Power of Data Scraping

Data Engineering Week 3

18 Mei 2023

Whisnumurty Galih Ananta

Lead of Data Engineering, CODER TUS

Whisnumurty Galih Ananta

Lead of Data Engineering at 

"Passionate Data Science and Machine Learning enthusiast, dedicated to continuous learning and growth by actively contributing to challenging projects"

- **Machine Learning Cohort at Bangkit ,** Feb 2024 - Now
- **Lead of Data Engineering at CODER,** Des 2022 - Now
- **Core Team at GDSC Tel-U,** Aug 2023 - Now

Find Me:

 github.com/whisnumurtyga

 linkedin.com/in/whisnumurtyga

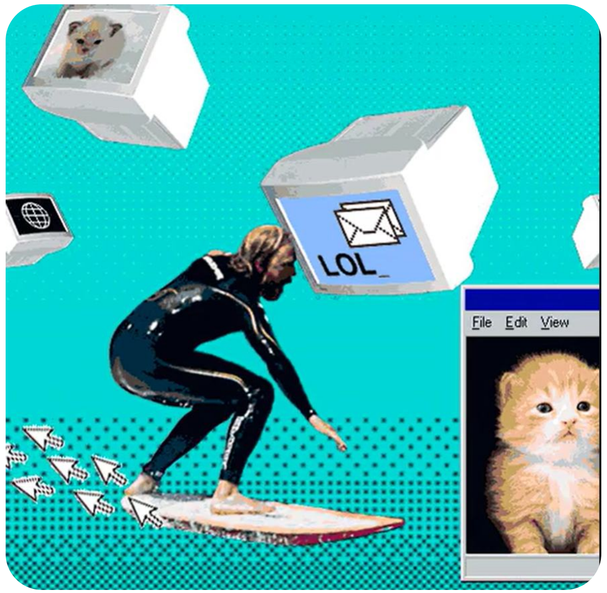
What ~~you~~ we will learn

1. What is Data Scraping
2. Type of Data Scraping
3. Scraping vs Crawling
4. How Data Scraping Works
5. Why Data Scraping is Powerful
6. Data Scraping Considerations
7. Data Scraping Mitigation
8. Hands-on

What is Data Scraping

Data scraping is the automatic extraction of information or data from various sources.

ex: Database, Web, Dokumen Spreadsheets, Text, Aplikasi



Type of Data Scraping

	Data Source	Method
 Web Scraping	specifically focuses on extracting data from websites and web pages.	automated techniques to fetch and parse the HTML code of web pages.
 Screen Scraping	process of capturing data from the visual display of computer screens.	scripts interact with the user interface of applications to capture and extract data from the screen

Scraping vs Crawling

Web scraping is about extracting the data from one or more websites.

While Crawling is about finding or discovering URLs or links on the web.

Web scraping is about extracting the data from one or more websites. While Crawling is about finding or discovering URLs or links on the web.

How Data Scraping Works

Select
Data Source

Access
the Data

Parsing
HTML

Select &
Extract Data

Data Cleaning
& Transformation

Storage
& Export

Monitoring
& Error Handling

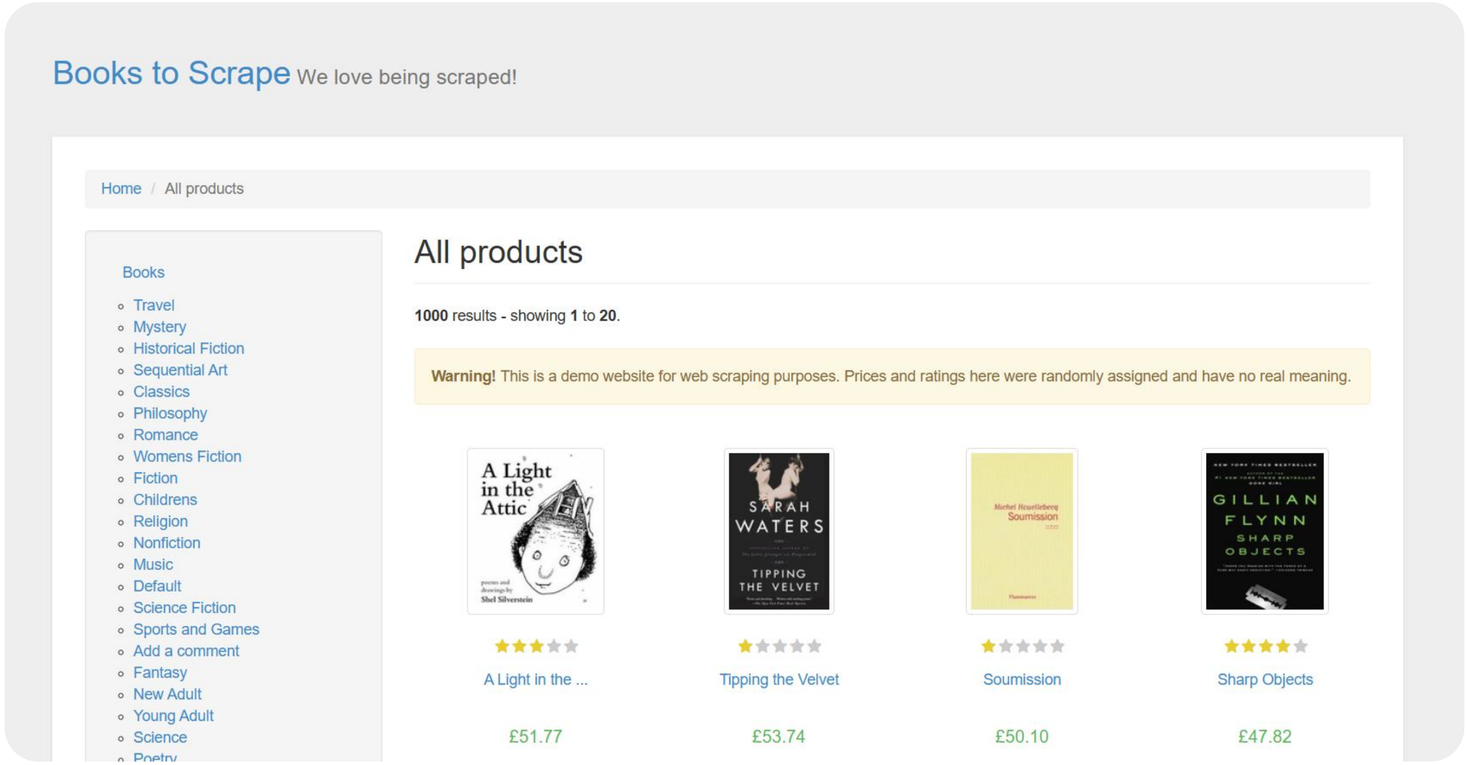
Legal
& Ethical

Consider
Scalability

Select Data Source

First, you identify the data source from which you want to extract information.

this could be a specific website, a database, a document, or any other data repository.



Access the Data

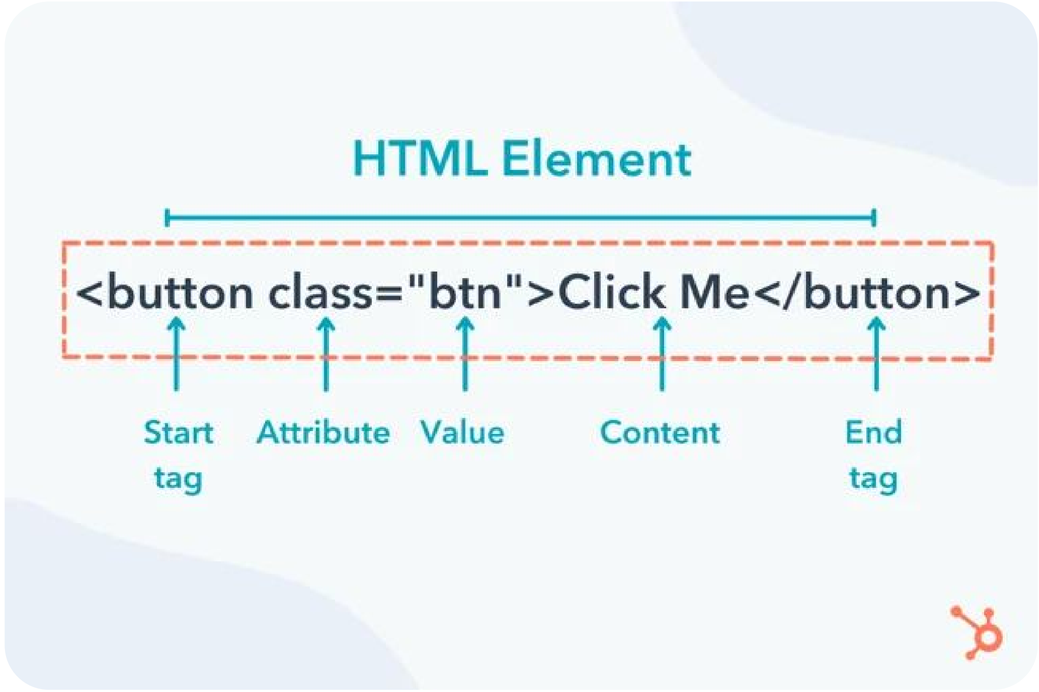
If the data source is a website, data scraping begins by sending HTTP requests to the website's server to retrieve the web page's HTML content.

You might need to establish a connection using the appropriate protocols or APIs for other sources like databases.

Parsing HTML

HTML content of the web page is fetched and then parsed. This parsing involves breaking down the HTML document into its constituent elements, like headings, paragraphs, tables, and links.

Tools like Beautiful Soup in Python or Cheerio in JavaScript are often used.



Select and Extract Data

Once the HTML is parsed, you define criteria to identify and extract the specific data you need.

this involves using selectors like CSS selectors or XPath to target the HTML elements that contain the data of interest.

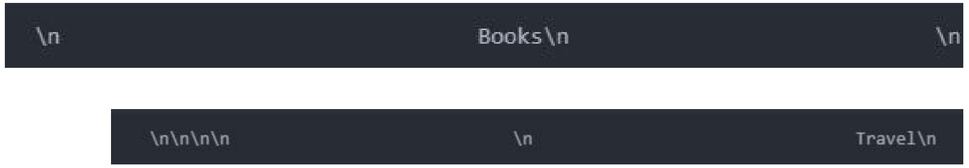
For instance, you could extract product prices, names, and descriptions from an e-commerce website.

```
1  for link in soup.find_all('a'):
2      print(link.get("href"))
```

```
1  a = []
2  Judul = soup.find_all('ul',class_ = "nav nav-list")
3  for judul in Judul:
4      a.append(judul.text.strip('\n'))
5  print(a)
```

Data Cleansing & Transformation

Raw data from scraping may contain unwanted elements, inconsistencies, or formatting issues.



Data cleaning involves removing or handling these anomalies to make the data usable, including removing HTML tags, converting data to a standardized format, and handling missing values.

```
split_text = a[0].split('\n\n\n')
cleaned_text = [part.strip() for part in split_text]
print(cleaned_text)
```

Storage & Export

After scraping and cleaning the data, you can choose to store it in a database, a spreadsheet, a text file, or any other suitable storage medium

Some scraping projects may involve real-time analysis or direct integration with other systems, so the data may not be stored locally.

Title	Rating	Price	Rating Nev	Review	Price New	Euro ke Rup
A Light in th	Three	51.77	3	bolehlah	Mahal	880090
Tipping the	One	53.74	1	gak dulu	Mahal	913580
Soumissio	One	50.1	1	gak dulu	Mahal	851700
Sharp Obje	Four	47.82	4	oke	Mahal	812940
Sapiens: A	Five	54.23	5	uapik	Mahal	921910
The Requie	One	22.65	1	gak dulu	Murah	385050
The Dirty Li	Four	33.34	4	oke	Mahal	566780
The Comin	Three	17.93	3	bolehlah	Murah	304810
The Boys ir	Four	22.6	4	oke	Murah	384200
The Black I	One	52.15	1	gak dulu	Mahal	886550
Starving He	Two	13.99	2	hmmm	Murah	237830
Shakespea	Four	20.66	4	oke	Murah	351220
Set Me Fre	Five	17.46	5	uapik	Murah	296820
Scott Pilgri	Five	52.29	5	uapik	Mahal	888930

Is Data Scraping Powerful?

Accurate, up-to-date data is a goldmine of knowledge and information for enterprises.

Depending upon how it was processed and analyzed, it can be used for a wide range of purposes.

For instance, you could extract product prices, names, and descriptions from an e-commerce website.

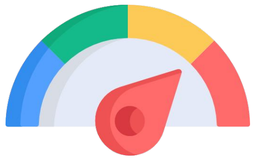
Use Case

- Brand, Product, Price Monitoring
  Using data scraping to gather up-to-the-minute data allows them to adjust and adapt strategies in real time.
- Consumer Sentiment Analysis
  The success of products and services can hinge on consumer perceptions. Scraping reviews, comments, as can gauge the pulse of the consumer.
- Lead Generation
  Automated data scraping improves online information gathering, lead generation, and lead profiling, improving marketing tactics for streamlined customer acquisition.

Data Scraping Considerations

Legal & Ethical

It’s essential to confirm whether you have the necessary rights and access to scrape data before you do so



Rate Limiting

Excessive scraping requests can overwhelm the source server.

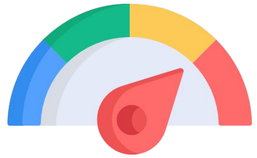
Data Privacy

Ensure any personal data is handled according to data privacy regulations.

Data Scraping Mitigation

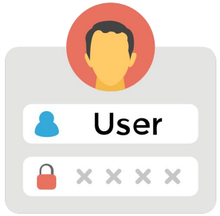
CAPTCHAs

requiring users to complete an automated test to “prove” they are a human visitor.



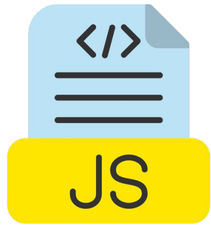
Rate Limiting

limiting certain types of network traffic to reduce strain and prevent bot activity.



Required Authentication

not allowing access to any unauthorized user or software.



Dynamic Website Content

web content that changes based on user behavior that can recognize and block scraping tools.

Refference

https://www.fortra.com/resources/guides/what-is-data-scraping-and-how-use-it
https://www.datamation.com/big-data/data-scraping/
https://blog.hubspot.com/website/html-elements
https://www.youtube.com/watch?v=TLosoD249NA

Let’s Hands On

Penugasan

A. jobstreet.co.id

Aqilah Jihan Nabila
I Made Wisnu Adi Sanjaya
Isni Khairul Fahmi

- Scrape by search keyword
- Job Title
- Company Data (Name, Field, Description, Total Employee, Review)
- Job Description (Loc, Position, Work Hour)
- Date Posted
- Qualifications
- Job Descriptions

B. quran.nu.or.id

Aisyah Nur Azizah Sajimin
Pinaringan Iman Santoso

- Surah Data (Number, Name in Latin and Arabic, Meaning of Surah, Number of Verses, Category of Surah (Makkiyah/ Madaniyah))
- Ayat Data (Number, Arabic, Latin, Indonesian Translation)

C. arxiv.org

Sun Kayla Zelikha Az Zahra
Aulia Wulandari
Daffa Farrel Giovany

- Just scrape research with PDFs that are free to access.
- Scrape (Title, Authors, Submitted Date, Announced Date, PDF Link).
- Input Keyword, Subject Field, and Search Term (Default field: Computer Science).
- Filter (Max size, Hide Abstract, Date).

Absensi

https://s.id/pods-w3

Feedback Form

https://s.id/ff-de23

Quote

”Menjadi hebat saja tidak cukup, namun jadilah hebat yang bermanfaat”

– Azzahra Putri Santi

Thank you