Assignment 7

Bhakti Sangoi

February 25, 2018

Libraries

```
library(png)
library(readxl)
```

GitHub

- https://github.com/sangoibhakti/NEU-DA5020.git

Questions

A. (50 Points) Pick at least 2 web scraping toolkits (either automated tools like Import.io or R packages such as rvest) and try to use them to extract data from the Yelp website. In particular, create a search in Yelp to find good burger restaurants in the Boston area. You must try out at least two toolkits, but you will use only one to actually extract and save the full data

Solution

Three web scraping toolkits tried are: 1) Instant Data Scarper 2) Grepsr 3) Import.io

Import.io is a great tool to learn and has lot of options to get an accurate dataset.Hence I continued this assignment using Import.io I have also used Instant Data Scarper because it is easy to use and quite fast.

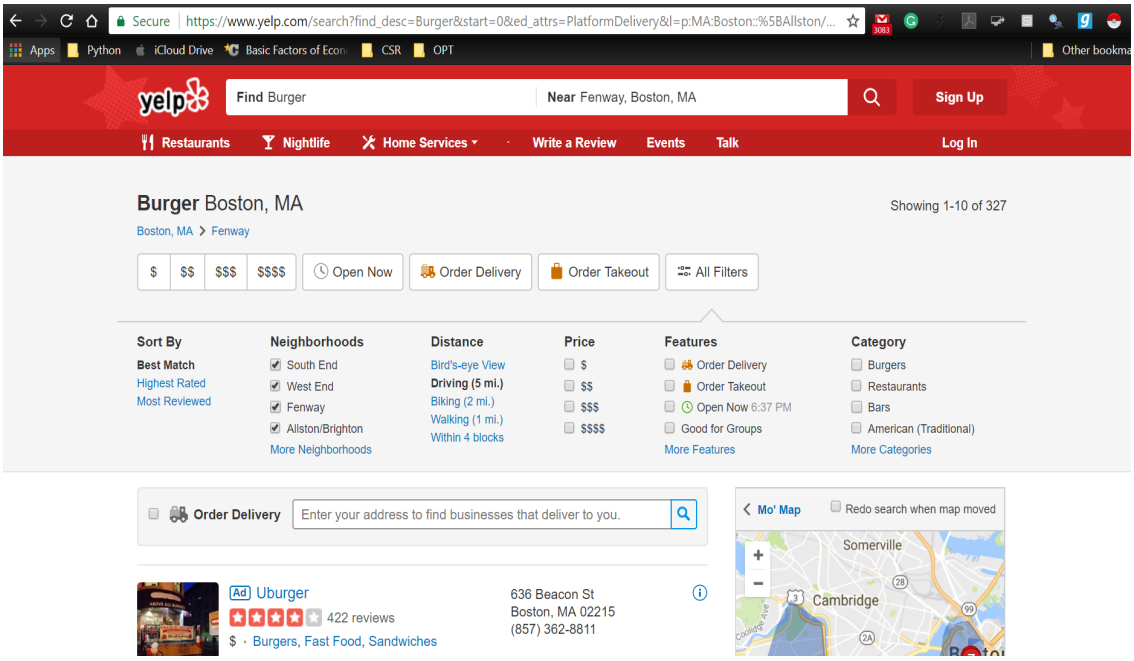
B. (20 points) Import the data you extracted into a data frame in R. Your data frame should have exactly 30 rows, and each row represents a burger restaurant in Boston.

Solution

Using Import.io

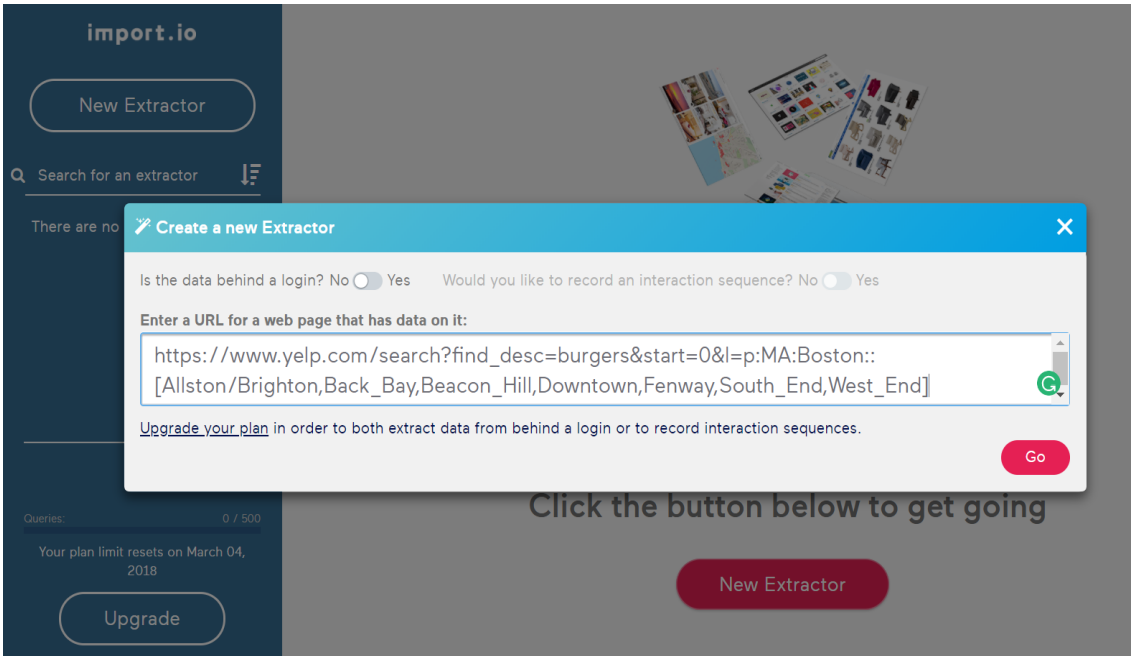
1) Open https://www.yelp.com/boston. Search for Burgers and limit Boston neighborhoods to Allston, Brighton, Back Bay, Beacon Hill, Downtown Area, Fenway, South End, and West End.

```
image1 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/yelp_filter.png")
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(image1,0,0,1,1)
```



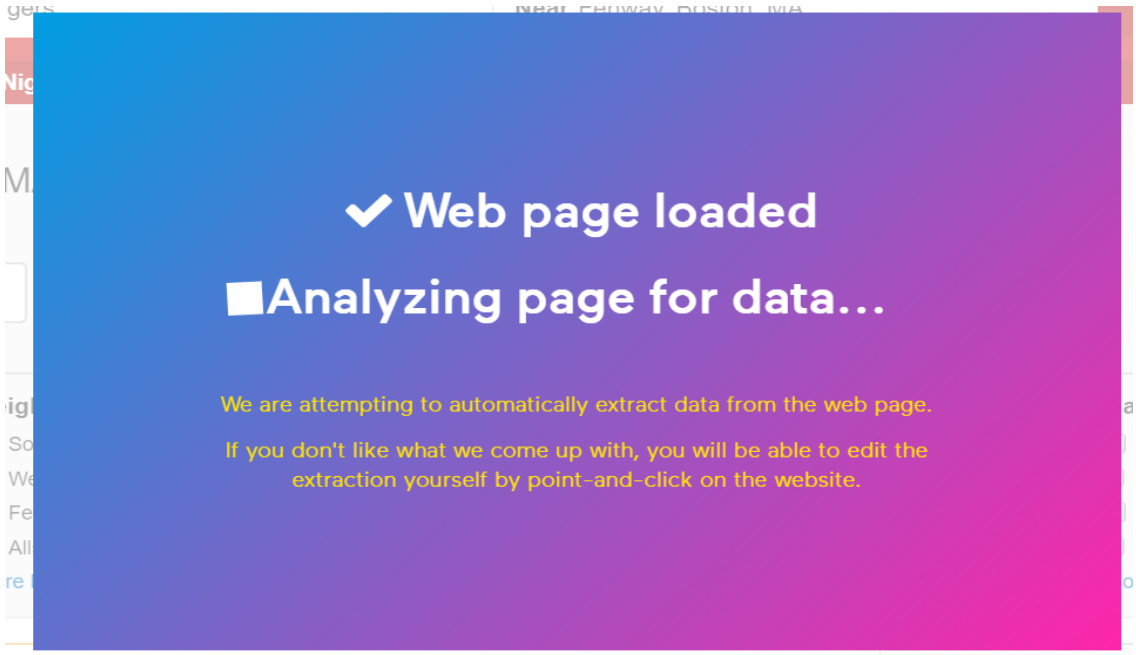
2) Copying the URL in Import.io toolkit

```
import_image2 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/import_
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(import_image2,0,0,1,1)
```



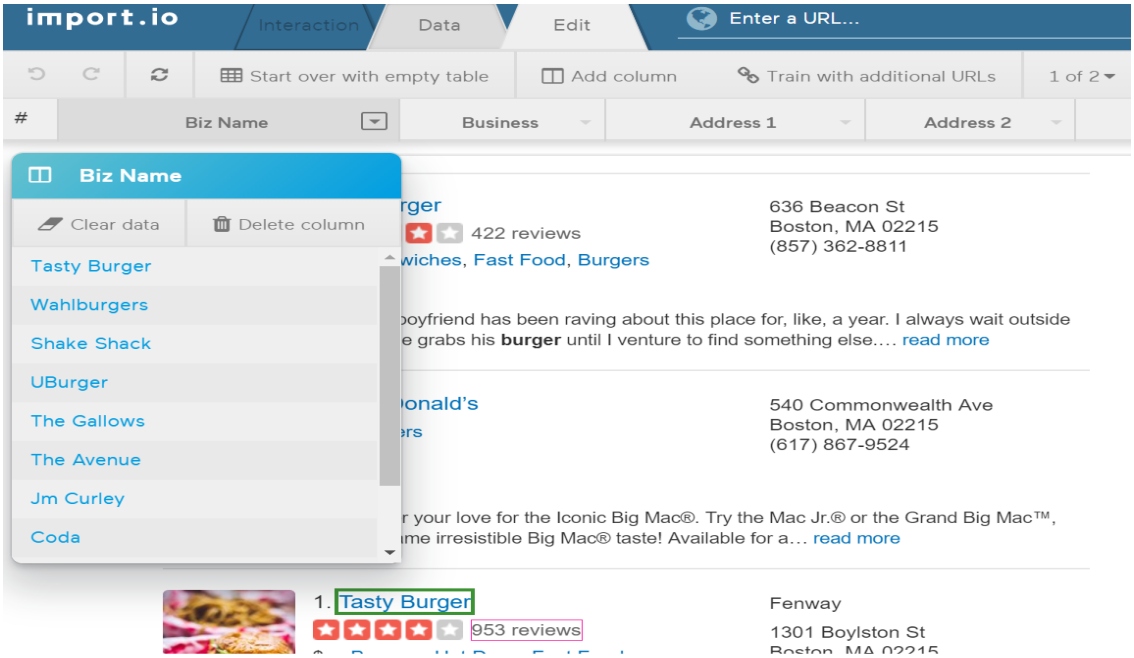
3) Data is loaded into the toolkit

```
import_image3 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/import_
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(import_image3,0,0,1,1)
```



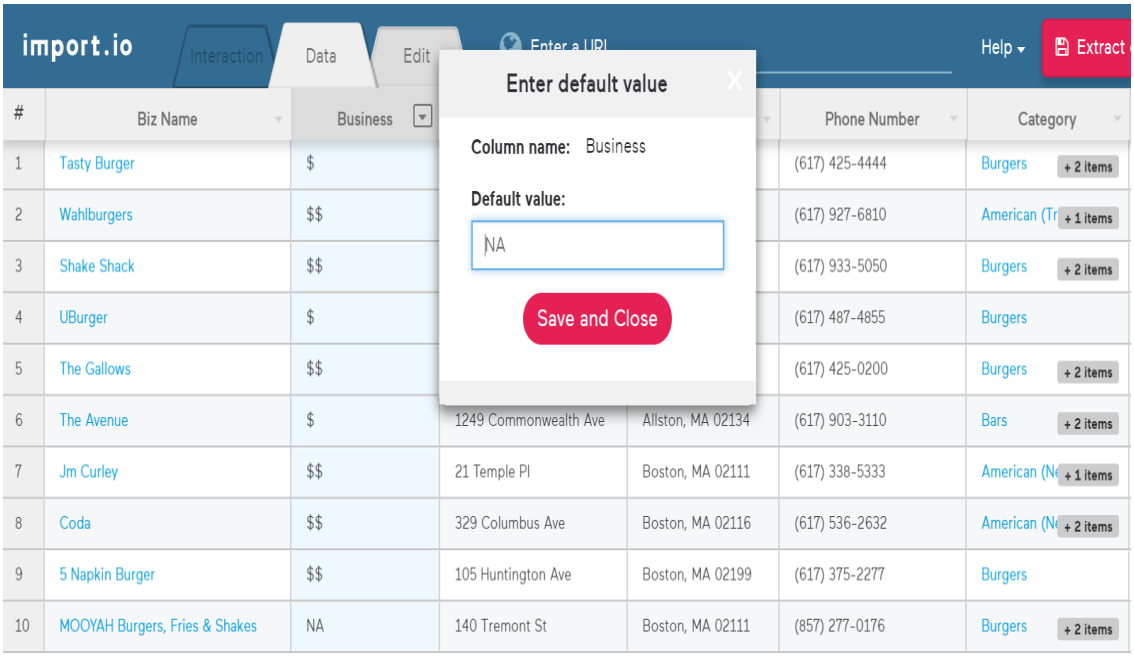
4) Selecting the required fields, deselecting Ads and editing column name

```
import_image4 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/import_
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(import_image4,0,0,1,1)
```



5)Entering NA's

```
import_image5 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/enterin
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(import_image5,0,0,1,1)
```



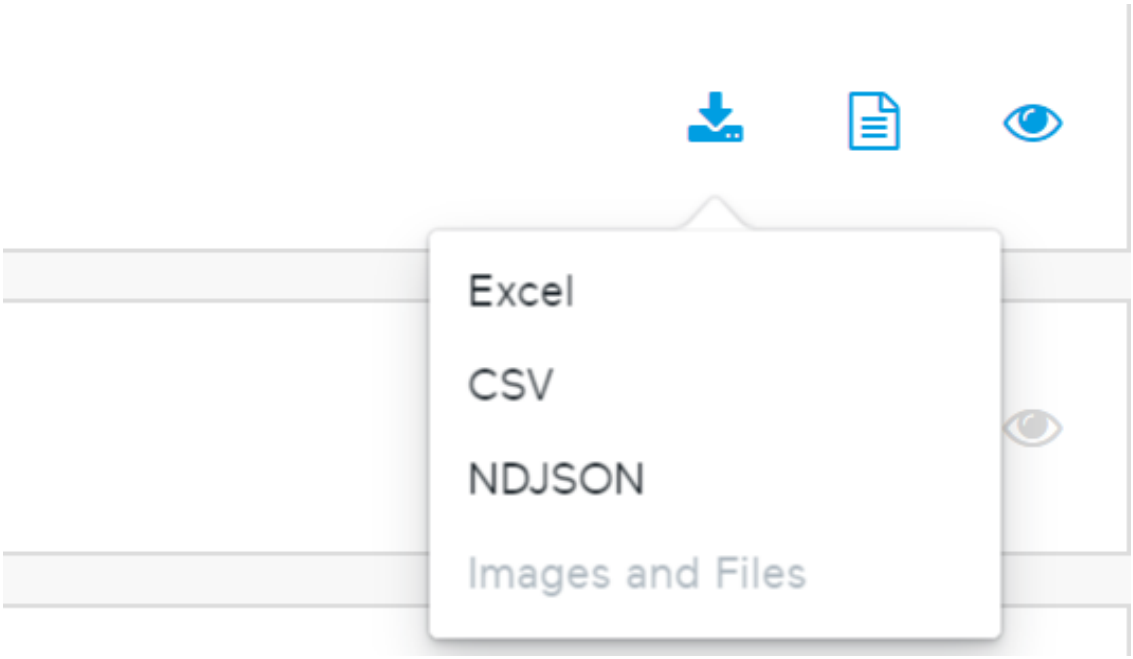
#	Biz Name	Business			Phone Number	Category
1	Tasty Burger	$			(617) 425-4444	Burgers + 2 items
2	Wahlburgers	$$			(617) 927-6810	American (Tr + 1 items
3	Shake Shack	$$			(617) 933-5050	Burgers + 2 items
4	UBurger	$			(617) 487-4855	Burgers
5	The Gallows	$$			(617) 425-0200	Burgers + 2 items
6	The Avenue	$	1249 Commonwealth Ave	Allston, MA 02134	(617) 903-3110	Bars + 2 items
7	Jm Curley	$$	21 Temple Pl	Boston, MA 02111	(617) 338-5333	American (Ne + 1 items
8	Coda	$$	329 Columbus Ave	Boston, MA 02116	(617) 536-2632	American (Ne + 2 items
9	5 Napkin Burger	$$	105 Huntington Ave	Boston, MA 02199	(617) 375-2277	Burgers
10	MOOYAH Burgers, Fries & Shakes	NA	140 Tremont St	Boston, MA 02111	(857) 277-0176	Burgers + 2 items

6)Pagination and Running the job with 3 URL's and having exact 30 rows

```
import_image6 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/import_
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(import_image6,0,0,1,1)
```

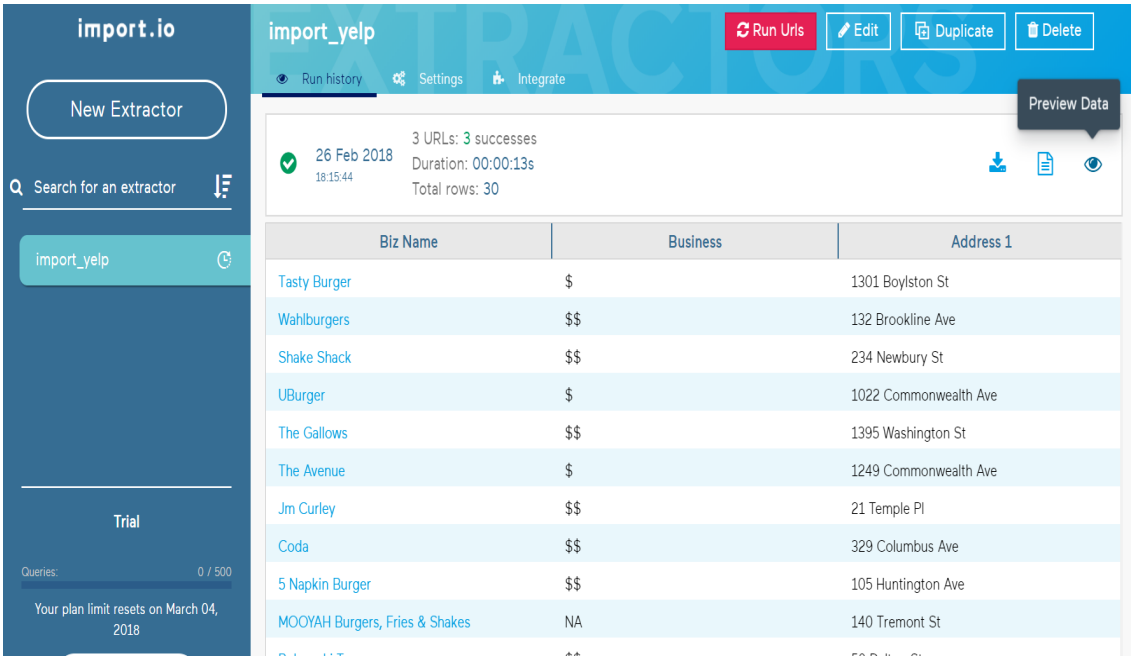
7) Download format

```
import_image7 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/import_
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(import_image7,0,0,1,1)
```



8) View the dataset in Import.io tool

```
import_image8 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/import_
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(import_image8,0,0,1,1)
```



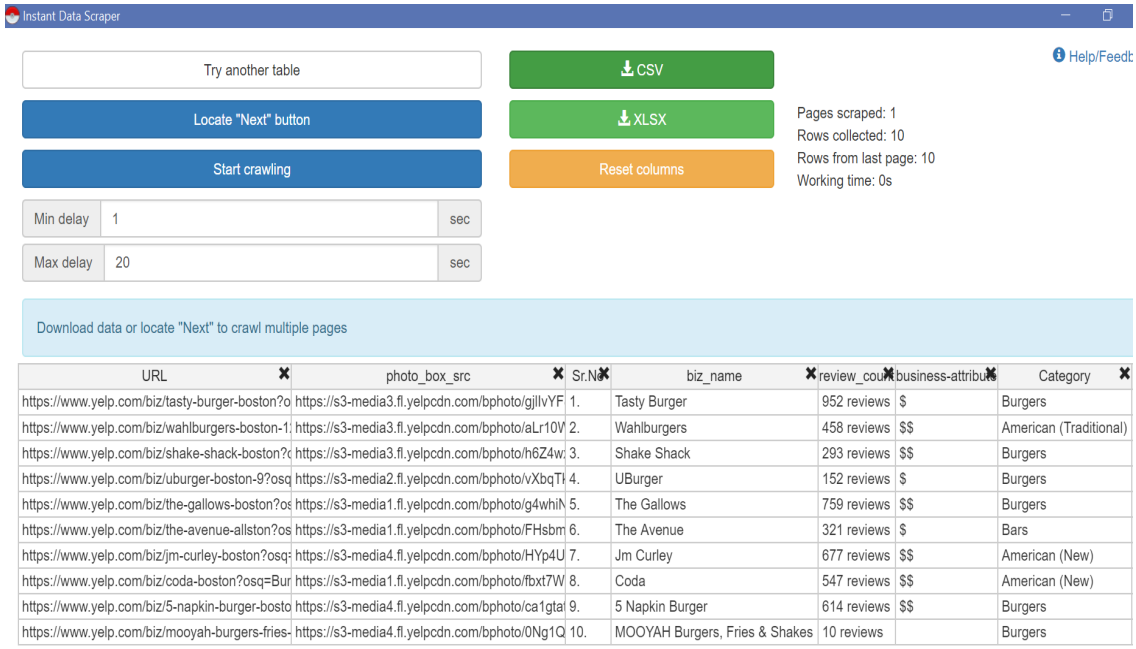
9) Viewing Dataset in R

```
importyelp_dataset <- read.csv("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/import.io/i
View(importyelp_dataset)
```

Using Instant Data Scraper

1) Renaming the column heading and selecting the required columns to be scraped.

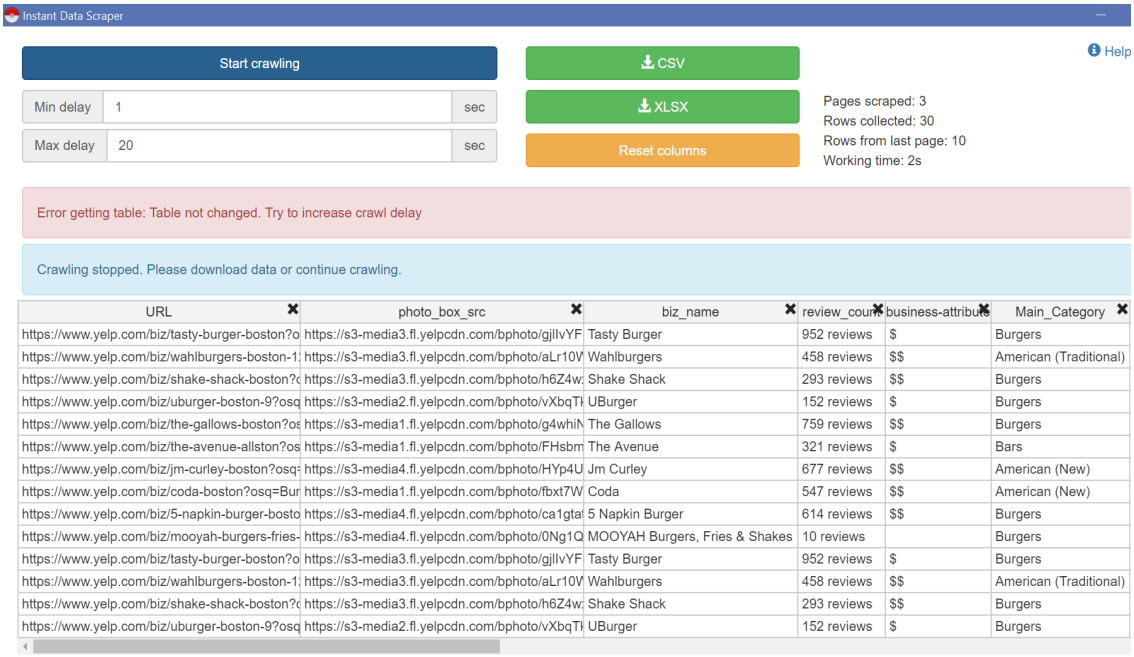
```
image2 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/instant_scraper1.png")
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(image2,0,0,1,1)
```



URL	photo_box_src	Sr.No	biz_name	review_count	business-attributes	Category
https://www.yelp.com/biz/tasty-burger-boston?o	https://s3-media3.fl.yelpcdn.com/bphoto/gjIlvYF	1.	Tasty Burger	952 reviews	$	Burgers
https://www.yelp.com/biz/wahlburgers-boston-1	https://s3-media3.fl.yelpcdn.com/bphoto/aLr10V	2.	Wahlburgers	458 reviews	$$	American (Traditional)
https://www.yelp.com/biz/shake-shack-boston?o	https://s3-media3.fl.yelpcdn.com/bphoto/h6Z4w	3.	Shake Shack	293 reviews	$$	Burgers
https://www.yelp.com/biz/uburger-boston-9?osq	https://s3-media2.fl.yelpcdn.com/bphoto/vXbqTI	4.	UBurger	152 reviews	$	Burgers
https://www.yelp.com/biz/the-gallows-boston?os	https://s3-media1.fl.yelpcdn.com/bphoto/g4whiN	5.	The Gallows	759 reviews	$$	Burgers
https://www.yelp.com/biz/the-avenue-allston?os	https://s3-media1.fl.yelpcdn.com/bphoto/FHsbm	6.	The Avenue	321 reviews	$	Bars
https://www.yelp.com/biz/jm-curley-boston?osq=	https://s3-media4.fl.yelpcdn.com/bphoto/HYp4U	7.	Jm Curley	677 reviews	$$	American (New)
https://www.yelp.com/biz/coda-boston?osq=Bur	https://s3-media1.fl.yelpcdn.com/bphoto/fbxt7W	8.	Coda	547 reviews	$$	American (New)
https://www.yelp.com/biz/5-napkin-burger-bosto	https://s3-media4.fl.yelpcdn.com/bphoto/ca1gta	9.	5 Napkin Burger	614 reviews	$$	Burgers
https://www.yelp.com/biz/mooyah-burgers-fries-	https://s3-media4.fl.yelpcdn.com/bphoto/0Ng1Q	10.	MOOYAH Burgers, Fries & Shakes	10 reviews		Burgers

2) Selecting how many pages to crawl and then scraping 3 pages for 30 rows.

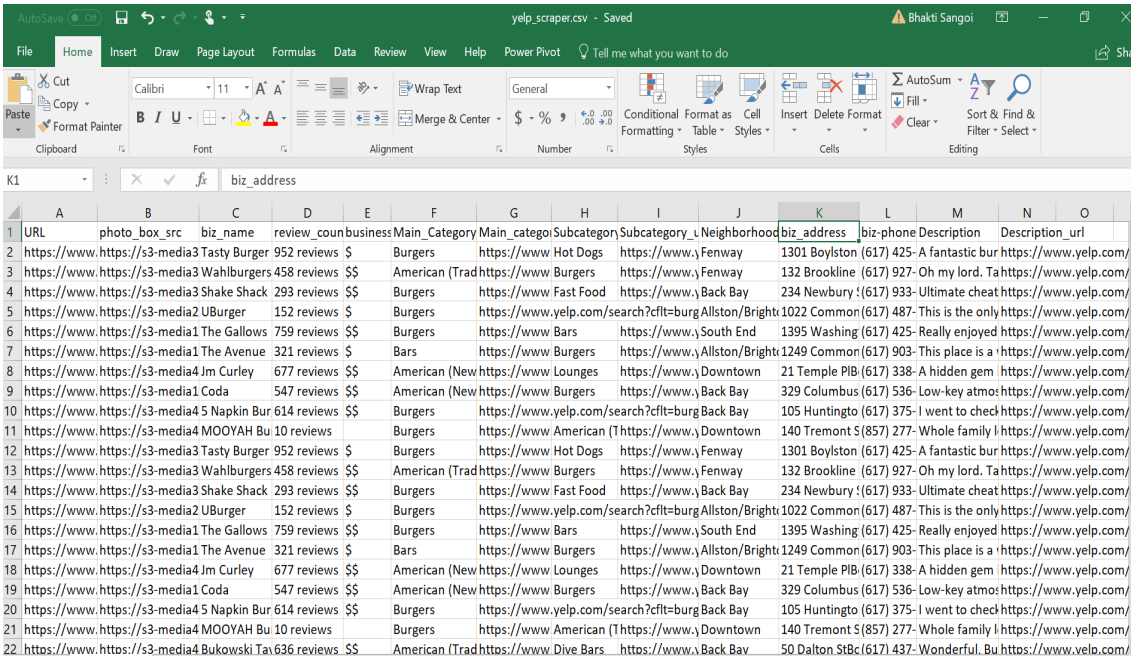
```
image3 <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/instant_scraper2.png")
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(image3,0,0,1,1)
```



URL	photo_box_src	biz_name	review_count	business-attributes	Main_Category
https://www.yelp.com/biz/tasty-burger-boston?o	https://s3-media3.fl.yelpcdn.com/bphoto/gjIlvYF	Tasty Burger	952 reviews	$	Burgers
https://www.yelp.com/biz/wahlburgers-boston-1	https://s3-media3.fl.yelpcdn.com/bphoto/aLr10W	Wahlburgers	458 reviews	$$	American (Traditional)
https://www.yelp.com/biz/shake-shack-boston?c	https://s3-media3.fl.yelpcdn.com/bphoto/h6Z4w	Shake Shack	293 reviews	$$	Burgers
https://www.yelp.com/biz/uburger-boston-9?osq	https://s3-media2.fl.yelpcdn.com/bphoto/vXbqTk	UBurger	152 reviews	$	Burgers
https://www.yelp.com/biz/the-gallows-boston?os	https://s3-media1.fl.yelpcdn.com/bphoto/g4whiN	The Gallows	759 reviews	$$	Burgers
https://www.yelp.com/biz/the-avenue-allston?os	https://s3-media1.fl.yelpcdn.com/bphoto/FHsbm	The Avenue	321 reviews	$	Bars
https://www.yelp.com/biz/jm-curley-boston?osq=	https://s3-media4.fl.yelpcdn.com/bphoto/HYp4U	Jm Curley	677 reviews	$$	American (New)
https://www.yelp.com/biz/coda-boston?osq=Bur	https://s3-media1.fl.yelpcdn.com/bphoto/fbxt7W	Coda	547 reviews	$$	American (New)
https://www.yelp.com/biz/5-napkin-burger-bosto	https://s3-media4.fl.yelpcdn.com/bphoto/ca1gta	5 Napkin Burger	614 reviews	$$	Burgers
https://www.yelp.com/biz/mooyah-burgers-fries-	https://s3-media4.fl.yelpcdn.com/bphoto/0Ng1Q	MOOYAH Burgers, Fries & Shakes	10 reviews		Burgers
https://www.yelp.com/biz/tasty-burger-boston?o	https://s3-media3.fl.yelpcdn.com/bphoto/gjIlvYF	Tasty Burger	952 reviews	$	Burgers
https://www.yelp.com/biz/wahlburgers-boston-1	https://s3-media3.fl.yelpcdn.com/bphoto/aLr10W	Wahlburgers	458 reviews	$$	American (Traditional)
https://www.yelp.com/biz/shake-shack-boston?c	https://s3-media3.fl.yelpcdn.com/bphoto/h6Z4w	Shake Shack	293 reviews	$$	Burgers
https://www.yelp.com/biz/uburger-boston-9?osq	https://s3-media2.fl.yelpcdn.com/bphoto/vXbqTk	UBurger	152 reviews	$	Burgers

3) Downloading in csv format and then Viewing the csv file

```
csv_image <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/yelp_csv.png")
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(csv_image,0,0,1,1)
```



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	URL	photo_box_src	biz_name	review_coun	business	Main_Category	Main_catego	Subcategor	Subcategory_	Neighborhood	biz_address	biz-phone	Description	Description_url	
2	https://www.	https://s3-media3	Tasty Burger	952 reviews	$	Burgers	https://www	Hot Dogs	https://www.	Fenway	1301 Boylston	(617) 425-	A fantastic bur	https://www.yelp.com/	
3	https://www.	https://s3-media3	Wahlburgers	458 reviews	$$	American (Trad	https://www	Burgers	https://www.	Fenway	132 Brookline	(617) 927-	Oh my lord. Ta	https://www.yelp.com/	
4	https://www.	https://s3-media3	Shake Shack	293 reviews	$$	Burgers	https://www	Fast Food	https://www.	Back Bay	234 Newbury S	(617) 933-	Ultimate cheat	https://www.yelp.com/	
5	https://www.	https://s3-media2	UBurger	152 reviews	$	Burgers	https://www.yelp.com/search?cflt=burg			Allston/Bright	1022 Common	(617) 487-	This is the only	https://www.yelp.com/	
6	https://www.	https://s3-media1	The Gallows	759 reviews	$$	Burgers	https://www	Bars	https://www.	South End	1395 Washing	(617) 425-	Really enjoyed	https://www.yelp.com/	
7	https://www.	https://s3-media1	The Avenue	321 reviews	$	Bars	https://www	Burgers	https://www.	Allston/Bright	1249 Common	(617) 903-	This place is a	https://www.yelp.com/	
8	https://www.	https://s3-media4	Jm Curley	677 reviews	$$	American (New	https://www	Lounges	https://www.	Downtown	21 Temple PlB	(617) 338-	A hidden gem	https://www.yelp.com/	
9	https://www.	https://s3-media1	Coda	547 reviews	$$	American (New	https://www	Burgers	https://www.	Back Bay	329 Columbus	(617) 536-	Low-key atmos	https://www.yelp.com/	
10	https://www.	https://s3-media4	5 Napkin Bur	614 reviews	$$	Burgers	https://www.yelp.com/search?cflt=burg			Back Bay	105 Huntingto	(617) 375-	I went to check	https://www.yelp.com/	
11	https://www.	https://s3-media4	MOOYAH Bu	10 reviews		Burgers	https://www	American (T	https://www.	Downtown	140 Tremont S	(857) 277-	Whole family l	https://www.yelp.com/	
12	https://www.	https://s3-media3	Tasty Burger	952 reviews	$	Burgers	https://www	Hot Dogs	https://www.	Fenway	1301 Boylston	(617) 425-	A fantastic bur	https://www.yelp.com/	
13	https://www.	https://s3-media3	Wahlburgers	458 reviews	$$	American (Trad	https://www	Burgers	https://www.	Fenway	132 Brookline	(617) 927-	Oh my lord. Ta	https://www.yelp.com/	
14	https://www.	https://s3-media3	Shake Shack	293 reviews	$$	Burgers	https://www	Fast Food	https://www.	Back Bay	234 Newbury S	(617) 933-	Ultimate cheat	https://www.yelp.com/	
15	https://www.	https://s3-media2	UBurger	152 reviews	$	Burgers	https://www.yelp.com/search?cflt=burg			Allston/Bright	1022 Common	(617) 487-	This is the only	https://www.yelp.com/	
16	https://www.	https://s3-media1	The Gallows	759 reviews	$$	Burgers	https://www	Bars	https://www.	South End	1395 Washing	(617) 425-	Really enjoyed	https://www.yelp.com/	
17	https://www.	https://s3-media1	The Avenue	321 reviews	$	Bars	https://www	Burgers	https://www.	Allston/Bright	1249 Common	(617) 903-	This place is a	https://www.yelp.com/	
18	https://www.	https://s3-media4	Jm Curley	677 reviews	$$	American (New	https://www	Lounges	https://www.	Downtown	21 Temple PlB	(617) 338-	A hidden gem	https://www.yelp.com/	
19	https://www.	https://s3-media1	Coda	547 reviews	$$	American (New	https://www	Burgers	https://www.	Back Bay	329 Columbus	(617) 536-	Low-key atmos	https://www.yelp.com/	
20	https://www.	https://s3-media4	5 Napkin Bur	614 reviews	$$	Burgers	https://www.yelp.com/search?cflt=burg			Back Bay	105 Huntingto	(617) 375-	I went to check	https://www.yelp.com/	
21	https://www.	https://s3-media4	MOOYAH Bu	10 reviews		Burgers	https://www	American (T	https://www.	Downtown	140 Tremont S	(857) 277-	Whole family l	https://www.yelp.com/	
22	https://www.	https://s3-media4	Bukowski Ta	636 reviews	$$	American (Trad	https://www	Dive Bars	https://www.	Back Bay	50 Dalton StBc	(617) 437-	Wonderful. Bu	https://www.yelp.com/	

4)Reading the csv file

```
yelp_dataset <- read.csv("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/yelp_scraper.csv")
View(yelp_dataset)
```

C. (30 Points) Write a report that compares the tools with a focus on cost, ease of use, features, and your recommendation. Discuss your experience with the tools and why you decided to use the one you picked in the end. Use screenshots of toolkits and your scraping process to support your statements. Also include a screenshot or an excerpt of your data in the report.

Solution

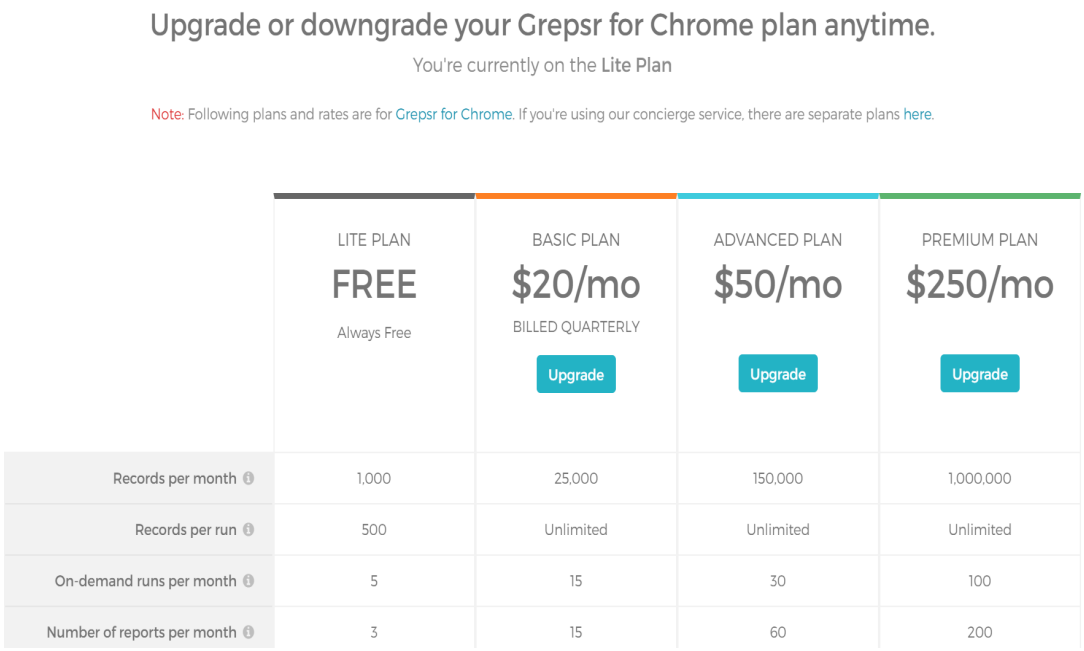
Import.io is the tool available on net whereas Instant Data Scraper and Grepsr both are found in Google chrome extension. All the 3 tools has different ways of using it.

Grepsr Toolkit

Cost:

It can be downloaded for free. It has different monthly plans. It has free plan which helps in creating 3 free reports per month.

```
grepsr_cost <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/grepsr_cost.png")
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(grepsr_cost,0,0,1,1)
```



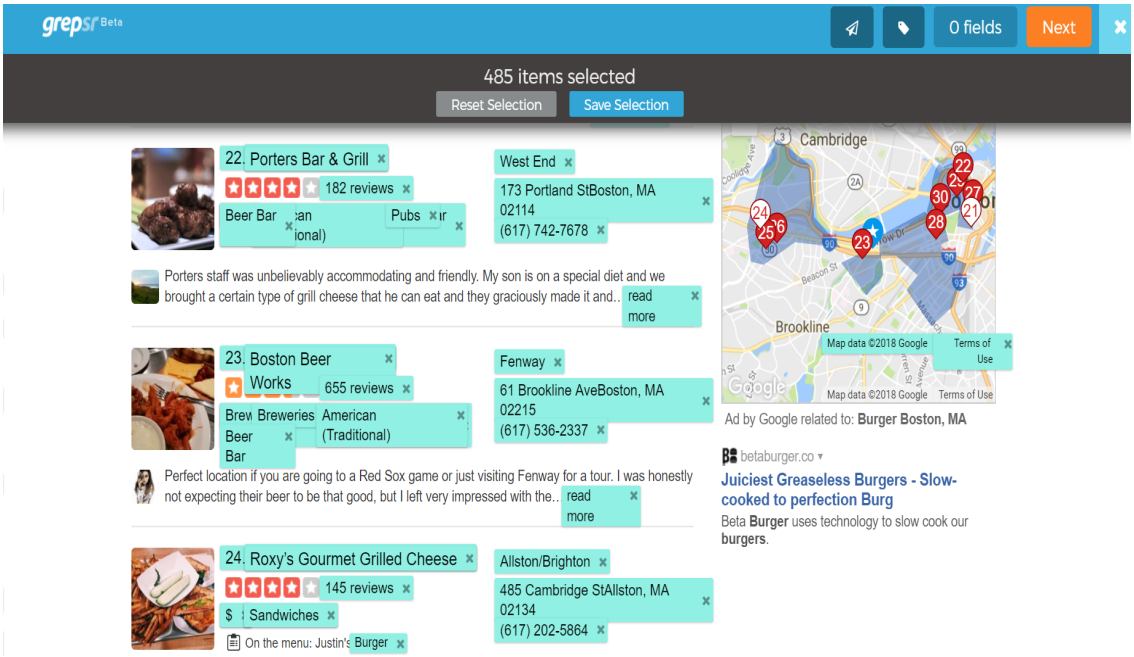
Ease of Use:

It is readily available. It is a quick google chrome extension. Also has a tour guide.

Features:

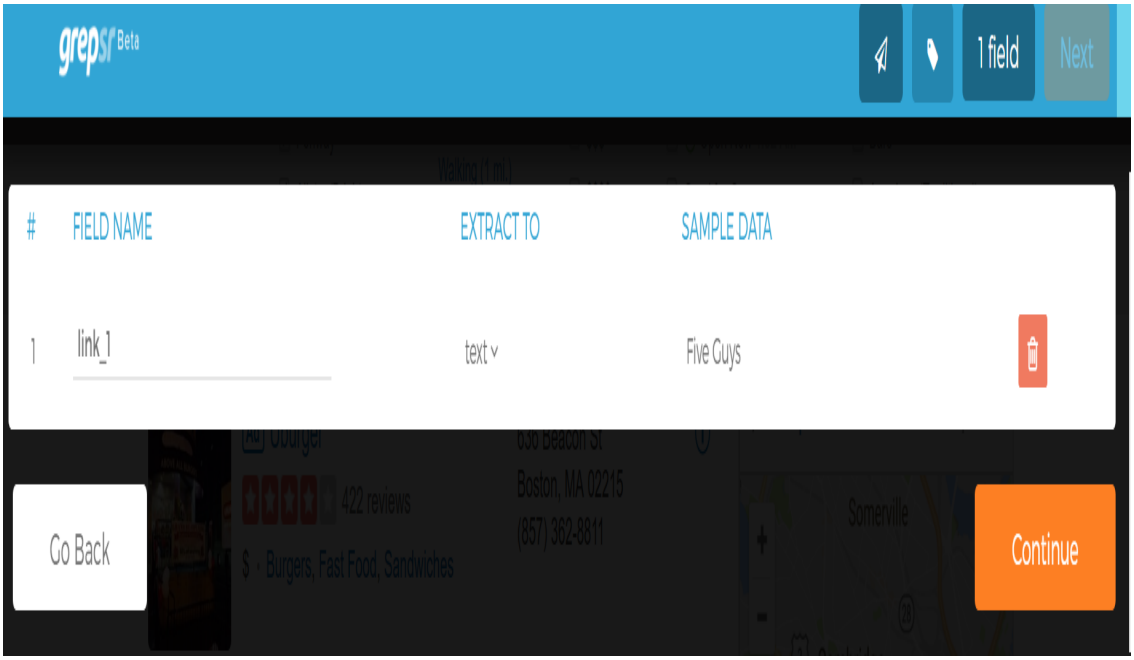
Once you select the page you want to scrap, you first select the tags that need to be extracted. Further it has different options of pagination like “Next link”“,”infinite scroll“” and “load more button”. Then you can extract and download the data in different formats such as csv,JSON, XML, Excel formats There are different downloading options by sending it via dropbox, google drive, dropbox. Scrap the data and then group it accordingly.

```
grepsr_selection <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/grepsr_selecti
grepsr_pagination <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/grepsr_pagina
grepsr_fileformat <- readPNG("C:/Users/sango/Documents/Desktop/R/Assignments/Assignment 7/grepsr_filefo
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(grepsr_selection,0,0,1,1)
```




```
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(grepsr_pagination,0,0,1,1)
```


```
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(grepsr_fileformat,0,0,1,1)
```



Instant Data Scarper

Cost

It can be downloaded for free. Activate the extension. There is no cost to it.

Ease of Use

It is readily available for free. It is very user friendly and all the features are easy to understand and use.

Features

Select the page you want to scrap, click on instant data scraper google extension. Extension will guess where the data is. Edit the column heading. There is an option to try another table button to guess again. To scrap another page, there is an option for "Locate next" Then start crawling the number of pages you want to scrap. Delete the unwanted fields anytime during scraping. Download it in csv or excel.

Import.io

Cost

It has a 7 day trial. Then It has different monthly/yearly plans.

Ease of Use

It has an online dashboard and reports to scrap the data of the url you want to. It has a lot of various features for advance data scraping. It is easy to understand and use. It is a very flexible tool.

Features

It has a feature of selecting whatever is required. Creating new columns as per our requirement. One can also add NA's in blanks which helps in cleaning up the dataset. You can also Combine columns which is similar to doing binding in R. It has an option of adding more URL to add data in same dataset after running the job once. It can be extracted in csv, excel and json.

Comparing

- Import.io has lot of features which is easy to use and gives a better result as required by the user. It can be done in few clicks. It's a perfect tool for students to learn Data Scraping.
- Grepsr is little time consuming in selecting what tags are needed for scraping. It was many different format options to save data. But pagination is tricky to understand. Both are google chrome extensions.
- Instand Data scraper is user friendly and very simple for beginners. It guesses the data not very well. There aren't many options as compared to import.io to edit your rows and columns. Can't insert NA's in blanks in both grepsr and instant data scraper. It can remove unwanted columns easily but not add columns.

Recommendation

- Grepsr should have better documentation or tour guide on how to use. It should be more user friendly. It should improve its options to scrap more pages together. Also it should have some guessing crawler data options which becomes faster and easy for user.
- Instand Data scraper is for beginners. Should have an option to add the columns atleast which gives better result as the user wants

D. (10 points) Within your report describe what you have derived about the URL for yelp pages. What are the differences between the three URLs? What are the parameters that determined your search query (Boston burger restaurants in 8 selected neighborhoods)? What is(are) the parameter(s) used for pagination? Without opening Yelp.com in the browser, what is your guess of the URL for the 7th page of Chinese restaurants in New York?

3 URL's used for scraping first three pages:

1) https://www.yelp.com/search?find_desc=Burger&start=0&l=p:MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_H
2) https://www.yelp.com/search?find_desc=Burger&start=10&l=p:MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_
3) https://www.yelp.com/search?find_desc=Burger&start=20&l=p:MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_

Difference between above three URL's

Difference between all the three links is the start tag. in case of page 1: start=0 ,in case of page 2: start=10 and in case of page 3: start=20.

Parameters

In the above URL, 3 parameters are seen. 1) "find_desc: Burger": This is the category of food served. It means finding restaurants that serve burger. 2) "start=0"= Page 1,"start=10"= Page 2, : This is Pagination. It leads to the page you have requested for. It can also mean "start=0" i.e. Page 1 has 10 search. next 10 are found in page 2 and next 10 in page 3. 3) MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_Hill,Downtown,Fenway,South_End,West_End: Filter on location. Finding restaurants which are only located in the above neighborhood.

URL for the 7th page of Chinese restaurants in New York.

https://www.yelp.com/search?find_desc=Chinese&start=60&l=p:NY:New_York