Northeastern University

Course:	DA5020
Assignment:	Web Scraping through Toolkits
Total Points:	100
Date Due:	Posted on Blackboard

Learning Objectives

In this assignment, you will learn how to:

- scrape data from HTML through a toolkit
- identify search parameters through a URL
- compare toolkits and write a report

Tasks

The objective of this assignment is to understand the structure of HTML pages and URL parameters. You will scrape data from a website we choose and write a report comparing different web scraping toolkits.

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.

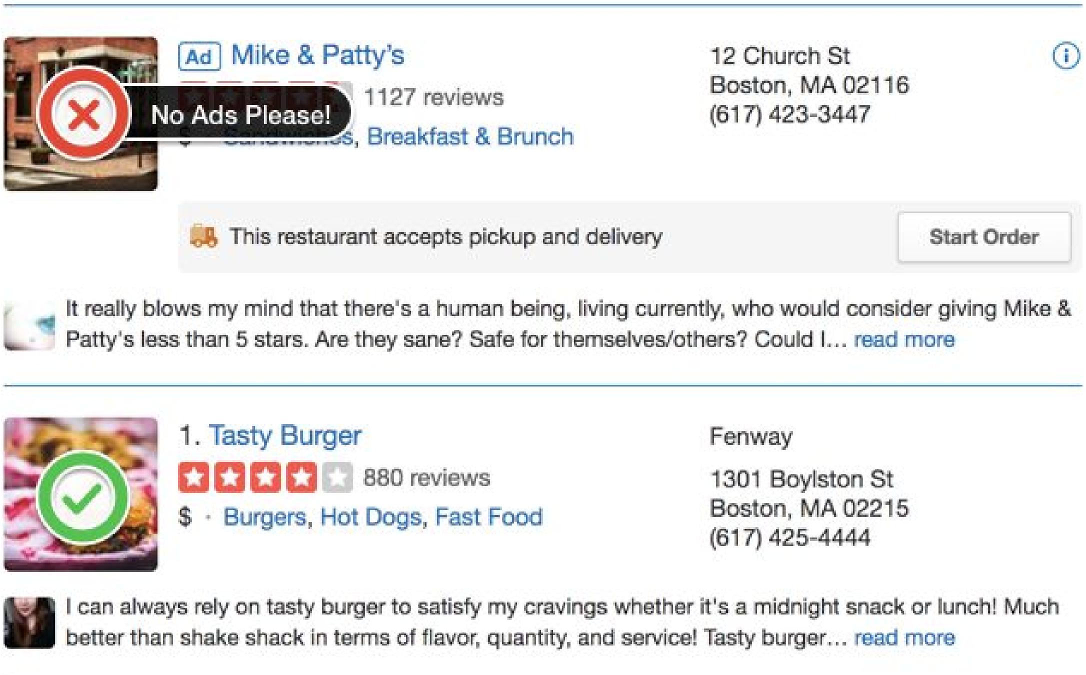
You are expected to:

a. Start at the website https://www.yelp.com/boston, create a search for Burgers.

b. Use the search filters to limit Boston neighborhoods to *Allston, Brighton, Back Bay, Beacon Hill, Downtown Area, Fenway, South End,* and *West End*.

c. Notice the URL format in your browser's location bar. Save the URL somewhere safe. You want to extract the first *three* pages of the search results. For each page notice the change in the URL and save the updated URLs, too. You will need to discuss these URLs later.

d. Extract information about restaurants appeared in the search results list, including but not limited to their name, address, service categories, review count, and review stars. *Do not scrape "Ad" items.*



Ad Mike & Patty's
1127 reviews
$ · Sandwiches, Breakfast & Brunch

12 Church St
Boston, MA 02116
(617) 423-3447

This restaurant accepts pickup and delivery

It really blows my mind that there's a human being, living currently, who would consider giving Mike & Patty's less than 5 stars. Are they sane? Safe for themselves/others? Could I... read more

1. Tasty Burger
880 reviews
$ · Burgers, Hot Dogs, Fast Food

Fenway

1301 Boylston St
Boston, MA 02215
(617) 425-4444

I can always rely on tasty burger to satisfy my cravings whether it's a midnight snack or lunch! Much better than shake shack in terms of flavor, quantity, and service! Tasty burger... read more

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.

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.

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*?

Deliverables

Complete the tasks and create a report in PDF format. Submit your report to blackboard. Within your report, include pictures, screenshots, data file extracts, charts, and anything else that shows your analysis of the toolkits.

You can write the PDF in any tool you like, but it is encouraged to write it with R markdown. You can learn how to insert images in R markdown documents here and here.