WebScrapeAI Documentation

WebScrapeAI offers a streamlined, efficient approach to web scraping, allowing users to easily extract data from websites with minimal setup. By providing detailed inputs and selecting your preferred output format, you can tailor the scraping process to your specific needs. Our platform is designed to simplify and enhance your data scraping experience using advanced AI technology. The following documentation outlines the process and capabilities of all the plans.

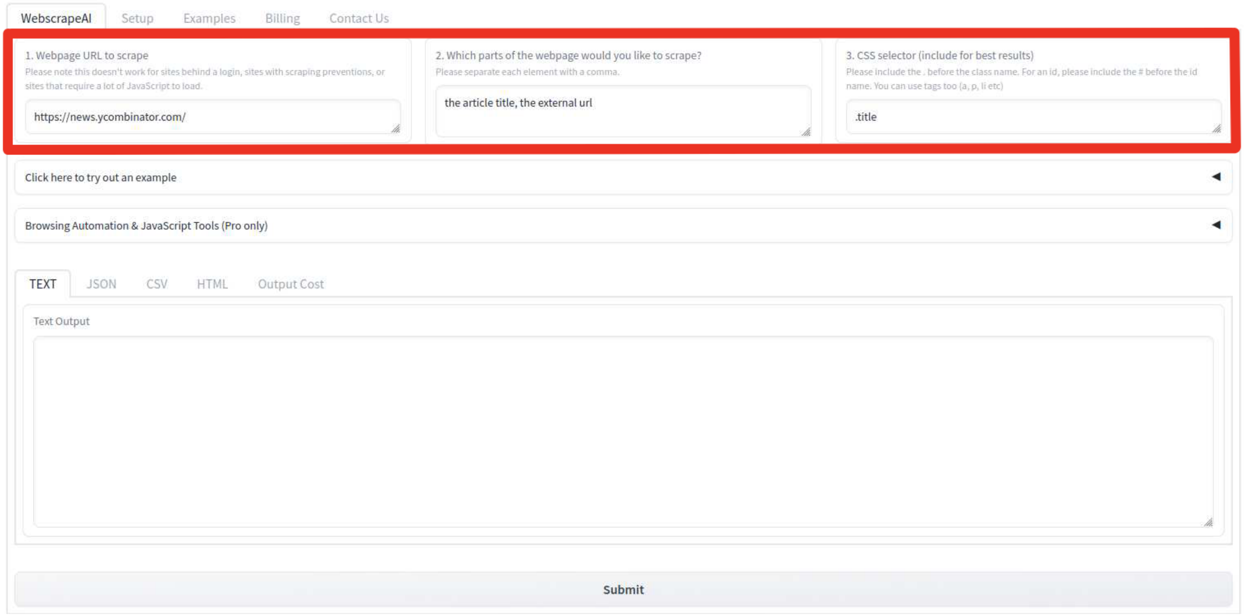
Plans Overview:

1. WebScrapeAI

The **WebScrapeAI** plan is crafted for users who need to extract data from a single URL at a time. This plan is perfect for straightforward, efficient web scraping tasks, providing a user-friendly interface and output flexibility.

To initiate a scraping task, you will need to provide the following information:

1. **URL of the Website:** The exact web address of the site you wish to scrape.
2. **Data Requirements:** Specify the data you want to extract, listed in a comma-separated format. This could include elements like product names, prices, descriptions, etc.
3. **CSS Selectors (Optional):** For enhanced accuracy, you can provide CSS selectors corresponding to the specific elements you're targeting. While optional, using CSS selectors is highly recommended for obtaining precise results.



1. Webpage URL to scrape

Please note this doesn't work for sites behind a login, sites with scraping preventions, or sites that require a lot of JavaScript to load.

https://news.ycombinator.com/

2. Which parts of the webpage would you like to scrape?

Please separate each element with a comma.

the article title, the external url

3. CSS selector (include for best results)

Please include the . before the class name. For an id, please include the # before the id name. You can use tags too (a, p, li etc)

.title

Click here to try out an example ◀
Browsing Automation & JavaScript Tools (Pro only) ◀
TEXT
JSON
CSV
HTML
Output Cost
Text Output
Submit

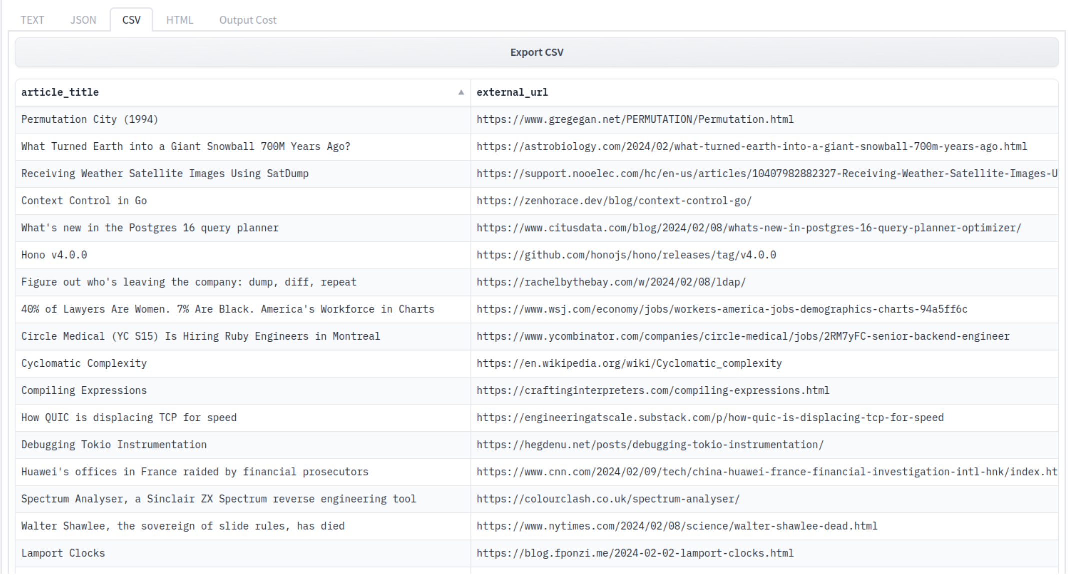
Once you have provided the necessary information, click the **Submit** button to activate our AI model. The AI will then proceed to scrape the specified data from the provided URL.

Output Formats

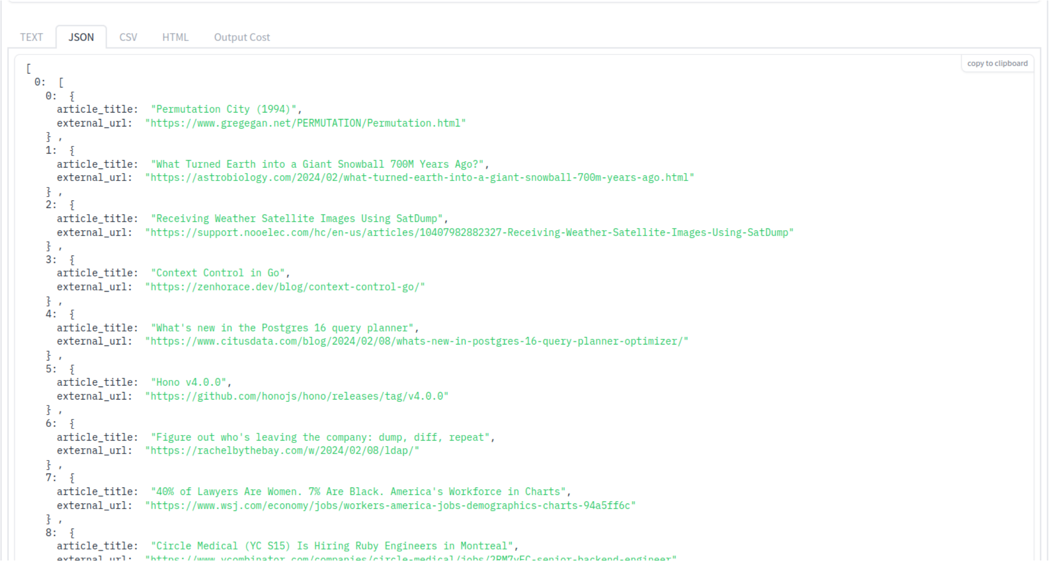
Upon completion of the scraping process, you can retrieve your data in one of the following formats, based on your preference and project requirements:

1. **CSV (Comma-Separated Values):** Ideal for spreadsheet applications and data analysis tools. Can be directly exported into a file, facilitating easy download and storage.
2. **JSON (JavaScript Object Notation):** Best for applications that require data interchange or further processing with scripting languages.
3. **Text:** Simple, unformatted text output for a wide range of uses.
4. **HTML:** Raw HTML format, useful for preserving the structural context of the scraped data.

Here are the snippets of the data extracted in CSV and JSON format:



article_title ▲	external_url
Permutation City (1994)	https://www.gregegan.net/PERMUTATION/Permutation.html
What Turned Earth into a Giant Snowball 700M Years Ago?	https://astrobiology.com/2024/02/what-turned-earth-into-a-giant-snowball-700m-years-ago.html
Receiving Weather Satellite Images Using SatDump	https://support.nooelec.com/hc/en-us/articles/10407982882327-Receiving-Weather-Satellite-Images-U
Context Control in Go	https://zenhorace.dev/blog/context-control-go/
What's new in the Postgres 16 query planner	https://www.citusdata.com/blog/2024/02/08/whats-new-in-postgres-16-query-planner-optimizer/
Hono v4.0.0	https://github.com/honojs/hono/releases/tag/v4.0.0
Figure out who's leaving the company: dump, diff, repeat	https://rachelbythebay.com/w/2024/02/08/ldap/
40% of Lawyers Are Women. 7% Are Black. America's Workforce in Charts	https://www.wsj.com/economy/jobs/workers-america-jobs-demographics-charts-94a5ff6c
Circle Medical (YC S15) Is Hiring Ruby Engineers in Montreal	https://www.ycombinator.com/companies/circle-medical/jobs/2RM7yFC-senior-backend-engineer
Cyclomatic Complexity	https://en.wikipedia.org/wiki/Cyclomatic_complexity
Compiling Expressions	https://craftinginterpreters.com/compiling-expressions.html
How QUIC is displacing TCP for speed	https://engineeringatscale.substack.com/p/how-quic-is-displacing-tcp-for-speed
Debugging Tokio Instrumentation	https://hegdenu.net/posts/debugging-tokio-instrumentation/
Huawei's offices in France raided by financial prosecutors	https://www.cnn.com/2024/02/09/tech/china-huawei-france-financial-investigation-intl-hnk/index.ht
Spectrum Analyser, a Sinclair ZX Spectrum reverse engineering tool	https://colourclash.co.uk/spectrum-analyser/
Walter Shawlee, the sovereign of slide rules, has died	https://www.nytimes.com/2024/02/08/science/walter-shawlee-dead.html
Lamport Clocks	https://blog.fponzi.me/2024-02-02-lamport-clocks.html



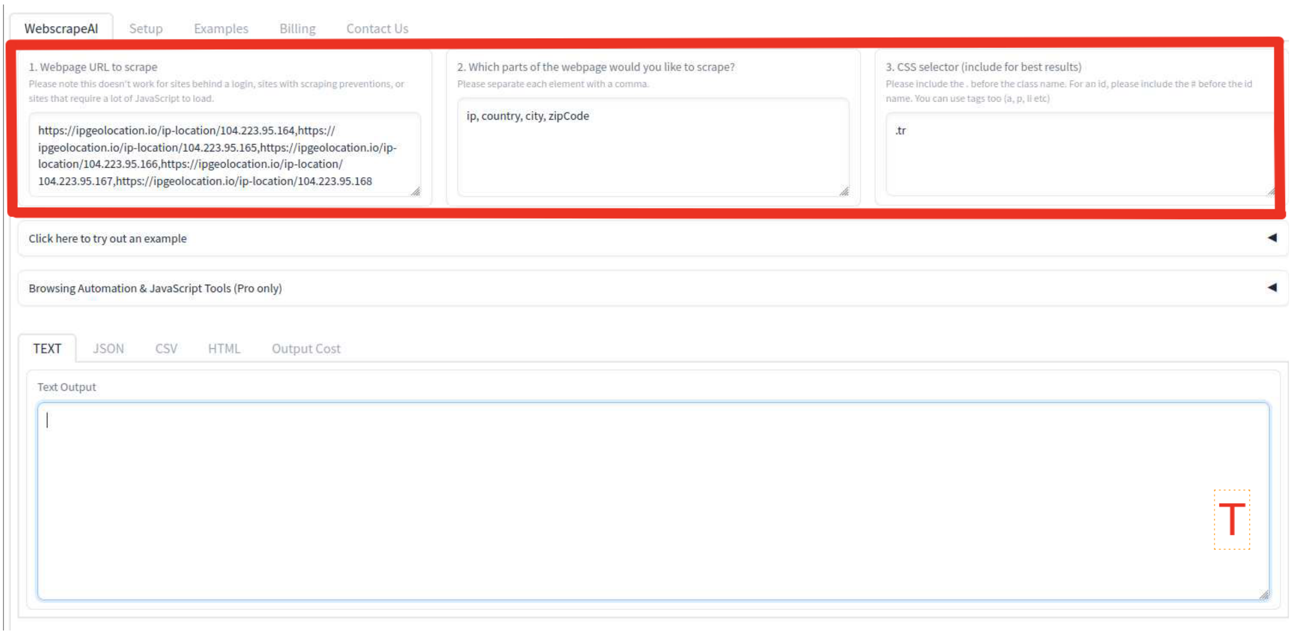
```
[
 0:  [
   0:  {
     article_title:  "Permutation City (1994)",
     external_url:  "https://www.gregegan.net/PERMUTATION/Permutation.html"
   } ,
   1:  {
     article_title:  "What Turned Earth into a Giant Snowball 700M Years Ago?",
     external_url:  "https://astrobiology.com/2024/02/what-turned-earth-into-a-giant-snowball-700m-years-ago.html"
   } ,
   2:  {
     article_title:  "Receiving Weather Satellite Images Using SatDump",
     external_url:  "https://support.nooelec.com/hc/en-us/articles/10407982882327-Receiving-Weather-Satellite-Images-Using-SatDump"
   } ,
   3:  {
     article_title:  "Context Control in Go",
     external_url:  "https://zenhorace.dev/blog/context-control-go/"
   } ,
   4:  {
     article_title:  "What's new in the Postgres 16 query planner",
     external_url:  "https://www.citusdata.com/blog/2024/02/08/whats-new-in-postgres-16-query-planner-optimizer/"
   } ,
   5:  {
     article_title:  "Hono v4.0.0",
     external_url:  "https://github.com/honojs/hono/releases/tag/v4.0.0"
   } ,
   6:  {
     article_title:  "Figure out who's leaving the company: dump, diff, repeat",
     external_url:  "https://rachelbythebay.com/w/2024/02/08/ldap/"
   } ,
   7:  {
     article_title:  "40% of Lawyers Are Women. 7% Are Black. America's Workforce in Charts",
     external_url:  "https://www.wsj.com/economy/jobs/workers-america-jobs-demographics-charts-94a5ff6c"
   } ,
   8:  {
     article_title:  "Circle Medical (YC S15) Is Hiring Ruby Engineers in Montreal",
     external_url:  "https://www.ycombinator.com/companies/circle-medical/jobs/2RM7yFC-senior-backend-engineer"
```

2. WebScrapeAI Bulk

It is an advanced offering from WebScrapeAI designed for users who require data extraction from multiple URLs simultaneously. This plan builds upon the foundation of our basic **WebScrapeAI** plan, introducing the capability to handle bulk scraping tasks with ease and efficiency.

To utilize the bulk scraping functionality, users must provide the following inputs:

1. **URLs of the Websites:** The web addresses to be scraped, listed in a comma-separated format. Ensure that all URLs are valid and accessible for scraping.
2. **Data Requirements:** Specify the data you wish to extract, formatted as a comma-separated list. This should be consistent across all URLs for optimal results.
3. **CSS Selectors (Optional):** Although optional, providing CSS selectors for the specific data elements you're targeting is recommended to achieve the highest accuracy.



1. Webpage URL to scrape

Please note this doesn't work for sites behind a login, sites with scraping preventions, or sites that require a lot of JavaScript to load.

https://ipgeolocation.io/ip-location/104.223.95.164,https://ipgeolocation.io/ip-location/104.223.95.165,https://ipgeolocation.io/ip-location/104.223.95.166,https://ipgeolocation.io/ip-location/104.223.95.167,https://ipgeolocation.io/ip-location/104.223.95.168

2. Which parts of the webpage would you like to scrape?

Please separate each element with a comma.

ip, country, city, zipCode

3. CSS selector (include for best results)

Please include the . before the class name. For an id, please include the # before the id name. You can use tags too (a, p, li etc)

.tr

Click here to try out an example
Browsing Automation & JavaScript Tools (Pro only)
TEXT
JSON
CSV
HTML
Output Cost
Text Output

After inputting the necessary information, click the **Submit** button. Our AI model will then begin the process of scraping the specified data from all provided URLs.

It's crucial that all pages from which data is being extracted contain the same type of information as specified in your data requirements. If a particular field is not found on a webpage, the AI model will return a **null** value for that field to maintain the consistency of the output format.

Output Formats

The **WebScrapeAI Bulk** plan offers the same output formats as the basic plan.

3. WebScrapeAI Pro:

WebScrapeAI Pro is the advanced plan designed for users seeking a comprehensive web scraping solution with additional capabilities. This plan extends the features of **WebScrapeAI** and **WebScrapeAI Bulk** with the integration of proxies, custom headers, and advanced JavaScript tools for a more powerful and tailored scraping experience.

Features

- **Proxies:** Use your own proxies for scraping to manage your IP footprint and bypass geo-restrictions.
- **Headers:** Forward custom headers to target websites for enhanced access control and personalization.
- **JavaScript Tools:** Execute advanced browsing instructions and JavaScript for dynamic websites and complex scraping tasks.

For advanced users who need to interact with web pages that require dynamic interaction, **WebScrapeAI Pro** offers JavaScript Tools. This feature allows users to input browsing instructions that the AI will execute before data collection begins.

How to Use:

1. **Enter JavaScript Instructions:** Provide instructions for page interactions, such as clicks, waits, and scrolls.
2. **Custom Headers:** Include any required headers to navigate the website.
3. **Proxies:** Enter any proxies you would like to use in the specified format.

JavaScript Instructions:

- `{"click": "#button_id"}`: Click on a specified element.
- `{"wait": 1000}`: Pause for a specified duration in milliseconds.
- `{"wait_for": "#element_id"}`: Wait for a specified element to become available.
- `{"scroll_y": 1000}`: Scroll vertically by the specified pixel amount.
- `{"fill": {"#input_id", "value_1"}}`: Fill in a specified input field with a value.
- `{"evaluate": "console.log('action')}`: Execute custom JavaScript code.

Headers

Custom headers can be forwarded to the target website. This is particularly useful for setting request headers like `User-Agent`, `Accept-Language`, or custom headers required by the website. Enter any headers in key-value format that you would like the scraper to use when making requests to the website.

Proxies

WebScrapeAI Pro users have the ability to use their own proxies. This feature is crucial for users who need to manage their scraping operations discreetly or access content from different geographical locations.

Enter the proxy details in the provided format:

```
<protocol>://<username>:<password>@<host>:<port>
```

Make sure to replace `<protocol>`, `<username>`, `<password>`, `<host>`, and `<port>` with your actual proxy details.

After configuring your JavaScript instructions, headers, and proxies, proceed with submitting your request just like in the basic and bulk plans. Click the **Submit** button to start the scraping process with all your specified parameters.

For further assistance you can contact our support team.