Web Scraping Workshop

William & Mary ACM

2/13/2025

Agenda

01 Announcements

02 Intro to Web Scraping

03 Interactive Web Scraping Demo

Hackathon Trips

Hackathon (n.) An event where people come together to make cool things fast, or just to learn!

We work to organize transportation and teams to build projects together!

Register for HooHacks today!
March 29th-30th
https://www.hoohacks.io/

Upcoming Events

CS Banquet (Tonight!!!)
February 13, 7:30pm - 9:30pm in Sadler Tidewater AB

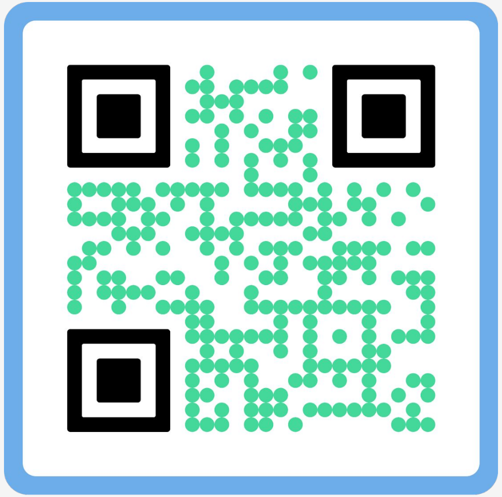
CS Department Townhall
February 28, 12pm - 1pm in McGlothlin St Hall, Room 020.

If you have any topics you'd like to see discussed at the town hall, please submit them through the form below by February 10th!

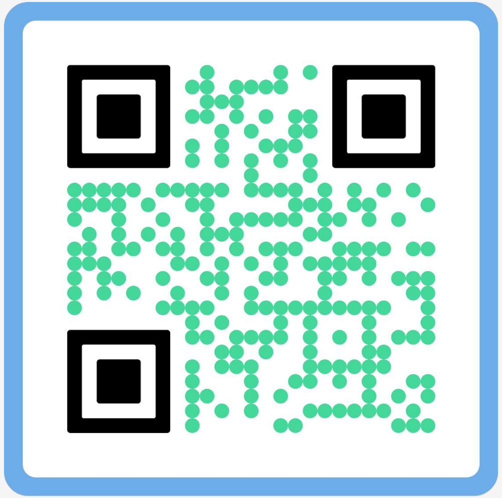
https://go.wm.edu/K2n8cc

Socials

Join our Groupme!



Follow us on Instagram!



Dues

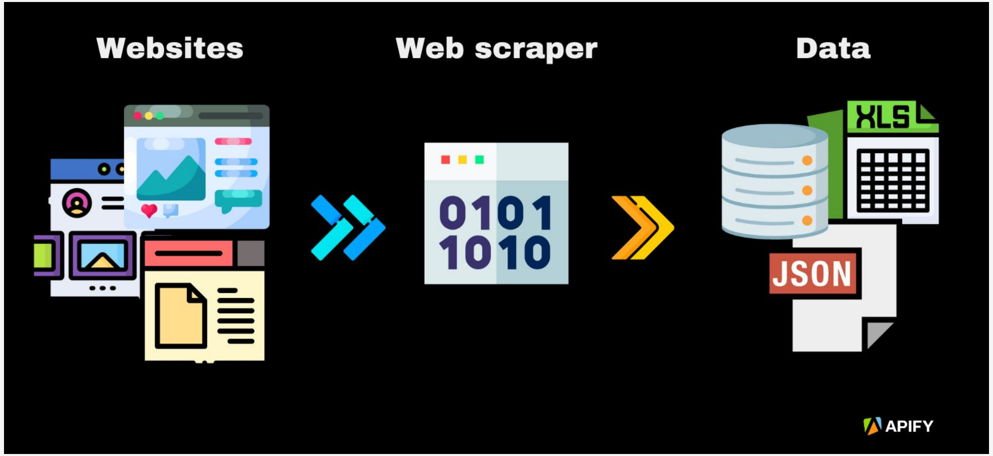
For $15, get access to:

- ACM Alumni Network
- ACM Dues-only member group chat and social events
- ACM T-shirt
- Other cool stuff

Talk to Lorenzo if you are interested!

What is Web Scraping?

Web Scraping is simply the process of extracting data from websites automatically.



Why Web Scrape?

Price Monitoring
I want to create a bot to monitor the price of my favorite pair of shoes and get the lowest price possible.

Data Analysis
I want to gather a large amount of data on weather data from an online database.

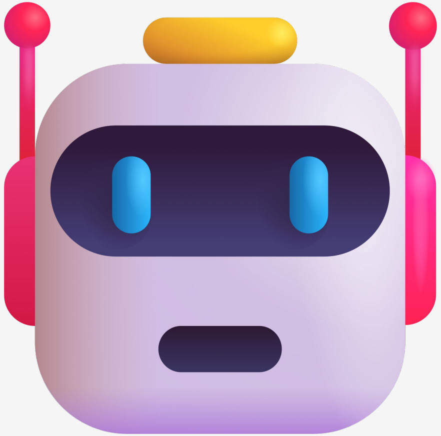
Email Marketing
I want to grab the emails of a bunch of journalists to email them about my newest book release. (Questionable)

How Does it Work?

Web scrapers effectively send out a "robot" to a specified URL to gather the raw HTML code that underlies a website.

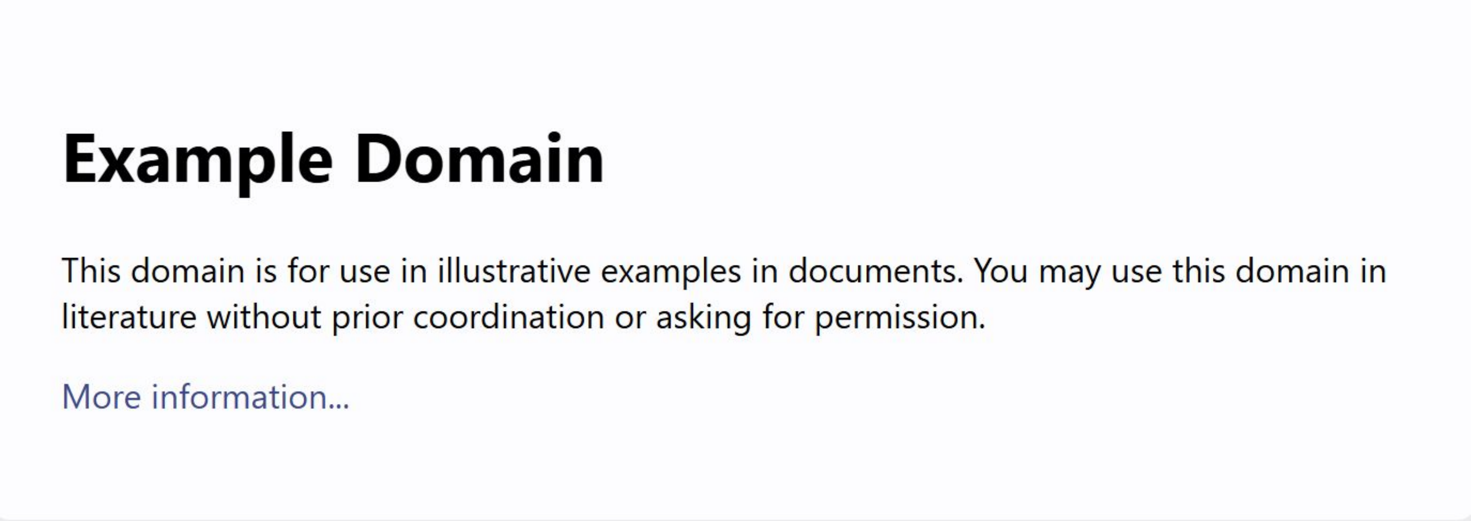
The programmer can then parse through this raw HTML code to gather the data that they desire.

More advanced web scrapers can gather the underlying css, javascript, etc.



How Does it Work?

https://example.com/

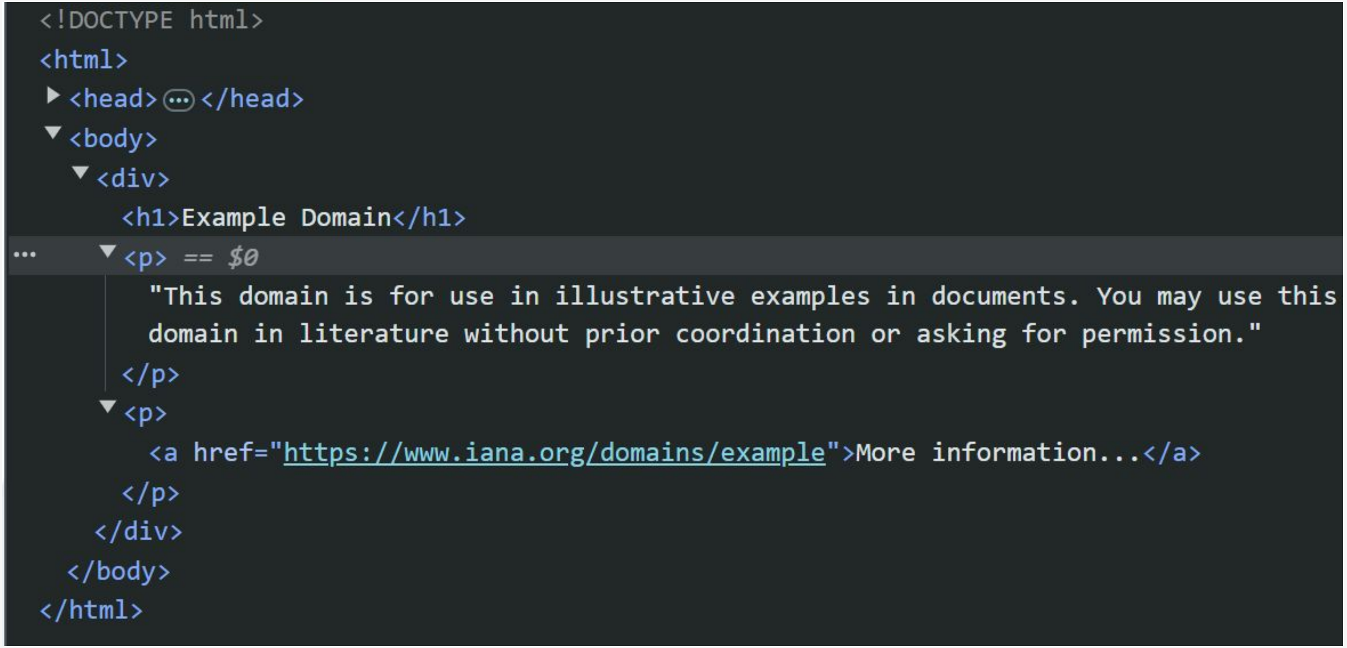


Example Domain

This domain is for use in illustrative examples in documents. You may use this domain in literature without prior coordination or asking for permission.

More information...

How Does it Work?



```
<!DOCTYPE html>
<html>
 ▶<head>(…)</head>
 ▼<body>
   ▼<div>
      <h1>Example Domain</h1>
…    ▼<p> == $0
        "This domain is for use in illustrative examples in documents. You may use this
        domain in literature without prior coordination or asking for permission."
      </p>
     ▼<p>
        <a href="https://www.iana.org/domains/example">More information...</a>
      </p>
    </div>
  </body>
</html>
```

Types of Web Scraping

Static

- Simply grabs the website's HTML code
- Best for websites that don't change much/load new information via JS
- Faster and more memory efficient
- Less complex, but much more limited

BeautifulSoup

scapy

Types of Web Scraping

Dynamic

- Loads and interacts with a web page
- Waits for JavaScript to execute
- Allows interaction with elements
- Useful for scraping modern websites with heavy JavaScript frameworks
- Slower and more complex

Legal/Ethical Concerns

Web Scraping itself is legal, but keep in mind…

- Violating Website Terms of Service
- Ignoring robots.txt Rules
- Copyright and Data Ownership Issues
- Overloading Servers
- Misuse of Data

Interactive Demo

You will need python and an appropriate IDE installed.

https://pypi.org/project/requests/

https://pypi.org/project/beautifulsoup4/

https://tinyurl.com/WMACM2025

Thanks for coming!

(Feel free to stay if you have any questions)

Next meeting:

Thursday, February 20th, 7pm in ISC 2280

You will learn about web-scraping and it's utilities, followed by a social activity so make sure to be there!