

Forum: [Hack Forums](#)

Thread: [Instagram bio parse - web requests](#)

mzaboss - September 19, 2016, 1:16 pm

Hello,

I need a small help.

I have a list of instagram usernames in a textfile.

i want to send multi web requests to instagram profiles and parse/filter a specific word.

Words example: Snapchat, snap, etc.

If one of those words found, then output whatever after word "Snapchat" in a textfile.

for example:

https://www.instagram.com/fbi_oo7/

check this profile bio.. you will notice " snapchat.....:sam_a1989"

I just want to save "sam_a1989"

BluePrint - September 19, 2016, 3:34 pm

I'll get you started, the code below will output into a text file called "hits.txt" any instagram urls which contain a match to a specified keyword.

My recommendations would be to use something like XPath to fine tune the keyword matches to just the bio section of the html, and perhaps Regex for removing any unwanted characters from the line where the string match exists.

Code #python 2.7

```
import requests
```

```
with open('instagram_urls.txt', 'r') as file_in:
    urls = file_in.readlines()
    for profile in urls:
        response = requests.get(profile)
        if 'snapchat' or 'other_keyword' in response.text:
            with open('hits.txt', 'a') as file_out:
                file_out.write(profile)
print("Complete")
```

I hope you find this useful ?

I made some alterations to the code this morning to now include the username in the output file as well. I've not edited the post above as you or others may find the code useful.

```
Code import requests
```

```
import re
```

```
keywords = ['snapchat','snap', 'twitt']
```

```
with open('urls.txt', 'r') as file_in:
    urls = file_in.readlines()
    for profile in urls:
        response = requests.get(profile)
        for keyword in keywords:
            if keyword in response.text:
                match = re.findall('{0}{.+}'.format(keyword), (response.text))
                string = (match[1]).encode('utf-8')
                with open('hits.txt', 'a') as file_out:
                    file_out.write('%s\t\t%s\n' % (profile.strip('\n'), string))
                break
print("Complete")
```

The output in the hits.txt file appears as:

https://www.instagram.com/fbi_oo7/ sam_a1989" />

My regex is not the greatest, so the usernames in the text file may require a little cleanup as you can see from example.

mzaboss - September 20, 2016, 4:39 am

Blueprint: Great job what you did. But how to make it read multi usernames in same time?

I am planning to check 1M lines of usernames.

BluePrint - September 20, 2016, 7:19 am

(Sep 20, 2016 04:39 AM)mzaboss Wrote: [\[CITE\]](#) Blueprint: Great job what you did. But how to make it read multi usernames in same time?

I am planning to check 1M lines of usernames.

multi-threading is possibly what you're after, I don't have much experience implementing it.

If you're considering parsing 1M user profiles, the script currently takes less a second to execute on a single profile, so you would be looking at approximately 27 hours of run time without multi-threading. I would have to do some research around multi-threading to know how much of a reduction in time you would save.

Either way you may need to implement proxy IP address rotation or a delay every so many seconds, as Instagram may ban your IP if you keep making requests to it every second. I might have some code lying about for IP rotation which I've previously used for web crawlers of mine.

Let me know your thoughts

Goodies - September 20, 2016, 8:14 am

There's just no highly reliable way of doing this. There are too many formats for putting your username in your bio. You'd have to capture the whole thing in order to be sure you got it...

mzaboss - September 20, 2016, 8:43 am

(Sep 20, 2016 07:19 AM)BluePrint Wrote: [\[CITE\]](#) multi-threading is possibly what you're after, I don't have much experience implementing it.

If you're considering parsing 1M user profiles, the script currently takes less a second to execute on a single profile, so you would be looking at approximately 27 hours of run time without multi-threading. I would have to do some research around multi-threading to know how much of a reduction in time you would save.

Either way you may need to implement proxy IP address rotation or a delay every so many seconds, as Instagram may ban your IP if you keep making requests to it every second. I might have some code lying about for IP rotation which I've previously used for web crawlers of mine.

Let me know your thoughts

Instagram doesn't block http requests, so its ok to send million requests. wont be effected. so no need for proxy support.

(Sep 20, 2016 08:14 AM)Goodies Wrote: [\[CITE\]](#) There's just no highly reliable way of doing this. There are too many formats for putting your username in your bio. You'd have to capture the whole thing in order to be sure you got it...

Yeah this is what I want.

SIM-1866423 - September 20, 2016, 11:57 am

(Sep 20, 2016 08:14 AM)Goodies Wrote: [\[CITE\]](#) There's just no highly reliable way of doing this. There are too many formats for putting your username in your bio. You'd have to capture the whole thing in order to be sure you got it...

Goodie check my solution(dynamic regex is possible with python fuu !!) ;D

PHP Code

```
<?php
content = soup.find("meta", {"property":"og:description"})["content"]
data = re.findall(("sn"+re.findall(r"sn(.+?):", content.lower())[0]+r"\S+"), content.lower())
if not data:
    data = re.findall(("sn"+re.findall(r"sn(.+?):", content.lower())[0]+r"\S+", content.lower())
print data[0].split(":")[1].strip()
```

PHP Code

```
<?php
// ?. ?????? ?????? ?????? ?????? ?????? ?? ?????? ????? ?? " snap?: mshall.11
Out : mshall.11
//?• Jizan Snap:mohs33n sayat.me/mohs33n
Out: mohs33n
```

??

mzaboss - September 20, 2016, 12:38 pm

This regex might help as well:

```
Code def findSnap(bio):
    regex = 'sn[a]*p[a-z ]*[^a-z0-9].*'
    snap = re.findall(regex, bio)
    if not snap:
        return None
    else:
        afterSnap = re.sub('sn[a]*p[a-z ]*[^a-z0-9]*', "", snap[0])
```

```
if afterSnap:
    afterSnap = re.findall('[\w_\.-]*',afterSnap)[0]
    sftS = afterSnap.split()
    if sftS:
        return sftS[0]
    return None
return None
```

But now I just need to make multi http requests and parse the responses with above useful regex.

Goodies - September 20, 2016, 1:07 pm

@b3mb4m

That is interesting, but what I'm talking about is that not everyone puts it the same way. For example:

```
"My snap is username1"
"Snap = username2"
"Add me on snappppp!!!!!!! username3"
```

Etc...

There are basically an infinite amount of variants.

Croc! - September 23, 2016, 6:39 am

(Sep 20, 2016 08:43 AM)mzaboss Wrote: [\[CITE\]](#) Instagram doesn't block http requests, so its ok to send million requests. wont be effected. so no need for proxy support.

How do you know Instagram doesn't block your IP from sending too many requests? I'm sure any site would do that if you scraped enough content from their site without going through their API.

SIM-1866423 - September 23, 2016, 6:57 am

(Sep 23, 2016 06:39 AM)Croc! Wrote: [\[CITE\]](#) How do you know Instagram doesn't block your IP from sending too many requests? I'm sure any site would do that if you scraped enough content from their site without going through their API.

Because of unconcern of course.I'm just start downloading usernames currently working just fine(1-2k request per 30 second)