Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.  
Scraping a web page involves fetching it and then extracting data from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Having fetched, extraction can take place. The content of a page may be parsed, searched and reformatted, and its data copied into a spreadsheet or loaded into a database. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be finding and copying names and telephone numbers, companies and their URLs, or e-mail addresses to a list (contact scraping).  
As well as contact scraping, web scraping is used as a component of applications used for web indexing, web mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), gathering real estate listings, weather data monitoring, website change detection, research, tracking online presence and reputation, web mashup, and web data integration.  
Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. As a result, specialized tools and software have been developed to facilitate the scraping of web pages. Web scraping applications include market research, price comparison, content monitoring, and more. Businesses rely on web scraping services to efficiently gather and utilize this data.  
Newer forms of web scraping involve monitoring data feeds from web servers. For example, JSON is commonly used as a transport mechanism between the client and the web server.  
There are methods that some websites use to prevent web scraping, such as detecting and disallowing bots from crawling (viewing) their pages. In response, web scraping systems use techniques involving DOM parsing, computer vision and natural language processing to simulate human browsing to enable gathering web page content for offline parsing.  
History  
After the birth of the World Wide Web in 1989, the first web robot, World Wide Web Wanderer, was created in June 1993, which was intended only to measure the size of the web.  
In December 1993, the first crawler-based web search engine, JumpStation, was launched. As there were fewer websites available on the web, search engines at that time used to rely on human administrators to collect and format links. In comparison, Jump Station was the first WWW search engine to rely on a web robot.

In 2000, the first Web API and API crawler were created. An API (Application Programming Interface) is an interface that makes it much easier to develop a program by providing the building blocks. In 2000, Salesforce and eBay launched their own API, with which programmers could access and download some of the data available to the public. Since then, many websites offer web APIs for people to access their public database.  
Techniques  
Web scraping is the process of automatically mining data or collecting information from the World Wide Web. It is a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding, artificial intelligence and human-computer interactions.  
Human copy-and-paste  
The simplest form of web scraping is manually copying and pasting data from a web page into a text file or spreadsheet. Sometimes even the best web-scraping technology cannot replace a human's manual examination and copy-and-paste, and sometimes this may be the only workable solution when the websites for scraping explicitly set up barriers to prevent machine automation.  
Text pattern matching  
A simple yet powerful approach to extract information from web pages can be based on the UNIX grep command or regular expression-matching facilities of programming languages (for instance Perl or Python).  
HTTP programming  
Static and dynamic web pages can be retrieved by posting HTTP requests to the remote web server using socket programming.  
HTML parsing  
Many websites have large collections of pages generated dynamically from an underlying structured source like a database. Data of the same category are typically encoded into similar pages by a common script or template. In data mining, a program that detects such templates in a particular information source, extracts its content, and translates it into a relational form, is called a wrapper. Wrapper generation algorithms assume that input pages of a wrapper induction system conform to a common template and that they can be easily identified in terms of a URL common scheme. Moreover, some semi-structured data query languages, such as XQuery and the HTQL, can be used to parse HTML pages and to retrieve and transform page content.  
DOM parsing  
By using a program such as Selenium or Playwright, developers can control a web browser such as Chrome or Firefox wherein they can load, navigate, and retrieve data from websites. This method can be especially useful for scraping data from dynamic sites since a web browser will fully load each page. Once an entire page is loaded, you can access and parse the DOM using an expression language such as XPath.  
Vertical aggregation  
There are several companies that have developed vertical specific harvesting platforms. These platforms create and monitor a multitude of "bots" for specific verticals with no "man in the loop" (no direct human involvement), and no work related to a specific target site.