

Beginner's Guide to Web Scraping

Based on [Compute Canada HSS's Web-Scraping Workshop](#)



Who are the GC Digital Fellows?

We are students from programs across the GC who integrate creative, critical, innovative use of technology in our scholarly and pedagogical practice. We work collaboratively to develop programs that support students, faculty, and staff as they develop their digital projects.

We offer **support for digital scholarship**

- Workshops
- Consultations / Office Hours
- Working Groups
- Online Resources

Checking-in

- [Google Chrome](#)
- [Data Miner's Scraper](#) (browser plug-in)
- Zoom orientation
 - It's okay to turn your video off! 😊
 - Please feel free to raise your questions in the chat! 👍
 - Time will be set aside for practice
- Slides will be made available to you after the workshop

Introductions!

Let's get to know each other a little!

Name,
gender pronouns,
program, and
what's brought you to this workshop today?

Goals of today's workshop

- Getting an understanding of what web-scraping is
- When is web-scraping useful and when it isn't
- Coming away with a method to scrape content
 - Website structures and HTML

What is web-scraping?

What is web-scraping?

- Method to extract information from websites
 - Transforming non-structured data to structured formats (e.g. .csv .xls)
- Manual process: Copy & pasting
 - Faster & less error-prone with small dataset
- Automated process
 - Saves time in the long-run

Automating

- Define information to look for and the (specific) type of information to extract
- How much of the website to scrape
 - Just the page you're on? The whole website?
- Frequency of scraping
 - Longitudinal project looking over a period of time? Snapshots of particular moment?

Is scraping your best option?

- Size of data
- Alternative ways to extract information?
 - Export information or download datasets directly
 - [Our World in Data](#)
 - APIs (or Application Programming Interface)
 - Extracts data in semi-structured/structured formats
 - Not always available on all sites

Scraping v. Crawling

Scraping

- Targeted; identifies and extract specific information
 - News headlines, captions of Instagram posts
- Focus on extracting *data*

Crawlers

- Indexes the web (e.g. Google for SEO); crawls through entire websites and all the links associated with the site
 - [Internet Archive](#) and their [waybackmachine](#)
- Focus on *indexing* - finding all the links so that they can be presented in search pages

When is web-scraping used?

- Online stores scoping competitor prices to adjust theirs
- Getting customer reviews and feedback to improve product
- Contact information off websites
- Online news outlets for headlines over a period of time
- Social media posts* about specific topics
- ...your projects!

Can I scrape everything, everywhere?

... it depends

- Intense scraping puts high demands on server; can cause sites to break
- robots.txt
 - Guidelines for how to scrape their site
 - <https://archive.org/robots.txt>
 - Remains advisory; no real legal means in enforcing, and not all bots (or the people behind the bots) obey
- Why are you scraping?
 - [Internet Archive made decision to ignore robots.txt since 2017](#)

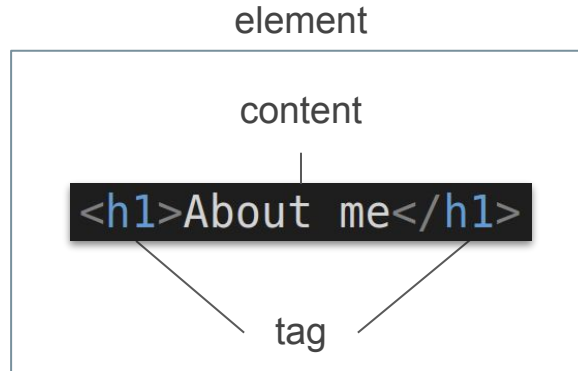
Can I scrape everything, everywhere?

- Can I take this data?
 - What's the terms of use?
- Is there a restriction to what I can do with the data?
 - Can you distribute your scraped data freely?

What is behind a webpage?

Website structure

- **HyperText Markup Language**
 - We must know where to point our tool (or script) to scrape
- Well-organized and descriptive structures makes things easier to scrape



paramajmera.github.io



Data Scraper - Easy Web Scraping

Available on Chrome

Offered by: dataminer.io

★★★★★ 601 | Productivity | 200,000+ users

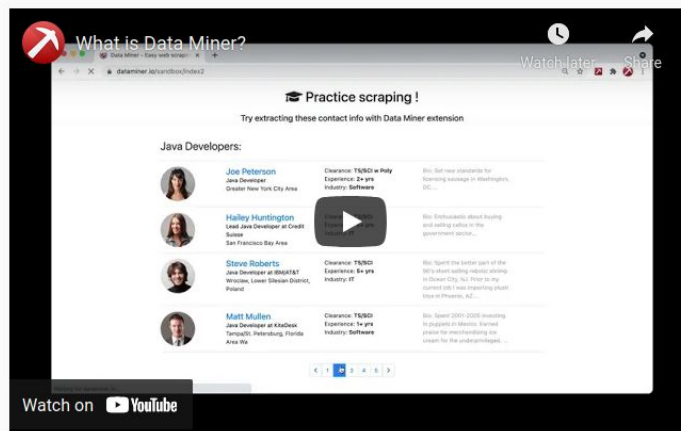
Overview

Privacy practices

Reviews

Support

Related



Hands-On

IMDb

- Row: Each movie listing
- Columns: Rank, Title, Year, IMDb Rating

Additional Challenge:

- Scrape the headlines of the BBC

Goodreads Quotes - Pagination

- Rows: Each quote listing
- Columns: Quote, Authors, Likes
- Scrape 3 pages of quotes automatically

Beyond Data Miner

Beyond Data Miner (or GUI tools)

- Additional and advanced features cost money
 - Large volume of data
- Not easily replicable or transparent process
 - Some journals are asking to see scripts
- Web-scraping scripts
 - Scrape and parse data
 - Some cleaning

Learning a new language

- Python
 - BeautifulSoup (bs4)
- R
 - rvest
- Choosing a language
 - Testing things out
 - Jupyter Notebook
 - Rstudio Cloud


Where to go from here?

- [Intro to HTML & CSS](#)
- [Intro to Python](#)
- [Intro to R](#)
- [Using Beautiful Soup \(bs4\) a Python package](#)
- [Using rvest a R package \(external\)](#)


Evaluations

cuny.is/GCDI-webevals

How can I get involved with GCDI?

GC Digital Initiatives


[About](#) [People](#) [Calendar](#) [Participate](#) [Resources](#) [News](#) 



Interdisciplinary Events, Workshops, & Speaker Series



Graduate Center Digital Initiatives (GCDI) brings together the work of leading scholars and technologists at the CUNY Graduate Center to pioneer new modes of inquiry that integrate digital tools and methods into the research, teaching, and service missions of the university.



Projects



Degrees & Certificates



Workshops



Guides & Tutorials



Join us for an event


How can I get involved with GCDI?

CUNYACADEMICCOMMONS


Google Custom Search Search

PeopleGroupsSitesPapersCoursesEventsNewsAbout

GC Digital Initiatives at the CUNY Graduate Center



HomeForumEventsPapersFilesDocsMembers 190



GC Digital Initiatives at the CUNY Graduate Center


Public Group active 3 days ago

This group accompanies the the GCDI website (<http://gcdi.commonsgc.cuny.edu/>). Here, members of the GC community can share news of recent events and new projects and can begin to build connections between projects.


Quick Link: <http://cuny.is/group-gcdi>

Recent group activity

RSS



Inés Vañó García started the topic CFP – Special Issue – Journal of Interactive Technology and Pedagogy (11/30/19) in the forum

 Digital Initiatives at the CUNY Graduate Center 3 days ago

The Journal of Interactive Technology and Pedagogy
Special Issue:

Group Admins

