Intro to data scraping and API’s

Complex Systems 530

Data scraping/web scraping

- Pulling data from the internet (web sites, social media, etc.)

- Involves: crawling/searching, extraction, parsing, reformatting

- Often two general approaches:

  - Directly scraping (note possibly rude! Your program/bot(s) will make requests from their server)

  - Use an API!

Data scraping python libraries

- BeautifulSoup
- Selenium

What is an API?

- **A**pplication **P**rogramming **I**nterface
- A way for programs/software to communicate
- Client/server - can be for web, operating system, databases, etc.
- Web APIs
    - APIs for either web browser or web server
    - Github API, Google API, Facebook API…

API Basics

- Calling an API: Request
- Response

Types of API’s

- SOAP (Simple Object Access Protocol) - stricter, mostly used for enterprise scale stuff
- REST (Representational State Transfer) - most common for public API’s, used broadly for web API’s. Flexible and relates nicely to HTTP syntax
- GraphQL - new one introduced by Facebook

https://realpython.com/python-api/

Let’s explore with code: the `requests` library

- https://colab.research.google.com/drive/1YjEq7b_ZxJxS9baRJ87fWAzKEa8zwrSd#scrollTo=rubnEA9sqXru

Rate limits

- If you use this approach for research, you will likely run into rate limits set by the API service

- Many fancier API packages have functions for this, but still often an issue

- Pause your code occaisionally to avoid bumping into these

- Most fancier libraries have functions to allow you to monitor how close you are to rate limits

Ethics

- Be nice to people...
- Don't make evil bots
- Ask for data nicely
- Reminder that while twitter and other social media/internet data is public, many users don't realize that this means their data can be used for research---they may disclose things (like health status) that they only expect their followers to see.
    - Be thoughtful about what user data you include in papers, etc.
- See here: https://howwegettonext.com/scientists-like-me-are-studying-your-tweets-are-you-ok-with-that-c2cfdfebf135 for a recent commentary on this