# **Day 1 Collecting digital trace data – Exercise cheat sheet**

|  |  |  |
| --- | --- | --- |
|  | **API** | **Web scraping (and browser automation)** |
| Task | Use the API provided by <https://api.congress.gov/> to get   1. Names and other information on members of congress from 2020 to 2023. Save the resulting data on your hard drive. 2. Congressional records from 2020 to 2023 (only includes URLs to records). Save the resulting data on your hard drive. 3. Think about, how you might be able to merge the two data sets | 1. Scrape title, author, date and the body of the articles.    1. Test scraping one of these articles.    2. Build a loop to scrape all the URLs. |
| Notes | * You have to register for the API, to get your API key here: <https://api.congress.gov/sign-up/> * The data should stay in list format, don’t bother trying to change the format or clean the data, we will do that tomorrow. * You can append lists using append() * You do not have to read the text of the congressional records * **Don’t forget to include a time delay between API requests!** | * The result should be a dataframe with five columns (title, author, date, body, url). * Tomorrow we will learn how to clean the text data. * **Don’t forget to include a time delay between each request to the server!** |
| R Packages | httr, tidyverse | tidyverse, httr |
| Other | - | SelectorGadget: <https://selectorgadget.com/> |
| Functions | httr:GET()  httr::content()  append()  *for loop* or *lapply* | read\_html() # read the website  html\_node() # target what you need  html\_nodes() # target what you need  html\_text() # extract what you need  html\_table() # extract what you need  html\_attr() # extract what you need |
|  |  |  |

|  |
| --- |
| 1. Install RSelenium using a guide (see below) 2. Use RSelenium to collect the URLs of 200 articles mentioning “Black Lives Matter” on <https://edition.cnn.com/>. |
| * In order to be able to use RSelenium, you can follow this guide: <https://www.youtube.com/watch?v=GnpJujF9dBw>. |
| rvest, rselenium, wdman, netstat, (tidyverse) |
| RSelenium:  selenium()  rsDriver() # create remote server  x$findElement() # target element  x$clickElement() # click that element  x$sendKeysToElement() # type something  x$getElementAttribute() # get smth. from element  x$executeScript() # run a script on the webpage  remote\_driver$server$stop() # stop the server  rvest:  read\_html()  html\_nodes()  html\_text() |
|  |