Web scraping

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 extracting 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. Once 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.

The history of web scraping dates back nearly to the time when the World Wide Web was born.

After the birth of the World Wide Web in 1989, the first web robot,[2] 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, JumpStation 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.

The algorithm to download videos:

```
import requests
chunk_size = 256
url=
"https://files3.lynda.com/secure/courses/786416/VBR_MP4h264_main_SD/786
416_00_01_WL30_Batteries.mp47_JlGPFuqEmh2WYnzy8KqP63YccR5B9TyiaLDUyJqMD
8vI0FKsEoIijfxnOyVetNrjDj4fDCUidEarb6NlvSpp8gq0tp4MNS2E_14bhH4B3Nj8ekt9
I0KwkiE6T4k39UnzitRKsZgt55bwl0MhP3aWp1j2U6IsuqTBssp_Gug0_ZtYFcWvzsH5BN5
EsA_A&c3.ri=3775645051486129780"
r = requests.get(url, stream=True)
with open("lynda.mp4","wd") as f:
    for chunk in r.iter_content(chunk_size=chunk_size):
        f.write(chunk)
```

relate to a website called Lynda.com which contains a lot of tutorials.

Source attribute(particular file that video is stored)

In python:

1-We will import requests library.

2-We will download video in chunks (small parts).

3-The size of the chunk will be 256 byte .. at a time to be downloaded we will be making multiple requests (r) instead of a single request.

4-write the URL of video.

5- Equal multiple requests(r) to requests.get , URL and stream

6-Equal the above to true means.. we tell the Lynda .com server to maintain a connection with the request

7- Open the name of the file is (.mp4)in right library mode as f for chunk in r.iter content

R.iter content will return an iterator which can let us keep on asking for the different parts of our video.

Art auditor content will require our chunk size . The chunk size will asked at a time .auto return content chunk size equal to charge size .Then we get that chunk and write that chunk inside our file .

The whole file will get downloaded.