

Mengambil Data Konten dari Situs Berita

Apa yang akan dipelajari?

- Pengantar web scraping
- Alur kerja web scraping
- Praktek web scraping pada static website
- Web Scraper dan RSelenium
- Praktek web scraping pada dynamic website

Pengantar Web Scraping

Apa itu web scraping?



“Kegiatan Mengambil data tertentu secara semi terstruktur dari sebuah website.”

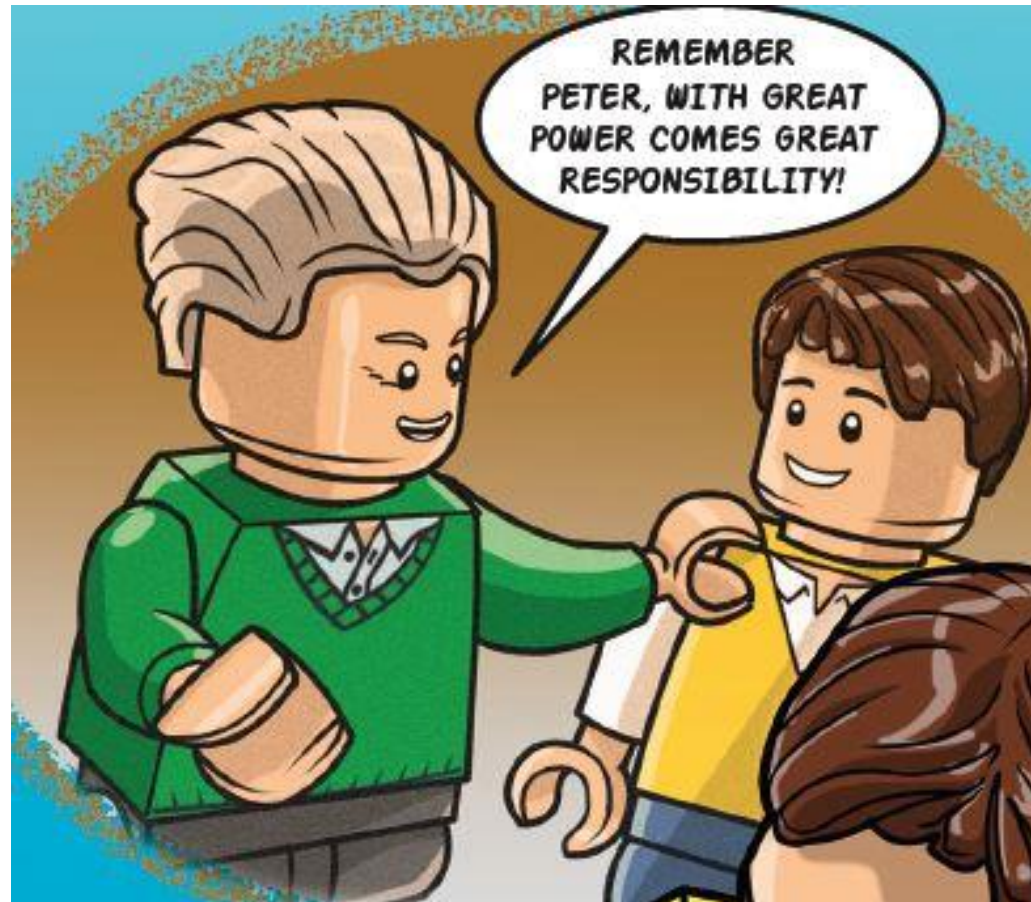
(Sumber: [Wikipedia](https://en.wikipedia.org/wiki/Web_scraping))

Scraping vs Crawling

Data Scraping	Data Crawling
Melibatkan kegiatan mengekstrak data dari berbagai sumber termasuk website	Merujuk pada mengunduh halaman website
Dapat dilakukan pada berbagai skala	Sebagian besar dilakukan pada skala besar
Deduplikasi belum tentu menjadi bagian	Deduplikasi merupakan bagian penting
Memerlukan crawler agent dan parser	Hanya memerlukan crawler agent

Sumber: promptcloud.com, 2012

Etika



FAQ of Scraping

- **Apakah kegiatan scraping illegal?**

“Bisa Ya dan Tidak, saat ini belum ada peraturan yang jelas terkait pengumpulan data dengan teknik scraping. Namun, aktivitas mencurigakan pada suatu situs seperti DDOS attack merupakan kegiatan ilegal”

- **Apa yang bisa saya lakukan untuk memastikan saya tidak mengekspos diri saya atau perusahaan dalam proyek scraping 'legal / etis'?**

“Jangan kumpulkan informasi personal (nama, alamat, email, dll). Pastikan setiap data melalui proses anonimisasi”

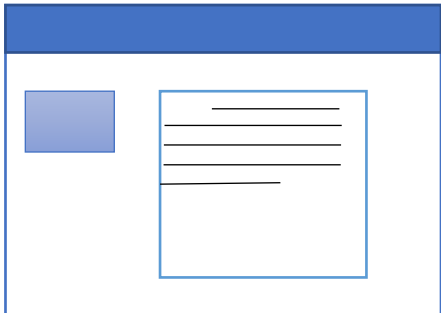
8 Commandments

1. Gunakan API publik jika ada
2. Hubungi admin website jika memungkinkan untuk memberitahu tujuan scraping dilakukan
3. Gunakan jeda pada program yang digunakan
4. Jangan kecanduan dengan data (ambil data secukupnya)
5. Jangan melakukan plagiasi
6. Jangan hanya tertarik dengan data. Berikan nilai tambah pada data yang ada
7. Sadarilah orang hanyalah orang → jika memungkinkan kita dapat meminta datanya secara langsung
8. Ingat Anda yang bertanggung jawab !!!

Alur Kerja Proses Scraping

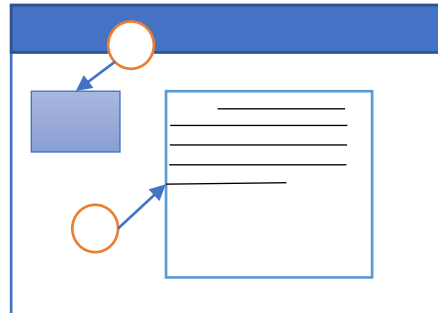
Alur Kerja

1



Menentukan
website target

2



Menentukan data
yang akan
diambil

3



Membangun
Program

Alur Kerja

4

```
<html>
<head>
</head>
<frameset rows="121,*" framespacing="0" border="0" frameborder="0">
  <frame name="header1" scrolling="no" src="header.html" marginheight="0"
  marginwidth="0" background="/hdrbg1.gif" target="_self">
</frame>
  <frame name="body" src="/css/body/body.html" marginheight="0"
  marginwidth="0" onload="setFrameSrc();"
  <frame name="body" src="/css/body/body.html" marginheight="0"
  marginwidth="0" onload="setFrameSrc();"
  <frame name="main1" src="/main.html" marginheight="0" marginwidth="0"
  noresize target="_self">
</frame>
</frameset>
</html>
```

Proses Scraping

5



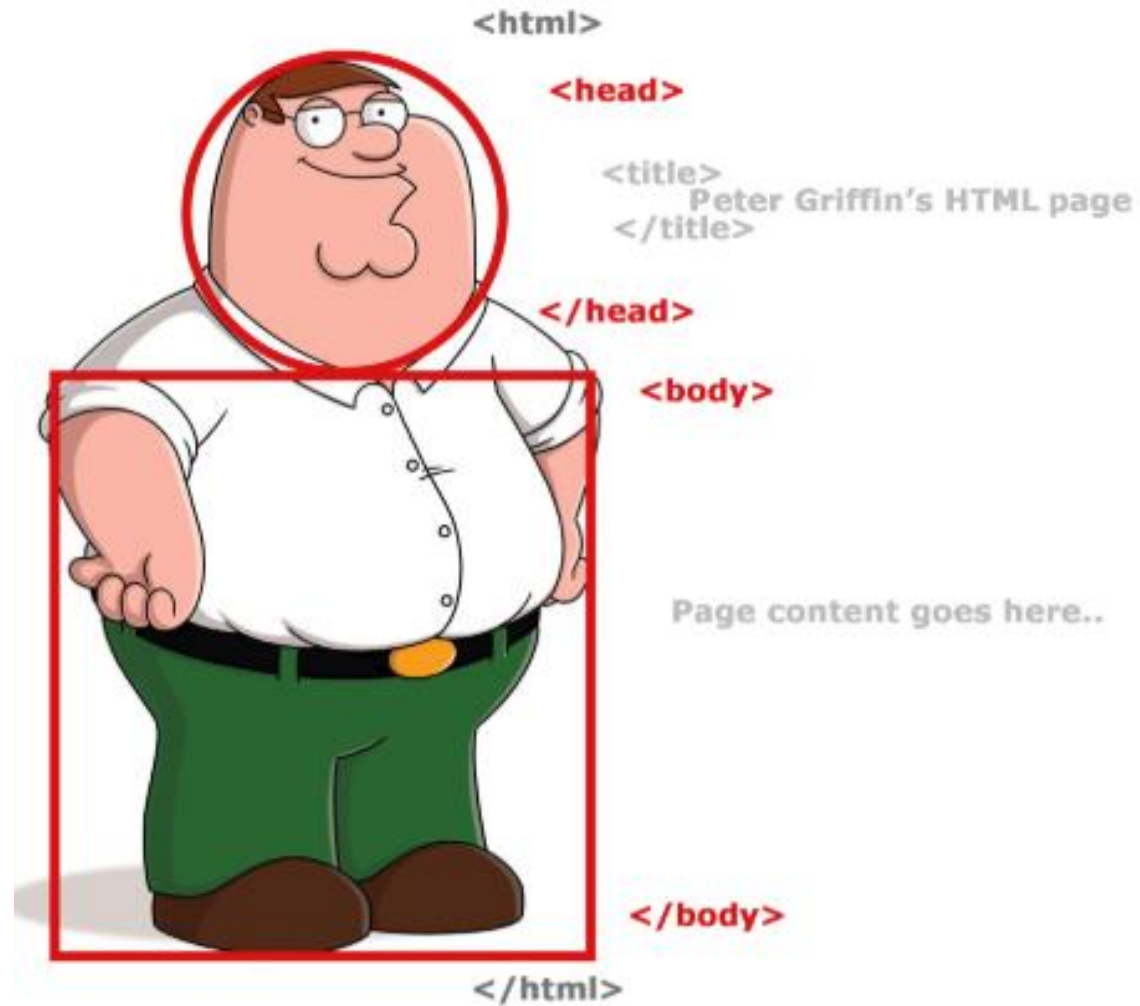
Simpan Data

Memahami Halaman Web

Terdapat 3 bahasa utama pembentuk sebuah website:

- **HTML** : Menentukan struktur dan konten sebuah website
- **CSS** : Menentukan gaya dan tampilan sebuah website
- **Javascript** : Memberikan fungsionalitas sebuah halaman web

HTML



CSS

Terdapat 2 konsep utama:

- **class** : mengatur style elemen-elemen umum

```
<p class="red-text" >Text 1</p>  
<p class="red-text" >Text 2</p>  
<p class="red-text" >Text 3</p>
```

- **id** : membedakan satu tag html dengan lainnya

```
<p id="special" >This is a special tag.</p>
```



Kita perlu menentukan secara tepat html tags, class, dan id lokasi data yang akan diambil

R Packages yang Digunakan



Web Scraping



Data Manipulation



Menyimpan Data

RVEST

- Packages yang digunakan untuk melakukan **web scraping** pada **website statis**.
- **Rvest** terinspirasi dari library **Beautiful Soup** yang ada di Python
- Rvest memungkinkan untuk digunakan bersama dengan operator pipe (**%>%**), sehingga sintaks lebih mudah dibaca.

RVEST

Fungsi – fungsi yang sering digunakan:

- **read_html(url)** : parsing halaman web
- **html_nodes("tag #id .class")** : memilih nodes lokasi data
- **html_text()** : mengambil teks dari node terpilih
- **html_attrs()** : mengambil atribut node (ex: link)

Selector Gadget

Selector gadget merupakan chrome extension yang digunakan untuk membantu mencari tag, id, atau class lokasi data yang akan diambil



SelectorGadget

Offered by: selectorgadget.com

★★★★★ 83 | [Developer Tools](#) | 100,000+ users

Overview

Reviews

Support

Related



Praktek

The screenshot displays the RStudio environment. The main window shows a data frame with 19 rows and 3 columns: 'judul', 'tanggal', and 'isi'. The data contains news articles related to the Hajj 2020 ban in Indonesia. A red hexagonal watermark with the 'rvest' logo is overlaid on the data frame.

The console window on the right shows the following R code and output:

```

D:/EnvStat/news_scraping/
+ str_remove_all("(?<=\\r\\n)(.?)?(?=\\r\\n\\r\\n)") %>%
+ str_remove_all("(?<=\\r\\n)(.?)?(?=\\r\\n\\r\\n)") %>%
+ str_remove_all("Baca juga: ") %>%
+ str_remove_all("[\\r\\n]")

[1] "Wakil Presiden Ma'ruf Amin menegaskan hak calon jemaah haji y
ang batal berangkat pada tahun ini tidak akan hilang. Hak keberang
katan akan diberikan dengan penyediaan slot pada musim haji tahun
2021." "Soal dana atau subsidi dari pengelolaan dana haji, itu mem
erintah juga mempersilakan calon jamaah menarik dana t
anah diatur dan merupakan bagian yang sudah menjadi hak darip
da jamaah haji itu. Jadi tidak akan hilang. Dan ketika di
n depan, mereka akan memperoleh haknya lagi," kata Ma'r
ekonferensi di Jakarta, pada Senin (8/6/2020). Ma'ruf men
emerintah juga mempersilakan calon jamaah menarik dana t
aji jika menghendaki hal itu. Sementara uang calon jemaah
ak menarik dananya, akan dikelola oleh Badan Pengelola Ke
(BPKH). "Kalau dia mau menarik, ya itu saya kira hak j
i kalau tidak menarik, dana itu akan dikelola oleh lemb
dah ditung oleh undang-undang yang memang sudah diberi
untuk mengelola dananya jamaah itu," ujar Maruf. Dia m
pembatalan keberangkatan calon jamaah haji tahun ini di
karena alasan keselamatan, baik untuk calon jamaah maupun
arakat Indonesia pada umumnya. Faktor keamanan di perjalanan
jadi pertimbangan Pemerintah Indonesia saat membatalkan keberang
katan jamaah haji di tengah pandemi virus corona (COVID-19). Menuru
Ma'ruf, perjalanan para jamaah menuju Tanah Suci dan kembali lag
i ke Indonesia memiliki risiko tinggi terhadap penularan COVID-19.
Pembatalan keberangkatan calon jamaah haji kali ini juga bukan yan
g pertama. Ma'ruf mencatat, pembatalan pernah terjadi karena alasa
n keamanan akibat adanya perang. "Ini memang terpaksa mundur karen
a tidak bisa berangkat karena ada alasan-alasan. Dulu pernah juga
(batal) karena alasan keamanan, terjadi perang. Itu juga tidak ad
a pemberangkatan jamaah haji," ujar Maruf."

```

Web Scraper dan RSelenium

Masalah Rvest?

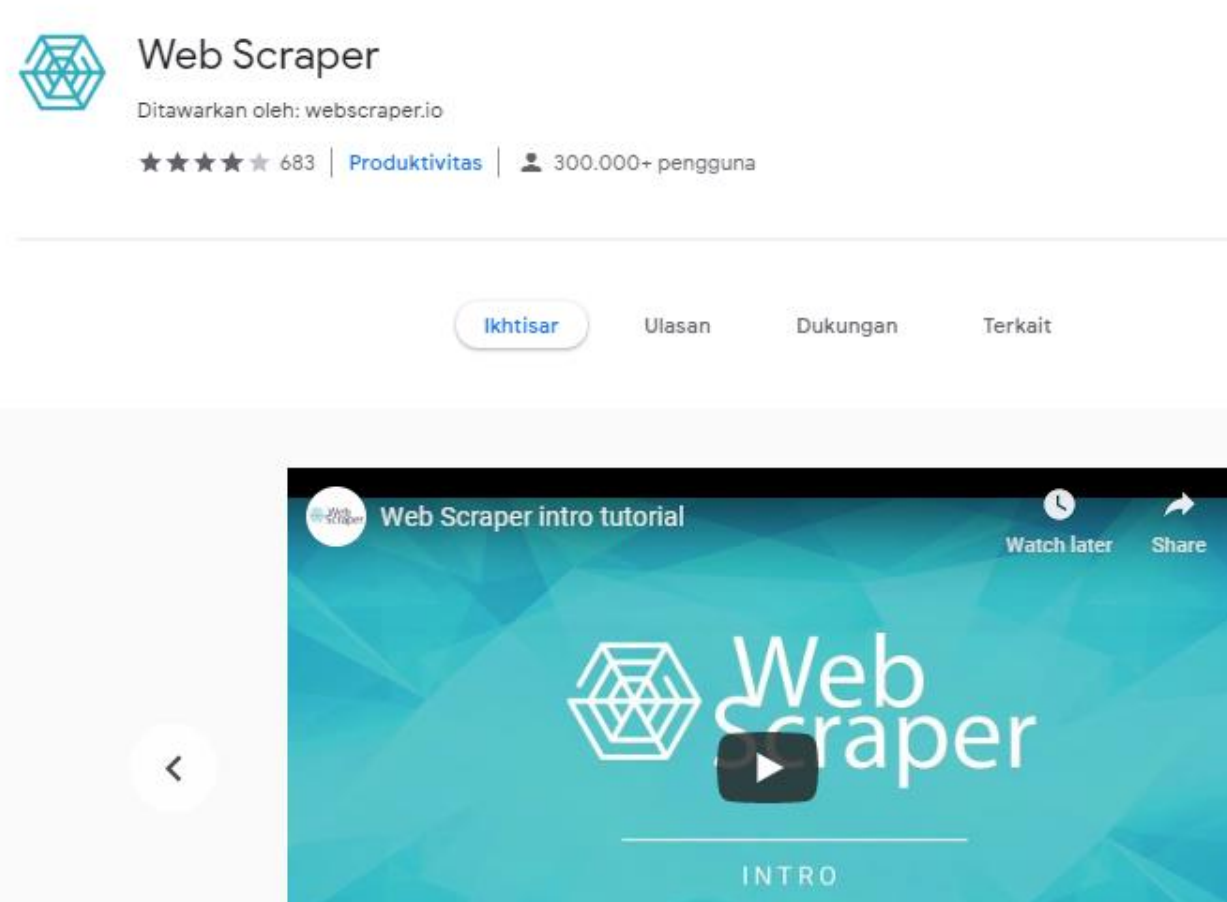
- Pada halaman web yang dimuat secara dynamic (ex: scroll untuk memunculkan konten), rvest hanya bisa menangkap halaman originalnya saja (tidak dengan halaman yang baru dimuat)
- Perlu dikombinasikan dengan aplikasi atau packages lainnya.

Tools Lainnya



Web Scraper

- Web scraper merupakan chrome extension untuk melakukan web scraping



RSelenium

- RSelenium merupakan packages yang digunakan untuk melakukan otomatisasi pada kegiatan web browsing
- RSelenium secara umum digunakan untuk kegiatan web testing
- RSelenium akan merender halaman web secara otomatis dan mengeksekusi semua kode Javascript, sehingga kita akan memperoleh halaman sumber yang telah terparsing.



Alur Kerja

1



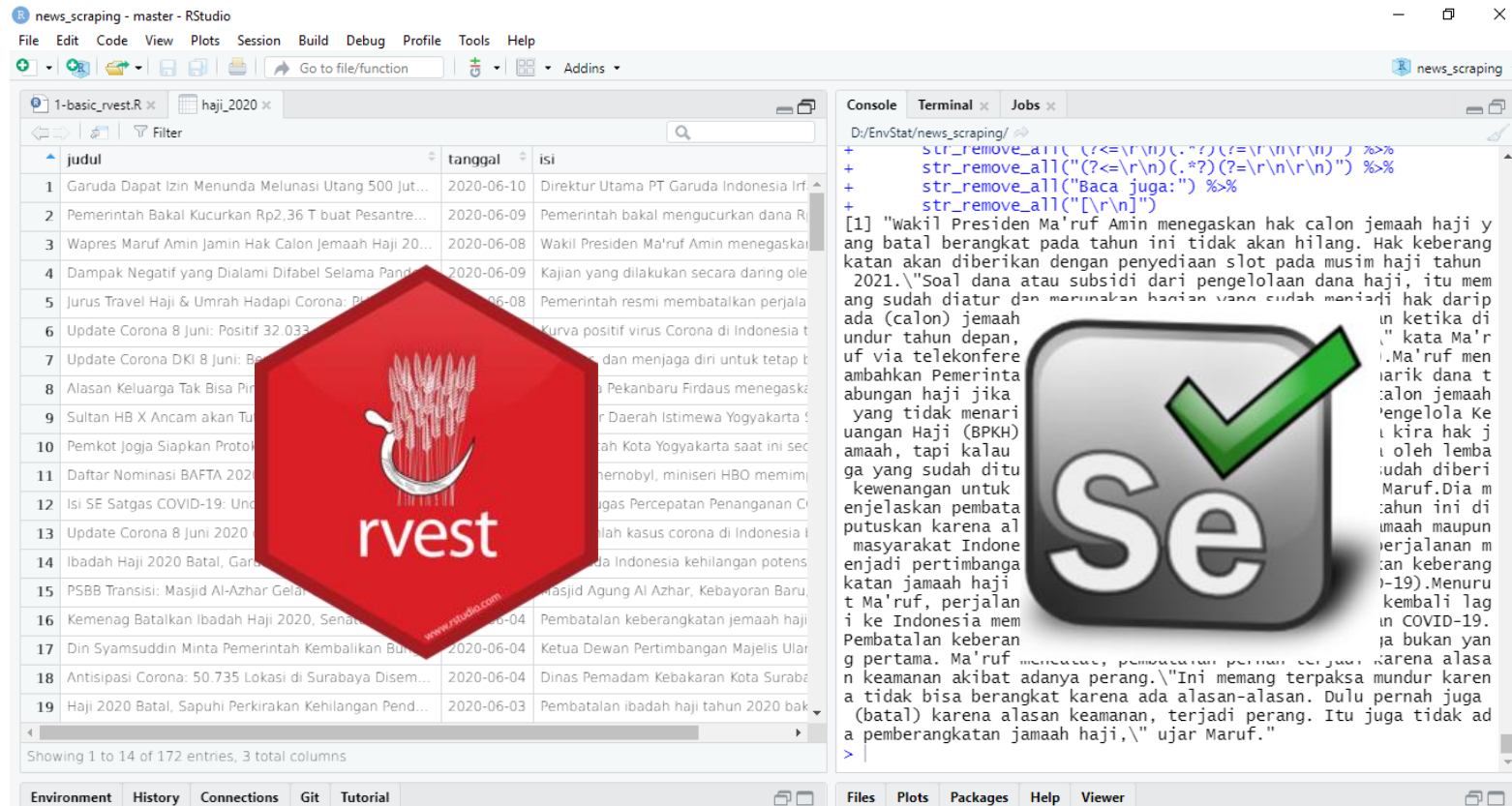
**Ambil tautan
berita**

2



**Scraping konten
berita**

Praktek



news_scraping - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

1-basic_rvest.R haji_2020

	judul	tanggal	isi
1	Garuda Dapat Izin Menunda Melunasi Utang 500 Jut...	2020-06-10	Direktur Utama PT Garuda Indonesia lrf...
2	Pemerintah Bakal Kucurkan Rp2,36 T buat Pesantre...	2020-06-09	Pemerintah bakal mengucurkan dana R...
3	Wapres Maruf Amin Jamin Hak Calon Jemaah Haji 20...	2020-06-08	Wakil Presiden Ma'ruf Amin menegaskan...
4	Dampak Negatif yang Dialami Difabel Selama Pand...	2020-06-09	Kajian yang dilakukan secara daring ole...
5	Jurus Travel Haji & Umrah Hadapi Corona: Di...	2020-06-08	Pemerintah resmi membatalkan perjala...
6	Update Corona 8 Juni: Positif 32.033		Kurva positif virus Corona di Indonesia t...
7	Update Corona DKI 8 Juni: Be...		dan menjaga diri untuk tetap b...
8	Alasan Keluarga Tak Bisa Pr...		Pekanbaru Firdaus menegaskan...
9	Sultan HB X Ancam akan Tu...		Daerah Istimewa Yogyakarta S...
10	Pemkot Jogja Siapkan Protok...		Kota Yogyakarta saat ini sec...
11	Daftar Nominasi BAFTA 202...		hernobyl, miniseri HBO memimi...
12	Isi SE Satgas COVID-19: Und...		Percepatan Penanganan C...
13	Update Corona 8 Juni 2020		kasus corona di Indonesia i...
14	Ibadah Haji 2020 Batal, Gar...		Indonesia kehilangan potens...
15	PSBB Transisi: Masjid Al-Azhar Gel...		Masjid Agung Al Azhar, Kebayoran Baru...
16	Kemenag Batalkan Ibadah Haji 2020, Senat...	2020-06-04	Pembatalan keberangkatan jemaah haji...
17	Din Syamsuddin Minta Pemerintah Kembalikan Bu...	2020-06-04	Ketua Dewan Pertimbangan Majelis Ular...
18	Antisipasi Corona: 50.735 Lokasi di Surabaya Disem...	2020-06-04	Dinas Pemadam Kebakaran Kota Suraba...
19	Haji 2020 Batal, Sapuhi Perkiraan Kehilangan Pend...	2020-06-03	Pembatalan ibadah haji tahun 2020 bak...

Showing 1 to 14 of 172 entries, 3 total columns

Environment History Connections Git Tutorial

news_scraping

Console Terminal Jobs

```

D:/EnvStat/news_scraping/
+ str_remove_all("(?<=\\r\\n)(.*)"") %>%
+ str_remove_all("(?<=\\r\\n)(.*)"") %>%
+ str_remove_all("Baca juga:") %>%
+ str_remove_all("[\\r\\n]")
[1] "Wakil Presiden Ma'ruf Amin menegaskan hak calon jemaah haji y
ang batal berangkat pada tahun ini tidak akan hilang. Hak keberang
katan akan diberikan dengan penyediaan slot pada musim haji tahun
2021.\\\"Soal dana atau subsidi dari pengelolaan dana haji, itu mem
ang sudah diatur dan merupakan bagian yang sudah menjadi hak darip
ada (calon) jemaah
in ketika di
\\\" kata Ma'r
Ma'ruf men
arik dana t
alon jemaah
pengelola Ke
kira hak j
oleh lemb
udah diberi
Maruf.Dia m
ahun ini di
maah maupun
perjalanan m
an keberang
-19).Menuru
kembali lag
in COVID-19.
ja bukan yan
g pertama. Ma'ruf menegaskan, pembatalan perma
n keamanan akibat adanya perang.\\\"Ini memang terpaksa mundur karen
a tidak bisa berangkat karena ada alasan-alasan. Dulu pernah juga
(batal) karena alasan keamanan, terjadi perang. Itu juga tidak ad
a pemberangkatan jemaah haji,\\\" ujar Maruf."
>

```

Files Plots Packages Help Viewer