DIGITAL TALENT SCHOLARSHIP

THEMATIC ACADEMY

Tema Pelatihan

Pertemuan #5 : Data Scraping & Crawling For Beginners To Data Analysis

KOMINFO

#JADIJAGOANDIGITAL

Badan Penelitian dan Pengembangan Sumber Daya Manusia

PROFIL PENGAJAR

Putu Rika Sahriana

Latarbelakang Pendidikan Pengajar

- Math Education, Ganesha University of Education

Riwayat Pekerjaan

- Math, Programming, Robot and IOT Teacher
- Data Analyst at KPMG (Virtual Internship)
- Data Science at Bisa AI (Internship)
- Data Freelancer

Contact Pengajar

Ponsel	: +6281999871090
Email	: rikasahriana28@gmail.com
Linkedin	: linkedin.com/in/rikasahriana

Learning Objective

In this course you will:

A. Introduction to Data Crawling & Data Scraping

B. Techniques for retrieving Data on the web and social media with R

C. Understand the basic principles of API

D. Import real time data using API

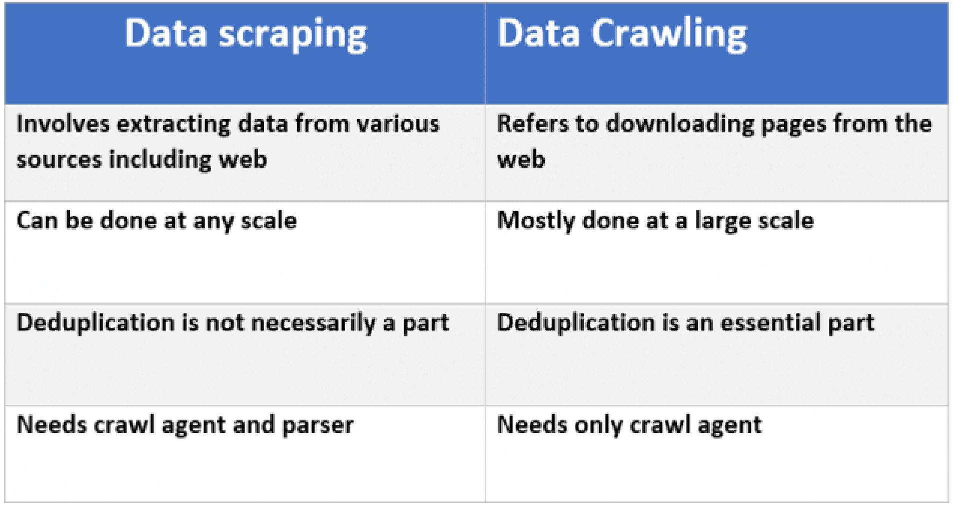
Introduction to Data Crawling & Data Scraping

Data Crawling

Crawling is a technique of collecting data on a website or social media by entering a Uniform Resource Locator (URL).

Data Scraping

Scraping is a technique of collecting data on a website or social media through the process of extracting information using Hypertext Transfer Protocol (HTTP).



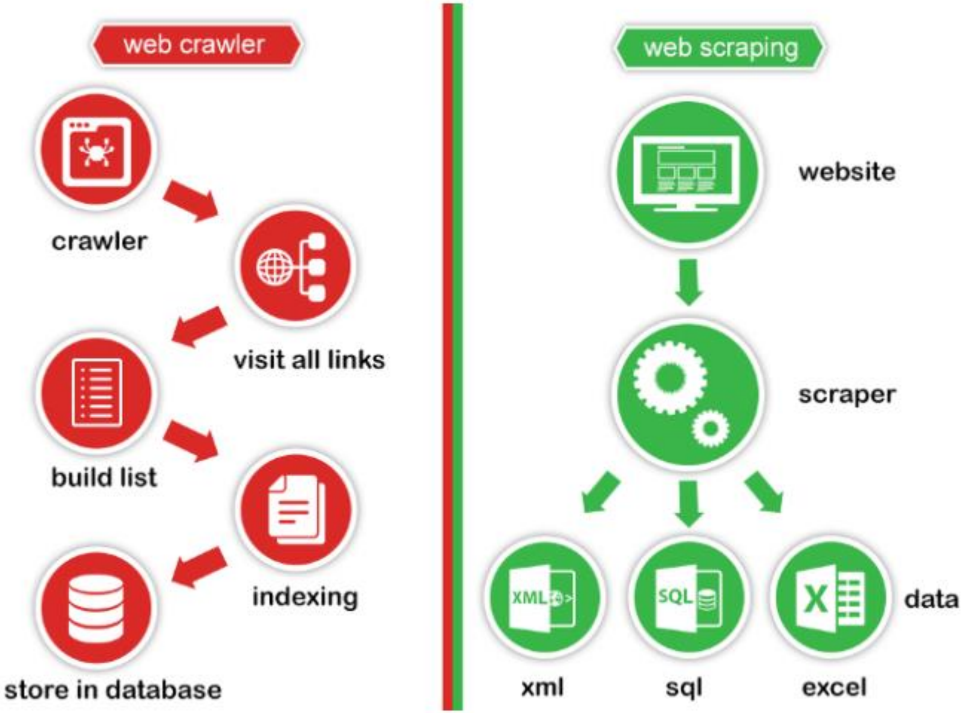
Data scraping	Data Crawling
Involves extracting data from various sources including web	Refers to downloading pages from the web
Can be done at any scale	Mostly done at a large scale
Deduplication is not necessarily a part	Deduplication is an essential part
Needs crawl agent and parser	Needs only crawl agent

sumber: promptcloud.com

Web Crawling vs Web Scraping Technique

Web Crawling	Web Scraping
Selective Crawling	Copy-Paste
Popularity	HTML Parsing
Focused Crawling	DOM Parsing
Distributed Crawling	Vertical Aggregation
Paralel Crawling	Xpath
Web Dynamic	Google Sheet
	Text Patern Machine

Introduction to Data Crawling & Data Scraping



sumber: prowebscraping.com

Introduction to Data Crawling & Data Scraping

Summary

We can assume that when we do web crawling we are actually doing web scraping. But on the contrary when we do web scraping we have not or do not do web crawling.

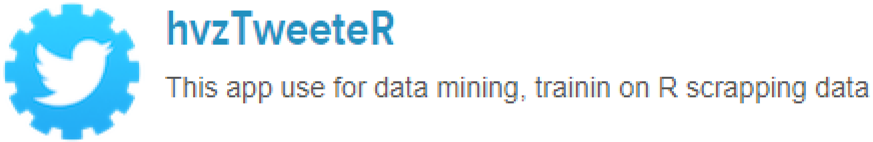
Then for the difference crawling is usually used for large-scale data. The implementation uses bots automatically and uses the Application Programming Interface (API).

While scraping is usually used for data that is relatively not too large and the process of retrieving data on HTML or XML elements using the HTTP protocol.

Techniques for retrieving Data on the web and social media with R

Step by Step to Retrieving Data from Twitter

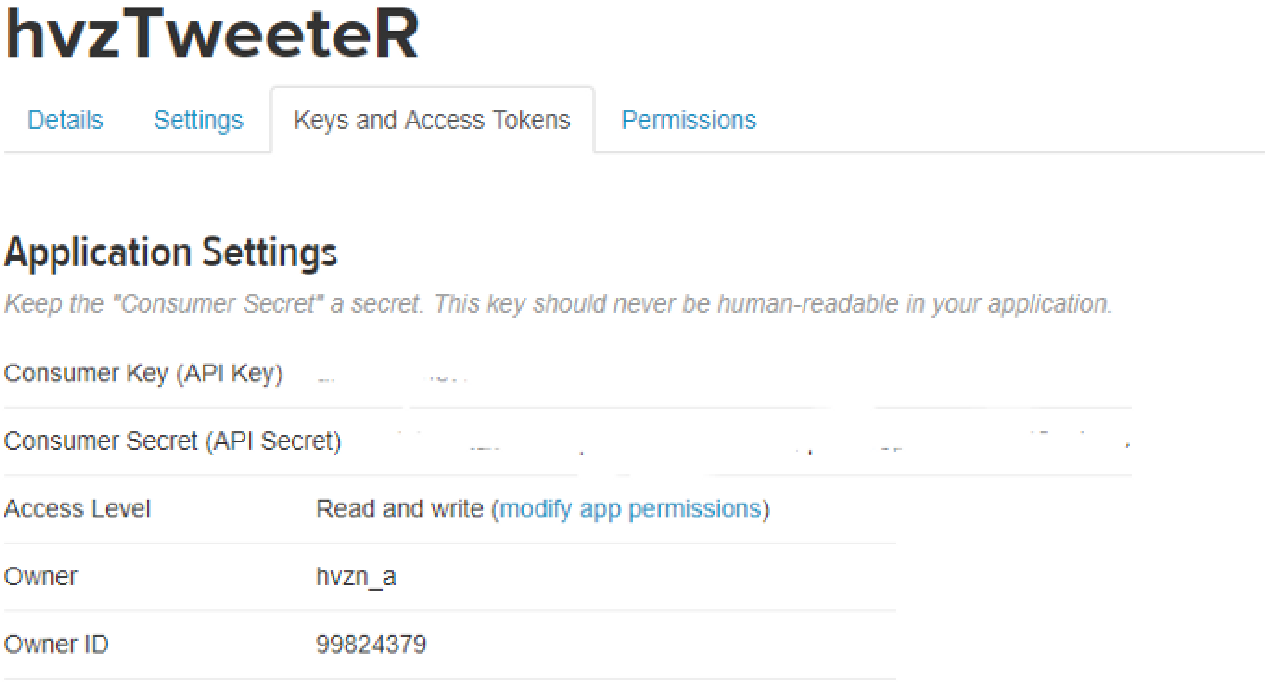
1. Membuat Twitter API : apps.twitter.com
2. Click Create New Apps > Fill everything according to your data > Open the App



hvzTweeteR

This app use for data mining, trainin on R scrapping data

Step by Step to Retrieving Data from Twitter



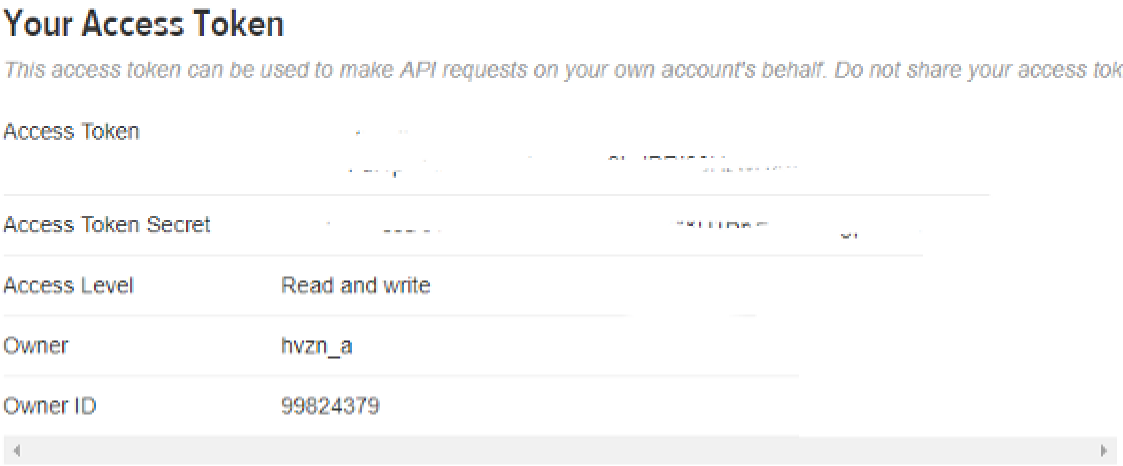
hvzTweeteR

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	
Consumer Secret (API Secret)	
Access Level	Read and write (modify app permissions)
Owner	hvzn_a
Owner ID	99824379

Step by Step to Retrieving Data from Twitter

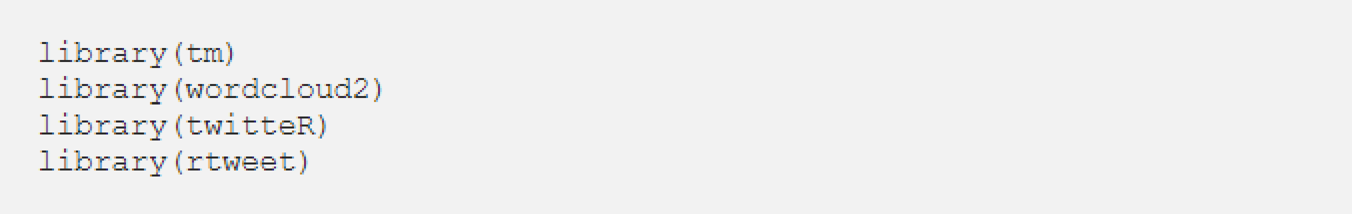


Your Access Token

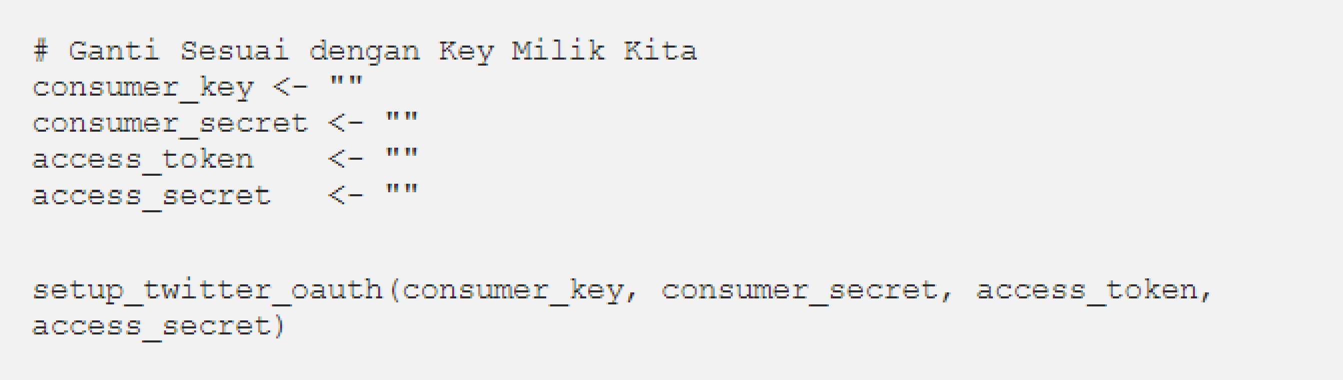
This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	
Access Token Secret	
Access Level	Read and write
Owner	hvzn_a
Owner ID	99824379

Step by Step to Retrieving Data from Twitter



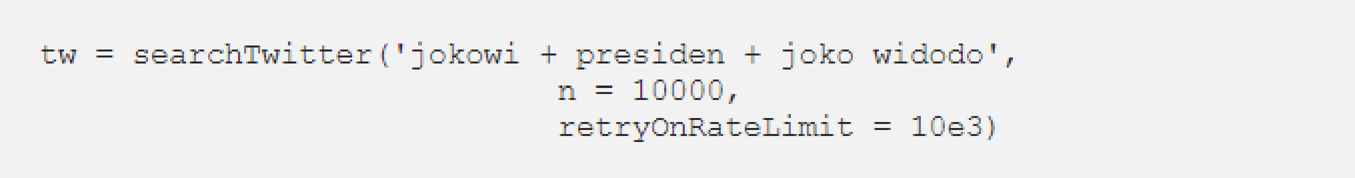
```
library(tm)
library(wordcloud2)
library(twitteR)
library(rtweet)
```



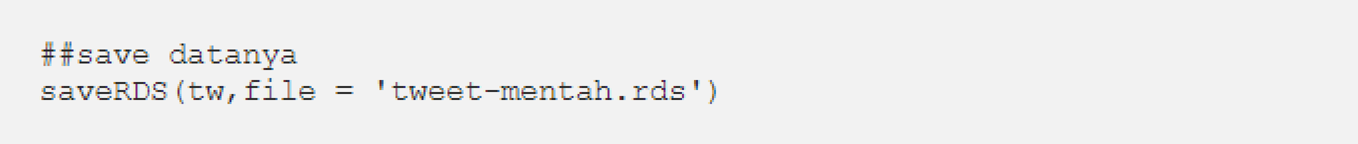
```
# Ganti Sesuai dengan Key Milik Kita
consumer_key <- ""
consumer_secret <- ""
access_token    <- ""
access_secret   <- ""


setup_twitter_oauth(consumer_key, consumer_secret, access_token,
access_secret)
```

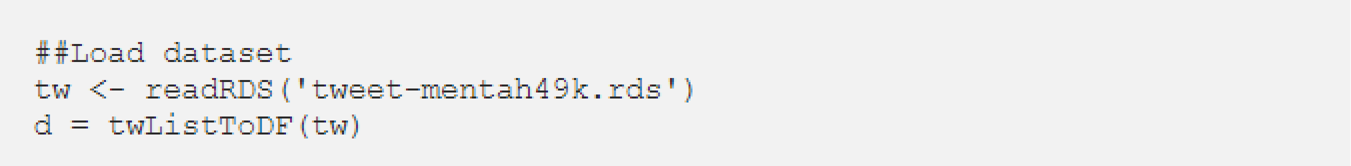
Step by Step to Retrieving Data from Twitter



```
tw = searchTwitter('jokowi + presiden + joko widodo',
                            n = 10000,
                            retryOnRateLimit = 10e3)
```



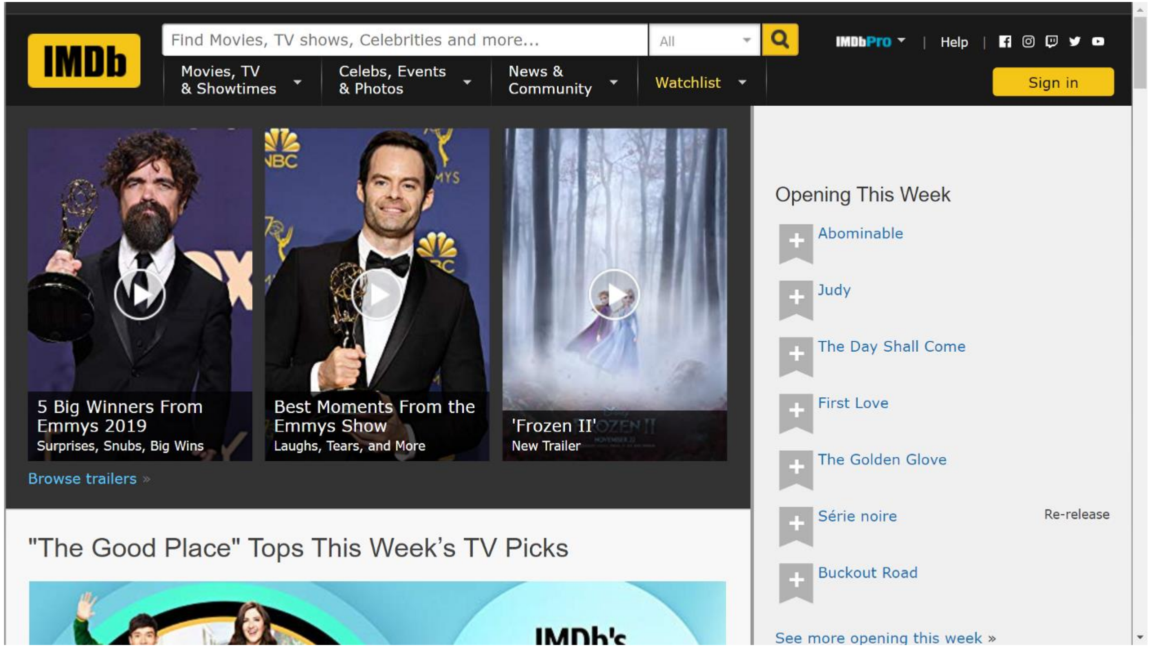
```
##save datanya
saveRDS(tw,file = 'tweet-mentah.rds')
```



```
##Load dataset
tw <- readRDS('tweet-mentah49k.rds')
d = twListToDF(tw)
```

Techniques for retrieving Data on the web and social media with R

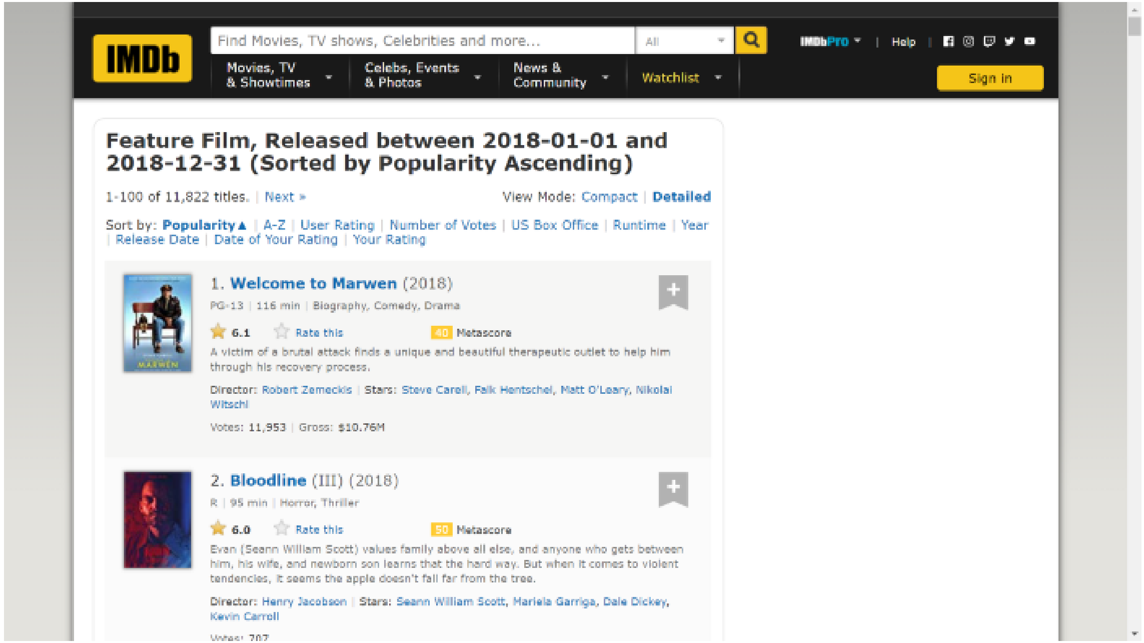
Step by Step to Retrieving Data from Web



Techniques for retrieving Data on the web and social media with R

Step by Step to Retrieving Data from Web

Then we want to see the data between gross (gross income) with runtime and genre



Feature Film, Released between 2018-01-01 and 2018-12-31 (Sorted by Popularity Ascending)

1-100 of 11,822 titles. | Next »

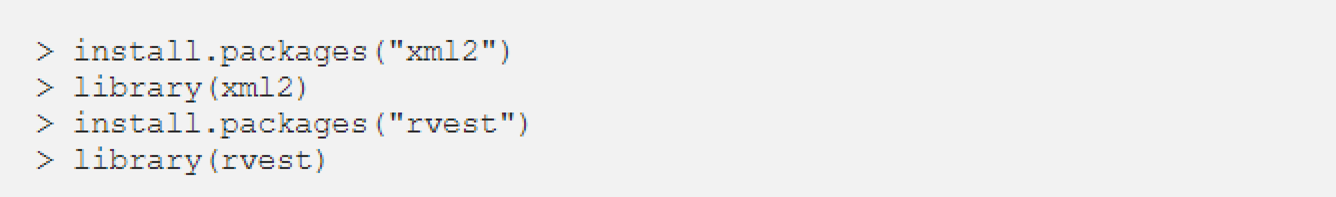
View Mode: Compact | Detailed

Sort by: Popularity ▲ | A-Z | User Rating | Number of Votes | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating

1. Welcome to Marwen (2018)
PG-13 | 116 min | Biography, Comedy, Drama
6.1 Rate this 40 Metascore
A victim of a brutal attack finds a unique and beautiful therapeutic outlet to help him through his recovery process.
Director: Robert Zemeckis | Stars: Steve Carell, Falk Hentschel, Matt O'Leary, Nikolai Witschl
Votes: 11,953 | Gross: $10.76M

2. Bloodline (III) (2018)
R | 95 min | Horror, Thriller
6.0 Rate this 50 Metascore
Evan (Seann William Scott) values family above all else, and anyone who gets between him, his wife, and newborn son learns that the hard way. But when it comes to violent tendencies, it seems the apple doesn't fall far from the tree.
Director: Henry Jacobson | Stars: Seann William Scott, Mariela Garriga, Dale Dickey, Kevin Carroll
Votes: 707

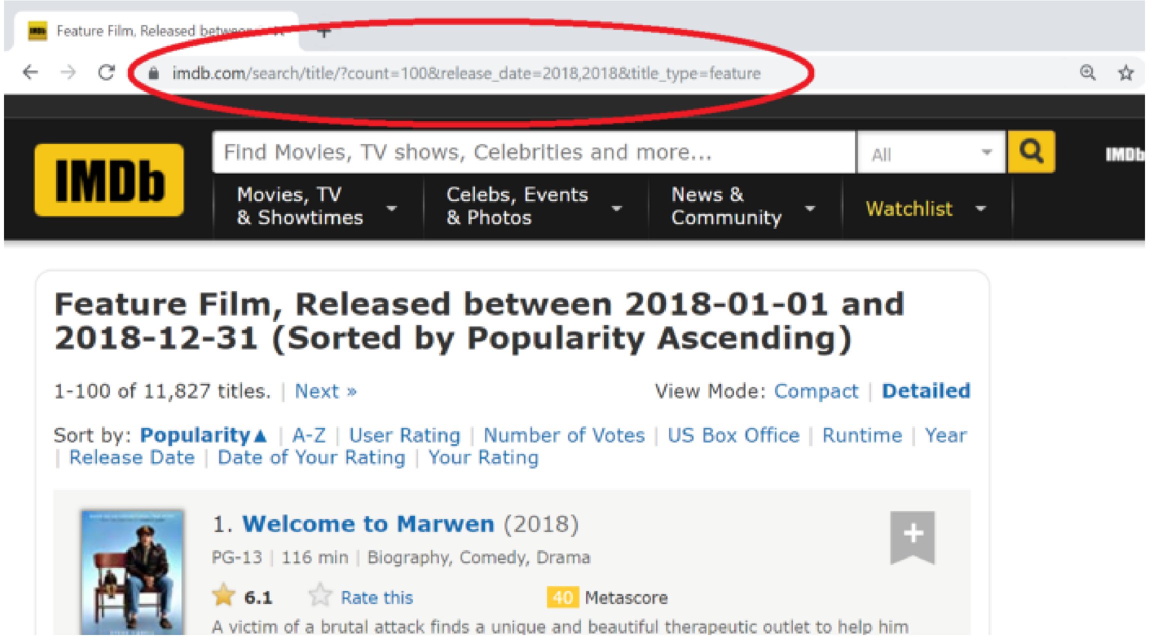
Step by Step to Retrieving Data from Web



```
> install.packages("xml2")
> library(xml2)
> install.packages("rvest")
> library(rvest)
```

Techniques for retrieving Data on the web and social media with R

Step by Step to Retrieving Data from Web


Feature Film, Released between ... ×
IMDb
Find Movies, TV shows, Celebrities and more...
All
Movies, TV & Showtimes
Celebs, Events & Photos
News & Community
Watchlist
IMDb

imdb.com/search/title/?count=100&release_date=2018,2018&title_type=feature

Feature Film, Released between 2018-01-01 and 2018-12-31 (Sorted by Popularity Ascending)

1-100 of 11,827 titles. | Next »

View Mode: Compact | **Detailed**

Sort by: **Popularity▲** | A-Z | User Rating | Number of Votes | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating

1. **Welcome to Marwen** (2018)

PG-13 | 116 min | Biography, Comedy, Drama

★ 6.1 ☆ Rate this 40 Metascore

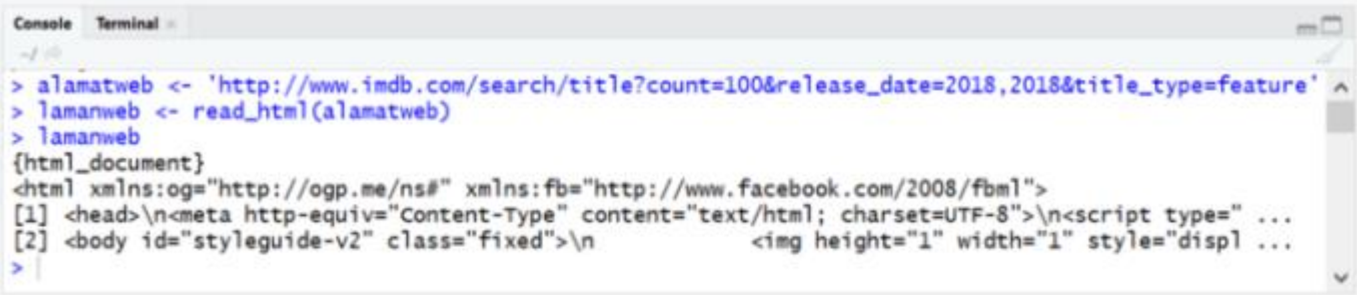
A victim of a brutal attack finds a unique and beautiful therapeutic outlet to help him

Techniques for retrieving Data on the web and social media with R

Step by Step to Retrieving Data from Web

```
> alamatweb <- '  https://www.imdb.com/search/title/?
count=100&release_date=2018,2018&title_type=feature'

> lamanweb <- read_html(alamatweb)
> lamanweb
```

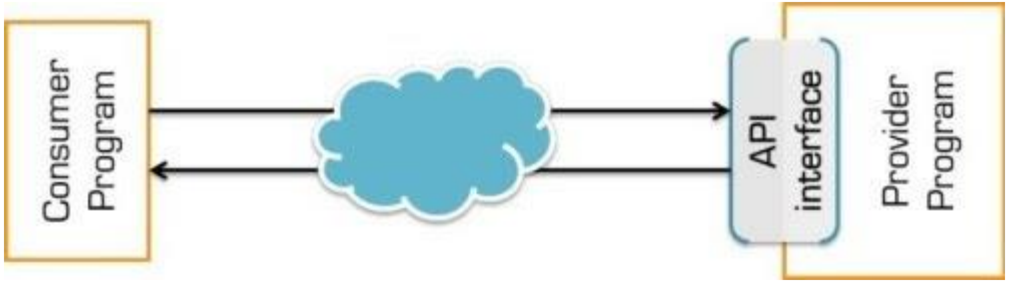


```
> alamatweb <- 'http://www.imdb.com/search/title?count=100&release_date=2018,2018&title_type=feature'
> lamanweb <- read_html(alamatweb)
> lamanweb
{html_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fbml">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n<script type=" ...
[2] <body id="styleguide-v2" class="fixed">\n              <img height="1" width="1" style="displ ...
> 
```

Understand the Basic Principles of API

What is API?

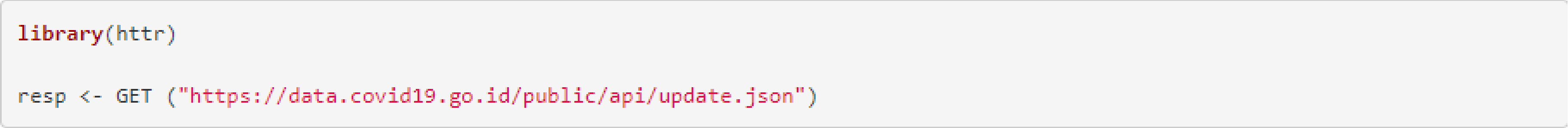
API or Application Programming Interface is an interface that can connect one application with another application. Thus, the API acts as an intermediary between different applications, either within the same platform or across platforms.



Import Realtime Data with API

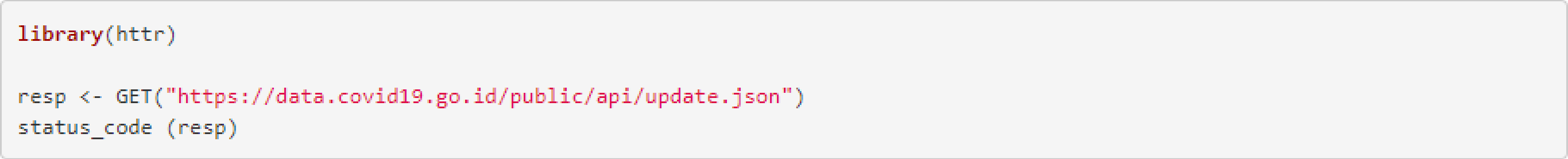
Accessing the covid.go.id API

Link : https://data.covid19.go.id/public/api/update.json.



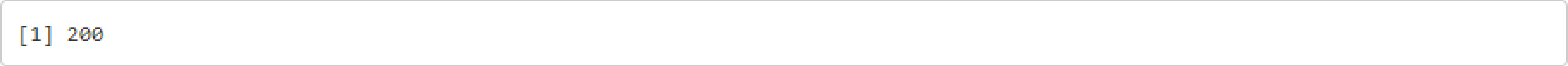
```
library(httr)

resp <- GET ("https://data.covid19.go.id/public/api/update.json")
```



```
library(httr)

resp <- GET("https://data.covid19.go.id/public/api/update.json")
status_code (resp)
```

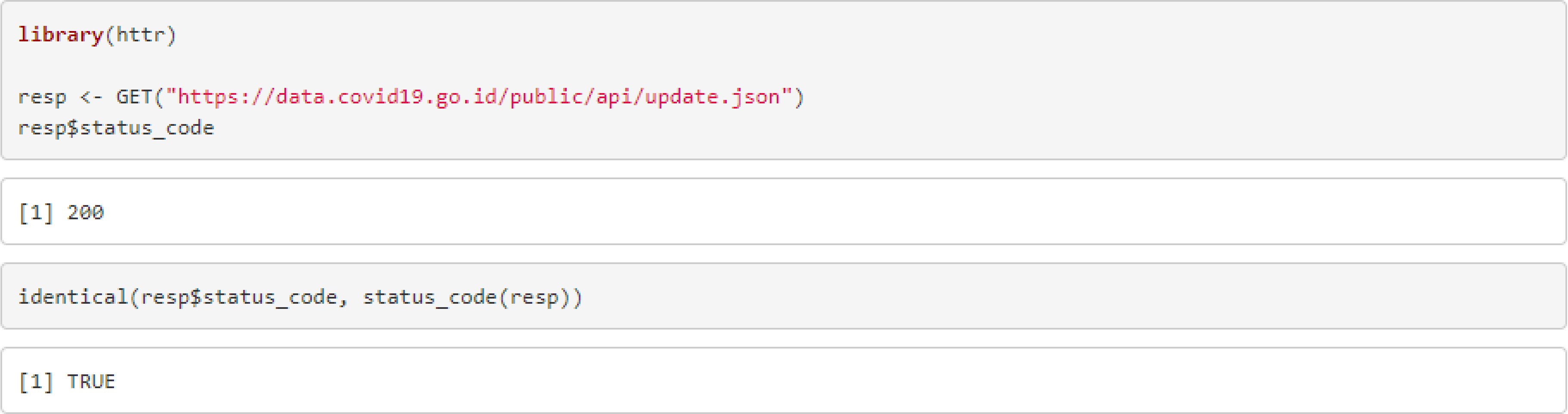


```
[1] 200
```

Import Realtime Data with API

Accessing the covid.go.id API

Link : https://data.covid19.go.id/public/api/update.json.



```
library(httr)

resp <- GET("https://data.covid19.go.id/public/api/update.json")
resp$status_code
```

```
[1] 200
```

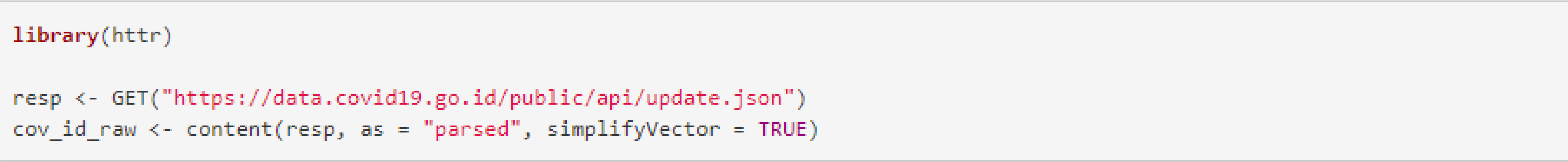
```
identical(resp$status_code, status_code(resp))
```

```
[1] TRUE
```

Import Realtime Data with API

Accessing the covid.go.id API

Link : https://data.covid19.go.id/public/api/update.json.



```
library(httr)

resp <- GET("https://data.covid19.go.id/public/api/update.json")
cov_id_raw <- content(resp, as = "parsed", simplifyVector = TRUE)
```

Referensi

https://rstudio-pubs-static.s3.amazonaws.com/288283_5e931dbca8ba4e72a43832cc5738b06c.html
https://www.sayonetech.com/blog/advanced-crawling-techniques/
https://www.radwarebotmanager.com/what-are-the-different-scraping-techniques/
https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/#h2_8
https://en.wikipedia.org/wiki/Web_crawler
https://en.wikipedia.org/wiki/Web_scraping
https://www.promptcloud.com/blog/data-scraping-vs-data-crawling/
https://www.quora.com/What-are-the-biggest-differences-between-web-crawling-and-web-scraping
https://stackoverflow.com/questions/4327392/what-is-the-difference-between-web-crawling-and-web-scraping

Quiz / Games : kahoot.it

Hands On Session

#JADIJAGOANDIGITAL

TERIMA KASIH

digitalent.kominfo

DTS_kominfo

digitalent.kominfo

digital talent scholarship