

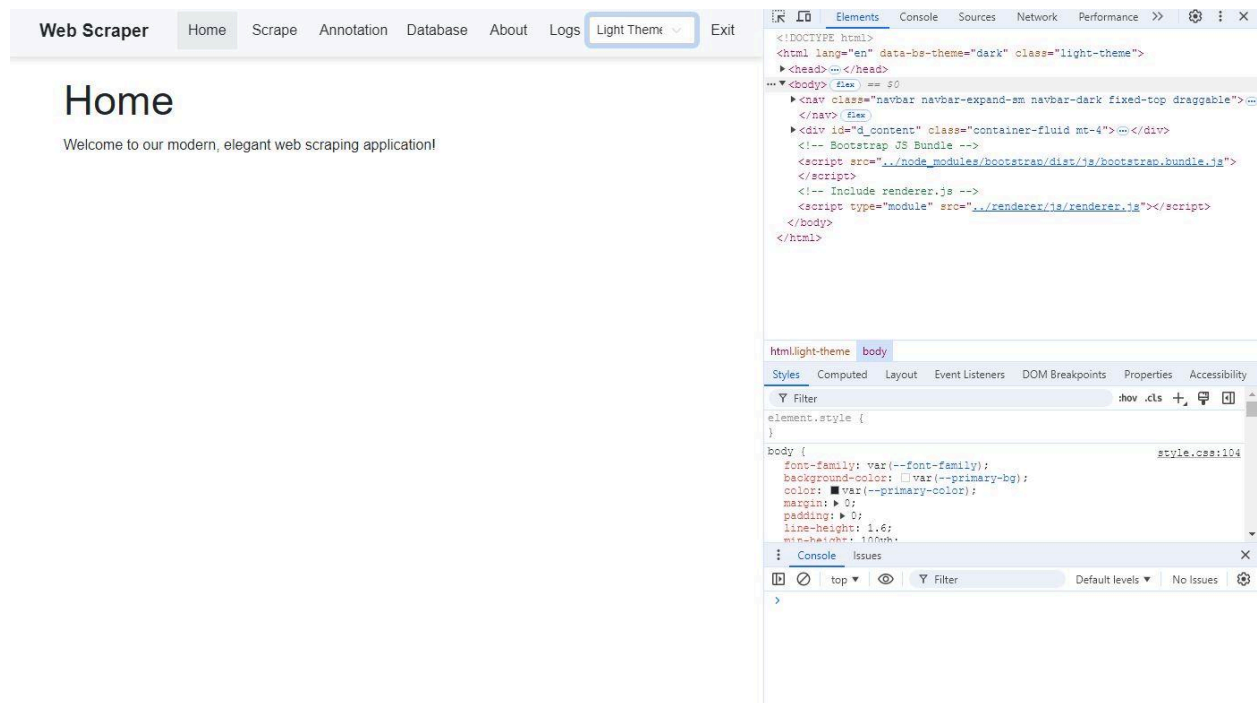
How to Use

J-Initiative Web Scraper

Features covered in this document:

1. Web Scraper
2. Annotation
3. Database
4. Logs

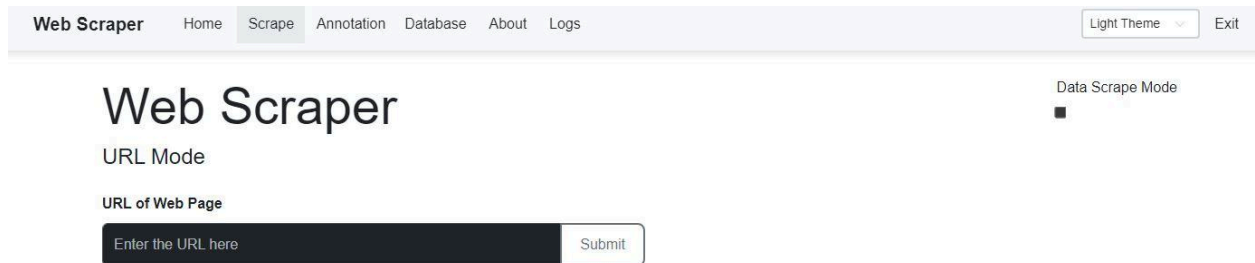
Introduction to the program



Welcome to the J-Initiative web scraper manual! This document is intended to allow anyone to be able to be caught up on our great web-scraping app and how to use it efficiently and effectively! By the end of this document, you should be able to use our top-tier web-scraping app and all of its features, regardless of your initial level of experience!

To begin, please open the web scraping app in order to follow along! Once the application opens you should see something quite similar to the picture provided above. The window on the right—an inspect element window—gives you information about the app and what's going on within it by hovering over and clicking through its different elements. For now though we can simply close it in order to clean up the space. Once you've done this you should be ready to get started. So, let's go!

Utilizing the web scraper feature



The screenshot shows the 'Web Scraper' application interface. At the top is a navigation bar with links: 'Web Scraper', 'Home', 'Scrape', 'Annotation', 'Database', 'About', and 'Logs'. On the right of the navigation bar are 'Light Theme' (with a dropdown arrow) and 'Exit'. Below the navigation bar, the main heading 'Web Scraper' is displayed. Underneath it, 'URL Mode' is indicated. A section labeled 'URL of Web Page' contains a dark input field with the placeholder text 'Enter the URL here' and a 'Submit' button. In the top right corner, there is a 'Data Scrape Mode' label next to a checked checkbox.

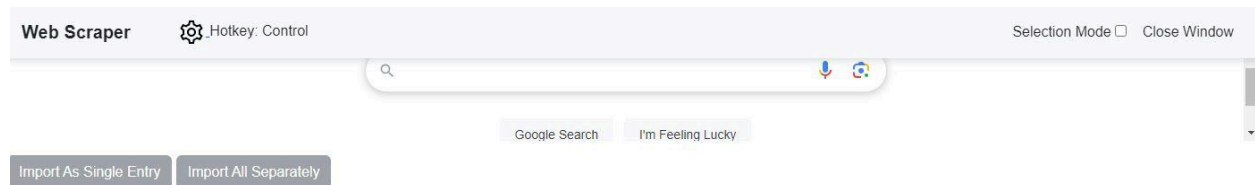
Naturally, the web scraping feature is going to be incredibly important in the J-Initiative web scraper app, so how do you access and use this feature? That's what we are here to go through now.

To begin, please click on "Scrape" at the top of the application in order to move to the Web Scraper section of the application. You should see something quite similar to the picture provided above.

As you can see, by default the application is in "URL Mode", which can be changed by clicking the checkbox below where it says "Data Scrape Mode" seen in the top right of the page. We will explore this mode more later.

First, we see that in URL Mode, you have only one real area to enter information, making the process quite simple. To start off with, decide on a website to scrape information from.

Utilizing the web scraper feature (cont.)



Next, copy the URL of this website and paste it into the textbox that says “Enter the URL here” and press the button that says “Submit”.

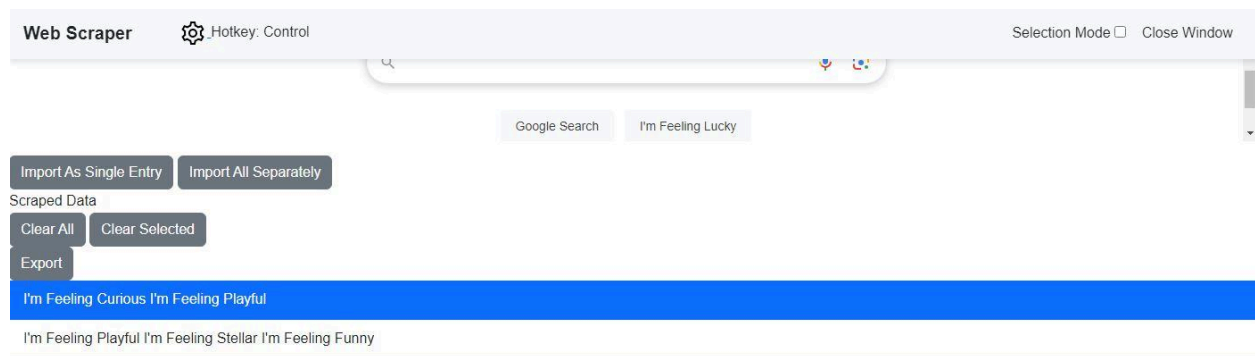
Next, a new page should open that, after loading, should contain your website! An example of what this might look like is seen above, where we have opened up <https://www.google.com> using our web scraper!

To take data from this page, identify what the “Hotkey” is. This should be indicated in the top left corner of the application. As you can see in the screenshot, our hotkey is the control key. To change this to another key, simply click the settings gear button next to the Hotkey text and then click the button you would like to make the new Hotkey. Click escape to cancel this if you enter into it accidentally.

The Hotkey is what you will hold down to select elements on the website to import.

If you would like to avoid using the Hotkey, simply click the checkbox next to where it says “Selection Mode” in the top right corner, which will simply treat it as if you are always holding down the Hotkey, meaning each click selects data. Otherwise, we will discuss below how to use the Hotkey.

Utilizing the web scraper feature (cont.)

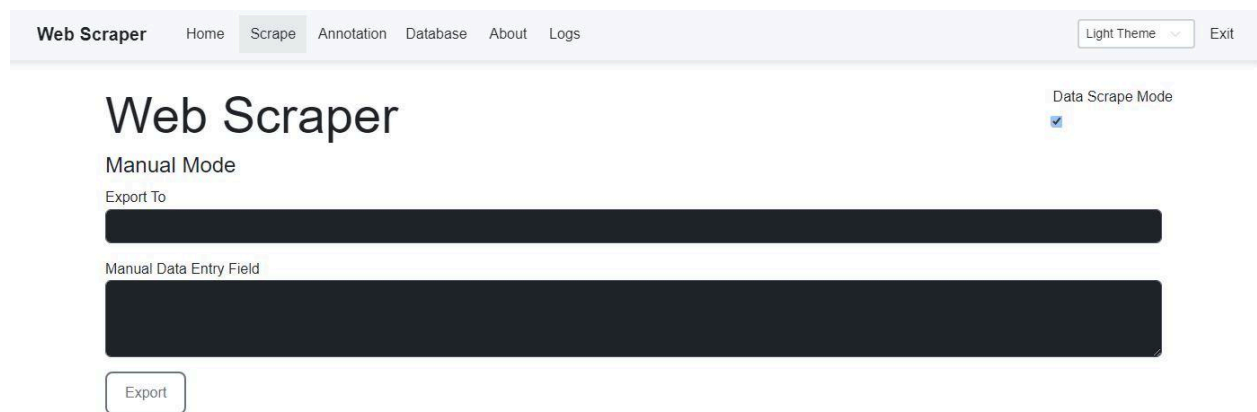


For example, here we have selected a number of the “I’m feeling” prompts that Google gives and selected “Import As Single Entry” in order to keep the selections together (as opposed to “Import All Separately” which separates selections into separate entries). To export this data in order to work on annotating it, simply click the button that says “Export”. This will take you back to the main web scraper page where the “Scraped Data” field will now exist and will be populated with that item which you scraped.

From here, you can choose to export this data to a Label Studio project by selecting your project from the drop down under where it says “Export To”.

Finally, simply click the button that says “Export” and the data will be moved to your Label Studio project.

Manual mode of the web scraper



The screenshot shows the 'Web Scraper' application interface. At the top is a navigation bar with links: 'Web Scraper', 'Home', 'Scrape', 'Annotation', 'Database', 'About', and 'Logs'. On the right of the navigation bar are 'Light Theme' and 'Exit' buttons. Below the navigation bar, the main heading is 'Web Scraper'. To the right of the heading is a 'Data Scrape Mode' checkbox, which is checked. Below the heading is the text 'Manual Mode'. Under 'Manual Mode' is an 'Export To' label followed by a dark input field. Below that is a 'Manual Data Entry Field' label followed by a larger dark input field. At the bottom left is an 'Export' button.

What if you want more control over what you get and where you put it? Well, then you should, instead of using the URL Mode, click on the button the checkbox under where it says “Data Scrape Mode” in order to use manual mode to have greater control over these steps.

Once clicking on the button, you should find yourself looking at something quite similar to the picture provided above.

Manual mode of the web scraper (cont.)

Once you are in this screen, you should notice two major areas for input; those being the: “Project to Export To” box and the “Manual Data Entry Field” box.

These two boxes do exactly what they sound like they do! Specifically, the “Project to Export To” box is where you can select the Label Studio project that you would like to export your scraped data to and the “Manual Data Entry Field” is where you can manually enter the data that you would like to be added to the project.

This area of the app is very helpful as it allows the user much greater control of what is exported.

For example, if you have scraped some data from a website and want to then move it to Label Studio, you can copy this information, add it into the manual data entry field and export it to your specified project.

In this way, you can export your data to Label Studio quite easily and with greater control over what is chosen and where it goes!

Annotation

Web Scraper Home Scrape **Annotation** Database About Logs Light Theme ↘ Exit

Label Studio

URL of Label Studio Project

Next, you may want to annotate data in order to provide connotations to words, in order to better analyze the data for biases in diction (and other such things).

To start off with, click on “Annotations” in the menu on top.

After doing so, you should be on a screen that looks very much like the picture seen above.

Here you can see that there is only one area to enter information, that being a textbox labeled “Enter the URL of the Label Studio Project”.

Annotation (cont.)

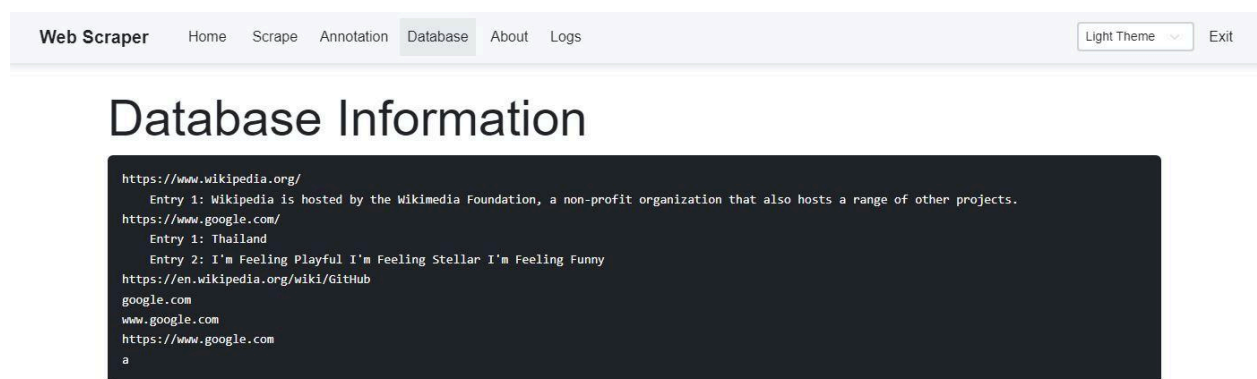
This textbox is where you will enter the URL for your label studio project, at which point you will be able to interact with your data using label studio.

This will allow you to take the data that you just scraped in the web scraping section of the application and actively label it, allowing for you to provide the wording with context that will be vital for any NLP model.

This area of the app is, as such, vital for anyone who wishes to do something with the data they just took.

So make sure you remember this incredibly helpful part of the app when you are using the other sections!

Database



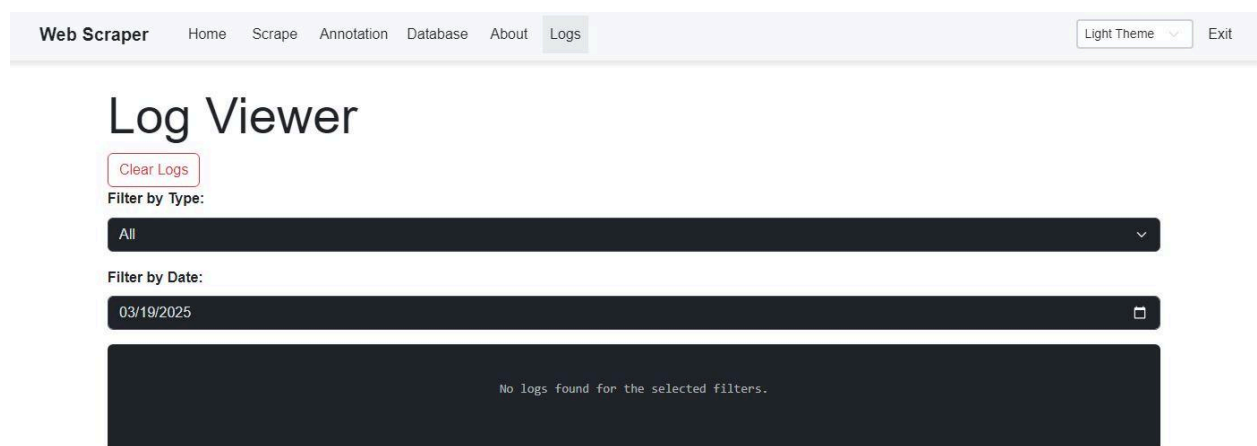
Next, click on “Database” at the top of the screen and you will likely see something like this.

This is the database section of the application, which shows the data that is stored in the database the application manages. When there is data to show it will replace the text that, in the picture above, has random, default information it prints out.

This database stores the data taken from scraped websites, and so this area is a user-friendly way of interacting with that stored data!

Make sure to remember the database for when you want to access saved data that you previously scraped!

Logs



Finally, click on “Logs” at the top in order to see the final area that we want to talk about today.

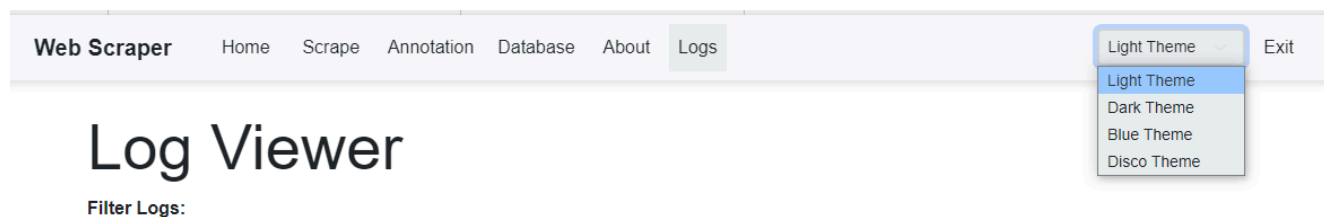
Here, you should see something quite similar to the picture above.

This page is where the application logs the events and the errors that it encounters. So this is where you can look if there are any issues, or to see what exactly is going on with the app!

At the top of the page, where in the picture it says “All”, you can click and select from “All”, “Info”, “Debug”, “Warn”, and “Error” to choose to either show all of the logs or just the logs about the application’s information, debugging logs, warning or errors specifically.

Other

Congratulations! You have learned what each of the major pages do and how to use them! The other two pages (“Home” and “About”) just give you some information about the application and don’t have any major functionality to know about.



One big thing to know that hasn’t been mentioned yet is the theme menu, which can be seen in the picture above.

Here you can choose from a set of different themes, which change the look and feel (but not the functionality) of the program, so feel free to play around with themes if you ever want to change how the program looks to be better tuned to your preferences!

Additionally, know that you may randomly receive a notification on the front page in the scenario that the app experiences serious errors or there is otherwise incredibly important information to relay to you. Notifications should be very rare, but may occur, so make sure not to ignore them if you receive one.

Aside from that, there shouldn't be much else to know. You are now officially an expert on the J-Initiative web scraping application!