ISA 401/501: Business Intelligence & Data Visualization

05: Scraping Multiple Webpages in R

Fadel M. Megahed, PhD

Enders Associate Professor
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
? Automated Scheduler for Office Hours

Fall 2022

Quick Refresher from Last Class

- ✅ Understand when can we scrape data (i.e., `robots.txt`)
- ✅ Scrape a webpage using R

Kahoot Competition #2

To assess your understanding and retention of the topics covered last week, you will **compete in a Kahoot competition (consisting of 5 questions)**:

- Go to https://kahoot.it/
- Enter the game pin, which will be shown during class
- Provide your first (preferred) and last name
- Answer each question within the allocated time-window (**fast and correct answers provide more points**)

Winning the competition involves having as many correct answers as possible AND taking the shortest duration to answer these questions. The winner 🏆 of the competition from each section will receive: $10 Starbucks gift card. Good luck!!!

P.S: The Kahoot competition will have **no impact on your grade**. It is a **fun** way of assessing your knowledge, motivating you to ask questions about topics covered that you do not have a full understanding of it, and providing me with some data that I can use to pace today's class.

Going Over Assignment 04 Solutions

Q1 Q2 Q3 Q4

Use the `robotstxt` package to examine whether https://www.yelp.com/biz/pattersons-cafe-oxford can be scraped per our discussion in class. You should store the output from the `robotstxt::paths_allowed()` function in an object with the name of `yelp_robots`

Going Over Assignment 04 Solutions

Q1 Q2 Q3 Q4

Go to the webpage https://docs.google.com/spreadsheets/d/e/2PACX-1vQ3uk9AJOMODxS9fUgX_4vnEMj-Di7ulkTXWzPUmaHvHbaIl63xmKmRu3VaBvOXrwQhtkOUlL9fxLMB/pubhtml?gid=1104208671&single=true which contains Miami University's Lost and Found Database and write an R script that will scrape the contents of the table in the spreadsheet. You should save the tibble/data.frame of the results (NOT a list) in an object with the name of `lost_found`

Going Over Assignment 04 Solutions

Q1 Q2 Q3 Q4

Go to https://www.miamioh.edu/fsb/academics/isa/about/faculty-staff/index.html and write an R script that will scrape the three column table containing faculty and staff information. You should save the tibble/data.frame of the results (NOT a list) in an object with the name of `isa_fac`

Going Over Assignment 04 Solutions

Q1 Q2 Q3 Q4

The most popular listings on Netflix are rated and reviews on ImDb are available at https://www.imdb.com/search/title/?companies=co0144901. Write an R script that will produce a tibble that contains the following:

- title, which you will save in a column titled `title`
- year/years of show, which you will save in a column titled `year`
- 1-2 sentence summary of show, which you save in a column titled `summary`

The tibble containing these three columns should have the name of `netflix`. Please make sure that you do not overwrite `netflix` at any point of the code.

Learning Objectives for Today's Class

- Understand when can we scrape data (i.e., `robots.txt`)
- Scrape multiple webpages using R
- Use loops and/or tidymodeling approaches to scrape data from multiple webpages

Web Scraping Demos (Cont.)

Cleaning Up the Output from your Non-Graded Class Activity

Activity My Solution

- Go to this database on plane crashes
- Scrape the HTML table. **Note the difference from text elements:**
  - The CSS selector for `html_elements()` will be different.
  - You will extract a table (in its **entirety**) and hence:
  - we will use `html_table()` instead of `html_text2()`
- Store the scraped data in an appropriate location on your computer (e.g., within the data folder for ISA 401)

Cleaning Up the Output from your Non-Graded Class Activity

Activity My Solution

Please refer to our discussion in class

Demo 2: Scraping all Plane Crashes 2013-2022

- We will build on the previous example and we will scrape all the plane crashes that were recorded in the plane crash database between 2013-2022.
- Then, we will create a single **tibble** for all crashes. It will contain the fields in the individual tables as well as the year of crash.
- Then, we will **export the results to a CSV** so that we can analyze that in a separate program if we wanted to.

Demo 3: Scraping the first 300 entries in IMDB

The most popular listings on Netflix are rated and reviews on ImDb are available at https://www.imdb.com/search/title/?companies=co0144901. Write an R script that will produce a tibble that contains the **following information for the first 300 entries**:

- title, which you will save in a column titled `title`
- year/years of show, which you will save in a column titled `year`
- 1-2 sentence summary of show, which you save in a column titled `summary`

Demo 4: Downloading the Lecture PDFs (if time allows)

If time allows, we will download all the ISA 401 lecture pdf-slides from GitHub using an R script.

Recap

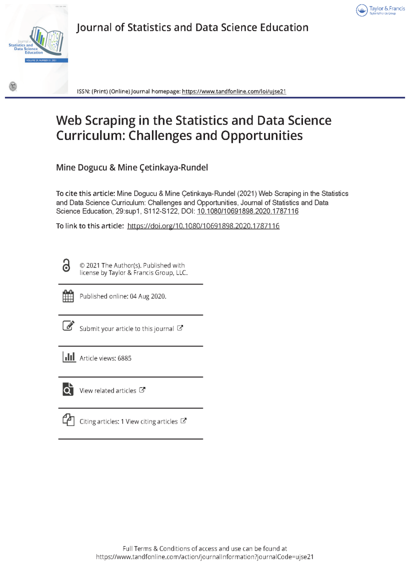
Summary of Main Points

By now, you should be able to do the following:

- Understand when can we scrape data (i.e., `robots.txt`)
- Scrape multiple webpages using R
- Use loops and/or tidymodeling approaches to scrape data from multiple webpages

Things to Do to Prepare for Next Class

- Go over your notes, read through the supplementary material (below), go through the self-paced tutorial and complete Assignment 05 on Canvas.



Journal of Statistics and Data Science Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ujse21

Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities

Mine Dogucu & Mine Çetinkaya-Rundel

To cite this article: Mine Dogucu & Mine Çetinkaya-Rundel (2021) Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities, Journal of Statistics and Data Science Education, 29:sup1, S112-S122, DOI: 10.1080/10691898.2020.1787116

To link to this article: https://doi.org/10.1080/10691898.2020.1787116

Published online: 04 Aug 2020.

Submit your article to this journal

Article views: 6885

View related articles

Citing articles: 1 View citing articles

- PDF of Published Paper
- ePub of Published Paper

- Selector Gadget
- Getting Started with rvest