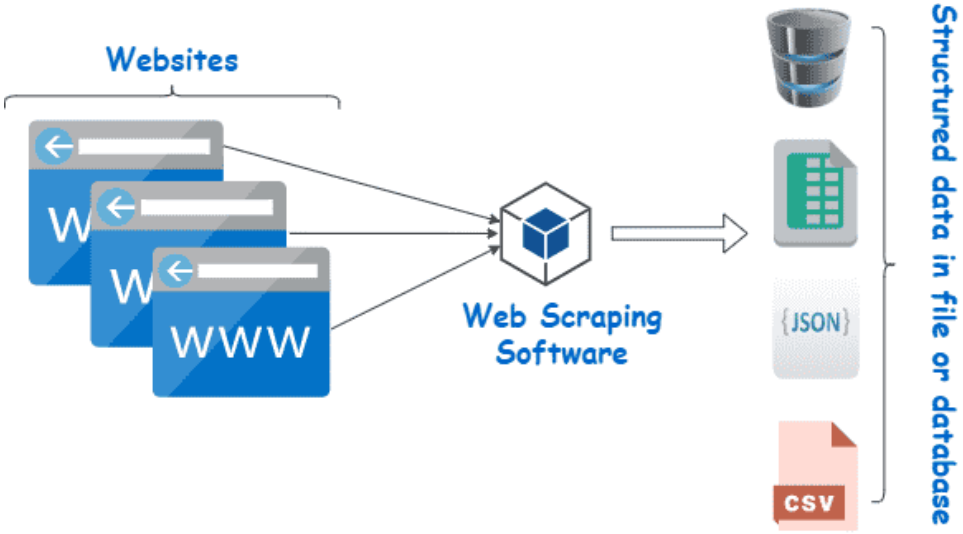
Web Scraping

- an automatic method to obtain large amounts of data from websites
    - data is unstructured data in HTML format
    - the data is then converted into structured data in a spreadsheet/database to be used in different applications
- web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser
- requires 2 parts: **crawler** & **scraper**

  Crawler
    - an AI algorithm that browses the web to search for the particular data required by following the links across the internet

  Scraper
    - a specific tool created to extract data from the website
    - procedure
        1. get necessary URLs
        2. load all the HTML code for those sites
            a. may extract all the CSS and JS elements
        3. obtains the required data from the HTML code
        4. outputs the data in the format specified by the user
            a. data is saved as an Excel, CSV, or JSON file



web scraping

Types of Web Scrapers

1. Self-built Web Scrapers
2. Browser Extensions Web Scrapers
3. Cloud Web Scrapers

Python Web Scraping Tools

Scrapy

- open-source web crawling framework in Python
- ideal for web scraping & extracting data using APIs

Beautiful Soup

- python library suitable for Web Scraping
- creates a parse tree that can be used to extract data from HTML on a website
- has multiple features for navigation, searching, and modifying the parse trees

Requests

- python library that interacts with APIs and web services to scrape websites and perform other HTTP-based tasks
- intuitive and easy to send HTTP requests
- supports a variety of HTTP methods
    - GET, PUT, DELETE, HEAD, OPTIONS, PATCH

lxml

- python's parsing library that works with both HTML and XML files
- ideals when extracting data from large datasets
- impacted by poorly designed HTML

Selemuim

- open-source browser automation tool that allows to automate processes
    - logging into a social media platform
- useful for websites that are written using javascript
- used for execution of test cases or test scripts on web applications
- able to initiate rendering web pages by running Javascript
- requires 3 components
    1. web browser
    2. driver for the browser
    3. selenium package