web scraping

Type	Note
Status	Completed

objs

- explain process and use of web scraping
- identify diff types of web scrapers

web scraping

- a way to fetch data from websites
  - many websites disallow users to save data for personal use
  - webscraping helps automate the data extraction process

basics

crawler

- our spider, an AI that browses the internet to index and search for content
- you usually first crawl the web or a specific site to discover the URLs which you then pass on to your scraper

scraper

- tool to accurately and quickly extract data from a webpage
- important parts of a scraper
  - data locaters (selectors) - finds the data that you want to extract from the HTML file
    - XPath
    - CSS selectors

- regex
- ...or a combination

differences

Web Scraping	Web Crawling
The tool used is **Web Scraper**.	The tool used **Web Crawler** or **Spiders**.
It is used for **downloading** information	It is used for **indexing** of Web pages
It **need not visit all the pages** of website for information.	It **visits each and every page**, until the last line for information.
It is done on both **small and large scale**.	It is mostly employed in **large scale**.
Application areas include Retail Marketing, Equity search, and Machine learning.	Used in search engines to give search results to the user.
Data de-duplication is not necessarily a part of Web Scraping.	Data de-duplication is and integral part of Web Scraping.
This **needs crawl agent and a parser** for parsing the response.	This only **needs only crawl agent**.
ProWebScraper, Web Scraper.io, ParseHub etc are the examples	Google, Yahoo or Bing do Web Crawling

real uses

price intelligence

- biggest use case for web scraping
- takes product info and price
- useful in
  - dynamic pricing
  - revenue optimisation
  - competitor monitoring
  - product trend monitoring

market research

- for business intelligence
  - market trend analysis
  - R&D
  - monitoring competitors

- etc

more uses

- ml
- financial data analysis
- social media analysis
- SEO monitoring

how to scrape

Step 1	find the URLs you want to scrape
Step 2	inspect page
Step 3	identify data to be extracted
Step 4	write code
Step 5	run code
Step 6	store data

guidelines

1. respect robots.txt
2. dont hit the server too frequently, set timeouts
3. spoof and rotate around different user agents
4. disguise requests by rotating IPs and Proxy
5. have different crawling patterns
6. scrape on off-peak
7. use the scraped data responsibly
8. canonical URLs
    a. duplicate URLs will have a canonical URL, which points to the parent or the original URL

theory

_____ is the biggest use case for web scraping. It involves the process of extracting product and pricing information from e-commerce websites, then turning it into intelligence is an important part of modern e-commerce companies that want to make better pricing/marketing decisions based on data.

- Market Research
- Price Intelligence
- Financial Data Analysis
- Social Media Analysis