EDM Laboratory

# GWE #2

## Data Scraping

Brief Explanation

EDM Laboratory

# Sebenernya apa sih data scraping itu?

Adalah proses buat kita **mengoleksi data dari sumber-sumber data** buat kita simpan dan olah lebih lanjut lagi. Nah biasanya karena data sifatnya kredensial kita sebagai user awam akan memanfaatkan data2 dari internet / website nih maka sering dinamakan sebagai **Web Scraping.**

Jadi bahkan copy paste wikipedia itu udah termasuk web scraping loh! Tapi kan kalau data kita banyak, sampai kapan kita mau scraping??? Maka dari itu sebenernya esensinya ialah adanya **automasi.** Bayangin nih, misal kalian mau ambil list hotel di Traveloka, ngga mungkin dong dicopas satu persatu datanya...

ps: sebenernya ada banyak tipe dari data scraping itu, tapi kita akan coba pelajari web scraping dulu ya

EDM Laboratory

# Sebenernya apa sih web scraping itu?

Kita tau.. pasti gamau belajar kalo gaada alesan kuatnya kenapa web scraping??? Yang lainnya emang ga lebih kepake? Gini gais, jadi dengan Web Scraping itu adalah **cara paling mudah dan paling luas kegunaannya** buat numbuhin portofolio kalian. Dengan web scraping kalian **bisa dapetin data apapun**.. literally apapun **asal itu ada di internet**.

Misal deh, report mining itu ambil data user yang akses website... otomatis kita harus punya akses ke website itu kan padahal bukan punya kita tuh.. sulitt.

So kita harap, dengan belajar web scraping ini gabakal ada alasan pas mau bikin analisis atau project tapi bilang “wah datanya gaada”. BREAK THE LIMIT GUYS!

eh btw.. ga semua website bolehin datanya diambil dan bentuk webnya itu beda2. So its gonna be little bit tricky but fun.

GWE #2

# Cara Web Scraping

# Cara Web Scraping

EDM Laboratory

**API**

*bisa dipelajari lebih lanjut pada matkul EAI (Sem 6) atau melalui Youtube Pak Faqih

API (Application Programming Interface) sebuah layer yang memungkinkan sistem untuk berkomunikasi. Unsur sebuah komunikasi pasti ialah bertukar data. **Beberapa** website memang menyediakan API agar bisa bertukar data dengan protokol yang telah diaturnya

Dimana data akan didapatkan melalui **method GET** dan akan kita miliki dalam **berbagai format seperti JSON, XML, Protobuf** dan sebagainya.

Salah satunya pada platform rapidapi.com atau google cloud

pros: mudah, tidak bermasalah jika tampilan web berganti

cons: harus tersedia dulu API nya, biasanya berbayar / based on kerjasama

GWE #2

# Cara Web Scraping

EDM Laboratory

**Extension**

Extension chrome ituh...

webscraper.io

kalo mau nyobain namanya ini

pros: mudah, plug and play

cons: pasti ada keterbatasannya, example cuma bisa format .csv / per page

EDM Laboratory

# Cara Web Scraping

**nah enih yang kita pelajari kali ini**

## Python Library

Taukan kalo py itu salah satu bahasa yang powerful. Tinggal import import kita bisa segalanya azek

Kita akan pakai library **BeautifulSoup & requests**

Dengan cara parsing HTML (tampilan web) yang ada di website. Jadi kita bisa ambil data yang terlihat di web itu.

pros: sangat fleksibel, mostly bisa web apapun

cons: complex, tampilan berubah bisa saja cara berubah

EDM Laboratory

# Jadi, Apa yang perlu kita ketahui dulu?

## 1. Python Library

-> Kumpulan kode kode yang udah di compile dan dapat digunakan kembali untuk tugas-tugas tujuannya.

- **requests** -> untuk mendapatkan kode HTML (berinteraksi dengan web) pada pada program Python. Bisa juga handle authentication lho
- **BeautifulSoup** -> Library untuk parsing struktur data terutama di HTML dan XML. Sehingga bisa mengetahui elemen2 yang ada di website.

maka dari itu, kalau belum punya librarynya coba di “pip install [nama library]” dulu ya di terminal

## 2. Website Structure

-> Kita akan mengambil data berdasarkan struktur HTML nya, maka penting untuk tau ada apa saja sih elemen yang digunakan untuk menyimpan data di web. Serta elemen seperti class, href dan sebagainya.

-> Dengan BeautifulSoup kita akan mengambil isi data berdasarkan nama elemennya.

## 3. Python Syntax

-> Percayalah cheat sheet nanti akan banyak kepakai. Terutama kita akan pakai untuk looping, format data, simpan data, data wrangling dan sebagainya.

# Jadi, Apa yang perlu kita ketahui dulu?

**4. HTTP Response Code**

-> Kita dapat mengetahui apakah web itu boleh diambil atau tidak / apakah kita berhasil terhubung melalui response code yang kita terima.

# Jadi, Apa yang perlu kita ketahui dulu?

EDM Laboratory

## 4. Alur

### Request

1. Menyambungkan ke web tujuan dengan method GET
2. Mempelajari struktur html website
3. Mengkonfigurasi agar dapat mengambil data dari website

→

### Parse

1. Inspect elemen HTML website
2. Memilih elemen yang akan kita ambil dan perlukan untuk tujuan data
3. Scrape text

→

### Store

1. Simpan data yang telah berhasil di scrape ke dalam file / database
2. Melakukan otomasi pengambilan data

# Let’s try