

## bash

```
1 scrapy startproject xxxPro  # 创建一个工程
2 cd xxxPro
3 scrapy genspider -t crawl xxx www.xxx.com
4 # run:
5 scrapy crawl sun
```

## settings.py

```
1 USER_AGENT = 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) \
2                                     Chrome/101.0.4951.64 Safari/537.36'
3 ROBOTSTXT_OBEY = False
4 LOG_LEVEL = 'ERROR'
5
6 ITEM_PIPELINES = {
7     'qiubaiPro.pipelines.QiubaiproPipeline': 300,
8 }
```

## items.py

```
1 import scrapy
2
3 class SunproItem(scrapy.Item):
4     title = scrapy.Field()
5     new_num = scrapy.Field()
6
7 class DetailItem(scrapy.Item):
8     new_id = scrapy.Field()
9     content = scrapy.Field()
```

## pipelines.py

```
1 class SunproPipeline(object):
2     conn = None
3     def open_spider(self, spider):
4         self.conn = spider.conn
5     def process_item(self, item, spider):
6         dic = {
7             'name': item['name'],
8             'desc': item['desc']
9         }
10        print(dic)
11        self.conn.lpush('movieData', dic)
12        return item
```

## redis Database shell

```
1 $ sadd name urls
2 -> (integer) 1
3
4 $ flushall
5 -> OK
```

## main.py

```

1 from scrapy.linkextractors import LinkExtractor
2 from scrapy.spiders import CrawlSpider, Rule
3 from sunPro.items import SunproItem, DetailItem
4 from redis import Redis
5 from moviePro.items import MovieproItem
6
7 class MovieSpider(CrawlSpider):
8     name = 'movie'
9     # allowed_domains = ['www.xxx.com']
10    start_urls = ['http://www.xxx.com']
11    # 链接提取器: 根据指定规则进行指定链接的提取
12    link = LinkExtractor(allow=r'type=4&page=\d+')
13    link_detail = LinkExtractor(allow=r'question/\d+/\d+\.html')
14
15    # 规则解析器
16    rules = (
17        Rule(link, callback='parse_item', follow=True)
18        Rule(link_detail, callback='parse_detail', follow=False)
19    )
20
21    conn = Redis(host='127.0.0.1', port=6379)
22
23    def parse_item(self, response):
24        tr_list = response.xpath('//*[@id="morelist"]/div//table/tr')
25        for li in li_list:
26            detail_url = 'https://www.xxx.com' + li.xpath('./div/a/@href').extract_first
27            ex = self.conn.sadd('urls', detail_url)
28            if ex == 1:
29                print('This url unaware to scrapy')
30                yield scrapy.Request(url=detail_url, callback=self.parse_detail)
31            else:
32                print('NDU')
33
34    def parse_detail(self, response):
35        item = MovieproItem()
36        item['name'] = response.xpath('/html//h1/text()').extract_first
37        item['desc'] = response.xpath('/html//p[5]/span/text()').extract
38        item['desc'] = ''.join(item['desc'])
39
40        yield item

```