Chapter 1: Introduction to Web Scraping

DR. LINDA MAHMOUDI

”يرفعِ اللهُ الذين أمنوا منكم والذين أوتوا العلم درجات“

Keywords

- web scraping,
- data extracting,
- data mining,
- crawler,
- data scraping
- web content extracting,
- data harvester,

What is web scraping?

- ✓ Web scraping is a technique to fetch data from websites
- ✓ It is the automation of the data extraction process from websites
- ✓ The construction of an agent to download, parse, and organize data from the web in an automated manner.
- ✓ Web scraping (or data scraping) is a technique used to collect content and data from the internet.

What is web scraping?

- This event is done with the help of web scraping software known as web scrapers.
  - They automatically load and extract data from the websites based on user requirements.
- This data is usually saved in a local file so that it can be manipulated and analyzed as needed.
- If you've ever copied and pasted content from a website into an Excel spreadsheet, this is essentially what web scraping is, but on a very small scale

Web scraping Tools

One of the many definitions of this concept

“A web scraping tool is a technology solution to extract data from web sites, in a quick, efficient and automated manner, offering data in a more structured and easier to use format, either for B2B or for B2C processes ».

Web scraping is the process of collecting and parsing raw data from the Web, and the Python community has come up with some pretty powerful web scraping tools.

Web scraping Tools

- Scraping processes may be written in different programming languages. The most popular are

  Java, Python, Ruby or Node.

- Nonetheless, some software companies have designed different tools that enable other people to use scraping techniques by means of attractive and powerful user interfaces.

What kinds of data can you scrape from the web?

If there's data on a website, then in theory, it's scrapable!

Common data types organizations collect include images, videos, text, product information, customer sentiments and reviews (on sites like Twitter, Yell, or Tripadvisor), and pricing from comparison websites.

There are some legal rules about what types of information you can scrape, but we'll cover these later on.

Scraping

- *Using tools to gather data you can see on a webpage*
- A wide range of web scraping techniques and tools exist.
- These can be as simple as copy/paste and increase in complexity to automation tools, HTML parsing, APIs and programming

HTTP

- *HyperText Transfer Protocol*
- Machine interchange information; Transported over the Internet to enable multi-media data exchange, aka WWW.
- The protocol defines aspects of authentication, requests, status codes, persistent connections, client/server request/response. etc.
- Access a server on port 80;
- The declarative Document Type Definition (HTML, XML, JSON, etc.)

HTML

- *HyperText Markup Language*
- The standard markup language on the Web
- As the web evolves so does the proliferation of technical wrappers surrounding the visible content of websites (text and data)

Parsing

The act of analyzing the strings and symbols to reveal only the data you need

Crawling

Moving across or through a website in an attempt to gather data from more than one URL or page

JSON

Javascript Open Notation

Readable text used to transmit data

objects consisting of attribute-value pairs


```
{
"firstName": "John",
"lastName": "Smith",
"isAlive": true,
"age": 25,
"address": {
"streetAddress": "21 2nd Street",
"city": "New York",
"state": "NY",
"postalCode": "10021-3100"
},
"phoneNumbers": [
{
"type": "home",
"number": "212 555-1234"
},
{
"type": "office",
"number": "646 555-4567"
}
],
"children": [],
"spouse": null
}
```

API

- Application Programming Interface
- A set of rules and protocols used to build a software application. In the context of Web
- Scraping an API is a method used to gather clean data from a website (i.e. data that is not wrapped in HTML, Javascript, bound in HTTP, etc.)

Techniques of Web Scraping:

There are two ways of extracting data from websites, the **Manual extraction** technique, and the **automated extraction** technique.

- **Manual Extraction Techniques:** Manually copy-pasting the site content comes under this technique. Though tedious, time taking and repetitive it is an effective way to scrap data from the sites having good anti-scraping measures like bot detection.
- **Automated Extraction Techniques:** Web scraping software is used to automatically extract data from sites based on user requirement.

Parsing:

Parsing means to make something understandable to be analyzing it part by part. To wit, it means to convert the information in one form to another form that is easy to that is easier to work on with.

HTML parsing means taking in the code and extracting relevant information from it based on the user requirement. Mainly executed using JavaScript, the target as the name suggests are HTML pages.

Parsing:

DOM Parsing: The Document Object Model is the official recommendation of the World Wide Web Consortium. It defines an interface that enables a user to modify and update the style, structure, and content of the XML document.

Web Scraping Software: Nowadays, many web scraping tools are available or are custom build on users need to extract required desiring information from millions of websites.

Legalization of Web Scraping

- The legalization of web scraping is a sensitive topic, depending on how it is used it can either be a boon or a bane.
- On one hand, web scraping with good bot enables search engines to index web content, price comparison services to save customer money and value. But web scraping can be re-targeted to meet more malicious and abusive ends.

Legalization of Web Scraping

- Web scraping can be aligned with other forms of malicious automation, named *“bad bots”*, which enable other harmful activities like *denial of service attacks*, *competitive data mining*, *account hijacking*, *data theft* etc.

- Legality of Web Scraping is a grey area that tends to develop as time goes on. Although the web scrapers technically increase the speed up data surfing, loading, copying, and pasting web scraping is also the key culprit behind the increases cases of copyright violation, violated terms of use and other activities that are highly disruptive to a company’s business.

Challenges to Web Scraping

Besides the challenge of the legality of web scraping, there are also other problems that pose a challenge to web scraping.

Data Warehousing: Data extraction at a scale will generate a large amount of information to be stored. If the data warehousing infrastructure is not properly built then the searching, storing and exporting of this data will become a cumbersome task. Hence, for large-scale data extraction, there needs to be a perfect data warehousing system without any flaws and faults.

Challenges to Web Scraping

Website Structure Changes: Every website periodically updates its user interface to improve its attractiveness and experience. This requires various structural changes too. Since the web scrapers are set up according to the code elements of the website at that time, they require changes too. So, they require changes weekly too to target the correct website for data scraping as incomplete information regarding the website structure will lead to improper scraping of data.

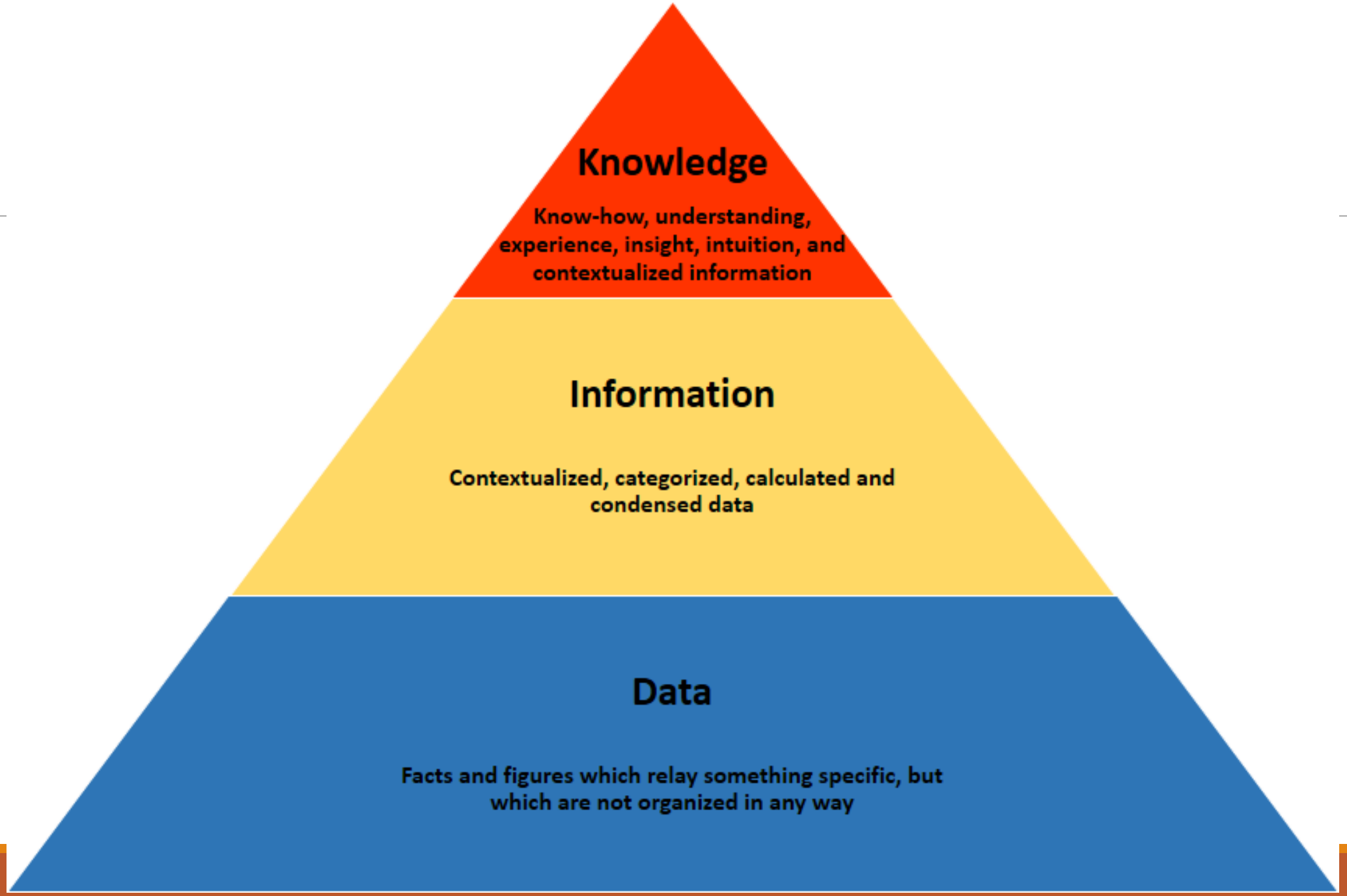
Challenges to Web Scraping

Anti-Scraping Technologies: Some websites use anti-scraping technologies that thwart away any scraping attempt. They apply a dynamic coding algorithm to prevent any bot intervention and use the IP blocking mechanism. It requires a lot of time and money to work around such anti-scraping technologies.

Quality of Data Extracted: Records that do not meet the quality of information required will affect the overall integrity of the data. Making sure that the Data Scraped meets the quality guidelines is a difficult task as it needs to be done in real-time.

Future of Data Scraping

- As there are some challenges and opportunities for data scraping, it can be fairly deemed that the unintended data-scraping practitioners are prone to create a moral hazard where they target the companies and retrieve their data.
- However, since we are on the verge of data transformation, data-scraping in combination with big data can provide the company's market intelligence and help them identify critical trends and patterns and identify the best opportunities and solutions.
- Hence, it won't be wrong to say that Data scraping can be upgraded to the better soon.


Knowledge
Know-how, understanding, experience, insight, intuition, and contextualized information
Information
Contextualized, categorized, calculated and condensed data
Data
Facts and figures which relay something specific, but which are not organized in any way

Data Scraping for data scientists

The web exposes for data scientists alot of interesting opportunities:

- There might be an interesting table on a Wikipedia page (or pages) you want to retrieve to perform some statistical analysis.
- Perhaps you want to get a list of reviews from a movie site to perform text mining, create a recommendation engine, or build a predictive model to spot fake reviews.
- You might wish to get a listing of properties on a real-estate site to build an appealing geo-visualization.

Data Scraping for Data Scientists

- You'd like to gather additional features to enrich your data set based on information found on the web, say, weather information to forecast, for example, soft drink sales.
- You might be wondering about doing social network analytics using profile data found on a web forum.
- It might be interesting to monitor a news site for trending new stories on a particular topic of interest.

What is Web Scraping used for?

1. Price Monitoring

Web Scraping can be used by companies to scrap the product data for their products and competing products as well to see how it impacts their pricing strategies. Companies can use this data to fix the optimal pricing for their products so that they can obtain maximum revenue.

2. Market Research

Web scraping can be used for market research by companies. High-quality web scraped data obtained in large volumes can be very helpful for companies in analyzing consumer trends and understanding which direction the company should move in the future.

3. News Monitoring

Web scraping news sites can provide detailed reports on the current news to a company. This is even more essential for companies that are frequently in the news or that depend on daily news for their day-to-day functioning. After all, news reports can make or break a company in a single day!

4. Sentiment Analysis

If companies want to understand the general sentiment for their products among their consumers, then Sentiment Analysis is a must. Companies can use web scraping to collect data from social media websites such as Facebook and Twitter as to what the general sentiment about their products is. This will help them in creating products that people desire and moving ahead of their competition.

5. Email Marketing

Companies can also use Web scraping for email marketing. They can collect Email ID’s from various sites using web scraping and then send bulk promotional and marketing Emails to all the people owning these Email ID’s.

The following list outlines some interesting real-life use cases:

- Many of **Google's products** have benefited from Google's core business of crawling the web. Google Translate, for instance, utilizes text stored on the web to train and improve itself.

- Scraping is being applied a lot in **HR and employee analytics**. The San Francisco-based hiQ startup specializes in selling employee analyses by collecting and examining public profile information, for instance, from LinkedIn

- **Banks and other financial institutions** are using web scraping for competitor analysis. For example, banks frequently scrape competitors' sites to get an idea of where branches are being opened or closed, or to track loan rates offered
- **Investment firms** also often use web scraping, for instance, to keep track of news articles regarding assets in their portfolio
- **Sociopolitical scientists** are scraping social websites to track population sentiment and political orientation.

In summary

- **Web scraping can be used to collect all sorts of data types:** From images to videos, text, numerical data, and more.
- **Web scraping has multiple uses:** From contact scraping and trawling social media for brand mentions to carrying out SEO audits, the possibilities are endless.
- **Planning is important:** Taking time to plan what you want to scrape beforehand will save you effort in the long run when it comes to cleaning your data.

In summary

- **Python is a popular tool for scraping the web:** Python libraries like Beautifulsoup, scrapy, and pandas are all common tools for scraping the web.
- **Don’t break the law:** Before scraping the web, check the laws in various jurisdictions, and be mindful not to breach a site’s terms of service.
- **Etiquette is important, too:** Consider factors such as a site’s resources—don’t overload them, or you’ll risk bringing them down. It’s nice to be nice!
- **Data scraping** is just one of the steps involved in the broader **data analytics process**.