

ISA 401: Business Intelligence & Data Visualization

06: Scraping Multiple Webpages

Fadel M. Megahed, PhD

Professor of Information Systems and Business Analytics
Farmer School of Business
Miami University

 @FadelMegahed


 fmegahed

 fmegahed@miamioh.edu

 Automated Scheduler for Office Hours

Fall 2024


Quick Refresher from Last Class

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

Non-Graded Assessment of Your Understanding

Scrape the Names and Positions of the current cabinet members from <https://www.whitehouse.gov/administration/cabinet/> and save the results into a **data frame** that contains two columns: (a) name, and (b) position.

Learning Objectives for Today's Class


- Scrape multiple webpages using .
- Use loops and/or tidymodeling approaches to scrape data from multiple webpages.

Web Scraping Demos (Cont.)

Demo 1: Scraping all Plane Crashes 2020-2024

- We will build on the previous example and we will scrape all the plane crashes that were recorded in the [plane crash database](#) between 2020-2024.
- Then, we will create a single **data frame** 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.

Practice Outside of Class


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  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`

Recap

Summary of Main Points

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

- Scrape multiple webpages using .
- Use loops and/or tidy modeling approaches to scrape data from multiple webpages.

Kahoot Competition # 1

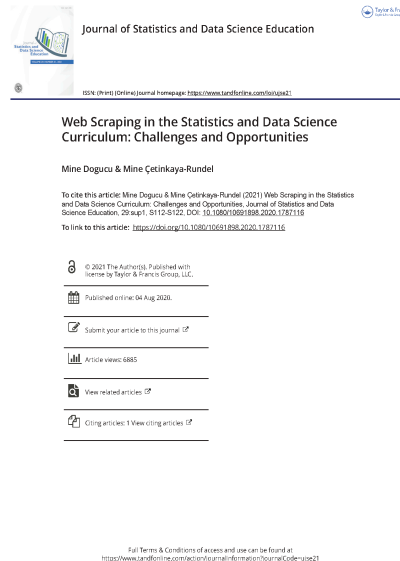
To assess your understanding and retention of the topics covered last week, you will **compete in a Kahoot competition (consisting of 16 questions)**:

- 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 20-second 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!!!

Things to Do to Prepare for Next Class

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



- PDF of Published Paper
- ePub of Published Paper
- Practical Web Scraping in R