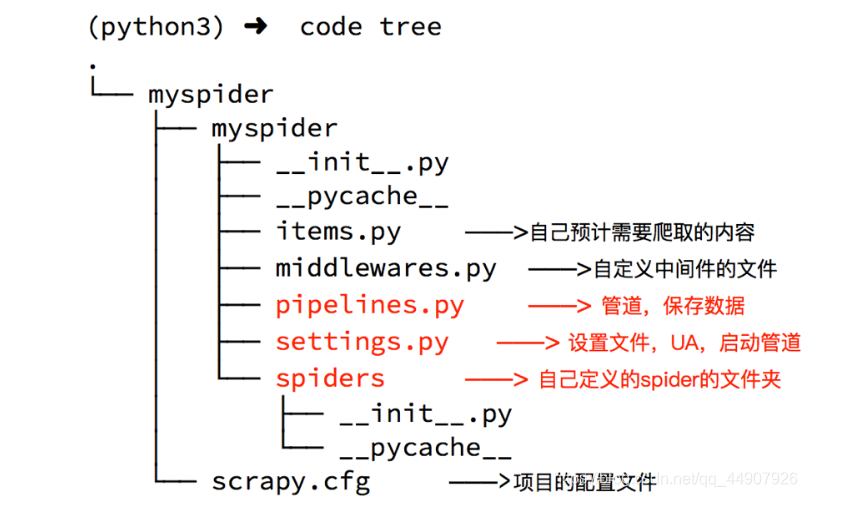
pip install scrapy 用清华源会快很多

scrapy startproject myspider 创建项目【解决“ImportError: DLL load failed while importing \_rust: 找不到指定的程序的问题】



cd myspider

scrapy genspider itcast itcast.cn 创建爬虫文件

scrapy crawl itcast 运行爬虫的时候注意启动的位置，是在项目路径下启动;

parse()函数中使用yield返回数据，注意:解析函数中的yield能够传递的对象只能是:Baseltem,Request,dict, None.

打开我们的shell交互式平台。再此项目中：输入命令scrapy shell [movie.douban.com/top250](https://link.juejin.cn/?target=https://movie.douban.com/top250" \o "https://movie.douban.com/top250)

从selector对象中提取电影名字使用selector对象的方法.extract()。这个方法可以提取到selector对象中data对应的数数据

response.xpath('//div[@class="info"]/div/a/span[1]/text()').extract()

item管道的典型用途是：

清理HTML数据验证已删除的数据（检查项目是否包含某些字段）检查重复项（并删除它们）将已爬取的item进行数据持久化