

Web Scraping

Ben Williams

October 9th 2020

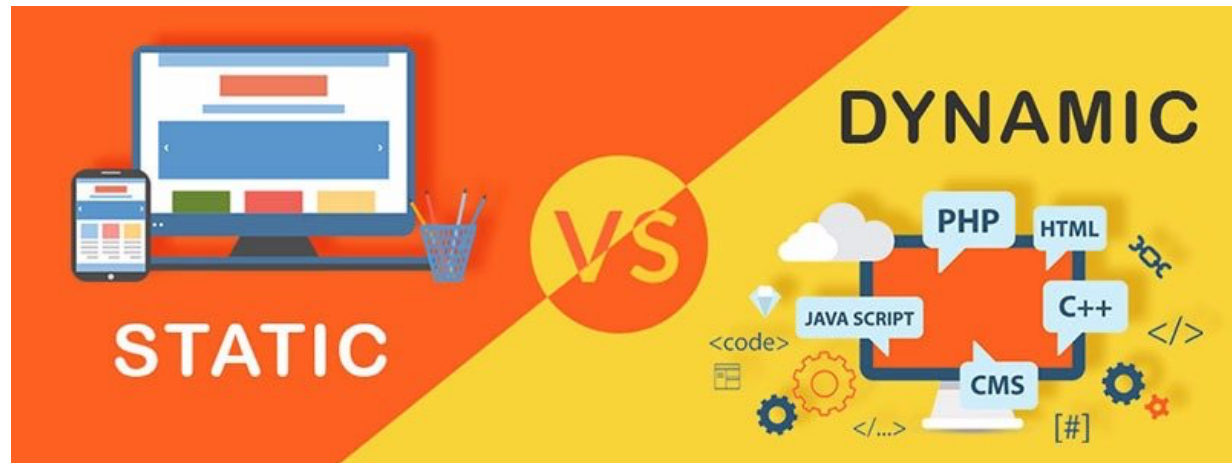


UNIVERSITY of
DENVER

DANIELS COLLEGE OF BUSINESS

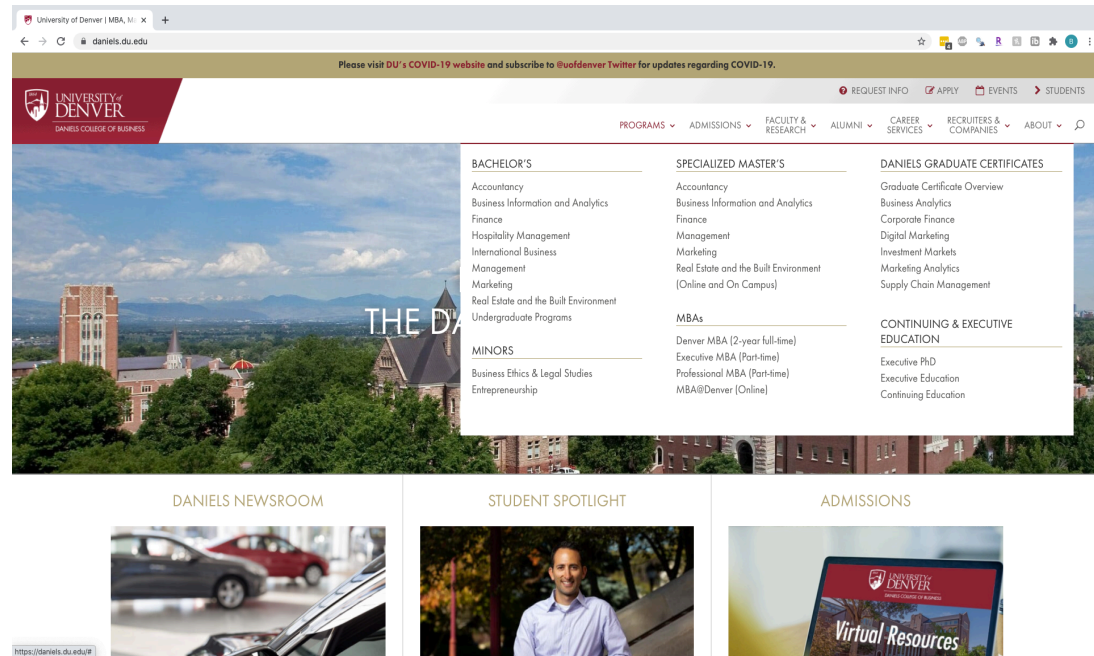
Non-Static Websites

- Dynamic Websites
- APIs

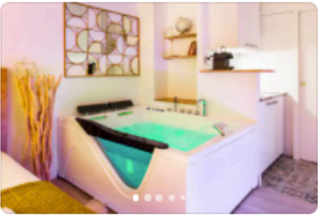


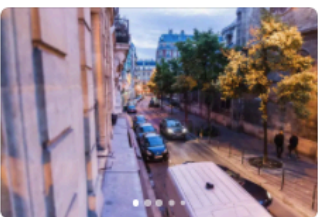
Dynamic Websites

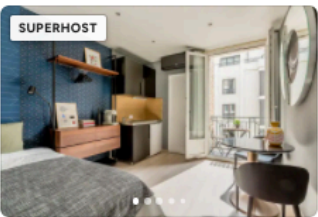
- Drop-downs
- Scrolling
- Pop-ups
- Inputting password






- 

Entire apartment in XVII Arrondissement
Cosy studio 2P- Jacuzzi - Batignolles / Pigalle
 2 guests · Studio · 1 bed · 1 bath
 Wifi · Kitchen
 ★ 4.62 (13) **\$130 / night**
- 

Entire apartment in XI Arrondissement
Appartement cosy 11ème Parmentier-République
 3 guests · 1 bedroom · 2 beds · 1 bath
 Wifi · Kitchen · Washer
 ★ 4.89 (9) **\$67 / night**
- 

SUPERHOST
 Entire apartment in XVII Arrondissement
Chic studio - Levallois/Champerret
 2 guests · Studio · 1 bed · 1 bath
 Wifi · Kitchen
\$47 / night
- 

SUPERHOST
 Entire apartment in Invalides - Ecole Militaire
Charmant studio neuf Invalides. 16m2. Clair.
 2 guests · Studio · 1 bed · 1 bath
 Wifi · Kitchen · Washer
 ★ 5.0 (4) **\$41 / night**



X

★ 4.77 (181 reviews)

you do decide to book this place. Take EXTRA care of yourself around this neighborhood. It is NOT s... [read more](#)



Response from Farah
February 2020

Dear Steven, thank you for your feedback. I'm glad to heard that you loved my place. I'm really sorry about what's happening to you. Please remind that my place is in a quite stree... [read more](#)



Trevor
January 2020

A perfect location



Nailah
January 2020

This was a great place for my first trip to Paris and I wish that I had been able to spend even more time here! I loved the thoughtful details in the apartment such as power converters for guest from different countries.



Kate
December 2019

We had a wonderful trip in Paris thanks to Farah. The apartment was cozy and stylish with everything we needed for our stay. It was nice to hang out late in the morning or come hom... [read more](#)

Season:  **2020 Outdoor**

Switch to **XC**

 [Meet List](#)

 [Teams](#)

 [Men's Top Athletes](#)

 [Women's Top Athletes](#)

 [Performance Lists](#)



 **5A Rankings**

Find a Team in 5A

Region 1

1 

- Canutillo
- El Paso
- El Paso Andress
- El Paso Austin
- El Paso Bowie
- El Paso Burges
- El Paso Chapin
- El Paso Irvin
- El Paso Jefferson

Region 2

09 

- ★ Frisco
- ★ Frisco Centennial
- ★ Frisco Heritage
- ★ Frisco Independence
- ★ Frisco Lebanon Trail
- ★ Frisco Liberty
- ★ Frisco Lone Star
- ★ Frisco Memorial
- ★ Frisco Reedy
- ★ Frisco Wakeland

Region 3

17 

- Cedar Park
- Leander Glenn
- Leander Rouse
- Marble Falls
- Pflugerville
- Pflugerville Connally
- Weiss

18 

- Bastrop

Region 4

25 

- Austin Crockett
- Austin LBJ
- Austin McCallum
- Austin Navarro (form...)
- Austin Northeast
- Austin Travis
- Dripping Springs
- Lockhart

26 



Canutillo

Eagles HS Canutillo, TX | Free Account

37 Followers

- Posts
- Records
- Rankings
- Custom Lists
- ★ Reports
- Training Log
- Photos

View Athletic.net Ad Free

2020 Outdoor Season Calendar ▾

- Outdoor
 - 2020 Outdoor
 - 2019 Outdoor
 - 2018 Outdoor
 - 2017 Outdoor
 - 2016 Outdoor
 - 2015 Outdoor
 - 2014 Outdoor
 - 2013 Outdoor
 - 2012 Outdoor
 - 2011 Outdoor
 - 2009 Outdoor
 - 2008 Outdoor
 - 2007 Outdoor
 - 2006 Outdoor
- Indoor

[Print Calendar](#)
[Download options ▾](#)

< Sep 27 - Oct 3, 2020 >

No Workouts Recorded

★ Upgrade Team [Learn More](#)

PrepSportswear

Price: \$29.99	Price: \$36.99	Price: \$19.99	Price: \$19.99

Athletes

Blog

Web-Crawling

- Automate movement through websites
- Navigate to website, then use techniques Ryan showed us
- Navigation done “remotely” via code

Example: Airbnb Plus

- Airbnb Plus: Airbnb differentiation program
- Hosts apply to be part of Plus program
- Variety of benefits once part of program
- Compare effect of Plus program introduction
- How to determine which listings are plus?
- Work with Karen Xie

Airbnb Plus



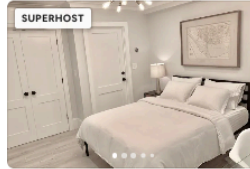
Cambridge | Add dates

300+ stays

Stays in Cambridge

Cancellation flexibility | Type of place | Price | More filters

Review COVID-19 travel restrictions before you book. [Learn more](#)

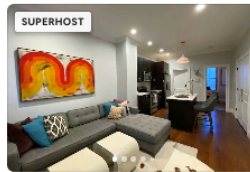


Entire apartment in Cambridge
Luxury studio w/ parking by MIT/Harvard/BU/Fenway

2 guests · 1 bedroom · 1 bed · 1 bath
Free parking · Wifi · Air conditioning

★ 4.87 (169)

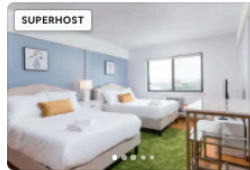
\$92 / night



Entire condominium in Jamaica Plain
Clean & Modern Apt Close to Public Transportation

2 guests · 1 bedroom · 1 bed · 1 bath
Wifi · Air conditioning · Kitchen

\$84 / night

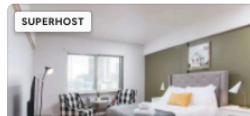


Entire apartment in Allston-Brighton
Gorgeous space in vibrant area, steps to the t617

4 guests · 1 bedroom · 2 beds · 1 bath
Wifi · Air conditioning · Kitchen

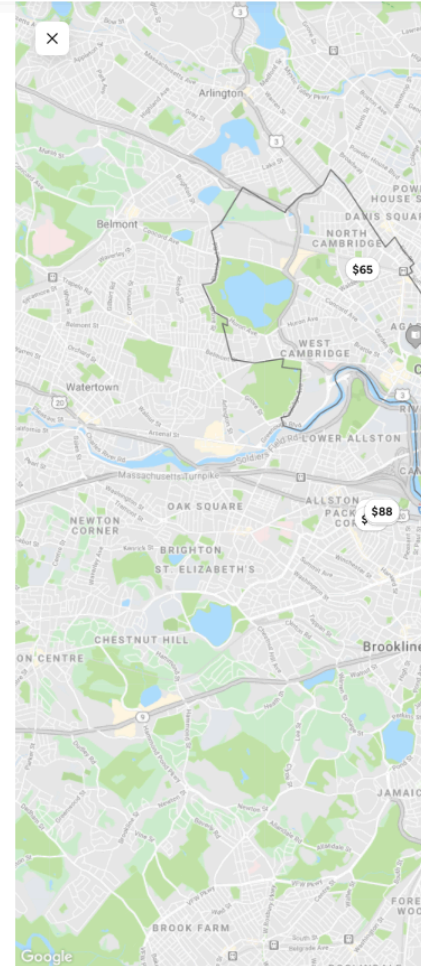
★ 4.92 (36)

\$88 / night



Entire apartment in Allston-Brighton
Chic & Comfort Boston Studio near Subway 616

2 guests · 1 bedroom · 1 bed · 1 bath
Wifi · Air conditioning · Kitchen



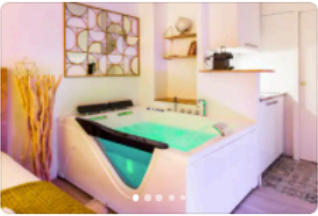
- 1) Identify main city page
- 2) Check if there are multiple listing pages
- 3) Scrape current page
- 4) Click on next page if applicable
- 5) Determine which listings have “plus” in their url

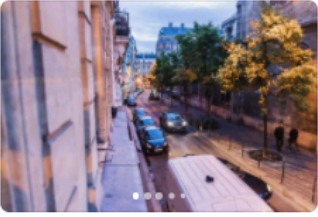
Listing ID Number

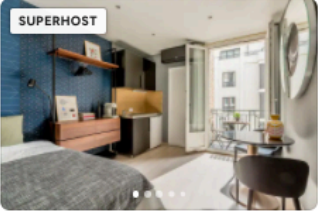
airbnb.com/rooms/plus/19624348?check_in=2020-10-21&check_out=2020-10-22&source_impression_id=p3_1602265346_SL3x7ho4ILtdeoyX


Plus Identifier



- 

Entire apartment in XVII Arrondissement
Cosy studio 2P- Jacuzzi - Batignolles / Pigalle
2 guests · Studio · 1 bed · 1 bath
Wifi · Kitchen
★ 4.62 (13) **\$130 / night**
- 

Entire apartment in XI Arrondissement
Appartement cosy 11ème Parmentier-République
3 guests · 1 bedroom · 2 beds · 1 bath
Wifi · Kitchen · Washer
★ 4.89 (9) **\$67 / night**
- 

SUPERHOST
Entire apartment in XVII Arrondissement
Chic studio - Levallois/Champerret
2 guests · Studio · 1 bed · 1 bath
Wifi · Kitchen
\$47 / night
- 

SUPERHOST
Entire apartment in Invalides - Ecole Militaire
Charmant studio neuf Invalides. 16m2. Clair.
2 guests · Studio · 1 bed · 1 bath
Wifi · Kitchen · Washer
★ 5.0 (4) **\$41 / night**



150 stays

Airbnb Plus stays in New Orleans

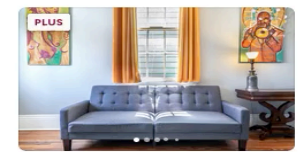
Cancellation flexibility Type of place Price Instant Book More filters -1



Entire apartment in Central City / Garden District
Restored Cottage Two Blocks from a Mardi Gras Pa...

4 guests · 2 bedrooms · 2 beds · 2 baths
 Wifi · Self check-in · Air conditioning

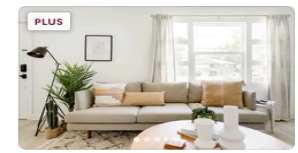
★ 4.96 (182) \$157 / night



Entire house in Mid-City District
Colorful, Historic Home near the French Quarter

4 guests · 2 bedrooms · 4 beds · 2 baths
 Wifi · Air conditioning · Washer · Kitchen

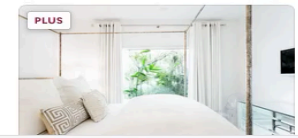
★ 4.89 (85) \$116 / night



Entire house in Gentilly Woods
Visit the French Quarter Near a Family Home With ...

8 guests · 4 bedrooms · 4 beds · 3 baths
 Indoor fireplace · Wifi · Air conditioning · Washer

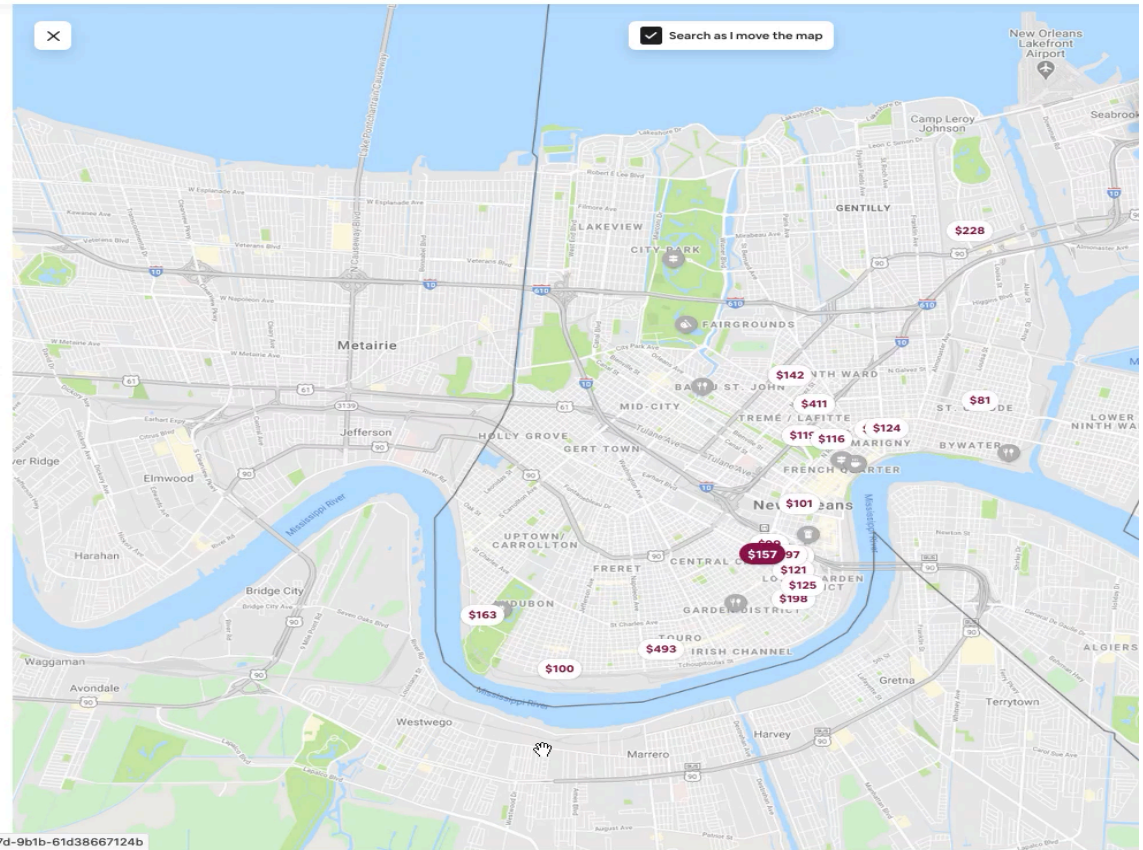
★ 4.90 (21) \$228 / night



Entire condominium in East and West Riverside
Penthouse with Private Outdoor Spaces and Balcony

4 guests · 2 bedrooms · 2 beds · 2.5 baths
 Wifi · Self check-in · Air conditioning

https://www.airbnb.com/rooms/plus/18124272?previous_page_section_name=1000&federated_search_id=d6557e0c-e09a-417d-9b1b-61d38667124b



```

for(i in 1:nrow(cities_url)){
  #city_urls <- list()
  #Navigate to listing url
  remDr$navigate(paste0("",cities_url[i,]))
  Sys.sleep(2)
  # Any pages to click through ?
  click_pages <- remDr$getPageSource()[[1]] %>%
  xml2::read_html() %>%
  html_nodes("._1bdke5s") %>%
  html_text() %>%
  as.numeric() %>%
  max()
}

```

Take a break!

Should we click through pages?

Dynamic Web-scraping

- Each situation is unique
- Requires trial and error
- Tools:
 - Selenium (python, R)



APIs

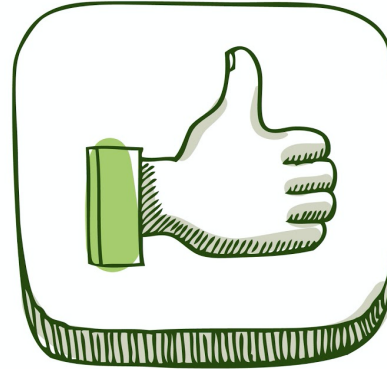
- Application Programming Interface
- “Easily” facilitated connection to apps, websites, etc.
- Another way to extract data from a website/platform



Some examples



APIs



- Pros:

- Can make data collection very smooth
- Popular APIs often have libraries/packages for common software (python, R)

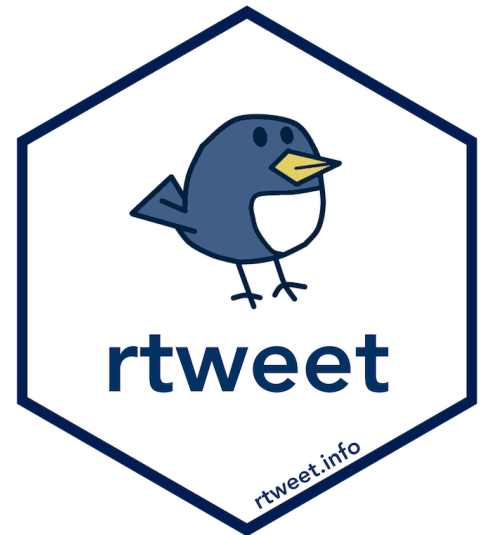
- Cons:

- Restricted Access (only a certain amount of data given per day)
- Data not in format of your choice

Example: Twitter



- What do you need?
 - A Twitter account!
- `rtweet` R package
 - Could use python as well



Example: Twitter

- What can I get?
 - Hashtags
 - Followers
 - Friends
 - Locations
 - Source (android, iPhone, etc)
- Basic: 18,000 tweets every 15 minutes from “rest” API
- More advanced: “streaming” API: much more data

Example: #fakenews

- Can we learn about the spread of #fakenews on Twitter?
- Scrape twice daily, look for #fakenews
- October 27th to December 11th 2019
- Over 170,000 unique tweets



Example: #fakenews

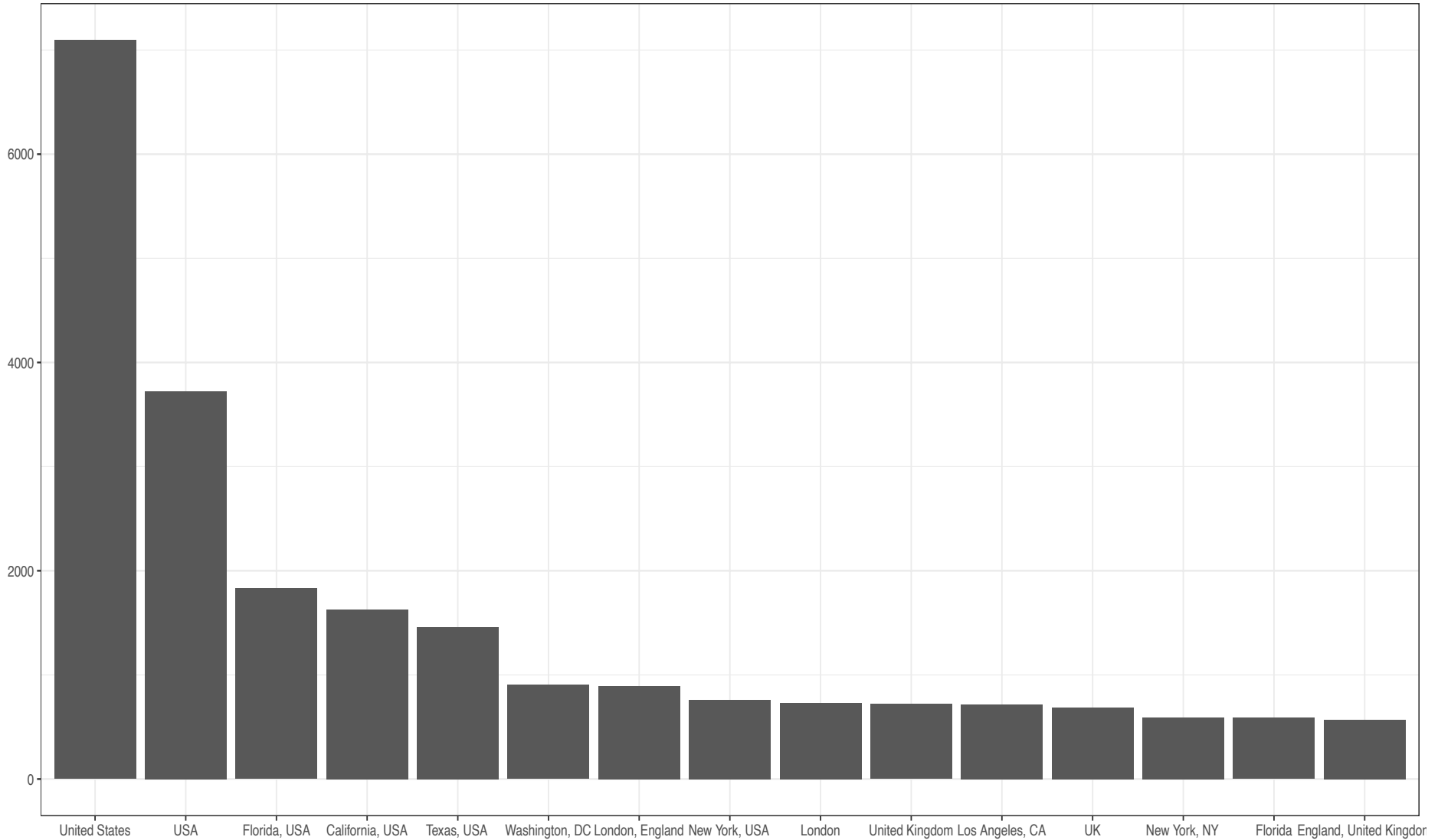
Search for tweets that use the hastag `#fakenews`

Simple code:

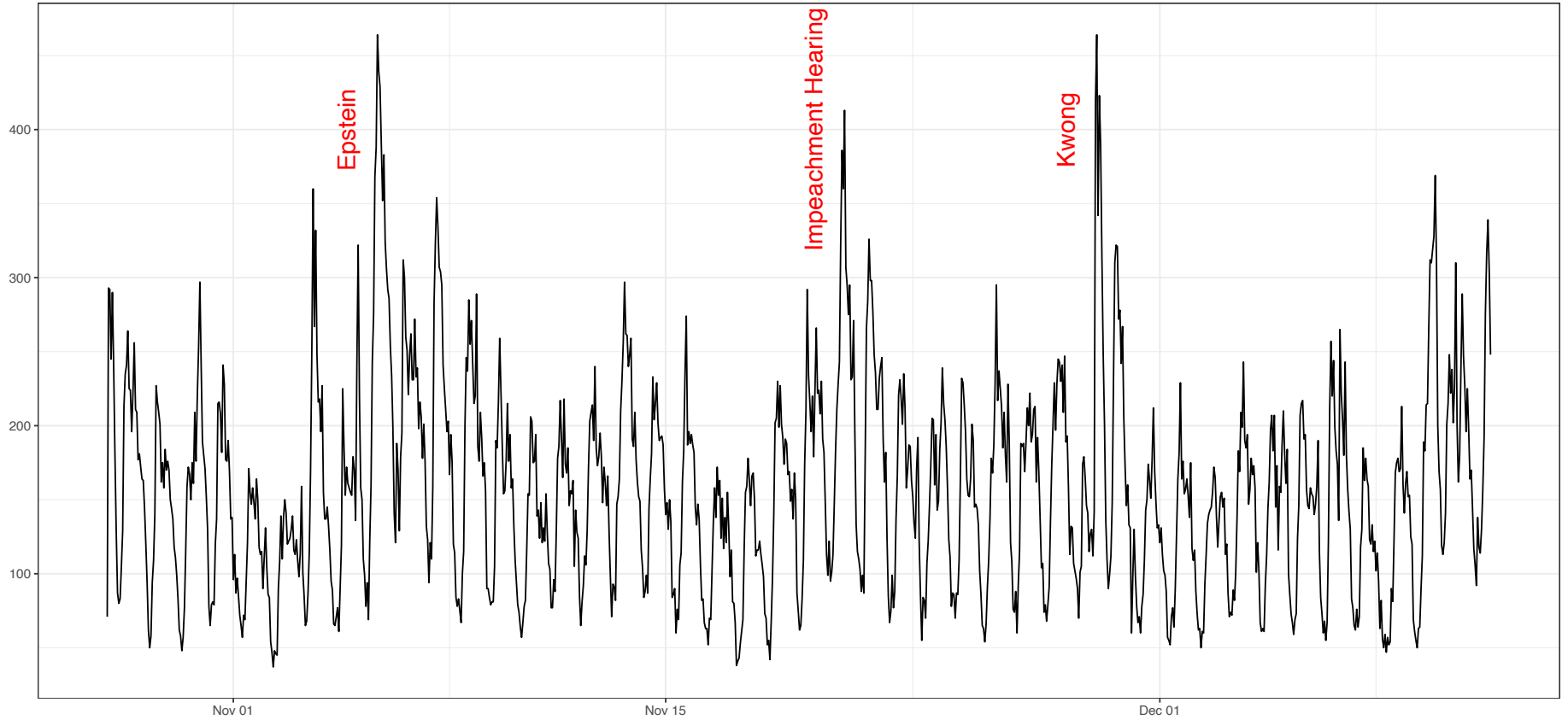
```
search_tweets(
```

```
"#fakenews", n = 18000, include_rts = FALSE, lang = "en")
```

Example: #fakenews



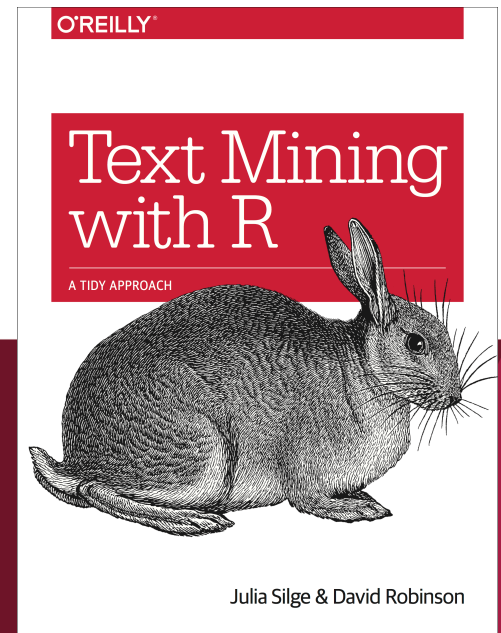
Example: #fakenews



After Scraping...



- Post-scraping analyses
 - Simple (sentiment analysis)
 - Complicated (machine learning)
- Many options, low hanging fruit
- Text Mining with R (Silge & Robinson)
tidytextmining.com



Take-aways

- Dream big about web-scraping!
- Different types of websites have different approaches
- Usually can find a way to scrape data
- Please do not hesitate to contact me for help/collaboration
- benjamin.williams@du.edu

