ProBytes software

SOFTBREAKS.COM

Job Portal Web Scraping

https://www.probytes.net

Softbreaks.com – Job Portal Web Scrapping

Accessing a sea of data out there has always been challenging. If something goes wrong there is a lot at stake. For the website Softbreaks.com, we did web scrapping from multiple websites to build a huge database for building a candidate and vacancy pool. Since data is changing constantly the crawling had to be done in specific intervals and that too with automation.

Requirement

Softbreaks.com is a job portal which required a huge pool of various categories of job and candidates related data for their website. The necessity demanded the construction of a stable system that can keep on crawl web pages and scrap data over and over again. After scrapping the data, it has to be converted to XML format and has to be stored in a database for various manipulation and analytical purposes.

Challenges

- Searching and extracting millions of HTML data and converting it to XML
- Retrying failed pages
- Skipping websites and links that do not allow crawling
- Frequent structural changes in the website
- To make a system that should have more than 90% efficiency
- Real time progress check window for clients
- Scrapped data were often dirty had to clean it before feeding it to the database
- Pages with had AJAX other multiple complications in it
- Canonical Issues
- Limited time frame
- Budget

Solution

We built a web scrapping system which could crawl any number of websites at the same time and can fetch the required data with ease and impeccable ability. Speed is another forte of this system as it can crawl and fetch much faster compared to other systems in the industry.

The scrapping system we made is scalable per requirements and cost-effective at the same Features of our robust system include,

- Very easy to create a scrapping project

- Ability to scrape data from multiple threads
- Can capture data from complex structures
- Saves time
- Can auto-execute project upon commands
- Data can be exported to any format

Investment

- 2 developers
- 1 designer
- 1 project lead
- 1 communication expert
- 1 year

What we did for the project

- Our developer has manually fed the system with links that need to be scrapped
- However, no two websites will have a similar structure. So specification of various websites has to be entered manually to the system so that it can fetch data accordingly.
- Websites have to be manually checked to see whether they allow scrapping or not.
- For repetitive scrapping from a page, a program was written in the system so that unchecked or links with error will be crawled after a certain interval of time.
- The system can crawl and scrape multiple websites. For us, the system had to crawl 16 websites simultaneously and upon ranking.
- Ranking can be set in the pipeline of the system.
- Developed a parser that will break down the scrapped data.
- If there is an issue with the link or even crawling, a notification will be sent and will restart the crawling process again from where it has stopped.
- Forking was done to spawn sub-processes from a parent process.
- An interface was created to review the data collected.
- Acquired data has been added to a database.

- Developed a program that will send mail to concerned people once a program has ended.

Application

- Clean data can be obtained.
- Can be used to create a data pool for manipulation and analytical purposes.
- A job portal needs a huge amount of data to survive the competition and that too in a small amount of time. For this data, scraping is the most cost-effective solution.
- Authoritative, relevant job-related data can be obtained.
- Automated data scrapping can add value to your business on many levels.

☎Feel free to contact us

➔ ***Pune***

➔ ***Kalas road, Vishrantwadi, Pune, Maharashtra-411015,***

➔ ***+91 81 1386 1000***

➔ ***info@probytes.net***