Introduction to Web Scraping with R

Legal Issues

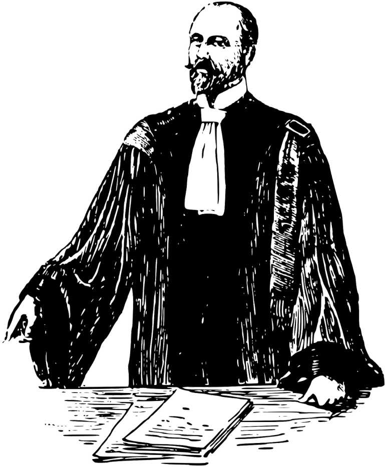
Simon Munzert | IPSDS

Is web scraping legal?

Is web scraping legal?

Disclaimer

Obviously, I am not a lawyer. Do not rely on any of my comments on this topic. If you are seriously worried about the legality of your scraping work, please consult a legal expert.



Is web scraping legal?

Scraping vs. crawling

- **Web scraping**: downloading data from a very specific page in a (semi-)automated manner
- **Web crawling**: automatically downloading webpage data, extracting hyperlinks, following links, downloading webpage data, ... (e.g.: Googlebot, BaiduSpider)

This course mostly is scraping, not crawling or web harvesting!

Is web scraping legal?

Scraping vs. crawling

- **Web scraping**: downloading data from a very specific page in a (semi-)automated manner
- **Web crawling**: automatically downloading webpage data, extracting hyperlinks, following links, downloading webpage data, ... (e.g.: Googlebot, BaiduSpider)

This course mostly is scraping, not crawling or web harvesting!

- web scraping per se is not illegal
- there is no unambiguous **yes** or **no** for specific applications in any country according to current jurisdiction
- so far, legal cases (especially in the US) often (but not always) dealt with commercial interest, crawling applications, and often (but not always) huge masses of data, e.g., *eBay vs. Bidder's Edge, AP vs. Meltwater, Facebook vs. Pete Warden*

Is web scraping legal?

Why web scraping can be problematic

- violation of copyrights, Terms of Service, consumption of bandwith

Is web scraping legal?

Why web scraping can be problematic

- violation of copyrights, Terms of Service, consumption of bandwith

Some counter-arguments

- data is publicly accessible
  → but the website as a "creative arrangement" might be copyrighted
- this is fair use → depends on your use
- it's the same what my browser does
  → not exactly, and many ToS prohibit automated uses of their data
- this is unfair—Google's business model is built on crawling the whole web
  → true, but you are not Google

(some arguments borrowed from Benoit Bernard; https://goo.gl/zCB3d6)

Is web scraping legal?

Some interesting comments by Pablo Hoffman, co-founder of Scrapinghub

- As long as they don't crawl at a disruptive rate, scrapers do not breach any contract (in the form of terms of use) or commit a crime (as defined in the Computer Fraud and Abuse Act).
- Website's user agreement is not enforceable as a browse-wrap agreement because companies do not provide sufficient notice of the terms to site visitors.
- Scrapers accesses website data as a visitor, and by following paths similar to a search engine. This can be done without registering as a user (and explicitly accepting any terms).

(found on https://goo.gl/jTt4ER)

Action from the other side

Things webpage administrators can do to prevent you from scraping massive amounts of data from their pages

- block your IP address
- identify your approximate geolocation from your IP address, then block
- move content exclusively to web services / APIs
- block bots with a particular user agent string (more on that later)
- challenge-response tests like **C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part (CAPTCHA)
- obfuscation of data
- frequent changes in HTML/CSS

Summary

Summary

- there is no unconditional "legal" or "illegal" status of web scraping
- your use of the data can violate the data owner's rights
- targeted scraping efforts with limited traffic usually do not cause any problems