J.Kimbrough_DATA-413_WebscrapingHW

2024-11-12

GitHub Link:

https://github.com/jk7830a/HW_kimbrough/tree/5d2cbbeea7a1064471e3484ab11ae3c039431876/AllHW/WebscrapingHW

Loading Libraries

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.4.2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

establishing html

```r
#establishing html
html <- read_html("https://www.american.edu/cas/mathstat/faculty/index.cfm#all-math-fac")
print(html)
```

```
## {html_document}
## <html lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body class="CS_Document">\n<span role="navigation" aria-label="Top of pa ...
```

—Web Scraping—

scraping the full-time faculty & staff NAMES

```r
# scraping the full-time faculty & staff NAMES
staff_names <- html %>% html_nodes(".profile-name span") %>%
  html_text()
print(staff_names)
```

```
##  [1] "Jeffrey Adler"      "Michael Baron"      "Maria Barouti"
##  [4] "Laura Bernhofen"    "Zois Boukouvalas"   "Stephen Casey"
##  [7] "Julia Chifman"      "Olga Cordero-Brana" "Andrea Correll"
## [10] "Kristina Crona"     "Elizabeth Curran"   "James Dickens"
## [13] "Donna Dietz"        "Kevin Duke"         "Artur Elezi"
## [16] "David Gerard"       "Jeff Gill"          "Mary Gray"
## [19] "Jeffrey Hakim"      "William Howell"     "Monica Jackson"
## [22] "Behzad Jalali"      "Zeynep Kacar"       "Aleka Kapatou"
## [25] "Michael Keynes"     "Joshua Lansky"      "Mike Limarzi"
## [28] "Jun Lu"             "Betty Malloy"       "Nimai Mehta"
## [31] "Jaime Miller"       "Chris Mitchell"     "Ahmad Mousavi"
## [34] "Richard Ressler"    "Michael Robinson"   "Hamid Semiyari"
## [37] "Rebecca Steiner"    "Hugo Van Dyke"      "Zeying Wang"
## [40] "Whiting Wicker"
```

scraping the full-time faculty & staff POSITIONS

```r
# scraping the full-time faculty & staff POSITIONS
staff_positions <- html %>% html_nodes("span+ small") %>% html_text()
print(staff_positions)
```

```
##  [1] "Professor"
##  [2] "Professor"
##  [3] "Director of the Online MS in Data Science Program and Senior Professorial Lecturer"
##  [4] "Senior Professorial Lecturer"
##  [5] "Assistant Professor"
##  [6] "Professor"
##  [7] "Associate Professor"
##  [8] "Senior Professorial Lecturer"
##  [9] "Senior Administrative Assistant"
## [10] "Associate Professor"
## [11] "Professorial Lecturer"
## [12] "Senior Professorial Lecturer"
## [13] "Senior Professorial Lecturer (Continuing)"
## [14] "Senior Professorial Lecturer"
## [15] "Associate Professor"
## [16] "Associate Professor"
## [17] "Distinguished Professor"
## [18] "Distinguished Professor"
## [19] "Professor and Department Chair, Math & Statistics"
## [20] "Professorial Lecturer"
## [21] "Deputy Provost & Dean of Faculty"
## [22] "Director, Quantitative Support"
## [23] "Adjunct Professorial Lecturer"
## [24] "Hurst Senior Professorial Lecturer (Continuing)"
## [25] "Senior Professorial Lecturer"
## [26] "Professor"
## [27] "Hurst Senior Professorial Lecturer (Continuing)"
## [28] "Associate Professor"
## [29] "Professor and Director of the Data Science Programs"
## [30] "Hurst Senior Professorial Lecturer (Continuing)"
## [31] "Hurst Senior Professorial Lecturer (Continuing)"
## [32] "Administrator-in-Residence"
## [33] "Assistant Professor"
## [34] "Hurst Senior Professorial Lecturer"
## [35] "Professor"
## [36] "Senior Professorial Lecturer"
## [37] "Professorial Lecturer"
## [38] "Senior Professorial Lecturer"
## [39] "Professorial Lecturer"
## [40] "Adjunct Professorial Lecturer"
```

scraping the full-time faculty & staff EMAILS (if given)

```
# scraping the full-time faculty & staff EMAILS (if given)
## one staff member listed does NOT have an email
staff_email_raw <- html %>% html_nodes(".profile-email span") %>% html_text()
print(staff_email_raw)
```

```
##  [1] "jadler@american.edu"   "baron@american.edu"    "barouti@american.edu"
##  [4] "bernhofe@american.edu" "boukouva@american.edu" "scasey@american.edu"
##  [7] "chifman@american.edu"  "corderob@american.edu" "acorrell@american.edu"
## [10] "kcrona@american.edu"   "curran@american.edu"   "jdickens@american.edu"
## [13] "dietz@american.edu"    "kduke@american.edu"    "aelezi@american.edu"
## [16] "dgerard@american.edu"  "jgill@american.edu"    "mgray@american.edu"
## [19] "jhakim@american.edu"   "whowell@american.edu"  "monica@american.edu"
## [22] "bjalali@american.edu"  "kacar@american.edu"    "kapatou@american.edu"
## [25] "keynes@american.edu"   "lansky@american.edu"   "limarzi@american.edu"
## [28] "lu@american.edu"       "malloy@american.edu"   "jmiller@american.edu"
## [31] "mitchell@american.edu" "mousavi@american.edu"  "rressler@american.edu"
## [34] "michaelr@american.edu" "semiyari@american.edu" "rsteiner@american.edu"
## [37] "hvandyke@american.edu" "zwang@american.edu"    "wicker@american.edu"
```

scraping the full-time faculty & staff PHONE #s (if given)

```
# scraping the full-time faculty & staff PHONE #s (if given)
## every staff member listed does NOT have a number
staff_phone_raw <- html %>% html_nodes(".profile-phone span") %>% html_text()
print(staff_phone_raw)
```

```
##  [1] "(202) 885-3361" "(202) 885-3130" "(202) 885-3132" "(202) 885-6806"
##  [5] "(202) 885-3126" "(202) 885-3686" "(202) 885-6527" "(202) 885-3182"
##  [9] "(202) 885-6627" "(202) 885-6804" "(202) 885-3142" "(202) 885-3119"
## [13] "(202) 885-6275" "(202) 885-3171" "(202) 885-3131" "(202) 885-1509"
## [17] "(202) 885-3154" "(202) 885-6471" "(202) 885-3151" "(202) 885-3687"
## [21] "(202) 885-3146" "(202) 885-3042" "(202) 885-3614" "(202) 885-3017"
## [25] "(202) 885-3685" "(202) 885-5950" "(202) 885-6472" "(202) 885-3681"
## [29] "(202) 885-1447" "(202) 885-3684"
```

—Cleaning Data—

staff member with NO email

```
# staff member with NO email
no_email_staff <- c("Nimai Mehta")
```

staff members with NO phone number

```
# distinguishing which staff member did NOT have a number
no_phone_staff <- c("Zois Boukouvalas",
                         "Andrea Correll",
                         "James Dickens",
                         "David Gerard",
                         "William Howell",
                         "Zeynep Kacar",
                         "Chris Mitchell",
                         "Rebecca Steiner",
                         "Zeying Wang",
                         "Whiting Wicker")
```

empty vectors for emails and phone numbers

```
# initialize empty vector to store all emails
## keeps the same length as `staff_names`
staff_email <- vector("character", length(staff_names))
```

```r
# initialize empty vector to store all phone numbers
## keeps the same length as `staff_names`
staff_phone <- vector("character", length(staff_names))
```

for loop for emails

```r
# assigns NA or the correct email to each staff/faculty
email_index <- 1 #(to track the current position in the `staff_email_raw` list)
for (i in seq_along(staff_names)) {
  if (staff_names[i] %in% no_email_staff)
    {
    # (will set email to NA if the staff/faculty is in the `no_email_staff` list)
    staff_email[i] <- NA
    } else {
    # (assign the next available email from the list)
    staff_email[i] <- staff_email_raw[email_index]
    email_index <- email_index + 1  # (moves to the next email)
    }
}
```

for loop for phone numbers

```r
# loop over each name to assign phone numbers correctly
phone_index <- 1  # (tracks the current position in the `staff_phone_raw` list)
for (i in seq_along(staff_names)) {
  if (staff_names[i] %in% no_phone_staff)
    {
    # (will set phone # to NA for staff/faculty without a phone number in the `no_phone_staff` list)
    staff_phone[i] <- NA
    } else {
    # (assigns the next available phone number [hopefully keeping the right order])
    staff_phone[i] <- staff_phone_raw[phone_index]
    phone_index <- phone_index + 1  # (moves to the next phone number)
    }
}
```

—Creating Data & CSV—

create the final data frame

```r
# create the final data frame
au_math_stat_dep_staff <- data.frame(
  staff_names,
  staff_positions,
  staff_email,
  staff_phone,
  stringsAsFactors = FALSE #(will keep my data as characters rather than make them factors)
)
```

view the data before making it csv

```
# view the data before making it csv [making sure it's right like on the html page]
view(au_math_stat_dep_staff)
```

make the csv

```
# make the csv
write.csv(au_math_stat_dep_staff, "J.Kimbrough_DATA-413_AUMathDepartmentFaculty.csv")
```