

Computational Social Science

Scraping the web II

Dr. Thomas Davidson

Rutgers University

October 2, 2025

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

The screenshot shows a web browser displaying a forum page on thecatsite.com. The URL in the address bar is https://thecatsite.com/threads/advice-on-cat-introductions-feeling-a-bit-lost.422344/. The page features a navigation bar with links for CAT ARTICLES, FORUMS, MEDIA, MEMBERS, LOG IN, REGISTER, and SEARCH. A yellow banner at the top of the forum section reads: "Too many ads? Join our community of cat lovers now to reduce ads by 90%! [Click here to join for free!](#)". The forum breadcrumb trail is: Forums > Cat Care Forums > Cat Behavior > Advice on Cat Introductions - Feeling a Bit Lost. The thread is by Fumama22, posted on Dec 22, 2020. The thread starter's profile shows a pink circle with the letter 'F', TCS Member status, and a "Kitten" tag. The thread content includes a greeting "Hi all," followed by a detailed introduction of a new cat named Florence, a 4-year-old spayed female from the local humane society. The post describes Florence's background, her relationship with a resident cat named Hawthorne, and the loss of a previous cat named Tennyson. The post concludes with information about Florence's introduction time and her current behavior. On the right side of the forum, there is a "LATEST POSTS" section listing several recent threads with their titles, latest posts, and timestamps.

the catsite

BROWSE BY CAT TOPICS CONNECT WITH US

CAT ARTICLES FORUMS MEDIA MEMBERS LOG IN REGISTER SEARCH

Too many ads? Join our community of cat lovers now to reduce ads by 90%! [Click here to join for free!](#)

Forums > Cat Care Forums > Cat Behavior >

Advice on Cat Introductions - Feeling a Bit Lost

Fumama22 · Dec 22, 2020

1 2 3 Next

Dec 22, 2020 #1

Hi all,

F
Fumama22
TCS Member
Thread starter
Kitten

Joined: Dec 22, 2020
Messages: 18
Reaction score: 12

I'm new to the forum and have been reading all of your excellent thoughts and posts on cat introductions. Our new cat is Florence, a 4 year old spayed female from the local humane society. Our resident cat is Hawthorne, a 10 1/2 year old neutered male. Hawthorne has always been an anxious and not particularly cuddly cat (he has moments, but is NOT a lap cat - he is still scared of my stopkicks after four years) and his friend and my beloved cat Tennyson passed away in August. They were not bonded, and fought occasionally, but generally speaking got along and did friendly nose touches.

We have had Florence for about 6 weeks. We lost one week (week #2) of introduction time because she was very sick with a URI (she is healthy now). She is declawed (something we found out after we knew we wanted her) and was surrendered by her family for "not getting along with their kids." They did not provide any additional information. From my perspective, she's one of the sweetest cats I've ever known. She has spots but she is quite confident now and good-natured.

LATEST POSTS

- Happy New Year!**
Latest: kate - 2 minutes ago
The Cat Lounge
- Adult Cats Fighting - Play, Territory, What to Do?**
Latest: cat nap - 9 minutes ago
Cat Behavior
- Thread Devoted To Sharing Your Favorite Albums And Songs - 2020**
Latest: rubymama - 13 minutes ago
The Cat Lounge
- post funny picture and memes here**
Latest: Wherewithalbreakalace - 14 minutes ago
Fur Pictures and Videos Only!
- Growing a Human - It's a girl!**
Latest: PoshPaws - 17 minutes ago

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_squish()  
  users <- thread %>% html_nodes(".message-name") %>%  
    html_text() %>% str_squish()  
  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
```

```
##   <chr>
```

```
## 1 "That's great, thank you pearl99 ! Our five minute visual acc~ Fur
```

```
## 2 "Furmama22 said: Thanks so much for commenting A ArtNJ . I th~ Art
```

```
## 3 "Thank you for your perspective! I'll keep that in mind."      Fur
```

```
## 4 "Furmama22 said: That's great, thank you pearl99 ! Our five m~ pea
```

```
## 5 "Furmama22 said: Thank you C calicosrspecial and pearl99 !! I~ cal
```

```
## 6 "Furmama22 said: When he does need to go in the room, I'll go~ Mam
```

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)
```

```
##
```

```
## 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
```

```
##   <chr>
```

```
## 1 "Thank you all for responding! Merry Christmas to all of you!~ Fur
```

```
## 2 "Sounds like a reason to be merry to me!" Mam
```

```
## 3 "Well I suppose it's always one step forward two steps back. ~ Fur
```

```
## 4 "AWWWWWWWW! She is adorable! And that really wasn't even a w~ Mam
```

```
## 5 "Thank you!" Fur
```

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"
```

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"))  
  {  
    print(next.page)  
    thread <- read_html(next.page)  
    posts <- rbind(posts, get.posts(thread))  
    next.page <- get.next.page(thread)  
    Sys.sleep(0.5) # wait 0.5 seconds  
  }  
  return(posts)  
}
```

How to scrape a web page

Paginate and scrape

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

[illegible]

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!!!!!! Parents are threatening to give away my cat because  
## [2] "Elder cat won't pee in litter box"  
## [3] "Thinking Of Taking In Sylvester's Sister, We Must Be Crazy?!"  
## [4] "Australian Bush Flowers: experiences?"  
## [5] "Hangin' In The Living Room... and Nowhere Else. Meds, Stress,  
## [6] "Extremely anxious about new cats and family dynamics"  
## [7] "Introducing 10 week kitten to 7 year old cat - doable?"  
## [8] "STAYING HOME"  
## [9] "Sort out Spraying facts & myths"  
## [10] "Introducing a new cat to our household ."  
## [11] "Resident cat aggressive towards new kitten despite attempts to  
## [12] "Pippin's curious and learned behaviors"  
## [13] "Why does kitten respect male but not female?"  
## [14] "Cat attacks. At a loss on what to do."  
## [15] "Commands to a cat"  
## [16] "Advice for getting a fearful, aggressive cat into a carrier?"
```

How to scrape a web page

Crawling a website

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

```
## [1] "/threads/help-parents-are-threatening-to-give-away-my-cat-beca
## [2] "/threads/elder-cat-wont-pee-in-litter-box.473076/"
## [3] "/threads/thinking-of-taking-in-sylvesters-sister-we-must-be-cr
## [4] "/threads/australian-bush-flowers-experiences.473153/"
## [5] "/threads/hangin-in-the-living-room-and-nowhere-else-meds-stres
## [6] "/threads/extremely-anxious-about-new-cats-and-family-dynamics.
## [7] "/threads/introducing-10-week-kitten-to-7-year-old-cat-doable.4
## [8] "/threads/staying-home.472955/"
## [9] "/threads/sort-out-spraying-facts-myths.473099/"
## [10] "/threads/introducing-a-new-cat-to-our-household.473096/"
## [11] "/threads/resident-cat-aggressive-towards-new-kitten-despite-at
## [12] "/threads/pippins-curious-and-learned-behaviors.469040/"
## [13] "/threads/why-does-kitten-respect-male-but-not-female.473078/"
## [14] "/threads/cat-attacks-at-a-loss-on-what-to-do.473040/"
## [15] "/threads/commands-to-a-cat.472994/"
## [16] "/threads/advice-for-getting-a-fearful-aggressive-cat-into-a-ca
```

How to scrape a web page

Crawling a website

Iterating over the first 5 pages of threads and collecting all threads in each set.

```
pagination.max <- 5
url <- forum.url
results <- tibble()

for (p in 1:pagination.max) {
  threads <- get.threads(url)
  for (t in 1:dim(threads)[1]) {
    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)
  }
  url <- get.next.page(read_html(url))
}
```


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.

```
write_csv(results, "../data/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 your 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

Advanced web scraping

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

Advanced webscraping

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
- ▶ It will 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/>

Advanced web scraping

Running Selenium using 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
from time import sleep
driver = webdriver.Chrome()
driver.get('https://www.sociology.rutgers.edu')
html = driver.page_source
sleep(15)
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>

Advanced webscraping

Using reticulate to obtain results using R

We can use reticulate to pass objects from Python to R. Note the Python objects are shown in the Environment tab.

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

Advanced web scraping

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))
```

Advanced web scraping

Browser automation

- ▶ Using browser automation you can perform other tasks that simulate a human user
 - ▶ Clicking buttons
 - ▶ Filing forms
 - ▶ Logging in
 - ▶ Saving images

Next week

- ▶ Online surveys and experiments
- ▶ Intro to RShiny
- ▶ Homework 2 released