted to JSON using ‘jsonify’.

Frontend

- **User Selects:** As the user selects the news category and presses the ‘Get News’ button, this makes the Event Listener active.
- **Axios**: As the Event Listener becomes active, it sends requests to the Flask backend using the Axios and retrieves the scraped data.
- **Dynamic Rendering**: JavaScript dynamically renders the data into the webpage, providing a seamless user experience.

Challenges and Solutions

CORS Issues

- **Challenge**: Cross-origin requests between the frontend and backend caused errors.
- **Solution**: Implemented Flask-CORS to allow requests from different origins.

Placeholder Images

- **Challenge**: Some scraped images were placeholders, reducing the quality of the output.
- **Solution**: Filtered out entries with missing or placeholder images.

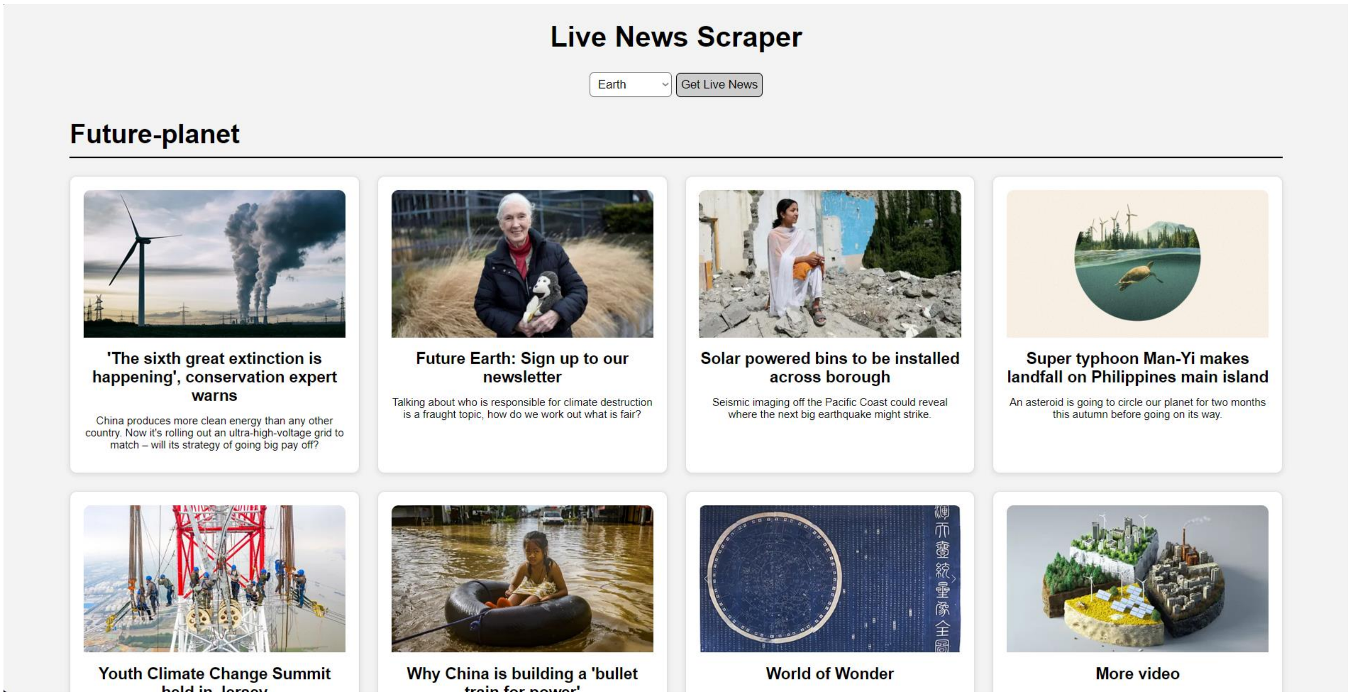
Future Improvements

- Add support for scraping from multiple news sources.
- Implement user authentication and personalized news feeds.
- Enhance the web interface with frameworks like React or Angular.
- Use a database to store and retrieve historical news data.

Conclusion

The News Scraper Web Application demonstrates the power of combining web scraping with web development. It provides users with a practical and interactive tool for staying updated on the latest news in their areas of interest.

Sample Images of the Project: -



Live News Scraper

Earth | Get Live News

Future-planet

'The sixth great extinction is happening', conservation expert warns

China produces more clean energy than any other country. Now it's rolling out an ultra-high-voltage grid to match – will its strategy of going big pay off?

Future Earth: Sign up to our newsletter

Talking about who is responsible for climate destruction is a fraught topic, how do we work out what is fair?

Solar powered bins to be installed across borough

Seismic imaging off the Pacific Coast could reveal where the next big earthquake might strike.

Super typhoon Man-Yi makes landfall on Philippines main island

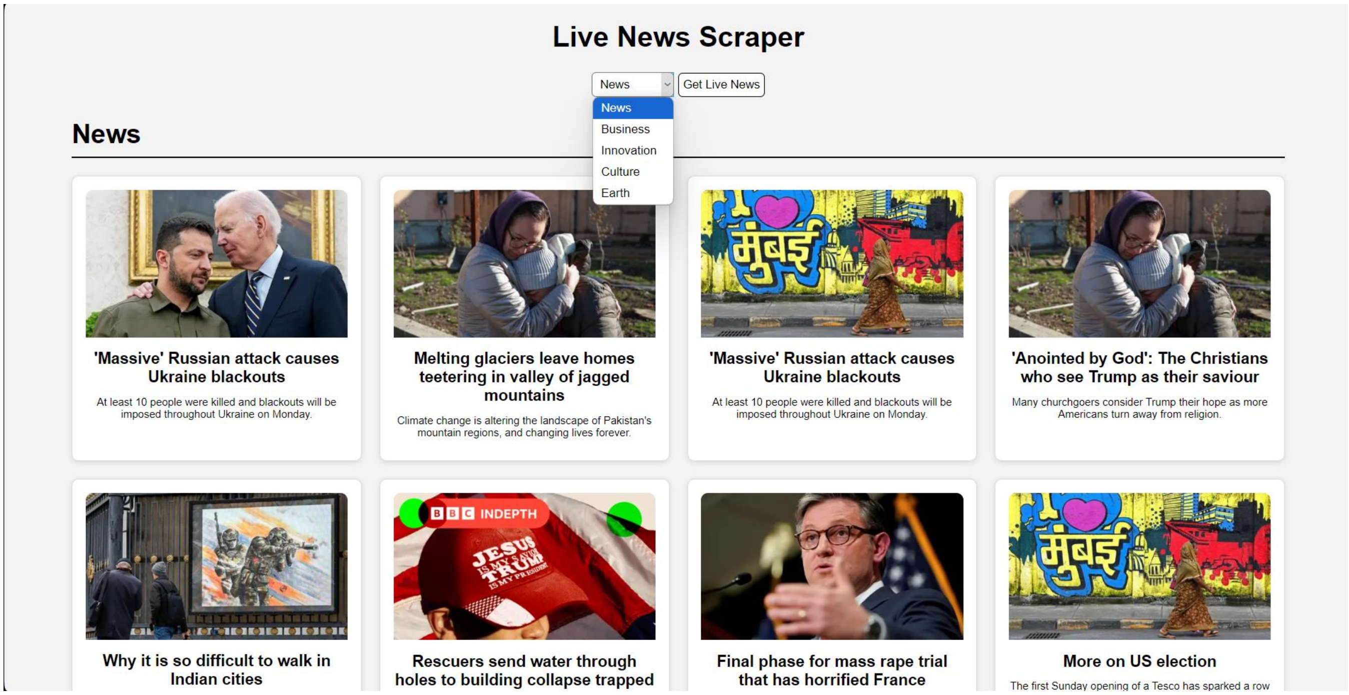
An asteroid is going to circle our planet for two months this autumn before going on its way.

Youth Climate Change Summit held in Jersey

Why China is building a 'bullet train for power'

World of Wonder

More video



Live News Scraper

News | Get Live News

News

'Massive' Russian attack causes Ukraine blackouts

At least 10 people were killed and blackouts will be imposed throughout Ukraine on Monday.

Melting glaciers leave homes teetering in valley of jagged mountains

Climate change is altering the landscape of Pakistan's mountain regions, and changing lives forever.

'Massive' Russian attack causes Ukraine blackouts

At least 10 people were killed and blackouts will be imposed throughout Ukraine on Monday.

'Anointed by God': The Christians who see Trump as their saviour

Many churchgoers consider Trump their hope as more Americans turn away from religion.

Why it is so difficult to walk in Indian cities

Rescuers send water through holes to building collapse trapped

Final phase for mass rape trial that has horrified France

More on US election

The first Sunday opening of a Tesco has sparked a row