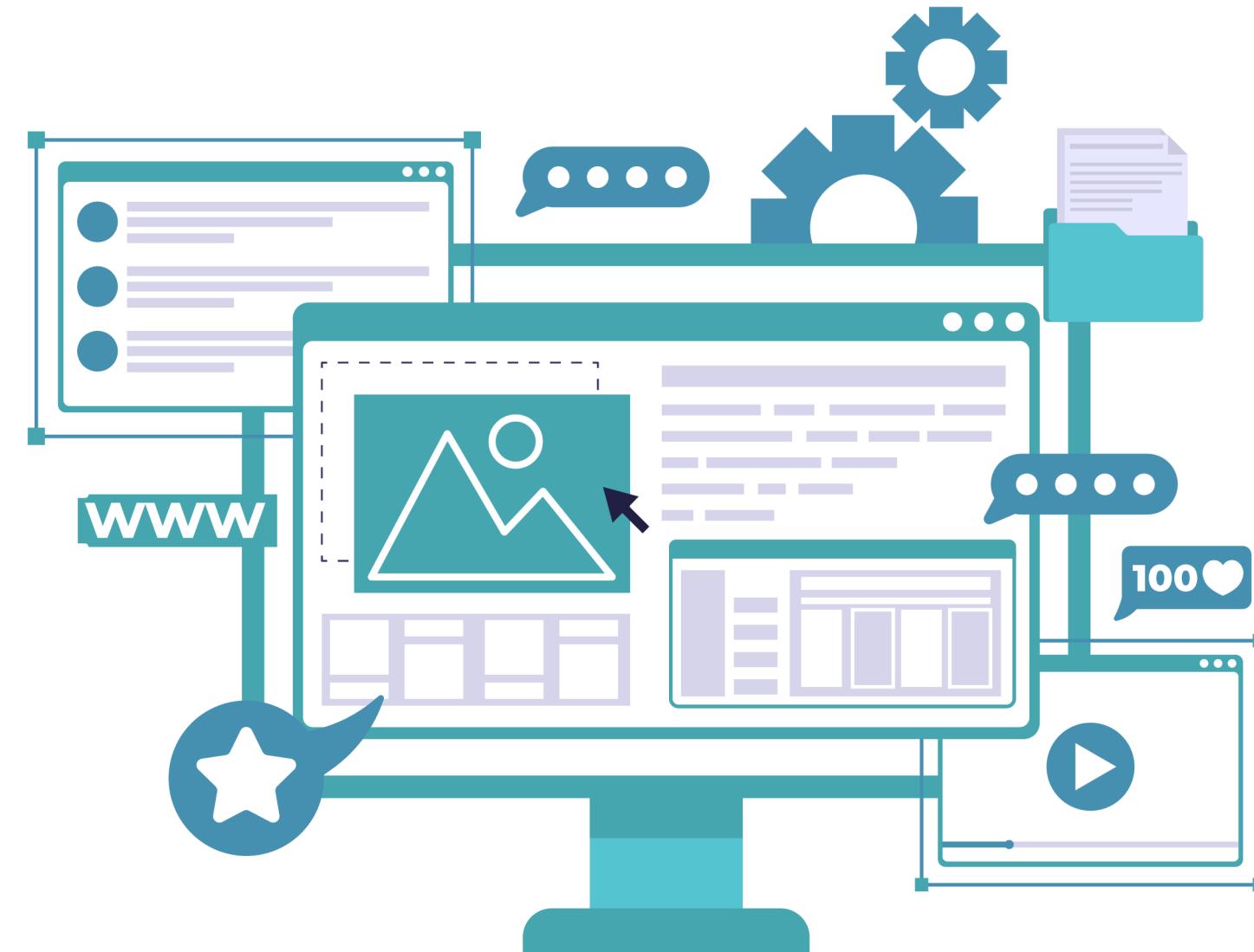


BIG DATA - RPL

WEB SCRAPING

Irfan Triadi Saputra, S.Tr.Kom





APA ITU WEB SCRAPING?

Web scraping adalah teknik yang digunakan untuk mengambil data dari situs web secara otomatis. Data yang diambil biasanya berupa teks, gambar, atau informasi lain yang ditampilkan di halaman web. Web scraping sering digunakan untuk mengambil data dari beberapa halaman secara cepat dan efisien tanpa harus melakukannya secara manual.

Web scraping sangat berguna ketika kita perlu mengambil informasi yang tidak disediakan secara resmi melalui API, seperti harga barang di situs e-commerce, berita terbaru, atau data dari forum.

PRINSIP KERJA WEB SCRAPING

Web scraping bekerja dengan mengirimkan permintaan HTTP ke situs web, mengunduh halaman HTML, lalu menguraikan konten HTML tersebut untuk mengekstrak informasi yang relevan. Proses ini bisa dilakukan secara manual atau otomatis dengan bantuan skrip atau tool scraping.



FUNGSI SCRAPING

- **Mengumpulkan Data dari Berbagai Sumber:** Mengambil data dari berbagai situs web untuk mengisi data lake atau data warehouse.
- **Menyediakan Data untuk Machine Learning dan AI:** Menyediakan data dalam jumlah besar untuk melatih model machine learning dan AI.
- **Memantau Perkembangan Pasar dan Kompetitor:** Mengamati harga, ulasan pelanggan, dan fitur produk kompetitor untuk analisis pasar.
- **Menyediakan Data untuk Analisis Tren:** Mengambil data dari situs berita, media sosial, dan blog untuk mendekripsi tren terbaru.
- **Mengumpulkan Data untuk Pemetaan Sosial dan Ekonomi:** Menggunakan data sosial untuk memahami perilaku masyarakat dan pola konsumsi.
- **Data Real-time untuk Sistem Rekomendasi:** Mengambil data harga dan ketersediaan layanan secara real-time untuk rekomendasi terbaik.
- **Penelitian dan Analisis Data Tekstual untuk Big Data:** Menggunakan data teks dari berbagai sumber untuk analisis NLP dan konten.
- **Pengambilan Data untuk Analisis Risiko:** Mengambil data dari berbagai sumber untuk penilaian risiko di sektor keuangan dan asuransi.

PROSES SCRAPING DAPAT DIBAGI MENJADI BEBERAPA TAHAP:

- **Mengirim Permintaan (Request):** Skrip mengirimkan HTTP request ke URL tertentu dan menerima respon dalam bentuk HTML dari server.
- **Mengurai HTML (Parsing):** HTML yang diterima diuraikan menggunakan library untuk mengambil elemen-elemen yang diinginkan.
- **Ekstraksi Data:** Setelah elemen-elemen HTML diidentifikasi, data diekstrak dari HTML tersebut.
- **Penyimpanan Data:** Data yang sudah diekstrak kemudian disimpan ke dalam format yang lebih mudah diolah, seperti CSV, database, atau JSON.

PERALATAN WEB SCRAPING

- **Bahasa Pemrograman:** Python adalah bahasa yang paling populer untuk web scraping karena mudah digunakan dan memiliki banyak library pendukung.
- **Library Parsing HTML:** Library yang digunakan untuk menguraikan dan mengekstrak elemen-elemen dari HTML, seperti BeautifulSoup atau lxml.
- **Library Pengiriman Permintaan (Request):** Library yang digunakan untuk mengirimkan HTTP request dan menerima respons dari server, seperti requests.

TOOLS UNTUK WEB SCRAPING:

Libraries: Beberapa library yang umum digunakan untuk web scraping adalah:

- **Python:** BeautifulSoup, Scrapy, Selenium, Requests
- **JavaScript:** Cheerio, Puppeteer

Browser Developer Tools: Alat ini digunakan untuk memahami struktur HTML dari halaman web yang ingin kita scraping.

TOOLS UNTUK WEB SCRAPING:

Libraries: Beberapa library yang umum digunakan untuk web scraping adalah:

- **Python:** BeautifulSoup, Scrapy, Selenium, Requests
- **JavaScript:** Cheerio, Puppeteer

Browser Developer Tools: Alat ini digunakan untuk memahami struktur HTML dari halaman web yang ingin kita scraping.

KEUNGGULAN BEAUTIFULSOUP4

- **Sintaks yang Mudah dan Intuitif:** Mudah dipelajari dan digunakan, bahkan bagi pemula.
- **Penanganan Struktur HTML yang Kompleks:** Mampu menangani struktur HTML yang tidak sempurna (misalnya, elemen HTML yang hilang atau berulang).
- **Navigasi Tag yang Fleksibel:** Mendukung berbagai metode untuk menavigasi elemen, seperti find(), find_all(), dan metode navigasi hierarkis lainnya.
- **Kompatibilitas dengan Parser:** BeautifulSoup4 mendukung beberapa parser seperti html.parser (standar Python), lxml, dan html5lib. Masing-masing parser memiliki kelebihan dalam kecepatan atau keakuratan parsing.
- **Ekosistem yang Kuat:** Bisa dengan mudah diintegrasikan dengan pustaka requests untuk mengambil data dari web atau dengan pandas untuk analisis data.

BEAUTIFULSOUP4

BeautifulSoup4 adalah pustaka Python yang digunakan untuk mengambil dan memanipulasi data dari halaman web (HTML atau XML). Pustaka ini memudahkan kita dalam parsing data dari struktur HTML yang kompleks, sehingga informasi tertentu (seperti teks, link, atau atribut) dapat diambil dengan mudah. BeautifulSoup4 biasanya digunakan bersama pustaka requests untuk melakukan web scraping.

LEGALITAS WEB SCRAPING

Sebelum melakukan web scraping, penting untuk memastikan bahwa tindakan ini sesuai dengan kebijakan situs yang ingin diambil datanya. Beberapa situs juga melarang scraping dalam syarat dan ketentuan penggunaannya.



TAHAPAN MELAKUKAN WEB SCRAPING

1. Menentukan Target

Pilih situs web yang datanya ingin Anda ambil. Pastikan Anda memahami struktur HTML situs tersebut. Gunakan alat inspeksi elemen di browser (seperti Inspect Element di Chrome) untuk mengidentifikasi elemen HTML yang ingin di-scrape.

2. Mendapatkan HTML dengan Mengirim Request

Gunakan library requests untuk mengambil konten halaman web yang akan di-scrape.

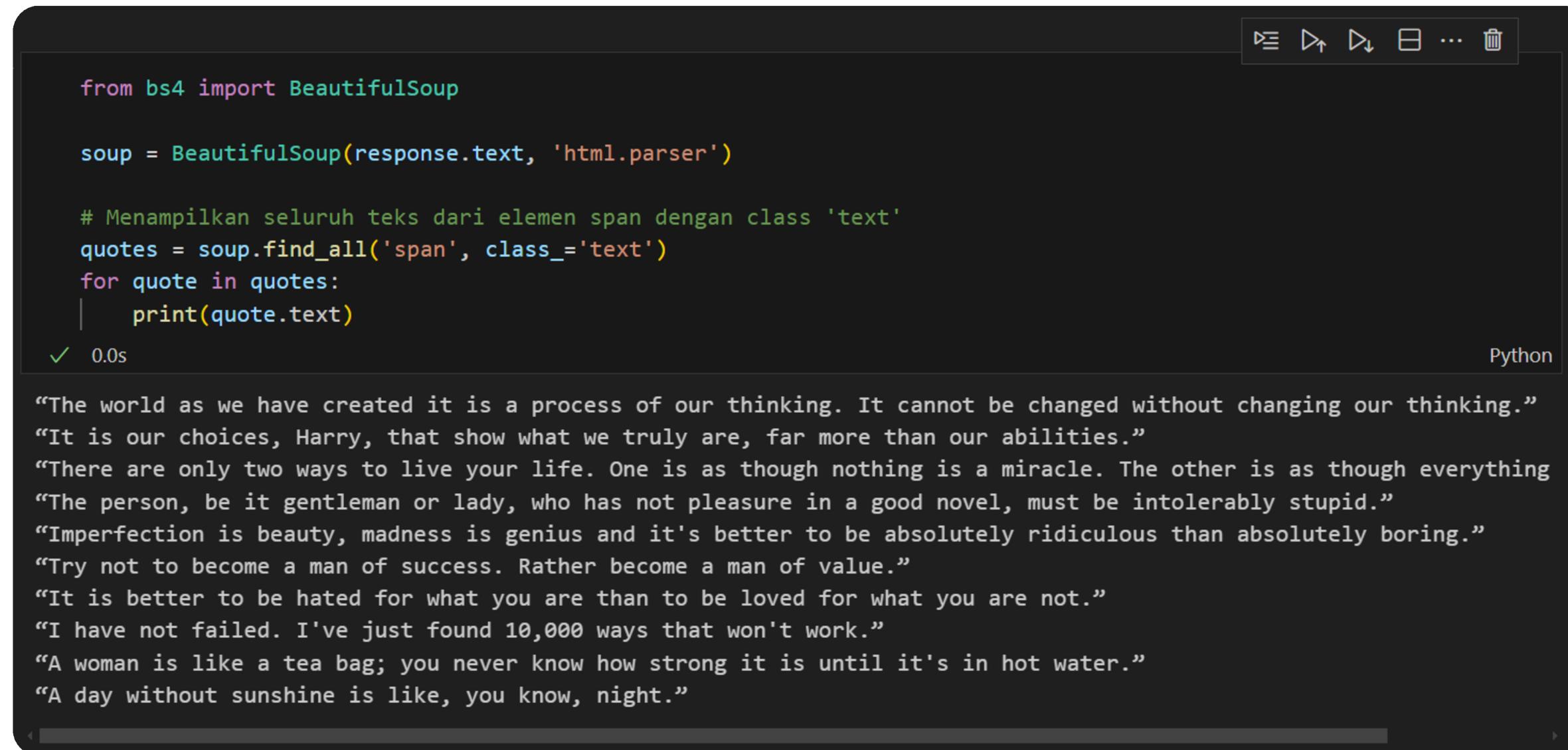
```
import requests

url = 'http://quotes.toscrape.com'
response = requests.get(url)
```

✓ 0.5s

3. Menguraikan HTML

Gunakan BeautifulSoup atau library lain untuk menguraikan HTML dan mengambil elemen-elemen yang diinginkan.



```
from bs4 import BeautifulSoup

soup = BeautifulSoup(response.text, 'html.parser')

# Menampilkan seluruh teks dari elemen span dengan class 'text'
quotes = soup.find_all('span', class_='text')
for quote in quotes:
    print(quote.text)
```

✓ 0.0s Python

The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.
It is our choices, Harry, that show what we truly are, far more than our abilities.
There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything
The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.
Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.
Try not to become a man of success. Rather become a man of value.
It is better to be hated for what you are than to be loved for what you are not.
I have not failed. I've just found 10,000 ways that won't work.
A woman is like a tea bag; you never know how strong it is until it's in hot water.
A day without sunshine is like, you know, night.

TANTANGAN DALAM WEB SCRAPING

Ada beberapa tantangan yang mungkin dihadapi ketika melakukan web scraping:

1. Struktur HTML yang Rumit

- Beberapa situs memiliki struktur HTML yang kompleks, sehingga sulit untuk menguraikan data yang dibutuhkan.
- Gunakan XPath atau CSS Selectors untuk mempermudah pemilihan elemen.

2. Perubahan Struktur Situs

- Situs dapat mengubah struktur HTML kapan saja, menyebabkan skrip scraping gagal. Anda perlu menyesuaikan skrip Anda secara berkala.

3. Anti-Scraping Measures

- Beberapa situs menggunakan teknik untuk mencegah scraping, seperti CAPTCHA, pemeriksaan IP, atau rate-limiting. Solusinya adalah menggunakan teknik seperti rotasi proxy atau bypass CAPTCHA.

4. Dynamic Content (JavaScript)

- Banyak situs menggunakan JavaScript untuk memuat konten secara dinamis. Dalam kasus ini, Anda bisa menggunakan tools seperti Selenium atau Playwright untuk melakukan scraping konten yang dimuat secara dinamis.