

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.



- 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?
- \diamondsuit Is the data available in another resource?





- 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.





- 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





Anatomy of a web page



```
.bold-paragraph {
                                             font-weight: bold;
<html>
                                             color: red
   <head>
       <link href="css file.css" rel="stylesheet" type="text/css" media="all">
   </head>
   <body>
       <div id="text-section1" class="box-around">
          Here's some bold text and
             <a href="https://library.gwu.edu" id="library-link">a link to
GW Libraries</a>
                                           href here is an
          attribute of the <a> tag
          Here's another paragraph, even bigger
           is a tag, or node
       </div>
       Stuff inside a table cell
         </body>
</html>
```

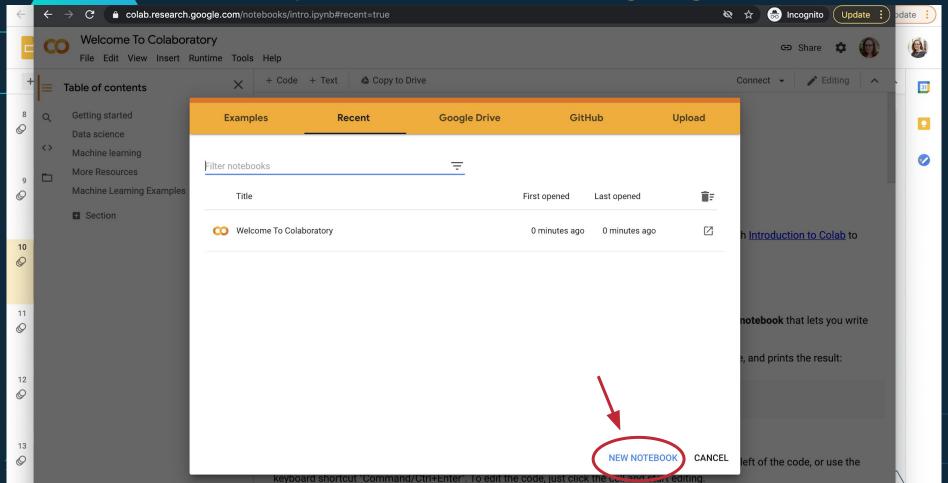


Scraping headlines from the GW Hatchet: https://www.gwhatchet.com

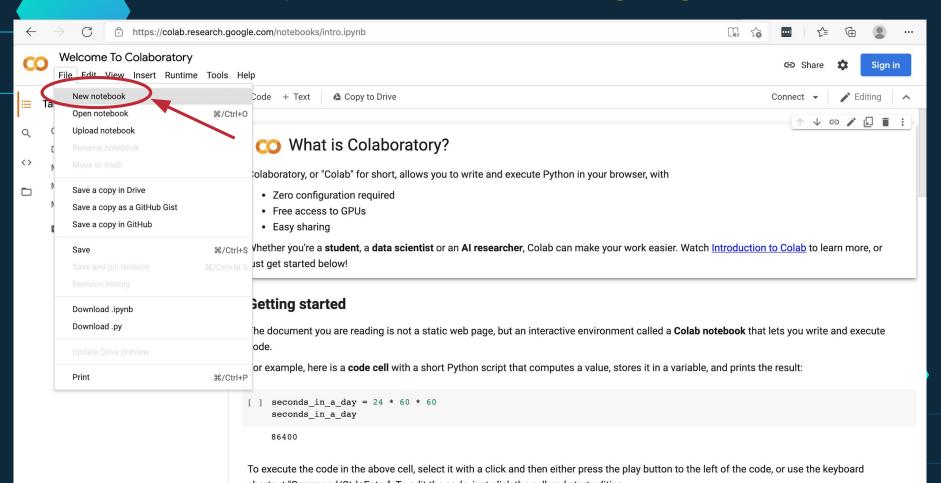
Using Google Colab: https://colab.research.google.com



https://colab.research.google.com



https://colab.research.google.com





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





- 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.





- 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.





- 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





- Dolsy Smith and Dan Kerchner, "Intro to Web Scraping" past workshop slides
- Walsh, Melanie. <u>Introduction to Cultural Analytics</u> <u>with Python</u>. (2021)
- Library Carpentry. "Intro to Web Scraping." 2018.





Thank you!

