How to run Scrapy Crawler Process parallel in separate processes? (Multiprocessing)

Asked 5 years, 6 months ago Modified 5 years, 6 months ago Viewed 4k times

2

I am trying to do `Multiprocessing` of my `spider`. I know `CrawlerProcess` runs the spider in a single process.

I want to run multiple times the same spider with different arguments.

I tried this but doesn't work.

How do I do multiprocessing?

Please do help. Thanks.

```
from scrapy.utils.project import get_project_settings
import multiprocessing
from scrapy.crawler import CrawlerProcess

process = CrawlerProcess(settings=get_project_settings())
process.crawl(Spider, data=all_batches[0])

process1 = CrawlerProcess(settings=get_project_settings())
process1.crawl(Spider, data=all_batches[1])

p1 = multiprocessing.Process(target=process.start())
p2 = multiprocessing.Process(target=process1.start())

p1.start()
p2.start()
```

python scrapy

Share Improve this question Follow

edited Aug 24, 2019 at 12:55

asked Aug 24, 2019 at 12:52
Juggernaut
806 1 11 16

1 Answer

Sorted by: Highest score (default)

1

You need to run each `scrapy` crawler instance inside a separate process. This is because `scrapy` uses twisted, and you can't use it multiple times in the same process.

Also, you need to disable the telenet extension, because `scrapy` will try to bind to the same port on multiple processes.

Test code:

```
import scrapy
from multiprocessing import Process
from scrapy.crawler import CrawlerProcess

class TestSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://blog.scrapinghub.com']

    def parse(self, response):
        for title in response.css('.post-header>h2'):
            print('my_data -> ', self.settings['my_data'])
            yield {'title': title.css('a ::text').get()}

def start_spider(spider, settings: dict = {}, data: dict = {}):
    all_settings = {**settings, **{'my_data': data, 'TELNETCONSOLE_ENABLED':
False}}
    def crawler_func():
        crawler_process = CrawlerProcess(all_settings)
        crawler_process.crawl(spider)
        crawler_process.start()
    process = Process(target=crawler_func)
    process.start()
    return process

map(lambda x: x.join(), [
    start_spider(TestSpider, data={'data': 'test_1'}),
    start_spider(TestSpider, data={'data': 'test_2'})
])
```

Share Improve this answer Follow

answered Aug 24, 2019 at 14:15

hernan
582 5 10

It shows that "This code is unreachable" for the map function. Is the indentation wrong in the example? How do I get this running? Many thanks! – Juggernaut Aug 24, 2019 at 15:39

The indentation is fine. Copy the text into a `file.py` and then run it with `python file.py` and it should run. I tested it on Python 3.7.3. – hernan Aug 24, 2019 at 16:21

I copied this above code to a file as is and tried to run it. It says `Can't pickle local object 'start_spider.<locals>.crawler_func'` – yalkris Apr 12, 2023 at 6:43

If you are getting the pickle error above - see the below question. Switching to multiprocess worked for me. stackoverflow.com/questions/72766345/... – Squiggs. Jul 14, 2023 at 15:49