一、选题为豆瓣电影Top250爬虫

二、方案设计：

1、整体思路：

1.利用requests库发送HTTP请求，获取网页内容

2.利用beautifulsoup4库解析HTML页面，提取电影信息

3.利用pymysql库连接MySQL数据库，存储爬取的数据

4.利用从网上获取的600个动态IP代理，以随机ip地址的方式应对豆瓣的反爬虫机制

2、面向对象设计：

1.定义Movie类，用于存储电影信息

2.定义Comment类，用于存储评论信息

虽然定义了两个类，但其功能并不是很多，主要用于模块化存储数据

三、代码模块的功能划分：

1.爬取模块：负责发送请求和解析页面

定义了get\_ip()和request\_with\_random\_ip()函数，用来获取随机IP地址和发送请求

2.数据处理模块：负责数据的提取和存储

定义了get\_movie\_info()和get\_comment\_info()函数，用来提取电影和评论信息

定义了crawl\_page()和crawl\_comments()函数，用来爬取页面数据，本质上是HTML解析

3.数据库模块：负责与MySQL数据库的交互

定义了save\_to\_mysql()、save\_movies\_to\_mysql()和save\_comments\_to\_mysql()函数，用来保存数据到MySQL数据库

4.主程序：负责程序的入口和控制流程

定义了main()函数，用来控制程序的执行流程

5.配置模块：负责配置数据库连接参数和爬虫设置

定义了config.py文件，用来配置数据库连接参数和爬虫设置

6.ip模块：负责IP代理的获取和管理

储存了600个IP地址，用于应对豆瓣的反爬虫机制

7.requirements.txt模块：负责依赖包的管理

定义了requirements.txt文件，用来管理项目依赖包

四、实现效果展示（包含演示视频或动图）