



Web Scraping with Python

February 11, 2021

GW Libraries and Academic Innovation

Slides: go.gwu.edu/scrapingpython

Laura Wrubel, Software Development Librarian
lwrubel@gwu.edu





Agenda

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





Ground rules

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





What is web scraping?

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



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.





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.
- ◇ 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?

- ◇ Even with public data, consult your advisor and potentially, GW's IRB.
- ◇ 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.





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>
```

css_file.css

bold-paragraph

**bold-paragraph {
font-weight: bold;
color: red**

**href here is an
attribute of the <a> tag**

<p> is a tag, or node





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 in a browser. The address bar displays `https://colab.research.google.com/notebooks/intro.ipynb`. The page title is "Welcome To Colaboratory". A menu bar at the top includes "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". The "File" menu is open, and the "New notebook" option is highlighted with a red circle and a red arrow pointing to it. Other options in the menu 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 the following text:

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- 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.



Be respectful and considerate

- ◇ Review the site's robots.txt and don't scrape out-of-bounds content
- ◇ Don't be deceptive (bypassing authentication, faking sessions)
- ◇ Don't overwhelm the site with requests. Insert pauses into your scraping code
- ◇ Do scraping during low-traffic times

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





Be respectful and considerate

- ◇ Don't get in the way of their business by interfering with orders or slowing it down
- ◇ Focus on publicly available content, not content behind authentication.
- ◇ 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

- ◇ Dolsy Smith and Dan Kerchner, "Intro to Web Scraping" past workshop slides
- ◇ Walsh, Melanie. [Introduction to Cultural Analytics with Python](#). (2021)
- ◇ Library Carpentry. "[Intro to Web Scraping](#)." 2018.





Thank you!

