Modul Praktikum: Data Scraping untuk Pemrosesan Bahasa Alami (NLP)

Pendahuluan

Pemrosesan Bahasa Alami (NLP) adalah cabang ilmu komputer yang berkaitan dengan interaksi antara manusia dan komputer melalui bahasa manusia. Salah satu tahapan penting dalam NLP adalah pengumpulan data teks yang relevan untuk dianalisis dan diproses. Dalam modul ini, Anda akan belajar bagaimana melakukan data scraping atau pengambilan data dari sumber-sumber teks online untuk digunakan dalam proyek NLP.

Tujuan

- Memahami konsep dasar data scraping.
- Menggunakan alat dan pustaka yang tepat untuk scraping data teks dari situs web.
- Menerapkan keterampilan scraping untuk mengumpulkan data teks yang dapat digunakan dalam proyek NLP.

Requriment

- Pengetahuan dasar dalam bahasa pemrograman, terutama Python.
- Koneksi internet yang stabil.
- Python 3.x dan beberapa pustaka yang diperlukan telah diinstal di komputer Anda.

Modul Praktikum

I. Pengenalan Data Scraping

1. **Apa itu Data Scraping?**

   Konsep dasar data scraping.

   Data scraping adalah teknik untuk mengekstrak data dari sebuah website atau sistem tertentu. Data scraping biasanya juga disebut dengan data extraction. Proses ini dapat dilakukan secara manual atau otomatis, namun ketika kita berbicara mengenai web scraping, umumnya kita membahas proses pengumpulan data secara otomatis menggunakan program atau bot.

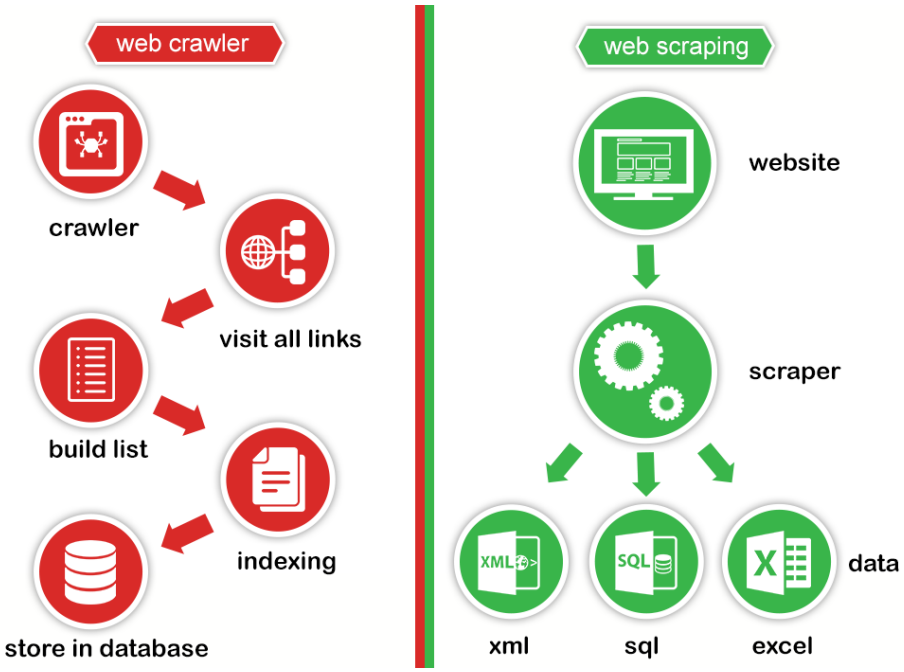
   Cara Kerja Data Scraping:

   1. Melakukan permintaan HTTP
   2. Menganalisis kode HTML
   3. Mengambil Data

   Perbedaan antara data scraping dan data crawling.

PARAMETER	DATA SCRAPING	DATA CRAWLING
Definition	Data extraction from multiple sources such as a local machine or database including web sites	Specific to downloading data from web sites and web pages
Features	- Can be performed at any scale - De duplication is not a necessity - Needs crawl agent as well as parser - Scarpers considers itself as search engine and bypasses the robots.txt file	- Mostly done at large scale - De duplication is essential - Needs only crawling agent - Crawler follows a robots.txt file
Examples	WebHarvy, Octoparse, import.io, Parsehub etc	Amazon Bot, Googlebot, Yahoo, Bingbot etc.
Use cases	Lead Generation for Marketing, Price Comparison, Academic Research etc	Machine learning training data collection, Price intelligence data collection, Fetching product data etc.
networkinterview.com (An Initiative By ipwithease.com)		

Crawling adalah proses pengambilan data dari website secara otomatis oleh mesin pencari, crawling biasanya menggunakan API untuk mengambil datanya, sementara scraping melibatkan ekstraksi data dengan memanipulasi kode HTML pada website, dan cenderung tidak menggunakan API dalam pengaplikasiannya.



2. **Pentingnya Data Scraping dalam NLP**

Scraping adalah teknik yang powerful yang dapat digunakan untuk mengumpulkan data teks dalam jumlah besar dari berbagai sumber. Data teks ini kemudian dapat digunakan untuk melatih model bahasa NLP.

II. Alat dan Pustaka untuk Data Scraping

1. **Python**
2. **Library**

   - BeautifulSoup dan requests untuk scraping data dari HTML.
   - Library scraping twitter seperti snscrape, tweet-harvest
   - GUI scrapper seperti Webharvy
   - Chrome Web Scraper
   - dll.

III. GUIDED

1. Crawling Twitter dengan tweet-harvest
   Link notebook :
   https://colab.research.google.com/drive/1lErgWDKVlieV684sIP9_gNitozirBKAK?usp=sharing
2. Scraping komentar playstore dengan googleplay scraper
   Link notebook:
   https://colab.research.google.com/drive/15Gr1RGe4N9tsxx8F3X6cS90hiIcbDfrT?usp=sharing
3. Scraping Tripadvisor menggunakan Chrome Web Scraper
4. Scraping artikel menggunakan BeautifulSoup
   Link notebook:
   https://colab.research.google.com/drive/1PZtamAJJaPDOJAnMTns37VXnSBqSM2-s?usp=sharing

IV. UNGUIDED

Pilihlah salah satu tools yang digunakan scraping/crawling pada praktikum hari ini, ambil topik tertentu dan ambil datanya (misal : twitter pake keywoard presiden ; komentar sebuah aplikasi playstore ; objek wisata tertentu tripadvisor ; kumpulan artikel tertentu dari website ; dll).

Upload kode program dan hasil dari scrapping ke github masing masing, dan kumpulkan link nya.

Referensi :

[1] https://cmlabs.co/id-id/seo-terms/data-scraping#:~:text=Apa%20Itu%20Data%20Scraping%3F,juga%20disebut%20dengan%20data%20extraction.

[2] https://dibimbing.id/en/blog/detail/apa-itu-web-scraping-definisi-manfaat-hingga-metode#:~:text=Definisi%20dan%20Fungsi%20Web%20Scraping,Apa%20itu%20web&text=Web%20scraping%20adalah%20proses%20mengumpulkan,otomatis%20menggunakan%20program%20atau%20bot.

[3] https://medium.com/@dede.brahma2/perbedaan-antara-crawling-dan-scraping-98e64e0c6439

[4] https://networkinterview.com/data-crawling-vs-data-scraping/

Penutup

Dalam modul praktikum ini, Telah dipelajari konsep dasar data scraping dan bagaimana teknik ini dapat digunakan dalam pemrosesan bahasa alami (NLP). Data scraping merupakan dasar pengambilan data untuk melatih model NLP. Selain itu, telah diperkenalkan dengan tools dan library yang digunakan untuk melakukan data scraping, seperti Python, BeautifulSoup, requests, dan berbagai alat khusus seperti tweet-harvest, Google Play Scraper, dan Chrome Web Scraper.