# **Accessing Social Content Using Web Scraper**

## Installation:

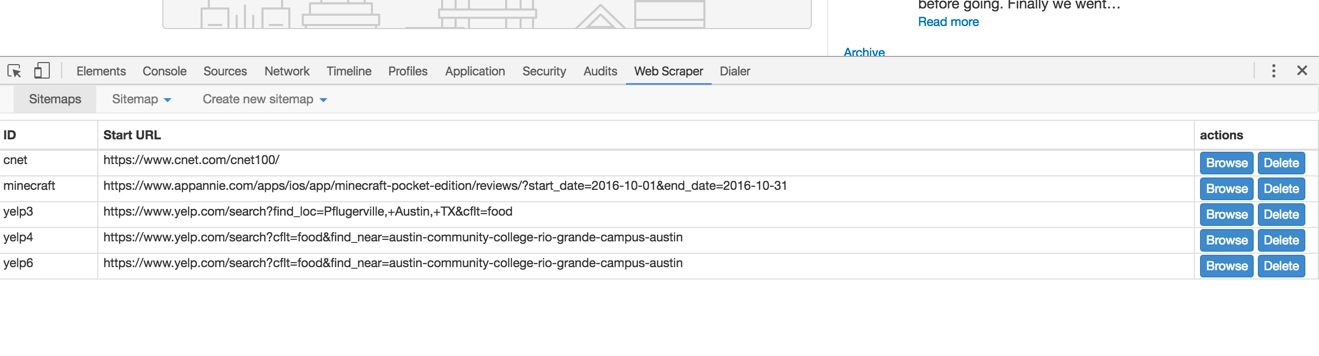
Web-scraper is available as an extension to the chrome browser. Use the following link to add it to your browser.

*Link*: <http://webscraper.io/>

After adding it, web-scraper can be accessed on your browser through developer tools. Based on your OS you can directly open the developer tools using following short cuts:

* Windows, Linux: Ctrl+Shift+I or F12
* Mac: Cmd+Opt+I
* Any OS: open Tools / Developer tools

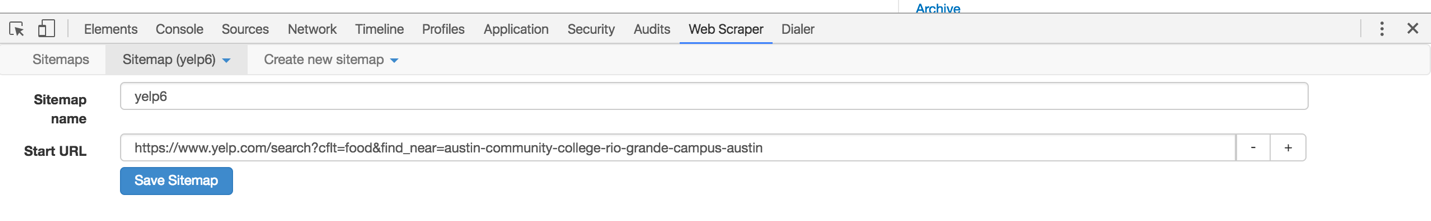
Click on the web-scraper within the developer tools. The window looks as follows:



If you are opening it for the first time you will not be able to see any sitemaps. It will be completely empty.

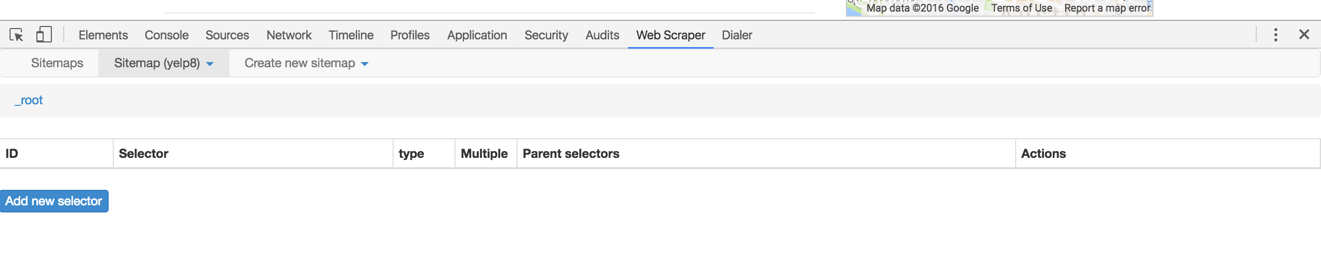
## **Demo with Yelp**

Click on the “Create new sitemap” and select “create sitemap”, a window will then be displayed as follows:

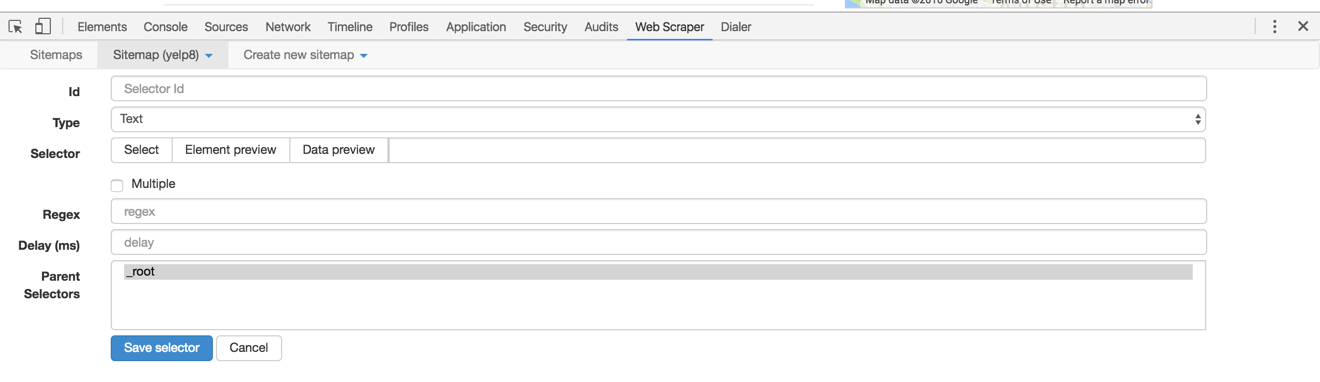


Provide a sitemap name and specify the URL (web address) of the initial page. This is the page with a list of restaurants. It is considered as a root or the starting point of navigation. Save the sitemap after providing the information.

Once you have saved the sitemap, an empty window with a button to add new selector will be displayed as follows:

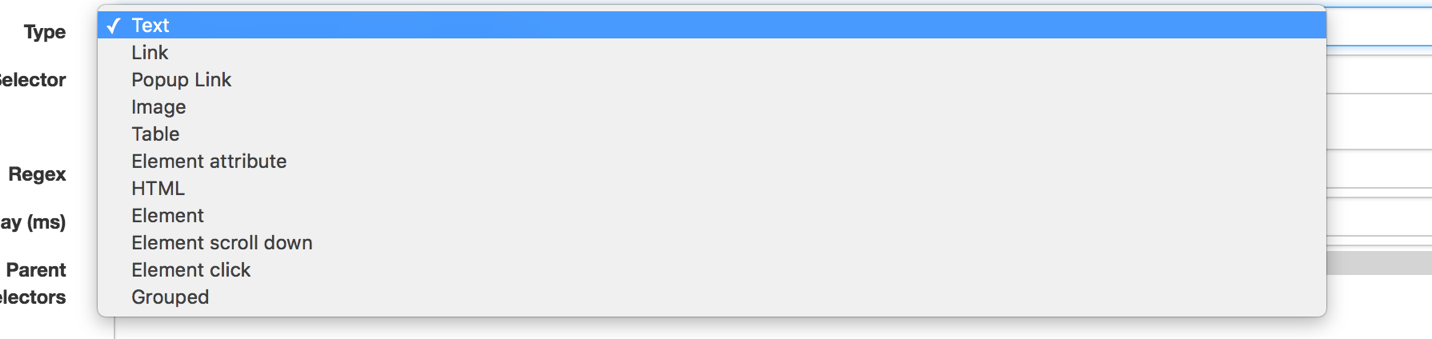


Click on the button to specify the action that you wish to perform on the root page. A window with various options is displayed as below:



Id: provide a unique name to the action that you wish to perform. Any name can be used.

Type: Type specifies what kind of action you want to do. There are various possibilities:



Text or image: to extract the text or image.

Link: helpful in navigating the pages

Popup Link: helpful in dealing with the popup windows which show up when clicked on.

Element: helpful in playing around the divisions. Extremely useful in the case of forum websites where all the reviews are similarly organized.

For the purpose of this demo, I use Element, Text, and Link types. A complete list of various uses of each element is present in the following link: <http://webscraper.io/documentation>

Since the idea is to scrape the reviews of good chunk of food places in a particular location, I have manually selected the location and the business type. Then I provided the link of that page as a root. My root webpage is the following:

<https://www.yelp.com/search?cflt=food&find_near=austin-community-college-rio-grande-campus-austin>

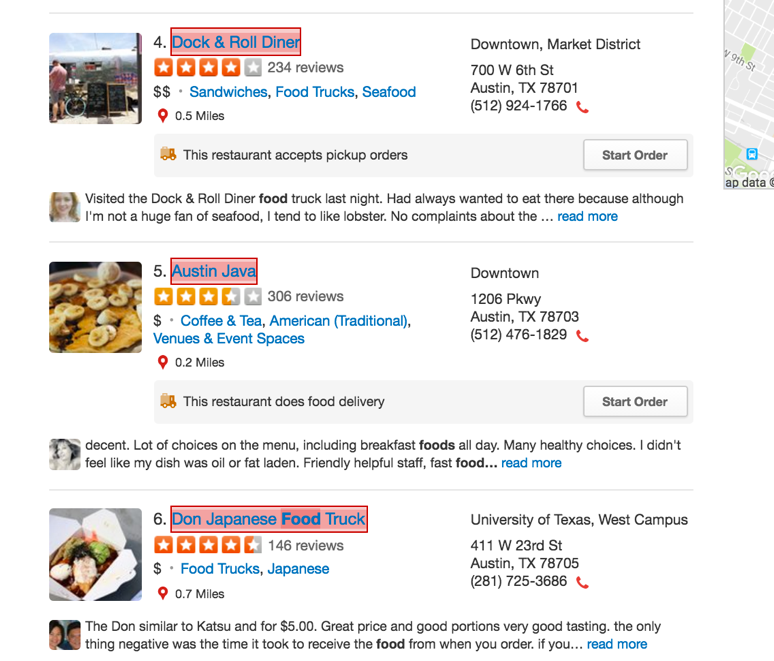
The above is just an example. You need to search based on your requirements. I want the scraper to visit the review page of each restaurant that has been listed and get the review text and demographic information of the reviewers. In order to do that we need to select the links of all the restaurants on the page.

## Selecting the Restaurant Links:

Select the sitemap that you have just created and click on the “add selector” button.

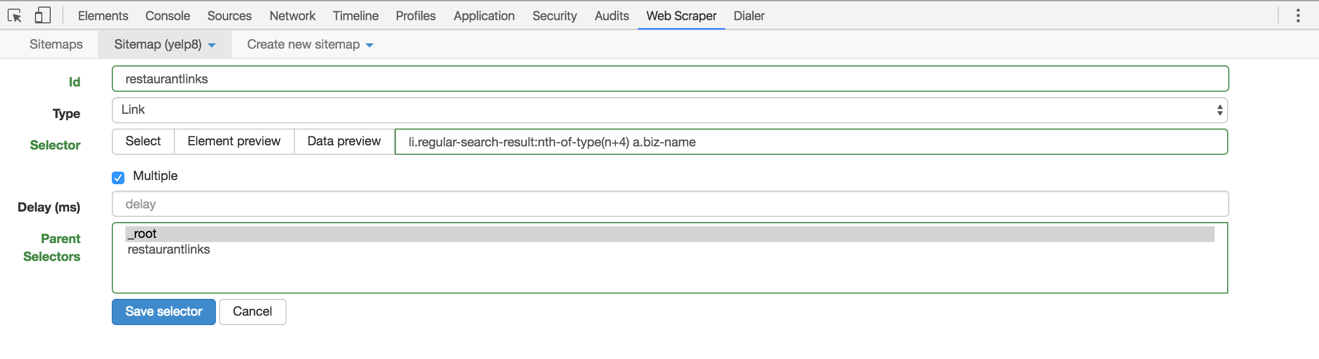
1. Give an “Id” of your choice.
2. Select the “Link” from types.
3. Check the multiple option (as we wish to get multiple links of restaurants)
4. Delay can be left empty.
5. Click on the “select” button in the selector bar. Now you will be able to mark the links on the web page. On the web page click on at least two restaurant links. It will automatically select the links of all the other links present in the page.

The web page with selected links looks as follows:

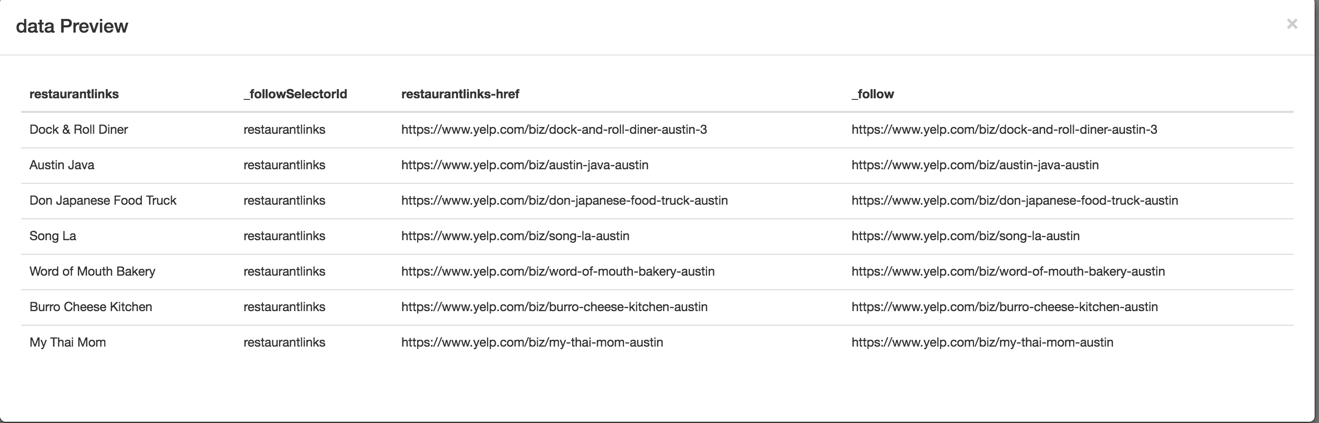


Once you have selected all restaurant links. Click on the done selecting button.

The selector view of the web-scraper looks as follows:



Click on the “Data preview” button to make sure that the tool has selected all the restaurant links. You will be able to see a pop up similar to this:



You can now close the pop and save the selector.



## Selecting the review elements:

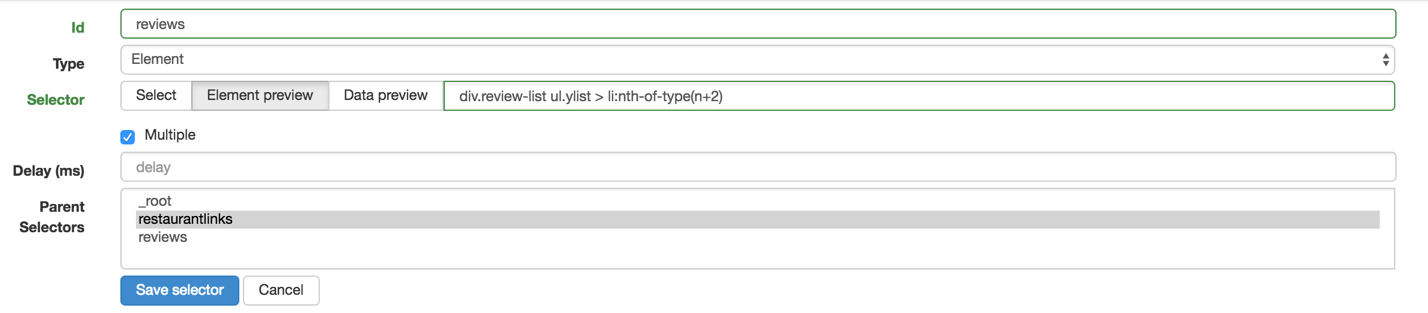
This sections helps you to perform actions on each of the links that you have just created. Within the web-scraper view, click on the “restaurantlinks” selector that you have created and go to a restaurant page in yelp.

Within the web-scraper:

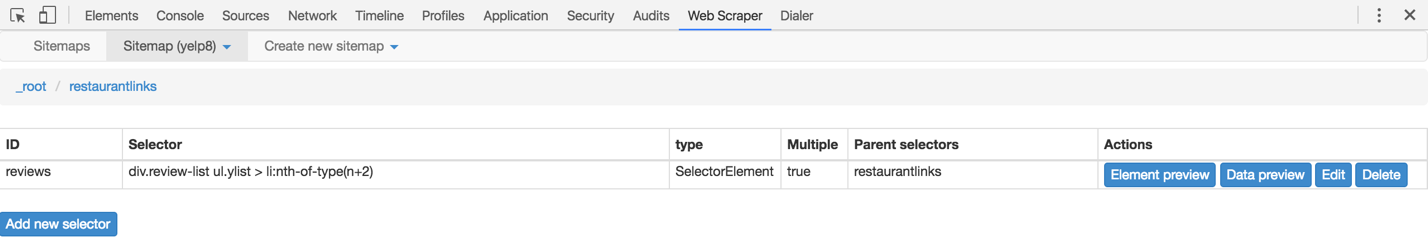
1. Click on the “Add Selector” button.
2. Fill in the “Id” with some name
3. Select the “Element” as type as each review from a user is displayed in a similar “div” tag.
4. Check “multiple” box as we have to select reviews of multiple user
5. Click on the “select” button and select at least two divisions of reviews with in the review page. The selected web page looks as follows:



When you have selected the divisions of reviews. You can save the selection. The selector with in the web-scraper looks as below.



You can now save the selector.

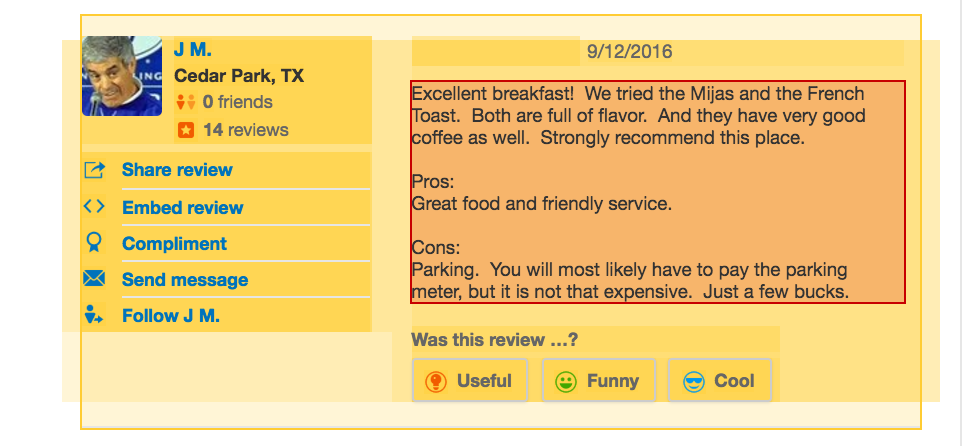


## Selecting the Review Text and User information:

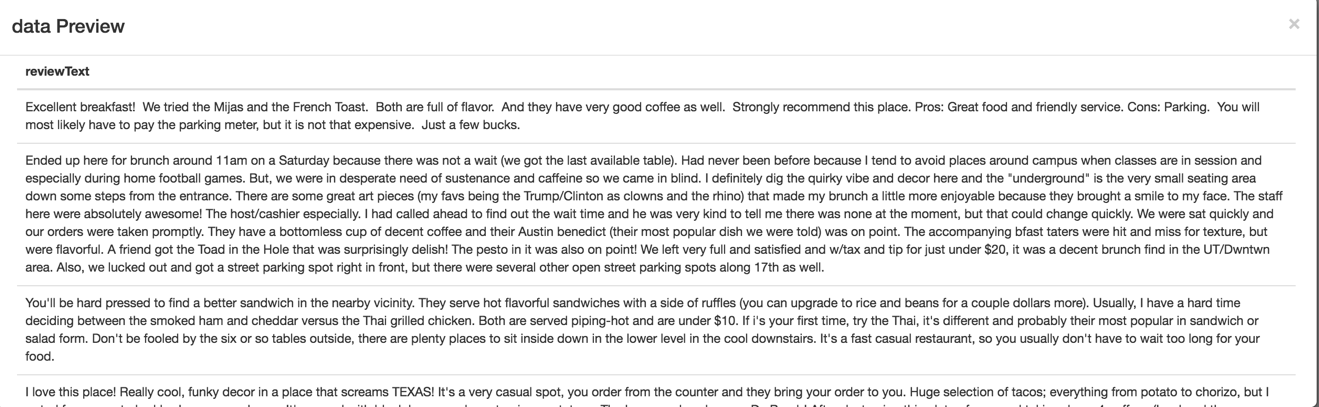
You can now select the parts of the “review element” that you want to extract. Click on the “reviews” element that you have just created and click on the “add new selector” button.

Selecting review text:

1. Fill in the Id
2. Select “text” as Type
3. Click on Multiple
4. Click on “select”. A part of the review element gets highlighted on the web page. Within this highlighted portion, select the review text. The selected portion of the webpage looks as follows:



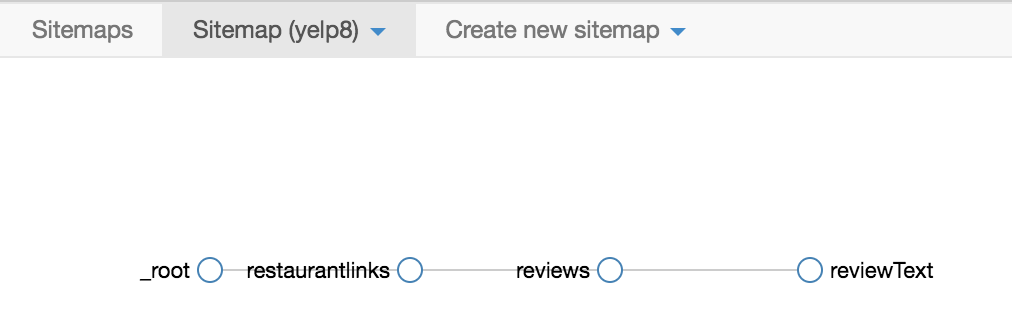
Once you have selected the text. Save the selection by clicking on “Done selecting”. Click on “Data Preview” to preview the review text the tool has selected. You will be able to see a pop as follows:



You can now close the preview and save the selector.

Similarly, you can select the name of the user, date of the review, etc.

At this stage you can look at the site map of the scraper, to confirm all the steps that you have added to the tool. The site map looks as follows:

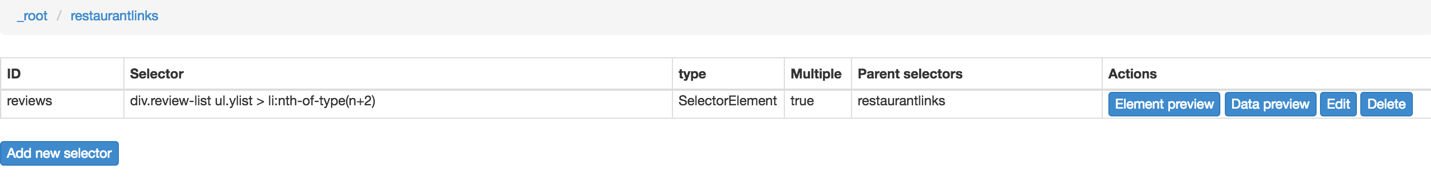


You can now go ahead and pagination to the reviews.

## Pagination to the reviews:

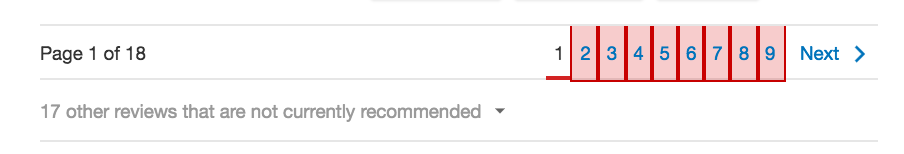
Each restaurant has at least two pages of reviews. So far, we have only enabled the functionality to scrape the reviews from the first or default display page. In this step, we will enable the web-scraper to scrape the reviews from other pages as well. The idea is simple. We will add the links of other pages and make the review element (which can get the review text) as a child of this node.

Click on the “selectors” from the “sitemap” and go to the “restaurantlinks” selector to enable pagination.

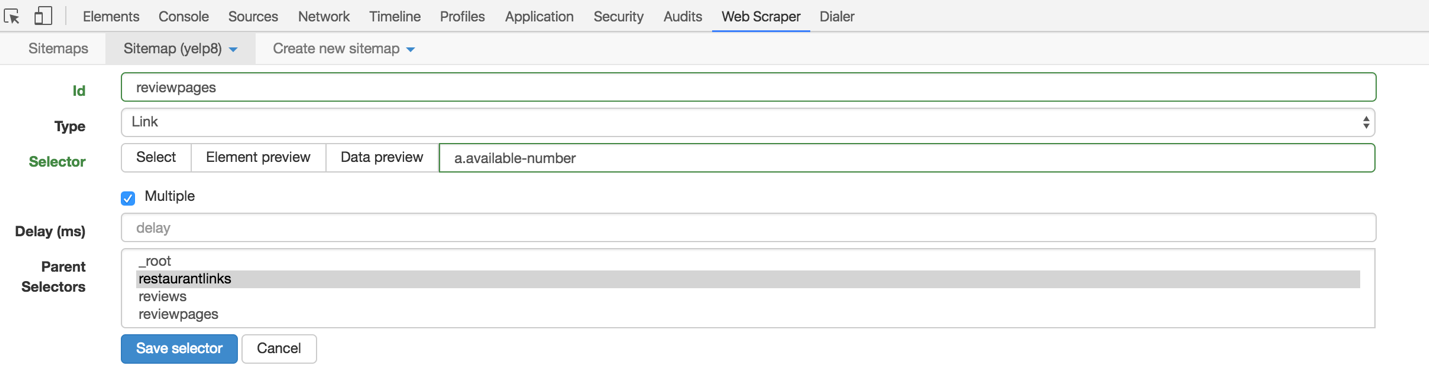


Click on the “Add new selector” button.

1. Fill in Id with some name like “reviewpages”
2. Select “Links” from Types
3. Check the “multiple” option
4. Click on selector and select the page numbers which have the link to navigate to the next page.



Once you are done with selecting the links, save the selection and the “reviewpages” selector.



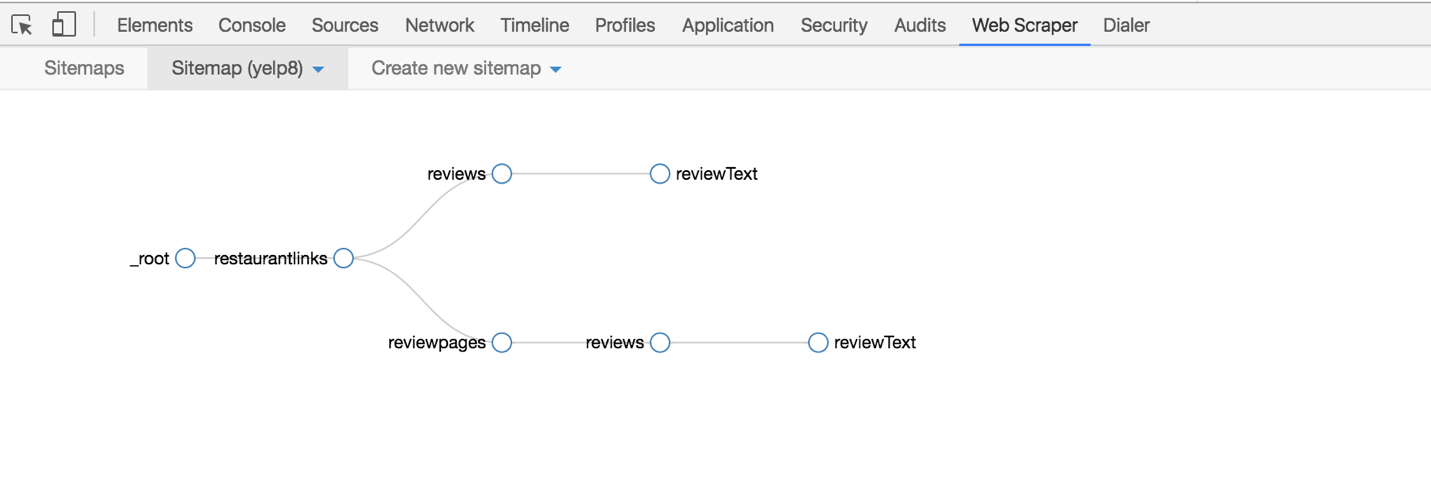


We will now make the “reviews” selector child of the “reviewpages” selector.

1. Click on the edit button of the reviews selector.
2. Select reviewpages from the “Parent Selectors” and save.

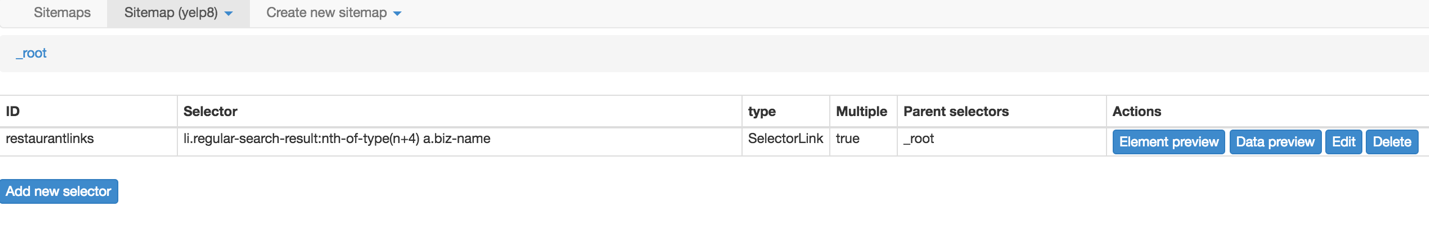


We can now view the “selector graph” to make sure the steps in navigation. The selector graph looks as follows:



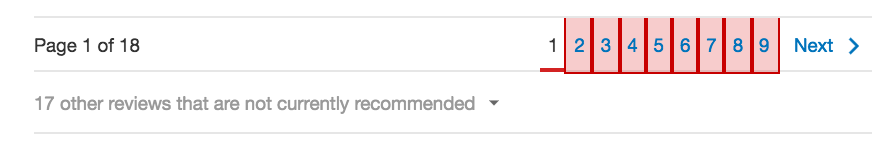
## Pagination of Restaurant Links:

In this section, we will try to scrape the reviews of restaurants listed in pages 2-9. The idea is similar to the pagination that we have done for reviews. Click on the selectors and go to the root selector in the web-scraper.

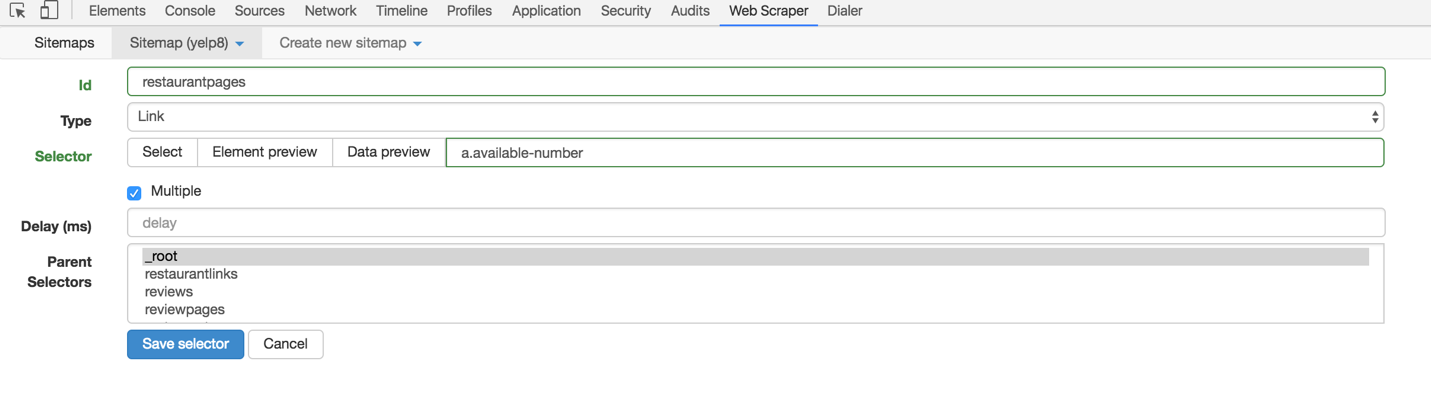


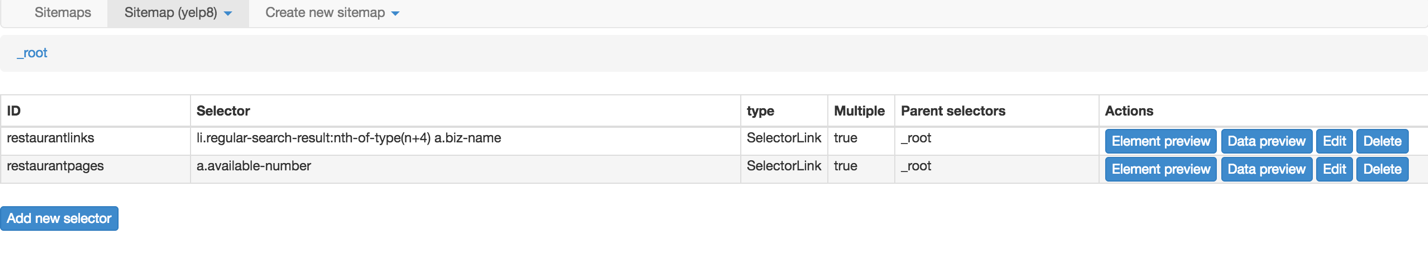
To add pagination to the restaurant links:

1. Click on the “Add new selector” button.
2. Fill in the box for “Id” with something like “restaurantpages”
3. Select “Links” from the Type
4. Check the “multiple” box.
5. Click on the select button and select the pages that you want to navigate.

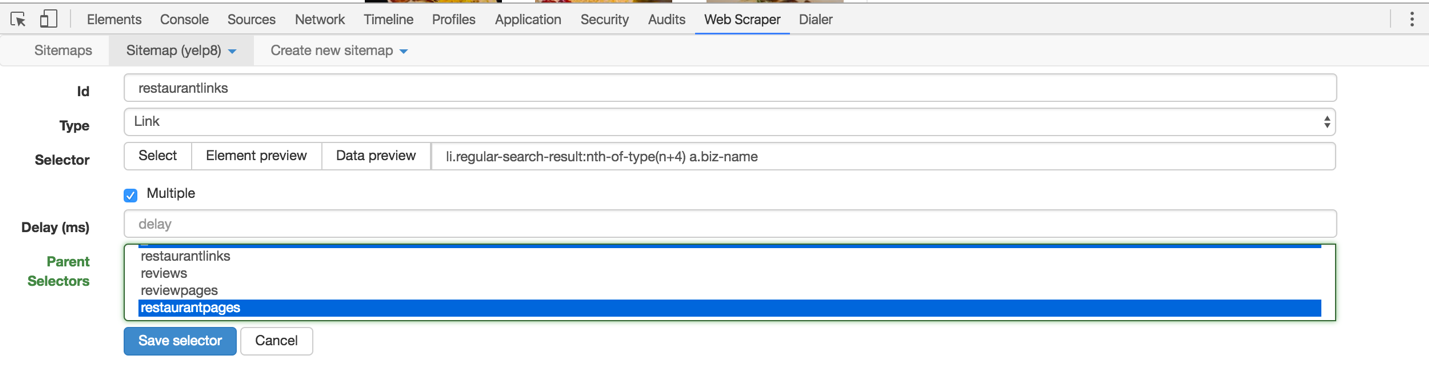


1. When you are done selecting, save the selection and the selector.

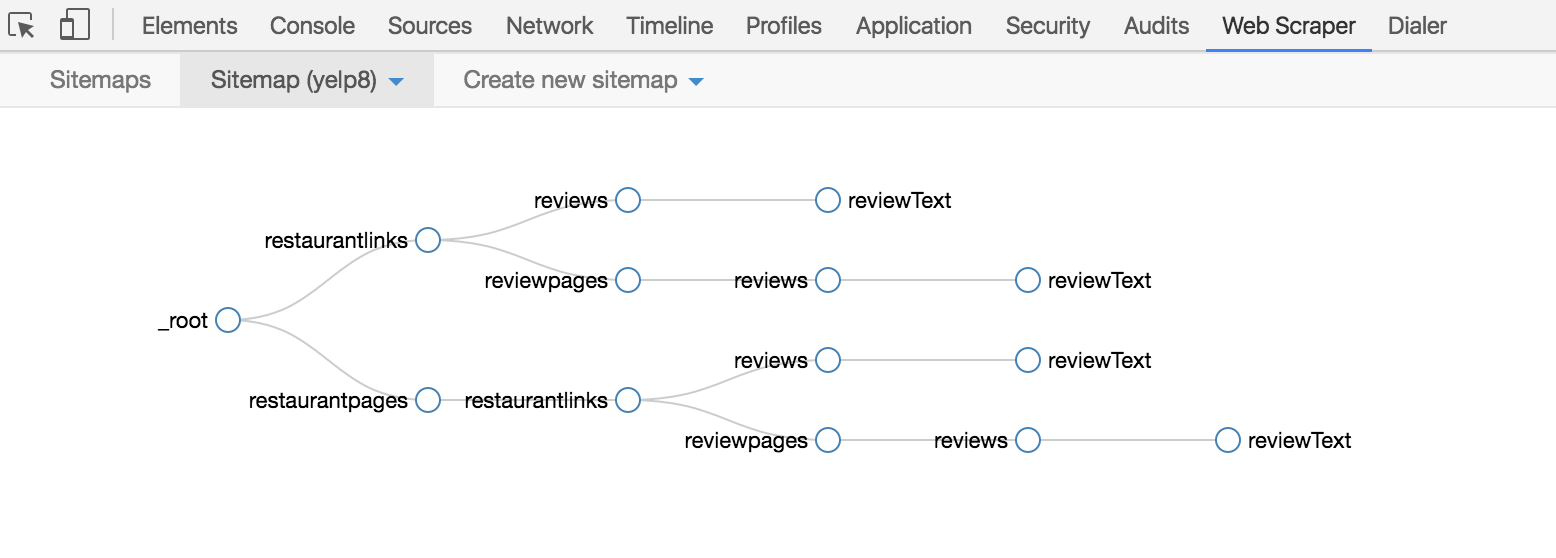




1. Now click on the edit button of the “restaurantlinks” selector and select “restaurantpages” from the parent selectors.



You can now save the selector. Once again to check the stages of scraping, we will go ahead and look at the selector graph.



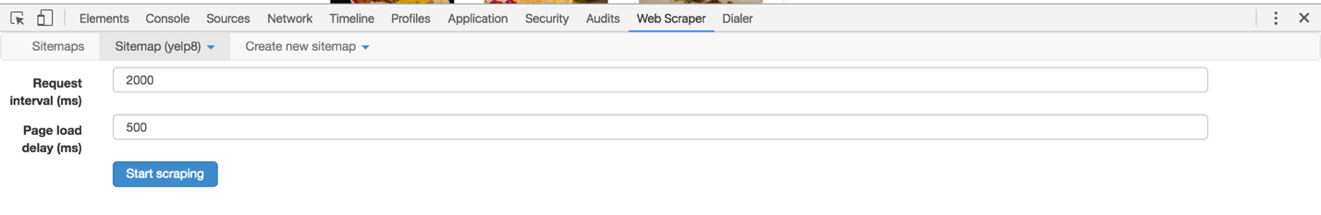
**Scraping ratings**

In the example above, I have not extracted ratings provided by users along with their reviews (but the homework requires you to do that). That would be another branch starting from reviews (parallel to reviewText), and HTML should be selected as type to extract the ratings. There will be a lot more html stuff in this column in addition to the rating (e.g., 4.0), but that extra content can be removed in Excel as shown in the assignment guidelines.

When you add the selector for rating, do a data preview, which should look like

<img class="offscreen" height="303" src="https://s3-media2.fl.yelpcdn.com/assets/srv0/yelp\_design\_web/9b34e39ccbeb/assets/img/stars/stars.png" width="84" alt="**4.0 star rating**">

Once you have completed the sitemap, you can start scraping by selecting scrape option and wait for it to finish.



You can use the default load times, but in case there is a problem, increase the page load time to something like 3000 ms.

Once the scraping is finished. You can export the csv by selecting the “export csv”. You can directly download the csv file by clicking on the “Download now!” link.