Web Scraping: Legality and How-To

CS50 for JDs

Web scraping, APIs, and crawling

APIs

- Request for data through official channels
- Wikipedia, Yelp, Spotify

Web crawling

- Linking pages together
- Google, Archive.org

Web scraping

- Mechanizing the gathering of data from public-facing websites

Web scraping: a conceptual step-by-step

By hand:

1. Go to web page.
2. Look at web page.
3. Find information.
4. Save information.
5. Go to next web page.

By program:

1. Request web page.
2. Parse web page.
3. Find information.
4. Save information.
5. Request next web page.

Accessing the Web through Code

urllib

- Built into Python
- Basic connections

requests

- https://requests.readthedocs.io/en/master/
- External library

Programming Etiquette

- Rate-limit requests
- Respect robots.txt and ToS
- No copyright infringement
- Noncommercial use
- Only access public information

Related Laws & Principles

- 1986: Computer Fraud and Abuse Act (CFAA)
- 1998: Digital Millennium Copyright Act (DMCA)
    - Fair use
- 2016: Better Online Ticket Sales Act
- Trespass to Chattels
- Contract Violation

Example: Trespass to Chattels

- eBay v. Bidder’s Edge (2000)
    - Access was unauthorized (ToS; IP blocking)
    - Access was only fraction of eBay’s traffic, but continuance could lead to a cumulative effect
    - Implicitly overruled in *Intel v. Hamidi* (2003)

Example: Terms of Service as Contracts

- Internet Archive v. Suzanne Shell (2007)
    - Each page links to ToS which specify $5,000 damages for each copied page
    - Copying site contents may constitute agreeing to ToS

- Southwest Airlines Co. v. BoardFirst, LLC (2007)
    - Browsewrap contract violation (compare clickwrap ToS)

Example: Fair Use

- Kelly v. Arriba Soft Corp. (2003)
  - Image search service providing thumbnails of copyrighted images
  - Ruled as fair use

- Associated Press v. Meltwater U.S. Holdings, Inc (2013)
  - News article aggregator with excerpts of AP news stories
  - Not held as fair use of AP’s copyrighted content

Example: CFAA

- Craigslist Inc. v. 3Taps Inc. (2013)
  - A cease-and-desist letter + IP blockage
  - Sufficient to serve as notice that further access is unauthorized

- QVC, Inc. v. Resultly, LLC (2015)
  - Scraping by Resultly crashed QVC servers
  - Lack of intent to harm means no violation of CFAA

- hiQ Labs, Inc. v. LinkedIn Corp. (2019)
  - Scraping publicly available LinkedIn data to sell to employers
  - Not unauthorized access if data is freely accessible