bash

```bash
scrapy startproject xxxPro   # 创建一个工程
cd xxxPro
scrapy genspider -t crawl xxx www.xxx.com
# run:
scrapy crawl sun
```

settings.py

```
USER_AGENT = 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) \
                                              Chrome/101.0.4951.64 Safari/537.36'
ROBOTSTXT_OBEY = False
LOG_LEVEL = 'ERROR'

ITEM_PIPELINES = {
    'qiubaiPro.pipelines.QiubaiproPipeline': 300,
}
```

items.py

```
import scrapy

class SunproItem(scrapy.Item):
    title = scrapy.Field()
    new_num = scrapy.Field()

class DetailItem(scrapy.Item):
    new_id = scrapy.Field()
    content = scrapy.Field()
```

pipelines.py

```
class SunproPipeline(object):
    conn = None
    def open_spider(self,spider):
        self.conn = spider.conn
    def process_item(self, item, spider):
        dic = {
            'name': item['name'],
            'desc': item['desc']
        }
        print(dic)
        self.conn.lpush('movieData',dic)
        return item
```

Redis Database order

```
# 原生的 scrapy 不能实现分布式和增量式爬虫，要实现此功能要结合 scrapy-redis 组件

$ sadd name urls
-> (integer) 1

$ keys *
-> (1) 'urls'
-> (2) 'movieData'

$ flushall
-> OK
```

main.py

```python
from scrapy.linkextractors import LinkExtractor
from scrapy.spiders import CrawlSpider, Rule
from sunPro.items import SunproItem, DetailItem
from redis import Redis
from moviePro.items import MovieproItem

class MovieSpider(CrawlSpider):
    name = 'movie'
    # allowed_domains = ['www.xxx.com']
    start_urls = ['http://www.xxx.com']
    # 链接提取器: 根据指定规则进行指定链接的提取
    link = LinkExtractor(allow=r'type=4&page=\d+')
    link_detail = LinkExtractor(allow=r'question/\d+/\d+\.html')

    # 规则解析器
    rules = (
        Rule(link, callback='parse_item', follow=True)
        Rule(link_detail, callback='parse_detail', follow=False)
    )

    # 创建 redis 链接对象
    conn = Redis(host='127.0.0.1', port=6379)

    def parse_item(self, response):
        tr_list = response.xpath('//*[@id="morelist"]/div//table/tr')
        for li in li_list:
            detail_url = 'https://www.xxx.com' + li.xpath('./div/a/@href').extract_first
            ex = self.conn.sadd('urls', detail_url)
            if ex == 1:
                print('This url unaware to scrapy')
                yield scrapy.Request(url=detail_url, callback=self.parst_detail)
            else:
                print('No Data Updata')

    def parse_detail(self,response):
        item = MovieproItem()
        item['name'] = response.xpath('/html//h1/text()').extract_first
        item['desc'] = response.xpath('/html//p[5]/span/text()').extrct
        item['desc'] = ''.join(item['desc'])

        yield item
```