AE 08: Scraping articles from the Cornell Review

Suggested answers

APPLICATION EXERCISE ANSWERS

MODIFIED
October 1, 2024

Packages

We will use the following packages in this application exercise.

- **tidyverse**: For data import, wrangling, and visualization.
- **rvest:** For scraping HTML files.
- **robotstxt:** For verifying if we can scrape a website.

```
library(tidyverse)
library(rvest)
library(robotstxt)
```

Data scraping

See the code below stored in `scrape-cornell-review.R`.

```
# load packages
library(tidyverse)
library(rvest)
library(robotstxt)

# check that we can scrape data from the cornell review
paths_allowed("https://www.thecornellreview.org/")

# read the first page
page <- read_html("https://www.thecornellreview.org/")

# extract desired components
titles <- html_elements(x = page, css = "#main .read-title a") |>
  html_text2()

authors <- html_elements(x = page, css = "#main .byline a") |>
  html_text2()

article_dates <- html_elements(x = page, css = "#main .posts-date") |>
```

```
  html_text2()

topics <- html_elements(x = page, css = "#main .cat-links") |>
  html_text2()

abstracts <- html_elements(x = page, css = ".post-description") |>
  html_text2()

post_urls <- html_elements(x = page, css = ".aft-readmore") |>
  html_attr(name = "href")

# create a tibble with this data
review_raw <- tibble(
  title = titles,
  author = authors,
  date = article_dates,
  topic = topics,
  description = abstracts,
  url = post_urls
)

# clean up the data
review <- review_raw |>
  mutate(
    date = mdy(date),
    description = str_remove(string = description, pattern = "\nRead More")
  )

# save to disk
write_csv(x = review, file = "data/cornell-review.csv")
```

Session information