

Forum:[Hack Forums](#)

Thread: [Scrape same class ID from several URL's and show the result on a webpage?](#)

Moseby - September 16, 2018, 1:57 pm

So I would like to scrape the same class ID from say about 100 URL's, put those values together (numbers) to a grand total, and show that grand total on a website.

Is it possible? Where do I start?

DrDoof - September 16, 2018, 2:08 pm

Of course it's possible.

What's your preferred language?

Are the URL:s behind a login or can they be accessible by anyone?

Do you have a website already running or do you want your application to also host the website?

How will you store the scraped data?

Moseby - September 16, 2018, 3:09 pm

(Sep 16, 2018 02:08 PM)DrDoof Wrote: [\[CITE\]](#) Of course it's possible.

What's your preferred language?

Are the URL:s behind a login or can they be accessible by anyone?

Do you have a website already running or do you want your application to also host the website?

How will you store the scraped data?

PHP.

I found out this "simple html dom" thing but not quite sure how to handle it.

To be honest, the URL's are Facebook-URL's, so yea they can be accessed without login.

Basically I want to scrape amount of members that says on the group page, like "1340 member in this group" from like 50 groups, put the numbers together to a total and show it on my website. I already have the website running, just need something to add up the group member numbers from all the URL's and show it.

DrDoof - September 16, 2018, 3:23 pm

(Sep 16, 2018 03:09 PM)Moseby Wrote: [\[CITE\]](#) PHP.

I found out this "simple html dom" thing but not quite sure how to handle it.

To be honest, the URL's are Facebook-URL's, so yea they can be accessed without login.

Basically I want to scrape amount of members that says on the group page, like "1340 member in this group" from like 50 groups, put the numbers together to a total and show it on my website. I already have the website running, just need something to add up the group member numbers from all the URL's and show it.

If [this](#) is what you are talking about then maybe something similar to this could work?

PHP Code

```
<?php
// get group page and parse it
```

```
$html = file_get_html('https://facebook.com/groups/<group_id>');
```

```
// the element on the fb page containing the number is a span with id count_text
```

```
$text = $html->find('span[id=count_text]', 0)->innertext;
```

```
// if the number is in the 1000s it will contain commas for seperation, no good
```

```
$text = str_replace(',', '', $text);
```

```
// parse the integer
```

```
$numMembers = intval($text);
```

nb: I haven't done php in years.

Moseby - September 16, 2018, 3:25 pm

(Sep 16, 2018 03:23 PM)DrDoof Wrote: [\[CITE\]](#) If [this](#) is what you are talking about then maybe something similar to this could work?

PHP Code

```
<?php
```

```
$html = file_get_html('https://facebook.com/groups/<group_id>');
```

```
$text = $html->find('span[id=count_text]', 0)->innertext;
```

```
$text = str_replace(',', '', $text);
```

```
$numMembers = intval($text);
```

That's probably it!

Looks nice, but doesn't "file_get_html" generate the template of that group page? I only want to fetch the content of the class, not the full page.

I see you are Swedish, so am I, would you like to help me in PM or something?

DrDoof - September 16, 2018, 3:30 pm

(Sep 16, 2018 03:25 PM)Moseby Wrote: [\[CITE\]](#) That's probably it!

Looks nice, but doesn't "file_get_html" generate the template of that group page? I only want to fetch the content of the class, not the full page.

I see you are Swedish, so am I, would you like to help me in PM or something?

In their docs it says:

PHP Code

```
<?php
```

```
// Create a DOM object from a URL
```

```
$html = file_get_html('http://www.google.com/');
```

So I think it is what we want. It requests the page and the parses it into a DOM object (\$html) which we can use to get values from the requested page.

I have classes early tomorrow morning so I have to go to bed now, but if you're ever stuck in the future send me a pm and I'll get back when I can. ?

Moseby - September 16, 2018, 3:32 pm

(Sep 16, 2018 03:30 PM)DrDoof Wrote: [\[CITE\]](#) In their docs it says:

PHP Code

```
<?php
```

```
// Create a DOM object from a URL
```

```
$html = file_get_html('http://www.google.com/');
```

So I think it is what we want. It requests the page and the parses it into a DOM object (\$html) which we can use to get values from the requested page.

I have classes early tomorrow morning so I have to go to bed now, but if you're ever stuck in the future send me a pm and I'll get back when I can. ?

I see, well I guess your code should work then, but the class id would contain like "12312 Members". First I need to separate those 2, so only the numbers are saved. Then I need to add all the numbers up and eventually I need to show that number on the webpage, how do I even keep it updated for the website all the time? I want it to be all auto after the code is done.

DrDoof - September 16, 2018, 3:43 pm

(Sep 16, 2018 03:32 PM)Moseby Wrote: [\[CITE\]](#) I see, well I guess your code should work then, but the class id would contain like "12312 Members". First I need to separate those 2, so only the numbers are saved. Then I need to add all the numbers up and eventually I need to show that number on the webpage, how do I even keep it updated for the website all the time? I want it to be all auto after the code is done.

Ah, yeah, either use split to split the string into '12312' and 'Members' or replace to replace the ' Members' with nothing. You'll have to store the numbers somewhere, preferably a database like MySQL. I would keep a variable in the database with the time of the last update and then every time the script is run (every time someone visits the page) I would see if it's been longer than X time and re-scrape all the pages from Facebook if it has. Alternatively you could put a scheduled script to do the re-scrapes every few minutes, though this could be a waste of resources if you're not having consistent views on your page. (Like at least a few views per minute all 24 hours a day.) gluck and gnight

Moseby - September 16, 2018, 3:48 pm

(Sep 16, 2018 03:43 PM)DrDoof Wrote: [\[CITE\]](#) Ah, yeah, either use split to split to split the string into '12312' and 'Members' or replace to replace the ' Members' with nothing. You'll have to store the numbers somewhere, preferably a database like MySQL. I would keep a variable in the database with the time of the last update and then every time the script is run (every time someone visits the page) I would see if it's been longer than X time and re-scrape all the pages from Facebook if it has. Alternatively you could put a scheduled script to do the re-scrapes every few minutes, though this could be a waste of resources if you're not having consistent views on your page. (Like at least a few views per minute all 24 hours a day.) gluck and gnight

Sounds good. Although I seem to get an error with your code. It says "Trying to get property 'plaintext' of non-object". I read that I can't get property of stuff that's not local?

If I hit you up in PM could you help me out tomorrow sometime like Discord?

Thanks

DrDoof - September 17, 2018, 1:38 am

(Sep 16, 2018 03:48 PM)Moseby Wrote: [\[CITE\]](#) Sounds good. Although I seem to get an error with your code. It says "Trying to get property 'plaintext' of non-object". I read that I can't get property of stuff that's not local?

If I hit you up in PM could you help me out tomorrow sometime like Discord?

Thanks

Sure, hmu.

The problems seems to be that the member count is not displayed when you're not logged in. See code:

PHP Code

```
<?php
$page = file_get_contents('https://www.facebook.com/groups/409279285782672/');
echo htmlspecialchars($page);
```

The html returned by above will not contain the member count.