NCSC Data Dives

Episode #1: What's Web Scraping?

By Andre Assumpcao | July 2023

What's web scraping?

Web scraping is the process of extracting data from websites using automated tools or programs to access web pages, read the page's source code or other structured data on those pages, and extract the desired information.

What are the skills/tools used for web scraping?

It generally requires programming skills in languages like Python, R, or JavaScript – which are the most popular languages to write the programs that visit web pages and extract the desired data. Familiarity with web technologies such as HTML, CSS and XPath selectors, and regular expressions helps navigate websites, extract data, and perform advanced pattern matching – e.g., to match dates, names, dollar amounts. These skills sound complex but are basic elements of building websites.

Web scraping is not web crawling – which is systematically browsing the web and following links to discover and index web pages. Web crawlers, also known as spiders or bots, are used by search engines to gather information about web pages, index them, and make them searchable. Web crawling aims to explore and map the structure of the web by visiting and indexing as many pages as possible. A web crawler program can contain a web scraper module to extract information at the same time as the crawler is indexing and discovering new web pages.

What are the common use cases for web scraping?

Market research, aggregation of content, price comparison, and academic research are some of the common use cases for web scraping. Web scraping can be done over public or proprietary data available on the web, thus it is important to know that web scrapers will not necessarily consider ethical and legal implications of collecting and using web data.

EXAMPLE

A web scraper can type out a username and password, getting past a court records login page, fill in search boxes and retrieve case filings that meet a certain set of keywords (e.g., "debt collection"+"John Doe"+ "2019" in different search or filter boxes would identify debt collection cases filed against John Doe in 2019).

What are the risks for sites/institutions who have their content scraped?

Web scraping poses risks, including performance impact (such as making a website run slower), data privacy concerns (such as collection of sensitive data), data security concerns (such as scrapers that access content behind a firewall or login and expose usernames and passwords), intellectual property infringement (such as collection of proprietary content), and misrepresentation of data (such as data extracted without context or altered and manipulated). ***Importantly*, it's very hard to prevent web scraping**; it's better to mitigate risks and establish clear guidelines and legal penalties for data misuse.

NCSC Data Dives is a forum designed to help the court community dive deeper into data analytics and gain valuable insights to help solve data problems. Through a combination of both interactive group and targeted individualized sessions, Data Dives focuses on past projects, innovations, and topics of interest. Visit **ncsc.org/datadives** to learn more.

NCSC
National Center for State Courts

How to mitigate risks of web scraping?

There are multiple ways to reduce the risk of web scraping: clear terms of use for your website; actionable legal guidelines for those who scrape and misuse your data; implement rate limiting technology to reduce the ability of web scrapers to quickly obtain data; implement database URL masking to prevent scraper to automatically visit pages that encode information in their web address; adopt CAPTCHA technology for accessing data; provide data in bulk form via official request (with or without a fee); offer data through an application programming interface (API) thus making web scraping moot.

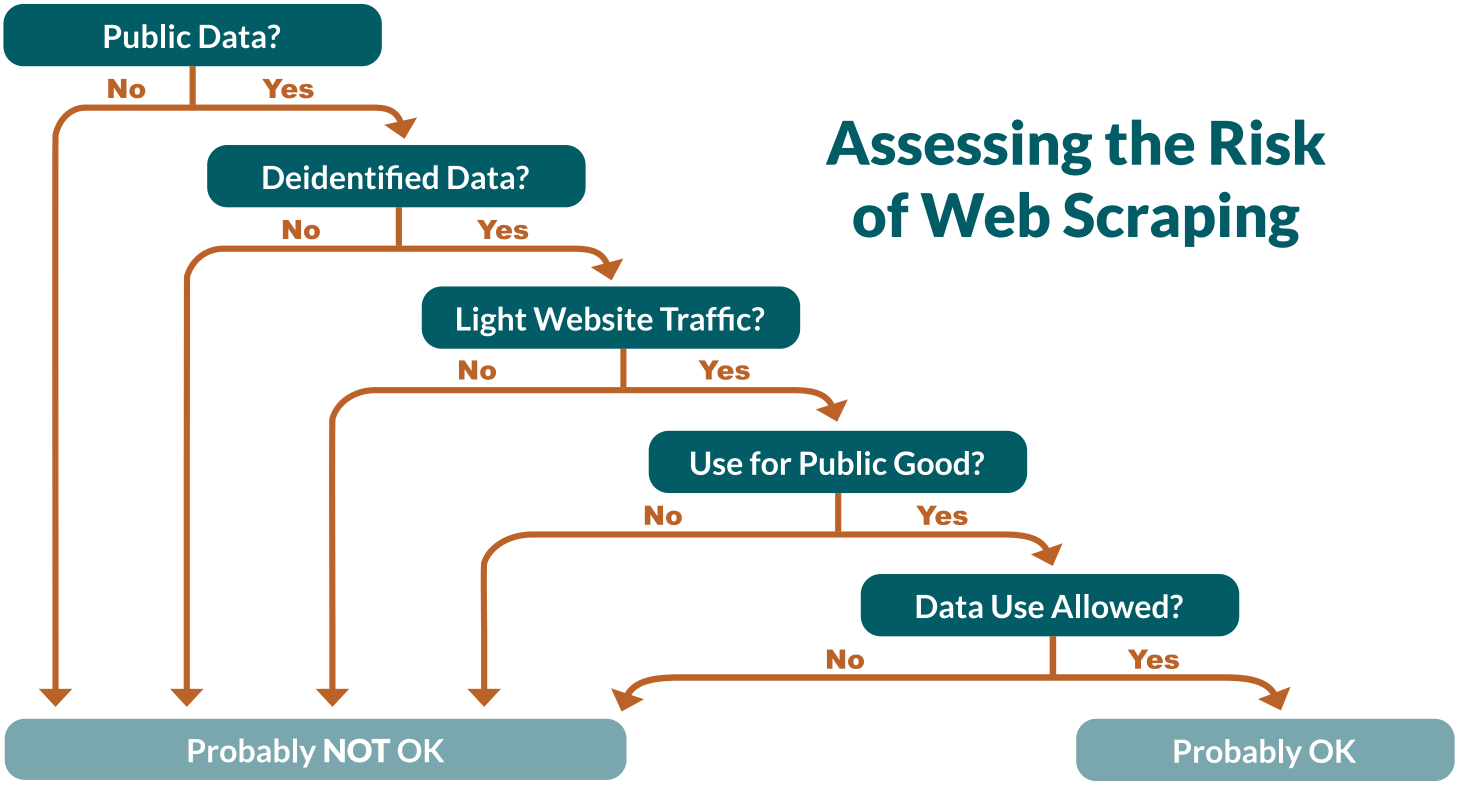
Are there situations in which web scraping is harmless?

Though harmless is a strong word, there are situations in which web scraping presents low risk and should be considered as a data collection tool. The factors influencing how risky web scraping is are: a) whether the data are proprietary or public; b) whether the data contain personally identifiable data or not; c) whether the website to be scraped has heavy or light data traffic; d) whether scraped data will be used for proprietary, commercial products or for research, public good purposes; e) whether the source code of the website prohibits or allows data scraping through the robots.txt hidden file guidelines. If all your answers to alternatives above is the latter option, then web scraping is probably okay; otherwise, it is probably not okay.

EXAMPLE

We had a publicly accessible website for public court records and were experiencing significant slowdowns and performance issues to the point we were getting complaints from the public and other agencies who also used it. We identified the root cause to be scraping of the data from a foreign country (India). While we came up with a more permanent solution, we prevented access to the site by blocking access from outside the US. We then implemented a CAPTCHA and URL masking to hide the case number or id so they couldn't just retrieve the case number from the URL to pull the next register of actions. For agencies who needed a solution without a CAPTCHA, we ended up creating a separate application that required a username and password to log in.

Assessing the Risk of Web Scraping



Public Data?

No

Yes

Deidentified Data?

No

Yes

Light Website Traffic?

No

Yes

Use for Public Good?

No

Yes

Data Use Allowed?

No

Yes

Probably **NOT** OK

Probably OK

Do you have suggestions of topics to cover in upcoming Data Dives? Please send them to **aassumpcao@ncsc.org**.

Learn more at

ncsc.org/datadives