

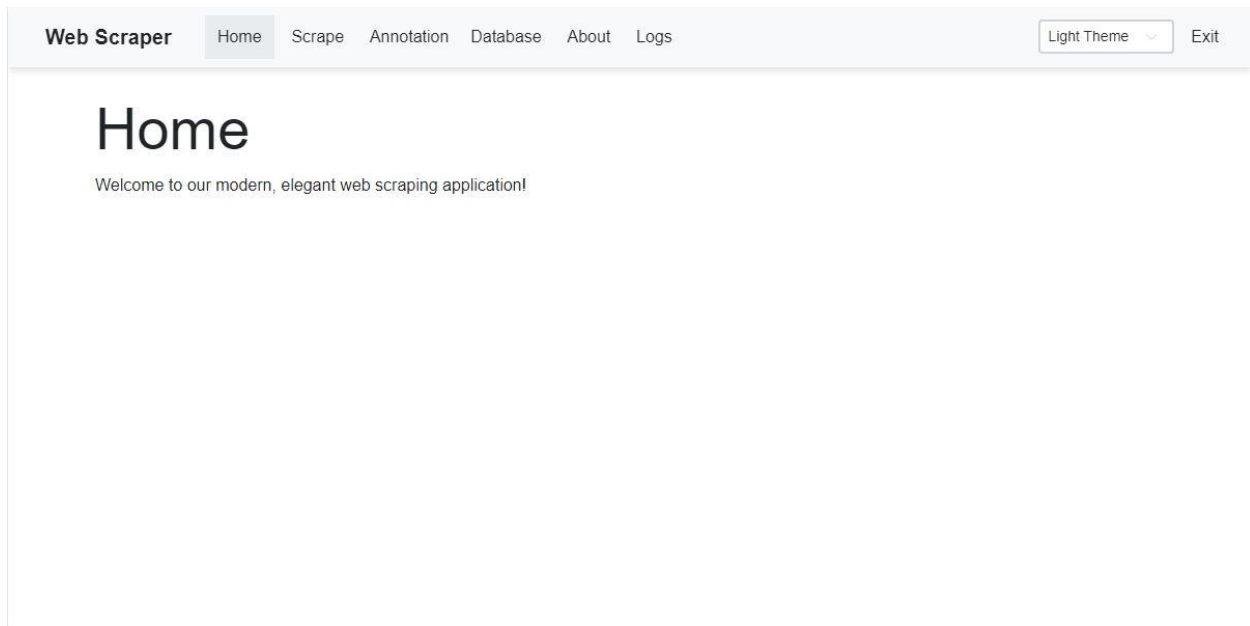
How to Use

J-Initiative Web Scraper

Features covered in this document:

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

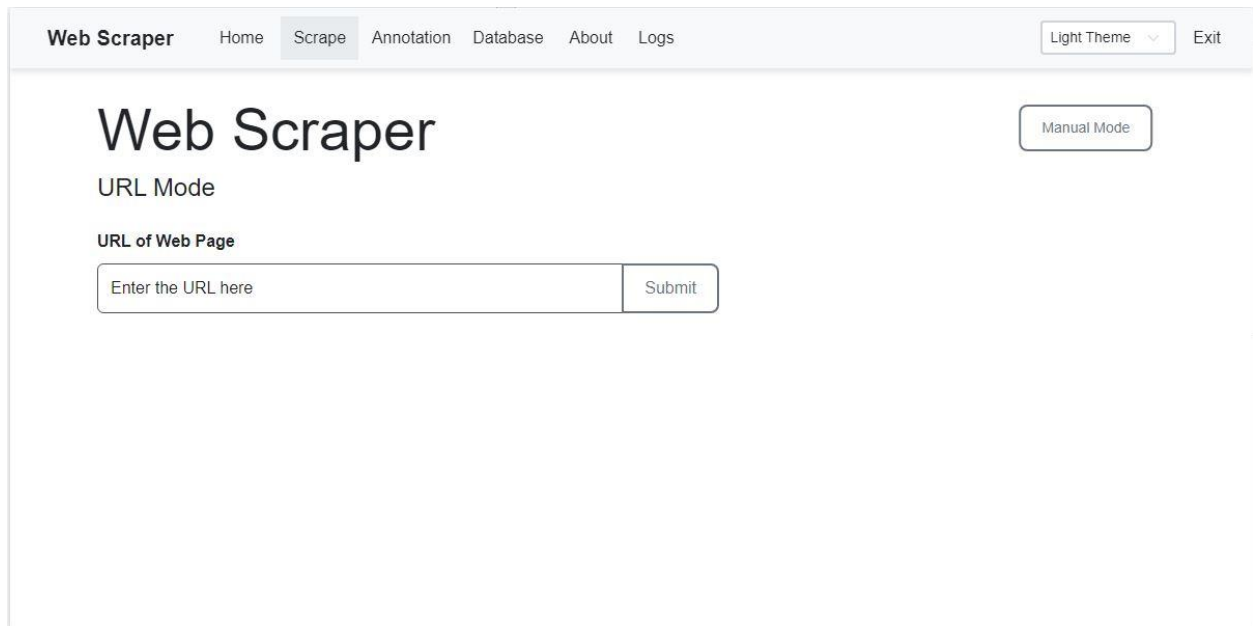
Introduction to the program



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. Once this is the case, 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, there is a navigation bar with the title 'Web Scraper' and links for 'Home', 'Scrape', 'Annotation', 'Database', 'About', and 'Logs'. On the right side of the navigation bar, there is a 'Light Theme' dropdown menu and an 'Exit' button. Below the navigation bar, the main content area has a large heading 'Web Scraper' and a subheading 'URL Mode'. To the right of the 'URL Mode' subheading, there is a button labeled 'Manual Mode'. Below the subheading, there is a section titled 'URL of Web Page' which contains a text input field with the placeholder text 'Enter the URL here' and a 'Submit' button.

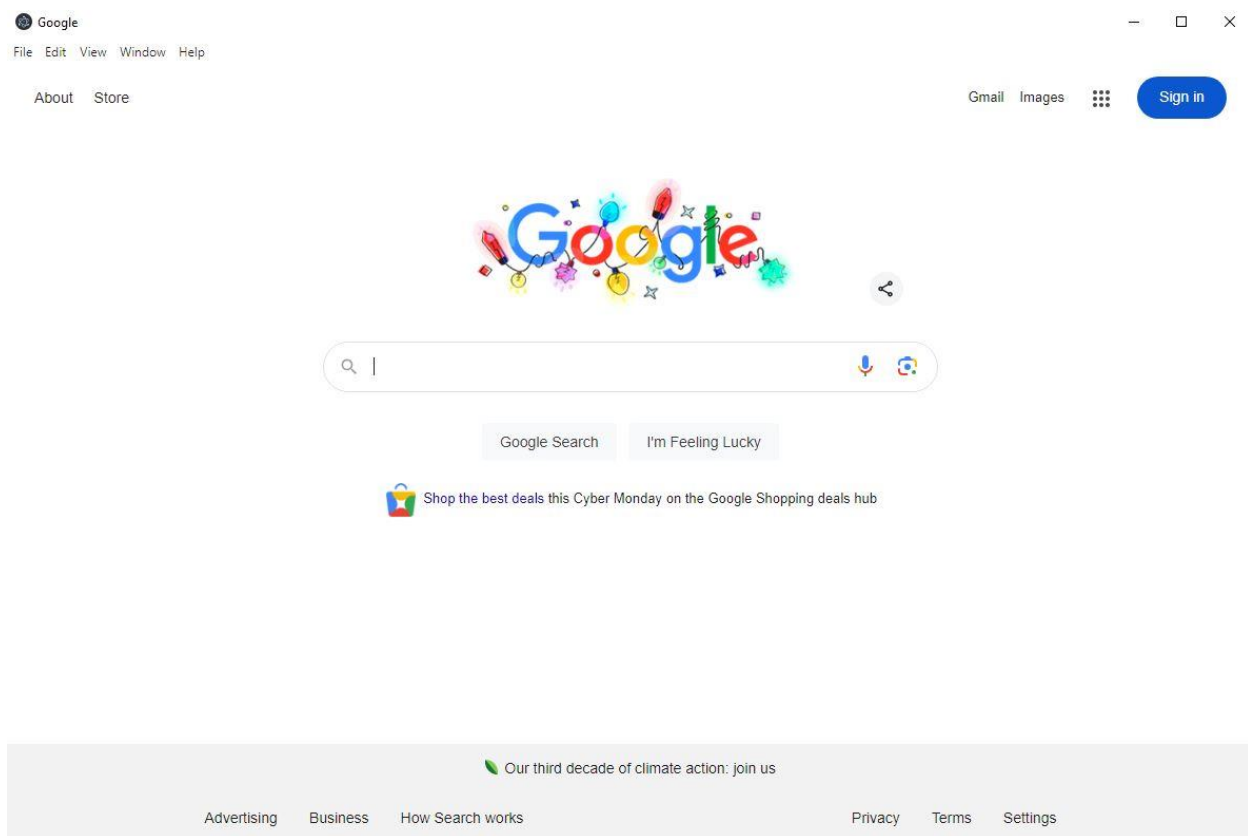
Naturally, the web scraping feature is going to be incredibly important in the J-Initiative web scraper app, so how do you access and use this feature? That's what we are here to go through now.

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 button labeled 'Manual 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.)

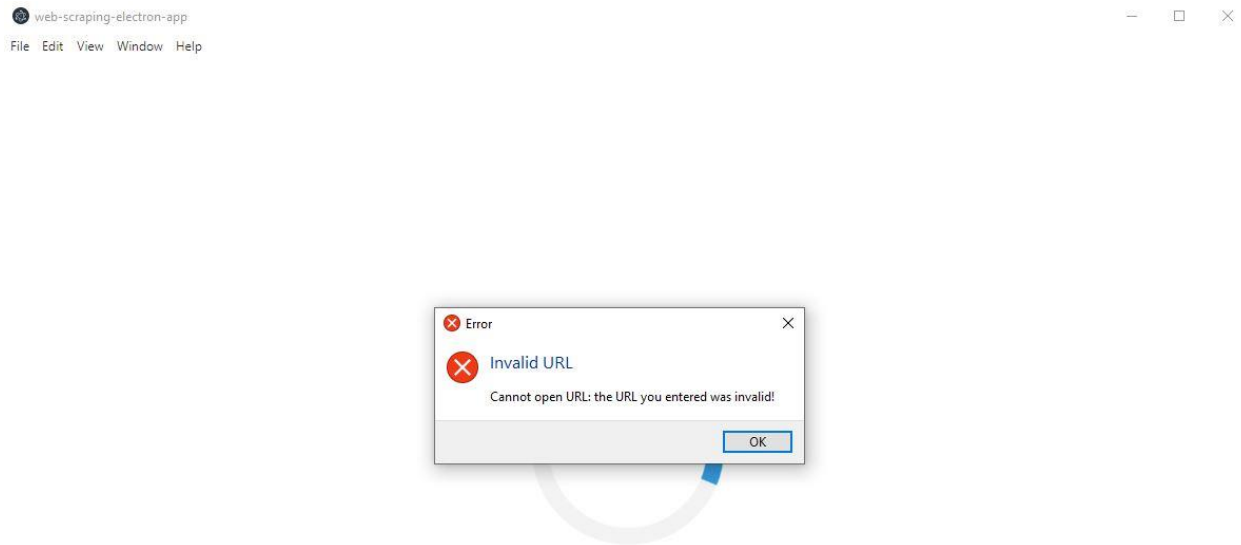


Next, copy the URL of this website and paste it into the textbox that says 'Enter the URL here' and press submit.

Next, a page should open that, after loading, should contain your page! An example of what this might look like is seen above, where we have opened up <https://www.google.com> using our web scraper! In your instance, you should likely see something similar to "Import Selected Data", which you can use to scrape data that you have highlighted and then export it to the label studio project using the "Export" button!

But what if I'm not seeing something like this? What if the page indicates that there is an issue with the URL? What does that mean?

Utilizing the web scraper feature (cont.)



If instead of a website loading, you receive the above message, you may be confused. Instead of opening the page that you wished to open, you have received an error indicating that something went wrong. But what?

This message indicates that the website URL that you entered into the textbox was invalid in some way—that is to say that the program was not able to open it. This is most likely due to a simple type, so make sure to check that your URL is entirely written correctly and is not missing any parts. Once you find the problem, simply fix it and click the 'Submit' button again, and you should see the website in all of its glory!

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' (which is highlighted), 'Annotation', 'Database', 'About', and 'Logs'. On the right side of the navigation bar are buttons for 'Light Theme' (with a dropdown arrow) and 'Exit'. Below the navigation bar, the main heading 'Web Scraper' is displayed on the left, and a 'URL Mode' button is on the right. Under the heading, the text 'Manual Mode' is shown. Below this, there is a label 'Project to Export To' followed by a text input field. Further down is a label 'Manual Data Entry Field' followed by a larger text input field. At the bottom left of the form area is a 'Submit' button.

Once you have successfully opened your page, you may find your formatted and raw data by going back to the main web scraping program (as opposed to the new page that contains the website) and clicking on either 'Formatted Data' for the formatted data from the website or on 'Raw Data' for the raw data from the website.

But from here, 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 that says 'Manual 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.

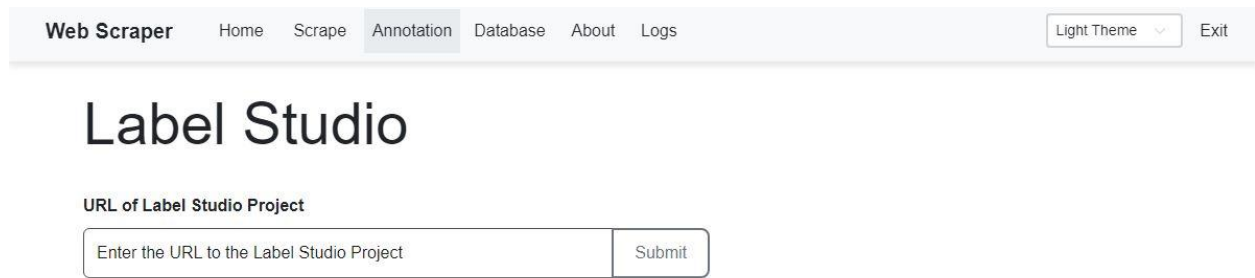
These two boxes do exactly what they sound like they do! Specifically, the project to export to box is where you can select the label studio project that you would like to export your scraped data to and the manual data entry field is where you can manually enter the data that you would like to be added 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!

Annotation



The screenshot shows a web application interface. At the top is a navigation bar with the following items: 'Web Scraper' (highlighted), 'Home', 'Scrape', 'Annotation' (highlighted), 'Database', 'About', and 'Logs'. On the right side of the navigation bar, there is a 'Light Theme' dropdown menu and an 'Exit' link. Below the navigation bar, the main heading 'Label Studio' is displayed. Underneath the heading, the text 'URL of Label Studio Project' is shown. Below this text is a form consisting of a text input field with the placeholder text 'Enter the URL to the Label Studio Project' and a 'Submit' button.

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.)

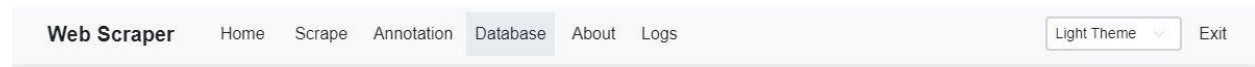
This textbox is where you will enter the URL for your label studio project, at which point you will be able to interact with your data using label studio.

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



Database Information

Website Info

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 local database the application manages. When there is data to show it will replace the text that, in the picture above, says "Website Info".

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 that when you want to access saved data that you scraped previously!

Logs