Practice3S24

Sofia Guttmann

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:

```r
library(rvest)
library(dplyr)

url <- "https://en.wikipedia.org/wiki/List_of_highest-grossing_films"
webpage <- read_html(url)

table <- html_table(html_nodes(webpage, "table")[[1]])

table <- table %>%
  na.omit() %>%
  select_if(function(x) any(!is.na(x)))

head(table)
```

```
# A tibble: 6 x 6
   Rank Peak  Title                        `Worldwide gross`  Year Ref
  <int> <chr> <chr>                        <chr>             <int> <chr>
1     1 1     Avatar                       $2,923,706,026     2009 [# 1][# 2]
2     2 1     Avengers: Endgame            $2,797,501,328     2019 [# 3][# 4]
3     3 3     Avatar: The Way of Water     $2,320,250,281     2022 [# 5][# 6]
4     4 1     Titanic                      T$2,257,844,554    1997 [# 7][# 8]
5     5 3     Star Wars: The Force Awakens $2,068,223,624     2015 [# 9][# 10]
6     6 4     Avengers: Infinity War       $2,048,359,754     2018 [# 11][# 12]
```

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:

The inclusion of race/ethnicity as a predictor in a predictive model for determining the length of stay in homeless services raises significant ethical considerations. There may be arguments for including race/ethnicity, such as addressing systemic disparities and ensuring equitable allocation of resources. For instance, certain racial or ethnic groups may face disproportionate barriers to housing stability due to historical discrimination, socioeconomic factors, or systemic inequalities. Many times, people of color have a harder time receiving loans as a result of systemic racism and prejudice. By including race/ethnicity as a predictor, policymakers and service providers can potentially identify and target interventions to address these disparities, ultimately aiming to improve outcomes for marginalized groups. This means that some groups may need to receive more aid than others to reach a level playing field in the outcome.

However, there are also ethical concerns associated with including race/ethnicity in predictive modeling. Doing so may perpetuate or exacerbate existing biases and stereotypes, leading to unfair treatment or stigmatization of certain racial/ethnic groups. Additionally, there is a risk of reinforcing systemic inequalities by using race/ethnicity as a proxy for other underlying factors, such as socioeconomic status or access to resources. This can lead to further marginalization of already vulnerable populations and contribute to systemic discrimination. Therefore, careful consideration must be given to the potential consequences of including race/ethnicity as a predictor in predictive models, weighing the potential benefits of addressing disparities against the risks of perpetuating bias and discrimination. Ultimately, a thorough ethical analysis should guide decision-making to ensure that predictive modeling in homeless services is conducted in a fair, transparent, and equitable manner, with a focus on promoting social justice and addressing systemic inequalities. For example, many doctors believe that black people are more or less prone to certain diseases as a result of their anatomy rather than the social factors that play into the results. Therefore, including these factors could allow for misinterpretation of data.

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:

```r
library(robotstxt)

weather_url <- "https://forecast.weather.gov/MapClick.php?lat=42.3879975&lon=-72.530450799

robots <- tryCatch({
  robotstxt::robotstxt(weather_url)
}, error = function(e) {
  print("Error parsing robots.txt file")
  NULL
})
```

```
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
```

```r
is_allowed <- tryCatch({
  robotstxt::paths_allowed(robots, user_agent = "*")
}, error = function(e) {
  print("Error checking if URL is allowed")
  NULL
})
```

```
[1] "Error checking if URL is allowed"
```

```r
print(is_allowed)
```

```
NULL
```

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
# Load the required libraries
library(rvest)

# Define the URL
weather_url <- "https://forecast.weather.gov/MapClick.php?lat=42.3879975&lon=-72.530450799

# Read the HTML content from the URL
page <- read_html(weather_url)

# Extract latitude
latitude <- page %>%
  html_nodes(xpath = '//*[@id="current_conditions-summary"]/div[1]/div/p[2]') %>%
  html_text()

# Extract longitude
longitude <- page %>%
  html_nodes(xpath = '//*[@id="current_conditions-summary"]/div[1]/div/p[4]') %>%
  html_text()

# Extract elevation
elevation <- page %>%
  html_nodes(xpath = '//*[@id="current_conditions-summary"]/div[1]/div/p[6]') %>%
  html_text()

# Print the extracted information
cat("Latitude:", latitude, "\n")
```

```
Latitude:
```

```r
cat("Longitude:", longitude, "\n")
```

```
Longitude:
```

```r
cat("Elevation:", elevation, "\n")
```

```
Elevation:
```