

Web-Crawling:

A Practical Introduction in Python

Start your server: <https://bit.ly/crawl-python-2021>

GitHub repository: <https://bit.ly/repo-crawl-2021>

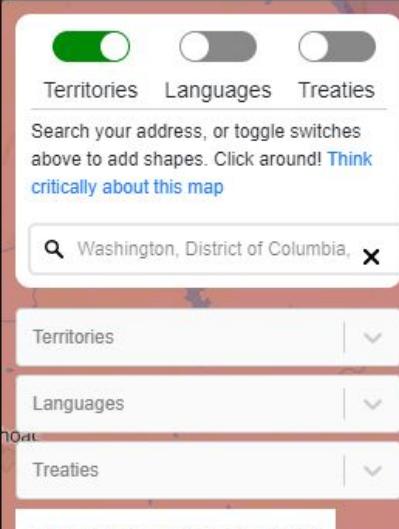
A Massive Data Institute (MDI) Workshop
By Jaren Haber, PhD, Postdoctoral Fellow
April 27, 2021 @ Georgetown University



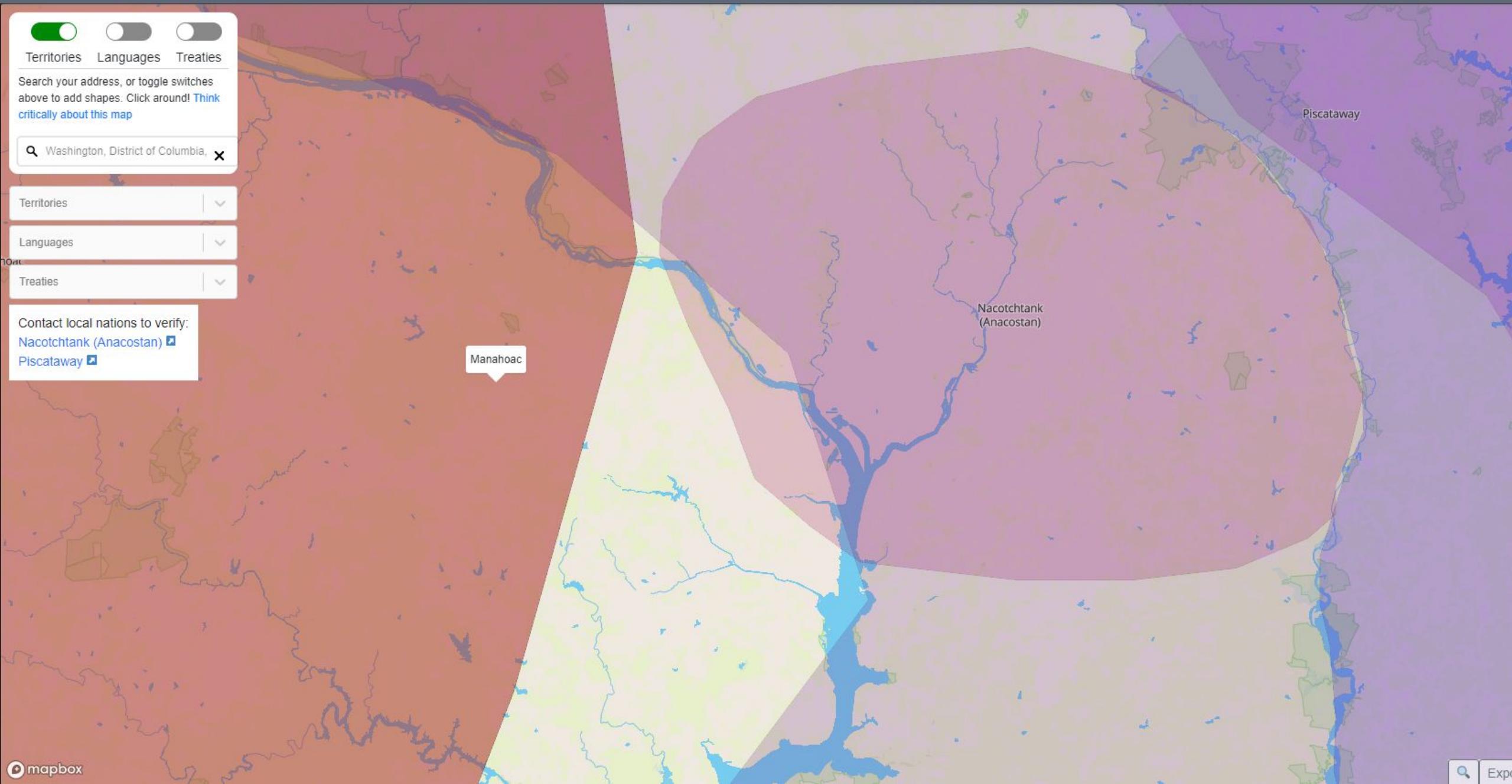


Land acknowledgement

- A few tribes of **Powhatan Confederacy**, Virginia side:
 - **Patawomeck** (i.e., Potomac): now based in Stafford County, VA, with ~2,300 enrolled members
 - **Doeg** (Tauxenent): above Aquia Creek, now extinct
 - **Pamunkey** and **Mattaponi**: Still on 17th century res.
- Tribe on Maryland side of DC area:
 - **Piscataway** (or Conoy): now in Southern Maryland with ~4,100 total population



Contact local nations to verify:
Nacotchtank (Anacostan)
Piscataway





Agenda

- Living in a digital era
- Researching the digital era
- How to scrape without hurting yourself
- A *Wild West* deep dive: Crawling charter school websites
- Hands-on practice



Living in a digital era

[Apartments](#)[Section 8 Waitlists](#)[Public Housing Waitlists](#)[Email Alerts](#)[FAQs](#)

Search Low Income Apartments And Waiting Lists

Search and Select City, State, or Zip

Affordable Housing Online is monitoring the federal government's response to the coronavirus disease (COVID-19) outbreak.

[An extensive list of coronavirus resources for low-income households can be found here.](#)

© [avalara.com](#)

See Nexus Laws by State

Find out how sales trigger nexus in each of the states you sell to.



[VISIT SITE](#)

Housing Waiting List News for April, 26th 2021

Falmouth, MA Section 8 HCV and Mainstream Section 8 HCV Waiting Lists

Open until May 31st, 2021

The Falmouth Housing Authority (FHA) Section 8 Housing Choice Voucher waiting list and Mainstream Section 8 Housing Choice Voucher waiting list applications are being accepted until May 31, 2021 at 4:30 pm ET.

[Learn How To Apply](#)

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][28][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.^[90] 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][28]}

RNA vaccines

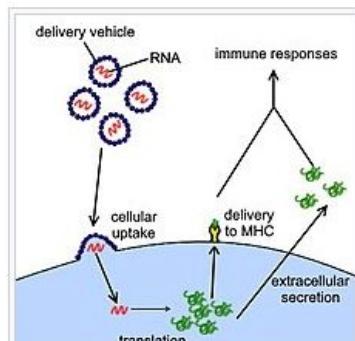


Diagram of the operation of an RNA vaccine. Messenger RNA contained in the vaccine enters cells and is translated into foreign proteins, which trigger an immune response.

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.^{[91][92][93][94]}

RNA vaccines were the first COVID-19 vaccines to be authorized in the United States and the European Union.^{[95][96]} As of January 2021, authorized vaccines of this type are the Pfizer–BioNTech COVID-19 vaccine^{[97][98][99]} 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

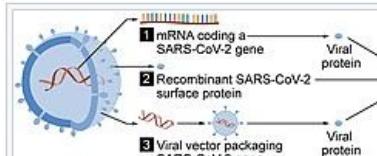
Inactivated vaccines consist of virus particles that have been grown in culture and then are killed using a method such as heat or formaldehyde to lose disease producing capacity, while still stimulating an immune response.^[115]

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

Subunit vaccines

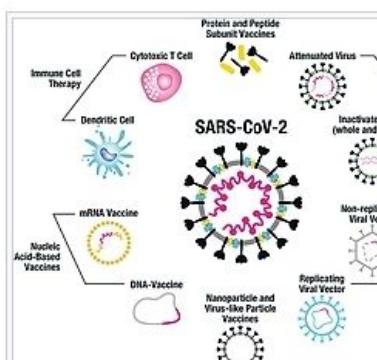
Subunit vaccines present one or more antigens without introducing whole pathogen particles. The antigens involved are often protein subunits, but can be any molecule that is a fragment of the pathogen.^[123]

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



Source: GAO | GAO-20-583SP

Conceptual diagram showing three vaccine types for forming SARS-CoV-2 proteins to prompt an immune response: (1) RNA vaccine, (2) recombinant SARS-CoV-2 surface protein, (3) viral vector vaccine



Vaccine platforms being employed for SARS-CoV-2. Whole virus vaccines include both attenuated and 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 receptor-binding domain. Multiple non-replicating viral vector vaccines have been developed, particularly focused on the spike protein using an adenovirus vector; while there has been less emphasis on replicating viral vector constructs.^[89]

Washington, DC 10-Day Weather Forecast ★ 🏠

68° RONALD REAGAN WASHINGTON NATIONAL AIRPORT STATION | CHANGE ▾

TODAY

HOURLY

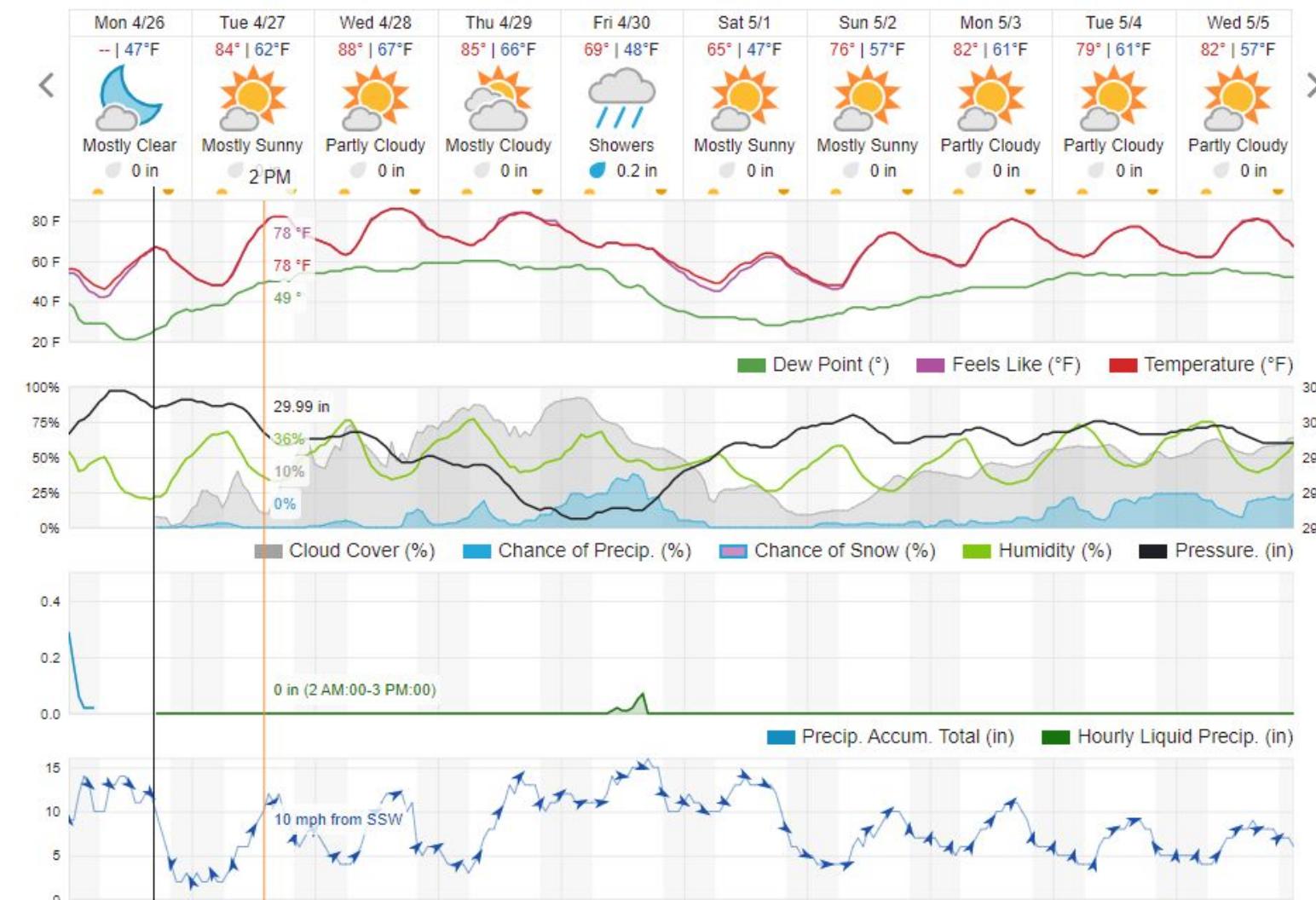
10-DAY

CALENDAR

HISTORY

WUNDERMAP

Customize



COVID-19 Information Center

Find a COVID-19 Vaccine

Who Can Get a COVID-19 Vaccine

CDC recommends that everyone 12 years and older get a COVID-19 vaccine.

People who can get a COVID-19 vaccine include:

- People living in long-term care facilities
- People 65 and older
- Individuals with certain medical conditions
- Healthcare workers
- People 12 and older

Facebook

Sponsored by: Facebook

Facebook Is Supporting The COVID-19 Vaccine Effort

Visit Facebook's COVID-19 Information Center for timely, reliable updates about COVID-19 and vaccines

See More



Listen to 'The Daily'
Why Russia is exporting so many vaccines.



Opinion: Listen to 'The Ezra Klein Show'
Noam Chomsky on anarchism, human nature and President Biden.



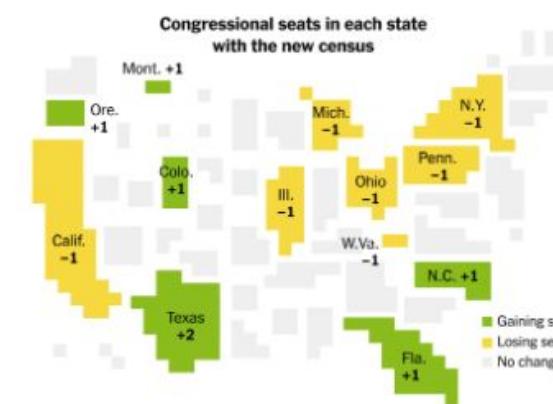
Got a Confidential News Tip?
The Times would like to hear from readers who want to share messages and materials with our journalists.

U.S. Population Over Last Decade Grew at Slowest Rate Since 1930s

- With immigration leveling off and a declining birthrate, the U.S. may be entering an era of substantially lower population growth, demographers said.
- The Census Bureau also reported changes to the political map: The South and the West gained population — and Congressional representation.

Which States Will Gain or Lose Seats in the Next Congress

Texas and Florida will get more House representatives. California and New York will lose slots. Here's how the 2020 census redistributes political power.



Washington Updates

- D.H.S. to review how it identifies and addresses extremism and white supremacy in its ranks.
- The Justice Department will investigate the Louisville police, Garland says.

New York will lose one House seat starting with the 2022 election, after coming up 89 people short on the Census.

LIVE

U.S. Expected to Share AstraZeneca Vaccine Doses With Other Nations

The Biden administration plans to share up to 60 million doses, as long as they clear a safety review, officials said. Here's the latest on Covid-19.

India's Fashion Artisans Face 'Extreme Distress' in Pandemic

The working conditions of the so-called karigars, who make handicrafts for luxury brands, have long been an issue. Now many have no job at all.



Karigars is an Urdu term for artisans who specialize in handicrafts like embroidery, beading and appliquéd. Atul Lalwani for The New York Times

Opinion

Gail Collins and Bret Stephens

Joe Biden Has Something Else He'd Like to Transform

"I can't remember the last time the conversation was so polarized."

Jessica Bruder

I Lived in a Van to Write the Book Behind 'Nomadland.' The Fear Is Real.

For people whose only home is a vehicle, "the knock" is a visceral threat.



The European Union may soon reopen to American tourists. Here's what to know.

United States > On Apr. 25 14-day change

New cases 33,662 -16% ↗

New deaths 282 -3% ↗

Exposure risk in your area >

Search for a county

Find specific information about the county where you live.

U.S. vaccinations >



Kathleen Kingsbury

Why The New York Times Is Retiring the Term 'Op-Ed'

'All in All, the Worst'



'Sway'
The C.I.A.'s Top Technologist Is Uncomfortable With Facebook

Daniela Gerson

Researching the digital era

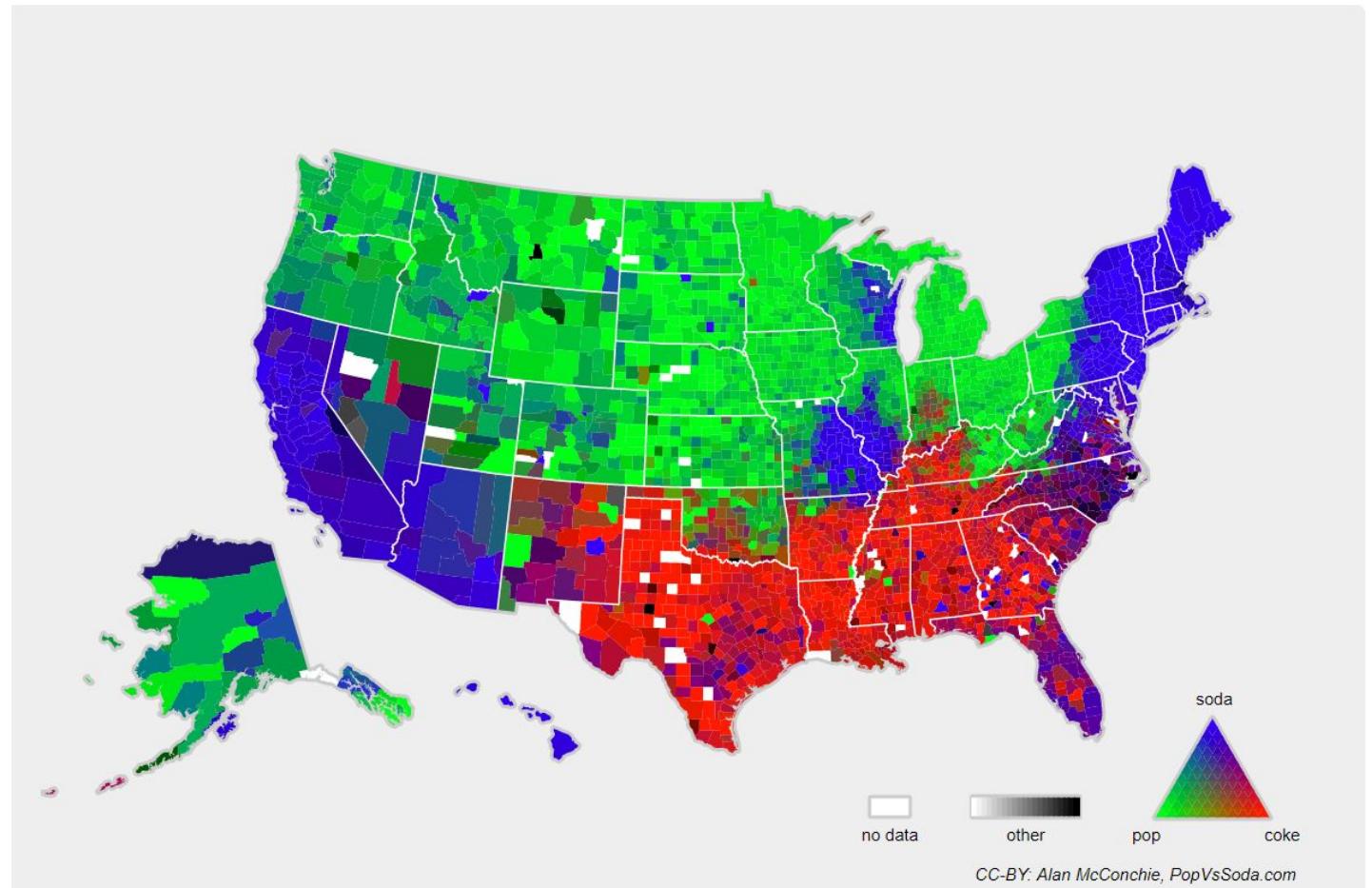


Social media → geography

Input: Tweet

Output: Latitude,
Longitude

POP vs SODA



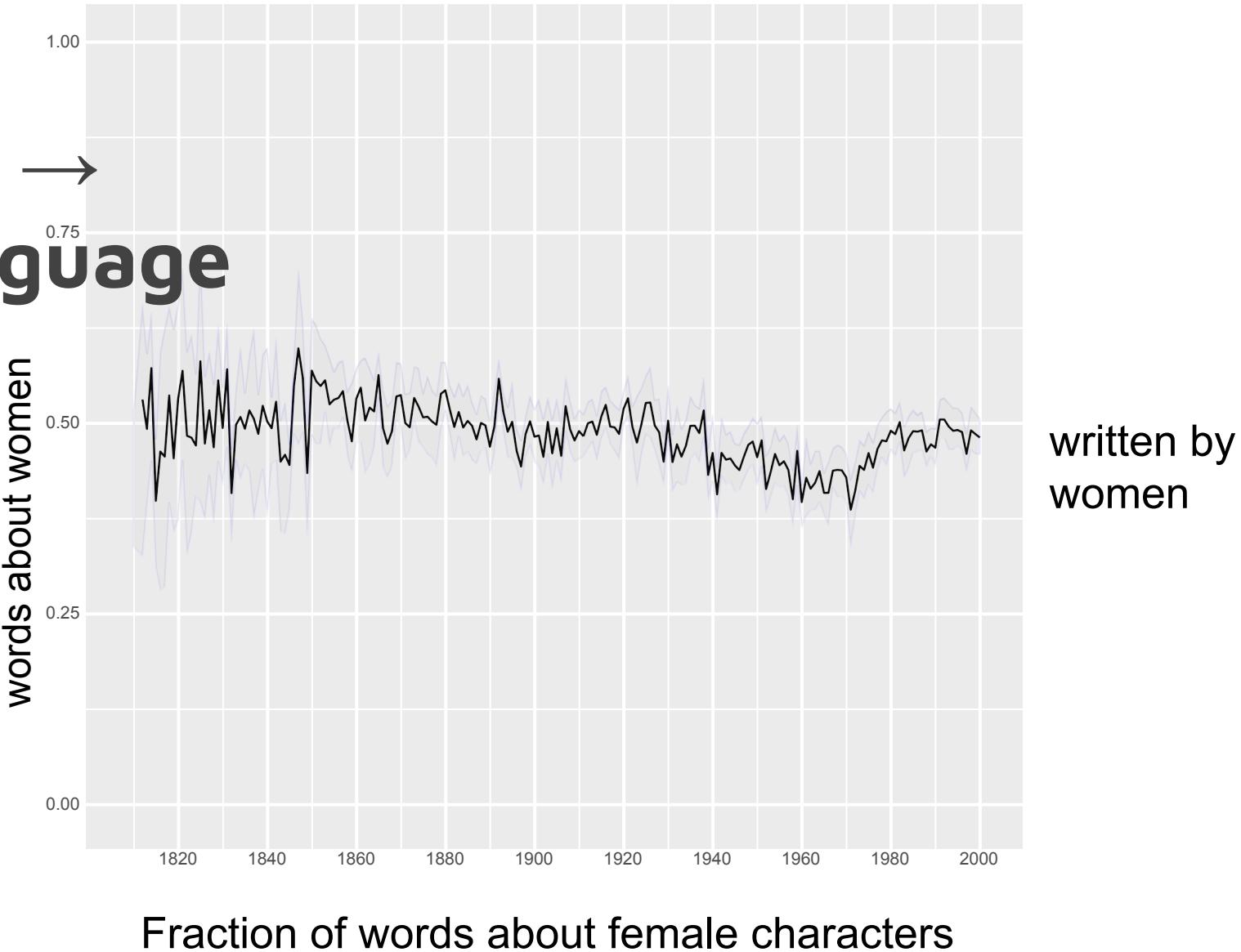
Wing and Baldridge (2011), "Simple supervised document geolocation with
geodesic grids" (ACL)



Digital books → gendered language

Input: Fictional
texts

Output: Proportion
words about
females



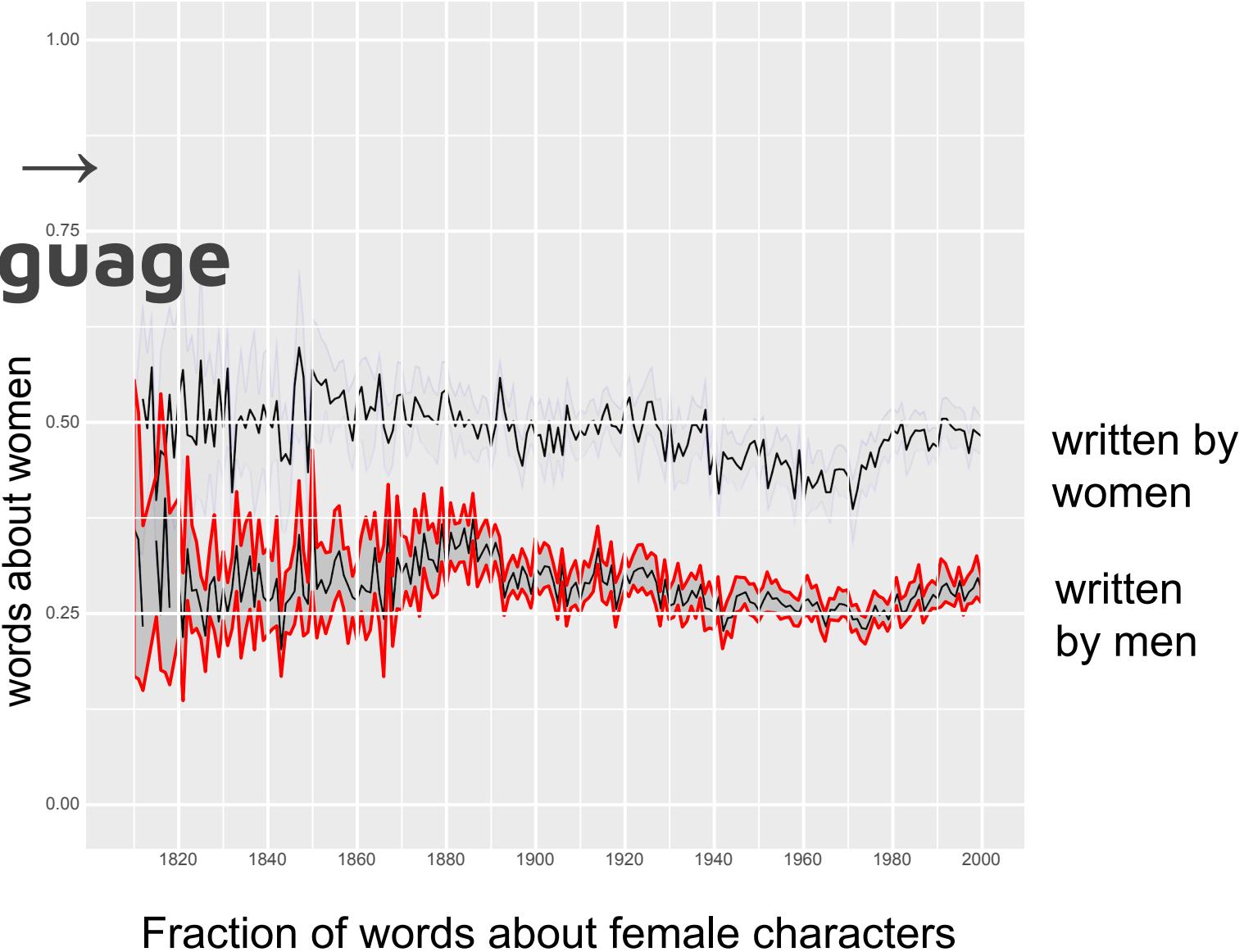
Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," (*Cultural Analytics*)



Digital books → gendered language

Input: Fictional
texts

Output: Proportion
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

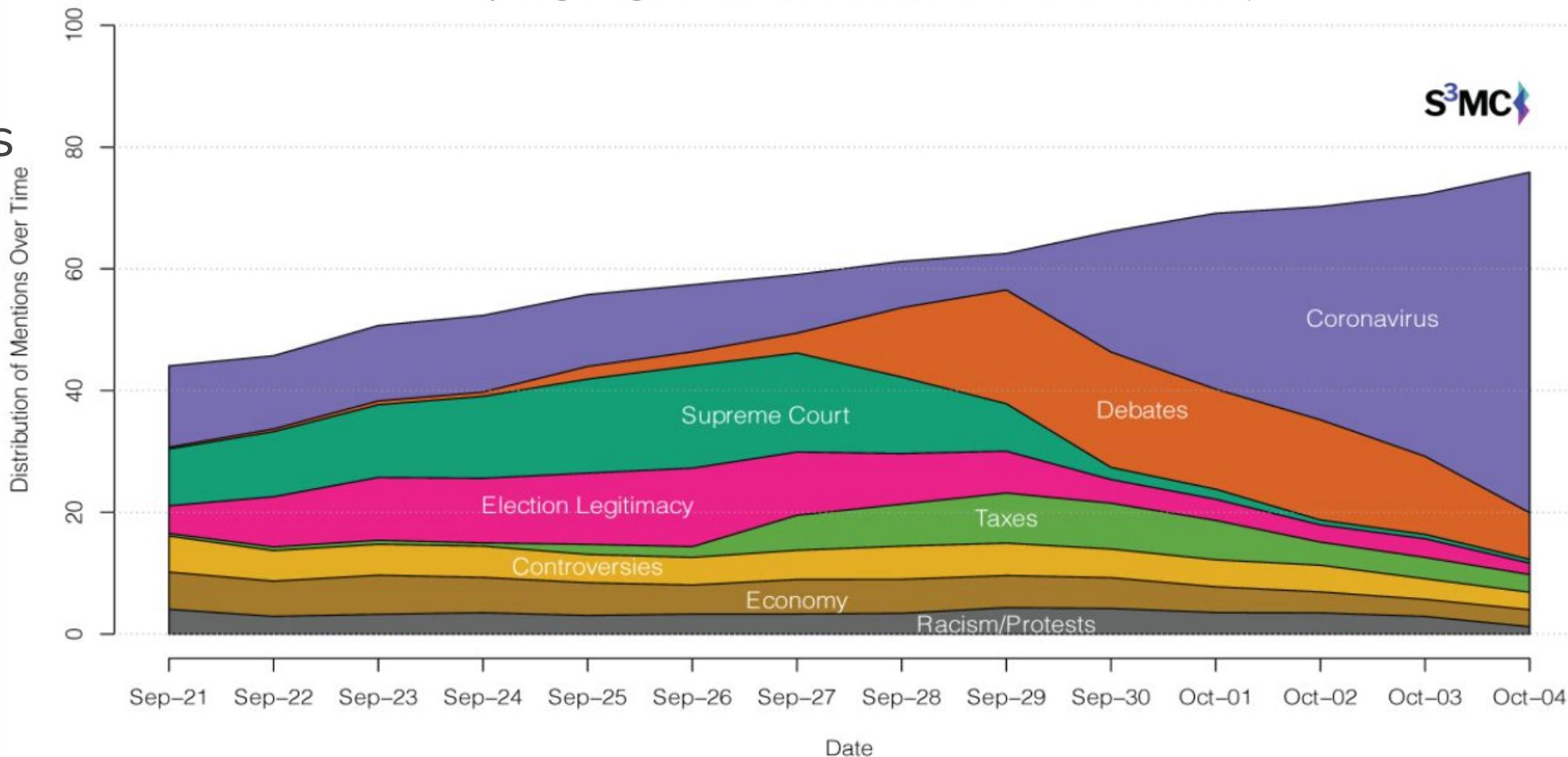
Trump's Most Frequently Mentioned Topics

7-day rolling averages on what Americans have seen, read or heard about Trump

Input: Speeches

Output: Topics
recalled by
public

S³MC

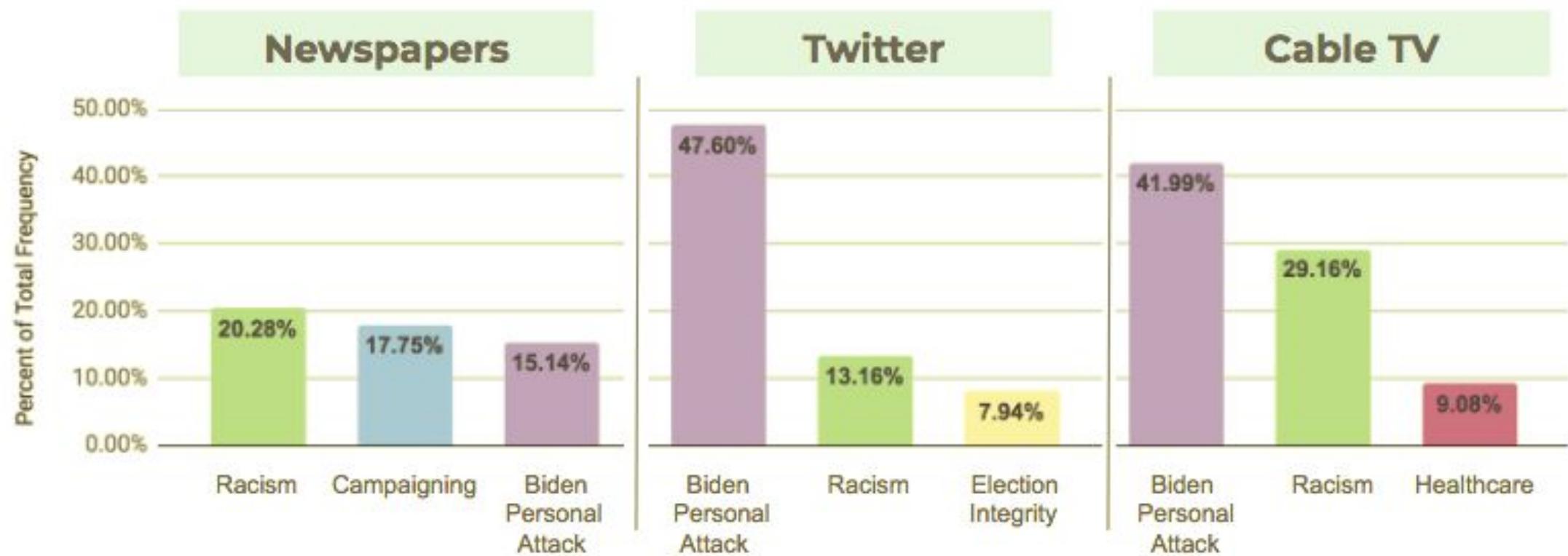


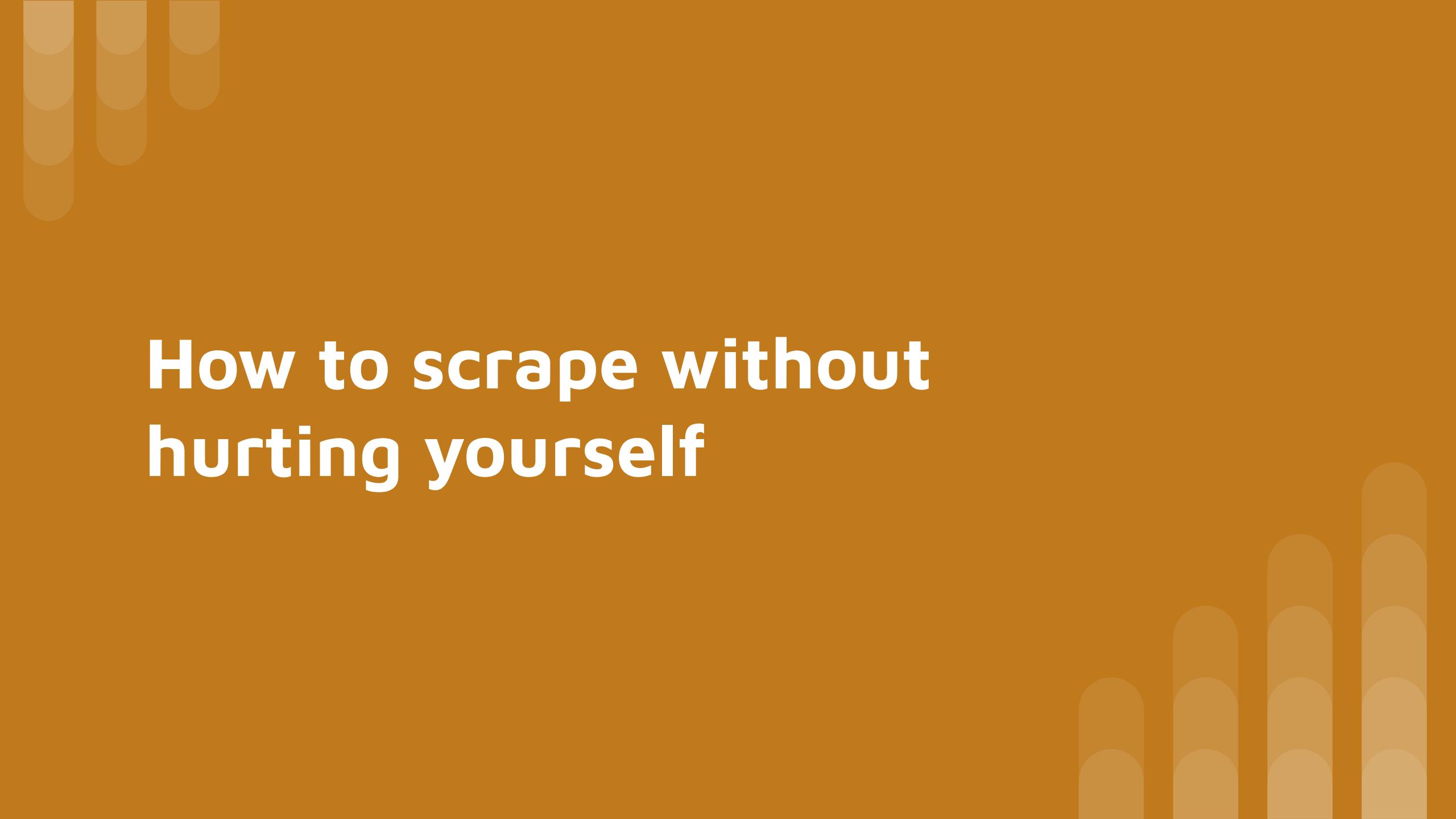


Digital media → spread of misinformation

Input: Articles, Tweets, segments

Output: Misinformation themes pushed by presidential candidates





**How to scrape without
hurting yourself**



Web-scraping decision hierarchy

Existing data

No scraping

A-Z Databases

Find the best library databases for your research.

[All Subjects](#)[All Database Types](#)[All Vendors / Providers](#)[Search for Database](#)[Go](#)

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 #

927 Databases found

A

[Abbreviations Online](#)

An electronic dictionary of medieval Latin abbreviations, a database designed for use in both learning and teaching medieval Latin paleography, can be used as a reference and research tool.



[ABELL \(Annual Bibliography of English Language and Literature\)](#)

Alternative Name(s) & Keywords: Literature Online

ABELL is part of the Literature Online (LION) database and covers all aspects and periods of English-language literature, linguistics, folklore, drama, and cultural studies. Indexes literary criticism, book reviews, periodical articles, critical editions of literary works, collections of essays, and doctoral dissertations published worldwide. Includes material in languages other than English. ABELL is linked to 325+ full text journals in the LION database.



New / Trial Databases

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

[Black Life in America](#) Trial

Trial expires April 30, 2021.

[more...](#)



[Independent and Revolutionary Mexican Newspapers](#) New

The open access collection, Independent and Revolutionary Mexican Newspapers, traces the evolution of Mexico during a pivotal period in Mexico's history. Comprising over 1,000 titles from Mexico's pre-independence, independence and revolutionary periods (1807-1929), the newspapers in this collection provide rare documentation of the dramatic events of this era.



English Reference Corpora

British National Corpus (bnc)	locked	eng-uk / 111,246,947	T / 4,054	📄
The Brown Corpus (brown)	green lock	eng-us / 1,172,053	T / 500	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 122,445,373	T / 26,203	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 108,622,152	T / 82,021	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 126,484,317	T / 25,993	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 129,459,423	T / 86,225	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 126,231,664	T / 90,243	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 130,081,788	T / 44,803	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 153	T / 23,975	📄
COCA - Corpus of Contemporary American English	green lock	eng-us / 139	T / 88,989	📄
Frown Corpus J (frown_j)	locked	eng-uk / 185,276	T / 80	📄
Georgetown University Multilayer Corpus	green lock	eng / 109,140	T / 126	📄
The Georgetown Multilayer Corpus 6	green lock	eng / 129,660	T / 148	📄
Penn Treebank CQPfied (auto tagged)	green lock	eng-us / 1,678,233	T / 2	📄
Penn Treebank CQPfied (gold tagged)	green lock	eng-us / 2,467,944	T / 3,153	📄

Spoken Language Corpora

HCRC Map Task Corpus (hcrcmap_2)	green lock	eng-sct / 188,898	T / 128	📄
Switchboard Corpus (switchboard)	green lock	eng-us / 1,159,308	T / 649	📄
TED talks English (ted)	green lock	eng / 5,159,586	T / 2,085	📄

Non-English Reference Corpora

Lancaster Corpus of Mandarin Chinese (lcm)	green lock	zho / 1,002,340	T / 500	📄
--	------------	-----------------	---------	---

Literary Corpora

Jane Austen Corpus (austen)	green lock	eng-uk / 423,669	T / 3	📄
Charles Dickens Corpus (dickens)	green lock	eng-uk / 3,407,085	T / 14	📄
Don Quijote (don_quijote_spa)	green lock	spa / 429,855	T / 1	📄
Tom Sawyer (tom_sawyer_eng)	green lock	eng-us / 86,747	T / 1	📄

Political Corpora

Bush and Kerry Presidential Debate (bush_kerry)	green lock	eng-us / 48,230	T / 2	📄
Inaugural Address Corpus (inaugural)	green lock	eng-us / 144,980	T / 56	📄
The Mueller Report Corpus (mueller)	green lock	eng-us / 228,799	T / 2	📄
German Bundestag Protocols (parlament)	green lock	deu / 36,723,139	T / 837	📄
State of the Union Corpus (1790-2017) (so)	green lock	eng-us / 1,937,728	T / 231	📄
State of the Union Corpus (1790-2020) (so)	green lock	eng-us / 2,081,385	T / 234	📄

Web Corpora

DECOW - Corpora from the Web - German	green lock	deu / 300,002,861	T / 198,608	📄
DECOW - Corpora from the Web - German	green lock	deu / 300,003,990	T / 231,532	📄
DECOW - Corpora from the Web - German	green lock	deu / 300,008,102	T / 325,463	📄
DE Web as Corpus - Part 1 (dewac01)	green lock	deu / 268,848,124	T / 289,824	📄
DE Web as Corpus - Part 2 (dewac02)	green lock	deu / 268,848,124	T / 288,223	📄
DE Web as Corpus - Part 3 (dewac03)	green lock	deu / 268,884,554	T / 290,941	📄
DE Web as Corpus - Part 4 (dewac04)	green lock	deu / 268,931,207	T / 289,139	📄
DE Web as Corpus - Part 5 (dewac05)	green lock	deu / 268,908,956	T / 288,386	📄
DE Web as Corpus - Part 6 (dewac06)	green lock	deu / 282,733,943	T / 305,382	📄
ENCOW2016 - Corpora from the Web (eng)	green lock	eng / 300,004,068	T / 225,073	📄
ENCOW2016 - Corpora from the Web (eng)	green lock	eng / 300,000,358	T / 222,658	📄
ENCOW2016 - Corpora from the Web (eng)	green lock	eng / 214,108,271	T / 163,651	📄
ENCOW2016 - Corpora from the Web (eng)	green lock	eng / 288,386	T / 288,386	📄
ENCOW2016 - Corpora from the Web (eng)	green lock	eng / 161,936	T / 161,936	📄
Corpus of Web-Based Global English - English	green lock	eng-us / 142,425,833	T / 106,385	📄
Corpus of Web-Based Global English - English	green lock	eng-us / 272,905,012	T / 168,771	📄
Russian Internet Corpus Sampler (rus)	green lock	rus / 5,231,112	T / 843	📄
Stanford Sentiment Analyzed Twitter Corpus (eng)	green lock	eng / 24,473,485	T / 2	📄

Newspaper Corpora

Arabic Treebank CQPfied (arabictb)	locked	ara / 168,722	T / 734	📄
Chinese Treebank 9.0 (chinese_treebank9)	green lock	zho / 2,080,333	T / 3,726	📄
New York Times - Arts Subcorpus (nyt_arts)	green lock	eng-us / 101,087,365	T / 118,433	📄
Slate Magazine Corpus (slate_alt)	green lock	eng-us / 4,929,752	T / 4,531	📄

Learner Corpora and Native Contradiction Corpora

Arabic Learner Corpus (arablearn)	locked	ara / 444,321	T / 1,585	📄
Hong Kong City University Corpus of English (eng-L2)	green lock	eng-L2 / 7,720,912	T / 11,170	📄
The Gyeon Korean EFL Learner Corpus (gak)	green lock	eng-L2 / 1,824,373	T / 16,111	📄
International Corpus of Learner English (icle)	green lock	eng-L2 / 2,808,577	T / 3,701	📄
International Corpus Network of Asian Learner English (icnae)	green lock	eng-L2 / 1,963,147	T / 9,836	📄
Louvain Corpus of Native English Student English (eng-uk/us)	green lock	eng-uk/us / 346,906	T / 388	📄
Spanish Learner Language Oral Corpora (SP)	green lock	spa-L2 / 372,567	T / 561	📄

Bible Corpora

The King James Bible Corpus (biblekjv)	green lock	eng-uk / 915,179	T / 66	📄
The Luther Bible Corpus (bibleluther)	green lock	deu / 812,322	T / 66	📄

World English Corpora

Corpus of Web-Based Global English - Hong Kong City University	green lock	eng-hk / 42,979,217	T / 43,936	📄
ICE Jamaica (ice_ja)	locked	eng-ja / 1,156,149	T / 500	📄
ICE Singapore (ice_sg)	locked	eng-sg / 1,163,008	T / 500	📄
National University of Singapore SMS Corpus (eng-sg)	green lock	eng-sg / 150,397	T / 10,117	📄

Historical Corpora

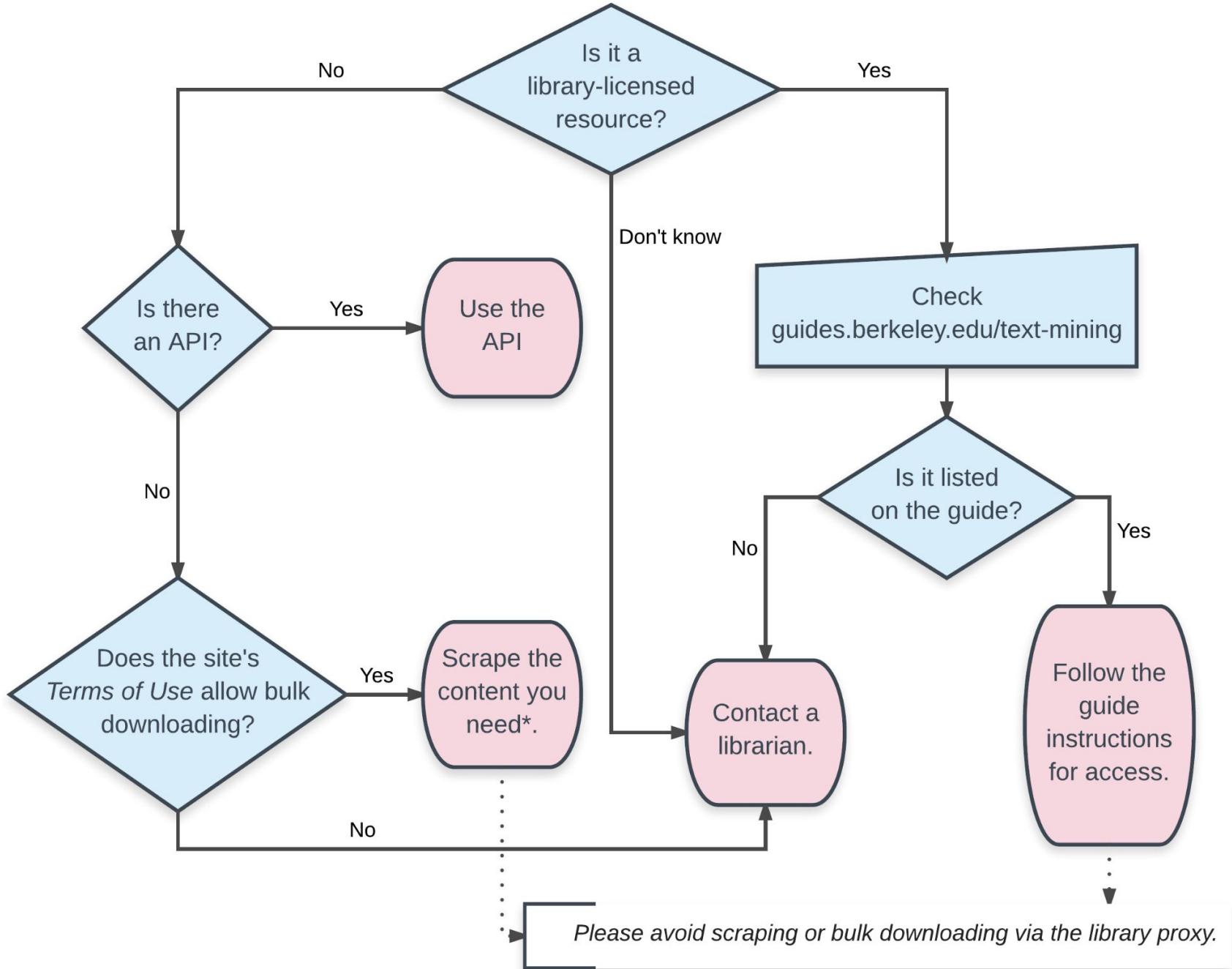
Ancient Chinese Corpus - Zuozhuan (acc)	green lock	zho-lzh / 194,258	T / 2	📄
Corpus of Historical American English (coh)	green lock	eng-us / 448,200,483	T / 116,773	📄
Georgetown University Historical Reddit Corpus (eng)	green lock	eng / 43,858,955	T / 557,579	📄
Penn Parsed Corpus of Middle English v2 (eng-enm)	green lock	eng-enm / 1,351,054	T / 56	📄
Sheffield Corpus of Chinese - Historical Chinese (zho)	green lock	zho / 14,282	T / 3	📄

Coptic Scriptorium Corpora

Besa Letters Corpus (besa)	green lock	cop / 1,907	T / 2	📄
----------------------------	------------	-------------	-------	---

Uncategorised

Chatino Zapotec Corpus (chazap)	green lock	zap / 721,976	T / 290,292	📄
Gradable Modal Expressions (for CQP, building)	green lock	eng / 1,000,000	T / 1,000,000	📄





Web-scraping decision hierarchy

Existing data

Web APIs

No scraping

Easy scraping



Web-scraping decision hierarchy

Existing data

Web APIs

Structured
HTML

No scraping

Easy scraping

Structured
scraping



How the web works





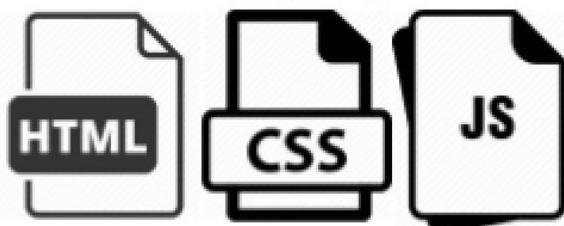
How the web works



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



Here you go, parse away (RESPONSE):



[main](#) [branch](#) [tags](#)[Go to file](#)[Add file](#)[Code](#)

	jhaber-zz moved solutions to separate notebooks
	day-1 moved solutions to separate notebooks
	day-2 added wget options, added BS text clea
	extra moved solutions to separate notebooks
	LICENSE Initial commit
	README.md move O'Reilly books to own section of R
	requirements.txt added wget options, added BS text clea

[README.md](#)[launch binder](#)

Web-Crawling: A Practical Introduction in Python

A workshop with the [Massive Data Institute](#), Georgetown University

Overview

When it comes to data collection, web-crawling (i.e., web-scraping, screen-scraping) is a common approach in our increasingly digital era--and a common stumbling block. With such a wide range of tools and languages available (Selenium, Requests, and HTML, to name just a few), developing and implementing a web-crawling pipeline is often a frustrating experience for researchers--especially those without a computer science background.

Whatever your background, this workshop will give you the foundation to use web-crawling in your research. We will tackle common problems including collecting web addresses/URLs (by automated Google search), downloading website copies (with wget), non-scalable website scraping (with requests), and scalable crawling of text (with scrapy).

[ff2ae80 24 minutes ago](#) [21 commits](#)

- [Back](#) [Alt+Left Arrow](#)
- [Forward](#) [Alt+Right Arrow](#)
- [Reload](#) [Ctrl+R](#)
- [Save as...](#) [Ctrl+S](#)
- [Print...](#) [Ctrl+P](#)
- [Cast...](#)
- [Send to your devices](#)
- [Create QR code for this page](#)
- [Translate to English](#)
- [Adobe Acrobat](#)
- [Black Menu for Google™](#)
- [Zotero Connector](#)
- [View page source](#) [Ctrl+U](#)
- [Inspect](#) [Ctrl+Shift+I](#)

About

An introduction to web-crawling/scraping for beginners with some Python know-how. Created for GU's Massive Data Institute in spring 2021 by Jaren Haber, PhD

[Readme](#)[GPL-3.0 License](#)

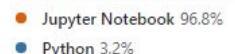
Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Languages



github.com/jhaber-zz/web-crawling-intro-2021

jhaber-zz / web-crawling-intro-2021

Code Issues Pull requests Actions Projects Wiki Security Insights ...

main Go to file Add file Code

jhaber-zz moved solutions to separate notebooks 25 minutes ago 21

day-1 moved solutions to separate notebooks 25 minutes ago

day-2 added wget options, added BS text cleaning 5 days ago

extra moved solutions to separate notebooks 25 minutes ago

LICENSE Initial commit 21 days ago

README.md move O'Reilly books to own section of README 21 hours ago

requirements.txt added wget options, added BS text cleaning 5 days ago

README.md

h1 600.5 × 90.59

Web-Crawling: A Practical Introduction in Python

A workshop with the **Massive Data Institute**, Georgetown University

Overview

When it comes to data collection, web-crawling (i.e., web-scraping, screen-scraping) is a common approach in our increasingly digital era--and a common stumbling block. With such a wide range of tools and languages available (Selenium, Requests, and HTML, to name just a few), developing and implementing a web-crawling pipeline is often a frustrating experience for researchers--especially those without a

Elements Console Sources Network Performance Memory Application Security Lighthouse

Styles Computed Layout Event Listeners

Filter :hov .cls +

```
<div class="Box md js-code-block-container Box--responsive">
  <div id="readme" class="Box md js-code-block-container Box--responsive">
    <div class="d-flex js-sticky js-position-sticky top-0 border-top-0 border-bottom p-2 flex-items-center flex-justify-between color-bg-primary rounded-top-2" style="position: sticky; z-index: 90; top: 0px !important;" data-original-top="auto">...</div> <flex>
      <div class="Popover anim-scale-in js-tagsearch-popover" hidden data-tagsearch-url="/jhaber-zz/web-crawling-intro-2021/find-definition" data-tagsearch-ref="main" data-tagsearch-path="README.md" data-tagsearch-lang="Markdown" data-hydro-click="{"event_type":"code_navigation.click_on_symbol","payload":{"action":"click_on_symbol","repository_id":355332923,"ref":"main","language":"Markdown","originating_url":"https://github.com/jhaber-zz/web-crawling-intro-2021","user_id":4977437}}" data-hydro-click-hmac="0e7810bcf9e7114e511649c9ac7d19749ee23e3ddae10b8df3899b9f9d7e1">...</div>
    <div data-target="readme-toc.content" class="Box-body px-5 pb-5">
      <article class="markdown-body entry-content container-lg" itemprop="text">
        ::before
        <p>...</p>
        <h1> == $0
          <a id="user-content-web-crawling-a-practical-introduction-in-python" class="anchor" aria-hidden="true" href="#web-crawling-a-practical-introduction-in-python">...</a>
          "Web-Crawling: A Practical Introduction in Python"
        </h1>
        <h2>...</h2>
        <h2>...</h2>
        <p>...</p>
        <p>...</p>
        <h2>...</h2>
        <ul>...</ul>
        <h2>...</h2>
        <p>...</p>
        <h2>...</h2>
        <p>...</p>
        <p>...</p>
        <pre>...</pre>
        <h2>...</h2>
        <h3>...</h3>
        <ul>...</ul>
        ... container.Box--responsive div.Box-body.px-5.pb-5 article.markdown-body.entry-content.container-lg h1 ...
    </div>
  </div>

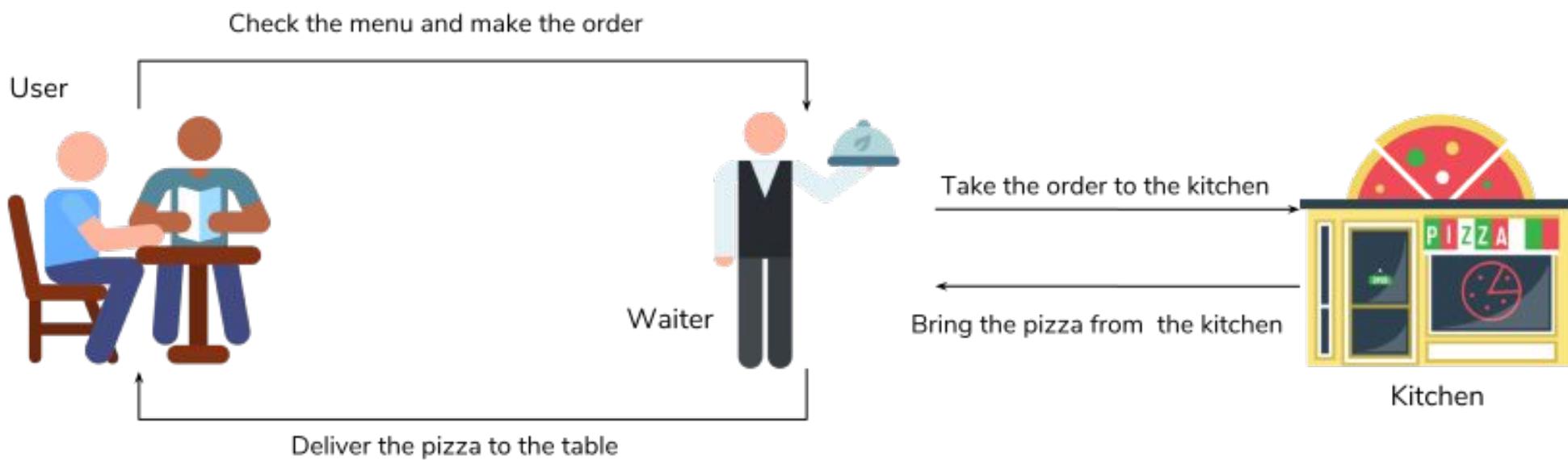
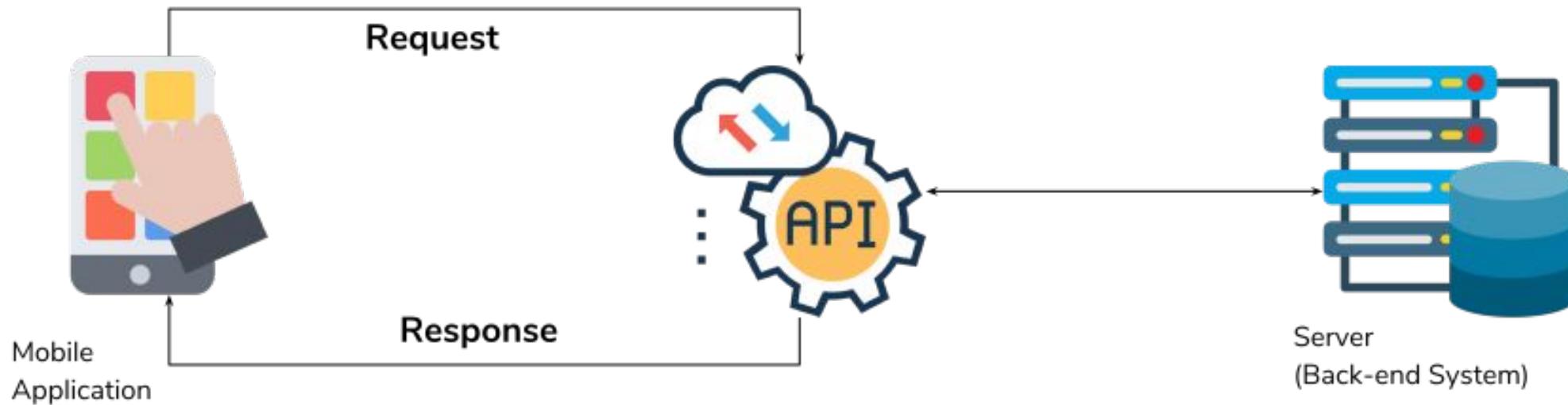
```

Console What's New

Highlights from the Chrome 90 update

- New CSS Flexbox debugging tools
- Debug and inspect CSS Flexbox with the new CSS Flexbox debugging tools.
- New Core Web Vitals overlay
- Visualize page performance with the new Core Web Vitals overlay.
- Report Trusted Web Activity issues

new





Web-scraping decision hierarchy

Existing data

Web APIs

Structured
HTML

Wild West

No scraping

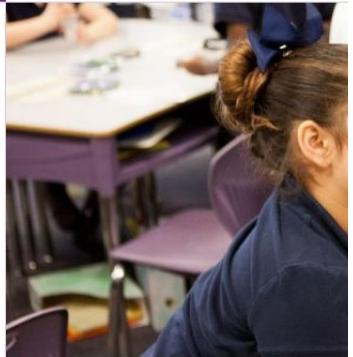
Easy scraping

Structured
scraping

- Directory
- Scrape URLs
- Scrape HTML
- Parse HTML
- Store output
- Filter content



A *Wild West* deep dive: Crawling charter school websites



Aspire Port

Welcome to the Aspire Po

Grades: K – 5

Sponsoring District: Stoc

CDS/Charter Numbers: CI
#1553

Enrollment: 410

Ethnicity:

Hispanic: 52.4%

Will Carleton Academy

[About](#) [Admissions](#) [Academics](#) [Athletics](#) [Student](#) [Parents](#) [Calendar](#) [Faculty](#) [Contact](#)
[Life](#)


Will Carleton Academy is a K-12 tuition-free public charter school in Hillsdale, Michigan founded in 1998. Will Carleton Academy devotes itself to serving the community as a charter school where parents can choose a traditional curriculum and educational atmosphere that is committed to both the intellectual development and the character formation of the students.

Our Mission Is



To Educate

- Back-to-basics curriculum
 - K-8th-Core Knowledge Sequence
 - 9th-12th: Michigan Merit/Colllege Preparatory Curriculum
 - Small classroom sizes at



To Enrich

- Advanced Placement class offerings on demand
- Dual Enrollment, Early/Middle College, and Vocational-Technical Programs
- Extensive extracurricular offerings at all grade



To Enlighten

- Orderly, disciplined environment
- Focus on the whole child, with the goal of helping to mold outstanding citizens
- K-12th Art and Music curriculum
- 4th-12th French



New School Hours for the 2018 - 2019 School Year

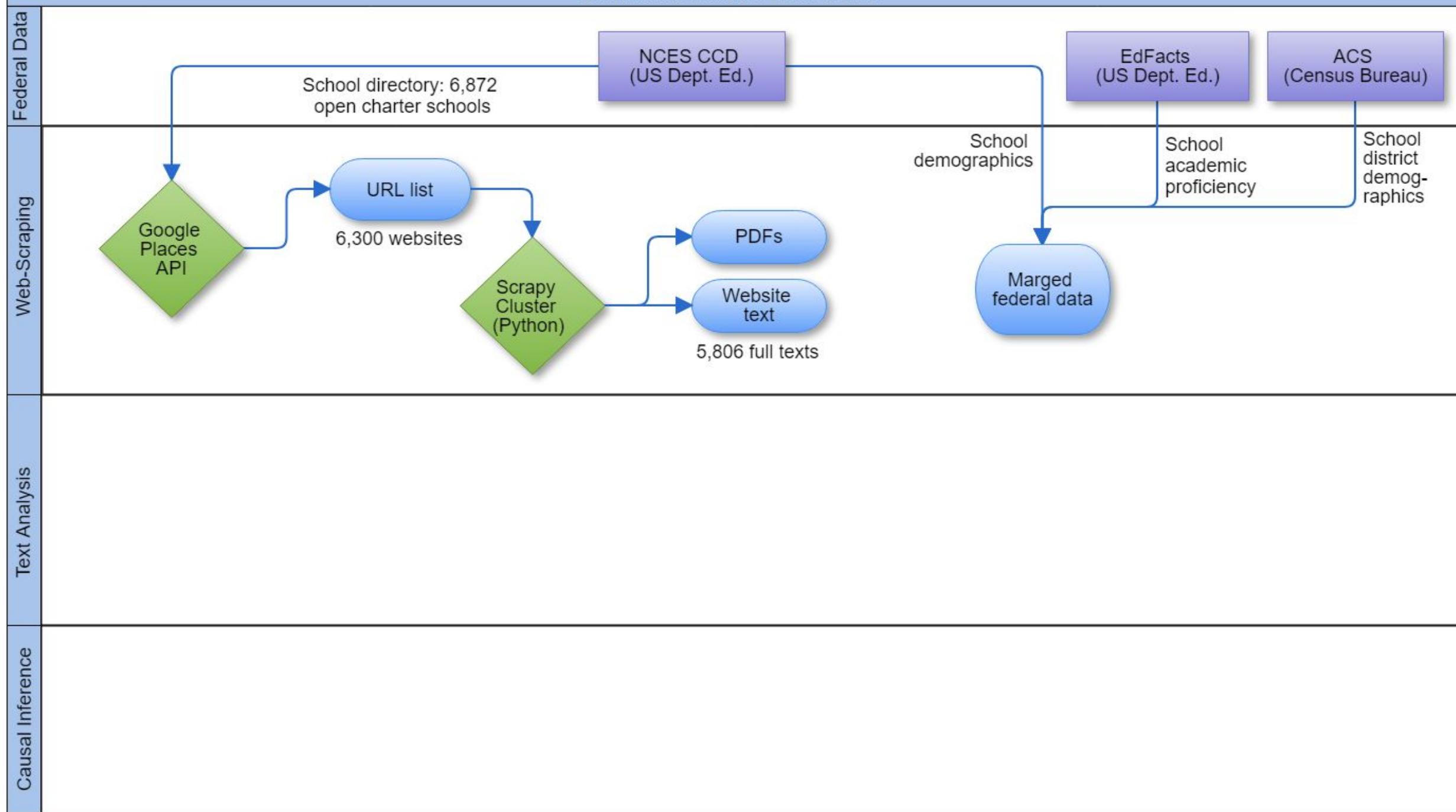
8:15 AM - 3:45 PM

Welcome to AIM College and Career Preparatory Academy

AIM is a GISD charter school that provides students the opportunity to accelerate their own learning. We are a 6th – 12th grade campus with a variety of programs. Middle school students have opportunities to test and promote to the next grade level. High school students attend face-to-face, hybridized and online courses. We also offer the HSEP (High School Equivalency Program) for qualifying students.



DATA COLLECTION WORKFLOW



charter school



Rating ▾ Hours ▾ More filters

Holyoke Community Charter School

3.8 ★★★★★ (17)

Charter school · 2200 Northampton St
Closing soon: 3:30 PM · (413) 533-0111



Website



Directions

Libertas Academy Charter School

3.7 ★★★★★ (18)

Charter school · 146 Chestnut St
Closing soon: 3:30 PM · (413) 342-5510



Website



Directions

Springfield Prep Charter School

4.0 ★★★★★ (10)

Charter school · 594 Converse St
Closing soon: 4:00 PM · (413) 231-2722



Website



Directions

Veritas Prep Charter School

3.7 ★★★★★ (28)

Charter school · 370 Pine St
Open until 4:45 PM · (413) 539-0055



Website



Directions

Hampden Charter School of Science - West

School · 511 Main St
Closing soon: 3:30 PM · (413) 278-5450



Website



Directions

Baystate Academy Charter Public School School

3.5 ★★★★★ (24)

Charter school · 2001 Roosevelt Ave



Website

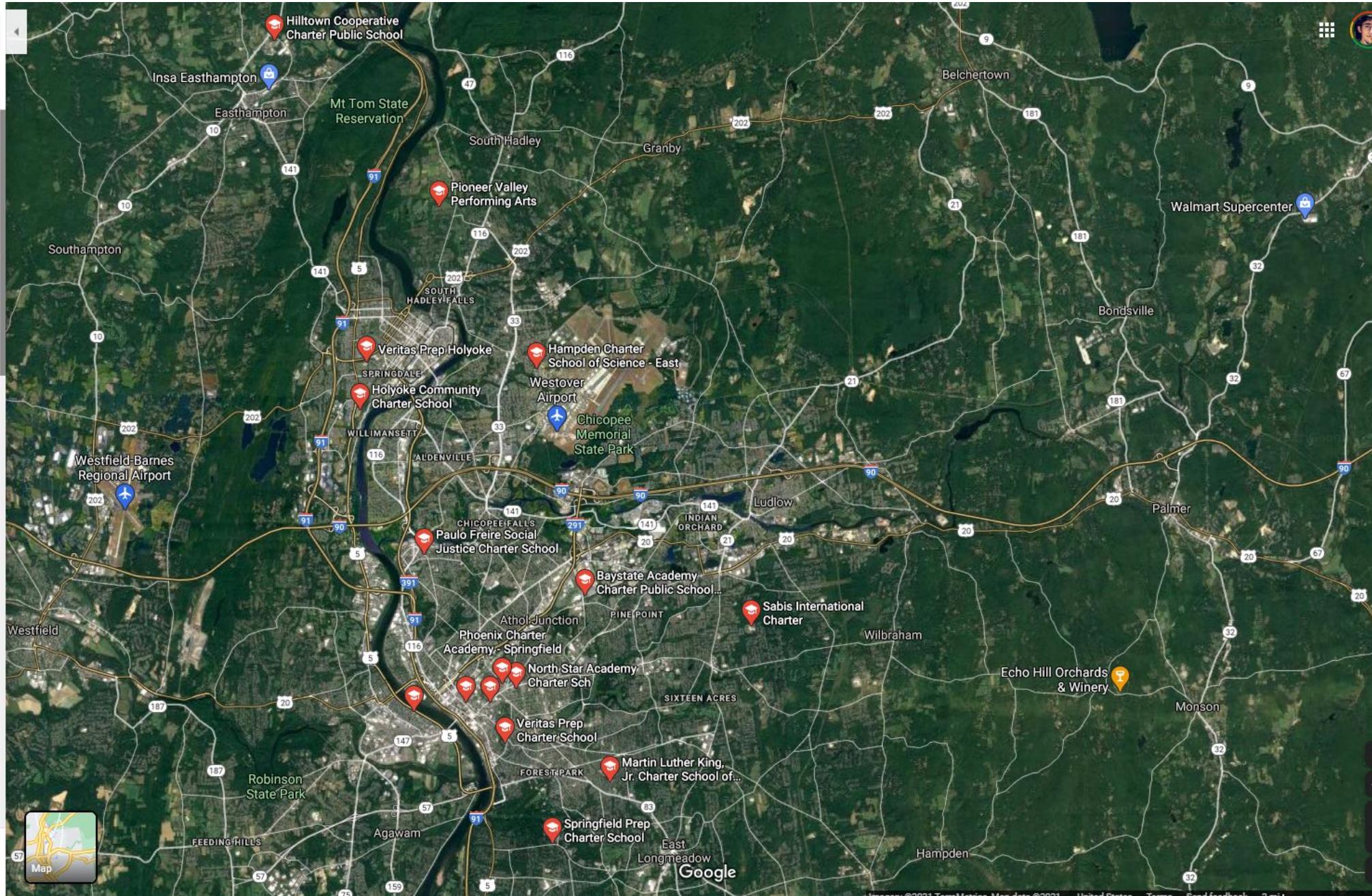


Directions

Showing results 1 - 16



Update results when map moves



Imagery ©2021 TerraMetrics, Map data ©2021

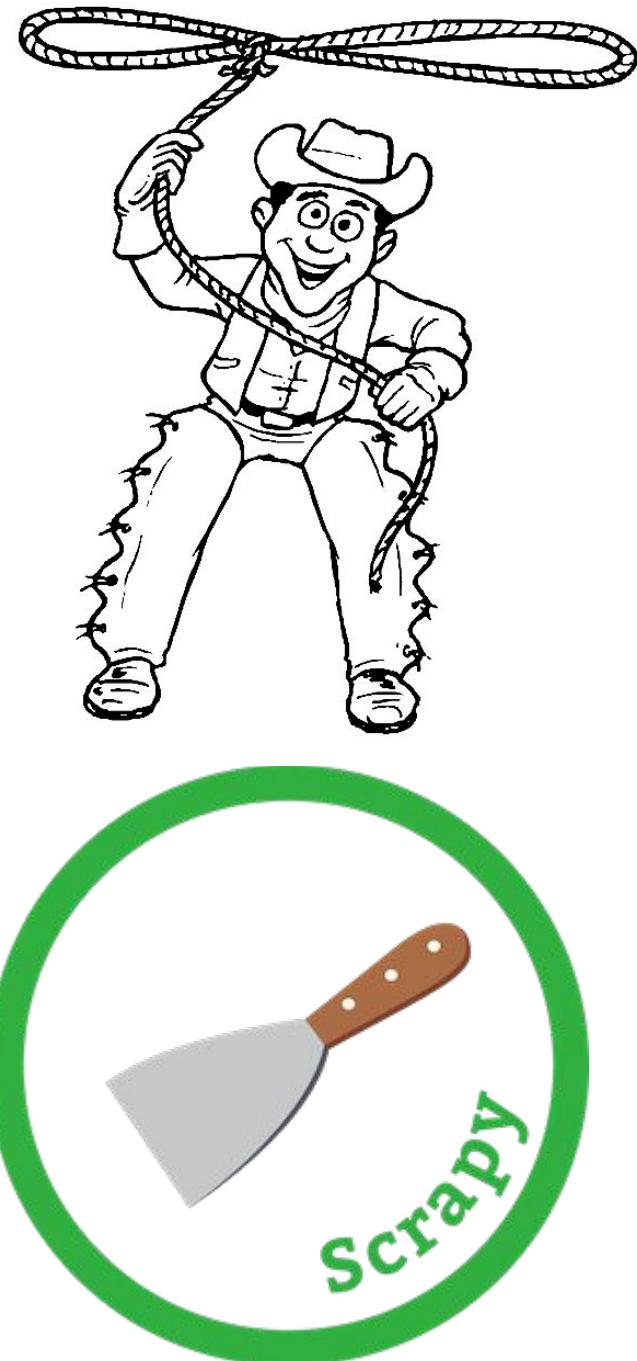
United States

Terms

Send feedback

2 mi

Scraping HTML





Con

```
def parsefile_by_ta
    """Cleans HTML
    with a random s

    random_string =
    soup = BeautifulSoup(
        inline_tags = [
            "strong", "samp", '"

    [s.extract() for
        non-visible tags
        for it in inlin
            [s.extract()
            visible_text =
            string, eliminate '"

    # Split text into
    visible_text =
    visible_text.split(
        visible_text =
        elements
    return(visible_
```



paragraph tags

delimiter

"em", "kbd",
ags to eliminate

nt]')]]) # Remove

with random

eadable text:

te empty



Mission and Vision

Stony Point Academy is committed to provide a rigorous preparatory program that ensures all students are ready for successful post-secondary pathway. This includes an academically rich curriculum with rigorous content. It includes an expectation that students will apply their knowledge through higher-order skills, and will develop the habits of mind and character traits known to support personal standards within a student-centered environment that emphasizes the Common Core Anchor Standards for College and Career Readiness (CCR) across all content areas at every grade level. It is with this preparation that students will become contributing members of the local and global communities.

Stony Point Academy





$f=0$ crawl approach: wget

```
wget {URL}
```

```
wget --no-parent --level 0 --no-check-certificate --recursive --adjust-extension  
--convert-links --page-requisites --wait=10 --random-wait --execute robots=off  
--follow-ftp --secure-protocol=auto --retry-connrefused --tries=12 --no-remove-listing  
--local-encoding=UTF-8 --no-cookies --default-page=default --server-response  
--trust-server-names --header="Accept:text/html" --user-agent=Mozilla  
--warc-file=bridge_warc --warc-cdx --warc-max-size=0.5G URL}
```

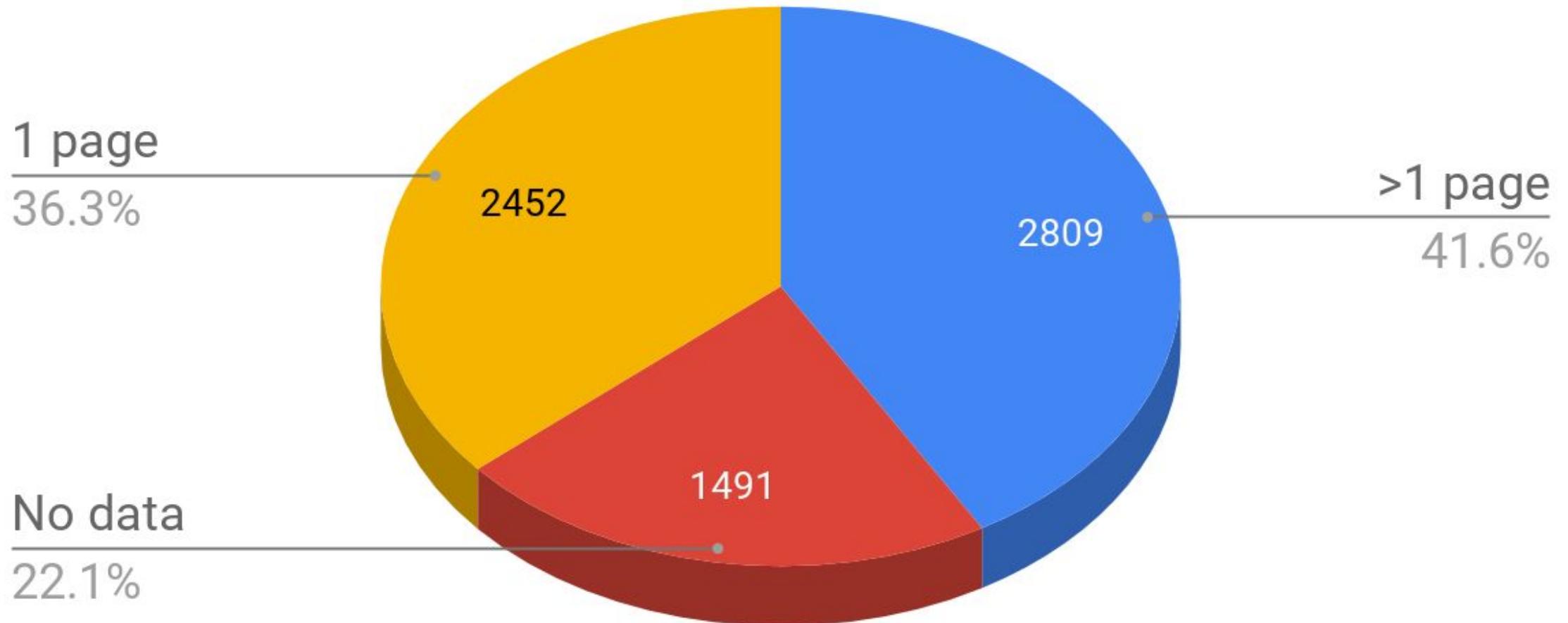
--recursive	Recursively download files and follow links
--no-parent	Follow links, but not beyond the last parent directory
--page-requisites	Grab all of the linked resources necessary to render the page (images, CSS, JS, etc.)
--adjust-extension	Append .html to the files when appropriate
--convert-links	Turn links into local links as appropriate
--execute robots=off	Turn off wget's automatic robots.txt checking
--user-agent	Mask the user agent and show wget like a browser



wget output

... xp=s\x92éëÊ00ÿ\r[ÖMission and VisionÈèüû]Dg"\x98\x95\xad-\x01ê\x1bÿìö?YüoeF2æAv\x150TS\x8bS\x82 ÄÝ¤\x98rY-#cÌ\x16µää\x8d8@^-L\x01û\x80sÍ\x89xp=s\x92éëÊ00ÿ\r[ÖÈèüû]Dg "\x98\x95\xad-\x01ê\x1bÿìö?Yüo ... x80sÍ\x89xp=s\x92éëÊ00ÿ\r[ÖBell ScheduleÈèüû]Dg "\x98\x95\xad-\x01ê\x1bÿìö?YüoeF2æAv\x150TS\x8bS\x82ÄÝ¤\x98rY-#cÌ\x16µää\x8d8@^-L\x01û\x80sÍ\x89xp=s\x92éëÊ00ÿ\r[ÖÈèüû]Dg"\x98\x95\xad-\x01ê\x1bÿìö ... â\x8d8@^-L\x01û\x80sÍ\x89xp=s\x92éëÊ00ÿ\r[ÖCollege & Career ReadinessÈèüû]Dg"\x98\x95\xad-\x01ê\x1 ... x16µää\x8d8@^-L\x01û\x80sÍ\x89xp=s\x92éëÊ00ÿ\r[ÖStony Point Academy is committed to provide a rigorous preparatory program that ensures all students are ready for successful post-secondary pathway. This includes an academically rich curriculum with rigorous content. It includes an expectation that students will apply their knowledge through higher-order skills, and will develop the habits of mind and character traits known to support personal standards within a student-centered environment that emphasizes the Common Core Anchor Standards for College and Career Readiness (CCR) across all content areas at every grade level. It is with this preparation that students will become contributing members of the local and global communities.Èèüû]Dg"\x98\x95\xad-\x01ê\x1bÿìö?YüoeF2æAv\x150TS\x8bS\x82ÄÝ¤\x98rY-#cÌ\x16µää\x8d8 @^-L\x01û\x80sÍ\x89xp=s\x9 2éëÊ00ÿ\r[ÖÈèüû]Dg "\x98\x95\xad-\x01ê\x1bÿìö?YüoeF2æAv\x150TS\x8bS\x82

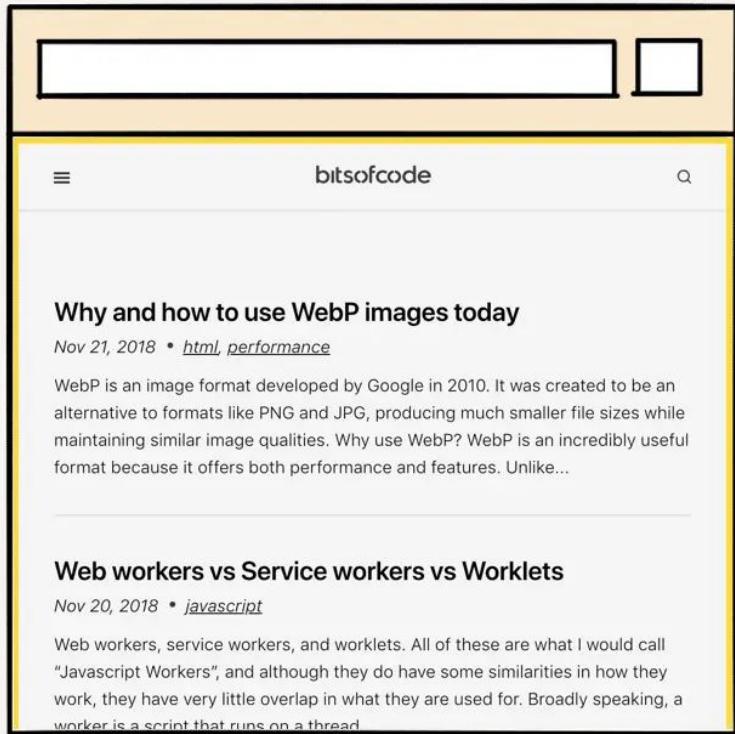
wget success rate



$f+1$ crawl approach: requests



$f+2$ crawl approach: Selenium



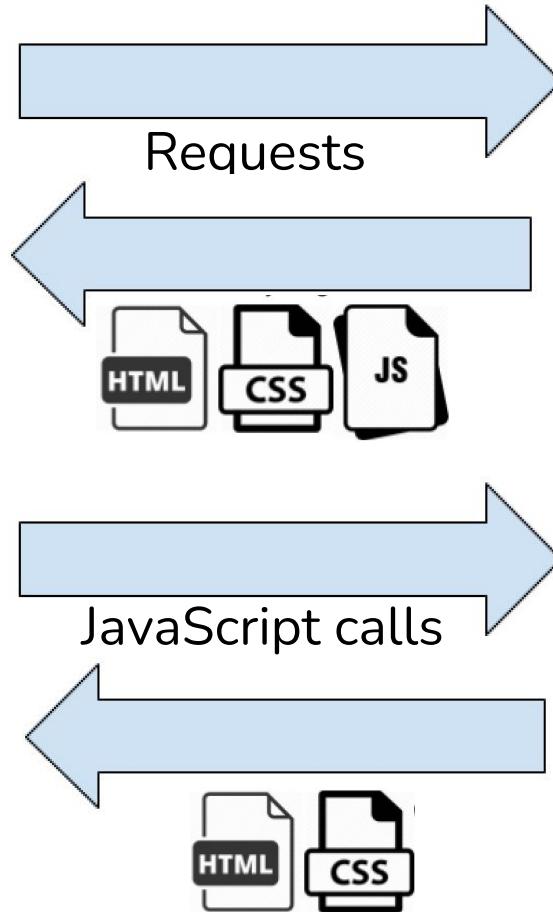
headful

```
<html lang="en">
  <head>
    <title> bitsofcode </title>
  </head>
  <body>
    <header>
      <h1> bitsofcode </h1>
    </header>
    <main>
```

headless



How interactive websites work





Selenium output

Home

Our School

...

Mission and Vision

Stony Point Academy is committed to provide a rigorous preparatory program that ensures all students are ready for successful post-secondary pathway. This includes an academically rich curriculum with rigorous content. It includes an expectation that students will apply their knowledge through higher-order skills, and will develop the habits of mind and character traits known to support personal standards within a student-centered environment that emphasizes the Common Core Anchor Standards for College and Career Readiness (CCR) across all content areas at every grade level. It is with this preparation that students will become contributing members of the local and global communities.

Stony Point Academy

Building Future Leaders

Select Language

Job Opportunities

Contact us

Website Issues



Limits of Selenium

```
[31mcrawler_10      | [0m 2018-04-05 21:42:10 [selenium.webdriver.remote.remote_connection]
DEBUG: GET http://172.18.0.2:4444/wd/hub/session/a9700948ceef7d04568452492f07e968/element/
0.786584017266313-80/attribute(href {"sessionId": "a9700948ceef7d04568452492f07e968", "name":
"href", "id": "0.786584017266313-80"}
[36;1mcrawler_2    | [0m TypeError: can't pickle thread.lock objects
[32mcrawler_5      | [0m 2018-04-05 21:41:37 [selenium.webdriver.remote.remote_connection]
DEBUG: GET http://172.18.0.2:4444/wd/hub/session/dcc13b350db90f2c33fccf3a6a15f850/element/
0.3070402879225935-111/attribute(href {"sessionId": "dcc13b350db90f2c33fccf3a6a15f850",
"name": "href", "id": "0.3070402879225935-111"}
[32;1mcrawler_9    | [0m 2018-04-05 21:42:29 [selenium.webdriver.remote.remote_connection]
DEBUG: Finished Request
[36mcrawler_6      | [0m {'appid': u'testapp',
[35;1mcrawler_1    | [0m 2018-04-05 21:42:44 [selenium.webdriver.remote.remote_connection]
DEBUG: GET http://172.18.0.2:4444/wd/hub/session/a0eaf0be3a4c7e187452951948b8a333/element/
0.1420010272851573-52/attribute(href {"sessionId": "a0eaf0be3a4c7e187452951948b8a333",
"name": "href", "id": "0.1420010272851573-52"}
```



$f+3$ crawl approach: Scrapy Cluster



Scrapy



Scrapy Cluster output

Home

Our School

...

Mission and Vision

Stony Point Academy is committed to provide a rigorous preparatory program that ensures all students are ready for successful post-secondary pathway. This includes an academically rich curriculum with rigorous content. It includes an expectation that students will apply their knowledge through higher-order skills, and will develop the habits of mind and character traits known to support personal standards within a student-centered environment that emphasizes the Common Core Anchor Standards for College and Career Readiness (CCR) across all content areas at every grade level. It is with this preparation that students will become contributing members of the local and global communities.

Stony Point Academy

Building Future Leaders

Job Opportunities

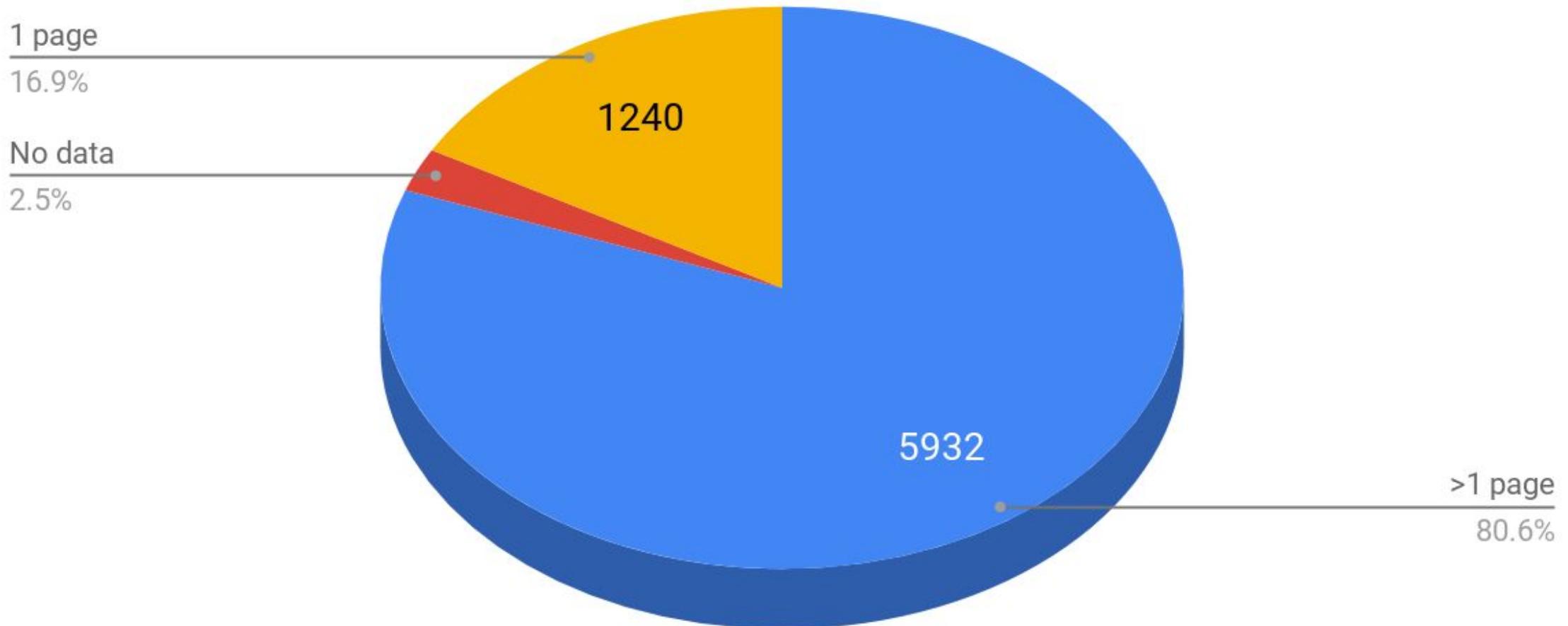
Contact us

Website Issues

Home

Our School

Scrapy Cluster success rate





Take-aways

- Web-crawling new corpora (still) the Wild West
- But don't "reinvent the flat tire"
- Cornerstone web-scraping tool (for Python): `scrapy`
 - Flexible, fast, strong community
 - Not a "black box" like Scrapy Cluster
- URL scraper, Scrapy spiders, etc.:
 - https://github.com/jhaber-zz/web_tools/



Ethical issues

- 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)*: Kind of
 - *LinkedIn v. HiQ*: Nope, web scraping a public right
- So is the whole web in the public domain?
 - It depends (contextual privacy; see *Bit by Bit*, Salganik 2017)

A photograph of a cowboy wearing a red shirt and a white cowboy hat, riding a dark horse. They are positioned on a rocky outcrop in a vast, arid landscape with large, layered rock formations under a clear sky.

Questions?

The background of the slide features a subtle, abstract design composed of several overlapping circles in shades of orange and light gray. These circles vary in size and are positioned in the upper right quadrant of the slide.

Time to get our hands dirty:

<http://bit.ly/crawl-python-2021>

Extra slides



Watch our video to learn about The Greene School

Innovative learning

The Greene School embraces an educational discipline of 'learn-by-doing.' Known as *EL Education (Expeditionary Learning)*, our approach is:

- Learning is active
- Learning is challenging
- Learning is meaningful
- Learning is public
- Learning is collaborative

Our values

Integrity
Compassion
Tolerance
Perseverance
Responsibility

Embracing nature

Our students grow academically and socially, become environmental stewards and actively manage the limited resources we share as a global community.

We provide direct experiences in nature along with the latest in environmental science and technology.



Local Urban Students Need More Support

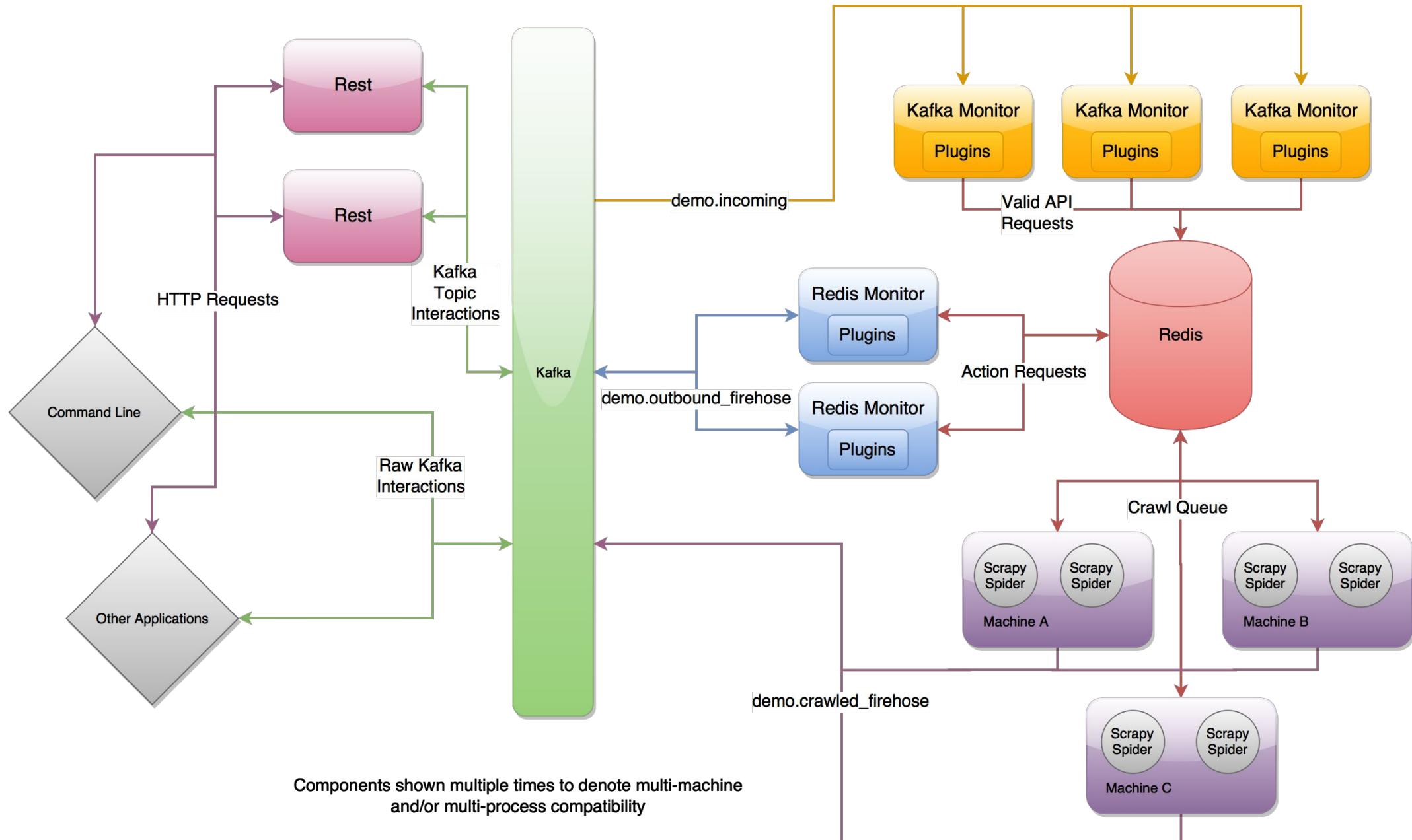
Consider the following statistics:

- Children entering kindergarten in Memphis scored in the 16th and 19th percentiles in math and language. This means that 80 percent of children entering kindergarten in the U.S. are better-prepared than children entering kindergarten in Memphis City Schools. Thirty-six out of 112 Memphis elementary schools reported scores of entering kindergarteners that are below 10 percent in at least one of these two subjects.
- 50% of urban children in Memphis will drop out of school before graduating from high school
- 96% of urban children in Memphis who do manage to graduate from high school will not meet college standards on all 4 subjects of the ACT test

All Children Can Learn in the Right Environment

The founders of CEG believe that all children can learn and be successful when given access to an achievement-oriented, structured, disciplined learning environment, beginning in Kindergarten and continuing through the eighth grade.

Scrapy Cluster Architecture Overview



ANALYTIC WORKFLOW

