Source: [What is Web Scraping and How to Use It? - GeeksforGeeks]

- Web scraping: automatic method to get a large amount of data from websites (usually in HTML format, which is then structured)
    - Many platforms have APIs that allow you to access their data
- 2 parts: crawler and scraper
    - Crawler: algorithm that browses the web for the particular data required (follwing links across the Internet)
    - Scraper: specific tool to extract data from website; can extract all the data on particular sites or the specific data that a user wants
- Scraping process
    - URLs are provided
    - Loads all the HTML code for those sites (might also do CSS and JS)
    - Obtains the required data from this HTML code
    - Outputs this data in the format specified by the user (usually CSV file or JSON)
- Types of web scrapers
    - Browser extensions Web Scrapers
        - Extensions that can be added to your browser
        - Easy to use, but can be limited by browser
    - Software Web Scrapers
        - Can be downloaded and installed on your computer
        - Have advanced features that are not limited by the scope of your browser
    - Cloud Web Scrapers
        - Run on the cloud (off-site server)
        - Allow your computer to focus on other tasks
    - Local Web Scrapers
        - Run on your computer using local resources
        - If it requires more CPU or RAM, then your computer will become slow
- Python is the most popular language for web scraping because it can handle most of the processes easily
- Python has a variety of libraries that were created specifically for Web Scraping

    - Scrapy → very popular open-source web crawling framework written in Python; ideal for web scraping as well as extracting data using APIs
    - Beautiful soup → highly suitable for Web Scraping; creates a parse tree that can be used to extract data from HTML on a website

Source: [Top 7 Data Scraping Tools in 2023 | Octoparse]

- Octoparse
    - Powerful web scraper with comprehensive features
    - Special auto-detection feature that auto-targets data for you
    - Built-in web scraping templates including Amazon, Yelp, and many popular website templates
- Zyte
    - Cloud-based web platform
    - Different types of tools — Scrapy Cloud, Smart Browser API, Automatic Extraction, and Splash
    - Provides different web services for different kinds of people, including the open-source framework Scrapy

Source: [API Portal (wikimedia.org)]

- Open access to multilingual information from Wikipedia and other Wikimedia projects
    - There are various APIs that provide access to content and data from Wikimedia projects
- APIs
    - Core REST API
        - Discover and interact with free knowledge from across Wikimedia projects.
        - https://api.wikimedia.org/wiki/Core_REST_API
    - Feed API
        - Discover and interact with free knowledge from across Wikimedia projects.
        - https://api.wikimedia.org/wiki/Feed_API

- Lift Wing API
  - Discover and interact with free knowledge from across Wikimedia projects.
  - https://api.wikimedia.org/wiki/Lift_Wing_API
- Page Description API
  - Discover and interact with free knowledge from across Wikimedia projects.
  - https://api.wikimedia.org/wiki/Page_Description_API
- Page Description API
  - Suggest links to add to an article on Wikipedia.
  - https://api.wikimedia.org/wiki/Link_Recommendation_API