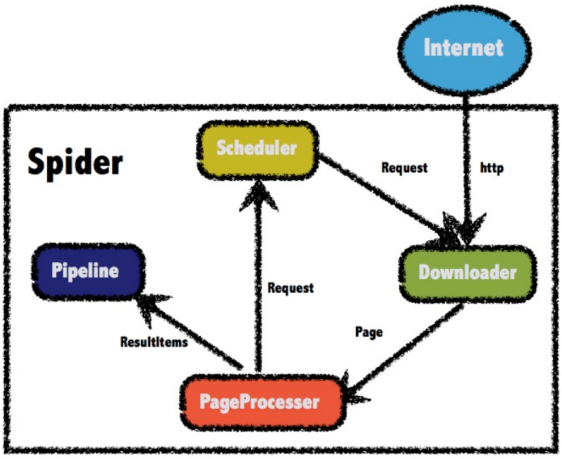
使用WebMagic爬取小说

一、环境搭建

- 开发工具： IDEA
- 爬虫工具：WebMagic 不使用Maven，将下载的jar包在IDE的项目中导入

二、编写代码

0.理解WebMagic处理流程



- 只需编写自己的PageProcessor提取网页信息，Pipline对信息进行持久化

1.MyPageProcessor类

- 继承自PageProcessor，实现process函数定义如何处理页面信息
- 本次爬虫所爬取的页面分为列表页和文章页，列表页需要抓取文章页url的地址并放入处理队列，文章页需要抓取文本内容而无需额外抓取页面内url地址

- 使用Xpath与正则表达式提取所在网页Html内的所需信息

```
public void process(Page page) {
    // 部分二：定义如何抽取页面信息，并保存下来

    //区分列表页和文章页
    if(page.getUrl().regex( s: "http://www.jianlaixiaoshuo.com/book/").match()){
        //获取当前文章页id
        String id = page.getUrl().regex( s: "[1-9]\\d*").toString();
        //更新最大页id
        if(Integer.parseInt(id) > maxPageId) maxPageId = Integer.parseInt(id);
        //存储该页标题
        page.putField( key: "title"+id,  field: page.getHtml().xpath("//h1/text()").toString() + "\n");
        //获取该页文本，分段，加换行并存储
        List<String> text = page.getHtml().xpath("//div[@id='BookText']/p/text()").all();
        String textString = "";
        for(int i = 0;i < text.size();i++){
            textString = textString + text.get(i) + "\n";
        }
        page.putField( key: "text"+id, textString);
    }else{
        //将列表页的所有文章链接加入请求队列
        page.addTargetRequests(page.getHtml().links().regex( s: "http://www.jianlaixiaoshuo.com/book/[1
        maxPageId = 0;
    }
```

2.MyPipline类

- 继承自Pipline，实现process函数定义如何持久化爬取信息

- 使用文件输出流将爬取内容转化为txt文件中的文本内容

```
public void process(ResultItems items, Task tasks){
    Map<String,Object> map = items.getAll();
    try{
        File f = new File( pathname: "C:\\Users\\Frey\\Desktop\\MyNovel.txt");
        FileOutputStream fop = new FileOutputStream(f,  append: true);
        String title;
        String text;
        for(int i = 1;i <= MyPageProcessor.maxPageId;i++) {
            title = "title" + String.valueOf(i);
            text = "text" + String.valueOf(i);
            if(map.get(title) != null){
                System.out.println(map.get(title).toString() + " √");
                fop.write(map.get(title).toString().getBytes());
                fop.write(map.get(text).toString().getBytes());
                fop.flush();
            }
        }
        //debug
        fop.close();
    }catch (IOException e) {
        System.out.print("Exception");
    }
```

3.测试与结果

- MyPageProcesser中可调整线程数与休眠时间，爬取时Console输出已经爬取到的文章页的标题

```
第二十四章 相赠
√
第二十五章 离别
√
第二十六章 好说话
√
第二十七章 点睛
√
第二十八章 财迷
√
第二十九章 狐魅
√
第三十章 暗室
√
```

- 在MyPipline中可定义输出文件地址，默认为`"C:\\Users\\Frey\\Desktop\\MyNovel.txt"`，需要根据用户修改地址

- 结果文件截图


MyNovel.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
第五百七十五章 于剑修如云处出拳
宁姚在斩龙崖之上潜心炼气。
陈平安没去凉亭那边，留在小宅屋内修行。
宁姚还有些疑惑，因为斩龙台那边明显灵气更为充沛，是整座宁府最佳修道之地。虽说陈平安不是剑修，
陈平安有些无奈，只是看着宁姚。
宁姚便撂下一句，难怪修行这么慢。
陈平安就更无奈了。
在北俱芦洲春露圃、云上城，宝瓶洲朦胧山这些山头，十年之内，跻身四境练气士，真不算慢了。
可惜在剑气长城，陈平安的修行速度，那就是装钱所谓的乌龟挪窝，蚂蚁搬家。
可哪怕是这位开山大弟子，不说她那练拳，只说那剑气十八停，自己这个当师父的，当年就算想要传授一
尤其是宁姚，当年提及阿良传授的剑气十八停，陈平安询问剑气长城这边的同龄人，大概多久才可以掌握
所以那会儿，陈平安甚至会觉得老大剑仙说自己有一份地仙资质，都只是在安慰人。
约莫两个时辰后，陈平安以内视洞天的修行之法、沉浸在木宅的那粒心念芥子，缓缓退出人身小天地，长
陈平安双手笼袖，斜靠廊柱，满脸笑意。
瞧瞧，我一眼相中的姑娘，用心修行起来，厉害不厉害？
在陈平安偷着乐呵的时候，老者无声无息出现在一旁，好像有些惊讶，问道：“陈公子瞧得见那些遗留在
陈平安赶紧站好，答道：“纳兰爷爷，只看得出些端倪，看不太真切。”
纳兰夜行点头笑道：“只说陈公子的眼力，已经不输咱们这边的地仙剑修了。”
陈平安轻声问道：“宁姚何时能够破开金丹瓶颈？”
纳兰夜行说道：“最少得等到下一场大战落幕吧。”
陈平安问道：“宁姚与他朋友每次离开城头，如今身边会有几位扈从剑师，境界如何？”