

# 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](mailto: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

## 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.**

Next Meetup: Oct 24th, 2020  
Speaker: DSI



# How Our Meetups Look Like



# Still about our Meetups



A composite image showing a video conference interface and a data analysis tool. The top right shows a video conference with multiple participants in a grid. The bottom half shows a screenshot of a software interface with a code editor, a variable table, and a variable distribution plot. The code editor contains R-like syntax related to data processing and visualization.



# More about R-Ladies Jakarta?

Email: [jakarta@rladies.org](mailto:jakarta@rladies.org) | Whatsapp Group | #rladiesjakarta #rladies #rstats

**R-Ladies Jakarta**  
@RLadiesJakarta  
Part of a worldwide organization promoting gender diversity in the R community. #rstats #rladies. tweets by @erikaris15

④ Jakarta Capital Region  
🔗 [meetup.com/r ladies-jakart...](https://meetup.com/r ladies-jakart...)  
Joined July 2019  
50 Following 151 Followers

Tweets Tweets & replies Media Likes

R-Ladies Jakarta @RLadies... · 23 Apr · Hi #rladies, how's your #workfromhome going?  
Let's keep improving your #R skill by learning new fun things in R. Let's spend some time to recreate a bubble chart that illustrates global #COVID19 cases reported to @WHO on 4/23/2020. Check [instagram.com/p/B\\_U1BfRDrot/](https://instagram.com/p/B_U1BfRDrot/) #rstats

**rladiesjkt** • 7 Posts 76 Followers 17 Following

**R-Ladies Jakarta**  
Official instagram account of R-Ladies Jakarta Community (twitter: @rladiesjakarta). Part of @rladiesglobal. #rstats #rladies. Posts by @erikaris15

www.meetup.com/rladies-jakarta/

Edit Profile New they\_say

R-Ladies Jakarta

Covid-19 Cases as Reported to WHO on 4/23/2020 Indonesia as of April 10, 2020

REACH OUT TO US!  
Twitter: @RLadiesJakarta  
Zulip: [rladies.global](https://rladies.global)  
(Email): [rladies@rladies.org](mailto:rladies@rladies.org)  
Meetup: <https://meetup.com/rladies-jakarta/>

Log in Sign up

**meetup**

**R-Ladies Jakarta**

Part of R-Ladies – 170 groups

**R-Ladies Jakarta**

Jakarta, Indonesia  
424 members · Public group  
Organized by R-Ladies G. and 3 others

Share: [Facebook](#) [Twitter](#) [LinkedIn](#)

Join this group

twitter @RLadiesJakarta

**R-Ladies Jakarta**  
RLadiesJakarta

Unfollow

@rladiesjakarta

@rladiesjkt

<https://meetup.com/rladies-jakarta>

@rladiesjakarta

# Web Scraping

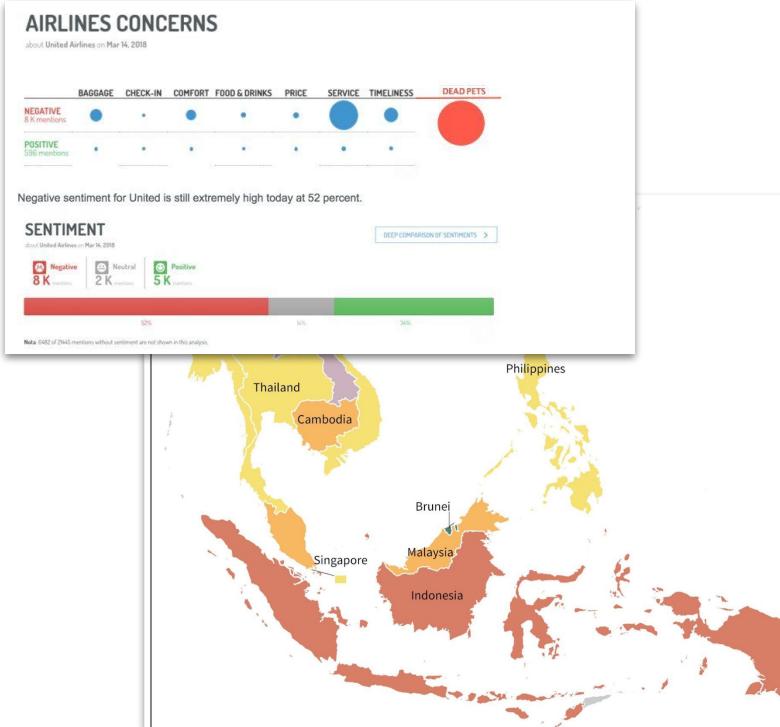
---

# The internet offers an abundance of information

The collage consists of five screenshots:

- tokopedia**: A screenshot of the Tokopedia marketplace showing a product category for "Paket Masak Praktis". It displays several food items with their names and prices: IONAS Saus Bolognais (Rp38.800), Empor Empon / Rimpang / Pack (Rp34.303), Sayur Sop / Pack (Rp20.880), and Paket Ma (Sayurbot) (Rp18.20).
- Jobs.id**: A screenshot of the Jobs.id job search platform for DKI Jakarta. It shows a search result for "Manajer Pemasaran" at Pelita Enamelware Industry Co. PT. The listing includes a company logo, address (Jakarta Raya), and a brief description about creating and implementing marketing concepts.
- Twitter**: A screenshot of a Twitter search for "#SaatnyaJokowiTurun" and "#PolisiAnarkis", both trending in Indonesia. It shows user profiles and tweets from users like @detikinet and @detikcom.
- detik.com**: A screenshot of the detik.com news website's home page. It features a video thumbnail of Johnny G Plate speaking, a headline about the Omnibus Law, and a section titled "Who to follow" with profiles of Muslim W... (@MWLOrg\_en) and others.
- Google Trends**: A screenshot of the Google Trends interface comparing search interest over time for terms like "demo", "ihsg", "buruh", and "omnibus law". The chart shows a significant spike in interest for "omnibus law" starting around September 27, 2020.
- detikNews**: A screenshot of the detikNews news channel on the detik.com platform. It lists categories like News, Finance, Hot, Inet, Sport, Oto, and Travel, along with headlines such as "Selundupkan Narkoba di Bra, Pramugari Malindo Air 3 Bulan Belajar Jadi Kurir" and "3 Hari Kericuhan di Bandung, 429 Demonstran Diringkus Polisi".

# 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.

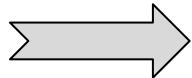
<https://medium.com/@octoparsewebscraping/web-scraping-in-the-big-data-solution-7d2804d41477>



- 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

5 results found in 2ms

- COVID-19 cases & deaths from CDC**  
Returns the total cases, deaths in US from CDC website
- COVID-19 news**  
Latest news across the world about COVID-19 situations crawled every hour
- COVID-19 cases by US states**  
Returns all United States of America and their Corona data.
- COVID-19 cases by countries**   
Returns data of all countries that has COVID-19.
- COVID-19 all cases**  
Returns all total cases, recovery, and deaths.

become the hub of gathering information  
about Covid 19



# Things You Must Know before Begin Scraping a Website

## I. Various Types of Website:

- A. Static
- B. Dynamic (lazy load)

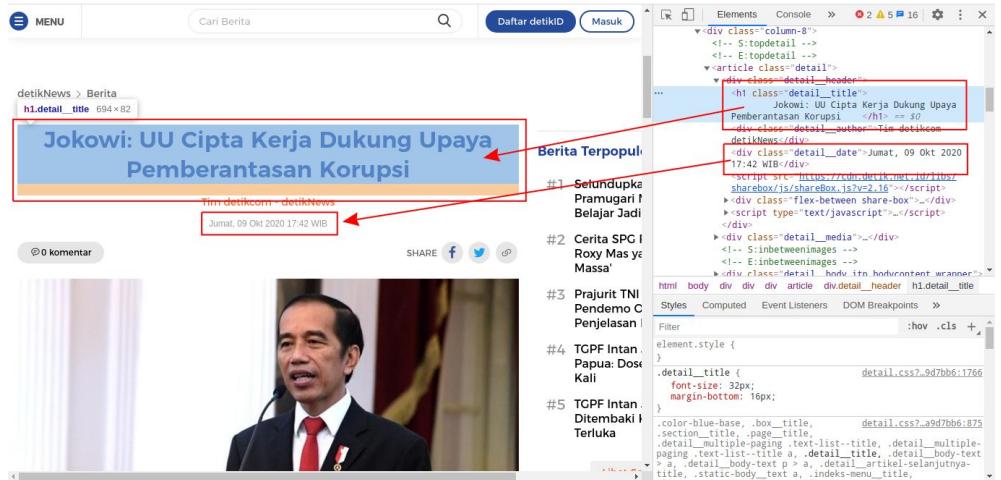
## II. Have a basic HTML knowledge.

- A. Components of a Website
- B. Document Object Model
- C. id, class, selector, xpath, dll.

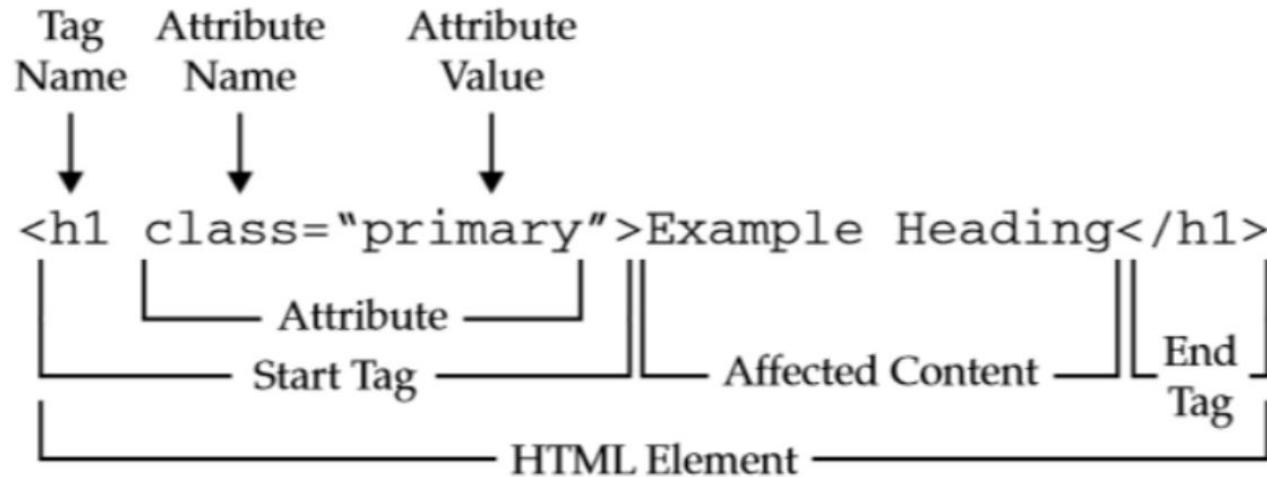


# Things You Must Know before Begin Scraping a Website

- 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

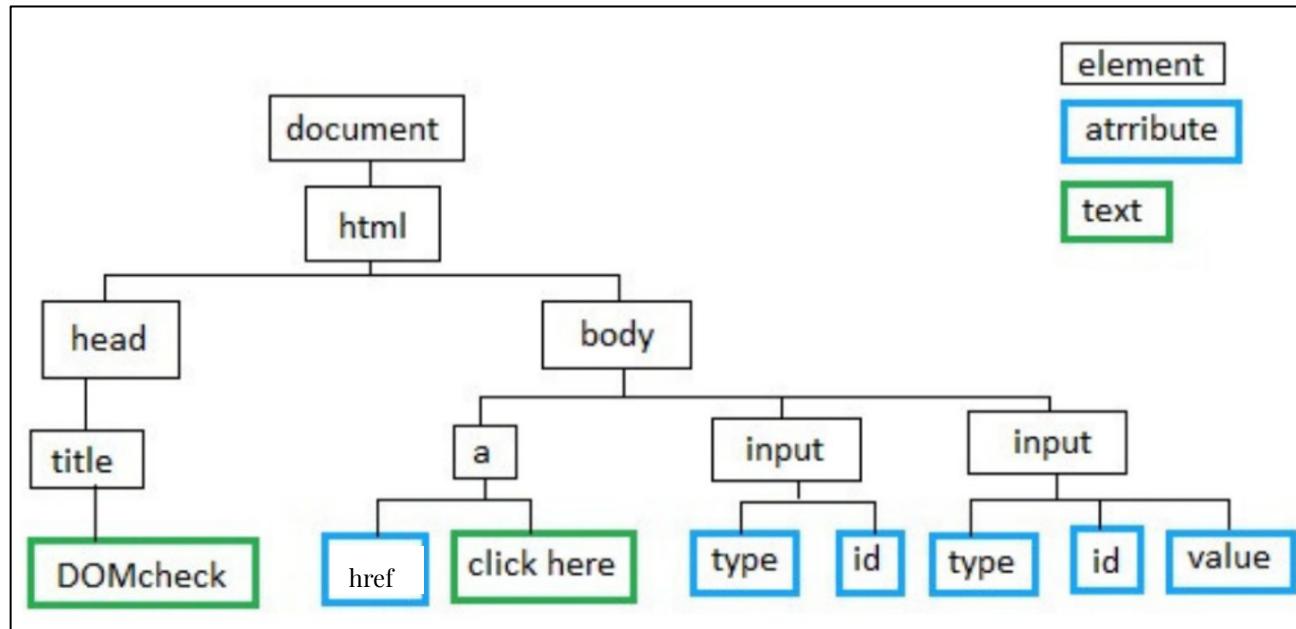


# HTML Tags

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
<b> ... </b>	Set ... in boldface
<i> ... </i>	Set ... in italics
<center> ... </center>	Center ... on the page horizontally
<ul> ... </ul>	Brackets an unordered (bulleted) list
<ol> ... </ol>	Brackets a numbered list
<li> ... </li>	Brackets an item in an ordered or numbered list
 	Forces a line break here
<p>	Starts a paragraph
<hr>	Inserts a horizontal rule
	Displays an image here
<a href="..."> ... </a>	Defines a hyperlink

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

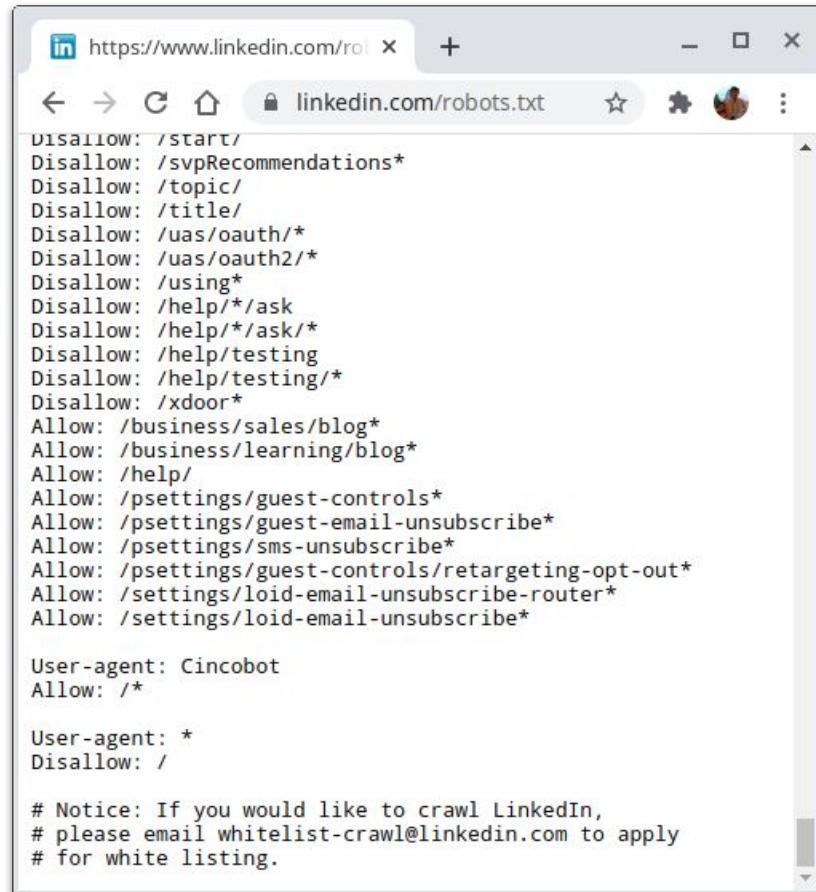
# Document Object Model (DOM)



# 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 screenshot shows a browser window with the URL <https://www.linkedin.com/robots.txt>. The page content displays the following text:

```
Disallow: /start/
Disallow: /svpRecommendations/*
Disallow: /topic/
Disallow: /title/
Disallow: /uas/oauth/*
Disallow: /uas/oauth2/*
Disallow: /using/*
Disallow: /help/*/ask
Disallow: /help/*/ask/*
Disallow: /help/testing
Disallow: /help/testing/*
Disallow: /xdoor*
Allow: /business/sales/blog/*
Allow: /business/learning/blog/*
Allow: /help/
Allow: /psettings/guest-controls/*
Allow: /psettings/guest-email-unsubscribe/*
Allow: /psettings/sms-unsubscribe/*
Allow: /psettings/guest-controls/retargeting-opt-out/*
Allow: /settings/loid-email-unsubscribe-router*
Allow: /settings/loid-email-unsubscribe*

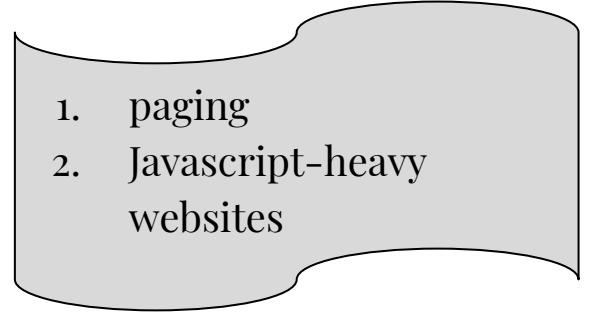
User-agent: Cincobot
Allow: /*

User-agent: *
Disallow: /

# Notice: If you would like to crawl LinkedIn,
# please email whitelist-crawl@linkedin.com to apply
# for white listing.
```

# 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

# 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.

The image shows a comparison between a browser's visual representation of a Wikipedia page and its underlying HTML code. On the left, a screenshot of a browser window displays the Wikipedia article for Eddie Van Halen. The page includes the title 'Eddie Van Halen', a sidebar with navigation links, the main content block with a bio about Edward Lodewijk Van Halen, and a 'Background information' section with details like birth date and place. A red arrow points from the 'fretboard' link in the bio section to the corresponding line of code in the browser's developer tools. On the right, a screenshot of the browser's 'view-source' tab shows the raw HTML code for the same page. The code is heavily annotated with red boxes highlighting specific sections: the bio block, the 'Background information' section, and the 'fretboard' link. The 'fretboard' link is specifically highlighted with a red box around the line of code: `<a href="/wiki/Fretboard" title="Fretboard">fretboard</a>". This demonstrates how web scraping tools can extract structured data from static HTML pages by identifying and extracting these annotated elements.`

# 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.



The screenshot shows a Jupyter Notebook interface with an R session. The code in cell [7] demonstrates how to extract data from a Wikipedia page about people on banknotes. The output cell displays a data frame named 'sample' containing information about Vladislav Arynba.

```
library(tidyverse)
library(rvest)

url = 'https://en.wikipedia.org/wiki/List_of_people_on_banknotes'

sample = url %>%
  read_html() %>%
  html_node('body #content #bodyContent #mw-content-text .mw-parser-output table') %>%
  html_table(fill = TRUE)
```

A data.frame: 1 × 6

Person	Years of Birth/Death	Reason for Honor	Denomination	Obverse or Reverse	In Circulation Since
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
Vladislav Arynba	1945-2010	1st President of Abkhazia (1994-2005)	500 apsars	Obverse	2018 (commemorative)

# Web Scrape a Dynamic Web

Beberapa website hanya menampilkan root element saja,

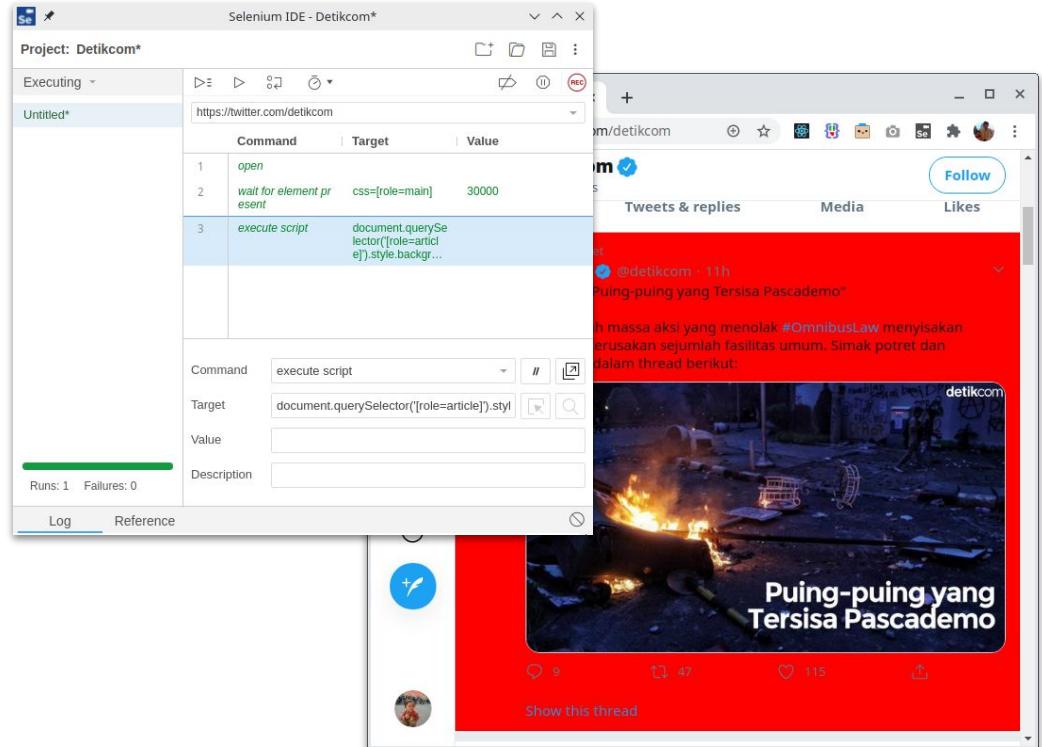
sementara konten-nya ditampilkan kemudian secara lazy-loaded dengan memanfaatkan XHR atau ajax.  
Beberapa contoh website yang bersifat dinamis adalah twitter dan instagram.

The screenshot shows a browser window with the Twitter homepage. A modal window from 'GoogleDevs Indonesia' is displayed, containing text about migrating to API 29 and a link to a guide. Below the modal, a dark-themed input field has the placeholder 'targetSdkVersion 29'. To the right, a separate window shows the raw HTML source code of the Twitter page, where the placeholder is also visible. A red box highlights the placeholder in the source code.

# 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](https://github.com/erikaris/bittalk/blob/main/basic_scraping.R)

# Demo of Linkedin Selenium

The screenshot shows the RStudio interface with the following details:

- Top Bar:** Shows multiple open files: Untitled1\*, Untitled2\*, lecture02.R, Untitled3\*, iqdf, linkedin\_selenium.Rmd, and job\_selenium\_script.R.
- Environment Tab:** Displays the message "Environment is empty".
- Code Editor (Left Panel):** Contains R code for setting up a remote Selenium driver and navigating to LinkedIn to search for jobs. The code includes imports for RSelenium, knitr, and formatR; connects to a local host on port 4445 using chrome; opens the browser; navigates to <https://www.linkedin.com/jobs>; finds elements using CSS selector "[name='keywords']"; and clears previous input.
- Console Tab:** Shows the command "R Script" and the path "~/".
- Terminal Tab:** Shows the command "Jobs".
- Bottom Status Bar:** Shows the number 146:1 and the file name "(Untitled) : R Script".

# Thank You