Multithreading in Scrapy using proxies

Asked 8 years, 7 months ago Modified 8 years, 7 months ago Viewed 2k times

1

I want to crawl about 4 million pages using scrapy. I am using storm proxies. Lets say Number of threads allowed on my account is 20. I want to ask -

Is multithreading means CONCURRENT_REQUESTS_PER_DOMAIN , in scrapy.

or there is an any other way to do that.

How can I effectively use those 20 threads

NOTE - In case I am not clear with my question , please leave a comment, and I will try to elaborate according to that.

python multithreading web-scraping scrapy

Share Improve this question Follow

edited Jun 20, 2020 at 9:12
Community Bot
1 1

asked Aug 13, 2016 at 12:06
sagar
745 3 14 33

Hi sagar, Can you share your scrapy storm proxy middleware, please? I am interested in doing something similar. – rolele Dec 30, 2016 at 23:16

@sagar yes can you please share? – kabus Jul 13, 2018 at 0:00

1 Answer

Sorted by: Highest score (default)

1

Straight out of the docs:

> `CONCURRENT_REQUESTS` - The maximum number of concurrent (ie. simultaneous) requests that will be performed by the Scrapy downloader.
>
> `CONCURRENT_REQUESTS_PER_DOMAIN` - The maximum number of concurrent (ie. simultaneous) requests that will be performed to any single domain.
>
> `CONCURRENT_REQUESTS_PER_IP` - The maximum number of concurrent (ie. simultaneous) requests that will be performed to any single IP. If non-zero, the CONCURRENT_REQUESTS_PER_DOMAIN setting is ignored, and this one is used instead. In other words, concurrency limits will be applied per IP, not per domain.

Answering your question directly

I suspect that that service only let's you scrape up to 20 threads overall, meaning it doesn't care what you are requesting so you should use `CONCURRENT_REQUESTS` set to 20 maximum (default is 16).

Each request is "kind of a thread". It's built on top of Twisted. In the eyes of the proxy service you are using, there's no way to tell the difference so every request will be a proxy thread!

Share Improve this answer Follow

answered Aug 14, 2016 at 15:11
Rafael Almeida
5,250 2 22 33

Thanks! Can you tell, how could I make 2 million request in a day, some best practices may be, I am currently able to make 40000 request using proxies – sagar Aug 14, 2016 at 16:47

Do you have any way to monitor how many threads are open at any given point in time? In the proxy system – Rafael Almeida Aug 14, 2016 at 17:00

not really, I had used a middleware and set request.meta[proxy] = proxy(that they have provided me) – sagar Aug 14, 2016 at 17:34

Please I am really stuck in increasing the efficiency of crawler, Is increasing CONCURRENT_REQUSTS is the only way to do that,, ? – sagar Aug 16, 2016 at 15:21

You can use multiple servers and deploy the spider on them, feed the crawlers with the urls from a database so they don't overlap links (scrapy redis does this well), I see no other way besides increasing the infrastructure of the machine you're running the crawler in (like more cores, more ram, better network). – Rafael Almeida Aug 16, 2016 at 16:21