



Web Scrapping Dengan Python

BY WIDODO · PUBLISHED FEBRUARY 3, 2021 · UPDATED MARCH 5, 2021

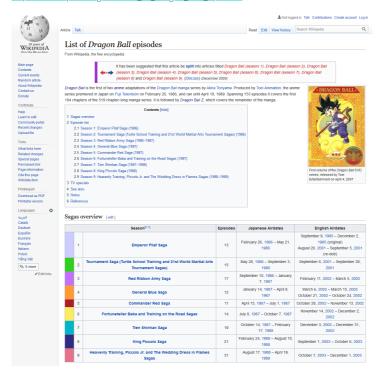
Web Scrapping Dengan Python

Web Scrapping merupakan proses untuk meng-ekstrak informasi (teks) dari website dan halaman online. Merupakan metode untuk mengambil informasi serta dataset untuk di analisis lebih lanjut.

Kali ini kita akan mencoba untuk melakukan ekstraksi informasi menggunakan python serta library Beautiful Soup. Beautiful Soup merupakan library yang banyak digunakan untuk melakukan parsing elemen html. Sederhananya, library ini digunakan untuk menerjemahkan elemen tag html untuk diambil isi teks-nya.

Mari Kita Coba

Pada praktek kali ini, mari kita coba untuk meng-ekstrak informasi mengenai daftar episode anime Dragon Ball. Setelah mencoba mencari informasi mengenai daftar episode anime Dragon Ball, kita dapatkan bahwa informasi tersebut dapat diambil dari website wikipedia di link berikut https://en.wikipedia.org/wiki/List_of_Dragon_Ball_episodes.



Dari sejumlah halaman yang ditampilkan, kita akan coba untuk meng-ekstrak tabel saga overview seperti gambar diatas.

Langkah Pertama

Pastikan library Beautiful Soup dan Request sudah terinstall, apabila belum jalankan perintah dibawah ini di console python. Library Requests berfungsi untuk mengambil elemen HTML dari sebuah alamat website yang diberikan, dan ini akan menjadi input bagi Beautiful Soup untuk melakukan parsing halaman.

pip install beautifulsoup4
pip install requests

KATEGORI

Big Data (1)

Cloud (4)

Coding (1)

Events (6)

Machine Learning (1)

NLP (1)

Opensource (1)

Python (4)

Teknologi Informasi (8)

ADLAS

API Application Programming Interface big data Cloud coding Internet JuaraGCP machine learning nip opensource python teknologi informasi Twitter visualisa:

data

Web Scrapping Dengan Python - Fani Widodo

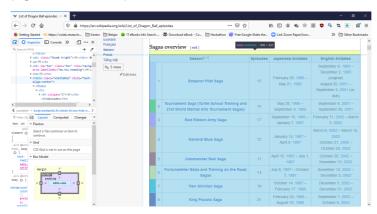
Setelah berhasil menginstall dua library diatas, lakukan import library untuk menguji apakah library tersebut sudah terinstall dengan baik dan selanjutnya bisa kita gunakan dalam kode python kita.

```
import pandas as pd
from bs4 import BeautifulSoup
from urllib.request import Request, urlopen
```

Langkah Kedua

Lakukan pengecekan melalui browser, pada elemen manakah tabel yang akan kita akstrak tersebut berada. Apabila menggunakan Firefox/ Chrome tekan CTR+Shift+I atau dengan melakukan klik kanan pada mouse dan pilih *Inspect Element*.

Dari informasi yang ditampilkan pada jendela *Inspect Element* tersebut diketahui bahwa tabel yang akan kita ekstrak berada pada elemen table class="wikitable".



Langkah Ketiga

Setelah mengetahui di elemen mana informasi tersebut akan diambil, saatnya kita melakukan pemanggilan dan ekstraksi dari halaman tersebut. Panggil halaman menggunakan library request:

```
alamat = "https://en.wikipedia.org/wiki/List_of_Dragon_Ball_episodes"
req = Request(alamat, headers={'User-Agent': 'Mozilla/5.0'})
```

Perintah diatas akan melakukan pemanggilan pada alamat wikipedia. Argumen headers digunakan agar panggilan kita seolah-olah berasal dari sebuah web browser (yaitu 'Mozilla/5.0'). Argumen headers ini sifatnya opsional, beberapa website tidak mensyaratkan hal tersebut tetapi ada beberapa website lainnya yang melakukan pengecekan agar halaman website mereka hanya bisa dibuka menggunakan browser yang *legitimate*.

Langkah Keempat

Setelah halaman web berhasil dibuka, jalankan Beautiful Soup untuk melakukan ekstraksi tabel yang akan kita ambil.

```
html = urlopen(req).read()
data = BeautifulSoup(html, 'html.parser')
table = data.find_all('table', ('class': 'wikitable'))[0]
```

Variabel *table* menyimpan hasil pengambilan elemen tabel pertama yang merupakan tabel yang ingin kita ambil dari halaman website tersebut. Apabila kita coba tampilkan variabel tersebut penggunakan fungsi print maka akan menghasilkan tampilan seperti berikut: print (table)

Selanjutnya kita akan coba simpan masing-masing baris dari tabel tersebut kedalam sebuah list bernama rows.

```
rows = table.findAll('tr')
```

Lakukan iterasi pada variabel row untuk memecah masing-masing kolom menjadi sebuah list yang kemudian disimpan kedalam variabel hasil.

```
hasil = []
for row in rows:
    cols = row.find_all('td')
    cols = [ele.text.strip() for ele in cols]
    hasil.append(cols)
```

Buat dataframe dari variabel list hasil dan tampilkan hasilnya.

```
df_hasil = pd.DataFrame(hasil)
df hasil
```

Maka akan menghasilkan tampilan seperti berikut ini.



Sudah lumayan menyerupai data yang diambil dari wikipedia bukan? Tetap ada beberapa hal yang harus diperbaiki dari dataframe yang telah dihasilkan, pertama adalah hapus kolom satu karena tidak menyimpan informasi apapun. Kemudian selanjutnya adalah memberi judul masing-masing kolom dari tabel kita. Terakhir adalah menghapus baris awal dari tabel kita karena terlihat disana tidak memiliki data/ none.

Jalankan perintah berikut untuk melakukan 3 hal tersebut.

```
df_hasil = df_hasil.drop([0], axis=1) # menghapus kolom pertama
df_hasil = df_hasil.drop([0], axis=0) # menghapus baris pertama
df_hasil.columns = ['No', 'Season', 'Episodes', 'Japanese Airdates',
'English Airdates'] # memberi judul kolom tabel
```

Maka dari perintah diatas akan menghasilkan dataframe seperti gambar dibawah ini.

lo	Season	Episodes	Japanese Airdates	English Airdates
1	Emperor Pilaf Saga	13	February 26, 1986 – May 21, 1986	September 9, 1995 - December 2, 1995 (original
2	Tournament Saga (Turtle School Training and 21	15	May 28, 1986 - September 3, 1986	September 6, 2001 - September 26, 2001
3	Red Ribbon Army Saga	17	September 10, 1986 – January 7, 1987	February 11, 2002 - March 5, 2002
4	General Blue Saga	12	January 14, 1987 - April 8, 1987	March 6, 2002 - March 15, 2002October 21, 2002
5	Commander Red Saga	11	April 15, 1987 – July 1, 1987	October 28, 2002 - November 13, 2002
6	Fortuneteller Baba and Training on the Road Sagas	14	July 8, 1987 - October 7, 1987	November 14, 2002 - December 2, 2002
7	Tien Shinhan Saga	19	October 14, 1987 - February 17, 1988	December 3, 2002 - December 31, 2002
8	King Piccolo Saga	21	February 24, 1988 - August 10, 1988	September 1, 2003 - October 6, 2003
9	Heavenly Training, Piccolo Jr. and The Wedding	31	August 17, 1988 - April 19, 1989	October 7, 2003 - December 1, 2003
	1 2 3 4 5 6 7	1 Emperor Pilal Saga 2 Tournament Saga (Turtle School Training and 21 3 Red Ribbon Army Saga 4 General Blue Saga 5 Commander Red Saga 6 Fortuneteller Baba and Training on the Road Sagas 7 Tien Shinhan Saga 8 King Piccolo Saga	1 Emperor Pilaf Saga 13 2 Tournament Saga (Tutle School Training and 21 15 3 Red Ribbon Army Saga 17 4 General Blue Sagaa 12 5 Commander Red Sagaa 11 6 Fontuneteller Baba and Training on the Road Sagaa 14 7 Tien Shinhaan Saga 19 8 King Piccolo Saga 21	1 Emperor Pilaf Saga 13 February 26, 1986 – May 21, 1986 2 Tournament Saga (Tutle School Training and 21 15 May 28, 1986 – September 3, 1986 3 Red Ribbon Army Saga 17 September 10, 1986 – January 7, 1987 4 General Bilus Sagas 12 January 14, 1987 – April 8, 1987 5 Commander Red Sagas 11 April 15, 1987 – April 8, 1987 – October 7, 1987 7 Tien Shinhaan Saga 19 October 14, 1987 – February 17, 1988 King Piccolo Saga 21 February 24, 1988 – August 10, 1988

Web Scrapping Dengan Python - Fani Widodo

Coba bandingkan hasilnya dengan halaman website wikipedia.

Season ⁽ⁿ⁻¹⁾		Episodes	Japanese Airdates	English Airdates	
	1	Emperor Pilaf Saga	13	February 26, 1986 – May 21, 1986	September 9, 1995 – December 2, 1995 (original) August 20, 2001 – September 5, 2001 (re-dub)
	2	Tournament Saga (Turtle School Training and 21st World Martial Arts Tournament Sagas)	15	May 28, 1986 – September 3, 1986	September 6, 2001 – September 26, 2001
	3	Red Ribbon Army Saga	17	September 10, 1986 – January 7, 1987	February 11, 2002 – March 5, 2002
	4	General Blue Saga	12	January 14, 1987 – April 8, 1987	March 6, 2002 – March 15, 2002 October 21, 2002 – October 24, 2002
	5	Commander Red Saga	11	April 15, 1987 – July 1, 1987	October 28, 2002 - November 13, 2002
	6	Fortuneteller Baba and Training on the Road Sagas	14	July 8, 1987 – October 7, 1987	November 14, 2002 – December 2, 2002
	7	Tien Shinhan Saga	19	October 14, 1987 – February 17, 1988	December 3, 2002 – December 31, 2002
	8	King Piccolo Saga	21	February 24, 1988 – August 10, 1988	September 1, 2003 – October 6, 2003
	9	Heavenly Training, Piccolo Jr. and The Wedding Dress in Flames Sagas	31	August 17, 1988 – April 19, 1989	October 7, 2003 – December 1, 2003

Langkah terakhir adalah menyimpan hasil tersebut kedalam file excel.

df_hasil.to_excel('data_serial_dragonball.xlsx', index=False)

Tags: python

YOU MAY ALSO LIKE...



in 1 100 3 1 penguatan yang dengan teknologi yang dengan tuk menjadi indonesia dan - prof bambang dunia

Membuat Wordcloud Dengan



Melakukan Visualisasi Data Kasus COVID19 di Indonesia

Python

Webinar Sejarah Internet Indonesia

MARCH 15, 2021

MARCH 1, 2021

JULY 20, 2020

Fani Widodo © 2022. All Rights Reserved.

Powered by 🕅 - Designed with the Hueman theme