PROJECT REPORT
WEB SCRAPPER

Omar ALLAL
Hamza BOUKHATEM
Skander RADHOUANE
Mohamed Ali BEN GHARBIA

Second deliverable: Existing solutions

Contents

I. Web scraping approaches.

There are 2 different approaches for web scraping depending on how does website structures their contents.

1. Approach 1: HTML method
2. Approach 2: API method

➢ HTML method:

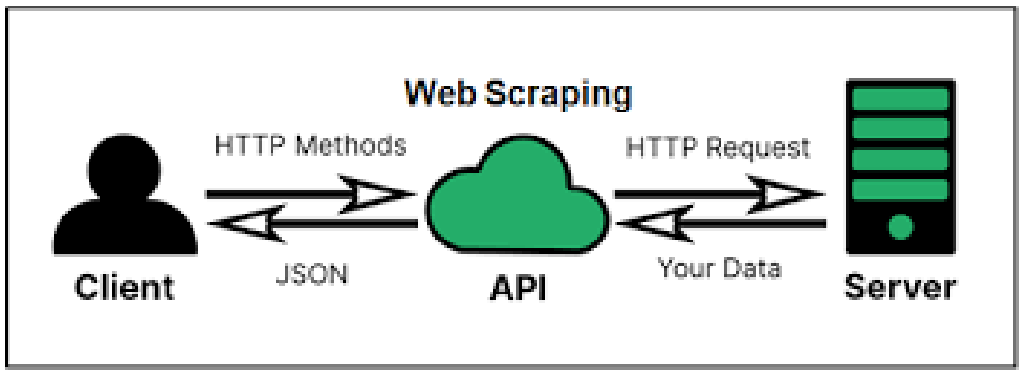
If a website stores all their information on the HTML front end, you can directly use code to download the HTML contents and extract useful information.

- Step 1: Inspect the website HTML that you want to crawl.
- Step 2: Access URL of the website using code and download all the HTML contents on the page.
- Step 3: Format the downloaded content into a readable format.
- Step 4: Extract out useful information and save it into a structured format.
- For information displayed on multiple pages of the website, you may need to repeat steps 2–4 to have the complete information.

Pros : simple and direct

Cons : you need to adjust your code according to the website's front-end structure.

- API method:



If a website stores data in API and the website queries the API each time when user visits the website, you can simulate the request and directly query data from the API.

- Step 1: Inspect the XHR network section of the URL that you want to crawl.
  XMLHttpRequest (XHR) is a JavaScript API to create AJAX requests. Its methods provide the ability to send network requests between the browser and a server.
- Step 2: Find out the request-response that gives you the data that you want.
- Step 3: Depending on the type of request (post or get) and the request header and payload, the request is simulated in the code and the data is retrieved from API. Usually, the data got from API is in a pretty neat format.
- Step 4: Extract out useful information that you need.
- For API with a limit on query size, you will need to use 'for loop' to repeatedly retrieve all the data.

Pros : more structured and stable data

Cons : more complicated especially if authentication or token is required

II. Web Scraper Desktop-based :

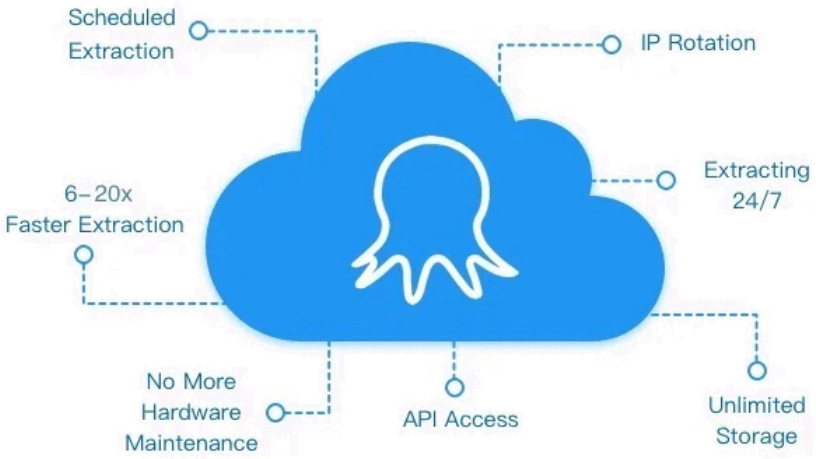
Web Scraper Desktop is a desktop application that allows users to extract data from websites using a visual scraping interface. It is a standalone application that can be downloaded and installed on a computer.

Example :“Octoparse”

Octoparse is not only a robust web scraping tool but also provides web scraping services for business owners and enterprises. Generally, the free version can meet the basic scraping needs.

Here are some main features :

- Device: Windows and macOS
- Data: It supports almost all types of websites for scraping, including social media, e-commerce, marketing, real-estate listing, etc.
- Function:
  - handle both static and dynamic websites with AJAX, JavaScript, cookies, etc.
  - extract data from a complex website that requires login and pagination.
  - deal with information that is not showing on the websites by parsing the source code.



Different modes:

- The Task Template Mode enables non-coding users to turn web pages into some structured data instantly. On average, it only takes about 6.5 seconds to pull down the data behind one page and allows you to download the data to Excel.
- The Advanced mode has more flexibility. This allows users to configure and edit the workflow with more options. Advance mode is used for scraping more complex websites with a massive amount of data.
- The brand-new Auto-detection feature allows you to build a crawler with one click. If you are not satisfied with the auto-generated data fields, you can always customize the scraping task to let it scrape the data for you.
- The cloud services enable large data extraction within a short time frame as multiple cloud servers concurrently are running for one task. Besides that, the cloud service will allow you to store and retrieve the data at any time.

III. Web Scraping Plugins and Extensions

They are software tools that can be added to web browsers to enable web scraping functionality. These plugins and extensions typically work by adding a scraping interface to the web browser, allowing users to select and extract data from web pages with a few clicks.

Example: “Webscraper.io” : https://webscraper.io

Web scraper has a Chrome extension and cloud extension. For the Chrome extension version, you can create a sitemap (plan) on how a website should be navigated and what data should be scrapped. The cloud extension can scrape a large volume of data and run multiple scraping tasks concurrently. You can export the data in CSV, or store the data into Couch DB.

Advantages:

- Ease of use: Webscraper.io has a user-friendly interface that makes it easy for users to create scraping agents without any coding knowledge.
- Customizability: Users can customize their scraping agents to extract specific data from websites, and the tool offers a wide range of options for data selection and export.
- Flexibility: Webscraper.io can be used to scrape data from any website that is accessible through a web browser, making it a versatile tool for a variety of applications.
- Timesaving: Webscraper.io can automate the process of data extraction, which can save time and effort for users who need to gather large amounts of data from multiple sources.

- Affordability: The free version of Webscraper.io offers basic scraping functionality, while the paid version is relatively affordable and offers additional features.

Limits:

- Dependence on website structure: Webscraper.io relies on the structure and formatting of websites, and any changes to the website's structure can disrupt the scraping process.
- Legality concerns: The use of web scraping tools like Webscraper.io may raise legal concerns if the data being scraped is protected by copyright or privacy laws.
- Technical limitations: Webscraper.io may not be able to extract certain types of data, such as images or multimedia files, and may encounter technical limitations in the scraping process.
- Reliance on internet connection: Webscraper.io requires an internet connection to access websites, and any interruptions or disruptions to the connection can affect the scraping process.
- Lack of support for certain websites: Webscraper.io may not be compatible with certain websites or may not be able to extract data from websites with complex or dynamic content.

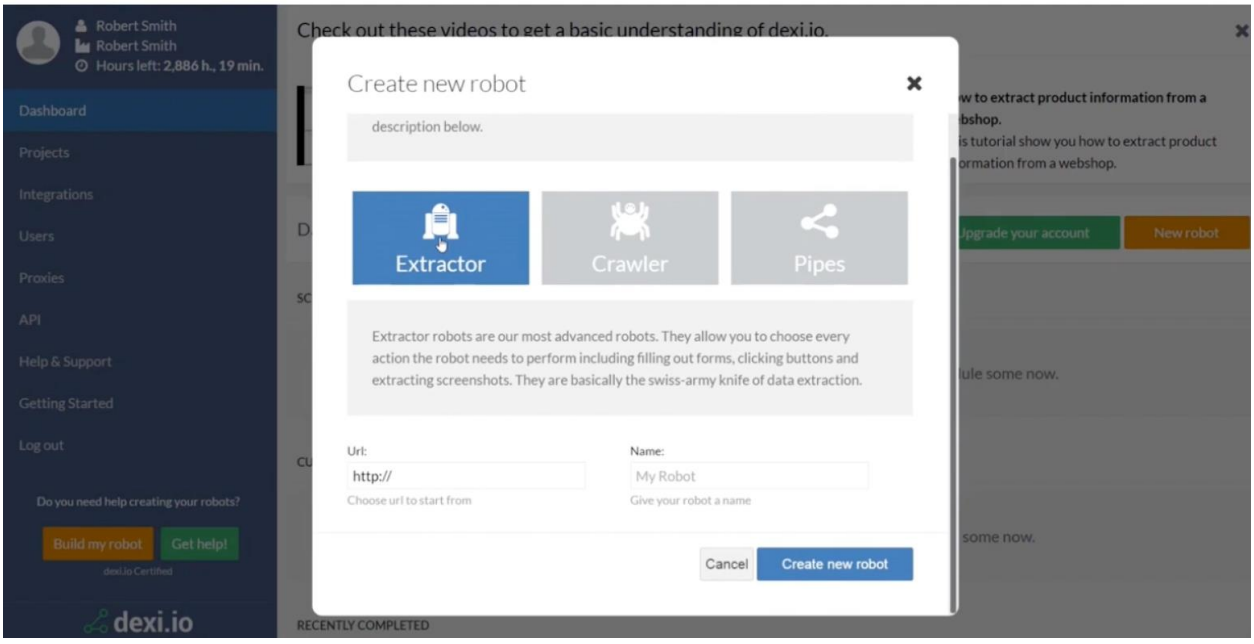
IV. Web-based Scraping Applications

They are software tools that are hosted online and allow users to extract data from websites without requiring any installation or setup on their local machines. These applications typically provide a user-friendly interface for creating scraping projects, selecting data to extract, and exporting the scraped data.

- Example 1: dexi

Dexi.io is intended for advanced users who have proficient programming skills. It has three types of robots for you to create a scraping task - Extractor, Crawler, and Pipes. It provides various tools that allow you to extract the data more precisely. With its modern feature, you will be able to address the details on any website. With no programming skills, you may need to take a while to get used to it before creating a web scraping robot. Check out their homepage to learn more about the knowledge base.

The freeware provides anonymous web proxy servers for web scraping. Extracted data will be hosted on Dexi.io's servers for two weeks before being archived, or you can directly export the extracted data to JSON or CSV files. It offers paid services to meet your needs for getting real-time data.



- Example 2: Webhose

The Webhose.io API makes data and meta-data easy to integrate, high-quality data, from hundreds of thousands of global online sources such as message boards, blogs, reviews, news, and more.

Webhose.io API, available either via query-based API or firehose, provides high coverage data with low latency, with an efficient dynamic capability to add new sources at record time.

Characteristics:

- Get standardized, machine-readable data sets in JSON and XML formats.
- Help you access a massive data feed repository without imposing any extra charges.
- Can perform granular analysis.

Advantages:

- Easy to use and consistent across data providers.

Limits:

- Has some learning curve.