Web scraping

How can we have information somewhere we need instead of just copy and paste from its source.

Coping and pasting from the original source will work for considerable smaller amounts of data. Suppose you have larger amount of data? In such situations, this will not work.

Here comes web scraping!

Instead of mining data manually web scraping allows to mine the data using automation methodologies. web scraping will extract the underlying HTML code with the data stored in it. Simply, using web scraping we can replicate the content of a whole website somewhere else.

Why we need scraping?

* Creating databases
* Data collection and Analytical purposes
* For Business Growth and Innovation
* Marketing and Research Purposes

Web scraping tools are software, programmed to extract the data. They can recognize the unique HTML site structure, Extract and transform content, store and scrape data and extract data from API.

Scraping tools available

* Puppeteer
* Cheerio
* Nightmare
* ScrapingBee

How a scraper works?

A screen shot of a computer screen

Description automatically generated

Automated Code component

A person sitting in a chair using a computer

Description automatically generated

Price Scraping Vs Content Scraping

Price Scraping – the purpose of the scraping tool to access pricing information.

Content Scraping – comprises large scale content theft from a given site.

Web Scraping Protection

HTML Fingerprint

IP Reputation

Behaviour Analysis

Progressive Challenges

// complete

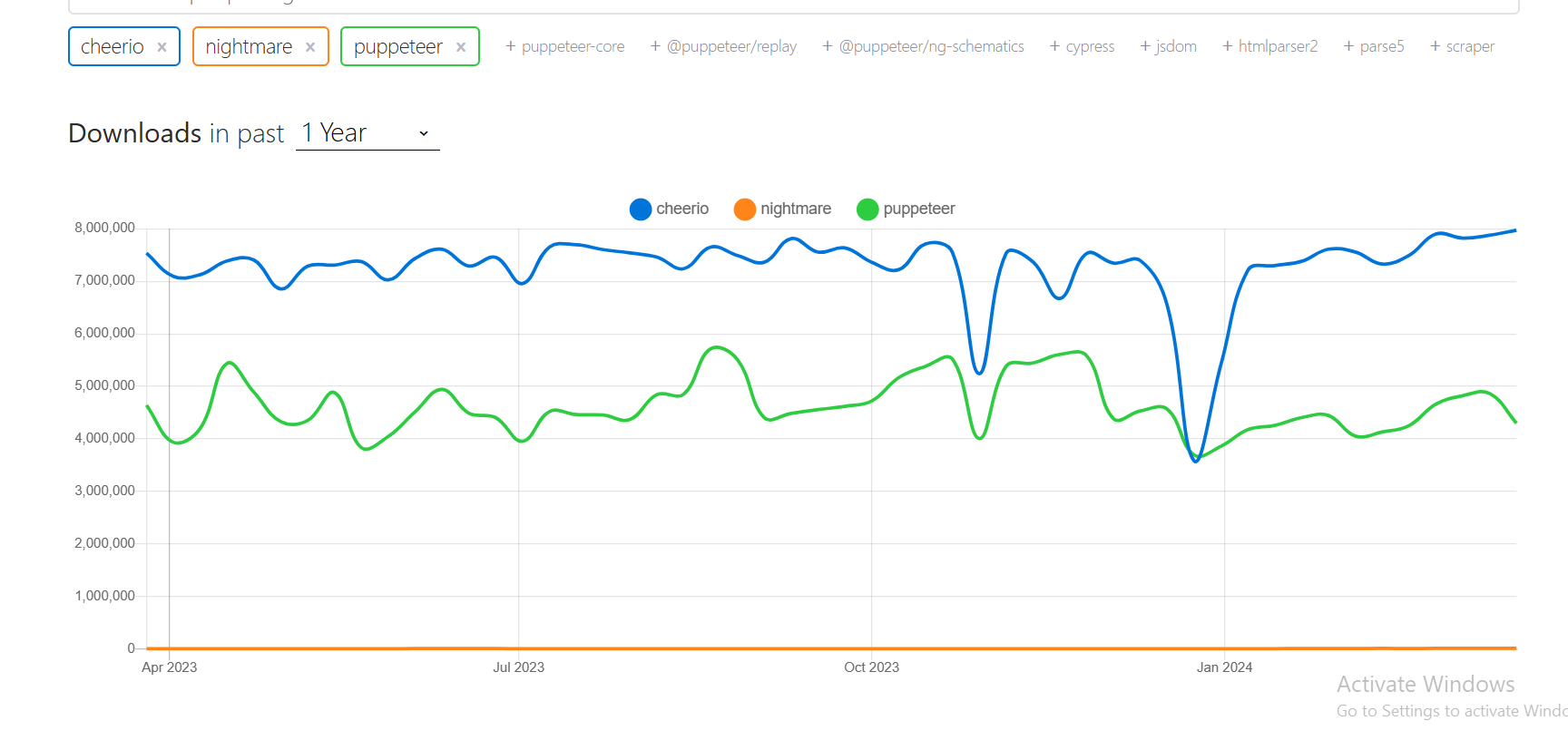
Selectors

CSS Selector

X-Path Selector

Web Scraping tool Comparison

Industry Usage



Ability of Scraping Complex sites

Cheerioo Vs Pupperteer

Cheerio – an HTML parser, enables you to take an HTML document and find the HTML elements via CSS selectors. Cheerio can’t interact with the site or access content behind scripts.