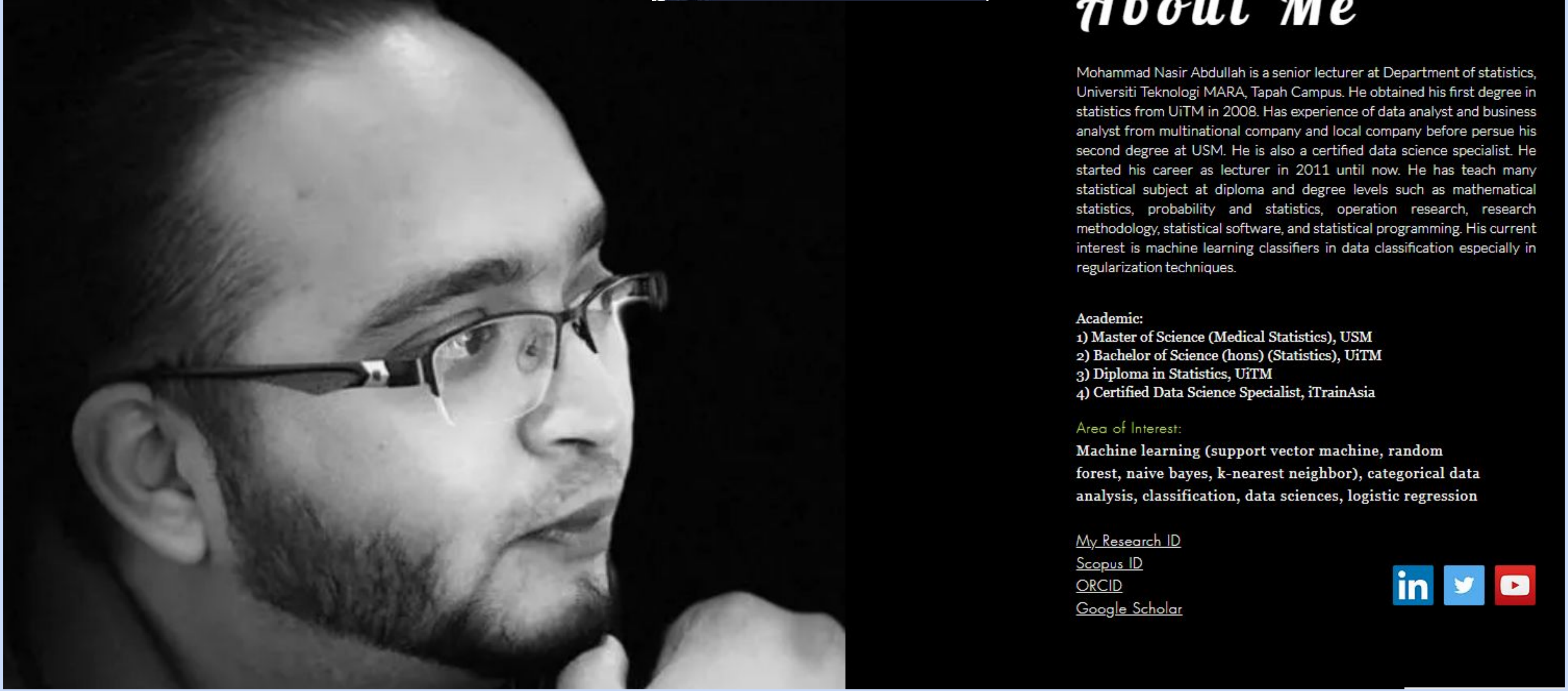
اونيۏرسيتي تيكنولوڬي مارا
UNIVERSITI
TEKNOLOGI
MARA

Carving for online data - the art of web scraping!

Mohammad Nasir Abdullah
Senior Lecturer
Department of Statistics,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA,
Tapah Campus, Perak.

https://nasirdrive1.wixsite.com/nasir916/

Mohammad Nasir Abdullah



About Me

Mohammad Nasir Abdullah is a senior lecturer at Department of statistics, Universiti Teknologi MARA, Tapah Campus. He obtained his first degree in statistics from UiTM in 2008. Has experience of data analyst and business analyst from multinational company and local company before persue his second degree at USM. He is also a certified data science specialist. He started his career as lecturer in 2011 until now. He has teach many statistical subject at diploma and degree levels such as mathematical statistics, probability and statistics, operation research, research methodology, statistical software, and statistical programming. His current interest is machine learning classifiers in data classification especially in regularization techniques.

Academic:
1) Master of Science (Medical Statistics), USM
2) Bachelor of Science (hons) (Statistics), UiTM
3) Diploma in Statistics, UiTM
4) Certified Data Science Specialist, iTrainAsia

Area of Interest:

Machine learning (support vector machine, random forest, naive bayes, k-nearest neighbor), categorical data analysis, classification, data sciences, logistic regression

My Research ID
Scopus ID
ORCID
Google Scholar

Introduction (1)

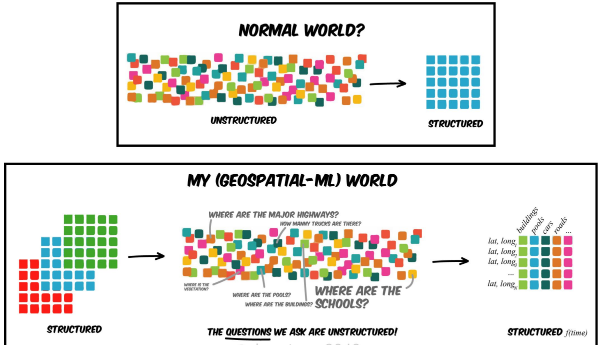
- Data and information on the web is growing exponentially.
- All of us today use Google as our 1st source of knowledge (including finding reviews about a place to understand a new term).
- With the amount of data available over the web, it opens new horizons of possibility for a data scientist.
- Now days, all the data are available on the internet, the only thing limiting us from using it is the ability to access it.

Introduction (2)

- If we wanted to access the information from the web, either:
  - Use whatever format the website uses
  - Copy and paste the information manually into a new document

This process can be very tedious when we want to extract a lot of information from a website!

What is web scraping?



A technique for converting the data present in unstructured format (HTML tags) over the web to the structured format which can easily be accessed and used.

Simply refers to the extraction of data from a website. This information is collected and then export into a format that is more useful to the user.

Example of web scraping

1. Export a list of product names and prices from Amazon onto an Excel spreadsheet.
2. Exporting stock prices to make better investing decisions.
3. Exporting data from YellowPages to generate leads.
4. Exporting data from store locators to create a list of business locations.
5. Exporting data from ecommerce sites for competitor analysis.

The list of what you can do with web scraping is almost EndLess! - After all, it about what you can do with the data you have collected and how valuable you can make it!

Why should learn web scraping?

1. Automate tasks. (Saving Time)
   a. Eg: monitoring data from different website (revenue gain from youtube, course dashboard, affiliate dashboard).
2. Getting Data. (Saving Time)
   a. Eg: Searching for article to read.
3. Testing Apps
   a. Eg: testing frameworks on apps, testing the landing page of the apps.
4. Actually Fun
   a. Feel “cool” once we reach the end results.

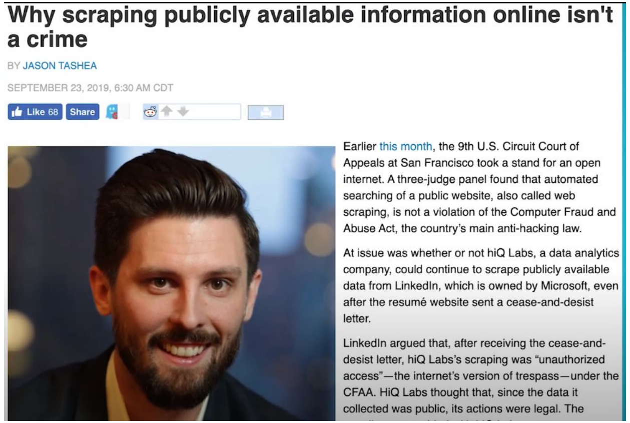
Is Web Scraping Legal?

Only for publicly available data!

1. User has made the data public.
2. No account required for access.
3. Not blocked by robots.txt file.

From Official API, (paid or free).

I am not a lawyer, this is not a legal advice!



Why scraping publicly available information online isn't a crime

BY JASON TASHEA

SEPTEMBER 23, 2019, 6:30 AM CDT

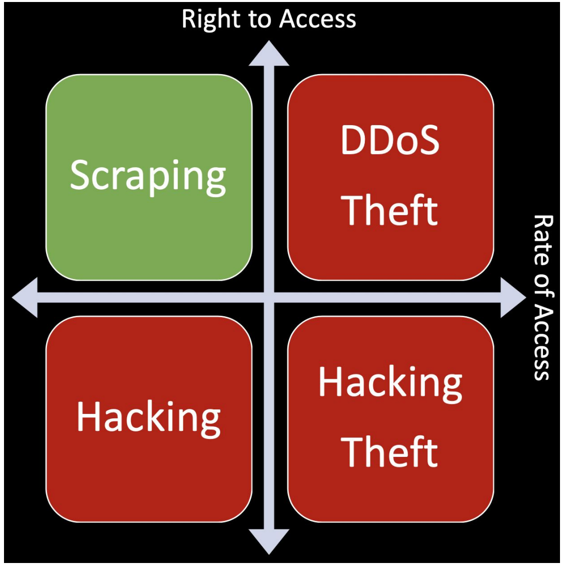
Earlier this month, the 9th U.S. Circuit Court of Appeals at San Francisco took a stand for an open internet. A three-judge panel found that automated searching of a public website, also called web scraping, is not a violation of the Computer Fraud and Abuse Act, the country's main anti-hacking law.

At issue was whether or not hiQ Labs, a data analytics company, could continue to scrape publicly available data from LinkedIn, which is owned by Microsoft, even after the resumé website sent a cease-and-desist letter.

LinkedIn argued that, after receiving the cease-and-desist letter, hiQ Labs's scraping was "unauthorized access"—the internet's version of trespass—under the CFAA. HiQ Labs thought that, since the data it collected was public, its actions were legal. The

We should read the term of service of the website. Still it not make web scraping illegal. haha..

Illegal cousins of data scraping



DDoS - Distributed Denial of Service (public data)

Theft - when you are accessing has financial value.

Hacking - you have gone way beyond recording data, try to reverse engineering the website.

Suggested guidelines to follow:

1. Use a conservative crawl rate. - don't hit servers to frequently.
2. Use an API if provided.
3. Don’t violate the terms of use of the site.
4. If scraping Public Data, Nothing to worry!
5. Content not under copyright?
6. Don’t gather sensitive user information.
7. Don’t re-publish the data that you scrape, without verifying the license

Example Legal Web Scraping

YouTube
Data API v3

Any API provider that allow you to scrap the data, and follow limitation of API.

Scrapy
WEB SCRAPING, CRAWLING
AND SPIDERS

Web crawling - this is google and any search engine do.

Ways to scrape data

- **Human Copy-Paste:**
    - A slow and efficient way of scraping data from the web. Involves human themselves analysing and copying the data to local storage.
- **Text pattern matching:**
    - Powerful approach to extract information from the web is by using regular expression matching facilities of programming languages. (click here for more information)
- (Application Programming Interface) **API Interface:**
    - Many website like Facebook, Twitter, LinkedIn, Instagram, Etc, provides public and private APIs which can be called using standard code for retrieving the data in prescribed format.
- (Document Object Model) **DOM Parsing:**
    - By using web browsers, programs can retrieve the dynamic content generated by client-side scripts. It is also possible to parse web pages into a DOM tree, based on which programs can retrieve parts of these pages.

Pre-requisites (the easiest way)

1) rvest package
2) selector gadget (https://selectorgadget.com/) - install it to your chrome extension

   https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoemdmbfginb

Let's try to scrape data from website

1) From outbreak.my - to scrape latest covid 19 data based on state
2) From isaham.my - to screen uptrend stock and find the best stock to entry
3) Twitter - using API
4) Pubmed, rplos, scopus database - using API

Thank you