



Web Scraping with Python

October 7, 2021

GW Libraries and Academic Innovation

Slides: go.gwu.edu/scrapingpython

Laura Wrubel, Software Development Librarian | lwrubel@gwu.edu






Agenda

- ◇ Intro to web scraping
- ◇ Hands-on web scraping using Google Colab and requests-html
- ◇ Legal and ethical considerations
- ◇ Best practices
- ◇ Other Python libraries
- ◇ Resources for learning and help





Ground rules

- ◇ We're all learners at different points on our journeys. Use welcoming and inclusive language.
 - ◇ Raise hand or use Zoom chat with questions.
 - ◇ Everyone makes Python errors: please let us know if you get stuck.
 - ◇ We will be recording this session.
- 



What is web scraping?

Extracting data from a web page using cut-and-paste, code, or another tool that parses the HTML.

Examples of web scraping in research

nature
human behaviour

LETTERS

PUBLISHED: 3 APRIL 2017 | VOLUME: 1 | ARTICLE NUMBER: 0079

Millions of online book co-purchases reveal partisan differences in the consumption of science

Feng Shi^{1†}, Yongren Shi^{2†}, Fedor A. Dokshin³, James A. Evans^{1,4*} and Michael W. Macy^{3*}

Passionate disagreements about climate change, stem cell research and evolution raise concerns that science has become a new battlefield in the culture wars. We used data derived from millions of online co-purchases as a behavioural indicator for whether shared interest in science bridges political differences or selective attention reinforces existing divisions. Findings reveal partisan preferences both within and across scientific disciplines. Across fields, customers for liberal or 'blue' political books prefer basic science (for example, physics, astronomy and zoology), whereas conservative or 'red' customers prefer applied and commercial science (for example, criminology, medicine and geophysics). Within disciplines, 'red' books tend to be co-purchased with a narrower subset of science books on the periphery of the discipline. We conclude that the political left and right share an interest in science in general, but not science in particular. This underscores the need for research into remedies that can attenuate selective exposure to 'convenient truth', renew the capacity for science to inform political debate and temper partisan passions.

In its quest for an objective understanding of the world¹, modern science has practised two distinct forms of political neutrality: as an apolitical 'separate sphere' detached from ideological debates, and as a 'public sphere' relevant to political issues but with balanced

perceived liberal bias in policies advocated by social scientists. For example, the conservative-funded scientific counter-movement in climate change research suggests the possibility of politically driven scientific polarization^{10,11}. When science becomes politicized, partisans tend to cast doubt on scientific consensus through questioning its inherent uncertainty^{12–14}. This process is manifest not only in conservative resistance to climate change, but in historically liberal resistance to consensus over the positive benefits of genetically modified organisms, vaccination, nuclear power and the safe storage of nuclear waste¹⁵.

Survey data show little overall change in public confidence in science since 1970, but beneath the surface there is a marked shift: conservatives in the Vietnam era were more confident in science than liberals, but today that pattern has reversed¹⁶ (Supplementary Fig. 1). Does public exposure to science play an integrative role by encouraging and informing empirical validation? Or has selective attention instead reinforced the 'Big Sort' of American politics^{17–19} — the tendency to cluster in like-minded communities?

Much previous research has used surveys to investigate political alignments of the producers of science (with a few exceptions^{20,21}). We focus instead on the consumers of science, using online co-purchases of books on science and politics as a behavioural indication of preferences held by customers who 'vote with their pocket-

Scraped book web pages on Amazon and Barnes & Noble sites for “Customers Who Bought This Item Also Bought”

Examples of web scraping in research

Research-Based Article

New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings

Geoff Boeing¹ and Paul Waddell¹

Abstract

Current sources of data on rental housing—such as the census or commercial databases that focus on large apartment complexes—do not reflect recent market activity or the full scope of the US rental market. To address this gap, we collected, cleaned, analyzed, mapped, and visualized eleven million Craigslist rental housing listings. The data reveal fine-grained spatial and temporal patterns within and across metropolitan housing markets in the United States. We find that some metropolitan areas have only single-digit percentages of listings below fair market rent. Nontraditional sources of volunteered geographic information offer planners real-time, local-scale estimates of rent and housing characteristics currently lacking in alternative sources, such as census data.

Keywords

big data, Craigslist, data science, GIS, housing, urban economics, web scraping

Introduction

It would be difficult to overstate the importance of the rental housing market in the United States, despite longstanding cultural attitudes and the long-term economic recession in many

sources. New York and San Francisco unsurprisingly have the first and third highest rent per square foot, and North Dakota comes in second, reflecting its recent oil industry boom and housing shortage. We assess affordability by calculating rent

Journal of Planning Education and Research
2017, Vol. 37(4) 457–476
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0739456X16664789
journals.sagepub.com/home/jpe
SAGE

Scraped 11 million
Craigslist rental
entries to analyze
spatial patterns

Web scraping as a research method

Quality & Quantity
<https://doi.org/10.1007/s11135-021-01164-0>



Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences

Alex Luscombe¹ · Kevin Dick² · Kevin Walby³

Accepted: 7 May 2021

© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Web scraping, defined as the automated extraction of information online, is an increasingly important means of producing data in the social sciences. We contribute to emerging social science literature on computational methods by elaborating on web scraping as a means of automated access to information. We begin by situating the practice of web scraping in context, providing an overview of how it works and how it compares to other methods in the social sciences. Next, we assess the benefits and challenges of scraping as a technique of information production. In terms of benefits, we highlight how scraping can help researchers answer new questions, supersede limits in official data, overcome access hurdles, and reinvigorate the values of sharing, openness, and trust in the social sciences. In terms of challenges, we discuss three: technical, legal, and ethical. By adopting “algorithmic thinking in the public interest” as a way of navigating these hurdles, research-

Advice, technology and tools

Work

Your story

Send your careers story to: naturecareerseditor@nature.com

```
}, {  
  init: function() {  
    var self = this;  
    this.element.html(can.view('///app/src/views/sign  
    this.element.parent().addClass('login-screen');  
  
    App.db.getSettings().then(function(settings) {  
      App.attr('settings', settings);  
      self.element.find('#login-remember').prop('c  
  
    App.db.getLoggedAccount().then(function(acco  
      if(account) {  
        self.options.attr('username', accou  
        self.options.attr('password', accou  
  
        if(account.attr('username') && acco  
          account.attr('usern
```

TOOLS THAT EASE DATA COLLECTION FROM THE WEB

Custom web scrapers are driving research — and collaborations.
By Nicholas J. DeVito, Georgia C. Richards and Peter Inglesby



Web scraping is an approach of last resort

- ◇ Does the site provide the data in a downloadable format (e.g. CSV, Excel)?
- ◇ Do they have an API for querying and receiving data?
- ◇ Would they share the data if contacted?
- ◇ Is the data available in another resource?

Chou, Sophie. [To scrape or not to scrape: technical and ethical challenges of collecting data off the web](#), 24 April 2016.





Anatomy of a web page



```
<html>
  <head>
    <link href="css_file.css" rel="stylesheet" type="text/css" media="all">
  </head>
  <body>
    <div id="text-section1" class="box-around">
      <p class="bold-paragraph" style="font-size:16">
        Here's some bold text and
        <a href="https://library.gwu.edu" id="library-link">a link to
        GW Libraries</a>
      </p>
      <p class="bold-paragraph extra-large">
        Here's another paragraph, even bigger
      </p>
    </div>
    <table id="table1">
      <tr>
        <td>Stuff inside a table cell</td>
      </tr>
    </table>
  </body>
</html>
```

```
.bold-paragraph {
  font-weight: bold;
  color: red
}
```

The <div> has a
unique id of
"text-section1"

href here is an
attribute of the <a> tag

<p> is a tag, or node
This <p> has class
"bold-paragraph"



Let's start scraping!

Scraping headlines from the GW Hatchet:
<https://www.gwhatchet.com>

Using Google Colab:
<https://colab.research.google.com>



https://colab.research.google.com

colab.research.google.com/notebooks/intro.ipynb#recent=true

Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

Table of contents

- Getting started
- Data science
- Machine learning
- More Resources
- Machine Learning Examples
- Section

+ Code + Text Copy to Drive

Connect Editing

Examples Recent Google Drive GitHub Upload

Filter notebooks

Title	First opened	Last opened	
Welcome To Colaboratory	0 minutes ago	0 minutes ago	

NEW NOTEBOOK CANCEL

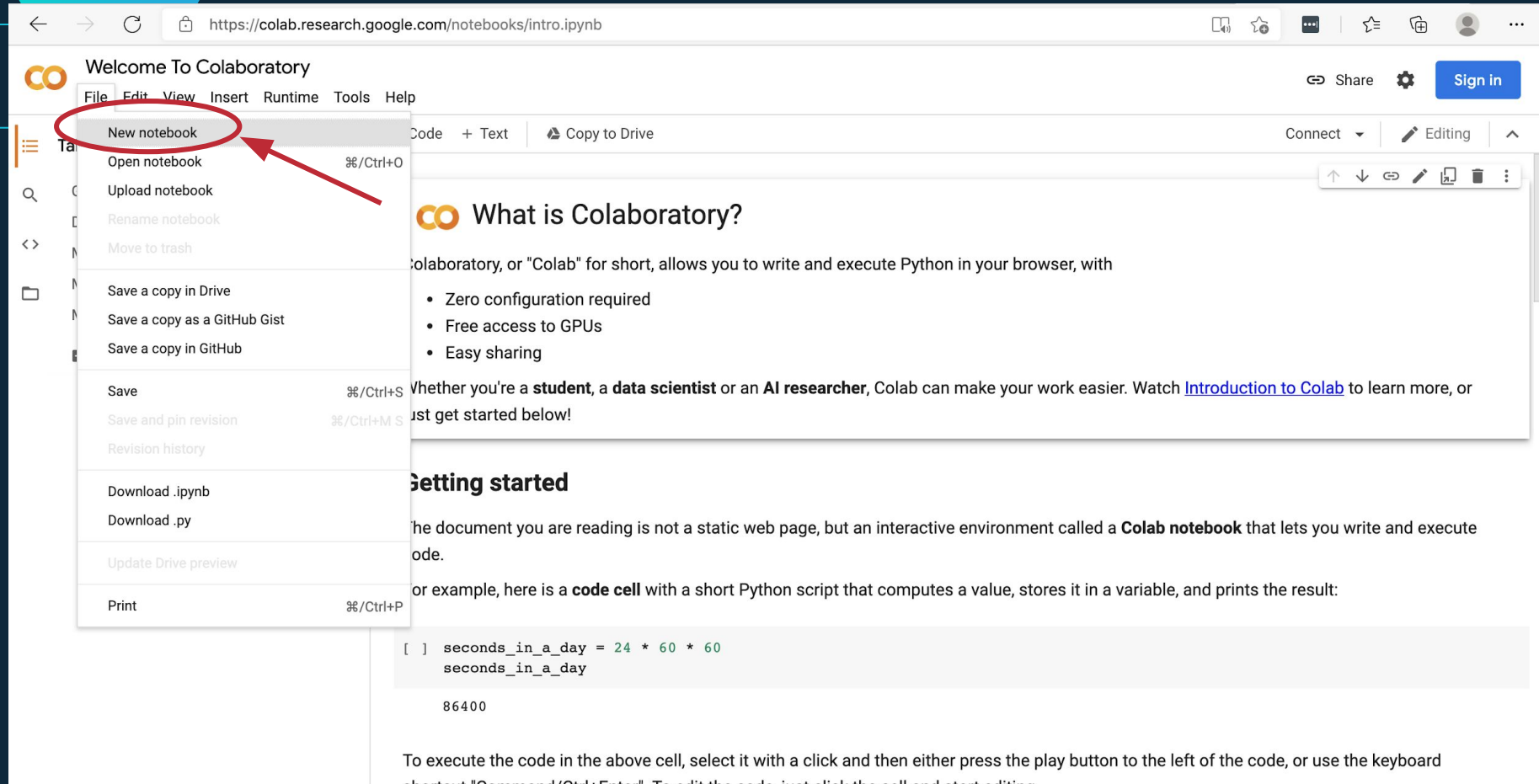
Introduction to Colab to

notebook that lets you write

, and prints the result:

left of the code, or use the

https://colab.research.google.com



The screenshot shows the Google Colaboratory web interface. The browser address bar displays `https://colab.research.google.com/notebooks/intro.ipynb`. The page header includes the Colab logo, a 'Welcome To Colaboratory' message, and a 'Sign in' button. A menu bar at the top contains 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. The 'File' menu is open, and a red circle and arrow highlight the 'New notebook' option. Other menu items include 'Open notebook', 'Upload notebook', 'Rename notebook', 'Move to trash', 'Save a copy in Drive', 'Save a copy as a GitHub Gist', 'Save a copy in GitHub', 'Save', 'Save and pin revision', 'Revision history', 'Download .ipynb', 'Download .py', 'Update Drive preview', and 'Print'. The main content area features a 'What is Colaboratory?' section with a list of bullet points: 'Zero configuration required', 'Free access to GPUs', and 'Easy sharing'. Below this, a paragraph explains that Colab allows writing and executing Python in the browser. A 'Getting started' section follows, stating that the document is an interactive environment called a 'Colab notebook'. It provides an example of a 'code cell' containing a Python script that calculates the number of seconds in a day (24 * 60 * 60), resulting in 86400. The code is shown in a light gray box with a green prompt character '['.

https://colab.research.google.com/notebooks/intro.ipynb

Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

New notebook

Open notebook

Upload notebook

Rename notebook

Move to trash

Save a copy in Drive

Save a copy as a GitHub Gist

Save a copy in GitHub

Save

Save and pin revision

Revision history

Download .ipynb

Download .py

Update Drive preview

Print

What is Colaboratory?

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch [Introduction to Colab](#) to learn more, or just get started below!

Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ] seconds_in_a_day = 24 * 60 * 60
    seconds_in_a_day

86400
```

To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command (Ctrl) Enter". To edit the code, just click the cell and start editing.



Public version of notebook code:

https://github.com/gwu-libraries/gwlibraries-workshops/blob/master/web-scraping-python/notebook_public.ipynb





Is it legal?

- ◇ *hiQ Labs, Inc. v. LinkedIn Corp.* (2019): Ninth Circuit Court of Appeals ruled that automated scraping of **publicly accessible data** likely does not violate the Computer Fraud and Abuse Act (CFAA).
- ◇ Only scrape publicly available sites and public data (not data behind authentication).
- ◇ Consider copyright of any resources.
- ◇ Check local legislation, especially outside U.S.

Fischer, C. and A. Crocker. [Victory! Ruling in hiQ v. LinkedIn Protects Scraping of Public Data.](#) Electronic Frontier Foundation, 10 Sep 2019.





Is it ethical?

- ◇ Consult your advisor and potentially, GW's IRB for human subjects research
- ◇ When data comes from people, consider possible harm.
 - Is your topic sensitive? (mental illness, financial status, health, interactions with law enforcement)
 - Are individuals vulnerable? (minors, patients)
 - Are you collecting personally identifiable info? (includes account names)
- ◇ Can you get the creator's permission for quotes?
- ◇ Include your ethical decision-making in your paper

Walsh, Melanie. "[User Ethics and Privacy Concerns](#)," [Introduction to Cultural Analytics with Python](#), 2021.





Be respectful and considerate

- ◇ Review the site's robots.txt and don't scrape out-of-bounds content
- ◇ Focus on publicly available content, not content behind authentication.
- ◇ Don't be deceptive (bypassing authentication, faking sessions)

Library Carpentry. "[Intro to Web Scraping](#)." 2018.





Be respectful and considerate

- ◇ Don't overwhelm the site with requests. Insert pauses into your scraping code.
- ◇ Do your scraping during low-traffic times
- ◇ Don't interfere with website's business
- ◇ Don't scrape library databases and online journals. This violates our license and potentially cuts off others' access to the content.

Library Carpentry. "[Intro to Web Scraping](#)." 2018.





Other Python libraries

- ◇ [BeautifulSoup](#): well-used entry-level library, with many introductory tutorials. Good for static web pages.
- ◇ [Selenium](#): Good for dynamic sites and for sites where you need to interact with buttons and menus to access content. Uses a headless browser in the background, so can be slower.
- ◇ [Scrapy](#): speedy, good for complex scraping operations. Used commercially.





More resources

- ◇ [Library catalog search](#) on ebooks about Python and web scraping (sorted with newest first).
- ◇ LinkedIn Learning, "[Web Scraping with Python](#)". Uses Scrapy, log in as a GWU user.
- ◇ Walsh, Melanie. "[Web Scraping](#)", *Introduction to Cultural Analytics with Python*.

Getting help

Make an appointment for a coding consultation:

<https://calendly.com/gwul-coding>





Credits

- ◇ Walsh, Melanie. [Introduction to Cultural Analytics with Python](#). (2021)
- ◇ Library Carpentry. "[Intro to Web Scraping](#)." 2018.





Thank you!

We'd love your feedback for next time!

<https://forms.gle/beXhYkP29EiojvyL6>

