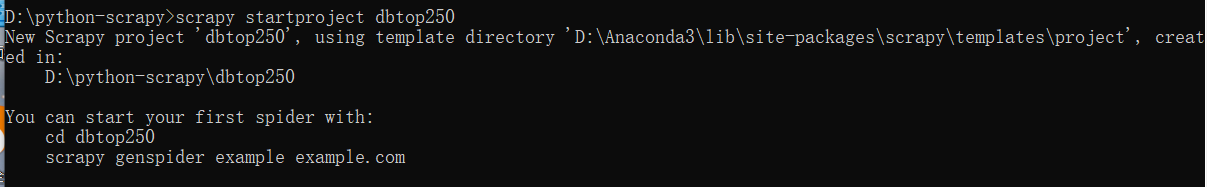
1. 创建项目

scrapy startproject 工程名

出现：



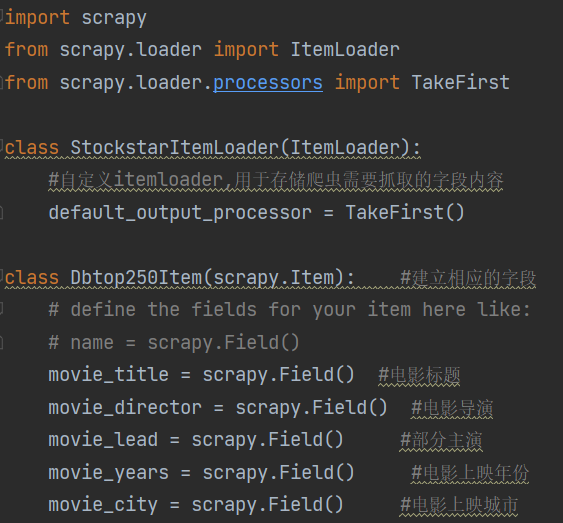
代表创建成功

创建 scrapy 项目以后，在 settings 文件中有这样的一条默认开启的语句。

POBOTSOXT\_OBEY = True

robots.txt 是遵循 Robot 协议的一个文件，在 Scrapy 启动后，首先会访问网站的 robots.txt 文件，然后决定该网站的爬取范围。有时我们需要将此配置项设置为 False。在 settings.py 文件中，修改文件属性的方法如下。

ROBOTSTXT\_OBEY=False

1. 明确爬虫目标

构建 item 模型（model）

1. 制作爬虫

在编写爬虫逻辑之前，需要在 stockstar/spider 子文件下创建 .py 文件，用于定义爬虫的范围，也就是初始 URL。接下来定义一个名为 parse 的函数，用于解析服务器返回的内容。  
  
首先在 CMD 中输入代码，并生成 spider 代码，如下所示(需先定位至相应子文件夹下)：  
scrapy genspider stock url

此时spider 子文件夹下会出现 stock.py 的文件

生成代码如下：

Start\_urls即最终爬取网址

再定义parse前需要编辑请求头等相关信息如下

重新定义函数start\_request来请求

For循环遍历获取start\_urls下的所有网址

最终爬虫逻辑，如下：

Yield item这一句映射方便后续保存

Return item返回最终数据

运行并保存数据

Json: scrapy crawl 工程名 -o 文件名.json



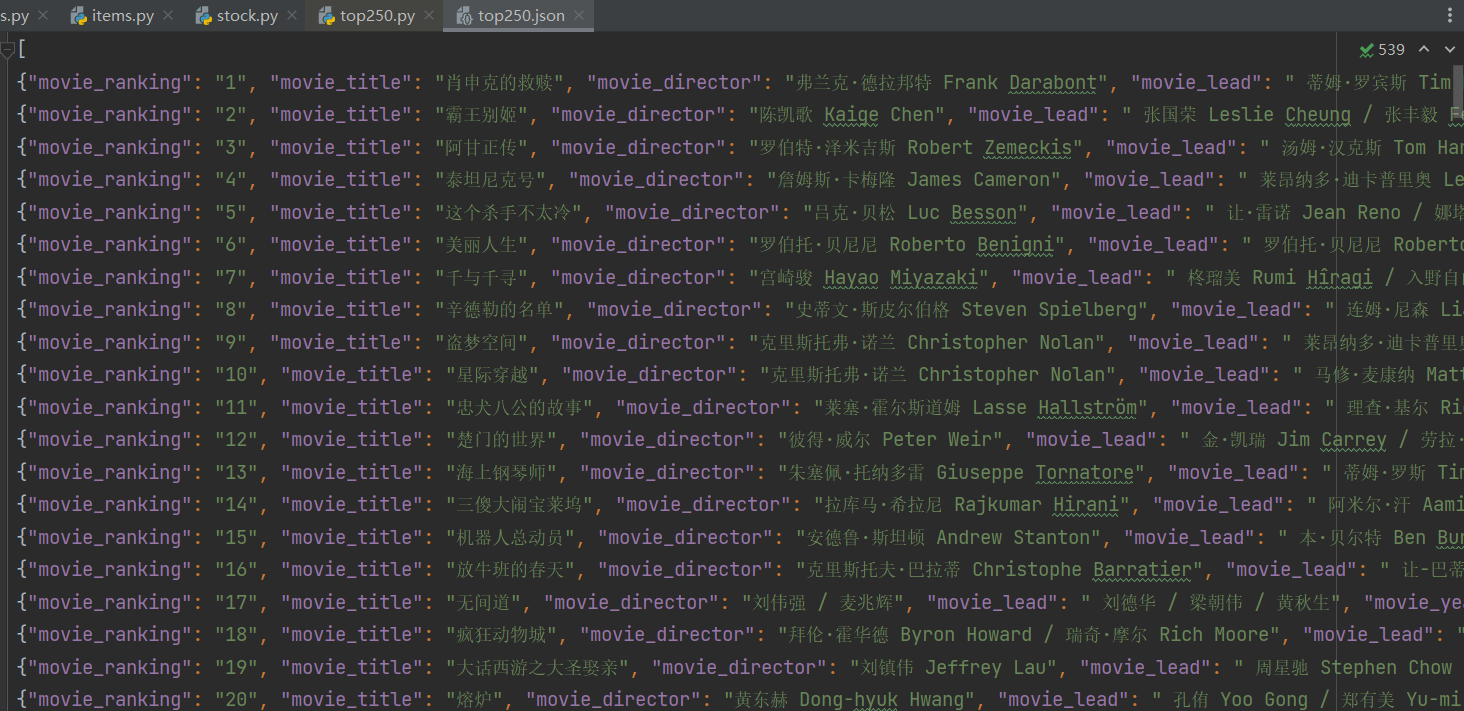
Csv: scrapy crawl 工程名 -o 文件名.csv



还有xml以及jsonl格式 只需修改文件后缀名即可

需要在settings.py中设置编码格式防止乱码

FEED\_EXPORT\_ENCODING = "utf-8"

最终保存JSON文件：

注：如果有数据无法正常显示，一般都是编码格式不匹配的问题