Web Scraping 101:
How You Can Benefit from The Internet

By: Erika Siregar - Cofounder of R-Ladies Jakarta
Presented at Bittalk, 10-10-2020

Hello, my name is Erika :)

- **Education:**
  - Bachelor of Applied Science from STIS
  - Master in Computer Science from Old Dominion University, US
- **What I am doing now**:
  - R-Ladies Jakarta : Cofounder
  - Jakarta Machine Learning: Head of Program
  - BPS: Data scientist, big data engineer and analyst
- **Connect with me:**
  - Email: erika.mukhlisina@gmail.com
  - GitHub: https://github.com/erikaris
  - Twitter: @erikaris
  - IG: @erikaris15

First, Let Me Introduce You to R-Ladies Jakarta

- komunitas **belajar bersama** untuk **perempuan dan gender minorities** yang ingin **meningkatkan kemampuan** dalam bahasa R maupun yang **baru mau mulai belajar R.**

- Worldwide organization → **part of R-Ladies Global** (https://rladies.org/)

Still About R-Ladies Jakarta

Goals:
promotes gender diversity in the R community via **meetups and mentorship** in a friendly and safe environment**.**

What do We Do in a Meetup?

- 15-mins Intro to R
- Delivering material, covering different topic each meetup.
- Hands-on + QnA
- Networking and mingling

Why You should Join R-Ladies Jakarta

Next Meetup: Oct 24th, 2020
Speaker: DSI

Why you should join R-Ladies?

- Welcomes members of all R proficiency levels.
  (it's **OK** to be a **newbie,** we'll help you with the installation)

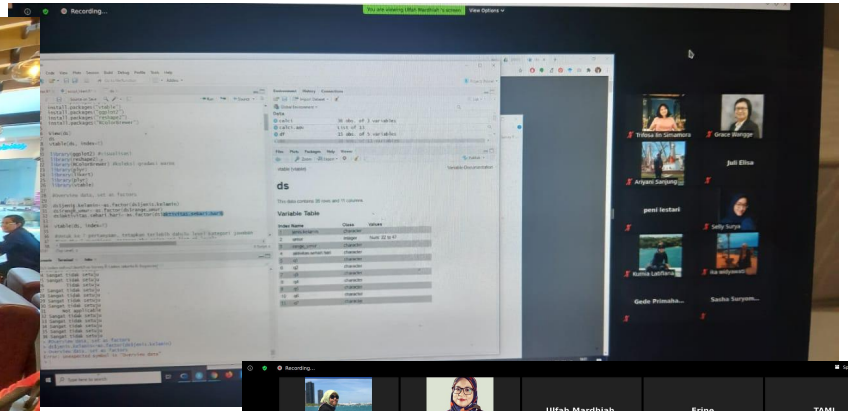
- Warm and friendly environment.

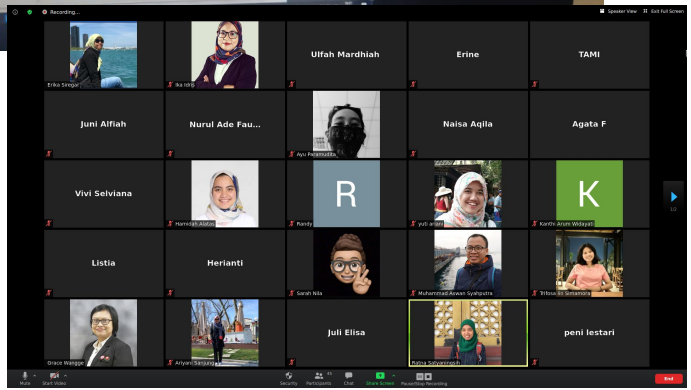
- No need to feel insecure.

- Konsepnya **bukan guru dan murid**, tapi **belajar, explore, dan mencoba scripting bersama**.

How Our Meetups Look Like

Still about our Meetups

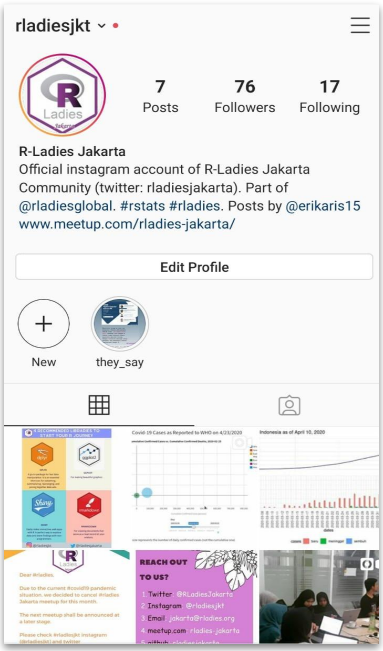




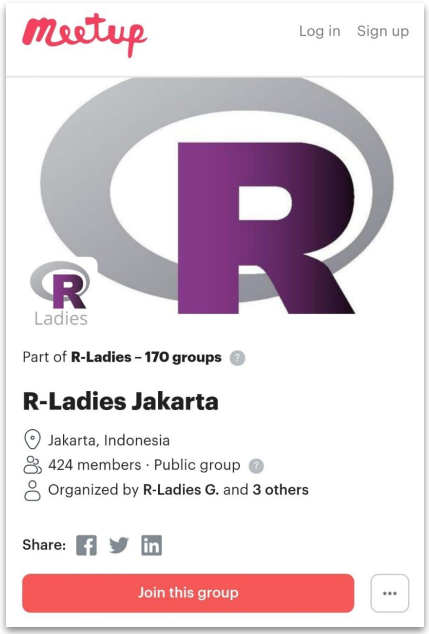
More about R-Ladies Jakarta?

Email: jakarta@rladies.org | **Whatsapp Group** | **#rladiesjakarta #rladies #rstats**

@rladiesjakarta



@rladiesjkt

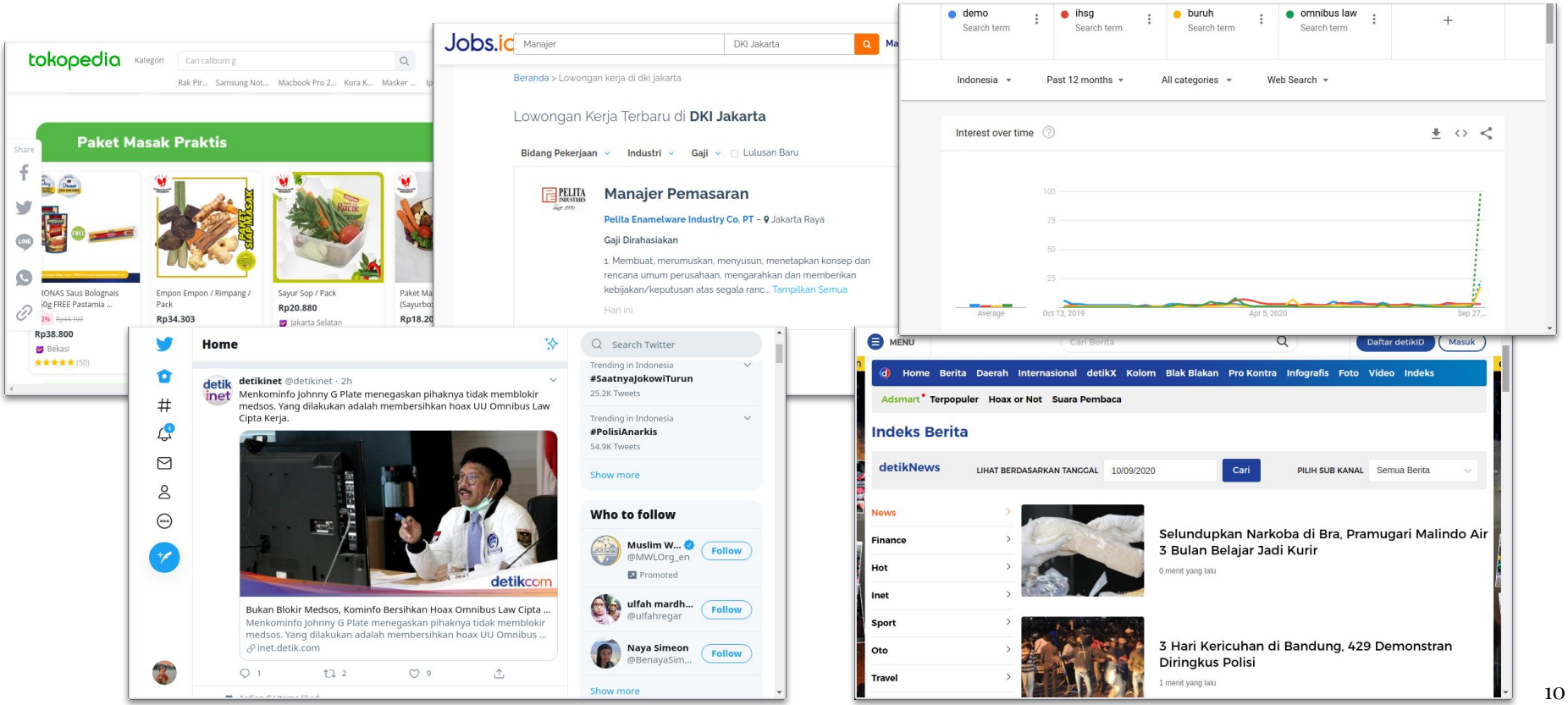


https://meetup.com/rladies-jakarta

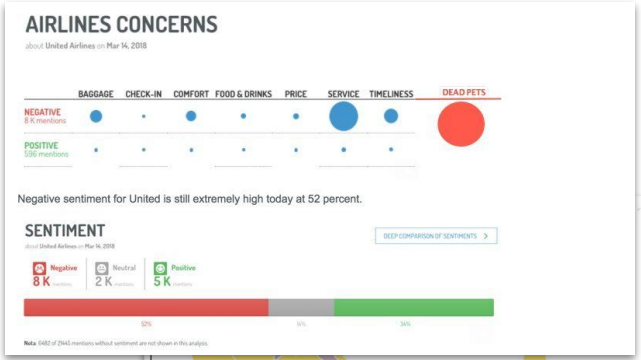
@rladiesjakarta

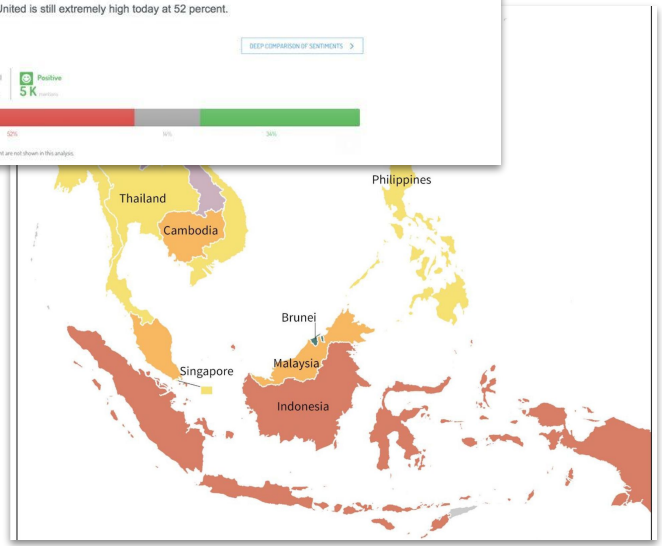
Web Scraping

The internet offers an abundance of information



don't just browsing it, scraping it





Why not take advantage from that abundance of data?

- Data-driven decision??
- Further analysis??
- dashboard??

What is Web Scraping

- Extracting data/information from a website and converting it into a format of your choice (HTML, JSON, CSV, etc.)
- Similar to manual copy and paste, but in a smarter way.
- Scraping the web is basically imitating human actions through a lines of script. You'll see why and how later.

Why do we scrape a web?

In 2020, the “ digital universe “ holds an estimated 40 trillion gigabytes or 40 zettabytes worth of information.

https://medium.com/@octoparsewebscraping/web-scraping-in-the-big-data-solution-7d2804d41477

- Easiest way to benefit from free available source of information
- Automate data collection from website (no copy and paste)
- effectively reduce manual work and the operation cost
- Speediness
- Get clean and structured data

Contoh-contoh informasi yang bisa kita peroleh dari webscraping

1. Hotel and Restaurant
2. Flight
3. E-commerce
4. Saham ↠ idx
5. Job vacancy
6. Car listing
7. Housing listing
8. Reviews listing
9. Social media
10. News website ↠ Kompas, detik, Liputan6, etc.

1. Monitoring price, etc.
2. Comparison
3. See trend
4. Sentiment Analysis
5. Reviews analysis
6. News monitoring
7. General business information
8. make more informed decisions

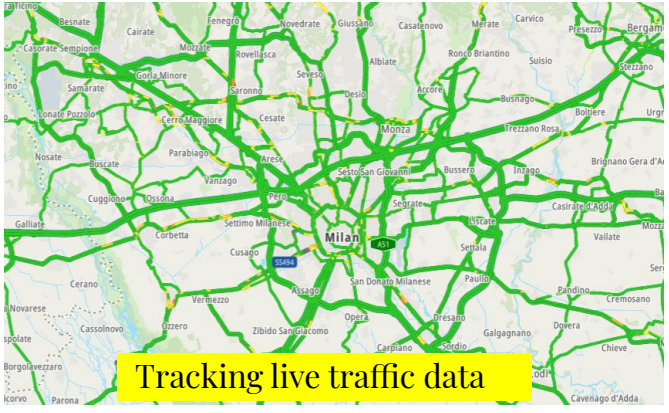
Web Scraping in The Time of Pandemic

Collecting all the things that can help us to fight the pandemic

- Tracking people's mobility
- Tracking people's sentiment towards the cases
- Gathering news about pandemic.
- Unemployment
- Poverty

Tracking number of cases

become the hub of gathering information about Covid 19

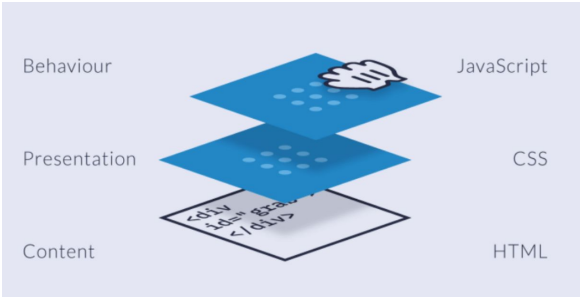


Tracking live traffic data

Things You Must Know before Begin Scraping a Website

I. Various Types of Website:
   A. Static
   B. Dynamic (lazy load)

II. Have a basic HTML knowledge.
   A. Components of a Website
   B. Document Object Model
   C. id, class, selector, xpath, dll.



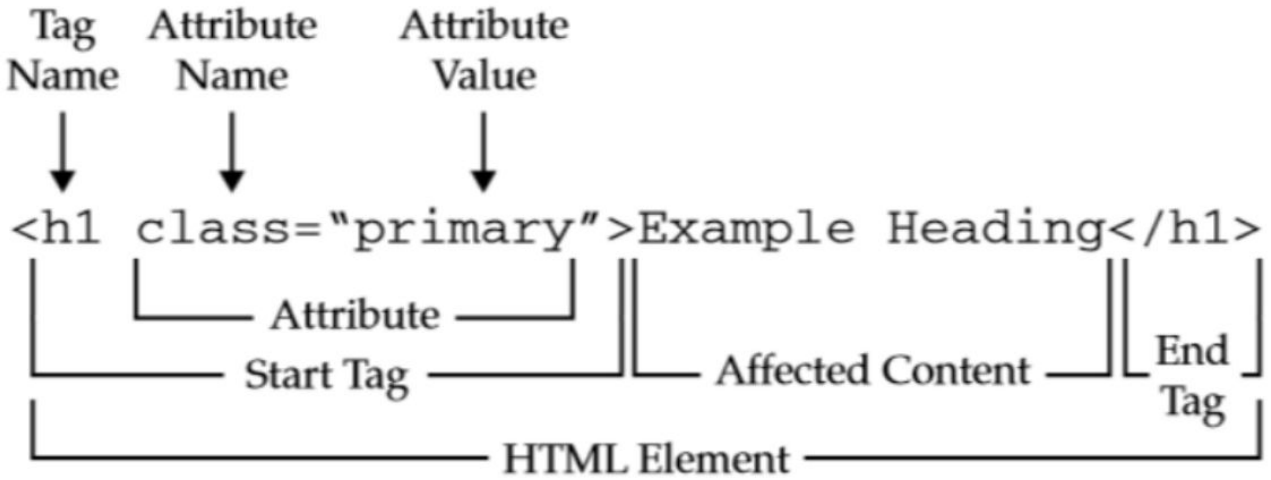
Things You Must Know before Begin Scraping a Website

IV. The web code and design can change anytime.

V. Be mindful in maintaining the number of requests

VI. If there is an API, use it

HTML Elements

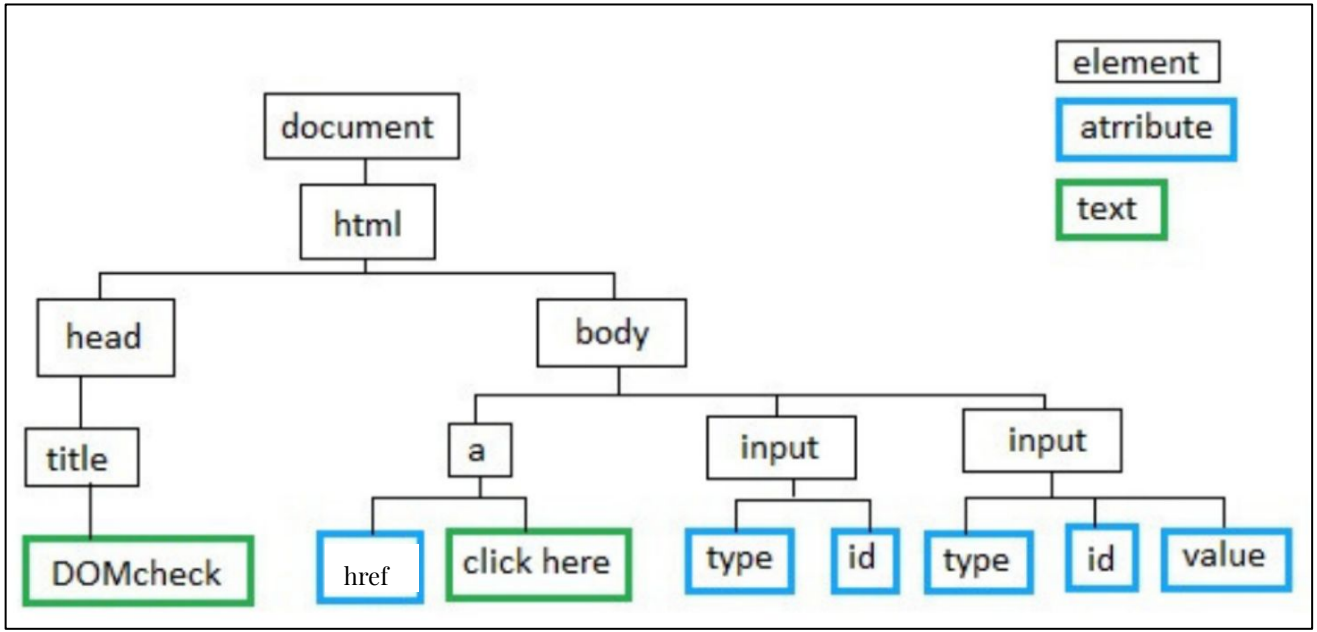


HTML Tags

Tag	Description
<html> ... </html>	Declares the Web page to be written in HTML
<head> ... </head>	Delimits the page's head
<title> ... </title>	Defines the title (not displayed on the page)
<body> ... </body>	Delimits the page's body
<h *n*> ... </h*n*>	Delimits a level *n* heading
...	Set ... in boldface
...	Set ... in italics
...	Center ... on the page horizontally
...	Brackets an unordered (bulleted) list
...	Brackets a numbered list
...	Brackets an item in an ordered or numbered list
 	Forces a line break here
<p>	Starts a paragraph
<hr>	Inserts a horizontal rule
<img src="...">	Displays an image here
 ... 	Defines a hyperlink

https://www.w3schools.com/TAGs/

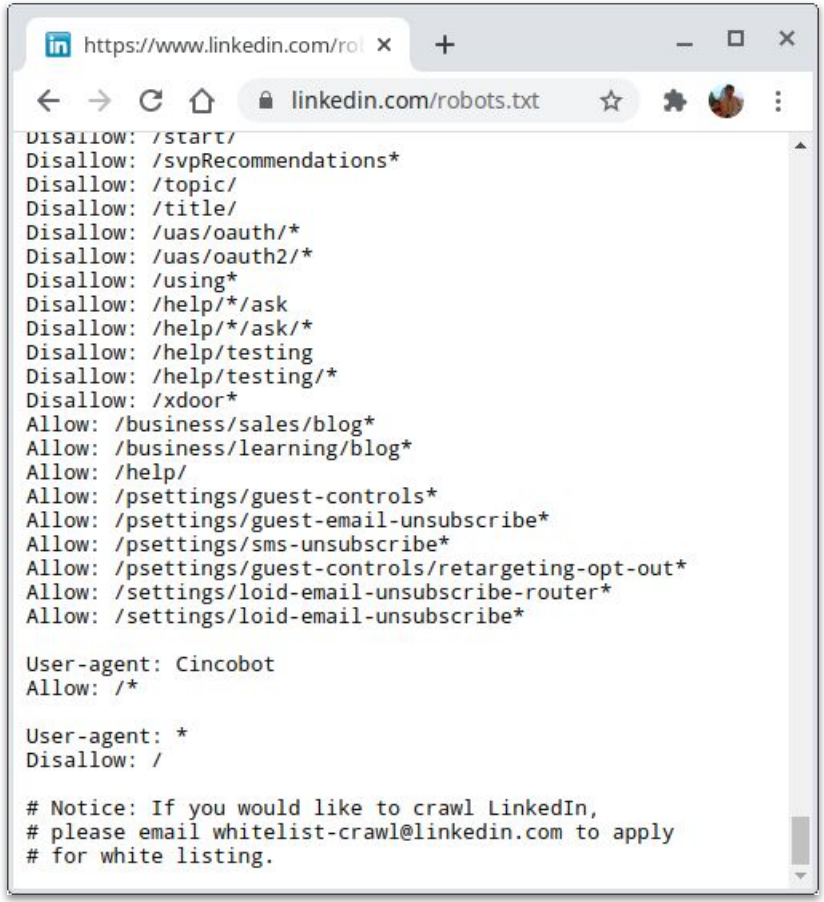
Document Object Model (DOM)



Ethic in Web Scraping

respecting robots

DISCLAIMER: The following is intended for the Big Data researchers who comply with the permissions from robots.txt, set the correct User Agent and do not violate the Terms of Service of the sites they scrape.



The Complexity of Web Scraping

1. **A lucky loiterer** ↠ low ↠ example: **wikipedia**
   a. The web pages you need to scrape have simple and clean markup without any JS.
   b. All the URLs to other websites and pages are direct
2. **A skilled professional** ↠ medium ↠ example: kompas, detik
   a. Partial JS rendering.
   b. Simple pagination.
   c. Simple URL creation rules.
3. **A Jedi Knight, may the Force be with you** ↠ high
   a. The page is fully built with JS.
   b. The URLs are formed using JS
   c. CAPTCHA is present.
   d. The website has an underlying API with complex rules of data transfer

1. paging
2. Javascript-heavy websites

https://towardsdatascience.com/big-data-what-is-web-scraping-and-how-to-use-it-74e7e8b58fd6

Pipeline

1. Explore website ↠ susunannya, komponen penyusunnya, behaviour, interaksi

2. Menentukan komponen yang akan di-scrape ↠ Remember DOM

3. Menentukan tools yang akan dipakai
    a. Non-scripting tools: e.g. Kofax Kapow, Octoparse, etc. ↠ limited GUI options, paid.
    b. **Scripting tools** ↠ self-made, customizable, free, communities
        i. Python ↠ Scrapy, beautiful soup,
        ii. R ↠ Rvest
        iii. Selenium ↠ Python Selenium, RSelenium

Pipeline (2)

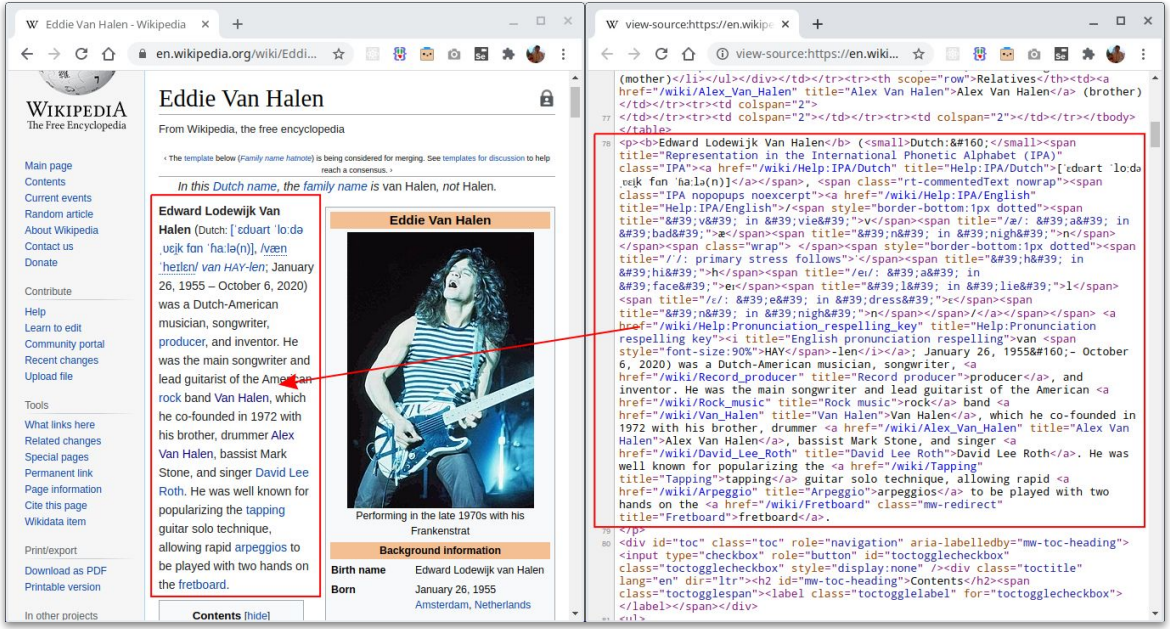
4. Develop script
   a. Send a “GET” request to the target website, and then parse the HTML accordingly.
   b. Fetch and parsing
   c. **Trial and error** ↠ reinspecting the web structures

5. Store the result ↠ file, database.

6. Further analysis ↠ visualisasi, sentiment analysis, dll.

Scrape a Static Webpage

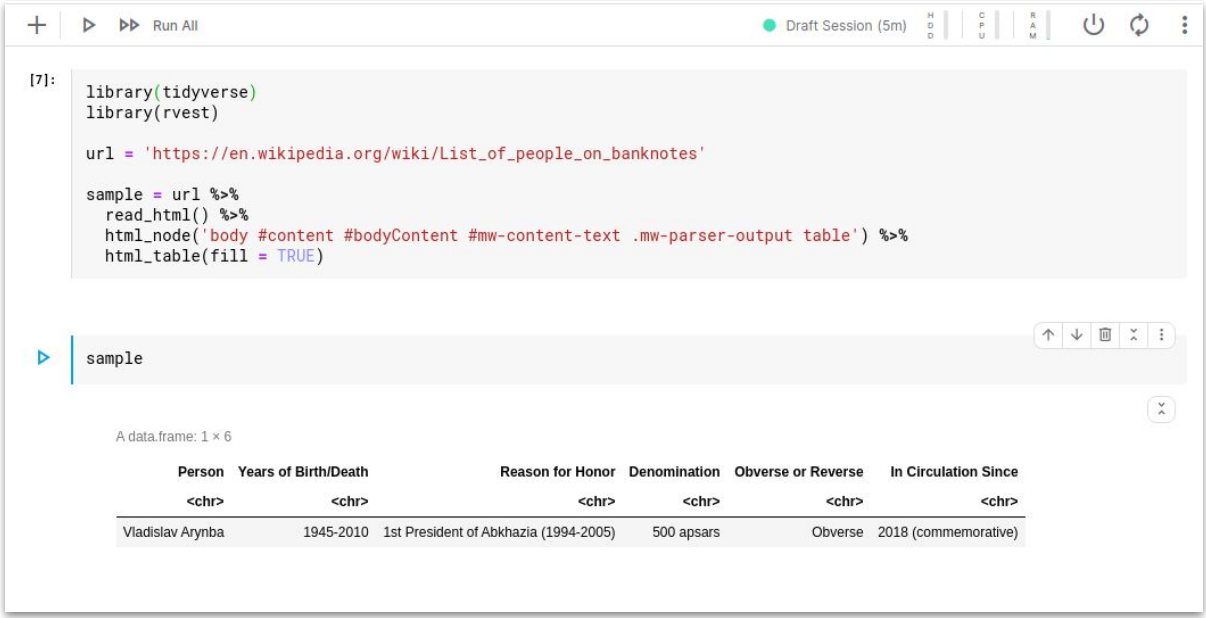
Beberapa website menampilkan halaman web persis seperti sumber yang diterima, seperti wikipedia.org, detik.com, kompas.com.



Web Scrape a Static Webpage (contd)

Kita dapat memparsing beberapa bagian konten web, misalnya judul, tanggal, image, related link, atau bahkan sebuah table.

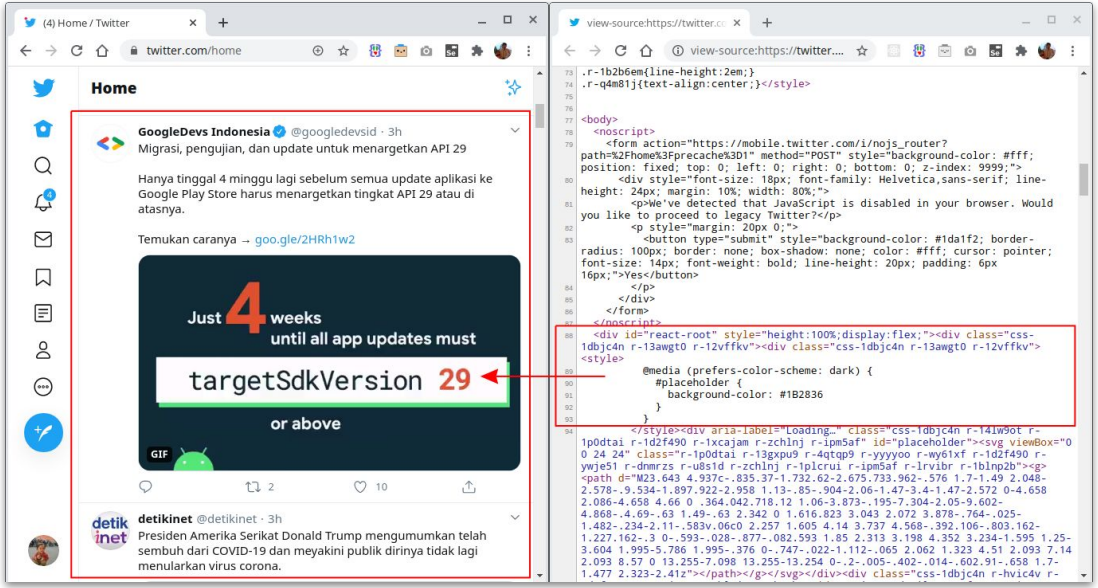
Beberapa contoh tools yang dapat digunakan adalah scrapy dan rvest.



Web Scrape a Dynamic Web

Beberapa website hanya menampilkan root element saja,

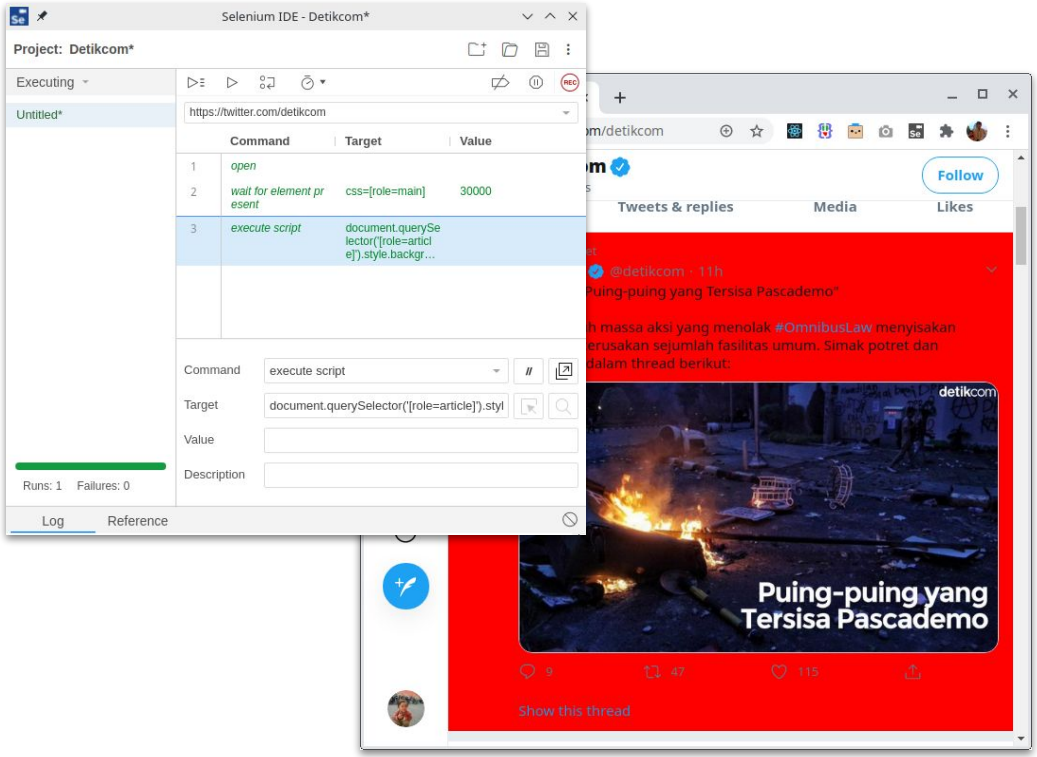
sementara konten-nya ditampilkan kemudian secara lazy-loaded dengan memanfaatkan XHR atau ajax. Beberapa contoh website yang bersifat dinamis adalah twitter dan instagram.



Web Scrape a Dynamic Web (contd)

Untuk mengambil konten dari website yang bersifat dinamis, kita perlu sebuah tools yang dapat berinteraksi dengan browser melalui, i.e. **chromedriver** atau **geckodriver**.

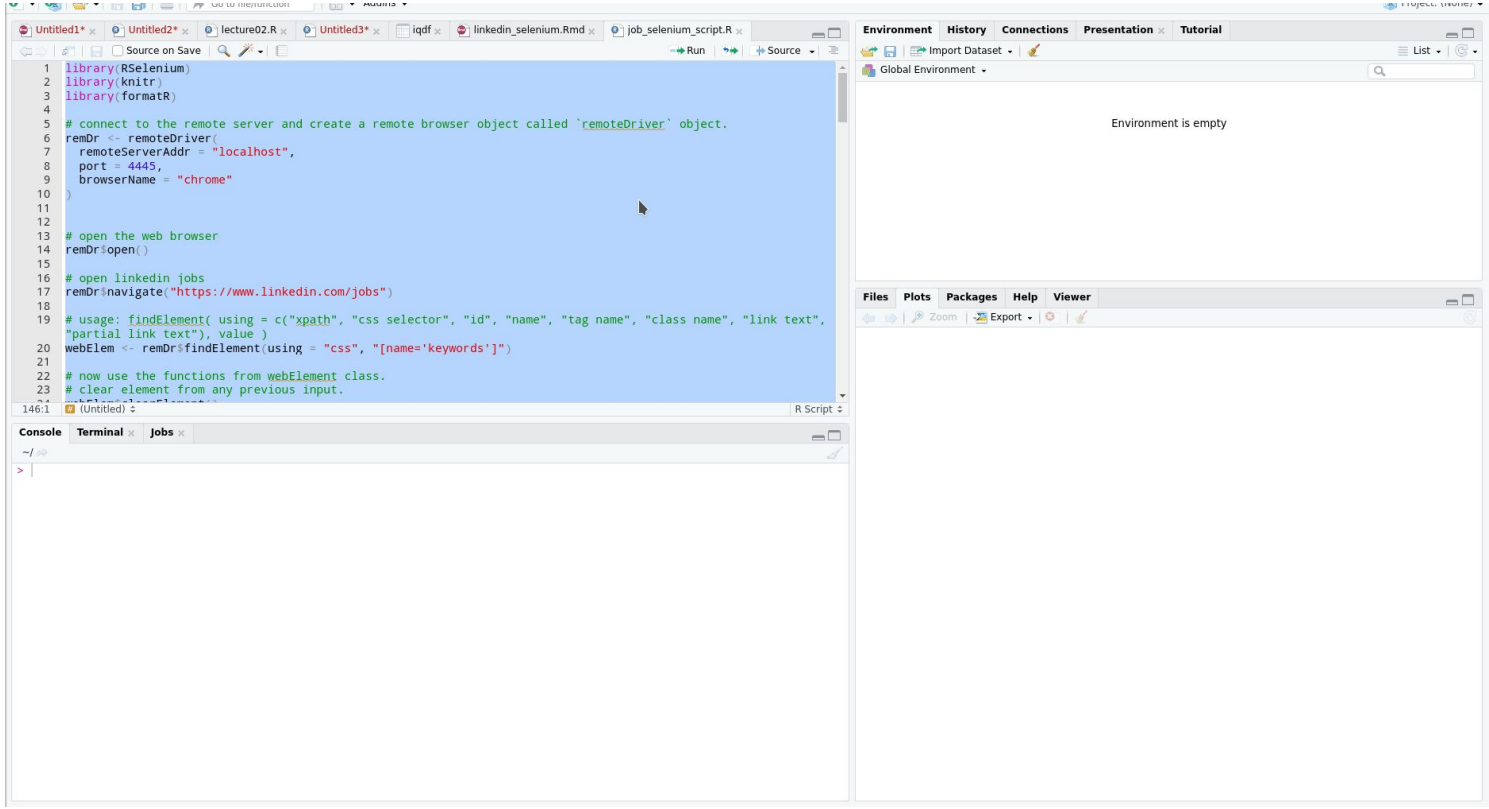
Salah satu tool yang dapat digunakan untuk keperluan tersebut adalah Selenium Webdriver.



Example of Simple Web Scraping (Static Web)

- Use R Library: rvest, tidyverse
- Data source: https://apps.who.int/bloodproducts/snakeantivenoms/database/SnakeAntivenomListFrm.aspx?@CountryID=23
- Script: https://github.com/erikaris/bittalk/blob/main/basic_scraping.R

Demo of Linkedin Selenium



Thank You