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'
```

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):
    def process_item(self, item, spider):
        if item.__class__.__name__ == 'DetailItem':
            print(item['new_id'], item['content'])
        else:
            print(item['new_num'], item['title'])
```

main.py

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

class SunSpider(CrawlSpider):
    name = 'sun'
    # 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)
    )

    def parse_item(self, response):
        tr_list = response.xpath('//*[@id="morelist"]/div//table/tr')
        for tr in tr_list:
            new_num = tr.xpath('./td[1]/text()').extract_first()
            new_title = tr.xpath('./td[2]/a[2]@title').extract_first()
            item = SunproItem()
            item['title'] = new_title
            item['new_num'] = new_num

            yield item

    def parse_detail(self,response):
        new_id = response.xpath('//span[2]/text()').extract_first()
        new_content = response.xpath('//tr[1]/text()').extract()
        new_content = ''.join(new_content)

        item = DetailItem()
        item['content'] = new_content
        item['new_id'] = new_id

        yield item
```