Proxy IP for Scrapy framework

Asked 11 years, 5 months ago Modified 6 years, 11 months ago Viewed 14k times

14

I am developing a web crawling project using **Python** and **Scrapy** framework. It crawls approax **10k web pages** from e-commerce shopping websites. whole project is working fine but before moving the code from testing server into production server i want choose a better proxy ip provider service, so that i dont have to worry about my **IP Blocking** or Denied access of websites to my spiders .

Until now i am using middleware in Scrapy to manually rotate ip from free proxy ip list available of various websites like this

Now i am confused about the options i should chosse

1. Buy **premium proxy list** from http://www.ninjasproxy.com/ or http://hidemyass.com/
2. Use **TOR**
3. Use **VPN Service** like http://www.hotspotshield.com/
4. Any Option better than above three

python proxy scrapy tor

Share Improve this question Follow

asked Oct 18, 2013 at 9:46

Binit Singh
983 4 15 35

Checkout this github.com/nabinkhadka/scrapy-rotating-free-proxies – Nabin May 24, 2020 at 11:39

3 Answers

Sorted by: Highest score (default)

10

Here are the options I'm currently using (depending on my needs):

- proxymesh.com - reasonable prices for smaller projects. Never had any issues with the service as it works out of the box with scrapy (I'm not affiliated with them)
- a self-build script that starts several EC2 micro instances on Amazon. I then SSH into the machines and create a SOCKS proxy connection, those connections are then piped through delegated to create normal http proxies which are usable with scrapy. The http proxies can either be loadbalanced with something like haproxy or you build yourself a custom middleware that rotates proxies

The latter solution is what currently works best for me and pushes around 20-30GB per day of traffic without any problems.

Share Improve this answer Follow

answered Oct 19, 2013 at 9:32

herrherr
698 1 9 26

does Amazon allow changing public IPs often? Didn't find any info on that... I'd like to spin up 20 instances and rotate their public IPs often (probably every minute) using APIs – Spaceman Jun 5, 2014 at 16:53

2 @herrherr could you share more on how to implement your second option. any guides for us to lookup on. much appreciated. thanks :) – Ming Jan 13, 2017 at 9:16

7

Crawlera is built specifically for web crawling projects. For example, it implements smart algorithms to avoid getting banned and it is used to crawl very large and high profile websites.

Disclaimer: I work for the mother company Scrapinghub, who also are core developers of Scrapy.

Share Improve this answer Follow

edited Jun 5, 2015 at 18:37

answered Oct 19, 2013 at 1:07

R. Max
6,710 1 28 35

2 It's just too expensive for a single developer. Their plans start at $99/month. – demisx Apr 19, 2020 at 18:57

0

If you don't want to use a paid service please consider just using a scrapy library that will automate rotating proxies for you: https://github.com/TeamHG-Memex/scrapy-rotating-proxies

You can have a look for a full tutorial on how to automate it here: https://tinyendian.com/articles/how-to-scrape-the-web-and-not-get-caught

Keep in mind, that when connecting through a proxy always imposes a performance penalty, but **10K** web pages that you mentioned is still well within your reach.

Share Improve this answer Follow

answered Apr 24, 2018 at 8:35

Karol Majta
246 3 7

1 If you don't want to always go and check for available free proxies, you can use this library github.com/nabinkhadka/scrapy-rotating-free-proxies. While running a spider, this library will automatically fetch fresh and newly available proxies. – Nabin May 2, 2020 at 5:31

Start asking to get answers

Find the answer to your question by asking.

Ask question

Explore related questions

python **proxy** **scrapy** **tor**

See similar questions with these tags.