How to choose the best open source website crawler?

Here are some tips to help you find the right tool for your needs.

Scalability

The web crawler should be scalable. If your data needs are growing, the crawling tool shouldn't slow you down. Your future data requirements should be covered.

Distributed web crawling

It means all downloaded pages have to be distributed among many computers (even hundreds of computers) in fraction of seconds.

Robustness

Robustness refers to the web scraper ability to not get trapped in a large number of pages.

Politeness

Crawlers must not harm the website. A web crawler should follow the rules website's robots.txt file and should have Crawl-Delay and User-Agent header.

Extensible

Web crawlers should be extensible in many terms. They have to handle new fetch protocols, new data formats, and etc.

Data delivery formats

Ask yourself what data delivery formats you need. Of course, the best choice is to find one that delivers data in multiple formats.

Data quality

The scraped data is initially unstructured data.Choose a software capable of cleaning the unstructured data and presenting it in a readable manner.