R-Ladies Jakarta 14th Meetup:

Intro to Web Scraping with R

Agenda

- Opening
- About R-Ladies Jakarata
- Web scraping 101: theory and concept
- Web Scraping: hands-on with R
- Closing (group picture)

The meeting will be recorded.

Please let us know if you have any concerns regarding this.

Hello, my name is Erika :)

Communities:

- Co-founder of R-Ladies Jakarta : **@rladiesjkt (IG)**
- Head of Program of Jakarta Machine Learning **@jkt.machinelearning (IG)**

Education:

- Bachelor of Applied Science from STIS
- Master in Computer Science from Old Dominion University, US

Work:

- BPS

Connect with me:

- Email: erika.mukhlisina@gmail.com
- IG: @erikaris15
- Linkedin: https://www.linkedin.com/in/erika-siregar

R-Ladies Team (Current Active Organizer)

Erika Siregar

Ulfah Mardhiah

Sinum Fariasi

Lutfia Nuzula

Grace Wangge

About R-Ladies Jakarta

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/)

Goals

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)

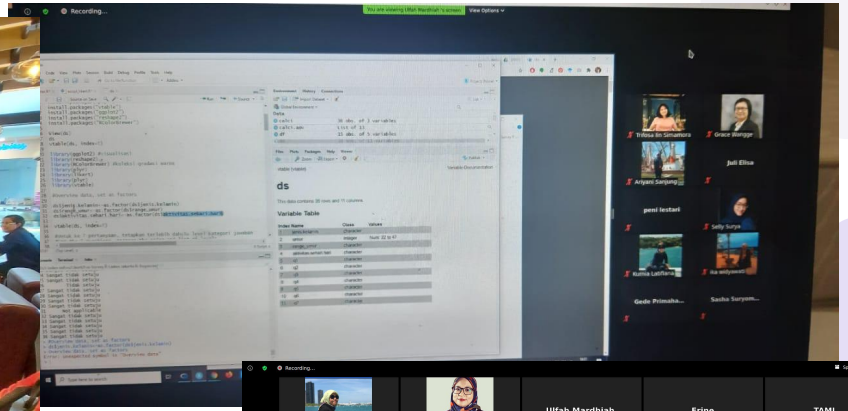
- Warm and friendly environment.

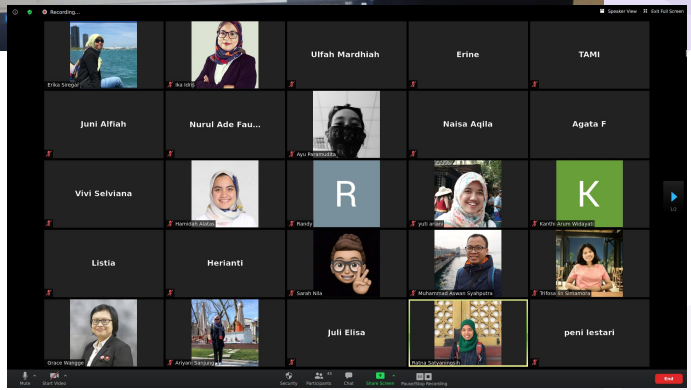
- No need to feel insecure.

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

How Our Meetups Look Like

Still about our Meetups

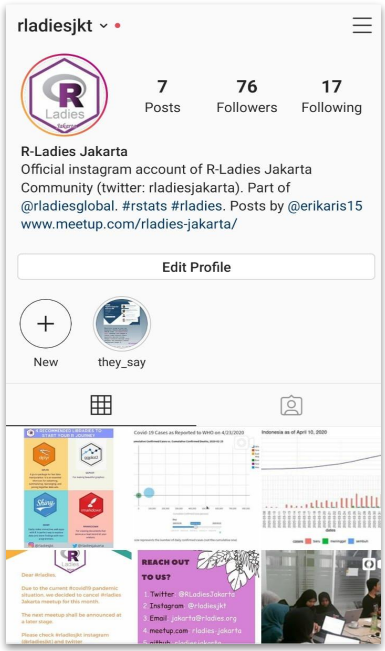




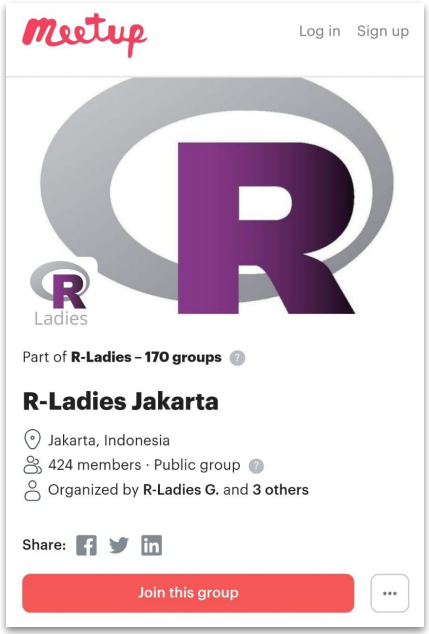
More about R-Ladies Jakarta?

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

@rladiesjakarta



@rladiesjkt



https://meetup.com/rladies-jakarta

@rladiesjakarta

Help Us to Grow

- Follow our social media
- Story and share about this event.
  - **Twitter**: mention **@rladiesjakarta**
  - **IG**: mention **@rladiesjkt**

  } **Hashtag: #rladiesjkt**
- Berikan feedback: https://forms.gle/MHfQwPkhdRJETHm56
- Be a volunteer

Also check our sister: R-Ladies Bogor:

- https://twitter.com/rladiesbogor
- https://www.youtube.com/channel/UCB7NOJYEOoT2IPLuImsx-3w

Web Scraping 101

Theory and Concept

Preparation

- Install required libraries:
  - install.packages('tidyverse')
  - install.packages('rvest')

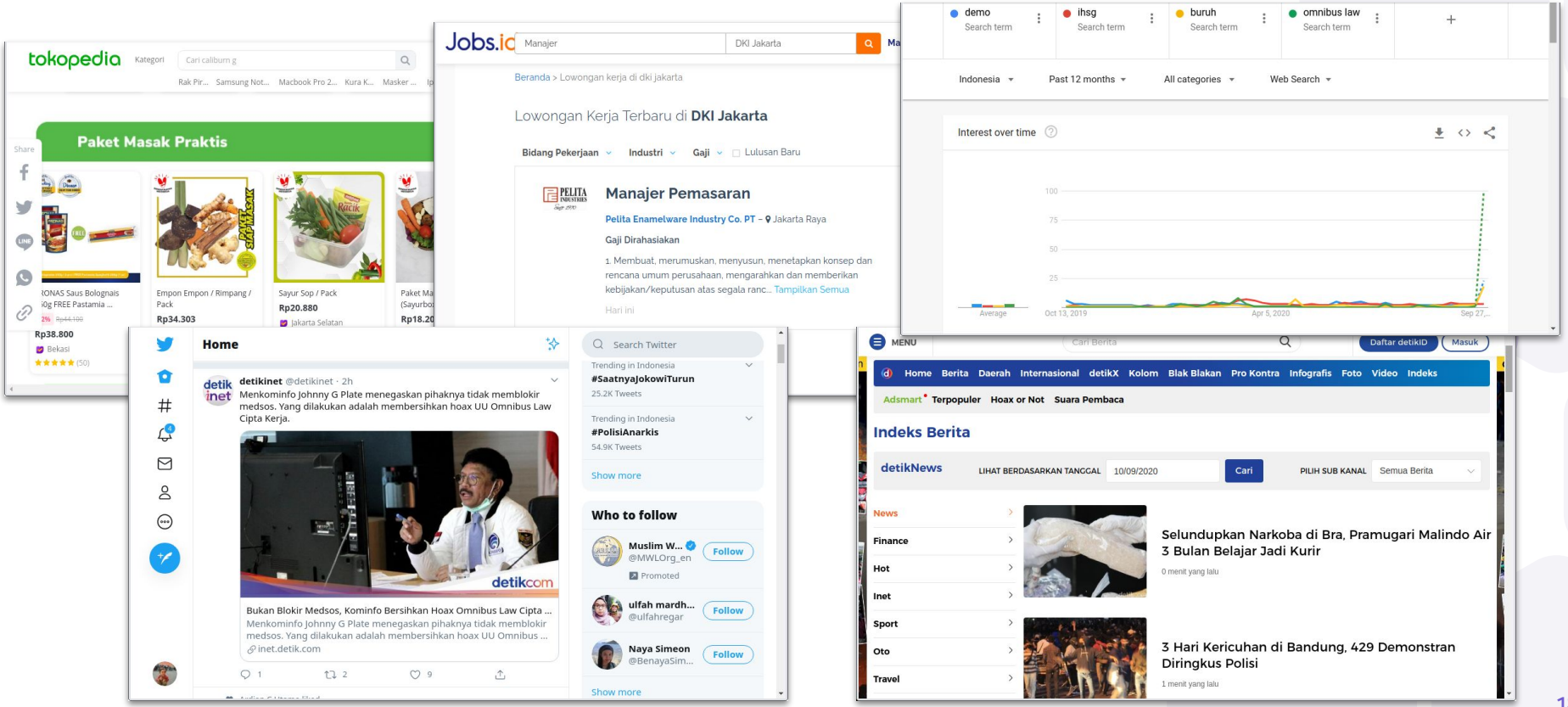
Self Check

- Pernah menggunakan R sebelumnya?
- Experience with web programming?
  - HTML
  - Inspect element
- Pernah melakukan web scraping sebelumnya?
- Pernah melakukan web scraping sebelumnya?

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.

The internet offers an abundance of information



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

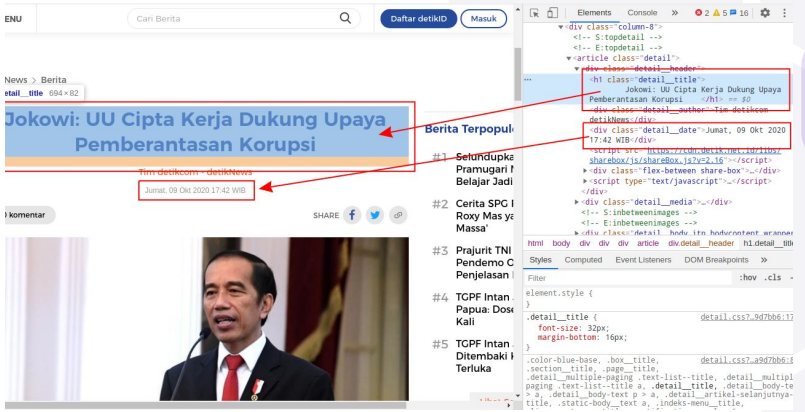
Contoh-contoh informasi yang bisa kita peroleh dari webscraping

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.

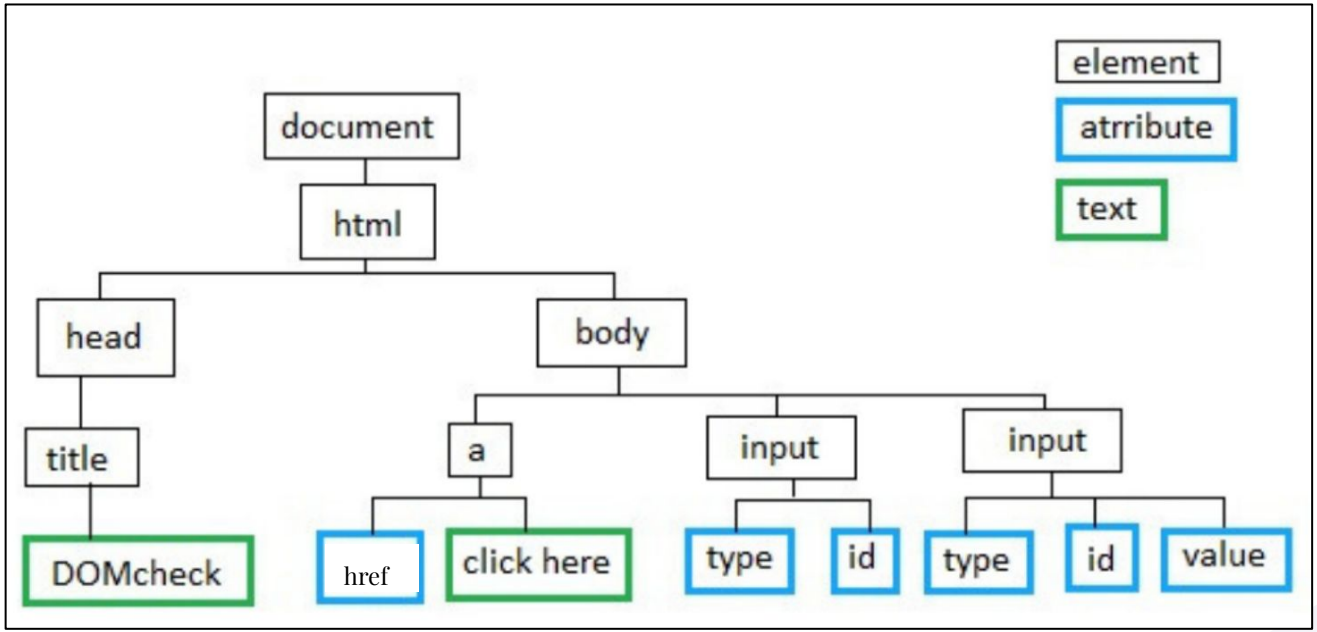
1. Monitoring price, etc.
2. Comparison
3. See trend
4. Sentiment Analysis
5. Reviews analysis
6. News monitoring
7. General business information
8. make more informed decisions

Things You Must Know before Begin Scraping a Website

- Different website has different complexity
  - Static
  - Dynamic (javascript heavy, lazy load)
- Have a basic HTML knowledge.
  - Components of a Website
  - Document Object Model
  - id, class, selector, xpath, dll.
- The web code and design can change anytime.
- Be mindful in maintaining the number of requests
- If there is an API, use it



Document Object Model (DOM)

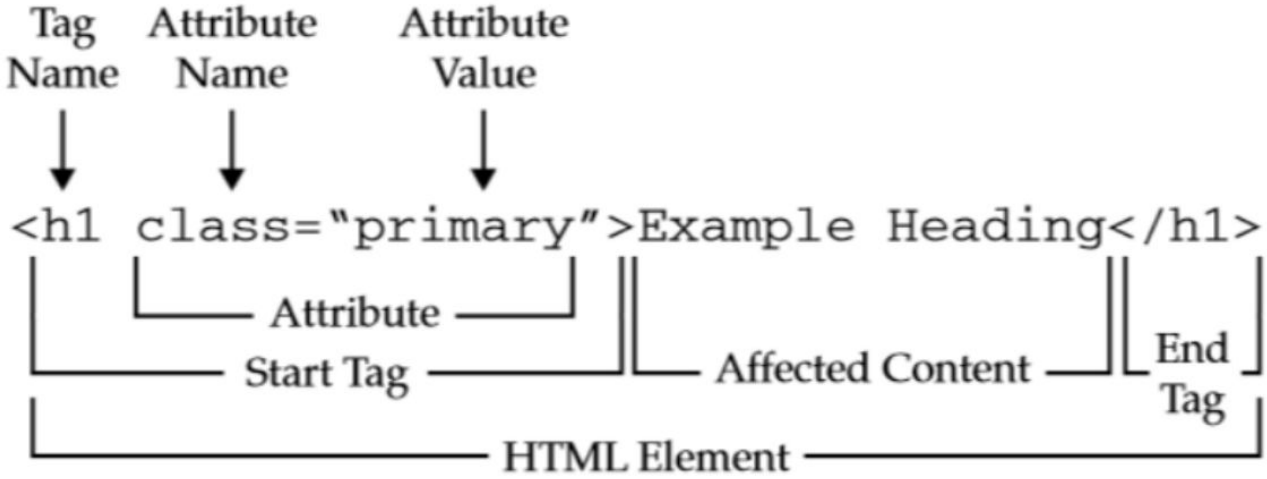


HTML Elements

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/

HTML Tags


Tag Name
Attribute Name
Attribute Value
<h1 class="primary">Example Heading</h1>
Attribute
Start Tag
Affected Content
End Tag
HTML Element

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.

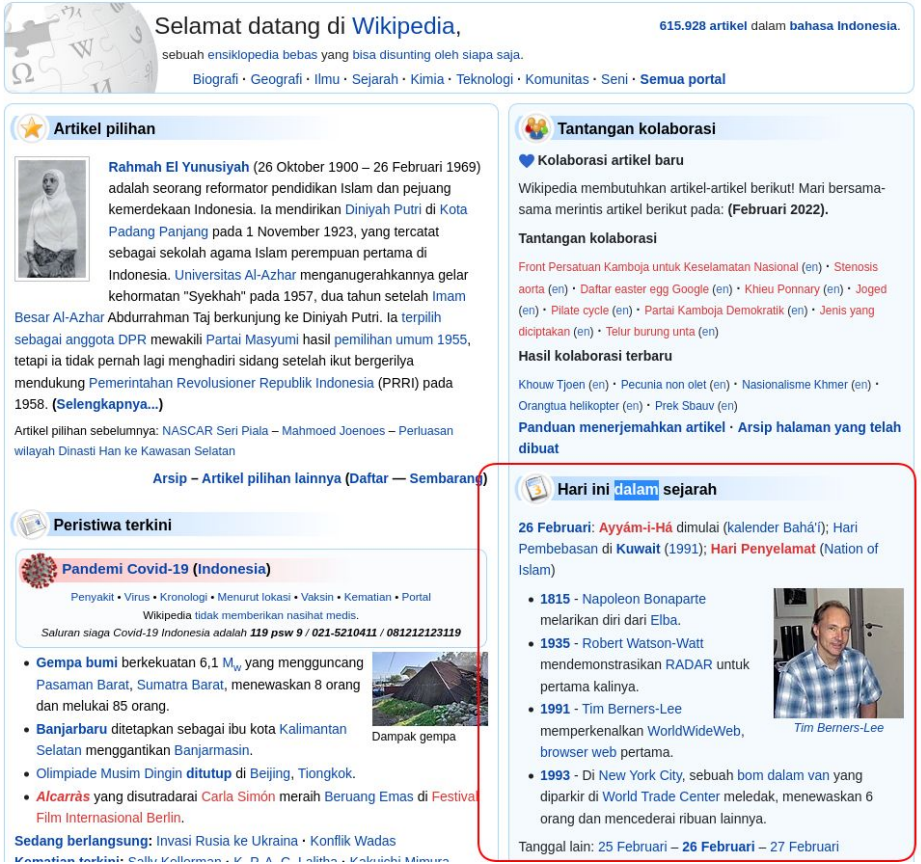
6. Further analysis → visualisasi, sentiment analysis, dll.

About rvest

- R library for scraping web pages
- check the documentation: https://rvest.tidyverse.org/
- important functions:
  - read_html() --> convert a website into an XML object.
  - html_elements() --> extract the relevant nodes from the XML object
  - html_text() --> extract the tagged data from the wanted nodes.
  - html_attrs() --> return a list of the attributes.

Web Scraping Hands On

Case 1: Simple HTML Page



Selamat datang di Wikipedia,

615.928 artikel dalam bahasa Indonesia.

sebuah ensiklopedia bebas yang bisa disunting oleh siapa saja.

Biografi · Geografi · Ilmu · Sejarah · Kimia · Teknologi · Komunitas · Seni · Semua portal

Artikel pilihan

Rahmah El Yunusiyah (26 Oktober 1900 – 26 Februari 1969) adalah seorang reformator pendidikan Islam dan pejuang kemerdekaan Indonesia. Ia mendirikan Diniyah Putri di Kota Padang Panjang pada 1 November 1923, yang tercatat sebagai sekolah agama Islam perempuan pertama di Indonesia. Universitas Al-Azhar menganugerahkannya gelar kehormatan "Syekhah" pada 1957, dua tahun setelah Imam Besar Al-Azhar Abdurrahman Taj berkunjung ke Diniyah Putri. Ia terpilih sebagai anggota DPR mewakili Partai Masyumi hasil pemilihan umum 1955, tetapi ia tidak pernah lagi menghadiri sidang setelah ikut bergerilya mendukung Pemerintahan Revolusioner Republik Indonesia (PRRI) pada 1958. (Selengkapnya...)

Artikel pilihan sebelumnya: NASCAR Seri Piala – Mahmoed Joenoes – Perluasan wilayah Dinasti Han ke Kawasan Selatan

Arsip – Artikel pilihan lainnya (Daftar — Sembarang)

Peristiwa terkini

Pandemi Covid-19 (Indonesia)

Penyakit • Virus • Kronologi • Menurut lokasi • Vaksin • Kematian • Portal

Wikipedia tidak memberikan nasihat medis.

*Saluran siaga Covid-19 Indonesia adalah **119 psw 9 / 021-5210411 / 081212123119***

- **Gempa bumi** berkekuatan 6,1 M_w yang mengguncang Pasaman Barat, Sumatra Barat, menewaskan 8 orang dan melukai 85 orang.
- **Banjarbaru** ditetapkan sebagai ibu kota Kalimantan Selatan menggantikan Banjarmasin.
- Olimpiade Musim Dingin **ditutup** di Beijing, Tiongkok.
- ***Alcarràs*** yang disutradarai Carla Simón meraih Beruang Emas di Festival Film Internasional Berlin.

Dampak gempa

Sedang berlangsung: Invasi Rusia ke Ukraina · Konflik Wadas

Kematian terkini: Sally Kellerman · K. P. A. C. Lalitha · Kakuichi Mimura

Tantangan kolaborasi

Kolaborasi artikel baru

Wikipedia membutuhkan artikel-artikel berikut! Mari bersama-sama merintis artikel berikut pada: **(Februari 2022).**

Tantangan kolaborasi

Front Persatuan Kamboja untuk Keselamatan Nasional (en) · Stenosis aorta (en) · Daftar easter egg Google (en) · Khieu Ponnary (en) · Joged (en) · Pilate cycle (en) · Partai Kamboja Demokratik (en) · Jenis yang diciptakan (en) · Telur burung unta (en)

Hasil kolaborasi terbaru

Khouw Tjoen (en) · Pecunia non olet (en) · Nasionalisme Khmer (en) · Orangtua helikopter (en) · Prek Sbauv (en)

Panduan menerjemahkan artikel · Arsip halaman yang telah dibuat

Hari ini dalam sejarah

26 Februari: **Ayyám-i-Há** dimulai (kalender Bahá'í); Hari Pembebasan di **Kuwait** (1991); **Hari Penyelamat** (Nation of Islam)

- **1815** - Napoleon Bonaparte melarikan diri dari Elba.
- **1935** - Robert Watson-Watt mendemonstrasikan RADAR untuk pertama kalinya.
- **1991** - Tim Berners-Lee memperkenalkan WorldWideWeb, browser web pertama.
- **1993** - Di New York City, sebuah bom dalam van yang diparkir di World Trade Center meledak, menewaskan 6 orang dan mencederai ribuan lainnya.

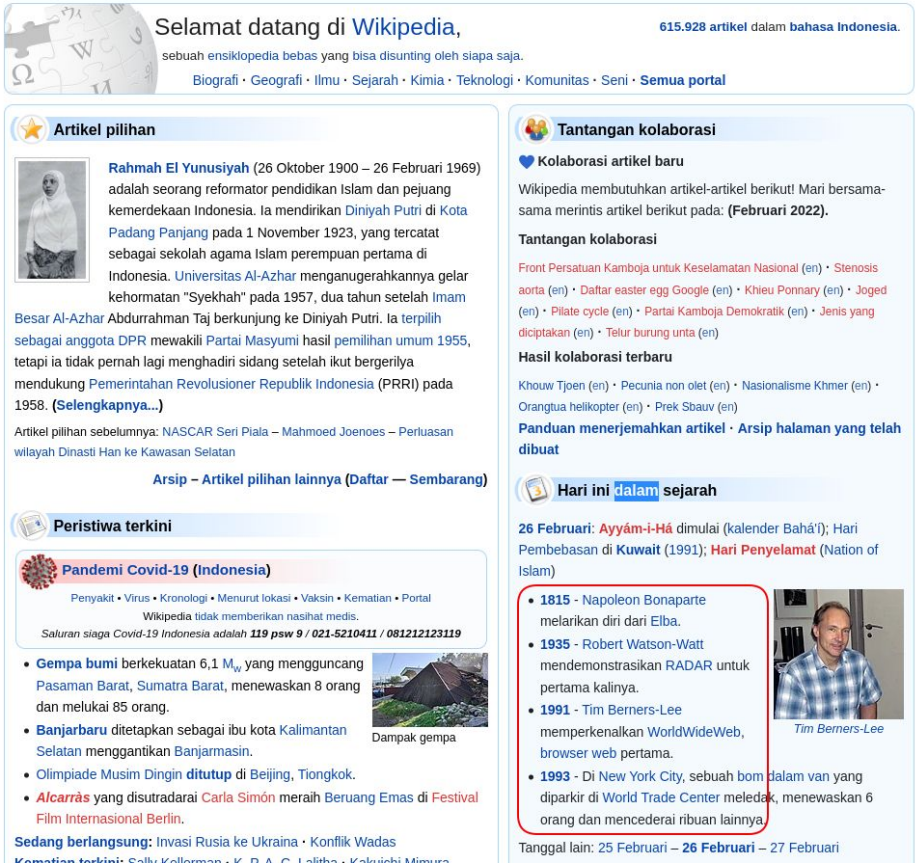
Tim Berners-Lee

Tanggal lain: 25 Februari – **26 Februari** – 27 Februari

https://id.wikipedia.org/wiki/Halaman_Utama

rvest_wikipedia.R

Case 2: Simple HTML Page (Multiple Elements)



Selamat datang di Wikipedia,

615.928 artikel dalam bahasa Indonesia.

sebuah ensiklopedia bebas yang bisa disunting oleh siapa saja.

Biografi · Geografi · Ilmu · Sejarah · Kimia · Teknologi · Komunitas · Seni · Semua portal

Artikel pilihan

Rahmah El Yunusiyah (26 Oktober 1900 – 26 Februari 1969) adalah seorang reformator pendidikan Islam dan pejuang kemerdekaan Indonesia. Ia mendirikan Diniyah Putri di Kota Padang Panjang pada 1 November 1923, yang tercatat sebagai sekolah agama Islam perempuan pertama di Indonesia. Universitas Al-Azhar menganugerahkannya gelar kehormatan "Syekhah" pada 1957, dua tahun setelah Imam Besar Al-Azhar Abdurrahman Taj berkunjung ke Diniyah Putri. Ia terpilih sebagai anggota DPR mewakili Partai Masyumi hasil pemilihan umum 1955, tetapi ia tidak pernah lagi menghadiri sidang setelah ikut bergerilya mendukung Pemerintahan Revolusioner Republik Indonesia (PRRI) pada 1958. **(Selengkapnya...)**

Artikel pilihan sebelumnya: NASCAR Seri Piala – Mahmoed Joenoes – Perluasan wilayah Dinasti Han ke Kawasan Selatan

Arsip – Artikel pilihan lainnya (Daftar — Sembarang)

Peristiwa terkini

Pandemi Covid-19 (Indonesia)

Penyakit • Virus • Kronologi • Menurut lokasi • Vaksin • Kematian • Portal

Wikipedia tidak memberikan nasihat medis.

*Saluran siaga Covid-19 Indonesia adalah **119 psw 9** / **021-5210411** / **081212123119***

- **Gempa bumi** berkekuatan 6,1 M_w yang mengguncang Pasaman Barat, Sumatra Barat, menewaskan 8 orang dan melukai 85 orang.
- **Banjarbaru** ditetapkan sebagai ibu kota Kalimantan Selatan menggantikan Banjarmasin.
- Olimpiade Musim Dingin **ditutup** di Beijing, Tiongkok.
- ***Alcarràs*** yang disutradarai Carla Simón meraih Beruang Emas di Festival Film Internasional Berlin.

Dampak gempa

Sedang berlangsung: Invasi Rusia ke Ukraina · Konflik Wadas

Kematian terkini: Sally Kellerman · K. P. A. C. Lalitha · Kakuichi Mimura

Tantangan kolaborasi

Kolaborasi artikel baru

Wikipedia membutuhkan artikel-artikel berikut! Mari bersama-sama merintis artikel berikut pada: **(Februari 2022).**

Tantangan kolaborasi

Front Persatuan Kamboja untuk Keselamatan Nasional (en) · Stenosis aorta (en) · Daftar easter egg Google (en) · Khieu Ponnary (en) · Joged (en) · Pilate cycle (en) · Partai Kamboja Demokratik (en) · Jenis yang diciptakan (en) · Telur burung unta (en)

Hasil kolaborasi terbaru

Khouw Tjoen (en) · Pecunia non olet (en) · Nasionalisme Khmer (en) · Orangtua helikopter (en) · Prek Sbauv (en)

Panduan menerjemahkan artikel · Arsip halaman yang telah dibuat

Hari ini dalam sejarah

26 Februari: **Ayyám-i-Há** dimulai (kalender Bahá'í); Hari Pembebasan di **Kuwait** (1991); **Hari Penyelamat** (Nation of Islam)

- **1815** - Napoleon Bonaparte melarikan diri dari Elba.
- **1935** - Robert Watson-Watt mendemonstrasikan RADAR untuk pertama kalinya.
- **1991** - Tim Berners-Lee memperkenalkan WorldWideWeb, browser web pertama.
- **1993** - Di New York City, sebuah bom dalam van yang diparkir di World Trade Center meledak, menewaskan 6 orang dan mencederai ribuan lainnya.

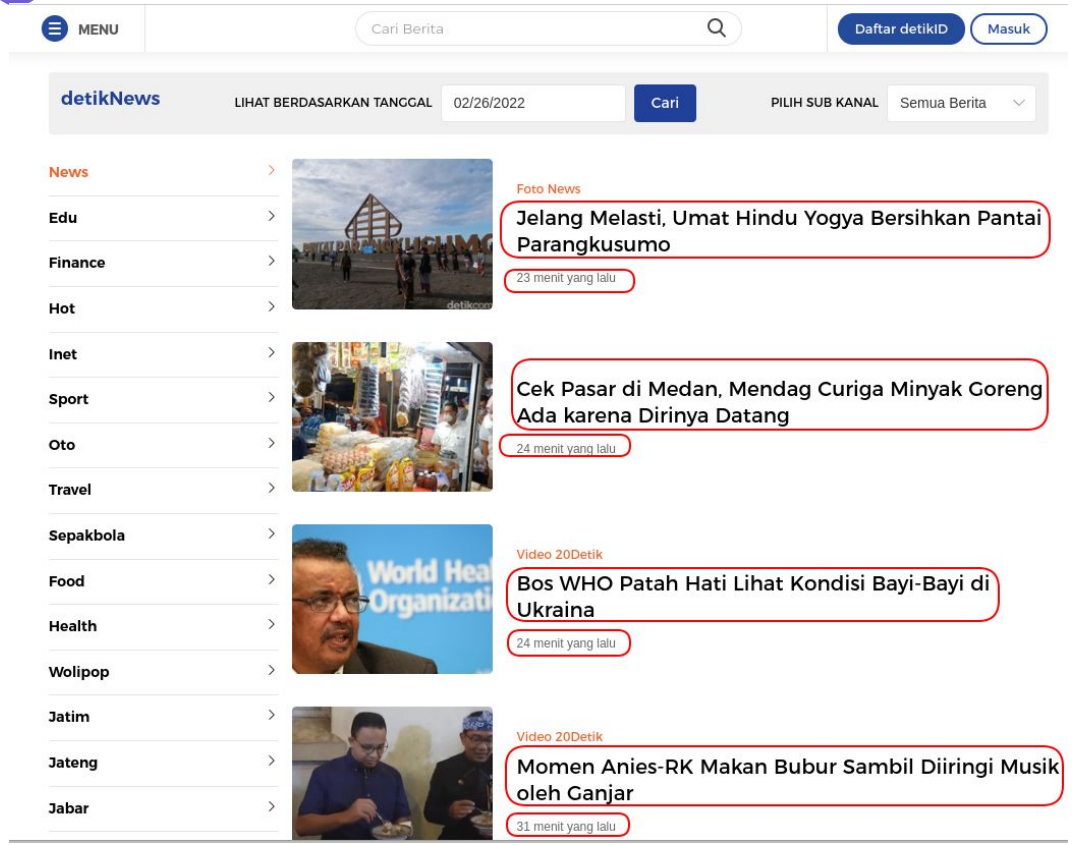
Tim Berners-Lee

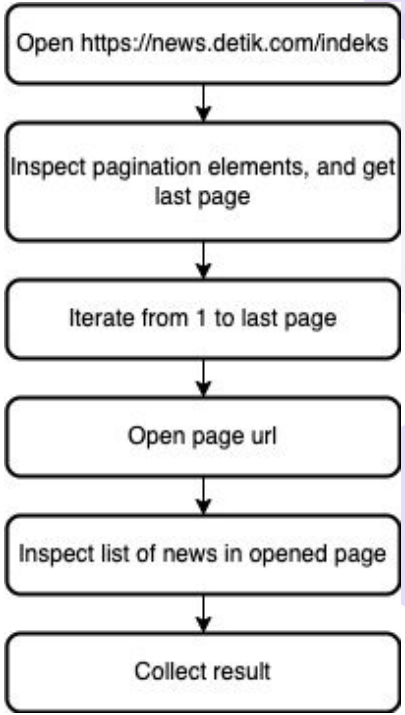
Tanggal lain: 25 Februari – **26 Februari** – 27 Februari

https://id.wikipedia.org/wiki/Halaman_Utama

rvest_wikipedia_multiple.R

Case 3: Indeks Berita Detik.com

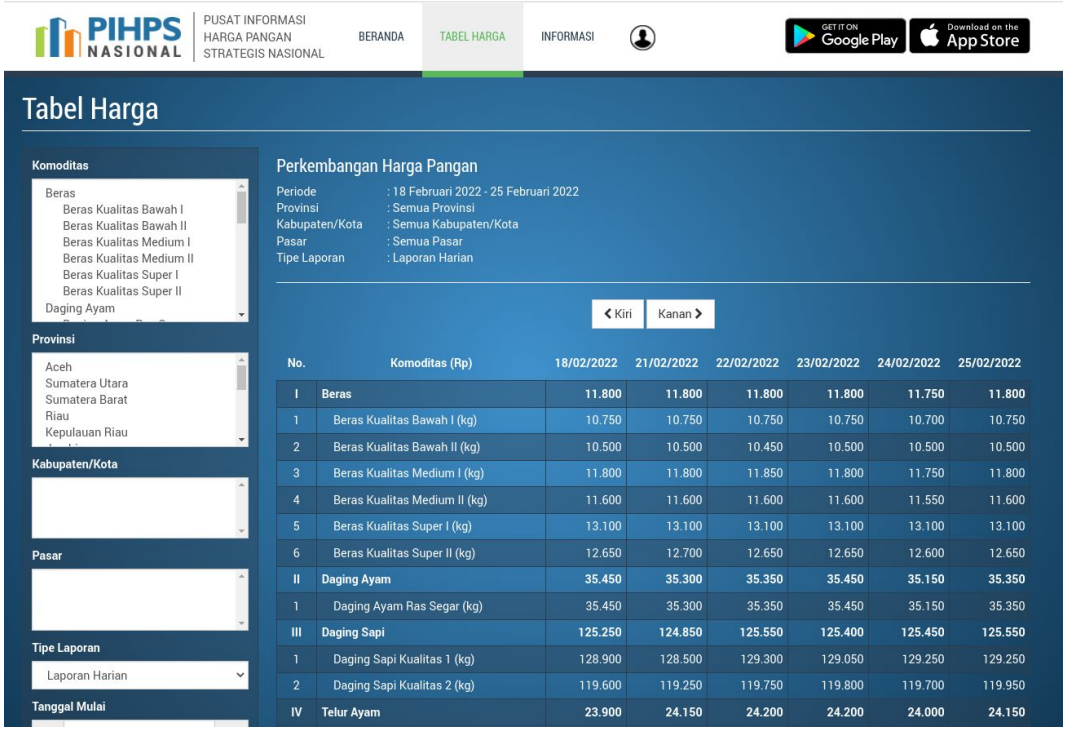




https://news.detik.com/indeks/

rvest_detik_index.R

Bonus Case: Harga Pangan



No.	Komoditas (Rp)	18/02/2022	21/02/2022	22/02/2022	23/02/2022	24/02/2022	25/02/2022
I	**Beras**	**11.800**	**11.800**	**11.800**	**11.800**	**11.750**	**11.800**
1	Beras Kualitas Bawah I (kg)	10.750	10.750	10.750	10.750	10.700	10.750
2	Beras Kualitas Bawah II (kg)	10.500	10.500	10.450	10.500	10.500	10.500
3	Beras Kualitas Medium I (kg)	11.800	11.800	11.850	11.800	11.750	11.800
4	Beras Kualitas Medium II (kg)	11.600	11.600	11.600	11.600	11.550	11.600
5	Beras Kualitas Super I (kg)	13.100	13.100	13.100	13.100	13.100	13.100
6	Beras Kualitas Super II (kg)	12.650	12.700	12.650	12.650	12.600	12.650
II	**Daging Ayam**	**35.450**	**35.300**	**35.350**	**35.450**	**35.150**	**35.350**
1	Daging Ayam Ras Segar (kg)	35.450	35.300	35.350	35.450	35.150	35.350
III	**Daging Sapi**	**125.250**	**124.850**	**125.550**	**125.400**	**125.450**	**125.550**
1	Daging Sapi Kualitas 1 (kg)	128.900	128.500	129.300	129.050	129.250	129.250
2	Daging Sapi Kualitas 2 (kg)	119.600	119.250	119.750	119.800	119.700	119.950
IV	**Telur Ayam**	**23.900**	**24.150**	**24.200**	**24.200**	**24.000**	**24.150**

https://hargapangan.id/tabel-harga/

https://github.com/erikaris/talks/blob/main/unipasby_202110/crawler2.R

Thank you

- Wildlife Conservation Society (WCS)
- R-Ladies Jakarta Team
- Semua yang telah hadir hari ini.

Semua materi tersedia di https://github.com/RLadiesJakarta
Recording: Youtube R-Ladies Jakarta
Join WAG: https://rladiesjakarta.github.io/#/registration
News & Events: https://www.instagram.com/rladiesjkt/