Been There, Scraped That.

Shengli Hu

Cornell NBA 6920

December 17, 2016

So... Why do you want to scrape data?

- Cool to have some data floating around
- For fun and profit
    - A clear question in mind?
    - What kind of data is needed?
    - Format? Structure? Granularity? Span? ...

How do we go about scraping data?

- Public processed datasets:
    - Wikipedia, Amazon, Reddit, AirBnB, NYCTaxi, IMDB, . . .
- Small static websites:
    - The American Presidency Project
- Large modern websites:
    - Application Programming Interface (API)
        - NOT for data scraping
        - RESPECT rate limit: check rate limit, sleep between calls
        - SAVE all raw data: disks are cheap, API calls are not

Babysitting

- How much data do you need?
  - Parallel programming
- There is no such thing as perfect data :/
  - robust error checking, email scripts
- Think about data format, hard
  - Flat files, databases, json?
- Contributions
  - data, code

Takeaways

- Start with your question in mind
- Think about what data you need
- Search for existing solutions (datasets, codes, tips)
- Start with small, manageable size, estimate how long it will take
- Keep logs and raw data
- Sanity checks