Web Scraping

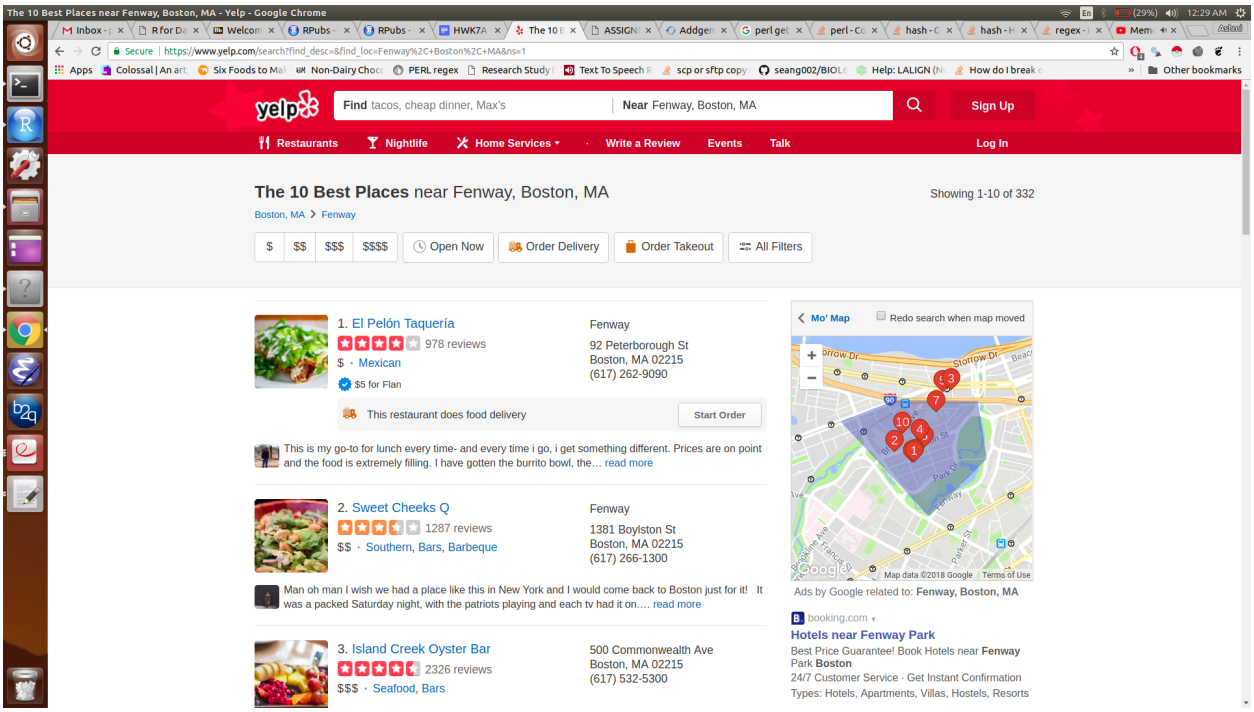
Ashmi

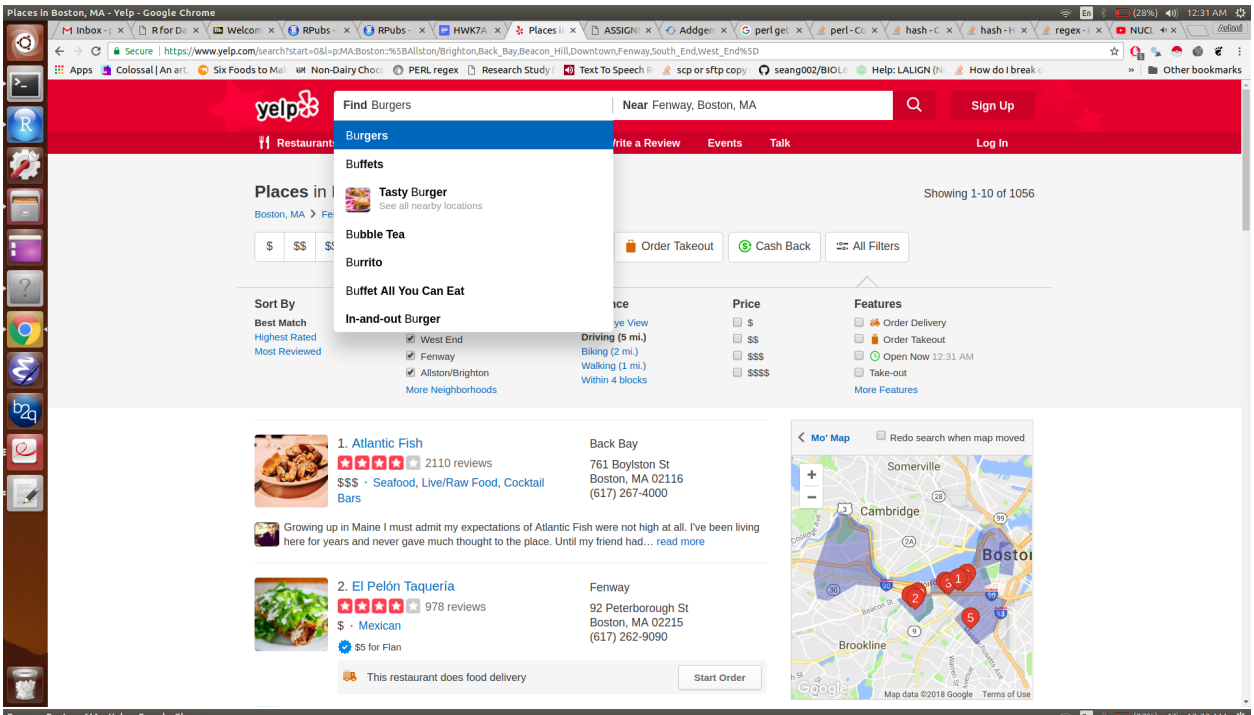
February 24, 2018

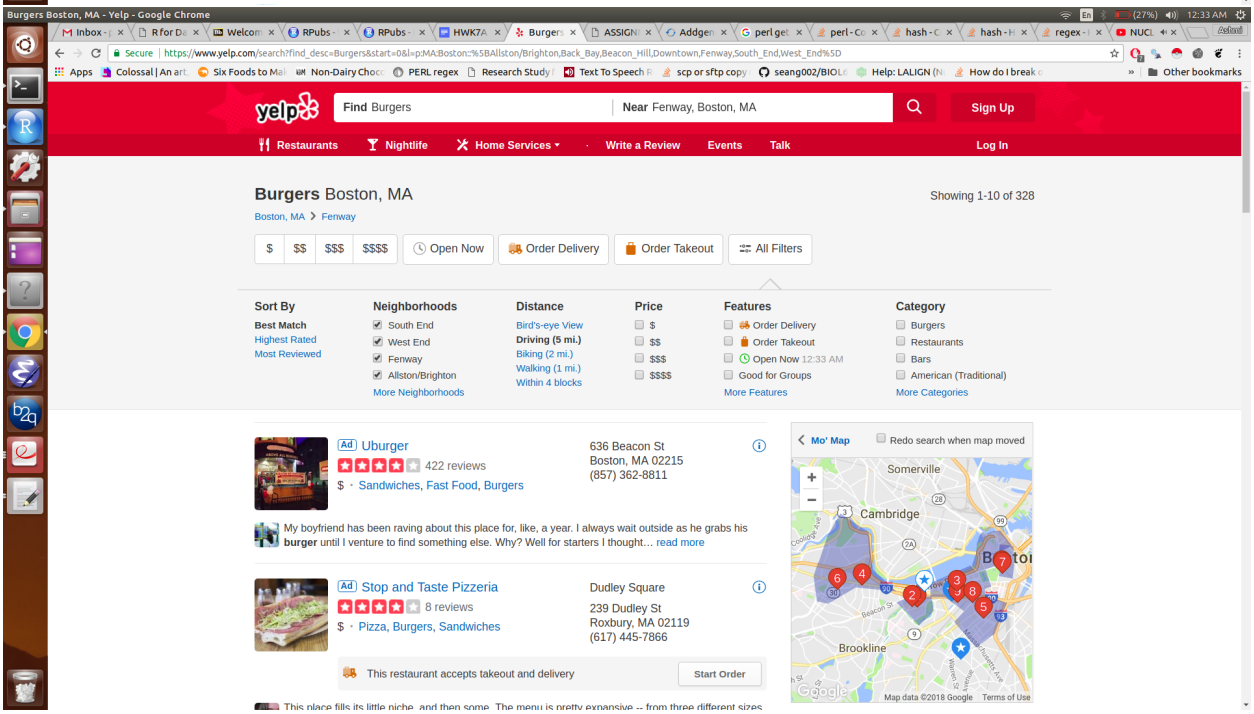
A. (50 Points) Pick at least 2 web scraping toolkits (either automated tools like Import.io or R packages such as rvest) and try to use them to extract data from the Yelp website. In particular, create a search in Yelp to find good burger restaurants in the Boston area. You must try out at least two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

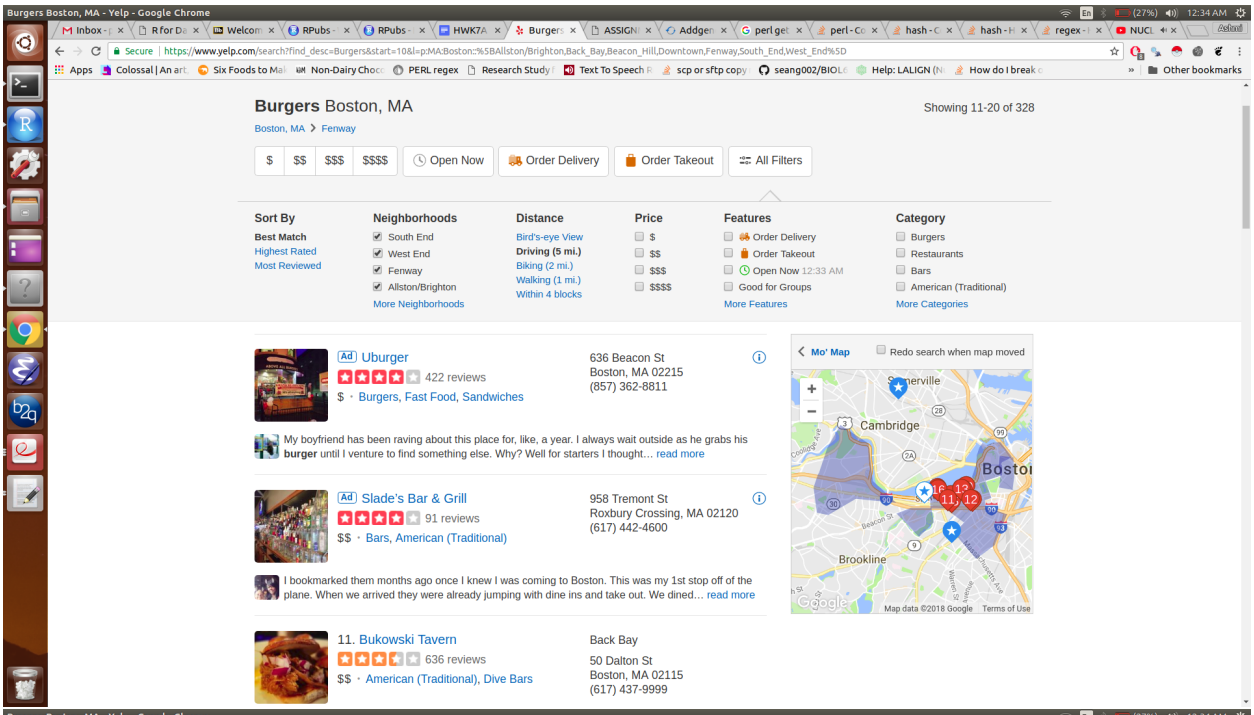
I used two web scraping toolkits: 1. Instant Data Scraper (chrome extension) 2. Rvest and SelectorGadget

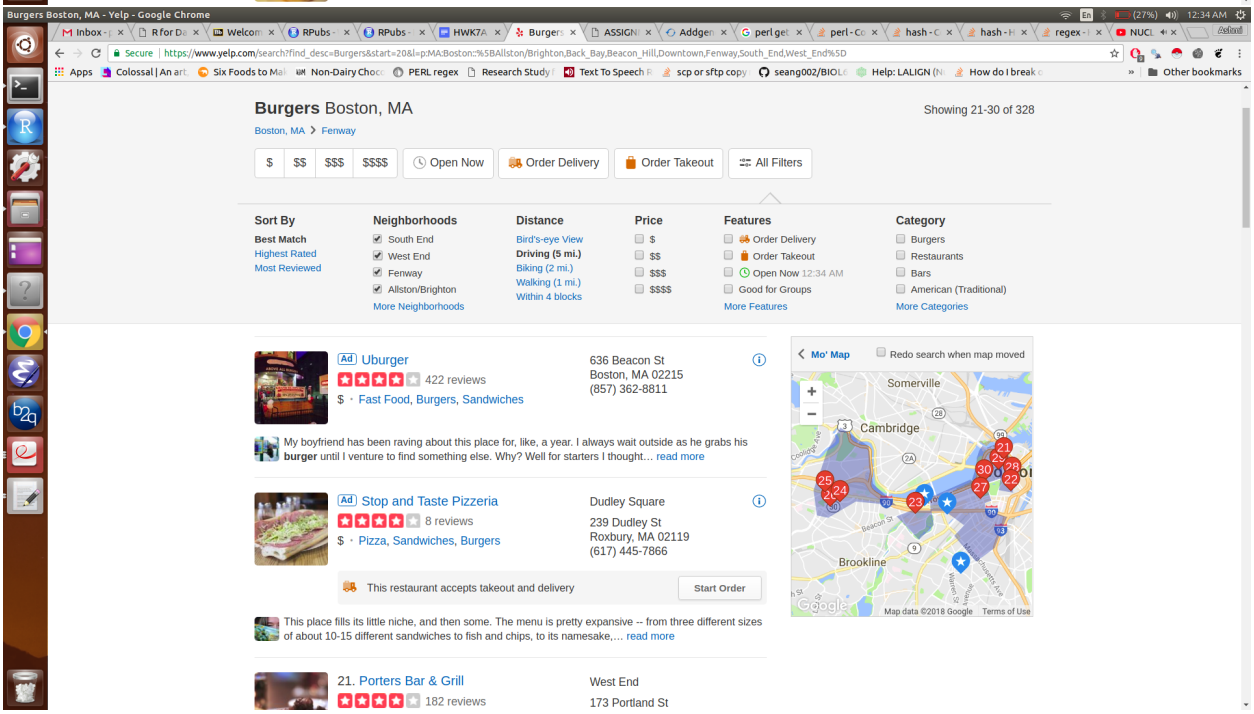
```
library(knitr)
knitr::include_graphics(c('yelp_default_restaurant_list.png','yelp_filter_burger.png','yelp_result_page
```

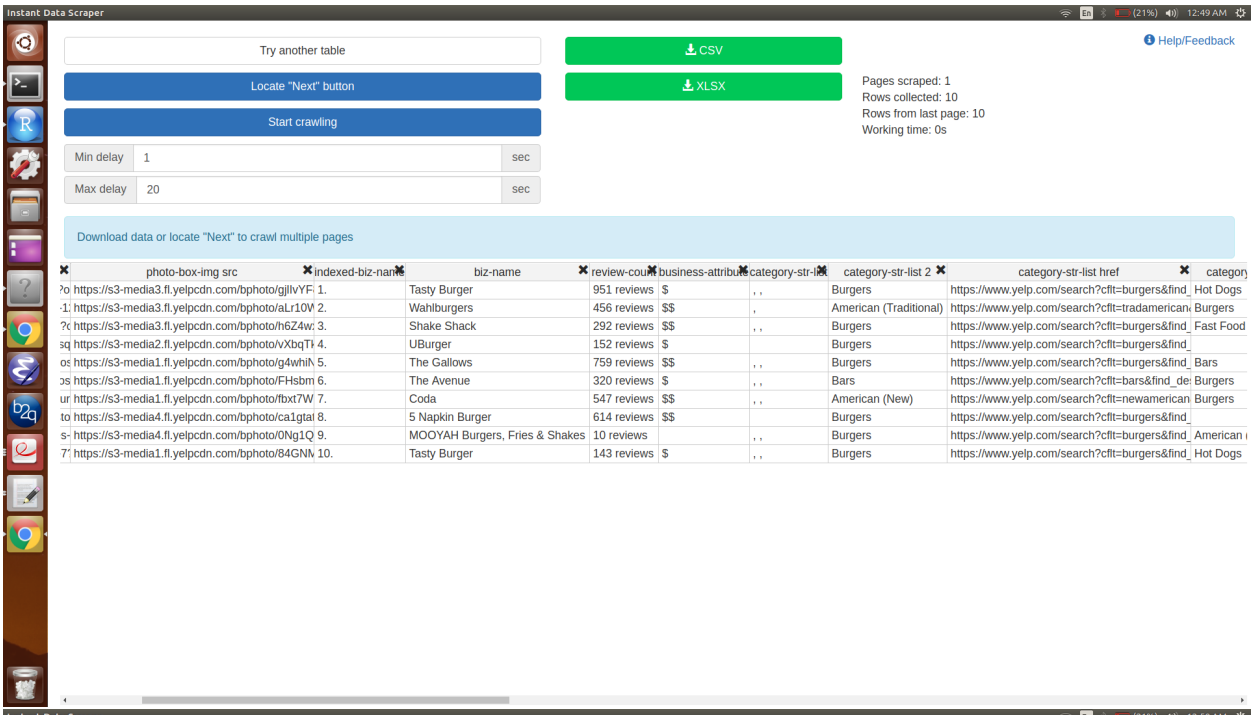


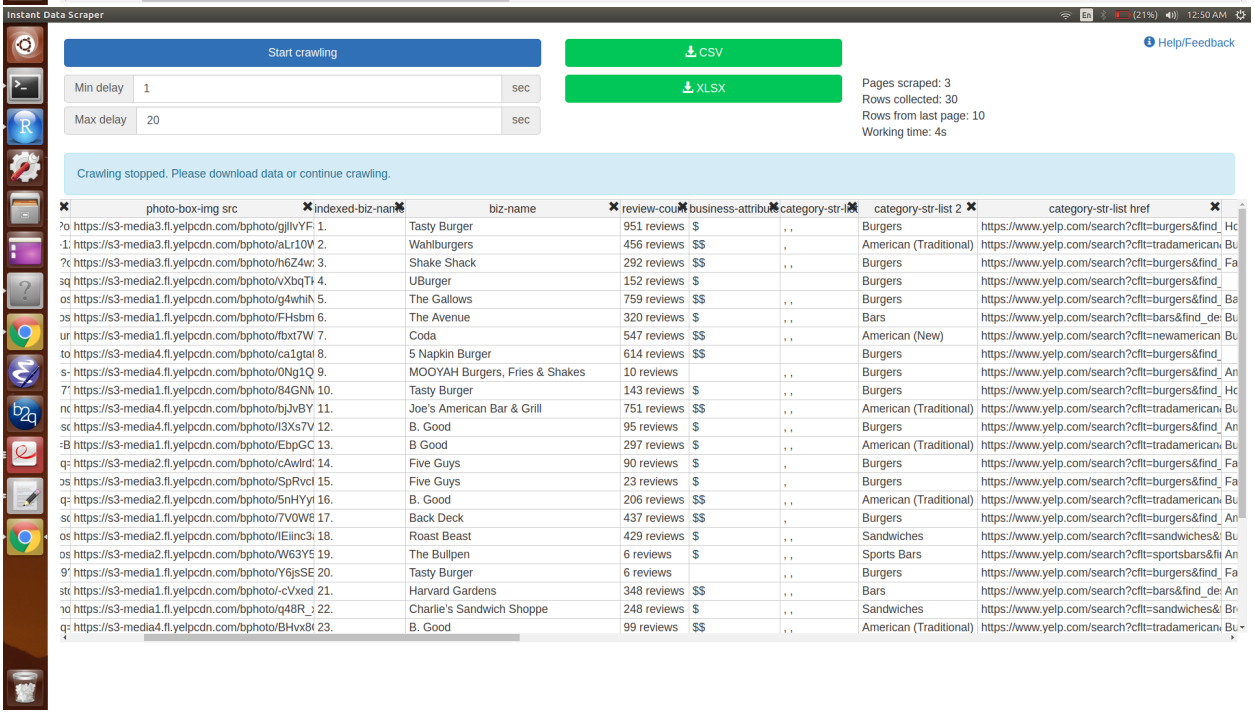

Places in Boston, MA - Yelp - Google Chrome
Find Burgers
Near Fenway, Boston, MA
Sign Up
Burgers
Buffets
Tasty Burger
See all nearby locations
Bubble Tea
Burrito
Buffet All You Can Eat
In-and-out Burger
Restaurants
Write a Review
Events
Talk
Log In
Showing 1-10 of 1056
Order Takeout
Cash Back
All Filters
Sort By
Best Match
Highest Rated
Most Reviewed
West End
Fenway
Allston/Brighton
More Neighborhoods
Driving (5 mi.)
Biking (2 mi.)
Walking (1 mi.)
Within 4 blocks
Price
Features
Order Delivery
Order Takeout
Open Now 12:31 AM
Take-out
More Features
1. Atlantic Fish
2110 reviews
$$$ · Seafood, Live/Raw Food, Cocktail Bars
Back Bay
761 Boylston St
Boston, MA 02116
(617) 267-4000
Growing up in Maine I must admit my expectations of Atlantic Fish were not high at all. I've been living here for years and never gave much thought to the place. Until my friend had… read more
2. El Pelón Taquería
978 reviews
$ · Mexican
$5 for Flan
Fenway
92 Peterborough St
Boston, MA 02215
(617) 262-9090
This restaurant does food delivery
Start Order
Mo' Map
Redo search when map moved
Somerville
Cambridge
Brookline
Map data ©2018 Google
Terms of Use

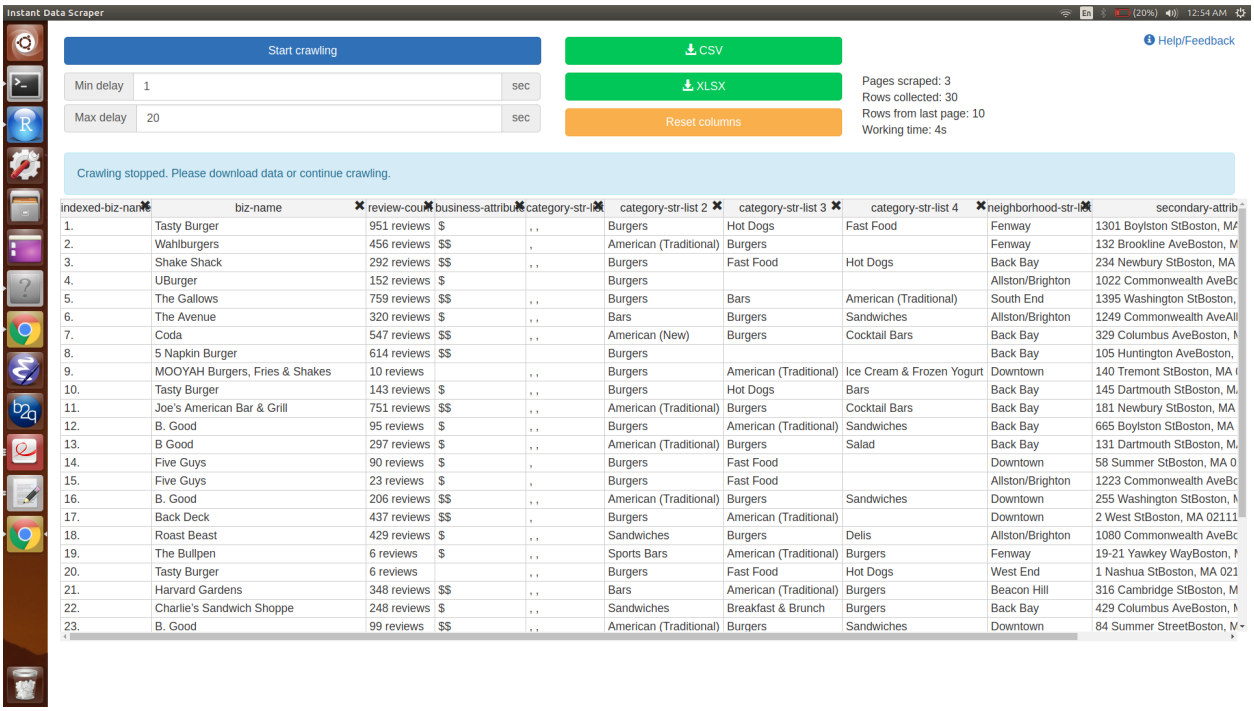

Burgers Boston, MA - Yelp - Google Chrome
Find Burgers
Near Fenway, Boston, MA
Sign Up
Restaurants
Nightlife
Home Services
Write a Review
Events
Talk
Log In
Burgers Boston, MA
Showing 1-10 of 328
Boston, MA › Fenway
$
$$
$$$
$$$$
Open Now
Order Delivery
Order Takeout
All Filters
Sort By
Best Match
Highest Rated
Most Reviewed
Neighborhoods
South End
West End
Fenway
Allston/Brighton
More Neighborhoods
Distance
Bird's-eye View
Driving (5 mi.)
Biking (2 mi.)
Walking (1 mi.)
Within 4 blocks
Price
Features
Order Delivery
Order Takeout
Open Now 12:33 AM
Good for Groups
More Features
Category
Burgers
Restaurants
Bars
American (Traditional)
More Categories
Ad Uburger
422 reviews
$ · Sandwiches, Fast Food, Burgers
636 Beacon St
Boston, MA 02215
(857) 362-8811
My boyfriend has been raving about this place for, like, a year. I always wait outside as he grabs his burger until I venture to find something else. Why? Well for starters I thought… read more
Ad Stop and Taste Pizzeria
8 reviews
$ · Pizza, Burgers, Sandwiches
Dudley Square
239 Dudley St
Roxbury, MA 02119
(617) 445-7866
This restaurant accepts takeout and delivery
Start Order
Mo' Map
Redo search when map moved
Somerville
Cambridge
Brookline
Map data ©2018 Google
Terms of Use


Burgers Boston, MA - Yelp - Google Chrome
Burgers Boston, MA
Showing 11-20 of 328
Boston, MA › Fenway
$ $$ $$$ $$$$ Open Now Order Delivery Order Takeout All Filters
Sort By
Best Match
Highest Rated
Most Reviewed
Neighborhoods
South End
West End
Fenway
Allston/Brighton
More Neighborhoods
Distance
Bird's-eye View
Driving (5 mi.)
Biking (2 mi.)
Walking (1 mi.)
Within 4 blocks
Price
$
$$
$$$
$$$$
Features
Order Delivery
Order Takeout
Open Now 12:33 AM
Good for Groups
More Features
Category
Burgers
Restaurants
Bars
American (Traditional)
More Categories
Ad Uburger
422 reviews
$ · Burgers, Fast Food, Sandwiches
636 Beacon St
Boston, MA 02215
(857) 362-8811
My boyfriend has been raving about this place for, like, a year. I always wait outside as he grabs his burger until I venture to find something else. Why? Well for starters I thought… read more
Ad Slade's Bar & Grill
91 reviews
$$ · Bars, American (Traditional)
958 Tremont St
Roxbury Crossing, MA 02120
(617) 442-4600
I bookmarked them months ago once I knew I was coming to Boston. This was my 1st stop off of the plane. When we arrived they were already jumping with dine ins and take out. We dined… read more
11. Bukowski Tavern
636 reviews
$$ · American (Traditional), Dive Bars
Back Bay
50 Dalton St
Boston, MA 02115
(617) 437-9999
Mo' Map
Redo search when map moved
Somerville
Cambridge
Brookline
Map data ©2018 Google
Terms of Use


Burgers Boston, MA - Yelp - Google Chrome
Burgers Boston, MA
Showing 21-30 of 328
Boston, MA › Fenway
$ $$ $$$ $$$$ Open Now Order Delivery Order Takeout All Filters
Sort By
Best Match
Highest Rated
Most Reviewed
Neighborhoods
South End
West End
Fenway
Allston/Brighton
More Neighborhoods
Distance
Bird's-eye View
Driving (5 mi.)
Biking (2 mi.)
Walking (1 mi.)
Within 4 blocks
Price
$
$$
$$$
$$$$
Features
Order Delivery
Order Takeout
Open Now 12:34 AM
Good for Groups
More Features
Category
Burgers
Restaurants
Bars
American (Traditional)
More Categories
Ad Uburger
422 reviews
$ · Fast Food, Burgers, Sandwiches
636 Beacon St
Boston, MA 02215
(857) 362-8811
My boyfriend has been raving about this place for, like, a year. I always wait outside as he grabs his burger until I venture to find something else. Why? Well for starters I thought… read more
Ad Stop and Taste Pizzeria
8 reviews
$ · Pizza, Sandwiches, Burgers
Dudley Square
239 Dudley St
Roxbury, MA 02119
(617) 445-7866
This restaurant accepts takeout and delivery
Start Order
This place fills its little niche, and then some. The menu is pretty expansive -- from three different sizes of about 10-15 different sandwiches to fish and chips, to its namesake,… read more
21. Porters Bar & Grill
182 reviews
West End
173 Portland St
Mo' Map
Redo search when map moved
Somerville
Cambridge
Brookline
Map data ©2018 Google
Terms of Use


Instant Data Scraper
Try another table
Locate "Next" button
Start crawling
Min delay 1 sec
Max delay 20 sec
CSV
XLSX
Help/Feedback
Pages scraped: 1
Rows collected: 10
Rows from last page: 10
Working time: 0s
Download data or locate "Next" to crawl multiple pages
photo-box-img src | indexed-biz-name | biz-name | review-count | business-attribute | category-str-list | category-str-list 2 | category-str-list href | category
1. Tasty Burger | 951 reviews | $ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_ | Hot Dogs
2. Wahlburgers | 456 reviews | $$ | , | American (Traditional) | https://www.yelp.com/search?cflt=tradamerican | Burgers
3. Shake Shack | 292 reviews | $$ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_ | Fast Food
4. UBurger | 152 reviews | $ | | Burgers | https://www.yelp.com/search?cflt=burgers&find_
5. The Gallows | 759 reviews | $$ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_ | Bars
6. The Avenue | 320 reviews | $ | , , | Bars | https://www.yelp.com/search?cflt=bars&find_de | Burgers
7. Coda | 547 reviews | $$ | , , | American (New) | https://www.yelp.com/search?cflt=newamerican | Burgers
8. 5 Napkin Burger | 614 reviews | $$ | | Burgers | https://www.yelp.com/search?cflt=burgers&find_
9. MOOYAH Burgers, Fries & Shakes | 10 reviews | | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_ | American
10. Tasty Burger | 143 reviews | $ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_ | Hot Dogs


Instant Data Scraper
Start crawling
Min delay 1 sec
Max delay 20 sec
CSV
XLSX
Help/Feedback
Pages scraped: 3
Rows collected: 30
Rows from last page: 10
Working time: 4s
Crawling stopped. Please download data or continue crawling.
photo-box-img src | indexed-biz-name | biz-name | review-count | business-attribute | category-str-list | category-str-list 2 | category-str-list href
1. Tasty Burger | 951 reviews | $ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
2. Wahlburgers | 456 reviews | $$ | , | American (Traditional) | https://www.yelp.com/search?cflt=tradamerican
3. Shake Shack | 292 reviews | $$ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
4. UBurger | 152 reviews | $ | | Burgers | https://www.yelp.com/search?cflt=burgers&find_
5. The Gallows | 759 reviews | $$ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
6. The Avenue | 320 reviews | $ | , , | Bars | https://www.yelp.com/search?cflt=bars&find_de
7. Coda | 547 reviews | $$ | , , | American (New) | https://www.yelp.com/search?cflt=newamerican
8. 5 Napkin Burger | 614 reviews | $$ | | Burgers | https://www.yelp.com/search?cflt=burgers&find_
9. MOOYAH Burgers, Fries & Shakes | 10 reviews | | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
10. Tasty Burger | 143 reviews | $ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
11. Joe's American Bar & Grill | 751 reviews | $$ | , , | American (Traditional) | https://www.yelp.com/search?cflt=tradamerican
12. B. Good | 95 reviews | $ | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
13. B Good | 297 reviews | $ | , , | American (Traditional) | https://www.yelp.com/search?cflt=tradamerican
14. Five Guys | 90 reviews | $ | , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
15. Five Guys | 23 reviews | $ | , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
16. B. Good | 206 reviews | $$ | , , | American (Traditional) | https://www.yelp.com/search?cflt=tradamerican
17. Back Deck | 437 reviews | $$ | , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
18. Roast Beast | 429 reviews | $ | , , | Sandwiches | https://www.yelp.com/search?cflt=sandwiches&
19. The Bullpen | 6 reviews | $ | , , | Sports Bars | https://www.yelp.com/search?cflt=sportsbars&fi
20. Tasty Burger | 6 reviews | | , , | Burgers | https://www.yelp.com/search?cflt=burgers&find_
21. Harvard Gardens | 348 reviews | $$ | , , | Bars | https://www.yelp.com/search?cflt=bars&find_de
22. Charlie's Sandwich Shoppe | 248 reviews | $ | , , | Sandwiches | https://www.yelp.com/search?cflt=sandwiches&
23. B. Good | 99 reviews | $$ | , , | American (Traditional) | https://www.yelp.com/search?cflt=tradamerican



indexed-biz-name	biz-name	review-count	business-attributes	category-str-list	category-str-list 2	category-str-list 3	category-str-list 4	neighborhood-str-list	secondary-attrib
1.	Tasty Burger	951 reviews	$	, ,	Burgers	Hot Dogs	Fast Food	Fenway	1301 Boylston StBoston, MA
2.	Wahlburgers	456 reviews	$$	,	American (Traditional)	Burgers		Fenway	132 Brookline AveBoston, M
3.	Shake Shack	292 reviews	$$	, ,	Burgers	Fast Food	Hot Dogs	Back Bay	234 Newbury StBoston, MA
4.	UBurger	152 reviews	$		Burgers			Allston/Brighton	1022 Commonwealth AveBo
5.	The Gallows	759 reviews	$$	, ,	Burgers	Bars	American (Traditional)	South End	1395 Washington StBoston,
6.	The Avenue	320 reviews	$	, ,	Bars	Burgers	Sandwiches	Allston/Brighton	1249 Commonwealth AveAll
7.	Coda	547 reviews	$$	, ,	American (New)	Burgers	Cocktail Bars	Back Bay	329 Columbus AveBoston, M
8.	5 Napkin Burger	614 reviews	$$		Burgers			Back Bay	105 Huntington AveBoston,
9.	MOOYAH Burgers, Fries & Shakes	10 reviews		, ,	Burgers	American (Traditional)	Ice Cream & Frozen Yogurt	Downtown	140 Tremont StBoston, MA (
10.	Tasty Burger	143 reviews	$	, ,	Burgers	Hot Dogs	Bars	Back Bay	145 Dartmouth StBoston, M
11.	Joe's American Bar & Grill	751 reviews	$$	, ,	American (Traditional)	Burgers	Cocktail Bars	Back Bay	181 Newbury StBoston, MA
12.	B. Good	95 reviews	$	, ,	Burgers	American (Traditional)	Sandwiches	Back Bay	665 Boylston StBoston, MA
13.	B Good	297 reviews	$	, ,	American (Traditional)	Burgers	Salad	Back Bay	131 Dartmouth StBoston, M
14.	Five Guys	90 reviews	$	,	Burgers	Fast Food		Downtown	58 Summer StBoston, MA 0
15.	Five Guys	23 reviews	$	,	Burgers	Fast Food		Allston/Brighton	1223 Commonwealth AveBo
16.	B. Good	206 reviews	$$	, ,	American (Traditional)	Burgers	Sandwiches	Downtown	255 Washington StBoston, M
17.	Back Deck	437 reviews	$$	,	Burgers	American (Traditional)		Downtown	2 West StBoston, MA 02111
18.	Roast Beast	429 reviews	$	, ,	Sandwiches	Burgers	Delis	Allston/Brighton	1080 Commonwealth AveBo
19.	The Bullpen	6 reviews	$	, ,	Sports Bars	American (Traditional)	Burgers	Fenway	19-21 Yawkey WayBoston, M
20.	Tasty Burger	6 reviews		, ,	Burgers	Fast Food	Hot Dogs	West End	1 Nashua StBoston, MA 021
21.	Harvard Gardens	348 reviews	$$	, ,	Bars	American (Traditional)	Burgers	Beacon Hill	316 Cambridge StBoston, M
22.	Charlie's Sandwich Shoppe	248 reviews	$	, ,	Sandwiches	Breakfast & Brunch	Burgers	Back Bay	429 Columbus AveBoston, M
23.	B. Good	99 reviews	$$	, ,	American (Traditional)	Burgers	Sandwiches	Downtown	84 Summer StreetBoston, M

#The screenshots are of web scraping using Instant Data Scraper

1. The first screenshot is the default yelp homepage without any filters.
2. The second is of the page where we filter the restaurants for burgers.
3. The third one is the first page of the search after we filtered Boston neighborhoods of Allston, Brighton, Back Bay, Beacon Hill, Downtown Area, Fenway, South End, and West End.
4. & 5. The fourth and fifth screenshots are of the second and third page results of our search.
5. After we click on the Instant Data Scraper extension in the chrome, we get a table with the data in the webpage in separated columns for each data type.
6. The seventh screenshot is after we located the next button on the webpage and started crawling to get the data of the first three pages in the search result.
7. This is the cleaned data after removing unwanted columns. Then I downloaded this data as a csv file which I loaded into R.

```
# Using rvest and selector gadget for web scraping
library(rvest)
```

```
## Loading required package: xml2
```

```
# The urls of the first three web page results
boston_burgers1 <- read_html("https://www.yelp.com/search?find_desc=Burgers&start=0&l=p:MA:Boston::%5BA
boston_burgers2 <- read_html("https://www.yelp.com/search?find_desc=Burgers&start=10&l=p:MA:Boston::%5B
boston_burgers3 <- read_html("https://www.yelp.com/search?find_desc=Burgers&start=20&l=p:MA:Boston::%5B

#Scrape the website for the restaurant names
rest_names <- boston_burgers1 %>%
  html_nodes(".indexed-biz-name span") %>%
  html_text()
rest_names
```

```
##   [1] "Tasty Burger"                  "Wahlburgers"
##   [3] "Shake Shack"                   "UBurger"
##   [5] "The Gallows"                   "The Avenue"
```

```
##  [7] "Jm Curley"                      "Coda"
##  [9] "5 Napkin Burger"                "MOOYAH Burgers, Fries & Shakes"
```

```
#Scrape the website for the restaurant addresses
rest_add <- boston_burgers1 %>%
  html_nodes(".natural-search-result address") %>%
  html_text()
rest_add
```

```
##  [1] "\n        1301 Boylston StBoston, MA 02215\n    "
##  [2] "\n        132 Brookline AveBoston, MA 02215\n    "
##  [3] "\n        234 Newbury StBoston, MA 02116\n    "
##  [4] "\n        1022 Commonwealth AveBoston, MA 02215\n    "
##  [5] "\n        1395 Washington StBoston, MA 02118\n    "
##  [6] "\n        1249 Commonwealth AveAllston, MA 02134\n    "
##  [7] "\n        21 Temple PlBoston, MA 02111\n    "
##  [8] "\n        329 Columbus AveBoston, MA 02116\n    "
##  [9] "\n        105 Huntington AveBoston, MA 02199\n    "
## [10] "\n        140 Tremont StBoston, MA 02111\n    "
```

```
#Scrape the website for the review count
review_count <- boston_burgers1 %>%
  html_nodes(".natural-search-result .rating-qualifier") %>%
  html_text()
review_count
```

```
##  [1] "\n            953 reviews\n    " "\n            458 reviews\n    "
##  [3] "\n            293 reviews\n    " "\n            152 reviews\n    "
##  [5] "\n            759 reviews\n    " "\n            321 reviews\n    "
##  [7] "\n            677 reviews\n    " "\n            547 reviews\n    "
##  [9] "\n            615 reviews\n    " "\n            11 reviews\n    "
```

```
#Scrape the website for price range
price_range <- boston_burgers1 %>%
  html_nodes(".natural-search-result .bullet-after") %>%
  html_text()
price_range
```

```
## [1] "\n            \n        $\n        "
## [2] "\n            \n        $$\n        "
## [3] "\n            \n        $$\n        "
## [4] "\n            \n        $\n        "
## [5] "\n            \n        $$\n        "
## [6] "\n            \n        $\n        "
## [7] "\n            \n        $$\n        "
## [8] "\n            \n        $$\n        "
## [9] "\n            \n        $$\n        "
```

```
#Scrape the website for service categories
ser_cat <- boston_burgers1 %>%
  html_nodes(".natural-search-result .category-str-list a") %>%
  html_text()
ser_cat
```

```
##  [1] "Burgers"                  "Hot Dogs"
##  [3] "Fast Food"                "American (Traditional)"
##  [5] "Burgers"                  "Burgers"
```

```
##  [7] "Fast Food"                 "Hot Dogs"
##  [9] "Burgers"                   "Burgers"
## [11] "Bars"                      "American (Traditional)"
## [13] "Bars"                      "Burgers"
## [15] "Sandwiches"                "American (New)"
## [17] "Lounges"                   "American (New)"
## [19] "Burgers"                   "Cocktail Bars"
## [21] "Burgers"                   "Burgers"
## [23] "American (Traditional)"    "Ice Cream & Frozen Yogurt"
```

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.

```
yelp <- read.csv('yelp_data.csv')
yelp <- as.data.frame(yelp)
View(yelp)
```

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.

Instant Data Scraper It is a Chrome extension which uses AI to detect tabular or listing type data on web pages. Such data can be scraped into CSV or Excel file without the need for any coding. The extension can also click on the "Next" page links or buttons and retrieve data from multiple pages into one file.

Cost: It is free to use

Ease of use: It is very easy to use. To use it, do the following: 1. Open the first page of listing results (products, directory, etc) in your browser 2. Activate the extension 3. Extension will guess where your data is. If not happy use "Try another table" button to guess again. 4. Download CSV or Excel from the first page if that is all you need. Or click to locate "Next" button to mark the "Next" link/button on a website. 5. Click "Start crawling" to start crawling through multiple pages a website. Extension will show statistics on what is being collected. 6. Download Excel or CSV file at any time during the crawl. 7. Clean up Excel or CSV files – it will most likely have some unwanted additional fields that were extracted from the page. Most likely column names will have to be renamed as well.

Features: 1. Try another table – It makes alternative tables if the initial table did not have the data you want. 2. Locate "Next" button – This marks the location of "Next" button on the website. This is used to scrape data from multiple pages into one file. 3. Crawl delay – It is the time in seconds before going to the next page. Default value is 1 second. It can be increased when pages load information dynamically. 4. CSV and XLSX file download buttons are there. They are active right away when any data is found.

Rvest It is a tool in R which is used to scrape information from web pages.

Cost: It is free to use.

Ease of use: It is not very easy to use if the user does not have previous programming experience. But after going through examples and with some practice, it can be mastered.

Features: 1. Create an html document from a url, a file on disk or a string containing html with read_html(). 2. Select parts of a document using css selectors: html_nodes(doc, "table td") 3. Extract components with html_tag() (the name of the tag), html_text() (all text inside the tag), html_attr() (contents of a single attribute) and html_attrs() (all attributes). 4. Parse tables into data frames with html_table(). 5. Extract, modify and submit forms with html_form(), set_values() and submit_form(). 6. Detect and repair encoding problems with guess_encoding() and repair_encoding(). 7. Navigate around a website as if you're in a browser with html_session(), jump_to(), follow_link(), back(), forward(), submit_form() and so on.

My recommendation: Instant Data Scraper It scrapes the data very neatly into columns. Even people who have no programming experience can use it to get the required data. All it takes is one click of the extension

in chrome on the webpage that we want to scrape and it automatically collects all the data in columns which is available for us to download as a csv or excel file. The extension runs completely in user's browser and does not send data to Web Robots.

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?

The urls: 1. https://www.yelp.com/search?find_desc=Burgers&start=0&cflt=burgers&l=p:MA:Boston::%5BAllston/Brighton,Back_Bay,Beacon_Hill,Downtown,Fenway,South_End,West_End%5D

2. https://www.yelp.com/search?find_desc=Burgers&start=10&cflt=burgers&l=p:MA:Boston::%5BAllston/Brighton,Back
3. https://www.yelp.com/search?find_desc=Burgers&start=20&cflt=burgers&l=p:MA:Boston::%5BAllston/Brighton,Back

The difference between the three URLs is &start= which is 0 in first, 10 in second and 20 in third url.

Parameters for search: 1. "search?find_desc=": This identifies the search terms used which in this case is burgers.

2. "&cflt=": This is a category filter that allows you to specify the categories which are to be searched.
3. "&l=": This is for location which is to be used in the search result
4. "%5": This is for neighborhood filters on the larger location

Parameter for pagination: 1.. "&start=": This identifies the start of the result of the query. First page starts with 0 results so it is 0, second page starts with 10th result so it is 10 and 3rd starts with 20th so it is 20.

My guess of the URL for the 7th page of Chinese restaurants in New York is: https://www.yelp.com/search?find_desc=Chinese&start=60&cflt=chinese&l=p:NY:NewYork:%5D