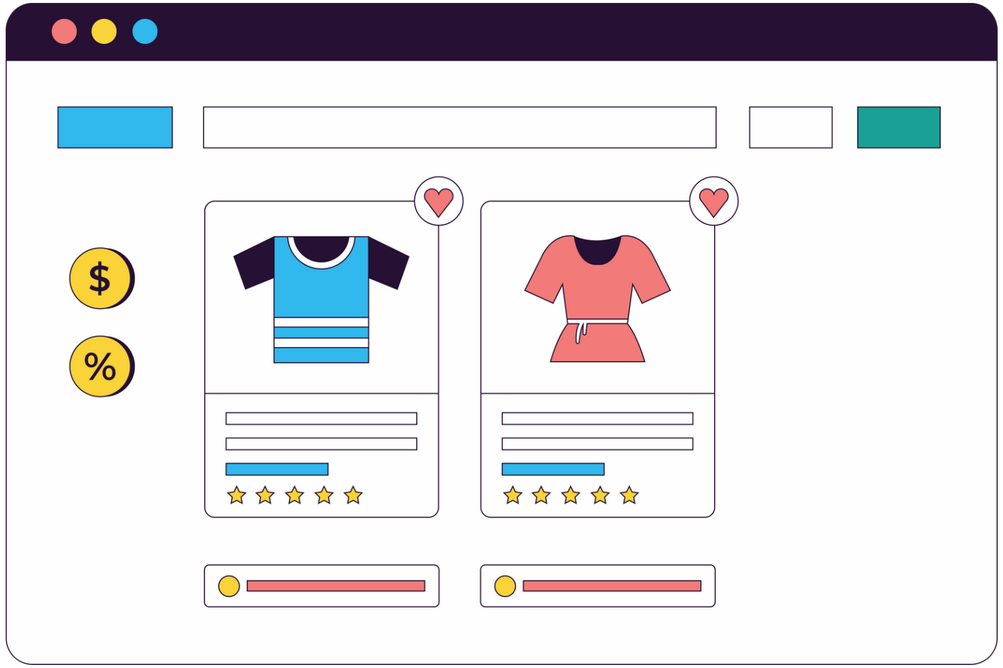
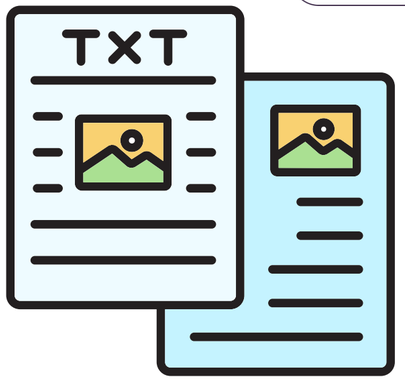
Introduction to Web Scraping & Text Extraction





Presentation by : Gurtaran Singh

What is

Web Scraping

and

Text Extraction

Web Scraping

- Web scraping is the **automated process of extracting data from websites**.
- With web scraping, **large amounts of data can be collected from multiple web pages** efficiently.
- Web scraping eliminates the need for manual data extraction, which can be time-consuming and labor-intensive. By automating the process, **web scraping saves considerable time and effort,** enabling quicker access to valuable data.

Legality and Ethics of Web Scraping

- 3rd Party APPs scam (mods)
  - Access to read emails and messages

- Data Breech NEWS
  - eg - in 2021 April

  530 million (53 Cr) people Facebook Data got Leaked

1. https://www.safetydetectives.com/blog/facebook-scraped-leak-report
2. https://www.bbc.com/news/technology-56772772
3. https://www.cnet.com/news/privacy/facebook-says-data-leak-is-from-old-vulnerability-that-no-longer-exists/
4. https://tech.hindustantimes.com/tech/news/truecaller-data-of-4-75-crore-indian-users-leaked-on-dark-web-report-71590562580077.html

Legality and Ethics of Web Scraping

For education purpose or college projects or for learning need we may use web scraping directly.

But for any commercial use or scraping private data we need to follow some steps.

Legality and Ethics of Web Scraping

- Respect **website terms of service** and honor any stated scraping restrictions.
- Obtain **proper consent** or ensure that the **data being scraped is publicly available**.
- Adhere to **copyright laws** and avoid scraping copyrighted material without permission.
- Respect privacy regulations and **avoid scraping personally identifiable information** without consent.
- **Attribute the source of scraped data when used publicly to maintain transparency.**
    - Open - https://en.wikipedia.org/wiki/India

Popular Web Scraping Tools and Libraries

- **BeautifulSoup:** A Python library for parsing HTML and XML documents, making it easy to navigate and extract data from web pages.
  - https://www.crummy.com/software/BeautifulSoup/bs4/doc/
  - https://colab.research.google.com/drive/1F8YhzNtJy3R74ha-DCj4YEieh8a7kgRX#scrollTo=izxwpB-dYTHb

- **Requests:** A versatile Python library for sending HTTP requests, used in combination with BeautifulSoup or other parsing libraries for web scraping.

You can also explore these tools and libraries:

- **Puppeteer:** A Node.js library that provides a high-level API for controlling headless Chrome or Chromium browsers, suitable for scraping JavaScript-rendered websites.
- **Selenium:** A powerful framework that automates browser interactions, enabling scraping of dynamic websites and handling JavaScript elements.
- **Scrapy:** A Python framework specifically designed for building web crawlers and scraping large-scale websites efficiently.
- **Octoparse:** A visual web scraping tool that allows non-programmers to extract data from websites through a user-friendly interface.
- **ParseHub:** A web scraping service that offers a point-and-click interface for creating scraping projects, suitable for both beginners and advanced users.
- **Scrapy Cloud:** A cloud-based platform for running and managing Scrapy spiders, providing scalability and ease of deployment for large-scale scraping tasks.

Web Scraping Workflow

POST PROCESSING

Filer and sort the useful information out of raw data

Extract the raw data

Get a website link

Determine the website from which you want to extract data. Identify the specific web pages or sections containing the desired information.

Check legality before scraping any data

Use any of above listed tools or libraries

Let's Code

OPEN GITHUB

https://github.com/gurtaransingh/scraping

Text Scraping from Images and PDFs

PDF

NEWS

THE MOST IMPORTANT NEWS FROM AROUND THE WORLD

WEDNESDAY, JULY 23, 2013 SINCE 1875 #204

Good news for travel industry

World Government allowed to cross all national borders without visa.

Starting today and forever. There is no need to apply for national or regional visa. All people of the world are free to visit any country and stay there for unlimited time. Travel agencies are predincting huge rise of their sales volumes.

Air carriers and aircraft manufacturers are even more optimistic about the future. CEO of the largest company said that "it is definitely a historical day".

Well, it's really a new level of freedom.

more on page 3...

New power saving technology invented.

British scientists unveil the 'world's first' laptop powered by light. Mobile computer owners are looking forward for first batch of the laptops.

Researchers claim that this kind of batteries can be extremely useful in space, where solar energy is much more powerful.

more on page 4...

Top stories

USA presidental election predictions.

Most experts think that it would be tight race for John Doe and John Smith. *more on page 2...*

"We are in the Golden Age of stock market"

says minister of economics at annual meeting.

more on page 7...

UFO defined!

Two Russian farmers caught an extraterrestial humanoid with fishing net.

more on page 12...

Sports

Soccer World Championship standings.

1. Germany
2. Brazil
3. France
4. Netherlands

more on page 24...

Three underappreciated swimmers.

Sometimes it's not so easy to predict winners in swimming. We found three 'dark horses' at started World Cup. *more on page 25...*

Text Scraping from Images and PDFs

Optical Character Recognition (OCR)

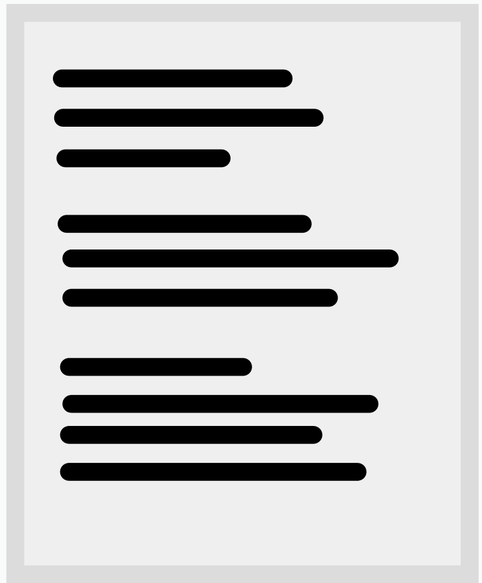
OCR technology converts scanned images or PDFs into machine-readable text.

Image Preprocessing

Images may require preprocessing techniques like noise reduction or contrast enhancement for better OCR results.

PDF Parsing

PDF parsing libraries extract text from PDF files by interpreting their structure and content.



Best Practices and Challenges in Data Extraction from PDFs and Images:

- **Quality of Source Material:** Ensure that the PDFs and images used for data extraction are of sufficient quality and clarity to yield accurate results.
- **Preprocessing Techniques:** Implement image preprocessing techniques like noise reduction, resizing, or contrast adjustment to enhance OCR accuracy.
- **OCR Selection:** Choose the appropriate OCR engine, such as Tesseract or Google Cloud Vision, based on the specific requirements and language support.
- **Language and Font Compatibility:** Verify that the OCR engine supports the language and font used in the PDFs or images to avoid character recognition errors.

Hello Hello HELLO HELLO HELLO Hello Hello HELLO Hello Hello Hello hello

Let's Code

OPEN GITHUB

https://github.com/gurtaransingh/scraping

Thank You