Social Data Science

Scraping the web II

Dr. Thomas Davidson

Rutgers University

September 29, 2021

Plan

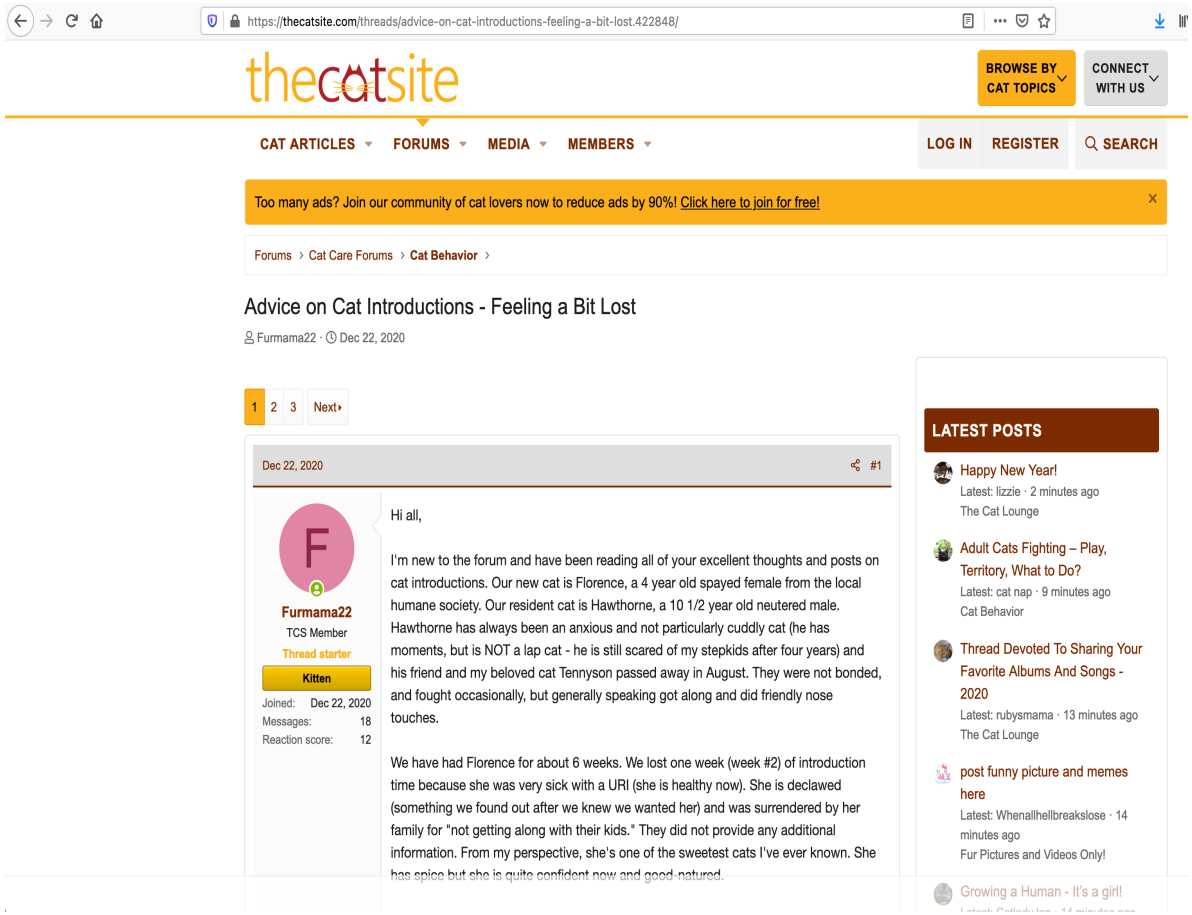
1. How to scrape a website in R, part II
2. Crawling websites using R
3. `selenium` and browser automation
4. Next week

How to scrape a web page

Using `rvest` to scrape HTML

```
library(rvest)
library(tidyverse)
library(stringr)
library(lubridate)
```

How to scrape a web page



How to scrape a web page

Using `rvest` to scrape HTML

We used `rvest` to read in this URL.

```
url <- "https://thecatsite.com/threads/advice-on-cat-introductions-feel
thread <- read_html(url)
```

How to scrape a web page

Creating a function to collect and store data

```
get.posts <- function(thread) {
  messages <- thread %>% html_nodes(".message-body") %>%
    html_text() %>% str_trim()
  users <- thread %>% html_nodes(".message-userDetails") %>%
    html_text() %>% str_trim() %>% str_split('\n') %>% map(1)
  timestamps <- thread %>% html_nodes(".u-concealed .u-dt") %>%
    html_attr("datetime") %>% ymd_hms(tz="EST")
  timestamps <- timestamps[-1] # remove first timestamp
  data <- as_tibble(cbind(messages, unlist(users), timestamps))
  colnames(data) <- c("message","user", "timestamp")
  return(data)
}
```

How to scrape a web page

Using the function

We then used this function to extract information from the forum.

```
results <- get.posts(thread)
tail(results)
```

```
## # A tibble: 6 x 3
##   message                                                        user
##   <chr>                                                          <chr>
## 1 "That's great, thank you \n pearl99\n! Our five minute~ Furmama22
## 2 "Furmama22 said:\n\n\n\nThanks so much for commenting ~ ArtNJ
## 3 "Thank you for your perspective! I'll keep that in min~ Furmama22
## 4 "Furmama22 said:\n\n\n\nThat's great, thank you \n pea~ pearl99
## 5 "Furmama22 said:\n\n\n\nThank you \nC\n calicosrspecia~ calicosrsp
## 6 "Furmama22 said:\n\n\n\nWhen he does need to go in the~ Mamanyt195
```

How to scrape a web page

Pagination

The next step is to figure out how we can navigate the different pages of the thread. Inspection of the HTML shows the CSS element `pageNav-jump` contains the relevant information.

```
thread %>% html_nodes(".pageNav-jump")
```

```
## {xml_nodeset (2)}
## [1] <a href="/threads/advice-on-cat-introductions-feeling-a-bit-lost
## [2] <a href="/threads/advice-on-cat-introductions-feeling-a-bit-lost
```

How to scrape a web page

Pagination

In this case I want both the links *and* the descriptions.

```
links <- thread %>% html_nodes(".pageNav-jump") %>%
  html_attr("href")
desc <- thread %>% html_nodes(".pageNav-jump") %>%
  html_text()
pagination.info <- data.frame(links, desc) %>%
  filter(str_detect(desc, "Next")) %>% distinct()
head(pagination.info)
```

```
##                                                                                                li
## 1 /threads/advice-on-cat-introductions-feeling-a-bit-lost.422848/pag
```

How to scrape a web page

Pagination

We can then use the base URL to get the link to the next page.

```
base <- "https://thecatsite.com"
next.page <- paste(base, pagination.info$links, sep = '')
print(next.page)
```

```
## [1] "https://thecatsite.com/threads/advice-on-cat-introductions-feel
```

How to scrape a web page

Pagination

Let's verify this works by using the `get.posts` function.

```
results <- get.posts(read_html(next.page))
results[1:5,]
```

```
## # A tibble: 5 x 3
##   message                                                        user
##   <chr>                                                          <chr>
## 1 "Thank you all for responding! Merry Christmas to all of ~ Furmama
## 2 "Sounds like a reason to be merry to me!"                    Mamanyt
## 3 "Well I suppose it's always one step forward two steps ba~ Furmama
## 4 "AWWWWWWWWW! She is adorable! And that really wasn't even~ Mamanyt
## 5 "Thank you!"                                                 Furmama
```

How to scrape a web page

Pagination function

```
get.next.page <- function(thread){
  links <- thread %>% html_nodes(".pageNav-jump") %>%
    html_attr("href")
  desc <- thread %>% html_nodes(".pageNav-jump") %>%
    html_text()
  pagination.info <- data.frame(links, desc) %>%
    filter(str_detect(desc, "Next")) %>% distinct()
  base <- "https://thecatsite.com"
  next.page <- paste(base, pagination.info$links, sep = '')
  return(next.page)
}
get.next.page(thread)
```

```
## [1] "https://thecatsite.com/threads/advice-on-cat-introductions-feel
```

How to scrape a web page

Testing the pagination function

We can easily use this function to paginate. In this case we use `get.next.page` to get the link to page 2, read the HTML for page 2, then use `get.next.page` to extract the link to page 3.

```
thread.2 <- read_html(get.next.page(thread))
page.3 <- get.next.page(thread.2)
page.3
```

```
## [1] "https://thecatsite.com/threads/advice-on-cat-introductions-feel
```

How to scrape a web page

Testing the pagination function

What happens when we run out of pages? In this case there is no link to the next page. The `get.next.page` function does not produce an error, but only returns the base URL.

```
get.next.page(read_html("https://thecatsite.com/threads/advice-on-cat-i
```

```
## [1] "https://thecatsite.com/threads/advice-on-cat-introductions-feel
```

How to scrape a web page

Improving the function

```
get.next.page <- function(thread){
  links <- thread %>% html_nodes(".pageNav-jump") %>%
    html_attr("href")
  desc <- thread %>% html_nodes(".pageNav-jump") %>%
    html_text()
  pagination.info <- data.frame(links, desc) %>%
    filter(str_detect(desc, "Next")) %>% distinct()
  if (dim(pagination.info)[1] == 1) {
    base <- "https://thecatsite.com"
    next.page <- paste(base, pagination.info$links, sep = '')
  return(next.page)
    } else {
    return("Final page")
  }
}
```

How to scrape a web page

Testing the pagination function

We now get this message when we try to paginate on the final page.

```
get.next.page(read_html("https://thecatsite.com/threads/advice-on-cat-i
```

```
## [1] "Final page"
```

How to scrape a web page

Paginate and scrape

```
paginate.and.scrape <- function(url){
  thread <- read_html(url)
  posts <- get.posts(thread)
  next.page <- get.next.page(thread)
  while (!str_detect(next.page, "Final page"))
  {
    thread <- read_html(next.page)
    posts <- rbind(posts, get.posts(thread))
    next.page <- get.next.page(thread)
    Sys.sleep(1) # wait 1 second
  }
  return(posts)
}
```

How to scrape a web page

Paginate and scrape

```
full.thread <- paginate.and.scrape(url)
dim(full.thread)
```

```
## [1] 489   3
```

```
print(head(full.thread))
```

```
## # A tibble: 6 x 3
##   message                                                       user
##   <chr>                                                         <chr>
## 1 "Hi all,\nI'm new to the forum and have been reading a~ Furmama22
## 2 "Furmama22 said:\n\n\n\nHi all,\nI'm new to the forum ~ calicosrsp
## 3 "Thank you SO much for taking the time to reply. I rea~ Furmama22
## 4 "I don't think I can add a thing to \nC\n calicosrspec~ Mamanyt195
## 5 "Thanks so much for your thoughts and comments! And th~ Furmama22
## 6 "You certainly came to the right place! And, in my exp~ Mamanyt195
```

How to scrape a web page

Crawling a website

- Now we have a function we can use to paginate and scrape the data from threads on the website
- The next goal is to write a crawler to traverse the website and retrieve information from all of the threads we are interested in.
- Fortunately, these threads are organized in a similar way
  - Each page contains 20 threads and links to the next page

How to scrape a web page

Crawling a website

```
get.threads <- function(url) {
  f <- read_html(url)
  title <- f %>% html_nodes(".structItem-title") %>%
    html_text() %>% str_trim()
  link <- f %>% html_nodes(".structItem-title a") %>%
    html_attr("href")  %>% str_trim()
  link <- data.frame(link)
  link <- link %>% filter(str_detect(link, "/threads/"))
  threads <- data.frame(title, link)
  return(threads)
}
```

How to scrape a web page

Crawling a website

```
forum.url <- "https://thecatsite.com/forums/cat-behavior.5/"

threads <- get.threads(forum.url)
```

How to scrape a web page

Crawling a website

```
print(threads$title)
```

```
##  [1] "help!"
##  [2] "Help! Cat keeps attacking me"
##  [3] "Introducing a new cat."
##  [4] "5 months old cat meowing non-stop"
##  [5] "Help and reassurance with cat introductions"
##  [6] "My cat keeps biting my ankles"
##  [7] "Featured\nEncouraging kitten to enjoy being held, tolerate bat
##  [8] "Help/Advice for introductions"
##  [9] "Looking for Advice on Integrating Nervous Cat into Existing Fo
## [10] "Kitten acts differently when 4 yo comes back from school"
## [11] "New kitten and older cat !"
## [12] "Tail Chewing"
## [13] "Cat intro - Managing aggression"
## [14] "My cat wont stop knocking things over, discipline has become a
## [15] "Kitten weeing on the bed"
## [16] "Separation Anxiety - Help!"
```

How to scrape a web page

Crawling a website

```
print(threads$link)
```

```
##  [1] "/threads/help.435228/"
##  [2] "/threads/help-cat-keeps-attacking-me.435213/"
##  [3] "/threads/introducing-a-new-cat.432924/"
##  [4] "/threads/5-months-old-cat-meowing-non-stop.434922/"
##  [5] "/threads/help-and-reassurance-with-cat-introductions.435222/"
##  [6] "/threads/my-cat-keeps-biting-my-ankles.434978/"
##  [7] "/threads/encouraging-kitten-to-enjoy-being-held-tolerate-baths
##  [8] "/threads/help-advice-for-introductions.428629/"
##  [9] "/threads/looking-for-advice-on-integrating-nervous-cat-into-ex
## [10] "/threads/kitten-acts-differently-when-4-yo-comes-back-from-sch
## [11] "/threads/new-kitten-and-older-cat.435194/"
## [12] "/threads/tail-chewing.435189/"
## [13] "/threads/cat-intro-managing-aggression.435169/"
## [14] "/threads/my-cat-wont-stop-knocking-things-over-discipline-has-
## [15] "/threads/kitten-weeing-on-the-bed.435133/"
## [16] "/threads/separation-anxiety-help.435146/"
```

How to scrape a web page

Crawling a website

Exercise: Write code to iterate over the first 5 pages of threads. You will need to use `get.threads`, `paginate.and.scrape`, and `get.next.page`. Store the results as a tibble in an object called `results`. Make sure to also retain the name of each thread. *Note that this may take a while to run. You should test it on a small subset to verify it works.*

```
# Complete code here
P <- 2
url <- forum.url
results <- tibble()

for (p in 1:P) {
  threads <- get.threads(url)
  for (t in 1:2) {
    page.url <- paste(base, threads$link[t], sep = '')
    new.results <- paginate.and.scrape(page.url)
    new.results$threads <- threads$title[t]
    results <- bind_rows(results, new.results)
```

How to scrape a web page

Storing the results

The results should consist of a few thousand messages and associated metadata. Save the results of this crawl to as a CSV file.

```
library(readr)
write_csv(results, "cat_crawl.csv")
```

How to scrape a web page

Data storage

- If you try to collect all the data you need before saving it, you run the risk of data loss if you script crashes
  - This risk increases as you collect more data
    - More memory on your computer is being used
    - Increased likelihood of encountering anomalies that cause errors
- Reasonable solutions
  - Continuously save results to disk (e.g. concatenate each thread to a CSV)
  - Store results in chunks (e.g. each thread in a new CSV)

How to scrape a web page

Data storage

- A more robust solution
  - Write output to a relational database
    - This helps to organize the data and makes it easier to query and manage, particularly with large datasets
    - I recommend PostgreSQL, a free open-source, SQL-compatible relational database

How to scrape a web page

Data storage

- If collecting a lot of data, I recommend use a server to run the code and to store scraped data
- Requirements
  - Access to a server ($)
    - But most universities have free computing clusters
  - Command line knowledge
  - Database knowledge
- It is beyond the scope of this class to cover this material, but I highly recommend you develop this infrastructure if you continue to work in this area

How to scrape a web page

Logging

- Log the progress of your webscraper
  - Simple option:
    - Print statements in code
  - Better option:
    - Use a log file
  - To keep yourself updated:
    - Use a Slack App to send yourself messages

How to scrape a web page

Javascript and browser automation

- Many websites use Javascript, which cause problems for web-scrapers as it cannot directly be parsed to HTML
- We can get around this by doing the following
  - Automatically to open a browser (e.g. Google Chrome)
  - Load a website in the browser
  - Read the HTML from the browser into R
- We can also use browser automation to click buttons, fill in forms, or enter login info

How to scrape a web page

Selenium

- Selenium WebDriver and the package `RSelenium` (https://github.com/ropensci/RSelenium) is the most popular approach
- **However**, `RSelenium` requires a complicated set up using a Docker container
  - This is a little technical and I've had trouble getting it to work
  - It may be easier to use `selenium` in Python then read the data into R
    - https://python-bloggers.com/2020/07/rvspython-3-setting-up-selenium-limitations-with-the-rselenium-package-getting-past-them/

How to scrape a web page

Using reticulate to run selenium in Python

This Python code uses selenium to open up a Chrome browser, visit a website, and collect the HTML. It then closes the browser.

```
from selenium import webdriver
driver = webdriver.Chrome()
driver.get('https://www.sociology.rutgers.edu')
html = driver.page_source
driver.close()
```

This will only work if the Chrome driver has been downloaded and is in your PATH. See https://chromedriver.chromium.org/getting-started

How to scrape a web page

Using reticulate to run selenium in Python

I saved the code in the previous chunk as a file called `get_html.py`. We can use `reticulate` to run the Python code then pass objects from Python to R. In this case we use Python to run selenium and get the HTML, then read it into R using `rvest`.

```
library(reticulate)

#py_install("selenium") # uncomment to install selenium.py
source_python('../code/get_html.py') # run python script

html.text <- read_html(py$html) %>% html_text()
```

How to scrape a web page

Inspecting the results in R

Reticulate allowed us to run a Python script then pass the results to R. We can then use the same commands as above to process it.

```
print(substr(html.text, 1, 35))
```

Questions