Unit 6: Data Collection

Scraping

Lesson 49

Derek Ruths

Overview of unit

Objectives:

- Understand why data collection is a time-consuming part of a project
- Get experience working with API collection
- Become familiar with standard data storage solutions: JSON, csv, and sqlite
- Understand how bias arises and why it's hard to detect

1. Overview
2. API-based collection
3. Reddit data collection
4. JSON
5. Scraping
6. Sampling

Lesson overview

Objectives

- Understand how scraping data collection works
- Know some of the key pitfalls with scraping data

Outline

- What is web scraping?
- Limitations
- Pitfalls

When can't we use an API?

Some websites don't expose an API

Some websites have an API that doesn't expose the info we want

given website instead of JSON/digestible form

info's available but it may not have an api

Crash course on HTML...

Scraping exercises...

Montreal Gazette Articles

McGill Academic Calendar

https://www.whosdatedwho.com/

The legality of web scraping

Very little clear law governing web scraping

Consider the website owner's perspective:

- Very little upside of having site scraped (visibility?)
- Unnecessary load on their servers
- Loss of intellectual property

abuse of resources company is fielding your requests

Unauthorized scraping might only acceptable for personal or research purposes

- Minimize server load
- Protect any data collected

if it's publically available anyway, then why not big question is purpose

Kinds of scraping

Static site scraping

what we've been doing

- Good for sites where the whole page loads at once
- Use wget, curl, requests, beautifulsoup

Dynamic site scraping (e.g., Goodreads)

social media w/ infinite scroll

- Necessary for pages that are "live"
- Use beautifulsoup, selenium

content changes after the page has loaded

+ need something on top that keeps the page constantly loading

tool/library that will drive a browser

Common challenges

Page walking/navigation

- Detecting the right links to visit
- Easy to miss page load errors

Extracting content

- Unreliable format/tags

Lesson wrap-up

Takeaways

- Web scraping is sometimes necessary
- It is not always legal … or polite
- Be a good citizen!

Up next

- Sampling