Week7

Vinisha

October 22, 2017

```
#Loading Libraries
library(rvest)

## Loading required package: xml2

library(magrittr)
library(tidyverse)

## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------
-

## filter(): dplyr, stats
## lag():    dplyr, stats

library(stringr)

# reading the HTML with all select parameters from Search results page 1


webpage1 <-
read_html("https://www.yelp.com/search?find_desc=burgers&start=0&l=p:MA:Bosto
n::%5BAllston/Brighton,Back_Bay,Beacon_Hill,Downtown,Fenway,South_End,West_En
d%5D")


#Getting the name of the places
name_page1 <- webpage1 %>% html_nodes(".indexed-biz-name span") %>%
html_text()
name_page1 <- as.data.frame(name_page1)

#Getting the location of the places

location_page1 <- webpage1 %>% html_nodes(".natural-search-result
.neighborhood-str-list") %>% html_text()
location_page1 <- as.data.frame(location_page1)
```

```
#getting the location of the page
address_page1 <- webpage1 %>% html_nodes(".natural-search-result address")
%>% html_text()
address_page1 <- as.data.frame(address_page1)

# getting the contact number
contact_page1 <- webpage1 %>% html_nodes(".natural-search-result .biz-phone")
%>% html_text()
contact_page1 <- as.data.frame(contact_page1)


# getting price range

price_page1 <- webpage1 %>% html_nodes(".natural-search-result .price-range")
%>% html_text()
price_page1 <- as.data.frame(price_page1)


# getting rating
rating_page1 <- webpage1 %>% html_nodes(".natural-search-result .rating-
large") %>% html_text()
rating_page1 <- as.data.frame(rating_page1)



#getting main review
review_page1 <- webpage1 %>% html_nodes(".natural-search-result .snippet")
%>% html_text()
review_page1 <- as.data.frame(review_page1)

# Combining results to form a single data frame
page1 <- as.data.frame(cbind(name_page1,location_page1,address_page1,
contact_page1, price_page1, rating_page1, review_page1, make.row.names = TRUE
))

#removing "\n" from entire data set
page1[] <- lapply(page1, gsub, pattern='\n', replacement='')


knitr::kable(head(page1[1:5,1:4]))
```

name_page1	location_page1	address_page1	contact_page1
Tasty Burger	Fenway	1301 Boylston StBoston, MA 02215	(617) 425-4444
Roast Beast	Allston/Brighton	1080 Commonwealth AveBoston, MA 02215	(617) 877-8690

The Avenue	Allston/Brighton	1249 Commonwealth AveAllston, MA 02134	(617) 903-3110
The Gallows	South End	1395 Washington StBoston, MA 02118	(617) 425-0200
Coda	Back Bay	329 Columbus AveBoston, MA 02116	(617) 536-2632

```
knitr::kable(head(page1[1:5,5:8]))
```

price_page1	rating_page1	review_page1	make.row.names
$		I can't say much about Tasty Burger - not because my experience was bad but because I am at a loss for words! I ordered the rise n' shine burger (with chipotle mayo) side of tater read more	TRUE
$		Wow! What a find. So glad that I tried this gem. I ordered the Sammy5 and was blown away. I had to order a second sandwich for lunch the next day. GREAT sandwiches - wonderful read more	TRUE
$		Burger were juicy & flavorful. Avocado fries were lightly breaded & there were a ton! Usually you get 4-5 slices of avocado but this was like 2 dozen. And the crispy cheese curds. read more	TRUE
$$		The food here was amazing. We had and few appetizers. all were good,but the Corn is a star and a must have when you eat here. Main course. the fried chicken was the best I ever read more	TRUE
$$		Same owners as the Salty Pig and just around the corner, you can tell that both of the places really want to do a good job with their food and drinks. Our bartender was great and read more	TRUE

```
webpage2 <-
read_html("https://www.yelp.com/search?find_desc=burgers&start=10&l=p:MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_Hill,Downtown,Fenway,South_End,West_End%5D")


#Getting the name of the places
name_page2 <- webpage2 %>% html_nodes(".indexed-biz-name span") %>%
html_text()
name_page2 <- as.data.frame(name_page2)
```

```r
#Getting the location of the places

location_page2 <- webpage2 %>% html_nodes(".natural-search-result
.neighborhood-str-list") %>% html_text()
location_page2 <- as.data.frame(location_page2)

#getting the location of the page
address_page2 <- webpage2 %>% html_nodes(".natural-search-result address")
%>% html_text()
address_page2 <- as.data.frame(address_page2)

# getting the contact number
contact_page2 <- webpage2 %>% html_nodes(".natural-search-result .biz-phone")
%>% html_text()
contact_page2 <- as.data.frame(contact_page2)


# getting price range

price_page2 <- webpage2 %>% html_nodes(".natural-search-result .price-range")
%>% html_text()
price_page2 <- as.data.frame(price_page2)


# getting rating
rating_page2 <- webpage2 %>% html_nodes(".natural-search-result .rating-
large") %>% html_text()
rating_page2 <- as.data.frame(rating_page2)


# getting review number
ratingnumber_page2 <- webpage2 %>% html_nodes(".rating-qualifier") %>%
html_text()
ratingnumber_page2 <- as.data.frame(ratingnumber_page2)


#getting main review
review_page2 <- webpage2 %>% html_nodes(".natural-search-result .snippet")
%>% html_text()
review_page2 <- as.data.frame(review_page2)


# Combining results to form a single data frame
page2 <- as.data.frame(cbind(name_page2,location_page2,address_page2,
contact_page2, price_page2, rating_page2, review_page2, make.row.names = TRUE
))
```

```
#removing "\n" from entire data set
page2[] <- lapply(page2, gsub, pattern='\n', replacement='')



knitr::kable(head(page2[1:5,1:4]))
```

name_page2	location_page2	address_page2	contact_page2
Tasty Burger	Back Bay	145 Dartmouth StBoston, MA 02116	(617) 425-4444
Corner Tavern	Back Bay	421 Marlborough StBoston, MA 02115	(617) 262-5555
B. Good	Back Bay	665 Boylston StBoston, MA 02116	(617) 927-8333
Saltie Girl	Back Bay	281 Dartmouth StBoston, MA 02116	(617) 267-0691
B Good	Back Bay	131 Dartmouth StBoston, MA 02116	(617) 424-5252

```
knitr::kable(head(page2[1:5,5:8]))
```

price_page2	rating_page2	review_page2	make.row.names
$		We were pleasantly surprised with the food, as we are not from the East Coast and have never heard of Tasty Burger. My husband ordered the Mushroom Burger and said that it is the read more	TRUE
$$		This place was everything we were hoping for. It wasn't too expensive and not too cheap. The ambiance was good. The high tops were a little uncomfortable but we got used to it. What read more	TRUE
$		Healthy and fast! I work in Copley and strolled over here on a hot Saturday afternoon. I was in the mood for something quick and refreshing, so I decided to give this spot a try. read more	TRUE
$$$		In a city with top notch restaurants at every corner, saltie girl manages to make it to the top of our favorites list. We chose nearly every single meal of our one week trip read more	TRUE
$		I love b good! Healthy food, fast, at a good price. I also like that it's a local Boston business and committed to sourcing food locally. My favorite thing to get is the kale read more	TRUE

```
webpage3 <-
read_html("https://www.yelp.com/search?find_desc=burgers&start=20&l=p:MA:Bost
on::%5BAllston/Brighton,Back_Bay,Beacon_Hill,Downtown,Fenway,South_End,West_E
```

```
nd%5D")


#Getting the name of the places
name_page3 <- webpage3 %>% html_nodes(".indexed-biz-name span") %>%
html_text()
name_page3 <- as.data.frame(name_page3)

#Getting the location of the places

location_page3 <- webpage3 %>% html_nodes(".natural-search-result
.neighborhood-str-list") %>% html_text()
location_page3 <- as.data.frame(location_page3)

#getting the location of the page
address_page3 <- webpage3 %>% html_nodes(".natural-search-result address")
%>% html_text()
address_page3 <- as.data.frame(address_page3)

# getting the contact number
contact_page3 <- webpage3 %>% html_nodes(".natural-search-result .biz-phone")
%>% html_text()
contact_page3 <- as.data.frame(contact_page3)


# getting price range

price_page3 <- webpage3 %>% html_nodes(".natural-search-result .price-range")
%>% html_text()
price_page3 <- as.data.frame(price_page3)


# getting rating
rating_page3 <- webpage3 %>% html_nodes(".natural-search-result .rating-
large") %>% html_text()
rating_page3 <- as.data.frame(rating_page3)


# getting review number
ratingnumber_page3 <- webpage3 %>% html_nodes(".rating-qualifier") %>%
html_text()
ratingnumber_page3 <- as.data.frame(ratingnumber_page3)


#getting main review
review_page3 <- webpage3 %>% html_nodes(".natural-search-result .snippet")
%>% html_text()
review_page3 <- as.data.frame(review_page3)
```

```
# Combining results to form a single data frame
page3 <- as.data.frame(cbind(name_page3,location_page3,address_page3,
contact_page3, price_page3, rating_page3, review_page3, make.row.names = TRUE
))

#removing "\n" from entire data set
page3[] <- lapply(page3, gsub, pattern='\n', replacement='')



knitr::kable(head(page3[1:5,1:4]))
```

name_page3	location_page3	address_page3	contact_page3
The Glenville Stops	Allston/Brighton	87 Glenville AveBoston, MA 02134	(617) 903-3638
Five Guys	Downtown	58 Summer StBoston, MA 02110	(617) 482-2244
Boston Beer Works	Fenway	61 Brookline AveBoston, MA 02215	(617) 536-2337
The Tip Tap Room	Beacon Hill	138 Cambridge StBoston, MA 02114	(857) 350-3344
Roxy□s Gourmet Grilled Cheese	Allston/Brighton	485 Cambridge StAllston, MA 02134	(617) 202-5864

```
knitr::kable(head(page3[1:5,5:8]))
```

price_page3	rating_page3	review_page3	make.row.names
$$		I underrated this pace. Perhaps it's gotten better over the years or maybe my taste has matured. I'll admit it, I was wrong - It's damn good. Beer selection; wine selection; food the□ read more	TRUE
$		Always a trusty spot around Downtown Crossing in Boston to take the kids to eat. They have clean bathrooms (protected by a code on the receipt) and a changing table. Basic American□ read more	TRUE
$$		Stopped here after a game at Fenway. The restaurant was pretty crowded, but we got seated right away and immediately had our drink orders taken. I ordered the seasonal flight, which□ read more	TRUE
$$		I walk by here all the time and it always seems busy, no matter what day or time.	TRUE

	There are always a ton of people standing/hanging out on the bar side and the tables are usually read more	
$	Food: Out of this world. I had the vegan buffalo tofu melt with a beer and it was amazing. Messy, delicious, "third date" food. The cheese was salty and the tofu was spicy. Yum!! My read more	TRUE

Observations using CSS tool from Google Chrome ; Selector Gadget and R Package "rvest"

URLs

https://www.yelp.com/boston

This URL describes the website Yelp.com for Boston area

https://www.yelp.com/search?find_desc=burgers&find_loc=Boston%2C+MA&ns=1

In this URL we are narrowing down our search to "Burgers" places in "Boston"

https://www.yelp.com/search?find_desc=burgers&start=0&l=p:MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_Hill,Downtown,Fenway,South_End,West_End%5D

In This URL we have filtered our search for specific loactions, Allston/Brighton, Back Bay, BeaconHill, Downtown, Fenway, South End WestEnd

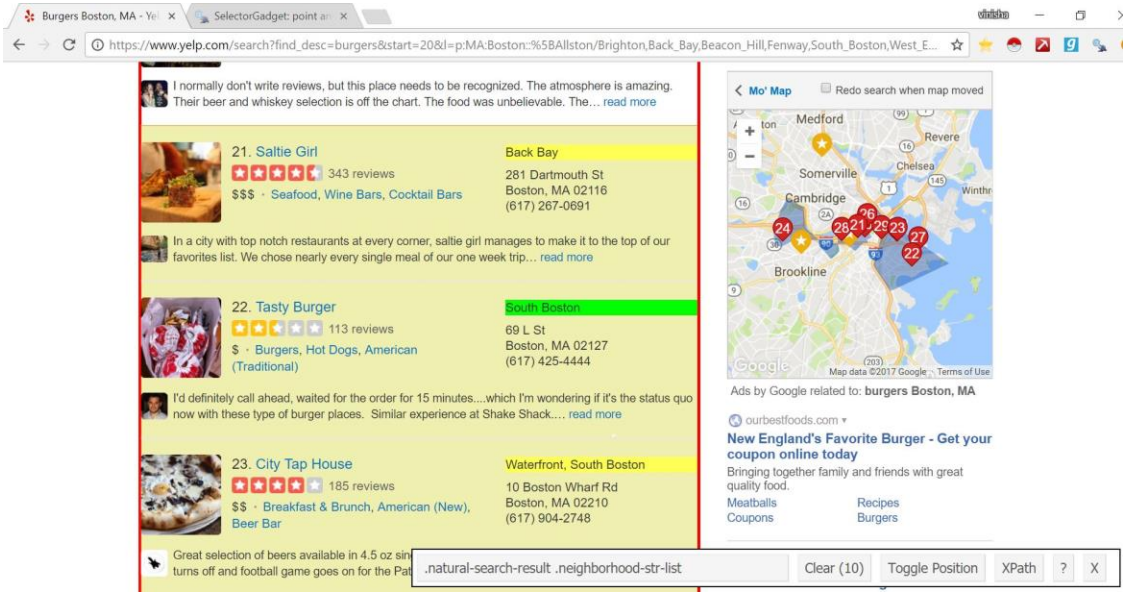
1. I notice to represent black space in "South End" the link recognizezs space with "_"
2. The search is described as find_desc = burgers (search for burgers), start=0&1 (starting with page 1), in MA, Boston. With areas separated by ","

https://www.yelp.com/search?find_desc=burgers&start=10&l=p:MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_Hill,Downtown,Fenway,South_End,West_End%5D

When we navigate to the 2nd page the change in the URL is start=10&1 that means the first page had 10 results and the new page will start with the 11th result and go up to 20

Using CSS Selector and Rvest

1. To scrap data from the web I used the Selector gadget (CSS) that helped me identify the css selector for that portion of the webpage.
2. Using the rvest package , first read the entire webpage using read_html().
3. Once we read the entire webpage we need to extract specific information, for that we use the selector gadget that helps in identifying the exact css selector to be used
4. Using html_nodes we can then scrape just the selected part of the webpage and converted it into html_text.
5. Then stored the data in the form of a dataframe using data.frame



Selector Gadget

The advantages of this tool and rvest is that the technique is quite simple however, we will have to extract information from each page and for each section of the webpage.

If multiple selectors are selected from the page, data for each observation is stored in just one column. Hence, to get data in more cleaner format, first data frame of single columns were made and then used a cbind to combine the columns into a single data frame.

Also we will have to first store the data of multiple pages into multiple data frames and then append into a single data frame.

PartD

Guessing the URl for 7th page of chinese restaurants in new york

https://www.yelp.com/search?find_desc=chinese_restaurants&find_loc=New_York%2C+NY&start=70

```
# scraped result using Instant Data Scraper

yelp <- as.data.frame(read.csv("Yelp_CSV_Instant_Data_Scraper.csv"))

knitr::kable(head(yelp, 5))
```

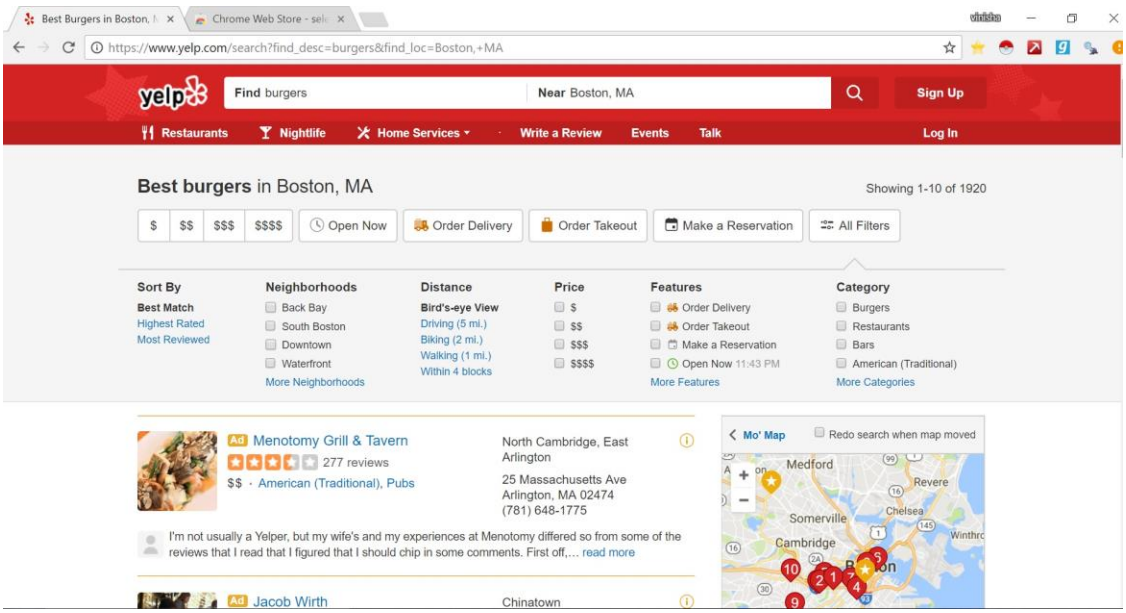
indexed.biz.name	biz.name	review.count	business.attribute	category.str.list	category.str.list.2	category.str.list.3	category.str.list.4	neighborhood.str.list	secondary.attributes	biz.phone	snippet
1	Tasty Burger	920 reviews	$	,							
	, Burgers	Hot	Dogs Fa	st Food Fenway	1	301 Boylston StBoston,	MA 02215 (617	) 425-4444 I can't say mu	ch about Tasty Burger - not because my ex	perience was bad	but because I am at a loss for words! Â I ordered the risen' shine burger (with chipotle may

											o) side of tater â¦
2	The Avenue , Bars	308 reviews Burgers	$ Sandwiches	, Allston/Brighton	249 Commonwealth AveAllston, MA 02134	(617) 903-3110	Burger were juicy & flavorful. Avocado fries were light	breaded & their	e were a ton! Usually you get 4-5 slices of avocado but this was like 2 dozen. And the crispy cheese curds.â¦
3	Roast Beast , Sandwiche	411 reviews Burgers	$ Delis	, Allston/Brighton	080 Commonwealth	ston, MA 02215	) 877-8690 Wow! What a	nd. So glad that I tried	red the Sam	d was blown

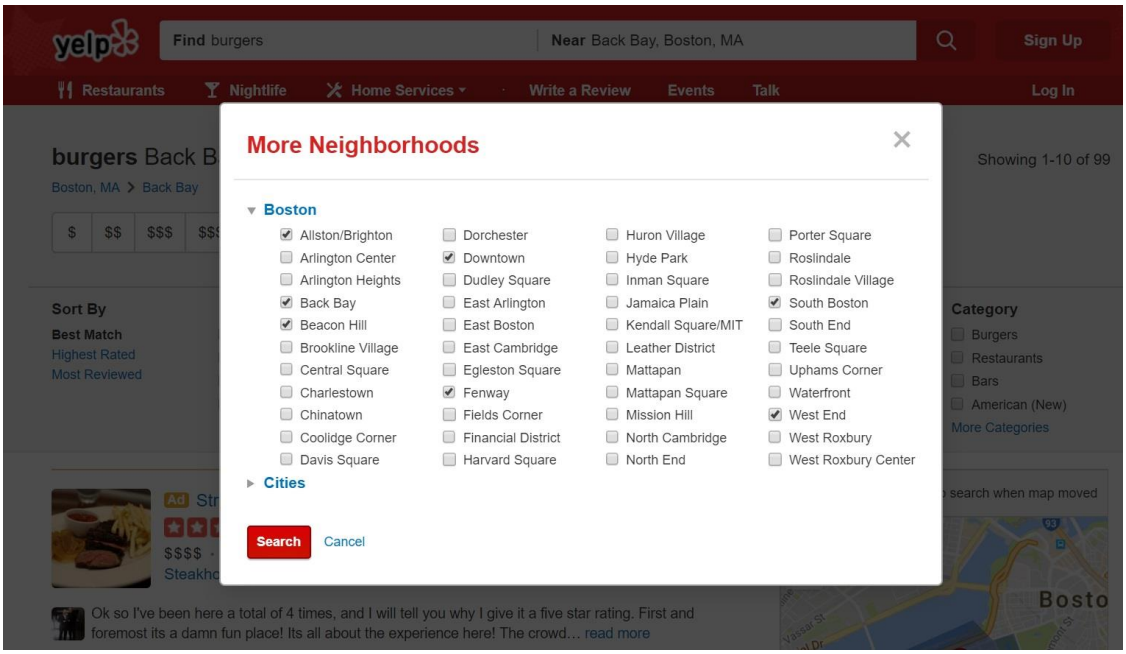
	s				h AveBo	(617	fi	this gem. I orde	my 5 an	away. I had to order a second sandwich for lunch the next day. Â GREAT sandwiches - wonderfulâ¦
4	Coda	534 reviews	$$	,						
	, American (	New) Bur	gers Co	cktail Bars Back B	ay 3	29 Columbus AveBoston,	MA 02116 (617	) 536-2632 Same owners as	the Salty Pig and just around the corner	, you can tell th	at both of the places really want to do a good job with their

											food and drinks. Â Our bartender was great and â¦
5	Drink	1397 reviews	$$$	,							
	, Lounges	American	(New) Bu	rgers Waterf	ront, South Boston 3	48 Congress StBoston,	MA 02228 (617	) 695-1806 This is probab	ly one of the most unique bars I've ever	been to even for	a speakeasy-esque bar. The place has no drink menu, basically u tell the bartender what u feel

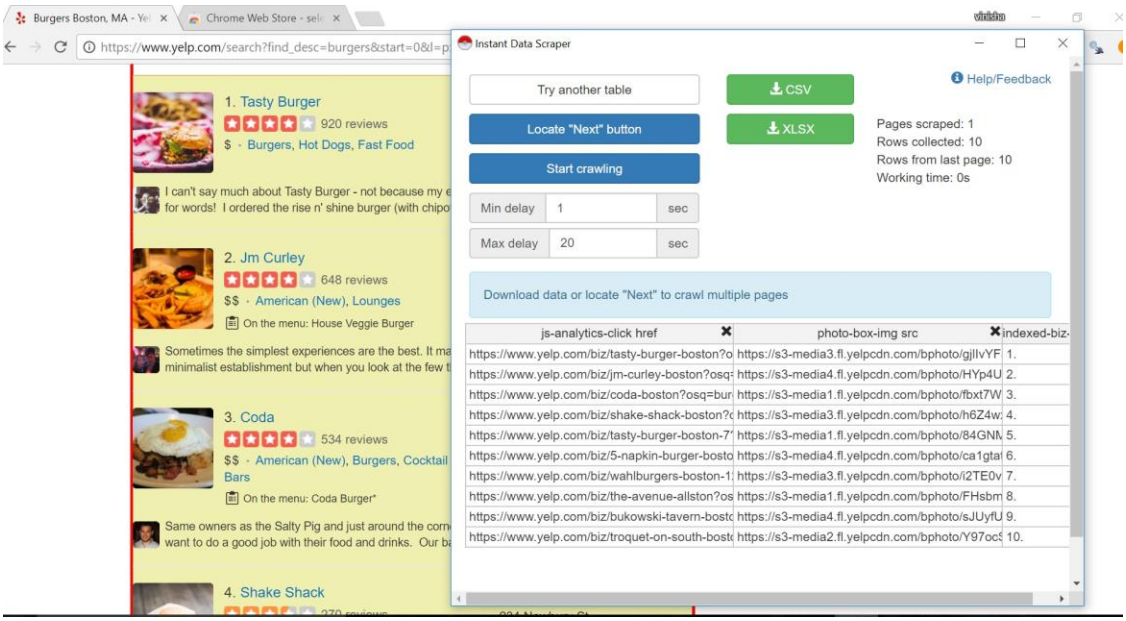
like
drin
king,
â¦



Search Screen at Yelp Website

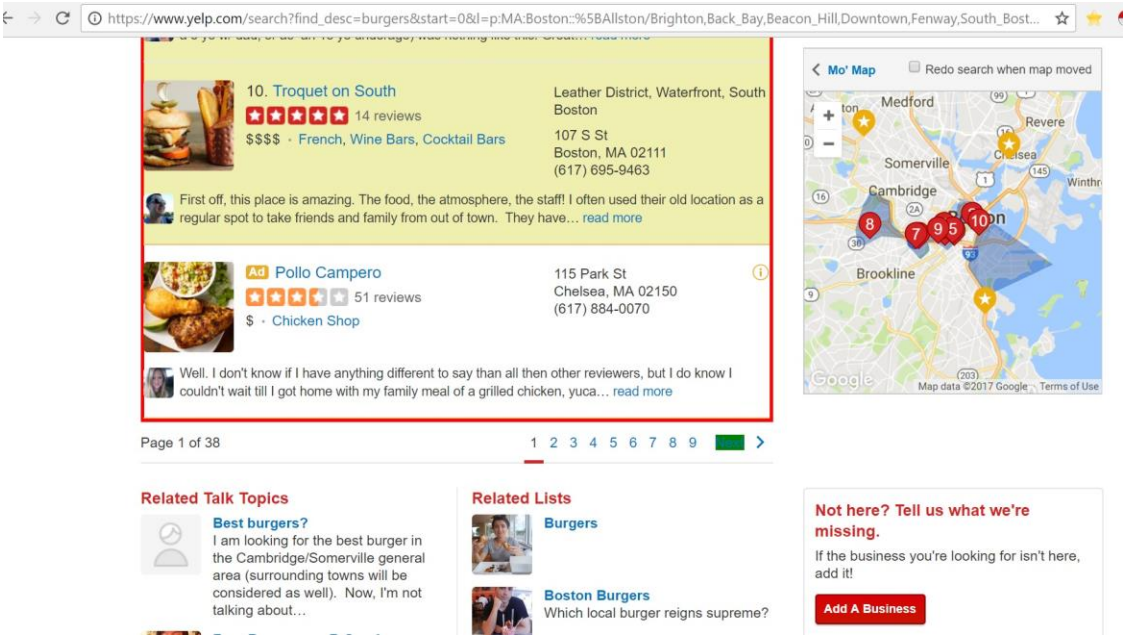


Adding Filter with respect to Location

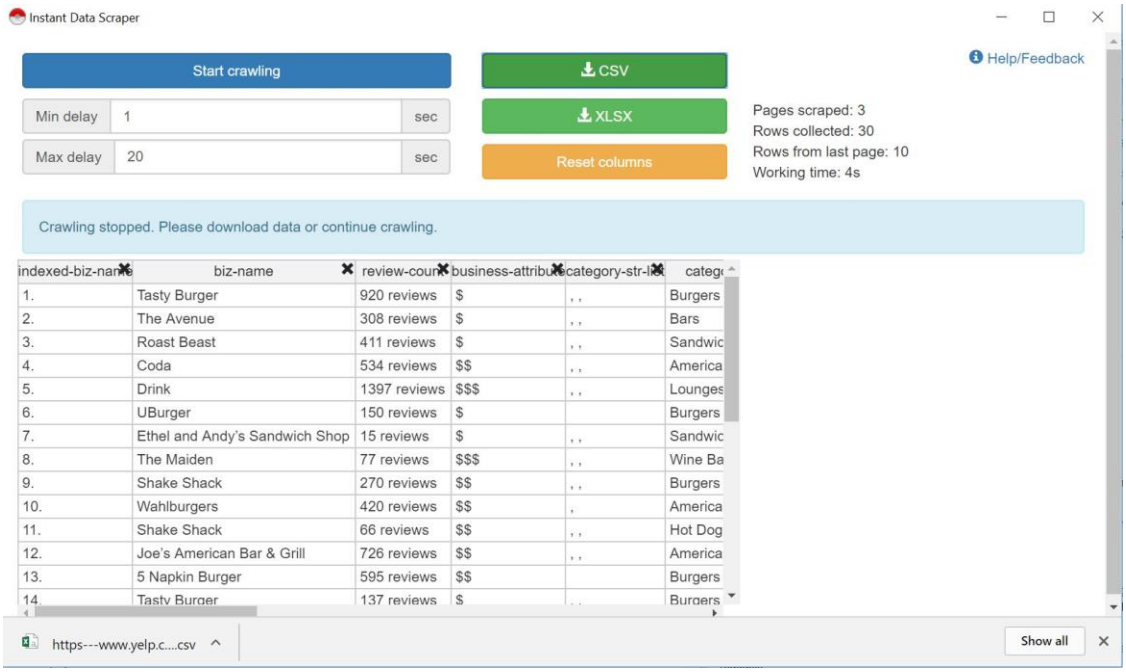


Launching the Instant data scraper Tool

Here we can see the format of the columns and the information that is being fetched from the entire web page. we have the liberty to take just the information that we need and discard the rest before scraping. Also there is option to locate the next button, which helps in fetching multiplt plages at once.



Locating the NEXT Button



start Crawling and Save as csv/xlsx format

Observations using Google Chrome Web Scraping Tool ; Instant Data Scraper

1 The advantage of using Instant webscraper took is that is extracts all the data into a single csv format file.

2 Navigating to multiple pages and stopping the scrawling at the required number of pages was an added feature to the tool.

3 However, there was a lot of junk data that came along with it as well. But there is provision of deleting the uncessary variable extracted from the web page

4 As a part of the unnecessary data, I noticed that the variables like Name, Category, Location had an added variable column of the http links associated with those individual variables.

5 These added variables if not removed in the final csv could require cleaning for the data to be come readable.

Overall Observation

The ease of each Tool depends on the requirement as each tool has its own functionality. The tools are used are open source as the cost associated with using these tools is nil. But time as a cost factor is highly relative on the user requirment.