Joshua Conte
6/15/17

DA5020 – Collect,Store,Retrieve Data
Assignment 6

Introduction

For this assignment I used 2 web scrapping toolkits.

The first one is a free software package called Web Scraper which is an extension for Chrome, it can be found at:
http://webscraper.io/

The second one is import.io. This one is a little expensive, $99 a year for students and $299 for non-students. The software can be found at:
https://www.import.io/standard-plans/

I used two websites to gauge how the software works. For the first website, I used Wikipedia and found a list and I extracted the data from it as a table, the website is:
https://en.wikipedia.org/wiki/List_of_genetic_disorders

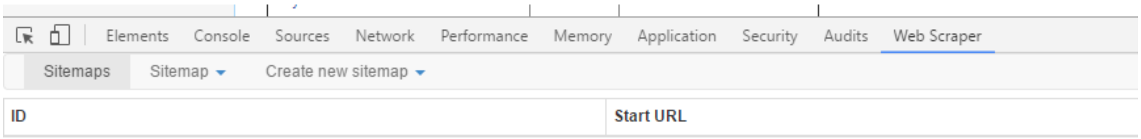
For the second website, I went to Zillow.com and tried extracting the information from a search of houses in the area I'm looking to buy a house:
https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1837-2756_mp/39.305413,-76.711264,39.192418,-76.8958_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/

Part 1: Web Scraper

Web Scraper is an extension for chrome browser made exclusively for web data scraping. You can setup a plan (sitemap) on how to navigate a website and specify the data to be extracted. The scraper will traverse the website according to the setup and extract the relevant data. It lets you export the extracted data to CSV. Multiple pages can be scraped using the tool making it all the more powerful. It can even extract data from dynamic pages that use Javascript and Ajax. All you need is to use Google Chrome.

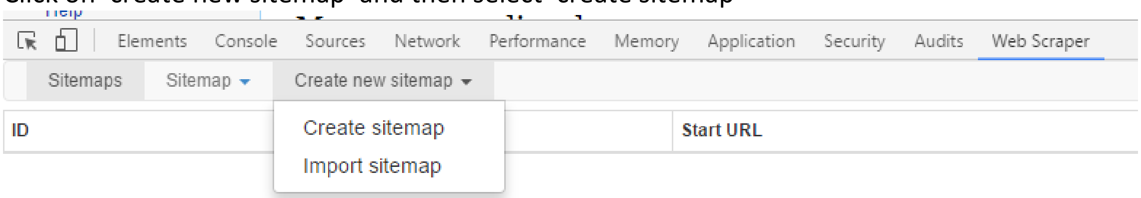
Getting Started with Tables:

After installation, open the Google chrome developer tools by pressing F12. (You can alternatively right click on the screen and select inspect element). In the developer tools, you will find a new tab named 'Web scraper' as shown in the screenshot below:

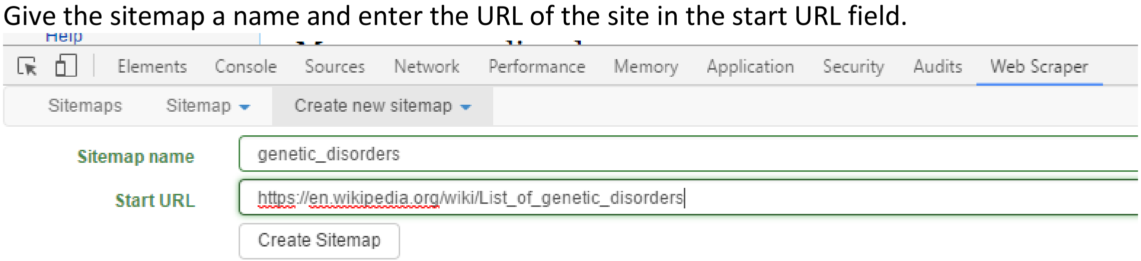


- To extract data open the website https://en.wikipedia.org/wiki/List_of_genetic_disorders
- Open developer tools by right clicking anywhere on the screen and then selecting inspect or by pressing F12 (as shown above)
- Click on the web scraper tab in developer tools

- Click on ‘create new sitemap’ and then select ‘create sitemap’



- Give the sitemap a name and enter the URL of the site in the start URL field.



Sitemap name: genetic_disorders

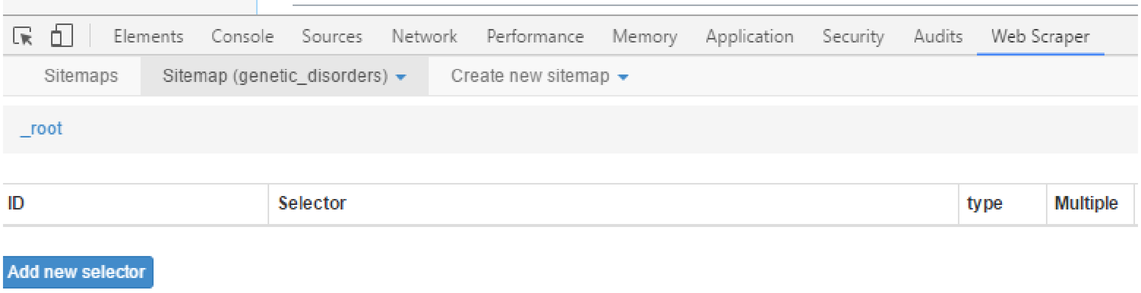
Start URL: https://en.wikipedia.org/wiki/List_of_genetic_disorders

- Click on ‘Create Sitemap’

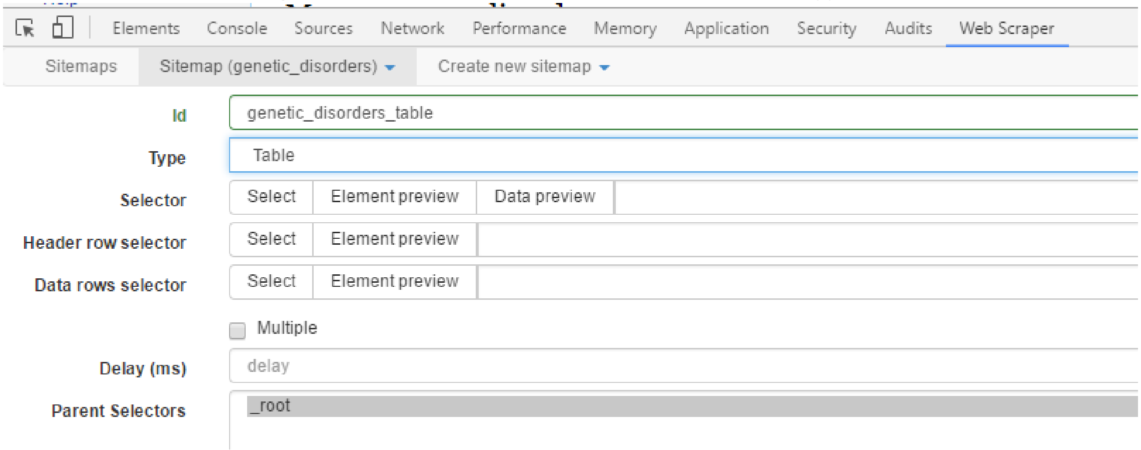
Scraping Elements of the Table:

I started with a table because I thought it would be the easiest to start with. According to the tutorial there are two ways to select the data, use the CSS selector by looking at the source file of the web page (CTRL+U) or use the selector tool to click and select any element on the screen. To use the selector:

- Click on the Sitemap and click on ‘Add new selector’.



- In the selector id field, give the selector a name and select table in Type.



Id: genetic_disorders_table

Type: Table

Selector: Select | Element preview | Data preview

Header row selector: Select | Element preview

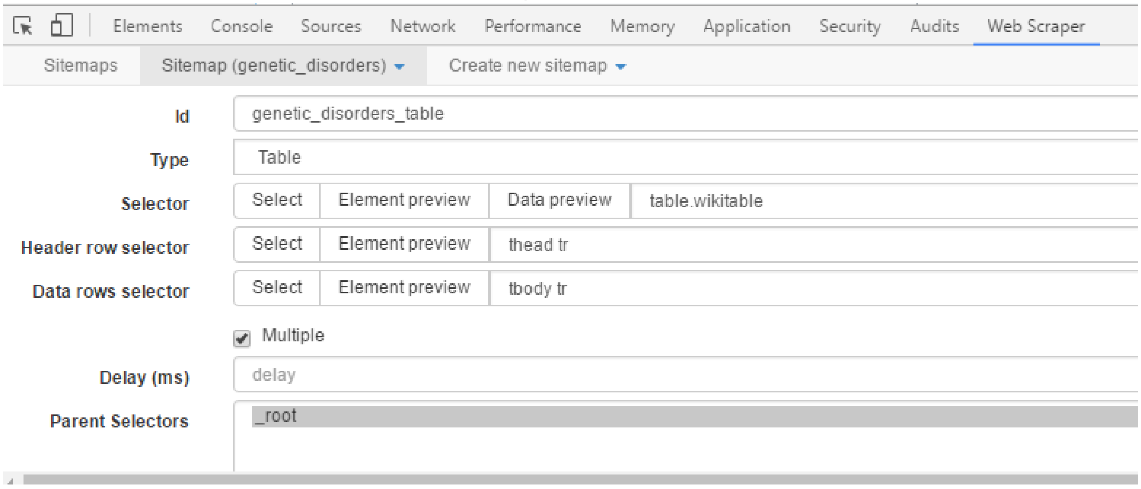
Data rows selector: Select | Element preview

☐ Multiple

Delay (ms): delay

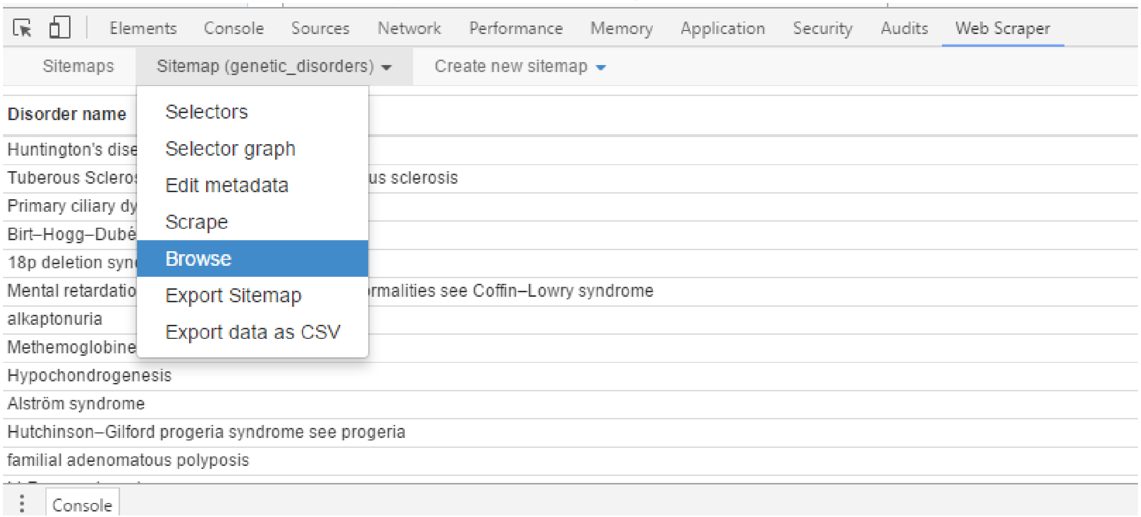
Parent Selectors: _root

- Click on the select button and select any element on the web page that you want to be extracted. When you are done selecting, click on ‘Done selecting’. check the ‘multiple’ checkbox to indicate that the element you want can be present multiple times on the page and that you want each instance of it to be scraped.



Id: genetic_disorders_table
Type: Table
Selector: table.wikitable
Header row selector: thead tr
Data rows selector: tbody tr
☑ Multiple
Delay (ms): delay
Parent Selectors: _root

- Now you can save the selector if everything looks good. To start the scraping process, just click on the sitemap tab and select ‘Scrape’. A new window will pop up and scrape the required data. If you want to stop the scraping process in between, just close this window and you will have the data that was extracted till then. Select Export data as CSV to get the data:



Disorder name
Huntington's dise
Tuberous Sclero: ... us sclerosis
Primary ciliary dy
Birt–Hogg–Dubé
18p deletion syn
Mental retardatio ... rmalities see Coffin–Lowry syndrome
alkaptonuria
Methemoglobine
Hypochondrogenesis
Alström syndrome
Hutchinson–Gilford progeria syndrome see progeria
familial adenomatous polyposis

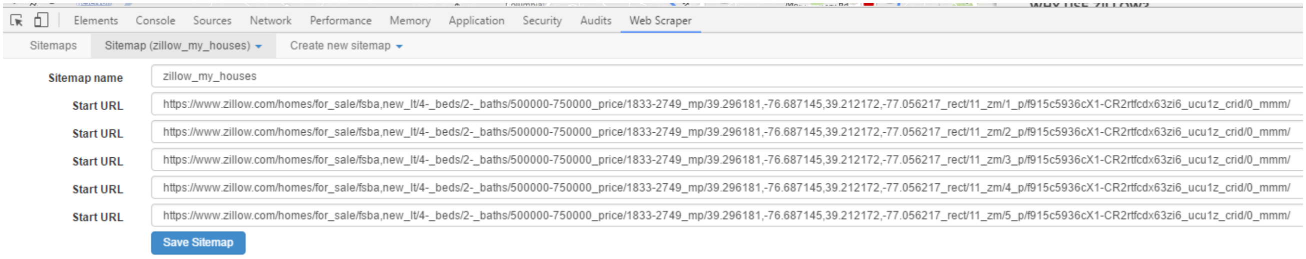
Once the info is in CSV form it’s easy to analyze in R, below is the output in R:

```
> # This imports a csv file that contains both numeric and character variable.
> # By default, the data is loaded as a list and data.frame.
> # I also added NA to all blank cells to make it easier to analyze and
> # stringsAsFactors = FALSE so I can remove levels from the data.
> if (!exists("gen.data")) {
+   gen.data <- read.csv(
+     " genetic_disorders.csv",
+     header = TRUE,
+     stringsAsFactors = FALSE,
+     na.strings = c("", "NA"),
+     row.names = NULL,
+     sep = ","
+   )
+ }
>
> class(gen.data)
[1] "data.frame"
> summary(gen.data)
 ï..Disorder.name   Mutation.type       Chromosome
 Length:183          Length:183          Length:183        
 Class :character    Class :character    Class :character  
 Mode  :character    Mode  :character    Mode  :character  
> head(gen.data)
                                                                                ï..Disord
er.name
1                                                                            Huntington's 
disease
2                               Tuberous Sclerosis Complex (TSC)\nsee Tuberous sc
lerosis
3                                                         Primary ciliary dyskinesi
a (PCD)
4                                                             Birtâ€“Hoggâ€“DubÃ© s
yndrome
5                                                                     18p deletion s
yndrome
6 Mental retardation with osteocartilaginous abnormalities\nsee Coffinâ€“Lowry s
yndrome
  Mutation.type Chromosome
1             T     4p16.3
2          <NA> TSC1, TSC2
3          <NA>       <NA>
4          <NA>         17
5             D        18p
6          <NA>       <NA>
```

The R file is also attached to this pdf as Conte_J_6A.r.

Getting Zillow Data:

Next, I wanted to see how this software would perform on more advanced tasks, like collecting information from Zillow. I used the url from the introduction and created a new sitemap as noted above. I did make one modification, to search more than one page I add all of the urls:



Sitemap name: zillow_my_houses

Start URL: https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1833-2749_mp/39.296181,-76.687145,39.212172,-77.056217_rect/11_zm/1_p/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/

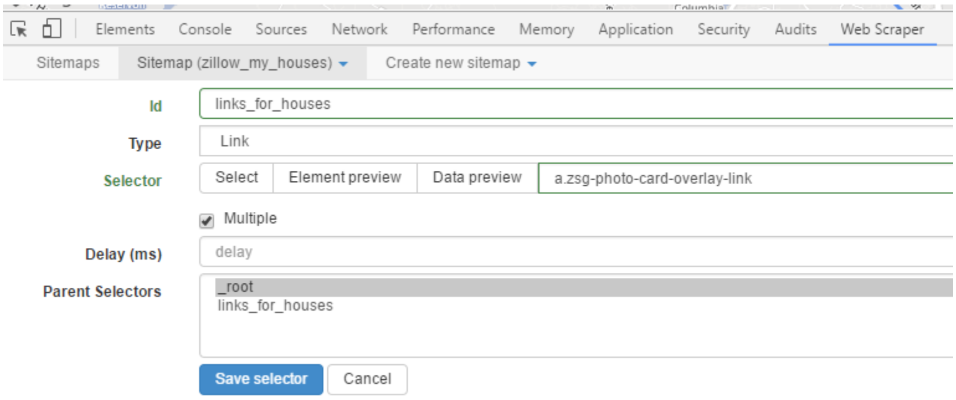
Start URL: https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1833-2749_mp/39.296181,-76.687145,39.212172,-77.056217_rect/11_zm/2_p/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/

Start URL: https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1833-2749_mp/39.296181,-76.687145,39.212172,-77.056217_rect/11_zm/3_p/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/

Start URL: https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1833-2749_mp/39.296181,-76.687145,39.212172,-77.056217_rect/11_zm/4_p/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/

Start URL: https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1833-2749_mp/39.296181,-76.687145,39.212172,-77.056217_rect/11_zm/5_p/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/

Then I created a new selector using the Type Link. I could not figure out how to scrape other information other than the links:



Id: links_for_houses

Type: Link

Selector: a.zsg-photo-card-overlay-link

☑ Multiple

Delay (ms): delay

Parent Selectors: _root
links_for_houses

Finally I scrapped the data and exported it as a CVS file. This is the R output:

```
> # This imports a csv file that contains both numeric and character variable.
> # By default, the data is loaded as a list and data.frame.
> # I also added NA to all blank cells to make it esaier to analyze and
> # stringsAsFactors = FALSE so I can remove levels from the data.
> if (!exists("zillow.data")) {
+   zillow.data <- read.csv(
+     "zillow_my_houses.csv",
+     header = TRUE,
+     stringsAsFactors = FALSE,
+     na.strings = c("", "NA"),
+     row.names = NULL,
+     sep = ","
+   )
+ }
>
> class(zillow.data)
[1] "data.frame"
> summary(zillow.data)
 ï..links_for_houses links_for_houses.href
 Mode:logical        Length:79
 NA's:79             Class :character
                     Mode  :character
> head(zillow.data)
  ï..links_for_houses
1                  NA
2                  NA
3                  NA
4                  NA
5                  NA
6                  NA
                                                                   links_f
or_houses.href
1        https://www.zillow.com/homedetails/8232-Elko-Dr-Ellicott-City-MD-21043/
36991645_zpid/
2             https://www.zillow.com/community/daniels-grove-at-patapsco-park/20
96741833_zpid/
3  https://www.zillow.com/homedetails/4902-Clearwater-Dr-Ellicott-City-MD-21043/
37031230_zpid/
4     https://www.zillow.com/homedetails/4643-Huntley-Dr-Ellicott-City-MD-21043/
37033478_zpid/
5 https://www.zillow.com/homedetails/4103-Sears-House-Ct-Ellicott-City-MD-21043/
53568409_zpid/
6  https://www.zillow.com/homedetails/10334-Pinehurst-Ct-Ellicott-City-MD-21042/
37028363_zpid/
```

The R file is also attached to this pdf as Conte_J_6B.r.

For some reason the scraper tool collected 2 columns, one with nothing in it and one with the links. This is easy to fix in R or excel, but I could not eliminate it in the scraping tool.

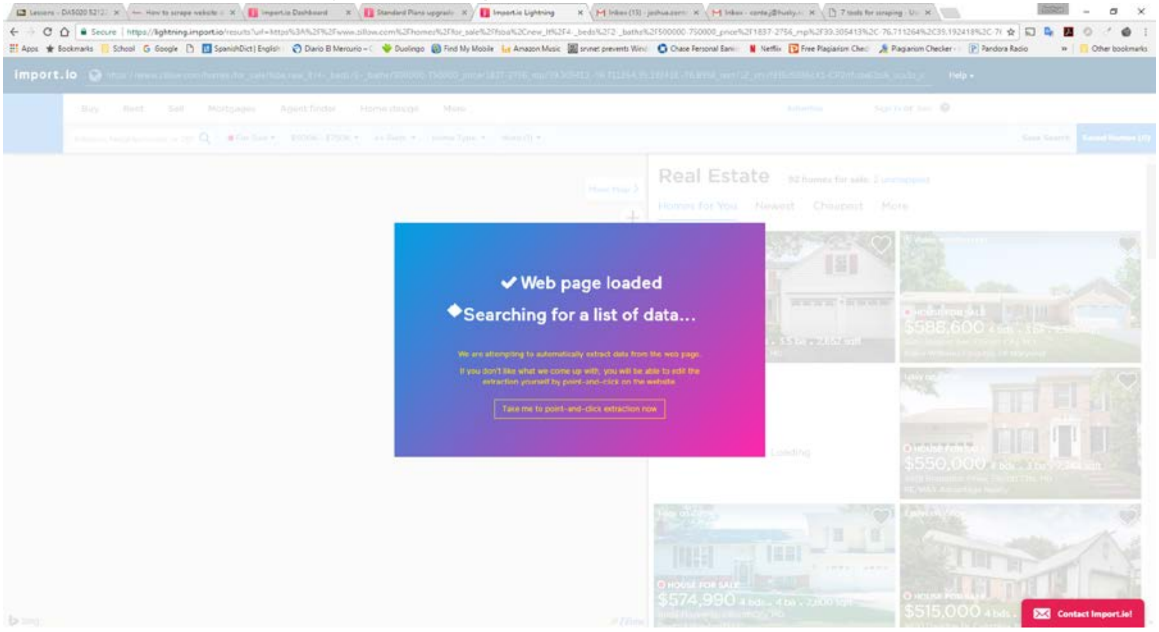
DA5020 – Collect,Store,Retrieve Data
Assignment 6

Part 2: import.io

Import.io uses highly sophisticated machine learning algorithms, to extract data automatically. The software gets great reviews and it looks really easy to use.

Getting Zillow Data:

I was a little anxious to try import.io on Zillow to see what info it could extract. It was easy to use, first you need to enter the URL that you want to go to:



✔ Web page loaded

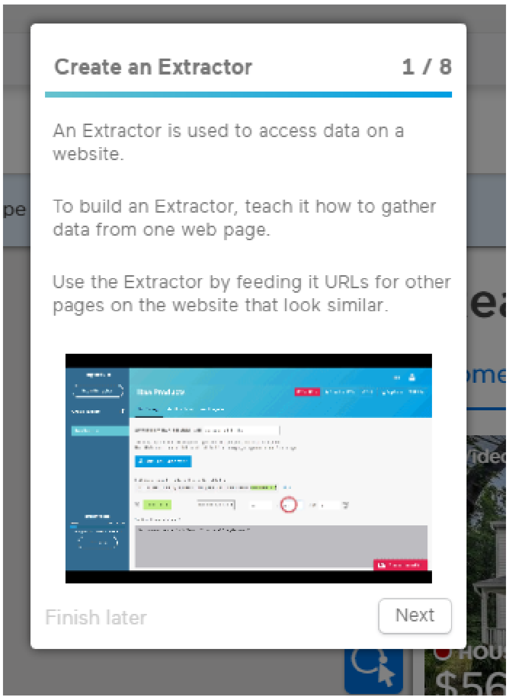
◆Searching for a list of data...

We are attempting to automatically extract data from the web page.

If you don't like what we come up with, you will be able to edit the extraction yourself by point-and-click on the website.

Take me to point-and-click extraction now

Then import.io will ask you to create an extractor and within 7 steps you can have everything you need:



Create an Extractor 1 / 8

An Extractor is used to access data on a website.

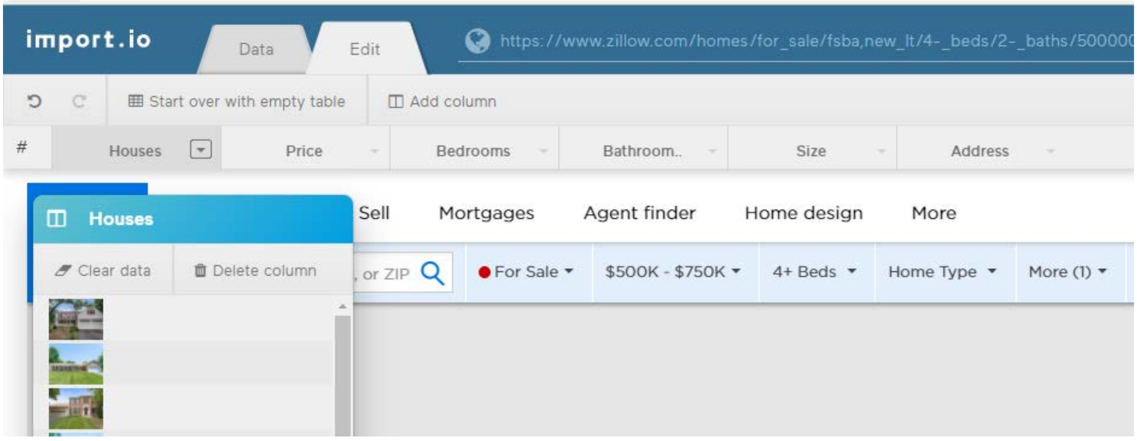
To build an Extractor, teach it how to gather data from one web page.

Use the Extractor by feeding it URLs for other pages on the website that look similar.

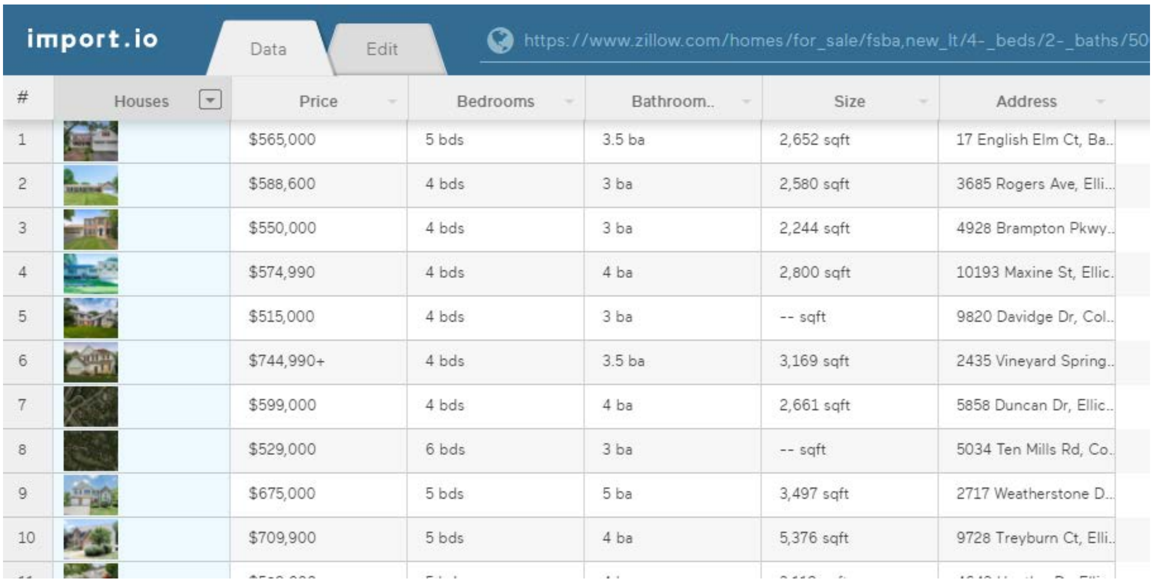
Joshua Conte
6/15/17

DA5020 – Collect,Store,Retrieve Data
Assignment 6

1) **Extract data into a column**
   a) Click on an item on the web page that you would like to extract.
   b) The item value will be extracted into the selected column.
   c) If you are trying to extract multiple items, keep clicking on items until all values are extracted.
2) **Selected column**
   a) The data that you are extracting into the selected column is highlighted on the web page and displayed in this floating window.
3) **Add column**
   a) Add a new column in order to extract additional properties from the items listed on the page.
4) **Columns of data**
   a) As you add more columns and extract more data into those columns, you can switch between columns by clicking on the column headings here.



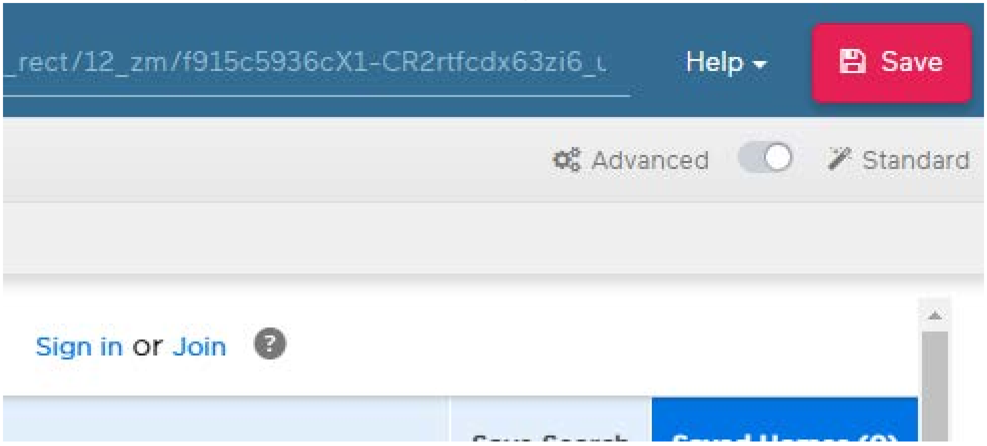
5) **The "Data" tab**
   a) The Data tab allows you to view all of the data that you are extracting from the web page in a single table.


import.io
Data
Edit
https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/50

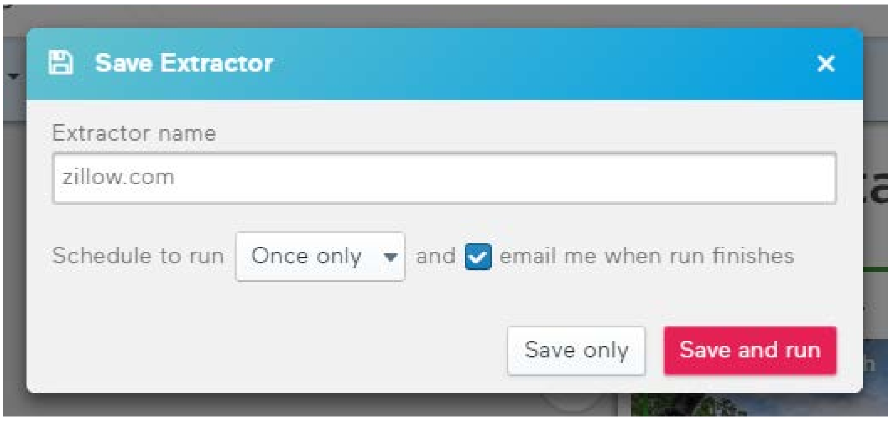
#	Houses	Price	Bedrooms	Bathroom..	Size	Address
1		$565,000	5 bds	3.5 ba	2,652 sqft	17 English Elm Ct, Ba..
2		$588,600	4 bds	3 ba	2,580 sqft	3685 Rogers Ave, Elli..
3		$550,000	4 bds	3 ba	2,244 sqft	4928 Brampton Pkwy..
4		$574,990	4 bds	4 ba	2,800 sqft	10193 Maxine St, Ellic..
5		$515,000	4 bds	3 ba	-- sqft	9820 Davidge Dr, Col..
6		$744,990+	4 bds	3.5 ba	3,169 sqft	2435 Vineyard Spring..
7		$599,000	4 bds	4 ba	2,661 sqft	5858 Duncan Dr, Ellic..
8		$529,000	6 bds	3 ba	-- sqft	5034 Ten Mills Rd, Co..
9		$675,000	5 bds	5 ba	3,497 sqft	2717 Weatherstone D..
10		$709,900	5 bds	4 ba	5,376 sqft	9728 Treyburn Ct, Elli..

6) **Advanced options**
   a) More advanced extraction options are available for particularly difficult websites.

7) **Save the Extractor**
   a) Once your Extractor is pulling the data that you want, save it.



b) It will give you an option of how many times you would like to run this extractor, i.e. once a day, once a week….

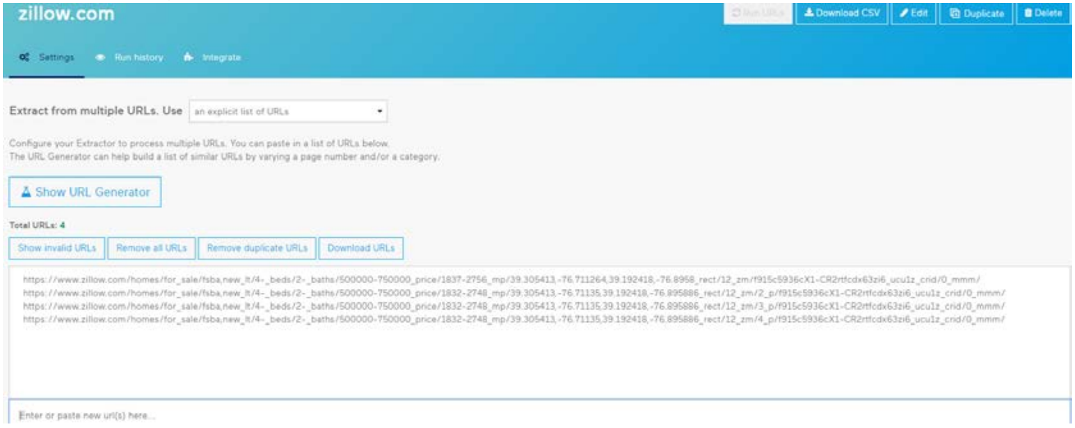


Save Extractor

Extractor name: zillow.com

Schedule to run: Once only and ☑ email me when run finishes

c) This will take you to the dashboard from where you can download the data and add additional URLs.



Extract from multiple URLs. Use: an explicit list of URLs

Configure your Extractor to process multiple URLs. You can paste in a list of URLs below.
The URL Generator can help build a list of similar URLs by varying a page number and/or a category.

Total URLs: 4

```
https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1837-2756_mp/39.305413,-76.711264,39.192418,-76.8958_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_uculz_crid/0_mmm/
https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1832-2748_mp/39.305413,-76.71135,39.192418,-76.895886_rect/12_zm/2_p/f915c5936cX1-CR2rtfcdx63zi6_uculz_crid/0_mmm/
https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1832-2748_mp/39.305413,-76.71135,39.192418,-76.895886_rect/12_zm/3_p/f915c5936cX1-CR2rtfcdx63zi6_uculz_crid/0_mmm/
https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-750000_price/1832-2748_mp/39.305413,-76.71135,39.192418,-76.895886_rect/12_zm/4_p/f915c5936cX1-CR2rtfcdx63zi6_uculz_crid/0_mmm/
```

Enter or paste new url(s) here...

d) Then you can run the URLs and download the data

The data saves as a csv file and it's easy to analyze in R, below is the output in R:

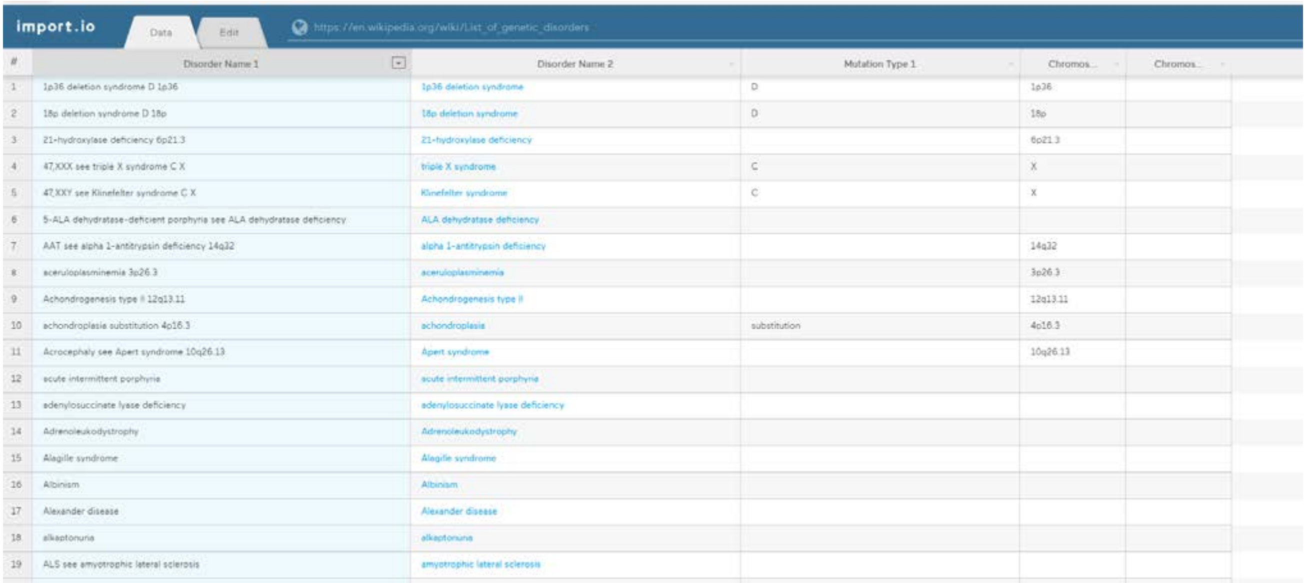
```
> # This imports a csv file that contains both numeric and character variable.
> # By default, the data is loaded as a list and data.frame.
> # I also added NA to all blank cells to make it esaier to analyze and
> # stringsAsFactors = FALSE so I can remove levels from the data.
> if (!exists("zillow.import.data")) {
+   zillow.import.data <- read.csv(
+     "zillow_import.csv",
+     header = TRUE,
+     stringsAsFactors = FALSE,
+     na.strings = c("", "NA"),
+     row.names = NULL,
+     sep = ","
+   )
+ }
>
> class(zillow.import.data)
[1] "data.frame"
> summary(zillow.import.data)
    ï..url              Houses           Houses_alt        Price              Bedro
oms
 Length:50          Length:50          Mode:logical   Length:50          Length:
50
 Class :character   Class :character   NA's:50        Class :character   Class :
character
 Mode  :character   Mode  :character                  Mode  :character   Mode  :
character
  Bathrooms             Size              Address
 Length:50          Length:50          Length:50
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
> head(zillow.import.data)

ï..url
1 https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-7500
00_price/1832-2748_mp/globalrelevanceex_sort/39.497087,-76.699762,38.999909,-76.
907473_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/
2 https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-7500
00_price/1832-2748_mp/globalrelevanceex_sort/39.497087,-76.699762,38.999909,-76.
907473_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/
3 https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-7500
00_price/1832-2748_mp/globalrelevanceex_sort/39.497087,-76.699762,38.999909,-76.
907473_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/
4 https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-7500
00_price/1832-2748_mp/globalrelevanceex_sort/39.497087,-76.699762,38.999909,-76.
907473_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/
5 https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-7500
00_price/1832-2748_mp/globalrelevanceex_sort/39.497087,-76.699762,38.999909,-76.
907473_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/
6 https://www.zillow.com/homes/for_sale/fsba,new_lt/4-_beds/2-_baths/500000-7500
00_price/1832-2748_mp/globalrelevanceex_sort/39.497087,-76.699762,38.999909,-76.
907473_rect/12_zm/f915c5936cX1-CR2rtfcdx63zi6_ucu1z_crid/0_mmm/
                                                             Houses Houses_alt
Price Bedrooms Bathrooms
1 https://photos.zillowstatic.com/p_e/IS233336df3r9e0000000000.jpg         NA   $
565,000    5 bds    3.5 ba
2 https://photos.zillowstatic.com/p_e/ISekc7qiyx2kv21000000000.jpg         NA   $
588,600    4 bds      3 ba
3 https://photos.zillowstatic.com/p_e/ISek0i6td01m9u0000000000.jpg         NA   $
550,000    4 bds      3 ba
4 https://photos.zillowstatic.com/p_e/ISyjp8ejp98wab0000000000.jpg         NA   $
574,990    4 bds      4 ba
5 https://photos.zillowstatic.com/p_e/IS27mc6rf6a1gd0000000000.jpg         NA   $
515,000    4 bds      3 ba
6 https://photos.zillowstatic.com/p_e/ISa5qwdt3tcevs0000000000.jpg         NA $7
44,990+    4 bds    3.5 ba
        Size                                  Address
1 2,652 sqft           17 English Elm Ct, Baltimore, MD
2 2,580 sqft          3685 Rogers Ave, Ellicott City, MD
```

```
3 2,244 sqft       4928 Brampton Pkwy, Ellicott City, MD
4 2,800 sqft          10193 Maxine St, Ellicott City, MD
5    -- sqft               9820 Davidge Dr, Columbia, MD
6 3,169 sqft 2435 Vineyard Spring Way, Ellicott City, MD
```

The R file is also attached to this pdf as Conte_J_6C.r.

Getting Tables:

I started a new extractor and entered the URL for the Wikipedia page and import.io automatically put everything from the table into columns, I did not have to do anything, I was quite amazed.



#	Disorder Name 1	Disorder Name 2	Mutation Type 1	Chromos...	Chromos...
1	1p36 deletion syndrome D 1p36	1p36 deletion syndrome	D	1p36	
2	18p deletion syndrome D 18p	18p deletion syndrome	D	18p	
3	21-hydroxylase deficiency 6p21.3	21-hydroxylase deficiency		6p21.3	
4	47,XXX see triple X syndrome C X	triple X syndrome	C	X	
5	47,XXY see Klinefelter syndrome C X	Klinefelter syndrome	C	X	
6	5-ALA dehydratase-deficient porphyria see ALA dehydratase deficiency	ALA dehydratase deficiency			
7	AAT see alpha 1-antitrypsin deficiency 14q32	alpha 1-antitrypsin deficiency		14q32	
8	aceruloplasminemia 3p26.3	aceruloplasminemia		3p26.3	
9	Achondrogenesis type II 12q13.11	Achondrogenesis type II		12q13.11	
10	achondroplasia substitution 4p16.3	achondroplasia	substitution	4p16.3	
11	Acrocephaly see Apert syndrome 10q26.13	Apert syndrome		10q26.13	
12	acute intermittent porphyria	acute intermittent porphyria			
13	adenylosuccinate lyase deficiency	adenylosuccinate lyase deficiency			
14	Adrenoleukodystrophy	Adrenoleukodystrophy			
15	Alagille syndrome	Alagille syndrome			
16	Albinism	Albinism			
17	Alexander disease	Alexander disease			
18	alkaptonuria	alkaptonuria			
19	ALS see amyotrophic lateral sclerosis	amyotrophic lateral sclerosis			

All I had to do was press save and run it. The data saves as a csv file and it's easy to analyze in R, below is the output in R:

```
> # This imports a csv file that contains both numeric and character variable.
> # By default, the data is loaded as a list and data.frame.
> # I also added NA to all blank cells to make it esaier to analyze and
> # stringsAsFactors = FALSE so I can remove levels from the data.
> if (!exists("genetic.import.data")) {
+   genetic.import.data <- read.csv(
+     "genetic_disorders_import.csv",
+     header = TRUE,
+     stringsAsFactors = FALSE,
+     na.strings = c("", "NA"),
+     row.names = NULL,
+     sep = ","
+   )
+ }
>
> class(genetic.import.data)
[1] "data.frame"
> summary(genetic.import.data)
   ï..url             Disorder.Name.1    Disorder.Name.2    Disorder.Name.2_link M
utation.Type.1
 Length:183          Length:183         Length:183         Length:183          L
ength:183
 Class :character   Class :character   Class :character   Class :character     C
lass :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character     M
ode  :character
 Chromosome.1        Chromosome.2   Chromosome.2_link
 Length:183         Mode:logical   Mode:logical  
 Class :character   NA's:183       NA's:183      
 Mode  :character                                
> head(genetic.import.data)
                                                   ï..url
1 https://en.wikipedia.org/wiki/List_of_genetic_disorders
2 https://en.wikipedia.org/wiki/List_of_genetic_disorders
3 https://en.wikipedia.org/wiki/List_of_genetic_disorders
4 https://en.wikipedia.org/wiki/List_of_genetic_disorders
5 https://en.wikipedia.org/wiki/List_of_genetic_disorders
6 https://en.wikipedia.org/wiki/List_of_genetic_disorders
                                                         Disorder.Name.1
Disorder.Name.2
1                                  1p36 deletion syndrome\n \nD\n \n1p36     1p3
6 deletion syndrome
2                                    18p deletion syndrome\n \nD\n \n18p      18
p deletion syndrome
3                                 21-hydroxylase deficiency\n \n \n6p21.3  21-hyd
roxylase deficiency
4                              47,XXX\n see triple X syndrome\n \nC\n \nX        
triple X syndrome
5                           47,XXY\n see Klinefelter syndrome\n \nC\n \nX       K
linefelter syndrome
6 5-ALA dehydratase-deficient porphyria\n see ALA dehydratase deficiency ALA deh
ydratase deficiency
                                      Disorder.Name.2_link Mutation.Type.1 Chrom
osome.1 Chromosome.2
1     https://en.wikipedia.org/wiki/1p36_deletion_syndrome               D
1p36           NA
2      https://en.wikipedia.org/wiki/18p_deletion_syndrome               D
18p           NA
3   https://en.wikipedia.org/wiki/21-hydroxylase_deficiency            <NA>
6p21.3           NA
4          https://en.wikipedia.org/wiki/Triple_X_syndrome               C
X           NA
5       https://en.wikipedia.org/wiki/Klinefelter_syndrome               C
X           NA
6 https://en.wikipedia.org/wiki/ALA_dehydratase_deficiency            <NA>
<NA>           NA
  Chromosome.2_link
1                NA
2                NA
3                NA
4                NA
5                NA
6                NA
```

Conclusion

Web Scraper was a good tool for nothing too complicated. It worked well with getting tables and links off of websites, but it was a little complicated to use at first. Once I watched the tutorials and read some of the documentation it was not too bad. This software works well for sites with linkable text, it can extract that information. However, for sites like Zillow, where the information is embedded in the thumbnail, it is not possible to extract the text, only the links. Overall, I think this is good for basic tasks, nothing too complicated, but it is free which is a bonus.

Import.io is awesome. It is simple to use and it extracts data very easily. For the wiki tables, I did not have to do anything and it automatically extracted everything I needed and information that I did not know I needed like urls of all the diseases (which was an added bonus). I also liked how it would get all of the information from Zillow, I could get the urls, price, number of beds and baths…etc. I was surprised how easy and efficient this program is. The only catch is that it is a little expensive, it starts at $299 ($99 for a student), so unless you plan on using this for a business or a major school project, I think it is a little too expensive for the casual scraper.

Overall, import.io is better. It’s easy to use and gets a lot of data with their features.