Practice3S24

BilalTariq

2024-02-22

Practice3 - Due Thursday, 2/22 by midnight to Gradescope

Reminder: Practice assignments may be completed working with other individuals.

Reading

The associated reading for the week is Sections 6.4, 8.5-8.7, and 8.9-8.10.

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook, course materials in the repo, labs, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

1 - Scraping Tables

The text example showed how to scrape tables from a Wikipedia page. We also saw how to scrape a table from basketball-reference.com in our lecture notes. For this exercise, your task is to:

- scrape a table of your choosing from a different website (yes, it can be a different Wikipedia page),
- clean it up (i.e. understandable variable names, etc. in a display), and
- display a few rows of it in a nice table.

You must be sure that scraping the table is allowed. Your code should show appropriate documentation of your steps.

Solution:

```
url <- "https://en.wikipedia.org/wiki/List_of_best-selling_video_games"

paths_allowed(url) #we're allowed to scrape
```

```
en.wikipedia.org
```

```
[1] TRUE
```

```
tables <- url %>%
  read_html %>%
  html_elements("table")

length(tables)
```

```
[1] 4
```

```
video_game_data <- tables %>%
  purrr::pluck(2)%>%
  html_table() %>%
  select(
    "Title", "Sales","Initial release date", "Developer(s)[b]"
  )
```

```r
video_game_table_cleaned <- clean_names(video_game_data) %>%
  rename(
    "Title" = title,
    "Sales" = sales,
    "Initial Release Date" = initial_release_date,
    "Developer(s)" = developer_s_b
  ) %>%
  slice(1:7)


kable(video_game_table_cleaned, booktabs = TRUE)
```

Title	Sales	Initial Release Date	Developer(s)
Minecraft	300,000,000	November 18, 2011[c]	Mojang Studios
Grand Theft Auto V	195,000,000	September 17, 2013	Rockstar North
Tetris (EA)	100,000,000	September 12, 2006	EA Mobile
Wii Sports	82,900,000	November 19, 2006	Nintendo EAD
PUBG: Battlegrounds	75,000,000	December 20, 2017	PUBG Studios
Mario Kart 8 / Deluxe	69,040,000	May 29, 2014	Nintendo EAD / Nintendo EPD (Deluxe)
Red Dead Redemption 2	61,000,000	October 26, 2018	Rockstar Games

2 - MDSR 8.6

Complete MDSR 8.6, which states: "A Slate article (http://tinyurl.com/slate-ethics) discussed whether race/ethnicity should be included in a predictive model for how long a homeless family would stay in homeless services. Discuss the ethical considerations involved in whether race/ethnicity should be included as a predictor in the model."

Solution:

Today, our world operates on our data being used (willingly or unwillingly). This results in rather unintended consequences that we couldn't have predicted, and these consequences escalate in their impact by the day. When it comes to using data, there is bound to be hidden nuance to a dataset that is not truly understood by the data scientist. In particular, when it comes to race, it is not clear what datasets are based on race or not. For example, when feeding data to machine learning models for image generation, there was a disproportionate amount of white males that popped up in those images but the models struggled with recreating faces of color because of the lack of data. Furthermore, as mentioned in the article, race is not just one category; it could be a factor in your ZIP code, in your income bracket, or how policed your neighbourhood is.

As an International Muslim Pakistani student, I have struggled first hand with this as well, which is why as an aspiring data scientist it's my firm belief that we should monitor data and raise questions about it which encourage discourse. If the past data is biased, we should work towards making that not the case, and we should foster data points that do not involve predicting how long a family would be homeless based of their race/ethnicity.

3 - Scraping Text with Weather Data

We want to get a tiny bit of practice with the web developer tools demo-ed in class for scraping in this exercise.

Go to the National Weather Service website and get a forecast page for a city of your choice (maybe your hometown, or Amherst, or a place you want to visit in the States, etc.).

> part a - Save the url of the page as `weatherurl`. Then, check that you are allowed to access the page for scraping.

Solution:

```
url <- "https://forecast.weather.gov/MapClick.php?lat=42.3909&lon=-72.5283"

weatherurl <- url

paths_allowed(weatherurl) #paths_allowed is true
```

```
forecast.weather.gov

Warning in request_handler_handler(request = request, handler = on_not_found, :
Event: on_not_found

Warning in request_handler_handler(request = request, handler =
on_file_type_mismatch, : Event: on_file_type_mismatch

Warning in request_handler_handler(request = request, handler =
on_suspect_content, : Event: on_suspect_content
```

```
[1] TRUE
```

In class, we accessed the table of current conditions and the extended forecast temperatures for the Amherst page via text. Above but near the table of current conditions is information about the local site the conditions are taken from. This includes the latitude, longitude, and elevation of the site.

> part b - Adjust the commands demonstrated in class (used to get the extended forecast temperature information) to get these 3 pieces of information off your chosen page. Print the information to the screen from the website.

```r
weather_location_info <- weatherurl %>%
  read_html() %>%
  html_elements(".smallTxt") %>%
  pluck(1) %>%
  html_text()

print(weather_location_info)
```

```
[1] "Lat: 42.2°NLon: 72.53°WElev: 246ft."
```