Web scraping

WICSS-Tucson

Thomas Davidson

Rutgers University

January 6, 2020

Plan

1. What is web-scraping?
2. When should I use it?
3. How to scrape a web page
4. Crawling websites
5. `selenium` and browser automation

What is web-scraping?

Terminology

- Web scraping is a method to collect and store data from websites
  - We use the code underlying a webpage to collect data (**scraping**)
  - The process is then repeated for other pages on the same website in an automated fashion (**crawling**)

What is web-scraping?

Challenges

- Different websites have different structures
- Websites can be internally inconsistent
- Some content is harder to collect (e.g. Javascript)
- Some websites limit or prohibit scraping

What is web-scraping?

Examples: Commercial use cases

- Search engines
  - Google scrapes websites to create a searchable index of the internet
- Price comparison
  - Kayak scrape airlines to compare flight prices, other websites do the same for hotels and rental cars
- Recruitment
  - Recruitment companies scrape LinkedIn to get data on workers

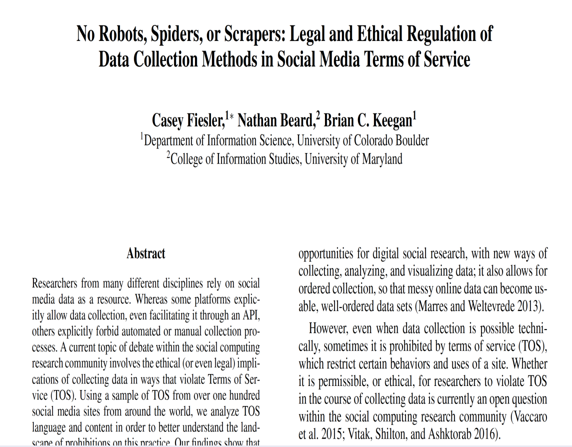
When should I use it?

Social scientific use cases

- Web-scraping is a useful tool to collect data from websites without APIs
  - Large social media platforms and other sites have APIs but smaller websites do not
    - Local newspapers, forums, small businesses, educational institutions, etc.
- Often we want to collect data from a single website
  - e.g. All posts written on a forum
- Sometimes we might want to collect data from many websites
  - e.g. All schools in a school district

When should I use it?

Ethical and legal considerations



No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service

Casey Fiesler,[1]* Nathan Beard,[2] Brian C. Keegan[1]

[1]Department of Information Science, University of Colorado Boulder

[2]College of Information Studies, University of Maryland

Abstract

Researchers from many different disciplines rely on social media data as a resource. Whereas some platforms explicitly allow data collection, even facilitating it through an API, others explicitly forbid automated or manual collection processes. A current topic of debate within the social computing research community involves the ethical (or even legal) implications of collecting data in ways that violate Terms of Service (TOS). Using a sample of TOS from over one hundred social media sites from around the world, we analyze TOS language and content in order to better understand the landscape of prohibitions on this practice. Our findings show that

opportunities for digital social research, with new ways of collecting, analyzing, and visualizing data; it also allows for ordered collection, so that messy online data can become usable, well-ordered data sets (Marres and Weltevrede 2013).

However, even when data collection is possible technically, sometimes it is prohibited by terms of service (TOS), which restrict certain behaviors and uses of a site. Whether it is permissible, or ethical, for researchers to violate TOS in the course of collecting data is currently an open question within the social computing research community (Vaccaro et al. 2015; Vitak, Shilton, and Ashktorab 2016).

When should I use it?

Ethical and legal considerations

- Fiesler, Beard, and Keegan (2020) review the legal cases related to web-scraping and analyze website terms of service
  - “In short, it is an unsettled question as to whether it is explicitly illegal (or even a criminal act) to violate TOS.”
  - No academic or journalist has ever been prosecuted for violating a website terms of service to collect data for research
- They analyze terms of service of over 100 social media websites
  - Terms of service are ambiguous, inconsistent, and lack context

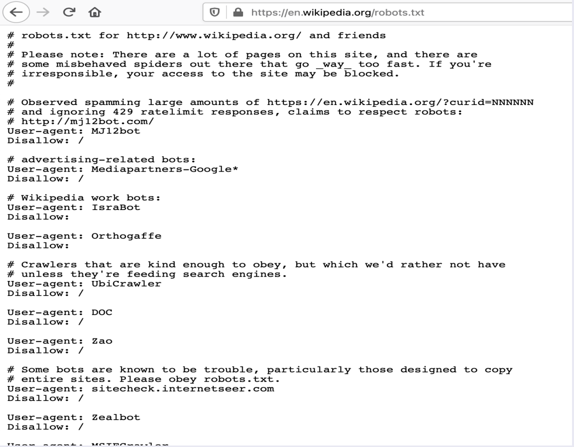
When should I use it?

Best-practices

- Only scrape publicly available data
  - i.e. You can access the page on the web without logging in
- Do not scrape copyright protected data
- Try not to violate website terms of service
- Do not burden the website
  - Limit the number of calls you make (similar to rate-limiting in APIs)
- Avoid using the data in a way that may interfere with business
  - g.g. Don't copy valuable data from a small business and share it on Github

How to scrape a web page

Start by looking up `robots.txt`



https://en.wikipedia.org/robots.txt

```
# robots.txt for http://www.wikipedia.org/ and friends
#
# Please note: There are a lot of pages on this site, and there are
# some misbehaved spiders out there that go _way_ too fast. If you're
# irresponsible, your access to the site may be blocked.
#

# Observed spamming large amounts of https://en.wikipedia.org/?curid=NNNNNN
# and ignoring 429 ratelimit responses, claims to respect robots:
# http://mj12bot.com/
User-agent: MJ12bot
Disallow: /

# advertising-related bots:
User-agent: Mediapartners-Google*
Disallow: /

# Wikipedia work bots:
User-agent: IsraBot
Disallow:

User-agent: Orthogaffe
Disallow:

# Crawlers that are kind enough to obey, but which we'd rather not have
# unless they're feeding search engines.
User-agent: UbiCrawler
Disallow: /

User-agent: DOC
Disallow: /

User-agent: Zao
Disallow: /

# Some bots are known to be trouble, particularly those designed to copy
# entire sites. Please obey robots.txt.
User-agent: sitecheck.internetseer.com
Disallow: /

User-agent: Zealbot
Disallow: /

User-agent: MSIECrawler
```

How to scrape a web page

Decoding `robots.txt`

- **`User-agent`** = the name of the scraper
  - **`*`** = All scrapers
- **`Allow: /path/`** = OK to scrape
- **`Disallow: /path/`** = Not OK to scrape
  - **`Disallow: /`** = Not OK to scrape any pages
- **`Crawl-Delay: N`** = Wait `N` miliseconds between each call to the website

How to scrape a web page

Terminology

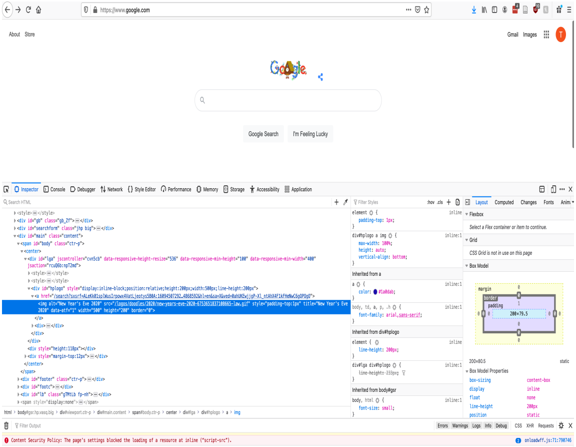
- A web-page is loaded using a **URL** (Uniform Resource Locator)
- The underlying code we are interested in is usually **HTML** (Hypertext Markup Language)
- Many websites use **CSS** (Cascading Style Sheets) to structure HTML
    - This will help us to find what we are interested in
        - See https://flukeout.github.io/ for an interactive tutorial on using CSS selectors
        - Chrome Plugin to help find CSS elements: https://selectorgadget.com/

How to scrape a web page

Inspecting HTML

- Open up a website and right click on any part of the screen
  - You should see an option titled `Inspect Element` or similar
  - This will allow you to view the code used to generate the page
    - You will see code related to the particular element you were hovering over when your right clicked

How to scrape a web page



```
<style>⋯</style>
<div id="gb" class="gb_Zf">⋯</div>
<div id="searchform" class="jhp big">⋯</div>
<div id="main" class="content">
  <span id="body" class="ctr-p">
    <center>
      <div id="lga" jscontroller="cvn5cb" data-responsive-height-resize="536" data-responsive-min-height="100" data-responsive-min-width="400"
      jsaction="rcuQ6b:npT2md">
        <style>⋯</style>
        <style>⋯</style>
        <div id="hplogo" style="display:inline-block;position:relative;height:200px;width:500px;line-height:200px">
          <a href="/search?sxsrf=ALeKk01splWusIrpowx4VatLjeotys5B0A:16094507292…4868592&hl=en&sa=X&ved=0ahUKEwjjgP-Xl_ntAhX4F1kFHeNwC6gQPQgO">
            <img alt="New Year's Eve 2020" src="/logos/doodles/2020/new-years-eve-2020-6753651837108665-law.gif" style="padding-top:1px" title="New Year's Eve
            2020" data-atf="1" width="500" height="200" border="0">
          </a>
          <div>⋯</div>
        </div>
      </div>
      <div style="height:118px"></div>
      <div style="margin-top:12px">⋯</div>
    </center>
  </span>
<div id="footer" class="ctr-p">⋯</div>
<div id="footc">⋯</div>
<div id="lb" class="gTMtLb fp-nh">⋯</div>
<span style="display:none">⋯</span>
```

html › body#gsr.hp.vasq.big › div#viewport.ctr-p › div#main.content › span#body.ctr-p › center › div#lga › div#hplogo › a › img

```
element {                                inline
  padding-top: 1px;
}
div#hplogo a img {                       inline:1
  max-width: 100%;
  height: auto;
  vertical-align: bottom;
}
Inherited from a
a {                                      inline:1
  color: ● #1a0dab;
}
body, td, a, p, .h {                     inline:1
  font-family: arial,sans-serif;
}
Inherited from div#hplogo
element {                                inline
  line-height: 200px;
}
div#lga div#hplogo {                     inline:1
  ~~line-height: 233px;~~
}
Inherited from body#gsr
body, html {                             inline:1
  font-size: small;
}
```

Flexbox

Select a Flex container or item to continue.

Grid

CSS Grid is not in use on this page

Box Model

margin 0 0 0 0; border 0 0 0 0; padding 1 0 0 0; 200×79.5

200×80.5 static

Box Model Properties

box-sizing	content-box
display	inline
float	none
line-height	200px
position	static

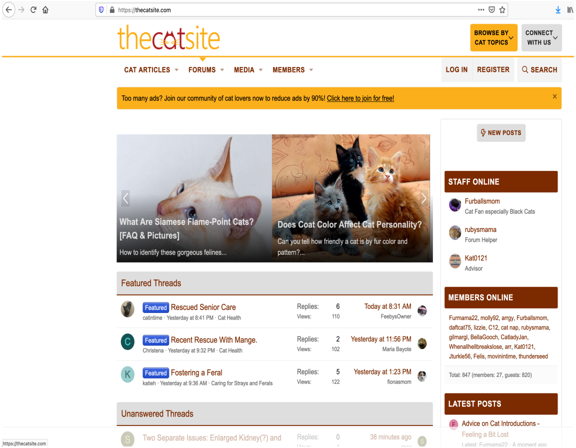
```
Content Security Policy: The page's settings blocked the loading of a resource at inline ("script-src").    2 onloadwff.js:71:790746
```

How to scrape a web page

Using rvest to scrape HTML

```
library(rvest)
library(dplyr)
library(stringr)
```

How to scrape a web page



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

What Are Siamese Flame-Point Cats? [FAQ & Pictures]

How to identify these gorgeous felines...

Does Coat Color Affect Cat Personality?

Can you tell how friendly a cat is by fur color and pattern?...

Featured Threads

Featured Rescued Senior Care catintime · Yesterday at 8:41 PM · Cat Health	Replies: Views:	6 110	Today at 8:31 AM FeebysOwner
Featured Recent Rescue With Mange. Christena · Yesterday at 9:32 PM · Cat Health	Replies: Views:	2 102	Yesterday at 11:56 PM Maria Bayote
Featured Fostering a Feral katieh · Yesterday at 9:36 AM · Caring for Strays and Ferals	Replies: Views:	5 122	Yesterday at 1:23 PM fionasmom

Unanswered Threads

Two Separate Issues: Enlarged Kidney(?) and	Replies:	0	38 minutes ago

STAFF ONLINE

Furballsmom
Cat Fan especially Black Cats

rubysmama
Forum Helper

Kat0121
Advisor

MEMBERS ONLINE

Furmama22, molly92, arrgy, Furballsmom, daftcat75, lizzie, C12, cat nap, rubysmama, gilmargl, BellaGooch, CatladyJan, Whenallhellbreakslose, arr, Kat0121, Jturkle56, Felis, movinintime, thunderseed

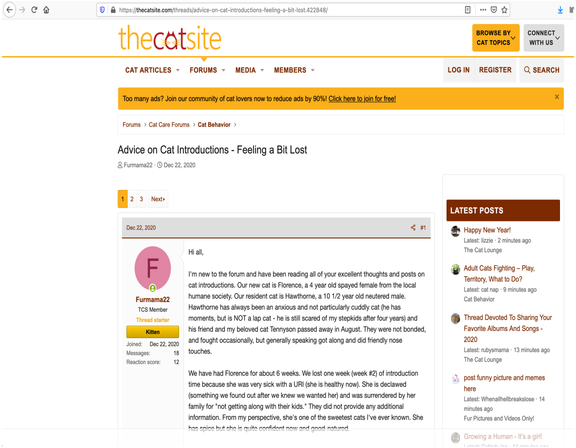
Total: 847 (members: 27, guests: 820)

LATEST POSTS

Advice on Cat Introductions - Feeling a Bit Lost

Latest: Furmama22 · A moment ago

How to scrape a web page



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

Furmama22 · Dec 22, 2020

Dec 22, 2020 #1

F
Furmama22
TCS Member
Thread starter
Kitten
Joined: Dec 22, 2020
Messages: 18
Reaction score: 12

Hi all,

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 stepkids 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 spice but she is quite confident now and good natured.

Happy New Year!
Latest: lizzie · 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: rubysmama · 13 minutes ago
The Cat Lounge

post funny picture and memes here
Latest: Whenallhellbreakslose · 14 minutes ago
Fur Pictures and Videos Only!

Growing a Human - It's a girl!

How to scrape a web page

Using rvest to scrape HTML

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

How to scrape a web page

Collecting messages

First, we parse the HTML to obtain the text of each message on the page. Here we use the CSS selector `.message-body`, which selects all elements with class `message-body`.

```
messages <- thread %>% html_nodes(".message-body") %>%
  html_text() %>% str_trim()
print(length(messages))
```

```
## [1] 20
```

```
print(substr(messages[[1]], 1, 50))
```

```
## [1] "Hi all,\nI'm new to the forum and have been reading"
```

How to scrape a web page

Getting user names

Next we collect the name of each user. User information is found by parsing the `.message-userDetails` node. This is followed by some string manipulation to extract the name.

```r
users <- thread %>% html_nodes(".message-userDetails") %>%
  html_text() %>% str_trim()
print(length(users))
```

```
## [1] 20
```

```r
users[[1]]
```

```
## [1] "Furmama22\nTCS Member\nThread starter\nYoung Cat"
```

```r
users <- thread %>% html_nodes(".message-userDetails") %>%
  html_text() %>% str_trim() %>% str_split('\n') %>% pluck(1)
users[[1]]
```

```
## [1] "Furmama22"
```

How to scrape a web page

Collecting timestamps

Finally, we also want to get the time-stamp of each message. While the forum only displays dates, we can actually get the full timestamp. Note how `time.u-dt` returns too much information, so `.u-concealed .u-dt` is selected instead.

```r
dates <- thread %>% html_nodes("time.u-dt")
length(dates)
```

```
## [1] 27
```

```r
dates <- thread %>% html_nodes(".u-concealed .u-dt")
length(dates)
```

```
## [1] 21
```

```r
dates <- dates %>% html_attr("datetime")
dates[[1]]
```

```
## [1] "2020-12-22T11:26:12-0800"
```

How to scrape a web page

Putting it all together

```r
get.posts <- function(thread) {
  messages <- thread %>% html_nodes(".message-body") %>%
    html_text() %>% str_trim()  %>%
    str_trunc(15, "right") # only get 1st 15 chars
  users <- thread %>% html_nodes(".message-userDetails") %>%
    html_text() %>% str_trim() %>%
    str_split('\n') %>% pluck(1)
  timestamps <- thread %>% html_nodes(".u-concealed .u-dt") %>%
    html_attr("datetime")
  timestamps <- timestamps[-1]
  df <- data.frame(messages, unlist(users), timestamps)
  colnames(df) <- c("message","user", "timestamp")
  return(df)
}
```

How to scrape a web page

Testing

We can now test the function to confirm it returns the information we are expecting. In this case, we want to see the first five posts.

```
results <- get.posts(thread)
results[1:5,]
```

```
##             message            user                timestamp
## 1 Hi all,\nI'm ...       Furmama22 2020-12-22T11:26:12-0800
## 2  Furmama22 sa... calicosrspecial 2020-12-22T13:13:04-0800
## 3  Thank you SO...       Furmama22 2020-12-22T14:01:53-0800
## 4  I don't thin...     Mamanyt1953 2020-12-22T18:21:00-0800
## 5  Thanks so mu...       Furmama22 2020-12-22T18:52:26-0800
```

How to scrape a web page

Pagination

Each thread is split into pages, each containing 20 messages. We want to be able to navigate through these pages.

```
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 = '')
```

How to scrape a web page

Pagination function

Let's create a function so we can easily repeat the process.

```
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

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

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

How to scrape a web page

Testing the pagination function

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

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

```
thread.4 <- read_html(get.next.page(thread.3))
pagination.4 <- get.next.page(thread.4)
print(pagination.4)
```

```
## [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

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

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

```
thread.4 <- read_html(get.next.page(thread.3))
pagination.4 <- get.next.page(thread.4)
print(pagination.4)
```

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

How to scrape a web page

Paginate and scrape

Now we can put these functions together to scrape the entire thread.

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

```
##                 message            user                timestamp
## 1   Hi all,\nI'm ...       Furmama22 2020-12-22T11:26:12-0800
## 2    Furmama22 sa... calicosrspecial 2020-12-22T13:13:04-0800
## 3    Thank you SO...       Furmama22 2020-12-22T14:01:53-0800
## 4    I don't thin...     Mamanyt1953 2020-12-22T18:21:00-0800
## 5    Thanks so mu...       Furmama22 2020-12-22T18:52:26-0800
## 6    You certainl...     Mamanyt1953 2020-12-22T19:06:23-0800
## 7    Furmama22 sa... calicosrspecial 2020-12-23T08:46:09-0800
## 8    This is grea...       Furmama22 2020-12-23T08:57:29-0800
## 9    Furmama22 sa...         pearl99 2020-12-23T09:17:28-0800
## 10   Furmama22 sa... calicosrspecial 2020-12-23T09:19:15-0800
## 11  Thank you \nC...       Furmama22 2020-12-23T14:23:36-0800
## 12   The thing th...           ArtNJ 2020-12-23T16:09:54-0800
## 13   Furmama22 sa...         pearl99 2020-12-23T16:31:12-0800
## 14   Thanks so mu...       Furmama22 2020-12-24T05:34:31-0800
```

How to scrape a web page

Crawling a website

We don't just want to collect a single conversation thread. We want to find all relevant threads and then apply the previous function. The function below allows us to collect links and titles for each thread within the sub-forum.

```r
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/" # This is
threads <- get.threads(forum.url)
```

How to scrape a web page

Crawling a website

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

```
##  [1] "Taking cat to the vet"
##  [2] "Need help introducing two cats"
##  [3] "Should We Get a Second Cat?"
##  [4] "?How do I help my cat feel more secure outside with the repair
##  [5] "Featured\nAdvice on Next Steps with Timid Cat"
##  [6] "Resident cat refuses to eat after hearing the new cat"
##  [7] "How high do enclosure walls need to be for 6 - 10 week old fos
##  [8] "3 Week business trip and I'm super stressed"
##  [9] "Advice on Cat Introductions - Feeling a Bit Lost"
## [10] "He is peeing everywhere :("
## [11] "Reintroducing lost kitty"
## [12] "I need help with a kitten I'm cat sitting for two weeks"
## [13] "One cat won't share a litter box, the other uses all of them h
## [14] "My cat is very vocal. I cannot figure him out..."
## [15] "New Cat - Very Timid"
## [16] "Siamese 5 year old male changed sleeping habits?"
```

How to scrape a web page

Crawling a website

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

```
##  [1] "/threads/taking-cat-to-the-vet.423384/"
##  [2] "/threads/need-help-introducing-two-cats.418696/"
##  [3] "/threads/should-we-get-a-second-cat.423393/"
##  [4] "/threads/how-do-i-help-my-cat-feel-more-secure-outside-with-th
##  [5] "/threads/advice-on-next-steps-with-timid-cat.423315/"
##  [6] "/threads/resident-cat-refuses-to-eat-after-hearing-the-new-cat
##  [7] "/threads/how-high-do-enclosure-walls-need-to-be-for-6-10-week-
##  [8] "/threads/3-week-business-trip-and-im-super-stressed.423394/"
##  [9] "/threads/advice-on-cat-introductions-feeling-a-bit-lost.422848
## [10] "/threads/he-is-peeing-everywhere.423300/"
## [11] "/threads/reintroducing-lost-kitty.423373/"
## [12] "/threads/i-need-help-with-a-kitten-im-cat-sitting-for-two-week
## [13] "/threads/one-cat-wont-share-a-litter-box-the-other-uses-all-of
## [14] "/threads/my-cat-is-very-vocal-i-cannot-figure-him-out.422683/"
## [15] "/threads/new-cat-very-timid.422782/"
## [16] "/threads/siamese-5-year-old-male-changed-sleeping-habits.42320
```

How to scrape a web page

Crawling a website

Exercise: Write a function to iterate through the 10 pages of threads, each time calling `get.threads` to collect all threads. Next, use `paginate.and.scrape` for each thread. Store the results in a single data frame.

```
# Complete function here
```

How to scrape a web page

Javascript and browser automation

- Many websites use Javascript
  - This can cause problems for web-scrapers as it cannot directly be parsed to HTML
- Rather than loading HTML directly into R, we can use R to automate a browser
  - Selenium WebDriver and the package `RSelenium` (https://github.com/ropensci/RSelenium) is the most popular approach

How to scrape a web page

What can RSelenium do?

- Extract HTML from Javascript-based websites
- Interact with web-based content
  - e.g., Click "OK" to a warning, complete a search box
- **However**, `RSelenium` currently requires a more complicated set up using a *Docker container* to work in R so I will not demo it today

How to scrape a web page

Data storage and logging

- If collecting a lot of data, use a server to run your code
- Store output in a database
  - This helps to organize the data and makes it easier to query and manage
- Keep a log file with a record of which pages you have scraped
  - You could use Slack to send progress updates

References

- Fiesler, Casey, Nate Beard, and Brian C Keegan. 2020. “No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service.” In Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, 187–96. AAAI.

Questions