Data (Web) Scraping

Kaustubh R. Kulkarni

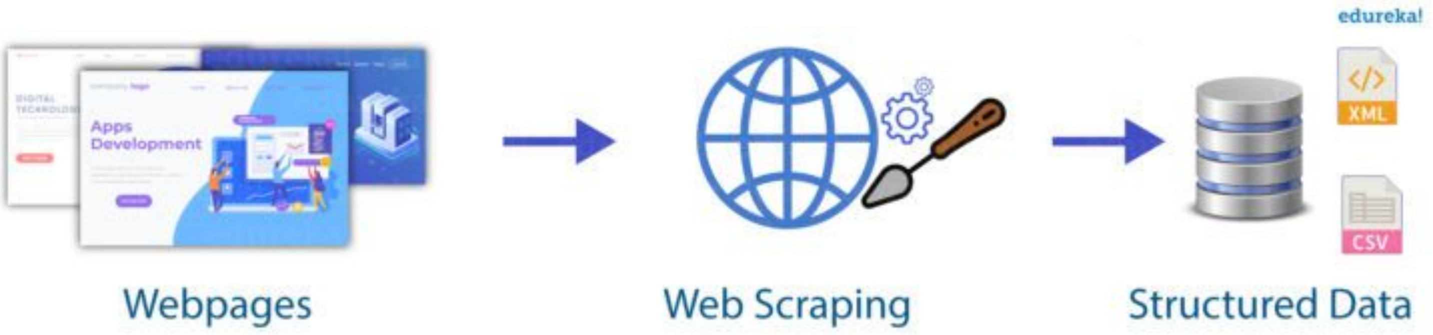
Assistant Professor
Department of Computer Engineering
K. J. Somaiya College of Engineering
Somaiya Vidyavihar University

Web Scraping

- Data scraping, also known as web scraping, is the process of importing information from a website into a spreadsheet or local file saved on your computer.

What is Web Scraping?

- Web scraping is an automated method used to extract large amounts of data from websites.
- The data on the websites are unstructured.
- Web scraping helps collect these unstructured data and store it in a structured form.
- There are different ways to scrape websites such as online Services, APIs or writing your own code.


edureka!
Apps Development
XML
CSV
Webpages
Web Scraping
Structured Data

Why is Web Scraping Used?

- **Price Comparison:** Services such as ParseHub use web scraping to collect data from online shopping websites and use it to compare the prices of products.
- **Email address gathering:** Many companies that use email as a medium for marketing, use web scraping to collect email ID and then send bulk emails.
- **Social Media Scraping:** Web scraping is used to collect data from Social Media websites such as Twitter to find out what's trending.
- **Research and Development:** Web scraping is used to collect a large set of data (Statistics, General Information, Temperature, etc.) from websites, which are analyzed and used to carry out Surveys or for R&D.
- **Job listings:** Details regarding job openings, interviews are collected from different websites and then listed in one place so that it is easily accessible to the user.

Uses of data scraping include:

- Research for web content/business intelligence
- Pricing for travel booker sites/price comparison sites
- Finding sales leads/conducting market research by crawling public data sources (e.g. Yell and Twitter)
- Sending product data from an e-commerce site to another online vendor (e.g. Google Shopping)

list's just scratching the surface.

Is Web Scraping Legal?

- The most prevalent misuse of data scraping is email harvesting – the scraping of data from websites, social media and directories to uncover people’s email addresses, which are then sold on to spammers or scammers.
- In some jurisdictions, using automated means like data scraping to harvest email addresses with commercial intent is illegal, and it is almost universally considered bad marketing practice.

Web Crawling v/s Web Scraping

- Often used interchangeably
- Web crawling is basically used to index the information on the page using bots aka crawlers. It is also called **indexing**.
- web scraping is an automated way of extracting the information using bots aka scrapers. It is also called **data extraction**.

Web Crawling	Web Scraping
Refers to downloading and storing the contents of a large number of websites.	Refers to extracting individual data elements from the website by using a site-specific structure.
Mostly done on large scale.	Can be implemented at any scale.
Yields generic information.	Yields specific information.
Used by major search engines like Google, Bing, Yahoo. **Googlebot** is an example of a web crawler.	The information extracted using web scraping can be used to replicate in some other website or can be used to perform data analysis. For example the data elements can be names, address, price etc

Components of a Web Scraper

- A web scraper consists of the following components −
- Web Crawler Module
- A very necessary component of web scraper, web crawler module, is used to navigate the target website by making HTTP or HTTPS request to the URLs. The crawler downloads the unstructured data (HTML contents) and passes it to extractor, the next module.
- Extractor
- The extractor processes the fetched HTML content and extracts the data into semi structured format. This is also called as a parser module and uses different parsing techniques like Regular expression, HTML Parsing, DOM parsing or Artificial Intelligence for its functioning.

- Data Transformation and Cleaning Module
  - The data extracted above is not suitable for ready use. It must pass through some cleaning module so that we can use it. The methods like String manipulation or regular expression can be used for this purpose. Note that extraction and transformation can be performed in a single step also.
- Storage Module
  - After extracting the data, we need to store it as per our requirement. The storage module will output the data in a standard format that can be stored in a database or JSON or CSV format.

Read Yourself Slide

Many web users have adopted techniques to help reduce the risk of email harvesters getting hold of their email address, including:

- Address munging: changing the format of your email address when posting it publicly, e.g. typing ‘patrick[at]gmail.com’ instead of ‘patrick@gmail.com’. This is an easy but slightly unreliable approach to protecting your email address on social media – some harvesters will search for various munged combinations as well as emails in a normal format, so it’s not entirely airtight.
- Contact forms: using a contact form instead of posting your email address(es) on your website.
- Images: if your email address is presented in image form on your website, it will be beyond the technological reach of most people involved in email harvesting.

The Data Scraping Future

- There are now data scraping **AI on the market that can use machine learning** to keep on getting better at recognising inputs which only humans have traditionally been able to interpret – like images.

Questions

?