**Web Scraper for Latest news**

**Abstract**

Our project title is “Web scraper for latest news”. Before discussing the functionalities of our project , let’s first understand what web scraping is.

Web Scraping is the process of collecting structured web data in an automated fashion. It’s also called web data extraction. In general, web data extraction is used by people and businesses who want to make use of the vast amount of publicly available web data to make smarter decisions.

Unlike screen scraping, which only copies pixels displayed on screen, web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.

If you’ve ever copy and pasted information from a website, you’ve performed the same function as any web scraper, only on a microscopic, manual scale. Unlike the mundane, mind-numbing process of manually extracting data, web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet’s seemingly endless frontier.

Web scraping is used in a variety of digital businesses that rely on data harvesting. Legitimate use can be : Search engine bots crawling a site, analyzing its content and then ranking it, price comparison sites deploying bots to auto-fetch prices and product descriptions for allied seller websites, market research companies using scrapers to pull data from forums and social media (e.g., for sentiment analysis).

Web scraping is also used for illegal purposes, including the undercutting of prices and the theft of copyrighted content. An online entity targeted by a scraper can suffer severe financial losses, especially if it’s a business strongly relying on competitive pricing models or deals in content distribution.

The basics of web scraping including working in 2 parts : a web crawler and a web scraper.The web crawler is the horse, and the scraper is the chariot. The crawler leads the scraper, as if by the hand, through the internet, where it extracts the data requested.

A web crawler, which we generally call a “spider,” is an artificial intelligence that browses the internet to index and search for content by following links and exploring, like a person with too much time on their hands. In many projects you first “crawl” the web or one specific website to discover URLs which then you pass on to your scraper.

A web scraper is a specialized tool designed to accurately and quickly extract data from a web page. Web scrapers vary widely in design and complexity, depending on the project. An important part of every scraper is the data locators (or selectors) that are used to find the data that you want to extract from the HTML file - usually xpath, css selectors, regex or a combination of them is applied.

Basically our project is one of the applications of web scraping. We are using web scraping to extract the latest news articles from different websites and show them on our web application.

Here’s a little something about our web application:

* Our web application has no sign in restrictions. It’s open for all.
* Anyone who visits our web application has an option to choose their own language in which they want their news article to be displayed. By default only english articles will be shown on our homepage.
* Further the system user has an option to categorically sort the articles and read articles according to that particular category.
* Also the user can search the articles by entering any particular keyword and the articles including that keyword will be displayed.
* If the user is kind of busy and just wants to get the highlights of the past 12 hours, then he/she also has access to that. Past 12 hours news headlines will be displayed to the user and the user can click the ‘read more’ link present below every article if he/she wishes to read the whole article in detail.
* Our news feed will get updated every 10 mins.
* Our web application also shows the weather forecast and keeps it updated.