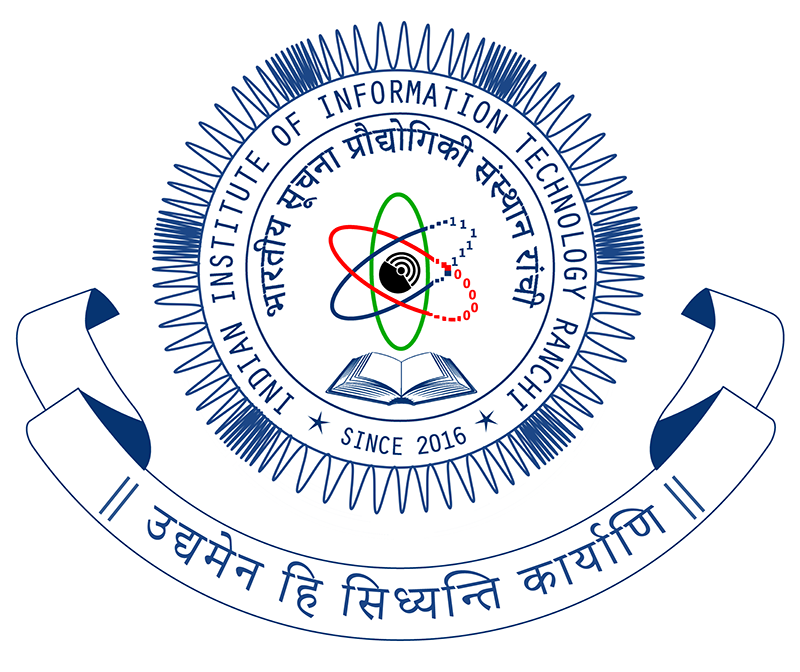
****

**Name:** Manas Tiwari

**Section:** A

**Roll No.:** 2023UG1031

**Subject:** Python Programming (CS-2001)

**Project Topic:** Live News Scrapping Web-Application

**Institute:** Indian Institute of Information Technology Ranchi

**Submitted To:** Dr. Shivang Tripathi

**Live News Scrapping Web Application**

**About the Project**

The **News Scraping 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 and clicks on the ‘Get News’ button to get news, 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 headline or metadata 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 script.js, which then iterate through the data and display it 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 Prospects**

* Add support for scraping from multiple news sources to increase the efficiency of getting correct and accurate data.
* Introduce a scheduler (e.g., Celery) to allow periodic scraping.
* Implement user authentication and personalized news feeds.
* AI and ML Integration: Automatically summarize long paragraphs or articles for users. Use machine learning models to recommend articles, extract sentiment, keywords, or summarize the scraped content.
* Enhance the web interface with frameworks like React or Angular.
* Introduce caching mechanisms to reduce redundant scraping and improve response time.
* Educational Use: Provide a user-friendly interface for students to practice scraping different websites.

**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: -**

