**开源软件基础大作业**

**豆瓣TOP250电影信息爬虫抓取**

**Douban TOP250 movie information crawler crawling**

学 生 姓 名： 孙鹏飞

学 号： 201892432

开源软件基础大作业项目报告

目 录

1 [引 言 3](#_Toc7959)

[2 项目创新点 4](#_Toc14054)

[3 项目设计 4](#_Toc31349)

[4 项目实现 5](#_Toc1672)

[5 项目测试 5](#_Toc14074)

[6 项目分析](#_Toc23762) 6

[7 项目不足和改进之处](#_Toc26162) 6

[8 项目总结](#_Toc14346) 6

[9 项目感悟 6](#_Toc7307)

# 引 言

开源软件基础是一门很全面的软件开发与测试方面课程，在这门课程中，我们不仅学习到了与开源软件项目有关的知识和理念，同时在课程的后半段学习了部分python语言知识和爬虫基础。

在本次大作业中，我从自身的角度出发，抓取豆瓣排名前250的电影信息，解决了假期“电影荒”的问题。

本报告包括项目背景以及意义，项目创新点，项目设计，项目实现，项目测试，项目分析，项目不足以及可改进之处，项目总结和感想各个方面，对本次大作业进行一定程度的总结和汇报。

1 项目背景及意义

近几年，随着社会的发展和进步，人们的物质文化需求也日益增强。为了方便大家寻找品质精良，值得一看的电影，我设计并制作了这一个爬虫，用来抓取豆瓣排名前250的优良电影。

# 2 项目创新点

可将信息存储于本地的xls表格中，方便浏览和存储。

# 3 项目设计

通过对豆瓣TOP250网站的爬取，获取TOP250的电影introduce信息，电影名，评价人数，豆瓣分数的信息，并且将爬取内容存放进xls表中方便查看浏览和更改。每一次运行程序都可以更新xls表格中的信息，使榜单内容为最新内容。

# 4 项目实现

使用requests包中的requests.get来获取豆瓣电影TOP250排行榜单中的信息，在浏览器中使用开发者模式查看网址，寻找网站中的信息，

name = li.xpath(".//a/span[@class='title'][1]/text()")[0]

director\_actor = "".join(li.xpath(".//div[@class='bd']/p/text()[1]")[0].replace(' ','').replace('\n','').replace('/','').split())

info = "".join(li.xpath(".//div[@class='bd']/p/text()[2]")[0].replace(' ','').replace('\n','').split())

rating\_score = li.xpath(".//span[@class='rating\_num']/text()")[0]

rating\_num = li.xpath(".//div[@class='star']/span[4]/text()")[0]

introduce = li.xpath(".//p[@class='quote']/span/text()")

之后将信息整合

movie = {'name': name, 'director\_actor': director\_actor, 'info': info, 'rating\_score': rating\_score, 'rating\_num': rating\_num, 'introduce': introduce[0]}

再统一保存到列表中

movies.append(movie)

# 5 项目测试

程序截图



Xls表格



# 6 项目分析

近几年，随着社会的发展和进步，人们的物质文化需求也日益增强。为了方便大家寻找品质精良，值得一看的电影，我设计并制作了这一个爬虫，用来抓取豆瓣排名前250的优良电影。

项目贴合大学生的物质文化需求，能够节省很多的时间，提供优质的电影名单。

# 7 项目不足和改进之处

7.1 项目的不足之处：

（1）爬取的内容过于多，大量的数据不利于快速阅读。

（2）没有进行数据的分析处理，只是单纯的罗列信息。

（3）可拓展性较差，只可以爬取TOP250网站

7.2 项目可改进之处：

（1）优化爬取的信息内容。

（2）对获取到的信息进行分析处理，最好有柱状图或者扇形图的示例

（3）爬取获得电影的图片，使内容更加的丰富

（4）增强可拓展性，可以用于爬取不同的网站来获取信息

# 8 项目总结

8.1 成员分工：

孙鹏飞 工作占比（100%）

# 9 项目感悟

本次开源软件基础课程大作业的制作，我在开源软件课程的学习基础之上，继续学习python基础知识、爬虫相关技术并自学网站框架的架构和制作，在积极巩固老师讲授课程内容的基础上，拓宽自身视角，为大作业的技术方面打下更坚实的基础。项目从选题开始，我们便从实际出发，选取较为切合自身的“豆瓣TOP250抓取”主题项目，并且通过测试，并在项目报告编写阶段不断发现错误，改进缺陷，得到最终的项目成果。虽然项目仍有很多的不足以及可以改进的地方，但是我认为我还是取得了一个比较满意的结果，在未来的学习生活中，我一定会更加努力地自学相关知识，提升自己的编码水平，争取在核心技术方面做出更多贡献。