

QSS20: Modern Statistical Computing

Unit 11: Web-scraping

Goals for today

- ▶ Recap of APIs
- ▶ Lecture on web-scraping
 - ▶ Living in a digital era
 - ▶ Researching the digital era
 - ▶ How to scrape/crawl without reinventing the flat tire
 - ▶ Intro to packages for wrangling web data
 - ▶ Ethics (briefly)
- ▶ Notebook for basic web-scraping

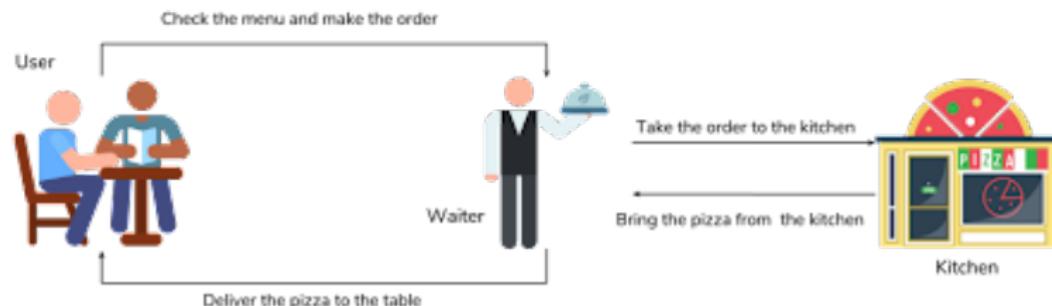
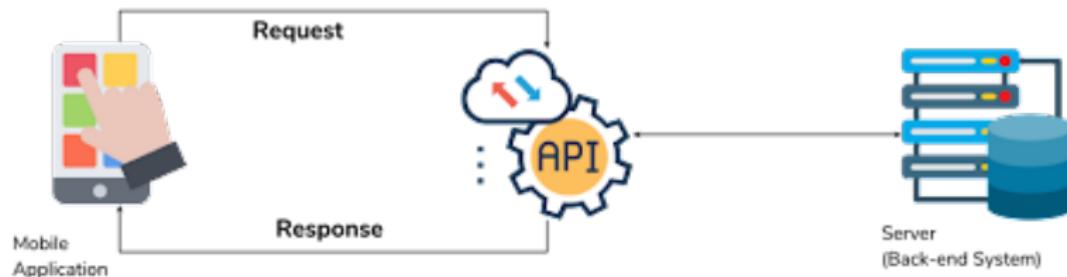
Goals for today

- ▶ **Recap of APIs**
- ▶ Lecture on web-scraping
 - ▶ Living in a digital era
 - ▶ Researching the digital era
 - ▶ How to scrape/crawl without reinventing the flat tire
 - ▶ Intro to packages for wrangling web data
 - ▶ Ethics (briefly)
- ▶ Notebook for basic web-scraping

Recap of APIs

What do you remember?

How APIs work



Source: Medium blog post

Recap of APIs

Tips:

- ▶ APIs are a "front door" to web-based data (visible, accessible, controlled), scraping is the "back door" (hard to find, often locked, unique rewards)
- ▶ APIs are everywhere! New York Times, Reddit, YouTube, Google Maps, Wikipedia, Twitter, etc.
- ▶ Basic workflow: Construct query → call API → check if it worked → if so, extract content
- ▶ Constructing queries usually means starting from a template, sniffing between '&'s, and updating
- ▶ try/except clauses often useful for catching failures without blocking

Useful commands:

```
creds = yaml.safe_load(stream) # load API credentials file
response = requests.get(query) # call API (nearly universal)
response_j = response.json() # get JSON of result (if it worked)
data = pd.DataFrame(response_j[colname]) # make usable
```

Where we are

- ▶ Recap of supervised machine learning
- ▶ **Lecture on web-scraping**
 - ▶ **Living in a digital era**
 - ▶ Researching the digital era
 - ▶ How to scrape/crawl without reinventing the flat tire
 - ▶ Intro to packages for wrangling web data
 - ▶ Ethics (briefly)
- ▶ Notebook for basic web-scraping

Learn about vaccines: Wikipedia

en.wikipedia.org/w/index.php?title=COVID-19_vaccine&oldid=950000000#Vaccine_types

Vaccine types

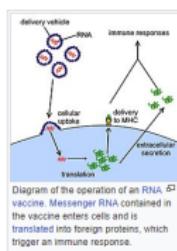
As of January 2021, nine different technology platforms – with the technology of numerous candidates remaining undefined – are under research and development to create an effective vaccine against COVID-19.^{[3][27]}

Most of the platforms of vaccine candidates in clinical trials are focused on the coronavirus spike protein and its variants as the primary antigen of COVID-19 infection.^[27] Platforms being developed in 2020 involved nucleic acid technologies (nucleoside-modified messenger RNA and DNA), non-replicating viral vectors, peptides, recombinant proteins, live attenuated viruses, and inactivated viruses.^{[14][27][28][34]}

Many vaccine technologies being developed for COVID-19 are not like vaccines already in use to prevent influenza, but rather are using "next-generation" strategies for precision on COVID-19 infection mechanisms.^{[27][30][34]}

Several of the synthetic vaccines use a 2P mutation to lock the spike protein into its prefusion configuration, stimulating an immune response to the virus before it attaches to a human cell.^[30] Vaccine platforms in development may improve flexibility for antigen manipulation and effectiveness for targeting mechanisms of COVID-19 infection in susceptible population subgroups, such as healthcare workers, the elderly, children, pregnant women, and people with existing weakened immune systems.^{[27][30]}

RNA vaccines



An RNA vaccine contains RNA which, when introduced into a tissue, acts as messenger RNA (mRNA) to cause the cells to build the foreign protein and stimulate an adaptive immune response which teaches the body how to identify and destroy the corresponding pathogen or cancer cells. RNA vaccines often, but not always, use nucleoside-modified messenger RNA. The delivery of mRNA is achieved by a coformulation of the molecule into lipid nanoparticles which protect the RNA strands and help their absorption into the cells.^{[9][32][33][34]}

RNA vaccines were the first COVID-19 vaccines to be authorized in the United States and the European Union.^{[9][34]} As of January 2021, authorized vaccines of this type are the Pfizer–BioNTech COVID-19 vaccine^{[37][38][39]} and the Moderna COVID-19 vaccine.^{[100][101]} As of February 2021, the CVnCoV RNA vaccine from CureVac is awaiting authorization in the EU.^[102]

Severe allergic reactions are rare. In December 2020, 1,893,360 first doses of Pfizer–BioNTech COVID-19 vaccine administration resulted in 175 cases of severe allergic reaction, of which 21 were anaphylaxis.^[103] For 4,041,396 Moderna COVID-19 vaccine dose administrations in December 2020 and January 2021, only 10 cases of anaphylaxis were reported.^[103] The lipid nanoparticles were most likely responsible for the allergic reactions.^[103]

Adenovirus vector vaccines

These vaccines are examples of non-replicating viral vector vaccines, using an adenovirus shell containing DNA that encodes a SARS-CoV-2 protein.^{[104][105]} The viral vector-based vaccines against COVID-19 are non-replicating, meaning that they do not make new virus particles, but rather produce only the antigen which elicits a systemic immune response.^[104]

As of January 2021, authorized vaccines of this type are the Oxford–AstraZeneca COVID-19 vaccine,^{[106][107][108]} the Sputnik V COVID-19 vaccine,^[109] Convidecia, and the Johnson & Johnson COVID-19 vaccine.^{[110][111]}

Convidecia and the Johnson & Johnson COVID-19 vaccine are both one-shot vaccines which offer less complicated logistics and can be stored under ordinary refrigeration for several months.^{[112][113]}

The Sputnik V COVID-19 vaccine uses Ad26 for the first dose, which is the same as the Johnson & Johnson vaccine's only dose, and Ad5 for the second dose. Convidecia uses Ad5 for its only dose.^[114]

Inactivated virus vaccines

Inactivated vaccines consist of virus particles that have been grown in culture and then are killed using a method such as heat or formaldehyde to lose disease producing capacity, while still stimulating an immune response.^[115]

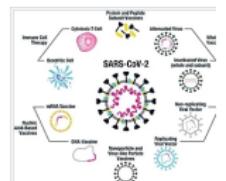
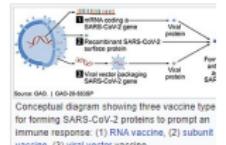
As of January 2021, authorized vaccines of this type are the Chinese CoronaVac,^{[116][117][118]} BBIBP-CoV,^[119] and WIBP-CoV; the Indian Covaxin; and the Russian CovIvac.^[120] Vaccines in clinical trials include the Valneva COVID-19 vaccine.^{[121][122]}

Subunit vaccines

Subunit vaccines present one or more antigens without introducing whole pathogen particles. The antigens involved are often protein subunits, but can be any molecule that is a fragment of the pathogen.^[123]

As of April 2021, the two authorized vaccines of this type are the peptide vaccine EpiVacCorona^[124] and RBD-Dimer.^[3] Vaccines with pending authorizations include the Novavax COVID-19 vaccine,^[125] SOBERANA 02 (a conjugate vaccine), and the Sanofi–GSK va

Other types



Vaccine platforms being employed for SARS-CoV-2. Whole virus vaccines include both attenuated inactivated forms of the virus. Protein and peptide subunit vaccines are usually combined with an adjuvant in order to enhance immunogenicity. The main emphasis in SARS-CoV-2 vaccine development has been on using the whole spike protein in its trimeric form or components of it, such as the RBD region. Multiple non-replicating viral vector vaccines have been developed, particularly focused on adenovirus, while there has been less emphasis on replicating viral vector constructs.^[99]

Find housing: Affordable Housing Online

affordablehousingonline.com



Section 8 Waiting Lists Apartment Waiting List

Search Low Income Apartments And Waiting Lists

Search and Select City, State, or Zip

Latest Affordable Housing News

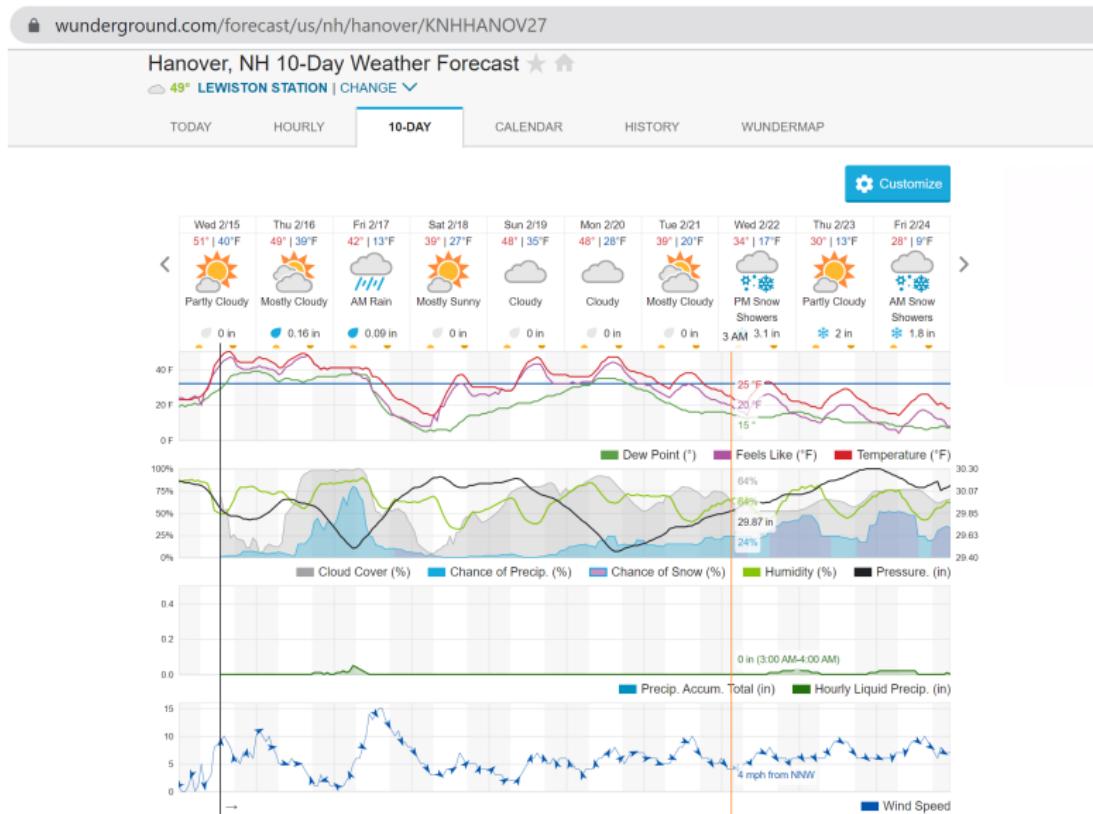


Waiting List News for October, 31st 2022

Vermillion County, Illinois
Opens on November 10th, 2022

The Vermillion Housing Authority (VHA) is accepting Section 8 Housing Choice Voucher waiting list applications from November 9, 2022 at 2:00 pm, until November 11, 2022 at 9:00 pm CT.

Check weather: [wunderground.com](https://wunderground.com/forecast/us/nh/hanover/KNHHANOV27)



Follow current events: New York Times

[nytimes.com](https://www.nytimes.com)

Wednesday, February 13, 2024
Today's Paper

U.S. INTERNATIONAL CANADA ESPAÑOL 🇺🇸

PLAY THE CROSSWORD ACCOUNT ▾
45°F 84° 30°
Nasdaq +0.1% ↑

World U.S. Politics N.Y. Business Opinion Science Health Sports Arts Books Style Food Travel Magazine Real Estate Cooking The Athletic Worcester Games

As Lawmakers Spar Over Social Security, Its Costs Are Rising Fast

New budget forecasts, set to be released today, are expected to show that spending for Social Security is rapidly outpacing growth in tax revenues.

President Biden seized political points this month by forcing G.O.P. leaders to profess that they will not seek cuts to the programs.

5 MIN READ

President Biden's I.R.S. nominees is poised to defend a plan to modernize the tax agency at a hearing today.

5 MIN READ

In a Bleak Russian Cemetery, a Sea of Crosses Signals the War's True Toll

A snowy plot is the final resting place for soldiers from the Wagner mercenary forces, a testament to Russia's soaring casualty count.

5 MIN READ

Losses Near Ukrainian Stronghold Cast Doubt on Russia's Offensive

Weeks of failed attacks around Volgograd have left two Russian brigades in tatters, but the fighting has also come at a cost for Ukraine.

2 MIN READ

Daily Briefing: War in Ukraine

A Ukrainian military official said his forces "fought back decently" in Volgograd.

As Russia and Ukraine expend ammunition at a staggering rate, the race to rearm takes on added urgency.

LIVE 4h 30m

Nicola Sturgeon, Scotland's Leader, Says She Will Step Down in Surprise Move

A longtime champion of Scottish

The New York Times

RESIDENT JOE BIDEN TARA JAPKOWSKI PROTECT AND STRENGTHEN MEDICARE PROTECT AND STRENGTHEN MEDICARE

Hector Retamal/Agence France-Presse via Getty Images

A Love Letter to Libraries, Long Overdue

The New York Times sent photographers to seven states to document the thrum and buzz in buildings once known for silence.

2 MIN READ

Asia Society Tokyo via Getty Images

Here's how a consistent sleep schedule might protect your heart.

5 MIN READ

Netflix is shunning live sports but embracing sports documentaries.

4 MIN READ

Opinion

'Nikki Haley Will Not Be the Next President': Our Columnists Weigh In

DAVID FRENCH

The Law Is Closing In on Trump

JESSICA GROSE

K M O D I L I S C

Where we are

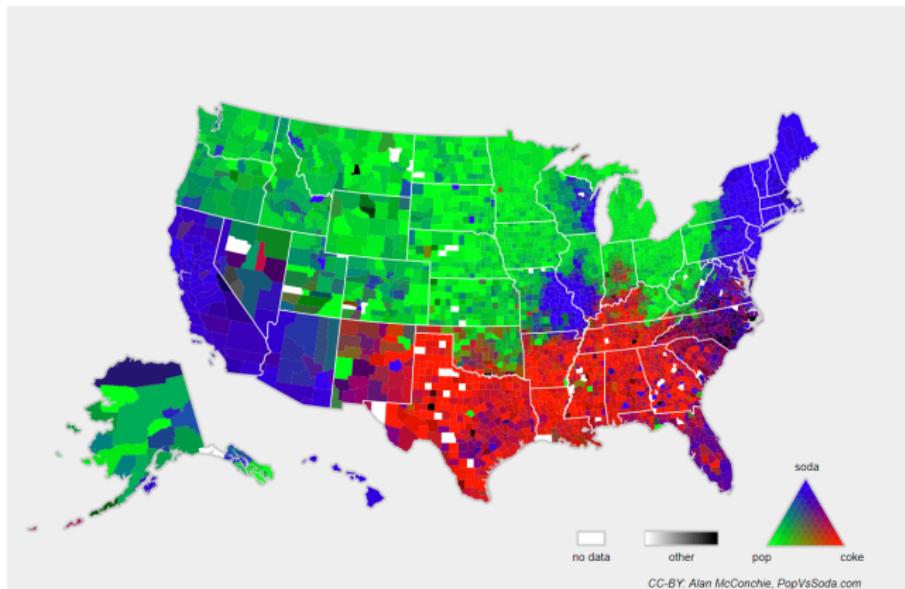
- ▶ Recap of supervised machine learning
- ▶ **Lecture on web-scraping**
 - ▶ Living in a digital era
 - ▶ **Researching the digital era**
 - ▶ How to scrape/crawl without reinventing the flat tire
 - ▶ Intro to packages for wrangling web data
 - ▶ Ethics (briefly)
- ▶ Notebook for basic web-scraping

Social media → geography

POP vs SODA

Input: Tweets

Output:
Latitude,
Longitude

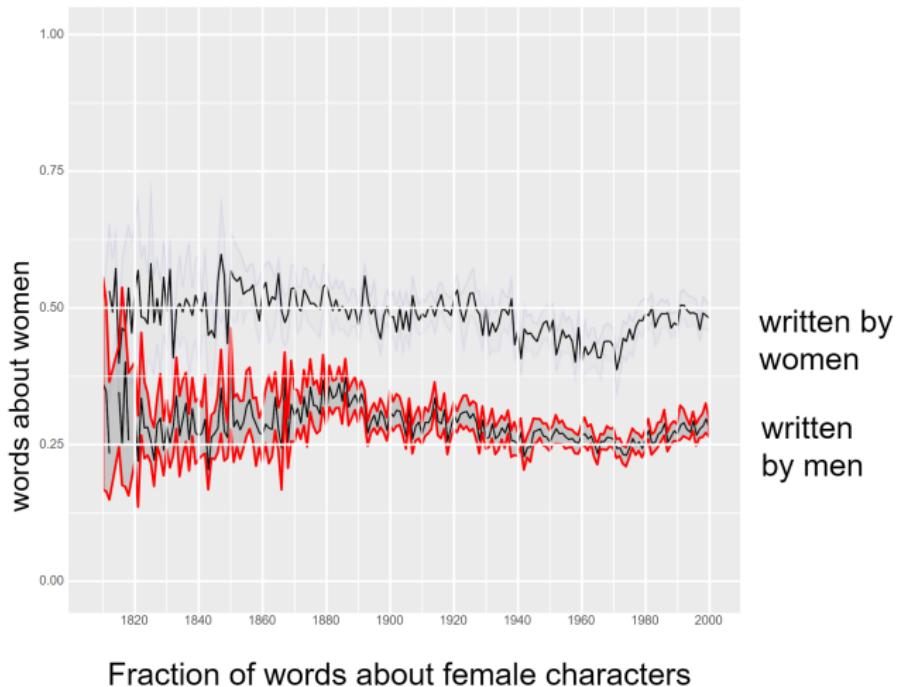


Benjamin Wing and Jason Baldridge (2011), "Simple supervised document geolocation with geodesic grids" (ACL)

Books → gendered language

Input:
Fictional texts

Output:
Prop. words
about females

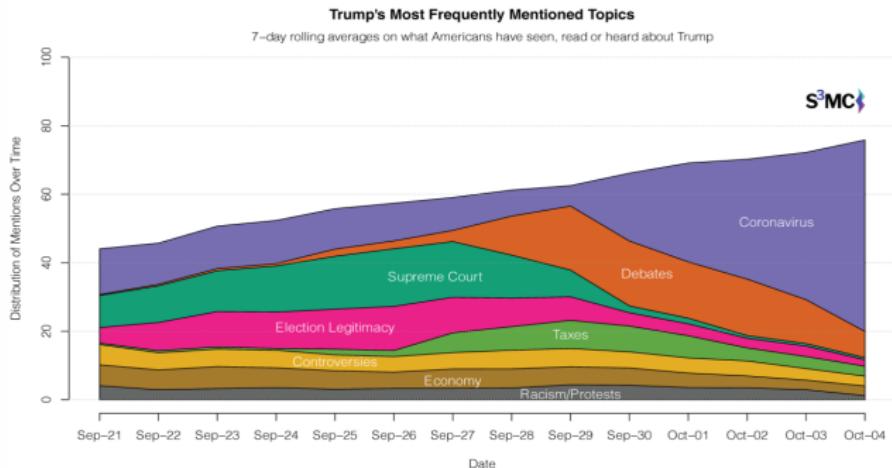


Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," (Cultural Analytics)

Presidential language → public attention

Input:
News articles

Output:
Topics recal-
led by public

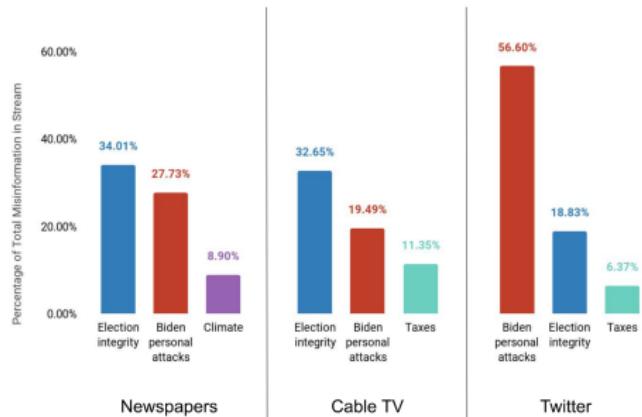


Jennifer Agiesta (2020), "In News about the Presidential Race, Coronavirus Overtakes Nearly All Else" (CNN Politics)

Digital media → misinformation over time

Input:
Articles,
tweets,
segments

Output:
Misinformation
themes pushed
by presidential
candidates



Jaren Haber, Lisa Singh, Ceren Budak, Josh Pasek, et al. (2021), Lies and presidential debates: How political misinformation spread across media streams during the 2020 election (HKS Misinfo. Review)

Obituaries → mortality by subgroups

Input:
Online
obituaries
from
Legacy.com



Output:
Models
of all-
cause
mortality;
predicted
trends

Where we are

- ▶ Recap of supervised machine learning
- ▶ **Lecture on web-scraping**
 - ▶ Living in a digital era
 - ▶ Researching the digital era
 - ▶ **How to scrape/crawl without reinventing the flat tire**
 - ▶ Intro to packages for wrangling web data
 - ▶ Ethics (briefly)
- ▶ Notebook for basic web-scraping

Existing data sources: Dartmouth Library

 researchguides.dartmouth.edu/az.php

Search books, articles, & more 

Research Support ▾ Borrow & Renew ▾ Libraries & Spaces ▾ Help ▾ Login 

Dartmouth Library / Research Guides / Selected Databases

Selected Databases

Find library databases for your research which have been selected by a subject librarian. The Selected Databases A-Z list is not intended to be a comprehensive list of databases.

All Subjects  Search A-Z list 

All A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #

352 DATABASES FOUND

A

A&E Portal 

Discover important art and architectural history scholarship from some of the world's finest publishers and museums. Search, read, cite, note, and more.
[more...](#)

AAPG/Datapages 

Online fulltext search platform for geoscientists working in energy fields (primarily oil, gas, and energy minerals) including material in environmental geosciences. Includes the AAPG Bulletin, Memoirs and Special Publications, as well as other AAPG digitized books and the publications of many regional geological societies.

ABI/INFORM Collection 

Indexes business and management journals worldwide; covers corporate strategies, marketing, product development, industry conditions, management techniques, business trends, management practice and theory, corporate strategy and tactics, and more. Includes the full text of articles from many publishers and links to full text options for other articles. 1923 to present.

NEW / TRIAL DATABASES

The following databases are newly acquired or being evaluated for a future subscription.

Docuseek sustainability collection  

The Sustainability Collection encompasses a wide array of disciplines and approaches to sustainability, including new approaches to urban design, the implications of energy choices, and new and traditional agricultural methods and food distribution strategies. The collection shows in a variety of ways and places how design, conservation, community, and legislative action are all crucial components of a sustainable future at both the local and global level.

Global environmental justice collection  

This is a curated collection of 35 documentaries that cover a wide range of subject areas, from Asian, environmental, and Indigenous studies to law, geography, anthropology, global health, policy, conservation biology, and more.

Existing data sources: Business

researchguides.dartmouth.edu/business/quicklinks 🔍 🔍 ⭐

Search books, articles, & more
Research Support Borrow & Renew Libraries & Spaces Help Login

Dartmouth Library / Research Guides / Dartmouth Library Guides / Business / Quick Links to Databases

Business: Quick Links to Databases

Find information resources on companies, industries, market research, entrepreneurship, M&A, hedge funds, private equity, finance, energy, nonprofits, healthcare, and real estate.

Search this Guide Search

Home
Quick Links to Databases
Library Catalog
Articles / News
Books
HBS Cases
Workshops

QUICK LINKS TO BUSINESS DATABASES

- Restricted to Tuck
- Tuck community:** Find passwords/instructions
- Off Campus Access: Use VDI or VPN First

A TO Z DATABASE DESCRIPTIONS

- A-F**
- G-M**
- N-Z**

Company Research

- Bloomberg
- Calcbench
- Capital IQ | Register
- D&B Hoovers
- EMIS Emerging Markets
- Intrinio | Register & Instructions
- Marketline Advantage
- Mergent Archives
- ORBIS
- PrivCo
- S&P's NetAdvantage
- Refinitiv Workspace | Register

Industry Research / Market Research

- AdSpender
- BCC Research
- Bizminer
- Business & Community Analyst -
- ArcGIS
- Claritas 360
- D&B Hoovers
- EASI Market Planner
- eMarketer
- EMIS Emerging Markets
- Film Industry Data

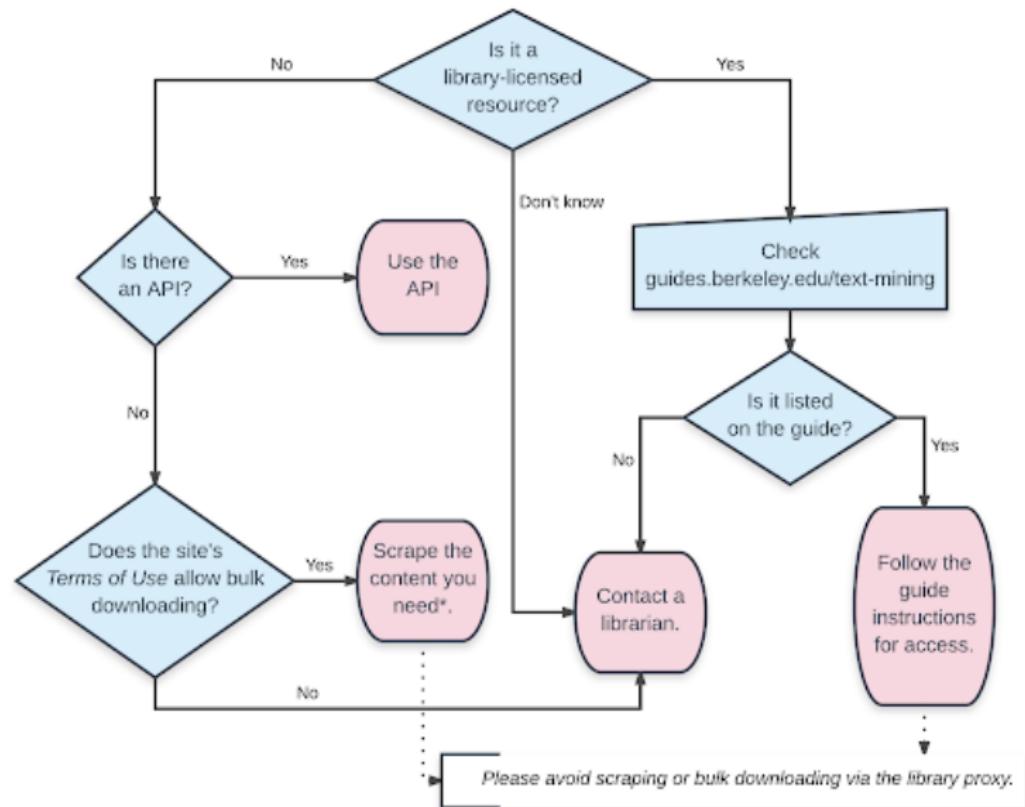
• **ABI/INFORM Collection**

Indexes business and management journals worldwide; covers corporate strategies, marketing, product development, industry conditions, management techniques, business trends, management practice and theory, corporate strategy and tactics, and more. Includes the full text of articles from many publishers and links to full text options for other articles. 1923 to present.

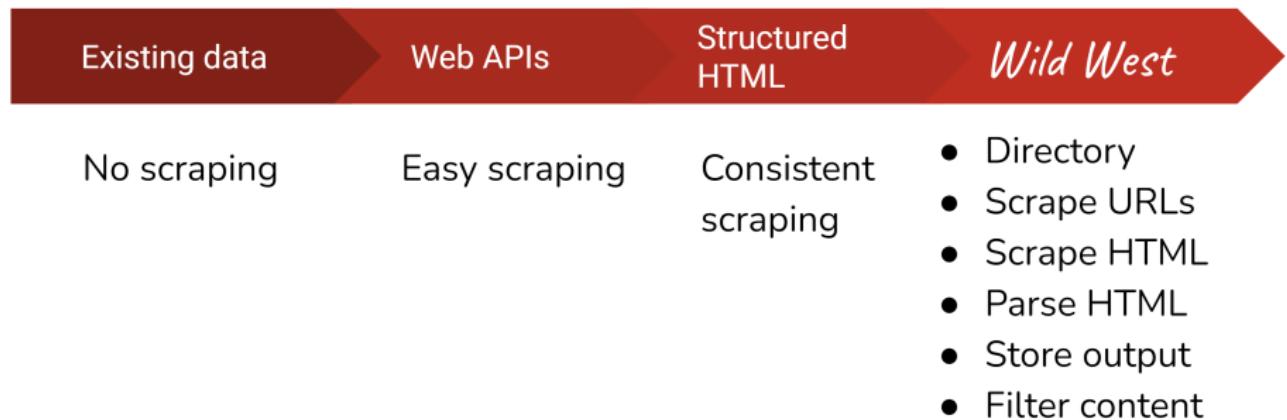
• **AdSpender**

Database of top level advertising expenditure, across 18 media and for 3+ million brands. Historical 10-year database of national summary spending trends, accessible by industry, parent company, and brand.

Data acquisition decision tree: Library perspective



Web-scraping decision hierarchy



How the web works



Please send me the files to display this website (REQUEST)



Here you go, parse away (RESPONSE)



Where we are

- ▶ Recap of supervised machine learning
- ▶ **Lecture on web-scraping**
 - ▶ Living in a digital era
 - ▶ Researching the digital era
 - ▶ How to scrape/crawl without reinventing the flat tire
 - ▶ **Intro to packages for wrangling web data**
 - ▶ Ethics (briefly)
- ▶ Notebook for basic web-scraping

```
# Simple page HTML grabbing
import requests

url =
'https://www.politifact.com/factchecks/2021/apr/02/citizens-united/citizens-united-calls-bidens-infrastructure-plan-g/'
response = requests.get(url)
html = response.text
```



BeautifulSoup

```
# Anchor extraction from HTML document
from bs4 import BeautifulSoup
from urllib.request import urlopen
with
urlopen('https://en.wikipedia.org/wiki/
Main_Page') as response:
    soup = BeautifulSoup(response,
'html.parser')
    for anchor in soup.find_all('a'):
        print(anchor.get('href', '/'))
```



Parsing with an exclusion list

```
1 soup = BeautifulSoup(open(HTML_file), "lxml")
2 inline_tags = ["b", "big", "i", "small", "tt", "abbr", "
    acronym", "cite", "dfn", "em", "kbd", "strong", "samp",
    "var", "bdo", "map", "object", "q", "span", "sub", "sup"
] # junk tags to eliminate
3
4 [s.extract() for s in soup(['style', 'script', 'head',
    'title', 'meta', '[document]'])] # Remove non-visible
    tags
5
6 for it in inline_tags:
    [s.extract() for s in soup("</" + it + ">")] # Remove
        inline tags
7
```



Scrapy

```
# Simple quote scraper
import scrapy
class SimpleSpider(scrapy.Spider):
    name = "simple"
    start_urls =
['http://quotes.toscrape.com/page/1/']

    def parse(self, response):
        page = response.url.split("/")[-2]
        filename = 'quotes-%s.html' % page
        with open(filename, 'wb') as f:
            f.write(response.body)
```

Terminology for scraping

- ▶ **Uniform Resource Locator (URL):** The address of information on the web and directions to get there. A URL points to resources—usually the files needed to show a website, but it can also point to raw data and such
- ▶ **Domain name:** A website identifier that begins a URL: for instance, in `https://www.example.com/` this is everything from 'https' to '.com/'
- ▶ **Web-scraping (i.e., 'screen-scraping')**: Extracting structured information from the files that make up websites (i.e., what's shown in web browsers), relying on their HTML, CSS, and sometimes JS files
- ▶ **Hyper-Text Markup Language (HTML):** The standard markup language for websites, the "nuts and bolts" of WHAT a website will display, including text
- ▶ **Cascading Style Sheets (CSS):** A technology used to format the layout of a webpage, i.e. HOW to make it pretty. For web-scraping purposes, this is usually less important than the webpage's HTML

Where we are

- ▶ Recap of supervised machine learning
- ▶ **Lecture on web-scraping**
 - ▶ Living in a digital era
 - ▶ Researching the digital era
 - ▶ How to scrape/crawl without reinventing the flat tire
 - ▶ Intro to packages for wrangling web data
 - ▶ **Ethics (briefly)**
- ▶ Notebook for basic web-scraping

Ethics (briefly)

Politeness good

Hacking bad

Ethics (briefly)

- ▶ How do I know what's OK to scrape?
 - ▶ Ask your IRB (they don't always know)
 - ▶ Follow the website's Terms of Service (usually)
- ▶ Am I legally liable if I break the Terms of Service?
 - ▶ Computer Fraud and Abuse Act (1986): Kindof
 - ▶ LinkedIn v. HiQ Labs (2019, 2022): Nope, web-scraping is a public right
- ▶ So is the whole web in the public domain?
 - ▶ It depends (contextual privacy; see Bit by Bit, 2017)

Where we are

- ▶ Recap of supervised machine learning
- ▶ Lecture on web-scraping
 - ▶ Living in a digital era
 - ▶ Researching the digital era
 - ▶ How to scrape/crawl without reinventing the flat tire
 - ▶ Intro to packages for wrangling web data
 - ▶ Ethics (briefly)
- ▶ **Notebook for basic web-scraping**

Notebook for basic web-scraping

https://github.com/jhaber-zz/QSS20_public/blob/main/activities/07_scraping_blank.ipynb

- ▶ Making requests
- ▶ Parsing HTML
- ▶ Scraping URLs

We covered

Concepts:

- ▶ Living in a digital era
- ▶ Researching the digital era
- ▶ How to scrape/crawl without reinventing the flat tire
- ▶ Intro to packages for wrangling web data
- ▶ Ethics (briefly)

Methods:

- ▶ Making requests
- ▶ Parsing HTML
- ▶ Scraping URLs (maybe self-study)

Last thing: Notecards

What's one thing you learned today and one lingering question or challenge?