SoDA 501_HWW3_Zhang

Lesley Zhang

2026-01-31

######PART 1##### #question 1 1. In the social sciences, what are two ethical or scientific risks of collecting data via web scraping (e.g., representativeness, privacy, terms of service, measurement error, scraping-induced missingness)? For each risk, briefly describe one practical mitigation strategy you would use in a reproducible workflow.

#answer One risk that web-scrapping often encounters is the unbalanced representation. For example, data available on web tend to show more information of individuals who are more publicly visible instead of representing the full population of interest. As a result, scraped datasets may systematically exclude less visible actors, leading to selection bias and limiting the generalizability of findings.

In addition, even when data are publicly accessible, web scraping can violate a website's terms of service or unintentionally collect sensitive personal information, such as email addresses, institutional affiliations, or time-stamped activity data. This raises ethical concerns related to privacy and consent and can also create legal or access risks for researchers, including IP blocking or account restrictions.

Researchers should clearly define the target population and document how the scraped sample may systematically differ from it, treating web data as a nonrandom sample rather than a complete census. In a reproducible workflow, this can be addressed by logging failed or missing scrape attempts, retaining metadata on coverage, and conducting sensitivity analyses that assess how results change when potentially overrepresented groups are downweighted or excluded. To address ethical and legal risks related to privacy and terms of service, researchers should limit data collection to information explicitly intended for public use, avoid scraping personally identifying or unnecessary fields, and adhere to platform-specific access rules such as rate limits. Documenting scraping decisions, access dates, and compliance considerations in code comments or a README, and sharing only derived or anonymized data when redistribution is restricted, further helps ensure that the research is both ethically responsible and reproducible.

######PART 2##### #question 3 and question 5

```
# ------------------------------------------------------------------------
# Setup
# ------------------------------------------------------------------------
# Install (if needed) and load the necessary libraries.

# install.packages(c("rvest", "dplyr", "ggplot2", "scholar", "stringr", "tibble"))
library(rvest)
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(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```
library(scholar)
library(stringr)
library(tibble)

# ------------------------------------------------------------------------------
# Part 2: Pulling Google Scholar Data (Citations Over Time)
# ------------------------------------------------------------------------------
# Goal:
# - For each professor, we will:
#   (1) Define the Google Scholar ID
#   (2) Pull a profile summary
#   (3) Pull publications (and view the first 5)
#   (4) Pull citation history by year
#   (5) Combine all citation histories into one table and plot them

# ------------------------------------------------------------------------------
# Step 1: Hard-code Google Scholar IDs
# ------------------------------------------------------------------------------
jwright_scholar_id   <- "DV5ECYgAAAAJ"
jedgerton_scholar_id   <- "LLcIlUkAAAAJ"
bdesmarais_scholar_id  <- "fRM8IN4AAAAJ"
cloyle_scholar_id  <- "IMUIrJMAAAAJ"
xcao_scholar_id <- "w18ZmkEAAAAJ"
slinn_scholar_id <- "I7Jx1fAAAAAJ"
rmcmanus_scholar_id   <- "3xe3Ck4AAAAJ"
bmukherjee_scholar_id   <- "6sS40fEAAAAJ"
dtavana_scholar_id  <- "j2a1_doAAAAJ"
vyadav_scholar_id  <- "vGjxl7YAAAAJ"

# ------------------------------------------------------------------------------
# Step 2: Pull Google Scholar profiles (sequentially)
# ------------------------------------------------------------------------------

scholars <- tibble::tibble(
  name = c(
    "Joe Wright",
    "Jared Edgerton",
    "Bruce Desmarais",
    "Cyanne Loyle",
    "Xun Cao",
    "Susanne Linn",
    "Roseanne McManus",
    "Bumba Mukherjee",
    "Daniel Tavana",
    "Vineeta Yadav"
```

```
  ),
  scholar_id = c(
    jwright_scholar_id,
    jedgerton_scholar_id,
    bdesmarais_scholar_id,
    cloyle_scholar_id,
    xcao_scholar_id,
    slinn_scholar_id,
    rmcmanus_scholar_id,
    bmukherjee_scholar_id,
    dtavana_scholar_id,
    vyadav_scholar_id
  )
)

# ------------------------------------------------------------------------------
# Step 2: Pull Google Scholar profiles (sequentially)
# ------------------------------------------------------------------------------

jwright_name   <- "Joe Wright"
jedgerton_name <- "Jared Edgerton"
bdesmarais_name<- "Bruce Desmarais"
cloyle_name    <- "Cyanne Loyle"
xcao_name      <- "Xun Cao"
slinn_name     <- "Susanne Linn"
rmcmanus_name  <- "Roseanne McManus"
bmukherjee_name<- "Bumba Mukherjee"
dtavana_name   <- "Daniel Tavana"
vyadav_name    <- "Vineeta Yadav"


jwright_profile    <- get_profile(jwright_scholar_id)
jedgerton_profile  <- get_profile(jedgerton_scholar_id)
bdesmarais_profile <- get_profile(bdesmarais_scholar_id)
cloyle_profile     <- get_profile(cloyle_scholar_id)
xcao_profile       <- get_profile(xcao_scholar_id)
slinn_profile      <- get_profile(slinn_scholar_id)
rmcmanus_profile   <- get_profile(rmcmanus_scholar_id)
bmukherjee_profile <- get_profile(bmukherjee_scholar_id)
dtavana_profile    <- get_profile(dtavana_scholar_id)
vyadav_profile     <- get_profile(vyadav_scholar_id)
```

```
cat("\n------------------------------\n")
```

```
##
## ------------------------------
```

```
cat("Google Scholar Profile Summaries\n")
```

```
## Google Scholar Profile Summaries
```

```r
cat("------------------------------\n")
```

```
## ------------------------------
```

```r
cat("\n", jwright_name, "\n", sep = "")
```

```
##
## Joe Wright
```

```r
print(jwright_profile)
```

```
## $id
## [1] "DV5ECYgAAAAJ"
##
## $name
## [1] "Joseph Wright"
##
## $affiliation
## [1] "Pennsylvania State University"
##
## $total_cites
## [1] 9699
##
## $h_index
## [1] 38
##
## $i10_index
## [1] 52
##
## $fields
## [1] "Authoritarianism"     "Comparative Politics" "Democratization"
## [4] "Foreign Aid"          "Political Science"
##
## $homepage
## [1] "http://sites.psu.edu/wright/"
##
## $coauthors
##  [1] "Erica Frantz"             "Abel Escribà-Folch"
##  [3] "Barbara Geddes"           "Simone Dietrich"
##  [5] "Covadonga Meseguer"       "David B. Carter"
##  [7] "Heather Boushey"          "Deniz Aksoy"
##  [9] "John Chin"                "George Derpanopoulos"
## [11] "Wonjun Song"              "Matthew S. Winters"
## [13] "Jia Li"                   "Xu Xu"
## [15] "Daehee Bak"               "Boliang Zhu"
## [17] "Matthew Charles Wilson"   "Margaret (Molly) Ariotti"
## [19] "Sophia McClennen"         "Elizabeth Stein"
##
## $available
## [1] 8
##
## $not_available
## [1] 0
```

```r
cat("\n", jedgerton_name, "\n", sep = "")
```

```
##
## Jared Edgerton
```

```r
print(jedgerton_profile)
```

```
## $id
## [1] "LLcIlUkAAAAJ"
##
## $name
## [1] "Jared F. Edgerton"
##
## $affiliation
## [1] "Assistant Professor, Pennsylvania State University"
##
## $total_cites
## [1] 654
##
## $h_index
## [1] 7
##
## $i10_index
## [1] 6
##
## $fields
## [1] "Artificial Intelligence" "Network Science"
## [3] "Deep Learning"           "International Security"
## [5] "Conflict Processes"
##
## $homepage
## [1] "http://jaredfedgerton.net/"
##
## $coauthors
##  [1] "Daniel Naftel"     "Skyler Cranmer"    "Jon Green"
##  [4] "Kelsey Shoub"      "Rongjun Qin"       "Erin Lin"
##  [7] "Sort by citations" "Sort by year"      "Sort by title"
## [10] "About Scholar"     "Search help"
##
## $available
## [1] 3
##
## $not_available
## [1] 1
```

```r
cat("\n", bdesmarais_name, "\n", sep = "")
```

```
##
## Bruce Desmarais
```

```r
print(bdesmarais_profile)
```

```
## $id
## [1] "fRM8IN4AAAAJ"
##
## $name
## [1] "Bruce A. Desmarais"
##
## $affiliation
## [1] "Professor, Department of Political Science, Penn State University"
##
## $total_cites
## [1] 4436
##
## $h_index
## [1] 30
##
## $i10_index
## [1] 47
##
## $fields
## [1] "political institutions"       "political methodology"
## [3] "computational social science" "network analysis"
## [5] "machine learning"
##
## $homepage
## [1] "http://brucedesmarais.com/"
##
## $coauthors
##  [1] "Skyler Cranmer"     "Jeffrey J. Harden"  "Frederick Boehmke"
##  [4] "Hanna Wallach"      "Philip Leifeld"     "Matthew J. Denny"
##  [7] "John A. Hird"       "Shankar Bhamidi"    "Michael S. Kowal"
## [10] "James David Wilson" "Raymond La Raja"    "Justin H. Kirkland"
## [13] "Vin Moscardelli"    "Mia Costa"          "Tobias Heinrich"
## [16] "Eugenia Giraudy"    "Matthew Burgess"    "Brian Schaffner"
## [19] "Juston Moore"       "Thomas M. Carsey"
##
## $available
## [1] 37
##
## $not_available
## [1] 0
```

```r
cat("\n", cloyle_name, "\n", sep = "")
```

```
##
## Cyanne Loyle
```

```r
print(cloyle_profile)
```

```
## $id
```

```
## [1] "IMUIrJMAAAAJ"
##
## $name
## [1] "Cyanne E. Loyle"
##
## $affiliation
## [1] "Professor, Pennsylvania State University & PRIO"
##
## $total_cites
## [1] 1618
##
## $h_index
## [1] 19
##
## $i10_index
## [1] 26
##
## $fields
## [1] "Conflict Studies"     "Human Rights"         "Transitional Justice"
##
## $homepage
## [1] "http://www.cyanneloyle.com/"
##
## $coauthors
##  [1] "[christian davenport]"         "Helga Malmin Binningsbø"
##  [3] "Kathleen Gallagher Cunningham" "Danielle F. Jung"
##  [5] "Benjamin Appel"                "Reyko Huang"
##  [7] "Jon Elster"                    "Scott Gates"
##  [9] "Christopher Michael Sullivan"  "Samuel E. Bestvater"
## [11] "Jessica Maves Braithwaite"     "R Joseph Huddleston"
## [13] "Anjanette (Angie) Raymond"     "Federica Carugati"
## [15] "Michael A. Rubin"              "Scott Shackelford"
## [17] "Abbey Stemler"                 "Jessica Steinberg"
## [19] "Haley Swedlund"                "Ilayda B. Onder"
##
## $available
## [1] 2
##
## $not_available
## [1] 3
```

```
cat("\n", xcao_name, "\n", sep = "")
```

```
##
## Xun Cao
```

```
print(xcao_profile)
```

```
## $id
## [1] "w18ZmkEAAAAJ"
##
## $name
## [1] "Xun Cao"
```

```
##
## $affiliation
## [1] "Penn State University"
##
## $total_cites
## [1] 2883
##
## $h_index
## [1] 25
##
## $i10_index
## [1] 31
##
## $fields
## [1] "political economy"      "climate change"         "environment and energy"
## [4] "conflicts"              "political geography"
##
## $homepage
## [1] "http://sites.psu.edu/xuncao/research"
##
## $coauthors
##  [1] "Aseem Prakash"           "Michael D. Ward"
##  [3] "Hugh Ward"               "Chuyu Liu"
##  [5] "Genia Kostka"            "Adam Luedtke"
##  [7] "Christian Breunig"       "Kristin M. Bakke"
##  [9] "James Piazza"            "Theodora-Ismene Gizelis"
## [11] "Brian Greenhill"         "helen v. milner"
## [13] "Bumba Mukherjee"         "Anthony A. Pezzola"
## [15] "Amanda Fidalgo"          "Andrew N. Kleit"
## [17] "Sort by citations"       "Sort by year"
## [19] "Sort by title"           "About Scholar"
##
## $available
## [1] 4
##
## $not_available
## [1] 1
```

```
cat("\n", slinn_name, "\n", sep = "")
```

```
##
## Susanne Linn
```

```
print(slinn_profile)
```

```
## $id
## [1] "I7Jx1fAAAAAJ"
##
## $name
## [1] "Suzanna Linn"
##
## $affiliation
## [1] "Professor of Political Science, Penn State University"
```

```
##
## $total_cites
## [1] 5217
##
## $h_index
## [1] 26
##
## $i10_index
## [1] 39
##
## $fields
## [1] "American politics"    "elections"            "public opinion"
## [4] "time series analysis"
##
## $homepage
## character(0)
##
## $coauthors
##  [1] "Amber Boydstun"          "Frank Baumgartner"
##  [3] "Janet Box-Steffensmeier" "Luke Keele"
##  [5] "Jonathan Nagler"         "Paul M. Kellstedt"
##  [7] "Tse-min Lin"             "Jim Granato"
##  [9] "Sort by citations"       "Sort by year"
## [11] "Sort by title"           "About Scholar"
## [13] "Search help"
##
## $available
## [1] 3
##
## $not_available
## [1] 0
```

```
cat("\n", rmcmanus_name, "\n", sep = "")
```

```
##
## Roseanne McManus
```

```
print(rmcmanus_profile)
```

```
## $id
## [1] "3xe3Ck4AAAAJ"
##
## $name
## [1] "Roseanne McManus"
##
## $affiliation
## [1] "Professor of Political Science, Pennsylvania State University"
##
## $total_cites
## [1] 1963
##
## $h_index
## [1] 14
```

```
##
## $i10_index
## [1] 14
##
## $fields
## [1] "International security" "US foreign policy"
##
## $homepage
## [1] "https://sites.psu.edu/roseannemcmanus/"
##
## $coauthors
##  [1] "Michael R. Kenwick"   "Vito D'Orazio"        "Timothy Nordstrom"
##  [4] "Jon CW Pevehouse"     "Kayla Kahn"           "Michael J. Soules"
##  [7] "Nick Dietrich"        "Andrew Kydd"          "Anne Spencer Jamison"
## [10] "Mikaela Karstens"     "Michael Goldfien"     "Michael F Joseph"
## [13] "Mark David Nieman"    "Sort by citations"    "Sort by year"
## [16] "Sort by title"        "About Scholar"        "Search help"
##
## $available
## [1] 3
##
## $not_available
## [1] 0
```

```
cat("\n", bmukherjee_name, "\n", sep = "")
```

```
##
## Bumba Mukherjee
```

```
print(bmukherjee_profile)
```

```
## $id
## [1] "6sS40fEAAAAJ"
##
## $name
## [1] "Bumba Mukherjee"
##
## $affiliation
## [1] "Professor, Political Science, Penn State University"
##
## $total_cites
## [1] 4606
##
## $h_index
## [1] 26
##
## $i10_index
## [1] 50
##
## $fields
## [1] "International Political Economy" "Civil Conflict"
## [3] "Statistical Methodology"
##
```

```
## $homepage
## [1] "https://sites.psu.edu/bumbamukherjee/"
##
## $coauthors
##  [1] "Benjamin E. Bagozzi" "Vineeta Yadav"       "Will H. Moore"
##  [4] "David Leblang"       "Ore Koren"           "Sergio Béjar"
##  [7] "helen v. milner"     "Daniel W. Hill, Jr." "Minnie Minhyung Joo"
## [10] "David Andrew Singer" "Brandon Bolte"       "Nguyen Khoi Huynh"
## [13] "Nathan Jensen"       "Xun Cao"             "Hugh Ward"
## [16] "Quan Li"             "Justin Esarey"       "Nicolás Schmidt"
## [19] "Anna Harvey"         "Alexandra Guisinger"
##
## $available
## [1] 2
##
## $not_available
## [1] 1
```

```
cat("\n", dtavana_name, "\n", sep = "")
```

```
##
## Daniel Tavana
```

```
print(dtavana_profile)
```

```
## $id
## [1] "j2a1_doAAAAJ"
##
## $name
## [1] "Daniel L. Tavana"
##
## $affiliation
## [1] "Assistant Professor, Penn State"
##
## $total_cites
## [1] 159
##
## $h_index
## [1] 6
##
## $i10_index
## [1] 3
##
## $fields
## [1] "voter behavior"         "survey research"        "elections"
## [4] "authoritarian politics"
##
## $homepage
## [1] "http://danieltavana.com/"
##
## $coauthors
## [1] "Christiana Parreira" "Rory Truex"          "charles harb"
## [4] "Courtney Freer"      "Sort by citations"   "Sort by year"
```

```
## [7] "Sort by title"       "About Scholar"       "Search help"
##
## $available
## [1] 1
##
## $not_available
## [1] 0
```

```
cat("\n", vyadav_name, "\n", sep = "")
```

```
##
## Vineeta Yadav
```

```
print(vyadav_profile)
```

```
## $id
## [1] "vGjxl7YAAAAJ"
##
## $name
## [1] "Vineeta Yadav"
##
## $affiliation
## [1] "Department of Political Science, Penn State University"
##
## $total_cites
## [1] 481
##
## $h_index
## [1] 11
##
## $i10_index
## [1] 11
##
## $fields
## [1] "Political Economy"
##
## $homepage
## character(0)
##
## $coauthors
## [1] "Sort by citations" "Sort by year"      "Sort by title"
## [4] "About Scholar"     "Search help"
##
## $available
## [1] 0
##
## $not_available
## [1] 1
```

```
# ------------------------------------------------------------------------
# Step 3: Pull Google Scholar publications (sequentially)
# ------------------------------------------------------------------------
```

```r
jwright_pubs    <- get_publications(jwright_scholar_id)
jedgerton_pubs  <- get_publications(jedgerton_scholar_id)
bdesmarais_pubs <- get_publications(bdesmarais_scholar_id)
cloyle_pubs     <- get_publications(cloyle_scholar_id)
xcao_pubs       <- get_publications(xcao_scholar_id)
slinn_pubs      <- get_publications(slinn_scholar_id)
rmcmanus_pubs   <- get_publications(rmcmanus_scholar_id)
bmukherjee_pubs <- get_publications(bmukherjee_scholar_id)
dtavana_pubs    <- get_publications(dtavana_scholar_id)
vyadav_pubs     <- get_publications(vyadav_scholar_id)

cat("\n------------------------------\n")
```

```
##
## ------------------------------
```

```r
cat("Recent Publications (first 5)\n")
```

```
## Recent Publications (first 5)
```

```r
cat("------------------------------\n")
```

```
## ------------------------------
```

```r
cat("\n", jwright_name, "\n", sep = "")
```

```
##
## Joe Wright
```

```r
print(head(jwright_pubs, 5))
```

```
##                                                                                 title
## 1                                  Autocratic breakdown and regime transitions: A new data set
## 2                                How dictatorships work: Power, personalization, and collapse
## 3 Do authoritarian institutions constrain? How legislatures affect economic growth and investment
## 4          Dealing with tyranny: International sanctions and the survival of authoritarian rulers
## 5                            How foreign aid can foster democratization in authoritarian regimes
##                         author                                  journal
## 1 B Geddes, J Wright, E Frantz                Perspectives on Politics
## 2 B Geddes, J Wright, E Frantz New York, NY: Cambridge University Press
## 3                     J Wright    American Journal of Political Science
## 4    A Escribà-Folch, J Wright          International studies quarterly
## 5                     J Wright    American journal of political science
##           number cites year
## 1 12 (2), 313-331  1989 2014
## 2                   995 2018
## 3 52 (2), 322-343   931 2008
## 4 54 (2), 335-359   494 2010
## 5 53 (3), 552-571   457 2009
##
```

```
## 1 12889078556620013812,6965286960134511402,690127994017231 29,11638530267385500476,858304667648725480
## 2
## 3
## 4
## 5
##          pubid
## 1 lYbyOjaXH8MC
## 2 5s9rAH04UEoC
## 3 b0M2c_1WBrUC
## 4 qjMakFHDy7sC
## 5 9yKSN-GCB0IC
```

```r
cat("\n", jedgerton_name, "\n", sep = "")
```

```
##
## Jared Edgerton
```

```r
print(head(jedgerton_pubs, 5))
```

```
##
## 1                                    Elusive consensus: Polarization in elite communication on the CO
## 2                          A quasi-experimental evaluation of the impact of public assistance on pris
## 3                                      Understanding trends in hate crimes against immigrants and His
## 4 Crater detection from commercial satellite imagery to estimate unexploded ordnance in Cambodian ag
## 5                                                          Analyzing participation in the 1994 gen
##                                                                      author
## 1                          J Green, J Edgerton, D Naftel, K Shoub, SJ Cranmer
## 2                                           J Luallen, J Edgerton, D Rabideau
## 3 M Shively, R Subramanian, O Drucker, J Edgerton, J McDevitt, A Farrell, ...
## 4                                            E Lin, R Qin, J Edgerton, D Kong
## 5                                H Nyseth Nzitatira, JF Edgerton, LC Frizzell
##                              journal           number cites year
## 1                    Science advances 6 (28), eabc2717   481 2020
## 2 Journal of Quantitative Criminology  34 (3), 741-773    42 2018
## 3                            Contract      2010, 10098    33 2014
## 4                            Plos one 15 (3), e0229826    20 2020
## 5           Journal of peace research  60 (2), 291-306    16 2023
##                    cid        pubid
## 1  4129238246377963800 qjMakFHDy7sC
## 2 14585942090835837924 zYLM7Y9cAGgC
## 3  8331876169121057160 Y0pCki6q_DkC
## 4 17231763820797827088 UeHWp8X0CEIC
## 5  5199440090935849786 5nxA0vEk-isC
```

```r
cat("\n", bdesmarais_name, "\n", sep = "")
```

```
##
## Bruce Desmarais
```

```r
print(head(bdesmarais_pubs, 5))
```

```
##                                                                                                  titl
## 1                                      Inferential network analysis with exponential random graph models
## 2 Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals
## 3                       Persistent policy pathways: Inferring diffusion networks in the American states
## 4                                                Complex dependencies in the alliance networ
## 5                        Testing for zero inflation in count models: Bias correction for the Vuong tes
##                                    author                                journal
## 1              SJ Cranmer, BA Desmarais                    Political analysis
## 2   P Leifeld, SJ Cranmer, BA Desmarais       Journal of Statistical Software
## 3   BA Desmarais, JJ Harden, FJ Boehmke     American political science review
## 4 SJ Cranmer, BA Desmarais, EJ Menninga Conflict management and peace science
## 5               BA Desmarais, JJ Harden                      The Stata Journal
##              number cites year                                  cid
## 1    19 (1), 66-86   671 2011                  6006786222391641107
## 2         83, 1-36   350 2018  4233572769257845403,5439272305831502570
## 3 109 (2), 392-406   242 2015 18103161049128175187,3713119990166792933
## 4  29 (3), 279-313   234 2012                   5850118939809938330
## 5  13 (4), 810-835   202 2013                  12921593961079245307
##          pubid
## 1 u5HHmVD_uO8C
## 2 9ZlFYXVOiuMC
## 3 4DMP91E08xMC
## 4 d1gkVwhDpl0C
## 5 5nxA0vEk-isC
```

```
cat("\n", cloyle_name, "\n", sep = "")
```

```
##
## Cyanne Loyle
```

```
print(head(cloyle_pubs, 5))
```

```
##                                                                                 title
## 1                      Armed conflict and post-conflict justice, 1946-2006: A dataset
## 2 Transitional injustice: Subverting justice in transition and postconflict societies
## 3                                    New directions in rebel governance research
## 4                                           Rebel justice during armed conflict
## 5     Justice during armed conflict: A new dataset on government and rebel strategies
##                                    author                        journal
## 1 HM Binningsbø, CE Loyle, S Gates, J Elster      Journal of Peace Research
## 2                      CE Loyle, C Davenport        Journal of Human Rights
## 3  CE Loyle, KG Cunningham, R Huang, DF Jung       Perspectives on Politics
## 4                                CE Loyle Journal of Conflict Resolution
## 5                    CE Loyle, HM Binningsbø Journal of conflict resolution
##            number cites year                                  cid
## 1 49 (5), 731-740   195 2012  8194982990194069316,17604960030813510272
## 2 15 (1), 126-149   125 2016   8624884034161998865,2346144168066242355
## 3 21 (1), 264-276   123 2023                   7197643048742268496
## 4 65 (1), 108-134   119 2021 10811109672491297131,18064199429841217126
## 5 62 (2), 442-466   108 2018                   5236977625418974050
##          pubid
## 1 u-x6o8ySG0sC
## 2 5nxA0vEk-isC
```

```
## 3 bEWYMUwI8FkC
## 4 e5wmG9Sq2KIC
## 5 KlAtU1dfN6UC
```

```
cat("\n", xcao_name, "\n", sep = "")
```

```
##
## Xun Cao
```

```
print(head(xcao_pubs, 5))
```

```
##
## 1                                                                                  Disputes, democracies, an
## 2 Greening the career incentive structure for local officials in China: Does less pollution increase
## 3                                                 Networks as channels of policy diffusion: Explain
## 4                                                      Networks of intergovernmental organiza
## 5                                                                                                  Dom
##                        author                                          journal
## 1 MD Ward, RM Siverson, X Cao          American journal of political science
## 2                 M Wu, X Cao Journal of Environmental Economics and Management
## 3                       X Cao                International Studies Quarterly
## 4                       X Cao                International Studies Quarterly
## 5               H Ward, X Cao                  Comparative Political Studies
##              number cites year
## 1   51 (3), 583-601   268 2007
## 2       107, 102440   240 2021
## 3   54 (3), 823-854   205 2010
## 4 53 (4), 1095-1130   199 2009
## 5 45 (9), 1075-1103   160 2012
##                                                                cid        pubid
## 1 9436046912141842014,7006577203637690885,6191265740987493681 u5HHmVD_uO8C
## 2                                    13343465540681945976 u_35RYKgDlwC
## 3                                     9251072475029652791 2osOgNQ5qMEC
## 4                                     5383673927489534701 u-x6o8ySGOsC
## 5                  11862212908679567538,10126636752852418171 4TOpqqG69KYC
```

```
cat("\n", slinn_name, "\n", sep = "")
```

```
##
## Susanne Linn
```

```
print(head(slinn_pubs, 5))
```

```
##                                                    title
## 1                                Taking time seriously
## 2 The decline of the death penalty and the discovery of innocence
## 3                      The dynamics of the partisan gender gap
## 4    The political (and economic) origins of consumer confidence
## 5  Repeated events survival models: the conditional frailty model
##                                author                                journal
## 1                  S De Boef, L Keele American journal of political science
```

```
## 2 FR Baumgartner, SL De Boef, AE Boydstun            Cambridge University Press
## 3 JM Box-Steffensmeier, S De Boef, TM Lin     American Political Science Review
## 4                 S De Boef, PM Kellstedt American Journal of Political Science
## 5         JM Box-Steffensmeier, S De Boef                Statistics in medicine
##               number cites year                  cid        pubid
## 1    52 (1), 184-200  1196 2008 11221633523384872300 Tyk-4Ss8FVUC
## 2                      683 2008  2722092506517403418 hC7cP41nSMkC
## 3    98 (3), 515-528   490 2004  9015462784536945332 Y0pCki6q_DkC
## 4    48 (4), 633-649   453 2004 12174132703337415332 W7OEmFMy1HYC
## 5 25 (20), 3518-3533   258 2006  8834796008477273355 YsMSGLbcyi4C
```

```
cat("\n", rmcmanus_name, "\n", sep = "")
```

```
##
## Roseanne McManus
```

```
print(head(rmcmanus_pubs, 5))
```

```
##                                                                                                                tit
## 1                                                The COW-2 International Organizations Dataset Version 3
## 2                    Tracking organizations in the world: The Correlates of War IGO Version 3.0 dataset
## 3                                The MID5 Dataset, 2011-2014: Procedures, coding rules, and descriptio
## 4 The Logic of "Offstage" Signaling: Domestic Politics, Regime Type, and Major Power-Protégé Relation
## 5                          Making it personal: The role of leader-specific signals in extended deterren
##                                                                 author
## 1             JC Pevehouse, R McManus, T Nordstrom, M Shannon, M Widmann
## 2                      JCW Pevehouse, T Nordstrom, RW McManus, AS Jamison
## 3 G Palmer, RW McManus, V D'Orazio, MR Kenwick, M Karstens, C Bloch, ...
## 4                                              RW McManus, K Yarhi-Milo
## 5                                                        RW McManus
##                                journal          number cites year
## 1                    Find this resource                     692 2016
## 2             Journal of Peace Research 57 (3), 492-503   262 2020
## 3 Conflict Management and Peace Science 39 (4), 470-482   200 2022
## 4            International Organization 71 (4), 701-733   138 2017
## 5               The Journal of Politics 80 (3), 982-995   109 2018
##                                                          cid        pubid
## 1                   8839044189782749212,15830011968727745174 Zph67rFs4hoC
## 2                                    12703616314352759668 MXK_kJrjxJIC
## 3 8563666843921224625,7473288045390631413,4024108108534268898 qxL8FJ1GzNcC
## 4                   14338509058313284013,4064813178228415493 YsMSGLbcyi4C
## 5                                    17782245686854228042 LkGwnXOMwfcC
```

```
cat("\n", bmukherjee_name, "\n", sep = "")
```

```
##
## Bumba Mukherjee
```

```
print(head(bmukherjee_pubs, 5))
```

```
##
```

```
## 1                            Decentralization and Accountability of Infrastructure Delivery in Develo
## 2                                          Decentralizing Anti-Poverty Program Delivery in Develo
## 3                                                                  Democratization and economi
## 4 Government partisanship, elections, and the stock market: Examining American and British stock retu
## 5                                  Corruption and Decentralization of Infrastructure Delivery in Develo
##                   author                                journal          number
## 1         PBD Mookherjee                      Economic Journal    116, 101-127
## 2           D Mookherjee           Journal of Public Economics     89, 675-704
## 3 HV Milner, B Mukherjee    Annual Review of Political Science 12 (1), 163-181
## 4 D Leblang, B Mukherjee American journal of political science 49 (4), 780-802
## 5              Mukherjee                      Economic Journal    116, 107-133
##   cites year                  cid        pubid
## 1   835 2006 14774776143133816367 vV6vV6tmYwMC
## 2   711 2005 10296943600175396955 35N4QoGY0k4C
## 3   305 2009 10554047441228252620 u5HHmVD_uO8C
## 4   281 2005  4766021211448773859 u-x6o8ySG0sC
## 5   260 2006  1244275812743458493 70eg2SAEIzsC
```

```
cat("\n", dtavana_name, "\n", sep = "")
```

```
##
## Daniel Tavana
```

```
print(head(dtavana_pubs, 5))
```

```
##                                                                                 title
## 1                                  Implicit attitudes toward an authoritarian regime
## 2                 Party proliferation and electoral transition in post-Mubarak Egypt
## 3                                Tunisia’s Parliamentary and Presidential Elections
## 4                              Ethnic political socialization and university elections
## 5 Cooptation in practice: measuring legislative opposition in an authoritarian regime
##                          author                          journal
## 1            R Truex, DL Tavana          The Journal of Politics
## 2                     DL Tavana       North Africa’s Arab Spring
## 3          D Tavana, A Russell Project on Middle East Democracy
## 4 C Parreira, DL Tavana, C Harb                   Party Politics
## 5             DL Tavana, E York                              OSF
##              number cites year                  cid        pubid
## 1 81 (3), 1014-1027    55 2019  3360068741651472142 qjMakFHDy7sC
## 2             51-67    35 2013 13820840156465843225 2osOgNQ5qMEC
## 3                      13 2014 16838240354251088819 W7OEmFMy1HYC
## 4   30 (3), 550-569     7 2024  4008036149451681787 Se3iqnhoufwC
## 5                       7 2020  9377810954808121289 eQOLeE2rZwMC
```

```
cat("\n", vyadav_name, "\n", sep = "")
```

```
##
## Vineeta Yadav
```

```
print(head(vyadav_pubs, 5))
```

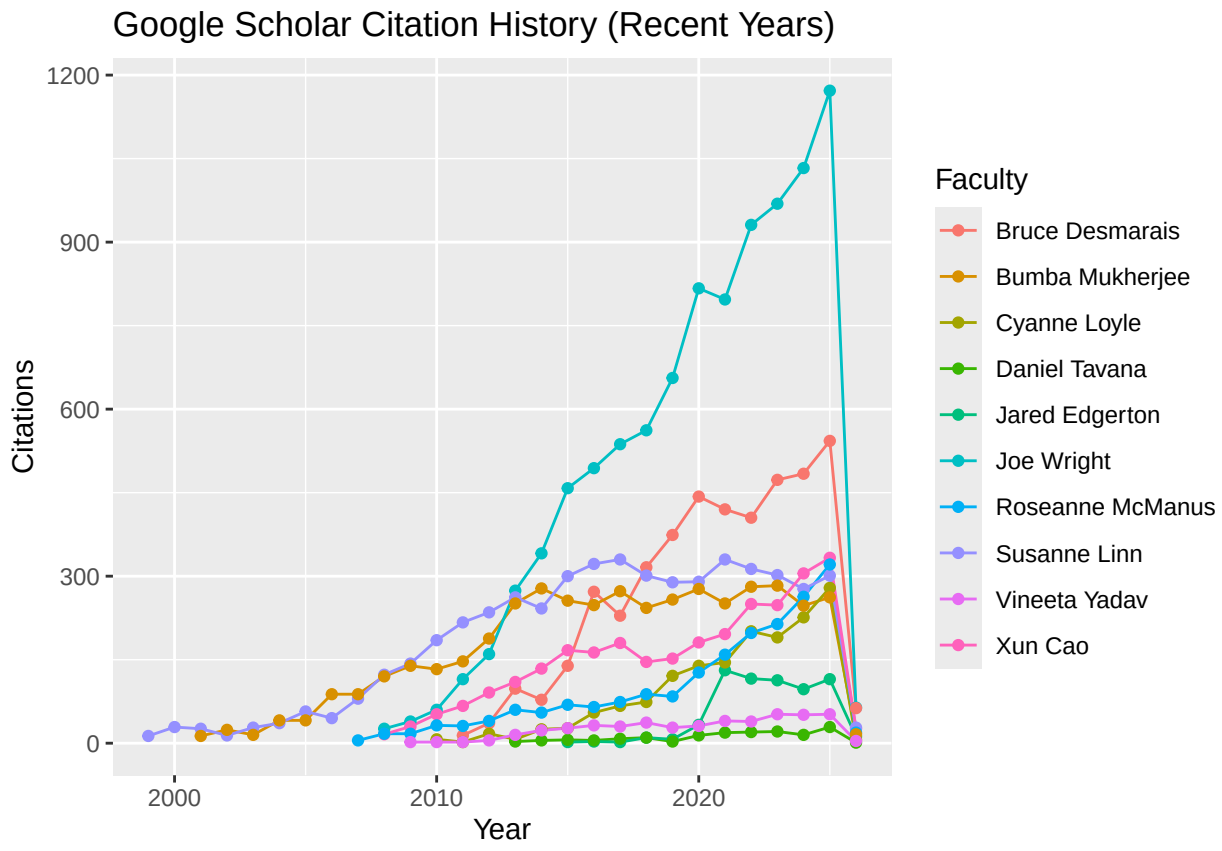
```
##                                                                                                                titl
## 1                          Political parties, business groups, and corruption in developing countrie
## 2                                                        The politics of corruption in dictatorship
## 3 Business lobbies and policymaking in developing countries: The contrasting cases of India and Chin
## 4                           Legislative institutions and corruption in developing country democracie
## 5                       Democracy, electoral systems, and judicial empowerment in developing countrie
##                  author                                                 journal
## 1               V Yadav                                Oxford University Press
## 2 V Yadav, B Mukherjee                               Cambridge University Press
## 3               V Yadav Journal of Public Affairs: An International Journal
## 4               V Yadav                           Comparative Political Studies
## 5 V Yadav, B Mukherjee                            University of Michigan Press
##               number cites year                cid        pubid
## 1                       127 2011  9662190811686073487 u5HHmVD_uO8C
## 2                        78 2016 11383904939677303609 roLk4NBRz8UC
## 3    8 (1-2), 67-82      54 2008 16526142858454422709 u-x6o8ySG0sC
## 4 45 (8), 1027-1058      46 2012  4244966706819684297 d1gkVwhDpl0C
## 5                        23 2014  9531708904108609377 UeHWp8X0CEIC
```

```r
# ------------------------------------------------------------------------------
# Step 4: Pull citation history (citations by year) and combine
# ------------------------------------------------------------------------------
jwright_ct    <- get_citation_history(jwright_scholar_id)    %>% mutate(name = jwright_name)
jedgerton_ct  <- get_citation_history(jedgerton_scholar_id)  %>% mutate(name = jedgerton_name)
bdesmarais_ct <- get_citation_history(bdesmarais_scholar_id) %>% mutate(name = bdesmarais_name)
cloyle_ct     <- get_citation_history(cloyle_scholar_id)     %>% mutate(name = cloyle_name)
xcao_ct       <- get_citation_history(xcao_scholar_id)       %>% mutate(name = xcao_name)
slinn_ct      <- get_citation_history(slinn_scholar_id)      %>% mutate(name = slinn_name)
rmcmanus_ct   <- get_citation_history(rmcmanus_scholar_id)   %>% mutate(name = rmcmanus_name)
bmukherjee_ct <- get_citation_history(bmukherjee_scholar_id) %>% mutate(name = bmukherjee_name)
dtavana_ct    <- get_citation_history(dtavana_scholar_id)    %>% mutate(name = dtavana_name)
vyadav_ct     <- get_citation_history(vyadav_scholar_id)     %>% mutate(name = vyadav_name)

citation_df <- bind_rows(
  jwright_ct,
  jedgerton_ct,
  bdesmarais_ct,
  cloyle_ct,
  xcao_ct,
  slinn_ct,
  rmcmanus_ct,
  bmukherjee_ct,
  dtavana_ct,
  vyadav_ct
)

# Print the combined citation data
print(head(citation_df, 10))
```

```
##    year cites       name
## 1  2008    26 Joe Wright
## 2  2009    39 Joe Wright
## 3  2010    60 Joe Wright
## 4  2011   115 Joe Wright
```

```
## 5  2012   160 Joe Wright
## 6  2013   274 Joe Wright
## 7  2014   341 Joe Wright
## 8  2015   458 Joe Wright
## 9  2016   494 Joe Wright
## 10 2017   537 Joe Wright
```

```r
# ------------------------------------------------------------------------------
# Step 5: Plot citations over time for each professor
# ------------------------------------------------------------------------------
ggplot(citation_df, aes(x = year, y = cites, color = name)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Google Scholar Citation History (Recent Years)",
    x = "Year",
    y = "Citations",
    color = "Faculty"
  )
```




```r
# ------------------------------------------------------------------------------
# Step 6: Median citations per year for each professor
# ------------------------------------------------------------------------------
median_cites <- citation_df %>%
  group_by(name) %>%
  summarize(median_cites = median(cites, na.rm = TRUE), .groups = "drop")
```

```r
print(median_cites)
```

```
## # A tibble: 10 x 2
##    name              median_cites
##    <chr>                    <dbl>
##  1 Bruce Desmarais          294
##  2 Bumba Mukherjee          216.
##  3 Cyanne Loyle              67
##  4 Daniel Tavana              9
##  5 Jared Edgerton            22.5
##  6 Joe Wright               494
##  7 Roseanne McManus          67
##  8 Susanne Linn             226
##  9 Vineeta Yadav             29
## 10 Xun Cao                  152
```

In this analysis, missing years are omitted rather than coded as zero citations. Treating missing years as zeros would implicitly assume that the scholar was active during those years but received no citations, which is often incorrect, especially for pre-career periods or gaps in observed activity. Omitting missing years avoids mechanically lowering the median citation measure, which is important because even a single zero can substantially affect the median when the number of observed years is limited.

#question 4

```r
# ------------------------------------------------------------------------------
# Part 1: Hard-code three Penn State faculty (social sciences broadly)
# ------------------------------------------------------------------------------
# These are the three faculty members we will use throughout the script.
# (We will repeat the same scraping steps for each person.)

# Joe Wright
jwright_name <- "Joe Wright"
jwright_dept <- "Political Science (College of the Liberal Arts)"
jwright_url  <- "https://polisci.la.psu.edu/people/jgw12/"

# Xun Cao
xcao_name <- "Xun Cao"
xcao_dept <- "Political Science (College of the Liberal Arts)"
xcao_url  <- "https://polisci.la.psu.edu/people/xuc11/"

# Bruce Desmarais
bdesmarais_name <- "Bruce Desmarais"
bdesmarais_dept <- "Political Science (College of the Liberal Arts)"
bdesmarais_url  <- "https://polisci.la.psu.edu/people/bbd5087/"

# Jared Edgerton
jedgerton_name <- "Jared Edgerton"
jedgerton_dept <- "Political Science (College of the Liberal Arts)"
jedgerton_url  <- "https://polisci.la.psu.edu/people/jared-edgerton/"

# Susanne Linn
slinn_name <- "Susanne Linn"
slinn_dept <- "Political Science (College of the Liberal Arts)"
```

```r
slinn_url  <- "https://polisci.la.psu.edu/people/sld8/"

# Cyanne Loyle
cloyle_name <- "Cyanne Loyle"
cloyle_dept <- "Political Science (College of the Liberal Arts)"
cloyle_url  <- "https://polisci.la.psu.edu/people/cel5432/"

# Roseanne McManus
rmcmanus_name <- "Roseanne McManus"
rmcmanus_dept <- "Political Science (College of the Liberal Arts)"
rmcmanus_url  <- "https://polisci.la.psu.edu/people/rum842/"

# Bumba Mukherjee
bmukherjee_name <- "Bumba Mukherjee"
bmukherjee_dept <- "Political Science (College of the Liberal Arts)"
bmukherjee_url  <- "https://polisci.la.psu.edu/people/sxm73/"

# Daniel Tavana
dtavana_name <- "Daniel Tavana"
dtavana_dept <- "Political Science (College of the Liberal Arts)"
dtavana_url  <- "https://polisci.la.psu.edu/people/daniel-tavana/"

# Vineeta Yadav
vyadav_name <- "Vineeta Yadav"
vyadav_dept <- "Political Science (College of the Liberal Arts)"
vyadav_url  <- "https://polisci.la.psu.edu/people/vuy2/"

# ------------------------------------------------------------------------------
# Step 1: Scrape Matt Golder (one complete example, step-by-step)
# ------------------------------------------------------------------------------
# 1) Read the PSU profile page
jwright_page <- read_html(jwright_url)

# (Optional) quick structure check / debugging
jwright_heads <- jwright_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

jwright_text <- jwright_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

# 2) Extract "Areas of Interest" (HTML) - interests only
jwright_areas <- jwright_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

# 3) Combine into one string (semicolon-separated)
jwright_interests <- paste(jwright_areas, collapse = "; ")

# 4) Count items
jwright_n_interest_items <- length(jwright_areas)
```

```r
# 5) Store results (tibble row)
jwright_row <- tibble(
  name = jwright_name,
  department = jwright_dept,
  url = jwright_url,
  scraped_interests = jwright_interests,
  n_interest_items = jwright_n_interest_items
)

xcao_page <- read_html(xcao_url)

xcao_heads <- xcao_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

xcao_text <- xcao_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

xcao_areas <- xcao_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

xcao_interests <- paste(xcao_areas, collapse = "; ")
xcao_n_interest_items <- length(xcao_areas)

xcao_row <- tibble(
  name = xcao_name,
  department = xcao_dept,
  url = xcao_url,
  scraped_interests = xcao_interests,
  n_interest_items = xcao_n_interest_items
)


bdesmarais_page <- read_html(bdesmarais_url)

bdesmarais_heads <- bdesmarais_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

bdesmarais_text <- bdesmarais_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

bdesmarais_areas <- bdesmarais_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

bdesmarais_interests <- paste(bdesmarais_areas, collapse = "; ")
bdesmarais_n_interest_items <- length(bdesmarais_areas)

bdesmarais_row <- tibble(
```

```
  name = bdesmarais_name,
  department = bdesmarais_dept,
  url = bdesmarais_url,
  scraped_interests = bdesmarais_interests,
  n_interest_items = bdesmarais_n_interest_items
)


jedgerton_page <- read_html(jedgerton_url)

jedgerton_heads <- jedgerton_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

jedgerton_text <- jedgerton_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

jedgerton_areas <- jedgerton_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

jedgerton_interests <- paste(jedgerton_areas, collapse = "; ")
jedgerton_n_interest_items <- length(jedgerton_areas)

jedgerton_row <- tibble(
  name = jedgerton_name,
  department = jedgerton_dept,
  url = jedgerton_url,
  scraped_interests = jedgerton_interests,
  n_interest_items = jedgerton_n_interest_items
)


slinn_page <- read_html(slinn_url)

slinn_heads <- slinn_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

slinn_text <- slinn_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

slinn_areas <- slinn_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

slinn_interests <- paste(slinn_areas, collapse = "; ")
slinn_n_interest_items <- length(slinn_areas)

slinn_row <- tibble(
  name = slinn_name,
```

```
  department = slinn_dept,
  url = slinn_url,
  scraped_interests = slinn_interests,
  n_interest_items = slinn_n_interest_items
)

cloyle_page <- read_html(cloyle_url)

cloyle_heads <- cloyle_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

cloyle_text <- cloyle_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

cloyle_areas <- cloyle_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

cloyle_interests <- paste(cloyle_areas, collapse = "; ")
cloyle_n_interest_items <- length(cloyle_areas)

cloyle_row <- tibble(
  name = cloyle_name,
  department = cloyle_dept,
  url = cloyle_url,
  scraped_interests = cloyle_interests,
  n_interest_items = cloyle_n_interest_items
)

rmcmanus_page <- read_html(rmcmanus_url)

rmcmanus_heads <- rmcmanus_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

rmcmanus_text <- rmcmanus_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

rmcmanus_areas <- rmcmanus_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

rmcmanus_interests <- paste(rmcmanus_areas, collapse = "; ")
rmcmanus_n_interest_items <- length(rmcmanus_areas)

rmcmanus_row <- tibble(
  name = rmcmanus_name,
  department = rmcmanus_dept,
  url = rmcmanus_url,
  scraped_interests = rmcmanus_interests,
```

```
  n_interest_items = rmcmanus_n_interest_items
)

bmukherjee_page <- read_html(bmukherjee_url)

bmukherjee_heads <- bmukherjee_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

bmukherjee_text <- bmukherjee_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

bmukherjee_areas <- bmukherjee_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

bmukherjee_interests <- paste(bmukherjee_areas, collapse = "; ")
bmukherjee_n_interest_items <- length(bmukherjee_areas)

bmukherjee_row <- tibble(
  name = bmukherjee_name,
  department = bmukherjee_dept,
  url = bmukherjee_url,
  scraped_interests = bmukherjee_interests,
  n_interest_items = bmukherjee_n_interest_items
)

dtavana_page <- read_html(dtavana_url)

dtavana_heads <- dtavana_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

dtavana_text <- dtavana_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

dtavana_areas <- dtavana_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

dtavana_interests <- paste(dtavana_areas, collapse = "; ")
dtavana_n_interest_items <- length(dtavana_areas)

dtavana_row <- tibble(
  name = dtavana_name,
  department = dtavana_dept,
  url = dtavana_url,
  scraped_interests = dtavana_interests,
  n_interest_items = dtavana_n_interest_items
)
```

```r
vyadav_page <- read_html(vyadav_url)

vyadav_heads <- vyadav_page %>%
  html_elements("h1, h2, h3, h4") %>%
  html_text(trim = TRUE)

vyadav_text <- vyadav_page %>%
  html_element("body") %>%
  html_text(trim = TRUE)

vyadav_areas <- vyadav_page %>%
  html_elements(xpath = "//h2[normalize-space()='Areas of Interest']/following-sibling::ul[1]/li") %>%
  html_text(trim = TRUE)

vyadav_interests <- paste(vyadav_areas, collapse = "; ")
vyadav_n_interest_items <- length(vyadav_areas)

vyadav_row <- tibble(
  name = vyadav_name,
  department = vyadav_dept,
  url = vyadav_url,
  scraped_interests = vyadav_interests,
  n_interest_items = vyadav_n_interest_items
)




# ------------------------------------------------------------------------------
# Step 5: Combine the scraped rows into one data frame and inspect
# ------------------------------------------------------------------------------
faculty_interest_df <- bind_rows(
  jwright_row, xcao_row, bdesmarais_row, jedgerton_row, slinn_row,
  cloyle_row, rmcmanus_row, bmukherjee_row, dtavana_row, vyadav_row
)

print(faculty_interest_df)
```

```
## # A tibble: 10 x 5
##    name             department          url   scraped_interests n_interest_items
##    <chr>            <chr>               <chr> <chr>                        <int>
##  1 Joe Wright       Political Science ~ http~ Comparative Poli~                5
##  2 Xun Cao          Political Science ~ http~ International Re~                2
##  3 Bruce Desmarais  Political Science ~ http~ American Politic~                2
##  4 Jared Edgerton   Political Science ~ http~ American Politic~                7
##  5 Susanne Linn     Political Science ~ http~ American Politic~                2
##  6 Cyanne Loyle     Political Science ~ http~ Comparative Poli~                4
##  7 Roseanne McManus Political Science ~ http~ International Re~                1
##  8 Bumba Mukherjee  Political Science ~ http~ International Re~                6
##  9 Daniel Tavana    Political Science ~ http~ Comparative Poli~                5
## 10 Vineeta Yadav    Political Science ~ http~ Comparative Poli~                1
```

```r
# ------------------------------------------------------------------------------
# Step 6: Quick plot (interest items captured per faculty member)
# ------------------------------------------------------------------------------
library(dplyr)
library(tidyr)
library(stringr)


interest_words <- faculty_interest_df %>%
  filter(!is.na(scraped_interests), scraped_interests != "") %>%
  separate_rows(scraped_interests, sep = ";") %>%
  mutate(
    scraped_interests = str_trim(scraped_interests),
    scraped_interests = str_to_lower(scraped_interests)
  ) %>%
  count(scraped_interests, sort = TRUE)

print(head(interest_words, 10))
```

```
## # A tibble: 10 x 2
##    scraped_interests                n
##    <chr>                        <int>
##  1 comparative politics             5
##  2 international relations          5
##  3 methodology                      4
##  4 american politics                3
##  5 comparative political economy     2
##  6 dictatorships                    2
##  7 government repression            2
##  8 human security                   2
##  9 political violence               2
## 10 civil military relations         1
```

```r
# install.packages("wordclound")
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```r
set.seed(123)

wordcloud(
  words = interest_words$scraped_interests,
  freq  = interest_words$n,
  min.freq = 1,
  max.words = 100,
  random.order = FALSE,
  colors = RColorBrewer::brewer.pal(8, "Dark2")
)
```

```
## Warning in wordcloud(words = interest_words$scraped_interests, freq =
## interest_words$n, : comparative politics could not be fit on page. It will not
## be plotted.
```

```
## Warning in wordcloud(words = interest_words$scraped_interests, freq =
## interest_words$n, : international relations could not be fit on page. It will
## not be plotted.
```

```
## Warning in wordcloud(words = interest_words$scraped_interests, freq =
## interest_words$n, : comparative political economy could not be fit on page. It
## will not be plotted.
```