Search

Follow us on RSS

Web Scraping

web scraping, screen scraping, data parsing and other related things

WEB SCRAPER TEST DRIVE! SOFTWARE FOR WEB SCRAPING

g+1 3

Nov 24, 2012
By Igor Savinkin
in Development
5 Comments
Tags: PHP, REGEX

Scraping in PHP with cURL

Like 0 Tweet 0 g+1 3 Share

In this post, I'll explain how to do a simple web page extraction in PHP using cURL, the 'Client URL library'.

The curl is a part of **libcurl**, a library that allows you to connect to servers with many different types of protocols. It supports the http, https and other protocols. This way of getting data from web is more stable with header/cookie /errors process rather than using simple file_get_contents(). If curl() is not installed, you can read here for Win or here for Linux.

Setting Up cURL

First, we need to initiate the cURL handle:

```
$curl = curl_init("http://testing-ground.scraping.pro/
        textlist");
```

Then, set **CURLOPT_RETURNTRANSFER** to **TRUE** to return the transfer page as a string rather than put it out directly:

```
curl_setopt($curl, CURLOPT_RETURNTRANSFER, TRUE);
```

Executing the Request & Checking for Errors

Now, start the request and perform an error check:

```
$page = curl_exec($curl);

if(curl_errno($curl)) // check for execution errors
{
    echo 'Scraper error: ' . curl_error($curl);
    exit;
}
```

Closing the Connection

To close the connection, type the following:

```
curl_close($curl);
```

Extracting Only the Needed Part and Printing It

After we have the page content, we may extract only the needed code snippet, under **id="case_textlist"**:

TAG CLOUD

ANALYTICS BIG DATA CAPTCHA CRAWLING DATA MINING GOOGLE HTTP IMPORT.IO JAVA JSON KIMONO PHP PROXY PYTHON REGEX SCRAPE-DETECTION SCRAPER SELENIUM SEO SERVICE SNIFFER STATISTICS VISUAL WEB RIPPER VISUALIZATION XPATH

Your email:

Enter email address...

SUBSCRIBE UNSUBSCRIBE

FEATURED

- OutWit Hub Review
- Helium Scraper Review
- Visual Web Ripper Review

BLOGROLL

- SQL Backup Blog
- W3 EDGE, Optimization Products for

```
$regex = '&lt;div id="case_textlist"&gt;(.*?)&lt;\/div&gt;/s';
if ( preg_match($regex, $page, $list) )
    echo $list[0];
else
    echo "Not found";
```

The Whole Scraper Listing

PHP

```php
<?php
$curl = curl_init('http://testing-ground.scraping.pro/textlist');
curl_setopt($curl, CURLOPT_RETURNTRANSFER, TRUE);

$page = curl_exec($curl);

if(curl_errno($curl)) // check for execution errors
{
    echo 'Scraper error: ' . curl_error($curl);
    exit;
}

curl_close($curl);

$regex = '/<div id="case_textlist">(.*?)<\/div>/s';
if ( preg_match($regex, $page, $list) )
    echo $list[0];
else
    print "Not found";
?>
```

This sample will guide you and give you further practice in daily web scraping.

« *How to leverage Web Scraping for SEO*

Handy Web Extractor 1.5 released »

5 Comments

P Guardiario

REPLY

NOV 28, 2012 @ 03:35

Good post. Instead of using regex though, I recommend parsing with phpquery and using css selectors.

Igor Savinkin

REPLY

NOV 28, 2012 @ 11:36

Guardiario, thank you. Yet this particular case is one having plain text (post: http://scraping.pro/test-drive-test-listing/), tricky to parse by an html dom parser or css selector. Feel free to check that post.

Michael Shilov

REPLY

NOV 28, 2012 @ 07:09

Hi! We used regex here in order to keep the example simple and do not dive into css or xpath. Thanks for the comment, though.

R Husbands

REPLY

JUN 18, 2013 @ 05:28

Great article. I found it very helpful. I am having an issue pulling information from a website and am convinced it is a problem with the regex, but not sure where the issue is. What is a good venue to seek assistance?

Thanks.

Michael Shilov

REPLY

JUN 18, 2013 @ 12:30

Hi R Husbands,

Try http://scrapetools.com/ you can quickly test regular expressions on a

WordPress

About the Author

my Linked in profile

website page there and even get the resulting PHP code.

BR,
Mike

Leave a Reply

Your Name

Your Email

Your Website

four × 5 =

POST COMMENT

Themify - Elemin theme is used in this blog.

↑