## 20200821信息安全实训

笔记本: 我的第一个笔记本

**创建时间**: 2020/8/21 18:19 **更新时间**: 2020/8/21 21:05

**作者:** 820410740@qq.com

全站爬虫 scrapy 爬虫框架 QuotesSpider类

```
1 hport scrapy
 3 class QuotesSpider(scrapy.Spider):
5
        nane = "quote"
        def start_requests(self):
8
             urls = [
                   "http://quotes.toscrape.com/page/1/",
"http://quotes.toscrape.com/page/2/",
9
10
11
              1
12
             for url in urls:
13
                   yield scrapy.Request(url=url, callback=self.parse)
14
15
        def parse(self, response):
             page = response.url.split("/")[-2]
16
             filename = 'quotes-xs.html' x page
with open(filename, 'wb') as f:
f.write(response.body)
17
18
19
20
             self.log('Saved file xs' x filename)
```

scrapy shell

css XPath

```
百月 (全年) 1902年 2012年 1903年
              scrapy
    2
    3 class LinkSpider(scrapy.Spider):
   5
            name = "links"
    6
   78
            start_urls = [
                   https://www.szlanyou.com',
   9
10
            1
   11
            def parse(self, response):
   12
                  for href in response.css('a::attr(href)').getall():
   13
                       print[href]
              scrapy
   3 class LinkSpider(scrapy.Spider):
  456
           name = "links"
  78
           start_urls = [
                 https://www.szlanyou.com',
  9
10
  11
           def parse(self, response):
  12
                for href in response.css('a'):
 13
                     yield {'link': href.css('::attr(href)').get(), }
   2 from urllib.parse import urljoin
   4 class LinkSpider(scrapy.Spider):
5
6 nanc = "links"
7
8 start_urls = [
                https://www.szlangou.com',
   10
   11
12
13
14
15
dd
dd
           def parse(self, response):
               for href in response.css('a'):
                   link = href.css('::attr(href)').get()
if "https://www.szlanyou.com" not link and link[0] == '/':
    link = urljoin('https://www.szlanyou.com', link)
    yield ('link': link, )
   16
17
18
                        yield response.follow(link, callback=self.parse)
```

指纹识别

消息摘要算法 MD5 SHA-1

```
hashlib
       3
                scrapy
the be
             scrapy.crawler import CrawlerProcess
selves.
       6 with open("./links.json", "r") as f:
             raw = f.read()
global
       9 raw = json.loads(raw)
e setti
      11 links = []
12 for i in raw:
             links.append(i["link"])
      13
cting t
      15 fgprts = []
      16 class MySpider(scrapy.Spider):
             name = "fgrprt"
ilt-in s
      19
             start_urls = links
      21
             def parse(self, response):
    md5 = hashlib.nd5()
                md5.update(response.body)
ch set
                 finger_print = md5.hexdigest()
                fgprts.append(
["url": response.url, "finger_print": finger_print]
e in P
      30 process = CraulerProcess()
/thon |
      31 process.crawl(MySpider)
      32 process.start()
         oo process.startt)
         36
         37 with open("./fgprts.json", "u") as f:
                  f.write(json.dumps(fgprts))
nechanicosptech@OspNo1:~/tutorial$
reasing (
                控制
                       柳图
                                   设备 帮助
           管理
                            規模
                       json
            z
                       hashlib
doesn't ex
            3
                       scrapy
            4 [ scrapy.crawler | crawlerProcess
the extrac
            6 with open('./fgprts.json') as f:
                   fgprts = json.loads(f.read())
dictionari
word in the
           9 urls = [i['url'] for i in fgprts]
           11 results = []
           12 class MySpider(scrapy.Spider):
           13
           14
                   name "check"
           15
           16
                   start_urls = urls
           17
           18
                   def parse(self, response):
           19
                        for i in fgprts:
                             if il'url'1 == response.url:
           20
           21
                                 finger_print = i['finger_print
               ٠,
           22
ote"):
```

```
CrawlerProcess
 6 with open('./fgprts.json', 'r') as f:
7  fgprts = json.loads(f.read())
 0
 9 urls = [i['url'] for i in fgprts]
11 results = []
                if i['url'] == response.url:
                    finger_print = il'finger_print'l
           md5.update(response.body)
           finger_print_now = md5.hexdigest()
                    'url': response.url,
                    'isDistorted': finger print == finger print now
33 process = craul(MySpider)
34 process.start()
36 with open('./results', 'w') as f:
       f.write('xs xs\n'x())
4 process.start()
6 with open('./results', 'w') as f:
7 for i in results:
           f.write['xs xs\n'x(i['url'], str(i['isDistorted']))]
              )
  31
  32 process = CraulerProcess()
     process.crawl(MySpider)
     process.start()
```