How can I implement custom proxy on Scrapy?

Asked 4 years, 5 months ago Modified 4 years, 5 months ago Viewed 167 times



I'm trying to implement custom scraper API but as a begging I think I'm doing wrong. But I follow their documentation to setup everything. Here is a documentation link





```
from scrapy import Spider
from scrapy.http import Request
from .config import API
from scraper_api import ScraperAPIClient
client = ScraperAPIClient(API)
class GlassSpider(Spider):
   name = 'glass'
   allowed_domains = ['glassdoor.co.uk']
    start urls =
[client.scrapyGet(url='https://www.glassdoor.co.uk/Job/russian-jobs-
SRCH_KE0,7.htm?fromAge=1')]
    def parse(self, response):
        jobs = response.xpath('//*[contains(@class, "react-job-listing")]')
        for job in jobs:
                                                                                    **
            job_url = job.xpath('.//*[contains(@class, "jobInfoItem
jobTitle")]/@href').extract_first()
            absulate_job_url = response.urljoin(job_url)
            yield Request(client.scrapyGet(url=absulate_job_url),
                           callback=self.parse_jobpage,
                           meta={
                               "Job URL": absulate job url
                        })
    def parse_jobpage(self, response):
        absulate_job_url = response.meta.get('Job URL')
        job description = "".join(line for line in response.xpath('//*)
[contains(@class, "desc")]//text()').extract())
        yield {
            "Job URL": absulate_job_url,
            "Job Description": job_description
```

That's the output I'm receiving.... Please what's wrong with my code. Please fix it for me. So I can follow and get the point. Thank you.

2020-10-01 23:01:45 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://api.scraperapi.com/? url=https%3A%2F%2Fwww.glassdoor.co.uk%2FJob%2F russian-jobs-

SRCH_KE0%2C7.htm%3FfromAge%3D1&api_key=bec9dd9f2be095dfc6158a7e609&scraper_sdk=python> (referer: None) 2020-10-01 23:01:45 [scrapy.spidermiddlewares.offsite] DEBUG: Filtered offsite request to 'api.scraperapi.com': <GET https://api.scraperapi.c om/?

url=https%3A%2F%2Fapi.scraperapi.com%2Fpartner%2FjobListing.htm%3Fpos%3D101%26ao%3D1044074%26s%3D 149%26quid%3D00000174e51ccd8988e2e5420e6

7cf0d%26src%3DGD_JOB_AD%26t%3DSRFJ%26vt%3Dw%26cs%3D1_94f59ee8%26cb%3D1601571704401%26jobLi stingId%3D3696480795&api_key=bec9d9f82b0955c61 5c8a7e639scraper_sdk=python>

python web-scraping scrapy

Share Improve this question Follow

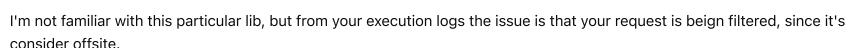


Highest score (default)

Sorted by:

1 Answer









[scrapy.spidermiddlewares.offsite] DEBUG: Filtered offsite request to 'api.scraperapi.com': <GET https://api.scraperapi.c om/?</pre> url=https%3A%2F%2Fapi.scraperapi.com%2Fpartner%2FjobListing.htm%3Fpos%3D101%26ao% $7 cf0 d\% 26 src\% 3 DGD_J0B_AD\% 26 t\% 3 DSRFJ\% 26 vt\% 3 Dw\% 26 cs\% 3 D1_94 f59 ee8\% 26 cb\% 3 D160157170440 in the contraction of the contraction$ 5c8a7e639scraper_sdk=python>



Since scraperapi will make your request go through their domain and that's outside of what you defined in your allowed_domains it's filtered as an offsite request. To avoid this issue you can remove this line entirely:

allowed_domains = ['glassdoor.co.uk']
or try include 'api.scraperapi.com' in it.
Share Improve this answer Follow

edited Oct 1, 2020 at 20:13

answered Oct 1, 2020 at 19:25

renatodvc 2,564 2 7 17

**

You are not familiar with but you reply very nicely and I tried your way and this time i got this error >>> ```raise AttributeError("Response content isn't text") AttributeError: Response content isn't text```` – booleantrue Oct 1, 2020 at 21:01 /

Start asking to get answers

Find the answer to your question by asking.

Ask question

 Explore related questions

 python
 web-scraping
 scrapy

 See similar questions with these tags.



