Network Programming

Web scraping

(Special Topics)

STT Terpadu Nurul Fikri
Henry Saptono
2019

Web scraping ?

- Arti kata 'Scraping' dari kamus berarti **mendapatkan sesuatu dari web**.
- Ada dua pertanyaan:
    1. Apa yang bisa kita dapatkan dari web ?
    2. Bagaimana cara mendapatkannya ?

Web scraping ?

- untuk pertanyaan pertama adalah '**data**'. Data yang berguna sangat diperlukan dari setiap proyek pemrograman
- untuk pertanyaan kedua agak sulit, karena ada banyak cara untuk mendapatkan data. Secara umum, kita dapat memperoleh data dari basis data atau file data dan sumber lainnya.

Web scraping ?

- Bagaimana jika kita membutuhkan sejumlah besar data yang tersedia secara online?
    - Salah satu cara mendapatkan data semacam itu adalah dengan mencari secara manual (mengklik di browser web) dan menyimpan (menyalin-paste ke spreadsheet atau file) data yang diperlukan.
    - Cara lain untuk mendapatkan data tersebut adalah menggunakan **web scraping**

Web scraping ?

- ***Web scraping is an automatic process of extracting information from web.***
- Web scraping, dikenal juga sebagai penambangan data web (***web data mining***) atau memanen web (***web harvesting***), adalah proses membangun **agen** yang dapat **mengekstraksi, menguraikan, mengunduh,** dan **mengatur (menyusun)** informasi yang berguna dari web **secara otomatis**

Web Crawling v/s Web Scraping

- Istilah Web Crawling dan Scraping sering digunakan secara bergantian karena konsep dasarnya adalah untuk mengekstraksi data. Namun, mereka berbeda satu sama lain.
- Web crawling pada dasarnya digunakan untuk mengindeks informasi pada halaman menggunakan bot alias ***crawler***. Ini juga disebut pengindeksan.
- Web scraping adalah cara otomatis mengekstraksi informasi menggunakan bot alias **scrapers**. Ini juga disebut ekstraksi data.

Uses of Web Scraping

- Penggunaan dan alasan untuk menggunakan web scraping tidak ada habisnya seperti penggunaan World Wide Web.
- Web scraping dapat melakukan apa saja seperti memesan makanan online, memindai situs web belanja online untuk Anda, dan membeli tiket pertandingan saat tersedia, dll. seperti yang bisa dilakukan manusia.
- Data untuk Penelitian - Para peneliti dapat mengumpulkan data yang berguna untuk tujuan pekerjaan penelitian mereka dengan menghemat waktu mereka dengan proses otomatis ini.

Components of a Web Scraper

- Scraper web terdiri dari komponen-komponen berikut :
  - **Web Crawler Module.**

    Komponen yang sangat penting dari web scraper, modul web crawler, digunakan untuk menavigasi situs web target dengan membuat permintaan HTTP atau HTTPS ke URL. Scraper mengunduh data yang tidak terstruktur (konten HTML) dan meneruskannya ke extractor, modul berikutnya.

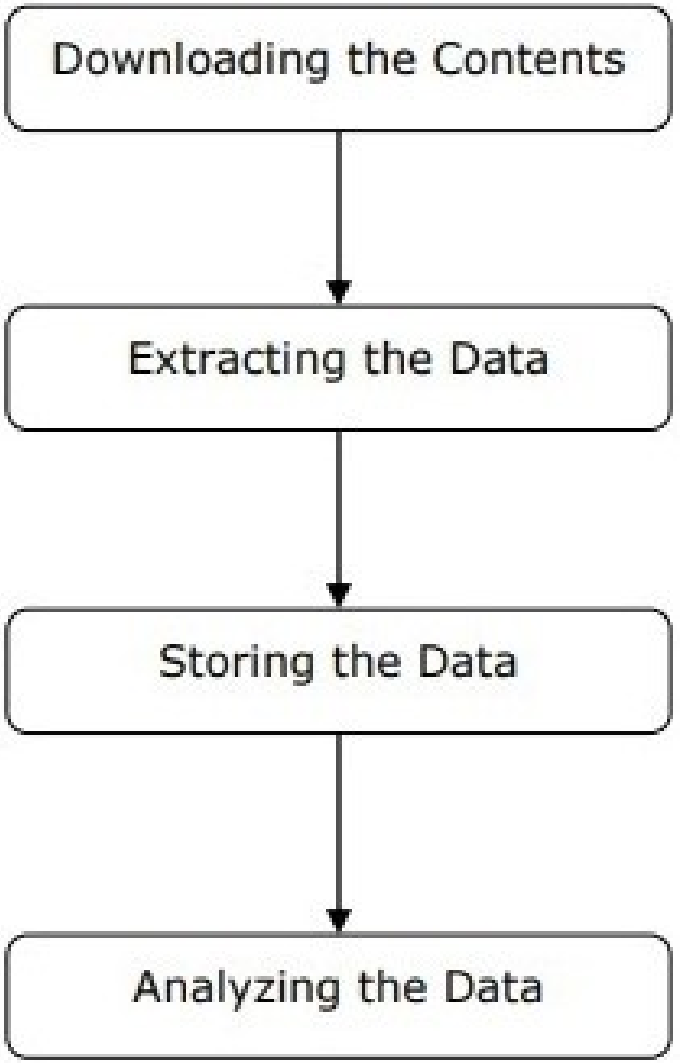
Components of a Web Scraper

- **Extractor**
  - Extractor memproses konten HTML yang diambil dan mengekstraksi data ke dalam format semi-terstruktur. Ini juga disebut sebagai modul parser dan menggunakan teknik parsing yang berbeda seperti ekspresi Reguler, Parsing HTML, parsing DOM atau Artificial Intelligence untuk fungsinya.
- **Data Transformation and Cleaning Module**
  - Data yang diekstraksi di atas belum siap digunakan. Itu harus melewati beberapa modul pembersih sehingga kita bisa menggunakannya. Metode seperti manipulasi string atau ekspresi reguler dapat digunakan untuk tujuan ini. Perhatikan bahwa ekstraksi dan transformasi dapat dilakukan dalam satu langkah juga.

Components of a Web Scraper

- **Storage Module**
  - Setelah mengekstraksi data, kita perlu menyimpannya sesuai kebutuhan kita. Modul penyimpanan akan menampilkan data dalam format standar yang dapat disimpan dalam database atau format JSON atau CSV.

Working of a Web Scraper



Web Scraping with python

- Python Modules for Web Scraping
  - Requests
    - pip install requests
  - Urllib3
    - pip install urllib3
  - BeautifulSoup4
    - pip install beautifulsoup4
  - Selenium
    - pip install selenium
  - Scrapy
    - pip install scrapy

Let's try to code