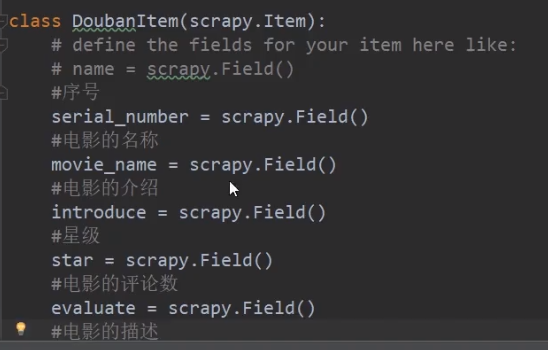
1新建项目  
`scrapy startproject douban`  
`scrapy genspider douban\_spider movie.douban.com`: 再切换到spiders目录下生成douban\_spider。正则表达式写在这里  
settings文件是配置信息

2

明确目标，items文件，填写需要获得的信息



3

编写爬虫文件douban\_spider，爬虫名与项目名不能重复

`scrapy crawl douban\_spider` 启动爬虫

需要更改useragent 从浏览器里面复制过来就好

使用main.py启动爬虫，而不是手动启动

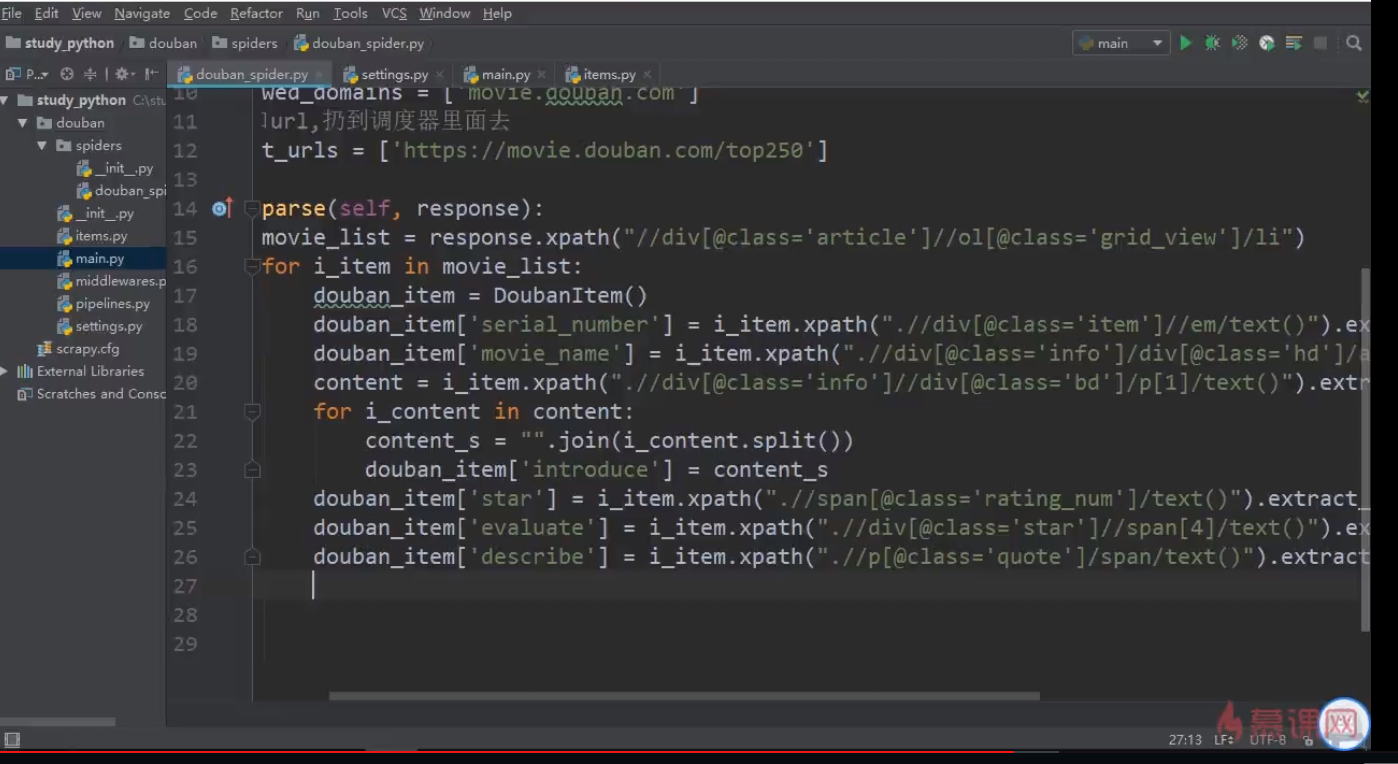
```

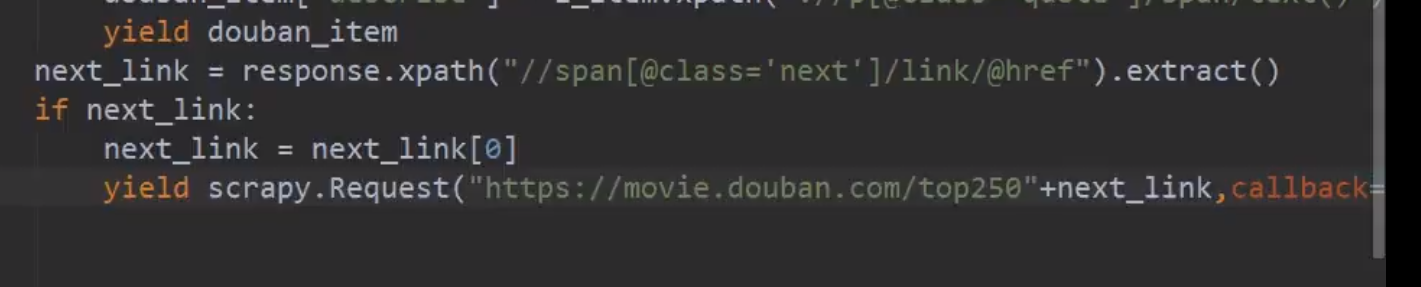
from scrapy import cmdline

cmdline.execute("scrapy crawl douban\_spider".split())

```

编写spiders文件





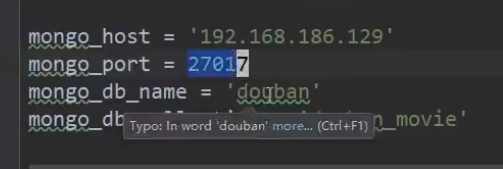
导出结果：

Scrapy crawl douban\_spider -o aaa.json

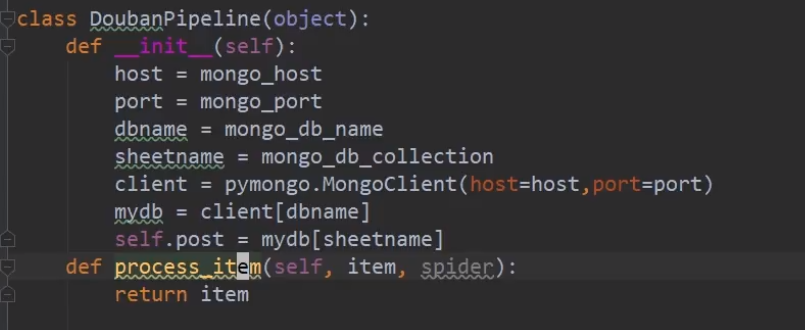
Scrapy crawl douban\_spider -o aaa.csv =🡺转换成UTF8-BOM格式后可以用excel打开

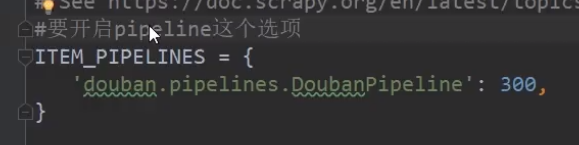
# 4 mongodb

链接mongodb。。setting.py



之后就可以在pipline里编辑 安装pymongo。。。。。需要在setting里打开itempiplines



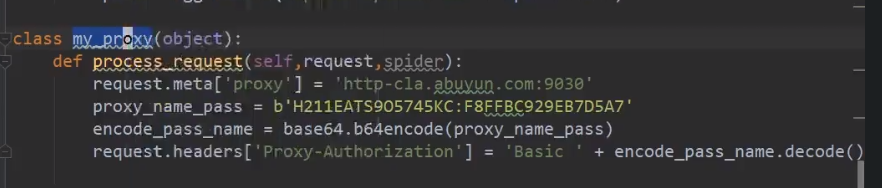


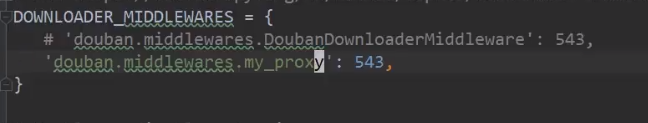
# # 爬虫的伪装

1. 代理ip
2. 随机useragent

**Middlewares 中编写代理ip**

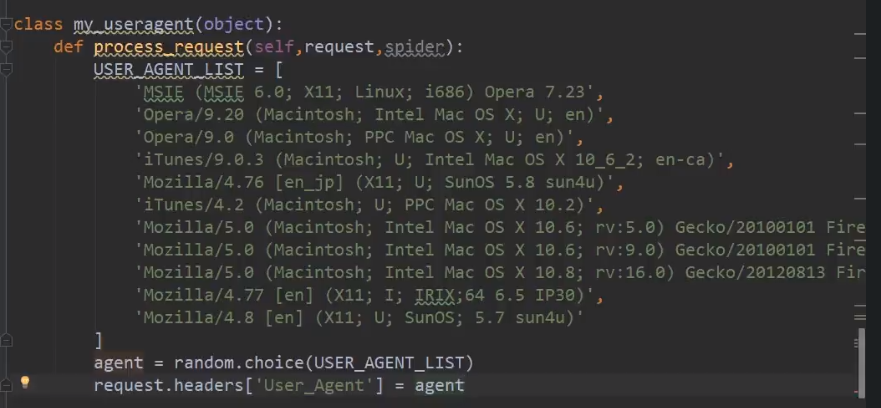
设置setting，下载中间件。优先级越小越高



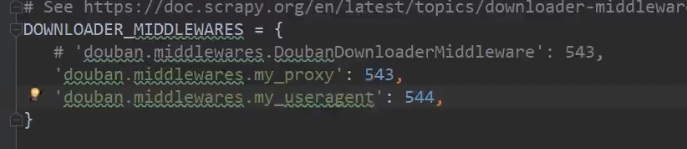


# 随机useragent

中间件py：先从晚上找useragent列表



同样在setting里开启useragent优先级不能相同



**注意**

