

Forum:[Hack Forums](#)

Thread: [What kind of web-scraper should I create?](#)

Crypt - October 19, 2020, 3:24 pm

I'm trying to come up with a web-scraping program I want to make but I just have no idea what information I should search for.

SecuroServ - October 19, 2020, 3:26 pm

You could do a fun project of scraping e.g. medium articles and their tags to train an AI, to then let an AI automatically tag documents.

Or anything else to train AI

Crypt - October 19, 2020, 3:37 pm

(Oct 19, 2020 03:26 PM)Lesley Wrote: [\[CITE\]](#) You could do a fun project of scraping e.g. medium articles and their tags to train an AI, to then let an AI automatically tag documents.

Or anything else to train AI

So store the tags on articles and then fitting certain characteristics let the AI decide what should be tagged in articles created?

Harley Quinn - October 20, 2020, 12:45 am

So the first web crawler i made did the following:

- Stored all links on website (with the same domain)
- Captured all mobile numbers, email addresses and social media links.
- Automatically downloaded all media, images, videos, etc.

Then displaying all this info nicely, then i built it so you can scrape any site using the same script.

It was a really fun project refining it. (The AI training sounds fun too)

Snowwy - October 20, 2020, 1:13 am

(Oct 20, 2020 12:45 AM)Harley Quinn Wrote: [\[CITE\]](#) - Automatically downloaded all media, images, videos, etc.Can I ask - do you happen to still have that script around somewhere by any chance?

Harley Quinn - October 20, 2020, 1:20 am

(Oct 20, 2020 01:13 AM)Snowwy Wrote: [\[CITE\]](#) Can I ask - do you happen to still have that script around somewhere by any chance?

I will have a look, it was on my old laptop. Is that something you are interested in, something to peel all media from an entire website?

G.P.D. - October 20, 2020, 4:32 am

(Oct 20, 2020 01:13 AM)Snowwy Wrote: [\[CITE\]](#) Can I ask - do you happen to still have that script around somewhere by any chance?

I could do this for you mate.

Send me a PM with the website you want to scape.

Snowwy - October 20, 2020, 4:36 am

(Oct 20, 2020 01:20 AM)Harley Quinn Wrote: [\[CITE\]](#) I will have a look, it was on my old laptop. Is that something you are interested in, something to peel all media from an entire website?Yeh more for

just research and knowledge tbh.

(Oct 20, 2020 04:32 AM)Whait Wrote: [\[CITE\]](#) I could do this for you mate.

Send me a PM with the website you want to scape.I don't want a particular website scaped - just want a script for it so I can play around with it if you have one?

Harley Quinn - October 20, 2020, 5:11 am

(Oct 20, 2020 04:36 AM)Snowwy Wrote: [\[CITE\]](#) Yeh more for just research and knowledge tbh.

I don't want a particular website scaped - just want a script for it so I can play around with it if you have one?

Sure, let me take a look and see if i can find it, i'll send the source code over. If i can't i'll try through something together and send that :)

Prawnster - October 20, 2020, 5:16 am

I made something in the past that basically just scanned like big social media sites for information like emails, phone numbers and sensitive information and then looked it up in a DB and if they had a leaked password, log it and later email them with some free email service.

Was a nice little one to teach more about the language I was choosing to use at the time :)

G.P.D. - October 20, 2020, 6:30 am

(Oct 20, 2020 04:36 AM)Snowwy Wrote: [\[CITE\]](#) Yeh more for just research and knowledge tbh.

I don't want a particular website scaped - just want a script for it so I can play around with it if you have one?

I don't have one for now but I can take my evening to create one.

I'll do it tonight or tomorrow.

Crypt - October 20, 2020, 12:20 pm

(Oct 20, 2020 05:16 AM)Mazl Wrote: [\[CITE\]](#) I made something in the past that basically just scanned like big social media sites for information like emails, phone numbers and sensitive information and then looked it up in a DB and if they had a leaked password, log it and later email them with some free email service.

Was a nice little one to teach more about the language I was choosing to use at the time :)

I really like this idea. I know how to open a specific web page and then look for emails but is there a better way to mass scan like tweets and look for emails and phone number?

This is a cool project.

Prawnster - October 20, 2020, 12:47 pm

(Oct 20, 2020 12:20 PM)Crypt Wrote: [\[CITE\]](#) I really like this idea. I know how to open a specific web page and then look for emails but is there a better way to mass scan like tweets and look for emails and phone number?

This is a cool project.

I just used google dorks as well as just massive amounts of checks (using public proxies) to facebook pages and shit to scrape all their users.

I don't remember the exact google dorks as it was such a long time ago, but it's an interesting one to learn as you get to learn working with site APIs, proper threading, proxy support, database usage if you wanna incorporate one of those and build up some form of database for each user, and also efficient scraping.

Goodluck mate!

Crypt - October 20, 2020, 12:55 pm

(Oct 20, 2020 12:47 PM)Mazl Wrote: [\[CITE\]](#) I just used google dorks as well as just massive amounts of checks (using public proxies) to facebook pages and shit to scrape all their users.

I don't remember the exact google dorks as it was such a long time ago, but it's an interesting one to learn as you get to learn working with site APIs, proper threading, proxy support, database usage if you wanna incorporate one of those and build up some form of database for each user, and also efficient scraping.

Goodluck mate!

Thanks man! This one interests me and I think I'd learn a lot from it. I'm going to work on this.

Prawnster - October 20, 2020, 1:15 pm

(Oct 20, 2020 12:55 PM)Crypt Wrote: [\[CITE\]](#) Thanks man! This one interests me and I think I'd learn a lot from it. I'm going to work on this.

Ayy glad to hear it mate, goodluck with it and keep us posted on how it goes :)

Snowwy - October 20, 2020, 2:46 pm

(Oct 20, 2020 05:11 AM)Harley Quinn Wrote: [\[CITE\]](#) Sure, let me take a look and see if i can find it, i'll send the source code over. If i can't i'll try through something together and send that :)Oh awesome! Thank you so much!

(Oct 20, 2020 06:30 AM)Whait Wrote: [\[CITE\]](#) I don't have one for now but I can take my evening to create one.

I'll do it tonight or tomorrow.Thanks man! Much appreciated!

Harley Quinn - October 20, 2020, 2:54 pm

(Oct 20, 2020 12:20 PM)Crypt Wrote: [\[CITE\]](#) I really like this idea. I know how to open a specific web page and then look for emails but is there a better way to mass scan like tweets and look for emails and phone number?

This is a cool project.

I have actually just done this with Facebook, got over 20,000 emails so far and will likely reach 100,000 by the end of the week!

Crypt - October 20, 2020, 2:57 pm

(Oct 20, 2020 02:54 PM)Harley Quinn Wrote: [\[CITE\]](#) I have actually just done this with Facebook, got over 20,000 emails so far and will likely reach 100,000 by the end of the week!

Curious how you do this. I'm still a noob to Python but don't you have to give the URL of each page you want to search and then have it search for emails on here? There's no way to mass search random pages is there?

Harley Quinn - October 22, 2020, 12:34 pm

(Oct 20, 2020 02:57 PM)Crypt Wrote: [\[CITE\]](#) Curious how you do this. I'm still a noob to Python but don't you have to give the URL of each page you want to search and then have it search for emails on here? There's no way to mass search random pages is there?

I'll tell you exactly how i did it if you want to Discord me? Its surprisingly easy!

Crypt - October 22, 2020, 12:37 pm

(Oct 22, 2020 12:34 PM)Harley Quinn Wrote: [\[CITE\]](#) I'll tell you exactly how i did it if you want to Discord me? Its surprisingly easy!

Yeah, send me your Discord!

Lindros - October 22, 2020, 3:58 pm

I made a twitter picture scraper. It was pretty simple.

G.P.D. - *October 24, 2020, 5:55 am*

(Oct 22, 2020 12:34 PM)Harley Quinn Wrote: [\[CITE\]](#) I'll tell you exactly how i did it if you want to Discord me? Its surprisingly easy!
I'm also intrested if you don't mind sharing it :)

Harley Quinn - *October 24, 2020, 6:42 am*

(Oct 24, 2020 05:55 AM)Whait Wrote: [\[CITE\]](#) I'm also intrested if you don't mind sharing it :)
Discord me buddy.