Web Scraping Assignment

AUTHOR
Jin Sook Song

Web Scraping

```
# Load the necessary packages
pacman::p_load(rvest, dplyr, tidyverse)

# Install pacman if not already installed
if (!require("pacman")) install.packages("pacman")
```

```
Loading required package: pacman
```

Problem 1.

v1. name, email, phone in separate columns

```
# URL to scrape
url <- "https://www.american.edu/cas/mathstat/faculty/"

# Read the HTML content of the page
webpage <- read_html(url)

# Extract all profile sections
profiles <- webpage %>% html_nodes(".profile-item")

# Extract name (only first line), email, and phone; store as tibble
faculty_data <- lapply(profiles, function(profile) {
  # Extract the full name/title/department block
  name_block <- profile %>% html_node(".profile-name") %>% html_text(trim = TRUE)

  # Extract only the first line (actual name)
  name <- strsplit(name_block, "\n")[[1]][1] %>% str_trim()

  # Extract email and phone
  email <- profile %>%
    html_node(".profile-email span[itemprop='email']") %>%
    html_text(trim = TRUE)

  phone <- profile %>%
    html_node(".profile-phone span[itemprop='telephone']") %>%
    html_text(trim = TRUE)

  # Return tibble row
  tibble(name = name, email = email, phone = phone)
}) %>% bind_rows()

faculty_data
```

```
# A tibble: 40 × 3
   name              email                phone
   <chr>             <chr>                <chr>
 1 Jeffrey Adler     jadler@american.edu  (202) 885-3361
 2 Michael Baron     baron@american.edu   (202) 885-3130
 3 Maria Barouti     barouti@american.edu (202) 885-3132
 4 Laura Bernhofen   bernhofe@american.edu (202) 885-6806
```

```
 5 Zois Boukouvalas   boukouva@american.edu <NA>
 6 Stephen Casey      scasey@american.edu   (202) 885-3126
 7 Julia Chifman      chifman@american.edu  (202) 885-3686
 8 Olga Cordero-Brana corderob@american.edu (202) 885-6527
 9 Andrea Correll     acorrell@american.edu <NA>
10 Kristina Crona     kcrona@american.edu   (202) 885-3182
# ℹ 30 more rows
```

```r
        # Save the data to a CSV file
        write_csv(faculty_data, "faculty.csv")

        # Read the saved CSV file
        faculty_check <- read_csv("faculty.csv")
```

```
Rows: 40 Columns: 3
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (3): name, email, phone

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
        # Check the data
        print(faculty_check)
```

```
# A tibble: 40 × 3
   name               email                phone
   <chr>              <chr>                <chr>
 1 Jeffrey Adler      jadler@american.edu  (202) 885-3361
 2 Michael Baron      baron@american.edu   (202) 885-3130
 3 Maria Barouti      barouti@american.edu (202) 885-3132
 4 Laura Bernhofen    bernhofe@american.edu (202) 885-6806
 5 Zois Boukouvalas   boukouva@american.edu <NA>
 6 Stephen Casey      scasey@american.edu   (202) 885-3126
 7 Julia Chifman      chifman@american.edu  (202) 885-3686
 8 Olga Cordero-Brana corderob@american.edu (202) 885-6527
 9 Andrea Correll     acorrell@american.edu <NA>
10 Kristina Crona     kcrona@american.edu   (202) 885-3182
# ℹ 30 more rows
```

v2. name-email-phone in one column

```r
        # URL to scrape
        url <- "https://www.american.edu/cas/mathstat/faculty/"

        # Read the HTML content of the page
        webpage <- read_html(url)

        # Extract all profile sections
        profiles <- webpage %>% html_nodes(".profile-item")

        # Extract and format as a single-column tibble
        faculty_data1 <- lapply(profiles, function(profile) {
          # Get name (first line only)
          name_block <- profile %>% html_node(".profile-name") %>% html_text(trim = TRUE)
          name <- strsplit(name_block, "\n")[[1]][1] %>% str_trim()

          # Get email
          email <- profile %>%
            html_node(".profile-email span[itemprop='email']") %>%
            html_text(trim = TRUE)
```

```
    # Get phone
    phone <- profile %>%
      html_node(".profile-phone span[itemprop='telephone']") %>%
      html_text(trim = TRUE)

    # Combine into one string
    entry <- paste(name, email, phone, sep = " - ")

    # Return a one-column tibble
    tibble(name_email_phone = entry)
  }) %>% bind_rows()

  faculty_data1
```

```
# A tibble: 40 × 1
   name_email_phone
   <chr>
 1 Jeffrey Adler - jadler@american.edu - (202) 885-3361
 2 Michael Baron - baron@american.edu - (202) 885-3130
 3 Maria Barouti - barouti@american.edu - (202) 885-3132
 4 Laura Bernhofen - bernhofe@american.edu - (202) 885-6806
 5 Zois Boukouvalas - boukouva@american.edu - NA
 6 Stephen Casey - scasey@american.edu - (202) 885-3126
 7 Julia Chifman - chifman@american.edu - (202) 885-3686
 8 Olga Cordero-Brana - corderob@american.edu - (202) 885-6527
 9 Andrea Correll - acorrell@american.edu - NA
10 Kristina Crona - kcrona@american.edu - (202) 885-3182
# ℹ 30 more rows
```

```
  # Save the data to a CSV file
  write_csv(faculty_data1, "faculty1.csv")

  # Read the saved CSV file
  faculty_check1 <- read_csv("faculty1.csv")
```

```
Rows: 40 Columns: 1
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): name_email_phone

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
  # Check the data
  print(faculty_check1)
```

```
# A tibble: 40 × 1
   name_email_phone
   <chr>
 1 Jeffrey Adler - jadler@american.edu - (202) 885-3361
 2 Michael Baron - baron@american.edu - (202) 885-3130
 3 Maria Barouti - barouti@american.edu - (202) 885-3132
 4 Laura Bernhofen - bernhofe@american.edu - (202) 885-6806
 5 Zois Boukouvalas - boukouva@american.edu - NA
 6 Stephen Casey - scasey@american.edu - (202) 885-3126
 7 Julia Chifman - chifman@american.edu - (202) 885-3686
 8 Olga Cordero-Brana - corderob@american.edu - (202) 885-6527
 9 Andrea Correll - acorrell@american.edu - NA
10 Kristina Crona - kcrona@american.edu - (202) 885-3182
# ℹ 30 more rows
```