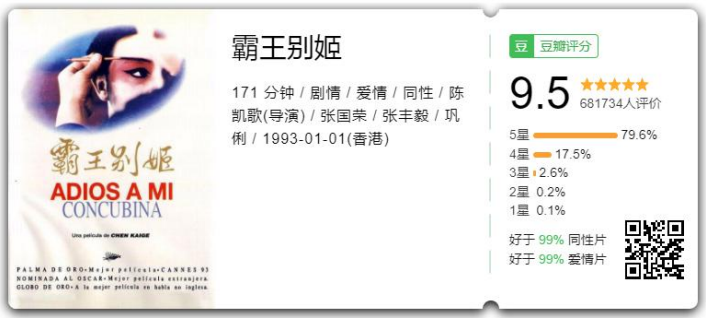
目录

1. 问题定义

1.1 问题背景

比起售票/售书 app 上对于电影或书籍的评分，豆瓣网的评分制度无疑更加值得信赖。人们吐槽一部烂片的方法是去豆瓣上给它打一星，指摘一本书籍的方法则是不仅打一星，还要在豆瓣上为它添加“浪费纸张”的标签。豆瓣评分越来越成为大众判断作品质量好坏的标杆，很多人习惯了在读书或观影前，先去豆瓣看看评分。8 分以上可以称为佳作，0-3 分则基本意味着这是一部骗钱之作。

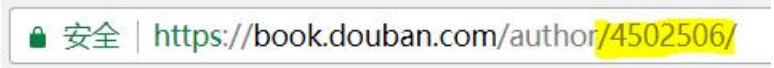


豆瓣上丰富的书籍、电影条目也成为爬虫学习者争相爬取的对象，豆瓣为这些开发者提供了豆瓣 API[1]，提供了获取图书信息的 API。尽管如此，豆瓣并未提供作家页面的爬取接口。可能正是因为这个原因，尽管网络上爬取豆瓣的教程丰富，但是并没有爬取豆瓣上作家信息的资源。

那么，如何利用豆瓣读书中丰富的作家信息，发现一些有趣的冷知识呢？

1.2 需求分析

豆瓣读书并不直接为用户提供作家链接，用户想要访问作家主页，可以在搜索框进行搜索，或是在图书信息页点击作家栏，跳转到作家主页。尽管作家主页的 URL 有统一格式，https://book.douban.com/author/XXXX/，（我们把 XXXX 这部分称作作家 ID）。



但是通过分析我们发现，作家 ID 并不是连续分配的，通过从 000001 到 999999 遍历作家 ID 这样的方式来请求作家页面得到的绝大多数都是无效网址。或许这也暗示了，豆瓣上的作家主页数量并不是很多，否则作家 ID 会分布得更加密集。

由此我们想到可以采用先爬取大量图书信息，通过图书页面跳转到作家主页，然后获取作家信息。

获取信息之后需要对数据进行分析，最后将分析结果以可视化的方式展示出来。

1.3 可行性分析

可爬取的资源非常丰富，又几乎没有什么反爬虫措施的网络小说门户可以说是爬虫初学者的“hello world”，比起这些网站，经常被爬虫学习者拿来练手的豆瓣的反爬措施做得更好。

未登录状态下爬取太过频繁，会要求登录，登录之后爬取太过频繁，会封账号（我被封了两次号，现在还没解封）。账号解封需要复杂的验证。有时会返回这样的页面：“你的操作很像机器人，请输入验证码证明你不是机器人”,比起这些，更狠的是返回一个假页面，里面都是乱码。

既然这么麻烦，豆瓣还能不能爬呢？

其实这些反爬措施，大致都是在检测同一个 IP 的频繁的、不自然的操作。只要不停的变换 IP，基本不会被反爬。文档开篇，祭出淘宝上 7 块钱可以用 24 小时的神器——PPTP 动态 ip，每隔 60 秒自动换一个 ip，几乎是畅通无阻。

2. 总体设计和详细设计

2.1 总体设计

项目应分为数据获取、数据清洗、数据分析、数据可视化四部分。

2.2 详细设计

2.2.1 数据爬取

使用 python 的 request、beautifulsoup、re 包，首先在豆瓣读书按标签推荐图书的页面，爬取不同标签下的图书信息，从图书页跳转到作者页，获取作者姓名、性别、国籍、生日信息，再从作者页跳转到作者作品页，爬取作者的代表作，最后将作者姓名、性别、国籍、生日信息、代表作信息一并存入 csv 文件中。

为了提高爬取速度，采用多线程爬虫。

2.2.2 数据清洗

一位作家的作品可能会反复在不同标签以及同一标签下的不同页数出现，导致爬取到的作家条目时有重复。在数据清洗这一步，用 wps 打开 csv 文件，选择“删去重复项”即可。

2.2.3 数据分析

通过爬取到的作家信息，希望获取如下知识：豆瓣上平均每部作品被打分最多的作家是谁？各个国家最受豆瓣用户追捧的作家分别是谁？谁是豆瓣读书上最冷门小众的作家。口碑最高的作家当中，男女比例如何？

2.2.4 数据可视化

Matplotlib 是 python 一个强大的库，我们用它绘制图形，展示分析结果。

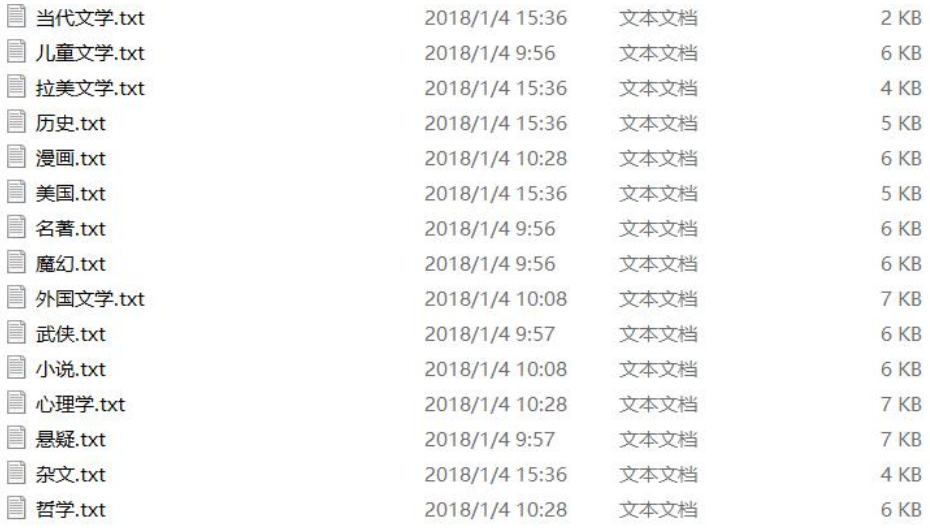
表格采用 excel 绘制。

3. 实现

3.1 get_book_by_label.py

选取拉美文学、美国、历史、杂文、当代文学小说、外国文学、漫画、武侠、哲学、心理学、儿童文学、魔幻、悬疑、名著标签，每个线程爬取一个标签下的前 800 本图书（每个页面有 20 本书，也就是爬取前 800 本书）

每获得一个图书 id，就检测它是否已经存在，如果不存在，就存入 txt 文件。最终得到 15 个存储该标签下的大约 800 本图书 ID 的 txt 文件。最终得到 8263 本书。



3.2 get_authorID.py

循环读取所有的 label.txt（如“当代文学.txt”）文件,遍历每本图书 ID，请求图书页面，返回报文用 beautifulsoup 处理之后找到存有作者主页的<a>标签，提取出 href，保存到 authorID.txt。对每一个 label.txt 的读取都新建一个 thread，多线程爬取提升速度。

3.3 getAuthorsCSV_multiThread.py

循环读取 author.txt 中的作家 ID，请求作家页面，用 beautifulsoup 解析页面，获取作者的姓名、性别、国家、生日、代表作，存入 author.csv。最终我们得到 591 位作家的信息。

3.4 authorAnalysis.py

分析 author.csv 中的数据并绘图。

4. 数据分析

4.1 谁是豆瓣平均得分最高的作家？

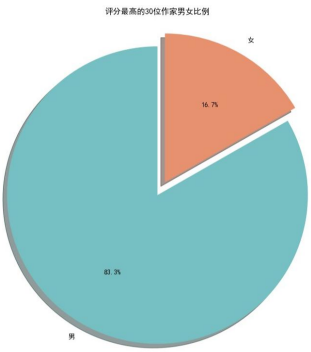
排名	作家	代表作平均评分	性别	国家
1	侯世达	9.4	男	美国
2	陈寅恪	9.4	男	中国
3	尾田荣一郎	9.4	男	日本
4	荒川弘	9.4	女	日本
5	乔治·R·R·马丁	9.366666667	男	美国
6	浦泽直树	9.366666667	男	日本
7	藤子·F·不二雄	9.366666667	男	日本
8	田中芳树	9.266666667	男	日本
9	阿兰·摩尔	9.22	男	英国
10	谭其骧	9.216666667	男	中国
11	安房直子	9.216666667	女	中国
12	阿瑟·柯南·道尔	9.2	男	英国
13	傅惟慈	9.183333333	男	中国
14	弗兰克·米勒	9.15	男	美国
15	仇鹿鸣	9.1	男	中国
16	西崛定生	9.1	男	日本
17	豪·路·博尔赫斯	9.083333333	男	阿根廷
18	伊曼努尔·康德	9.083333333	男	德国
19	富坚义博	9.066666667	男	とがし よしひろ
20	司马迁	9.05	男	中国
21	汪曾祺	9.05	男	中国
22	路易斯·塞尔努达	9.016666667	男	西班牙
23	松本零士	9.016666667	男	日本
24	艾萨克·阿西莫夫	9.016666667	男	美国
25	阿图尔·叔本华	9.016666667	男	德国
26	手塚治虫	9.016666667	男	日本
27	叶广芩	9	女	中国
28	安东尼·德·圣·埃克苏佩里	9	男	法国
29	托芙·扬松	9	女	芬兰
30	盐野七生	9	女	日本

谁是豆瓣评分最高的作家，想必每一位文学爱好者都会有一个自己猜想的答案吧。我原猜是曹雪芹，因为他只有一部《红楼梦》，一部即是中国古典文学的巅峰，一部书养活了多少红学家。但爬取的结果并不是曹雪芹。因为豆瓣没有曹雪芹的作者主页。曹雪芹无豆瓣主页，最大赢家竟是——侯世达，美国人，迷之译名让人误以为他是旅居海外的中国学者（其实不是），他何以位居榜首呢。原因在于他收录在豆瓣的作品也只有一部，也是一部封神——《哥德尔、艾舍尔、巴赫》，评分 9.4，不过看过评论区之后还是决定怀着朝圣的心情重新拜读。

屈居第二的是陈寅恪先生，我对先生并不了解，只把这段简介摘抄下来——中国现代最负盛名的集历史学家、古典文学研究家、语言学家、诗人于一身的百年难见的人物，与叶企孙、潘光旦、梅贻琦一起被列为清华大学百年历史上四大哲人，与吕思勉、陈垣、钱穆并称为“前辈史学四大家”。先后任职任教于清华大学、西南联大、广西大学、燕京大学、中山大学等。陈寅恪之父陈三立是“清末四公子”之一、著名诗人。祖父陈宝箴，曾任湖南巡抚。夫人唐筼，是台湾巡抚唐景崧的孙女。因其身出名门，而又学识过人，在清华任教时被称作“公子的公子，教授之教授”。

最令人意外的是，尾田荣一郎，以全 9 分+的系列神作《海贼王》打败了司马迁、汪曾祺、叔本华这些文学大家，登上第三宝座。《海贼王》是漫画史上又一部堪称热血旗帜的划时代作品，如今的热血漫画就是属于尾田荣一郎《海贼王》的时代。

4.2 上榜作家的性别分布

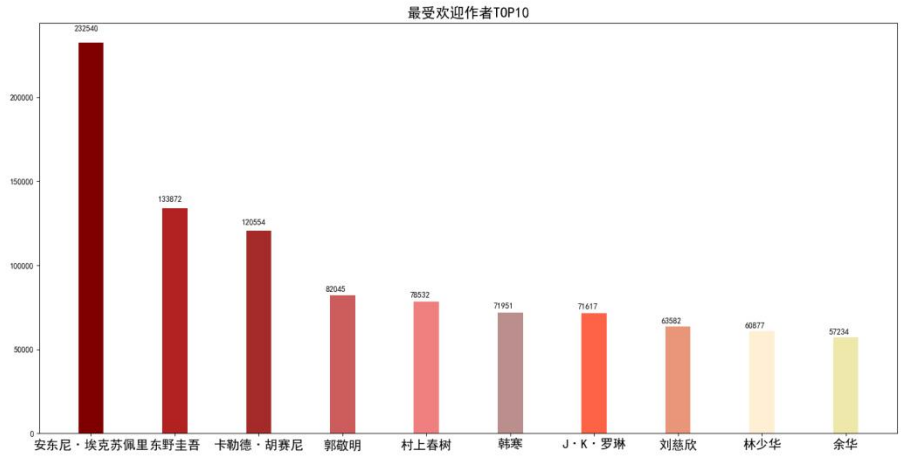


在平均作品得分最高的作家中，女性只占了 16.7%，女性中排第一的还是漫画家荒川弘，看来二次元是豆瓣用户中一股强大的势力，第二名安房直子是日本儿童文学作家，代表作《风与树的歌》评分 9.3。第三名为满族作家叶广岑，代表作《采桑子》9.0 分。

为什么上榜的男性作家几乎是女性作家的五倍之多呢？上榜作家的平均年龄几乎要被拉回大清朝了（考虑到里面还有司马迁），在他们的时代，女性普遍没有被赋予平等的接受教育的机会，赋予男孩更多雄心而赋予女孩更多温驯的社会观念还没有遭到像今天这样多的质疑。

今天我们似乎生活在一个更平等的时代了，但是我们仍然能注意到，根深蒂固的性别偏见仍然存在。即使是在大学里，诸如“我觉得像你这样的女孩以后留校就挺好”，“没想到咱们这个课女同学表现的都这么积极，我太意外了”这种潜意识认为女性在社会上不如男性有竞争力、男生占据着课堂主导权的观念仍然在困扰着我们。或许培养出下一位伟大的女性文学家，我们的社会还有很长的路要走。

4.3 豆瓣热门作家 top10



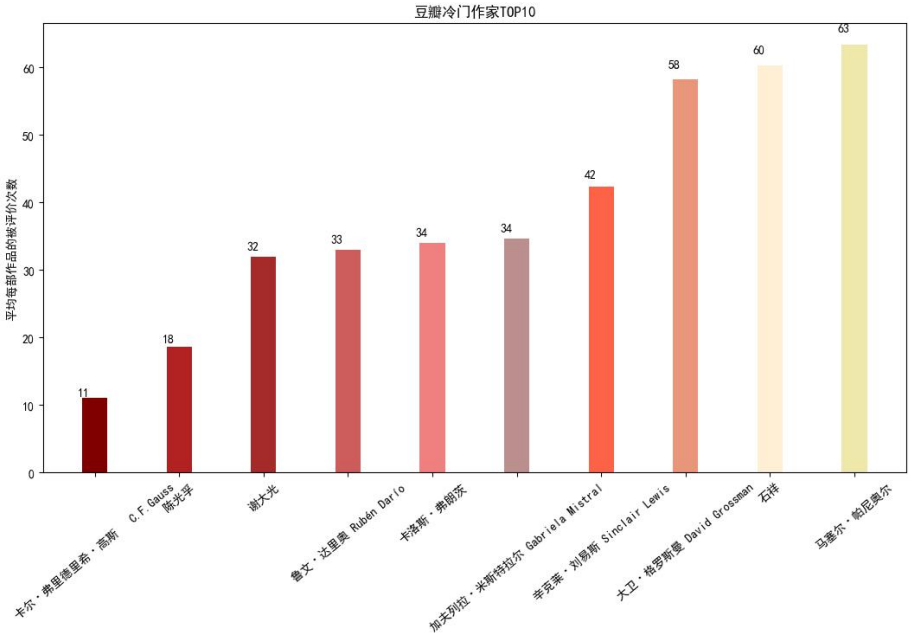
安东尼•德•圣埃克苏佩里凭借通俗读物《小王子》荣登被评价次数最多的作家第

一位。《小王子》有多受中国读者追捧，从它记录在豆瓣的 460 种版本就可见一斑。看似是用孩童般的口吻叙述的童话，却处处引发成年人的共鸣，最让它畅销的原因在于，它浅显易读，看似寓意深刻的地方实则道理浅显。通俗的即是大众的。

林少华先生作为译者和村上春树一同上榜，可见林译版的《挪威的森林》多么深入人心了。

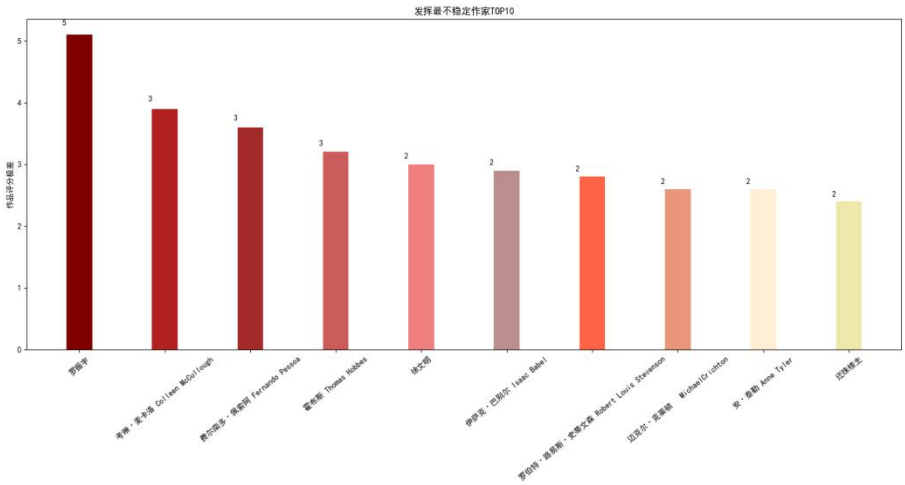
两位总是被放在一起比较的青年作家郭敬明和韩寒也一同上榜了，不过现在他们都成了商人。想当年韩寒连发“韩三篇”，嬉笑怒骂，少年意气，不免感慨。

4.4 豆瓣冷门作家 TOP10



冷门作品多是因为曲高和寡，“数学王子”高斯著有《算术探究》，评价人数不足没有评分。陈光孚译著有《拉丁美洲短篇小说选》、著有《魔幻现实主义》。谢大光作为编辑，编选了《俄罗斯散文经典》、《日本散文经典》等系列丛书。鲁文达理奥是尼加拉瓜天才诗人。大家有兴趣可以百度后面的作家，反正这个榜单上的人除了第一位以外我都不认识。

4.5 发挥最不稳定作家 TOP10



排名第一的罗振宇是自媒体视频脱口秀《罗辑思维》主讲人，得到 App 创始人 2013 年出版的《逻辑思维》7.1 分成绩尚可，而 2017 年出版的《终身学习》却被 93.8%的人打了一星，总分低到 2.3 分。2017 年双十一这天，罗振宇的新书《终身学习》迎来了它一生中的巅峰，以豆瓣 9.0 分摘下了本天度的“非虚构之王”的桂冠，然而众多豆瓣大 V 同一时段的密集四星五星评分收到之一，之后又被人爆料是收钱给好评带节奏，大量用户涌入为这本书打一星，仅用了几个小时把这本书从 9.0 分刷到了 2.3 分。[3]

澳大利亚女作家考琳·麦卡洛写出了畅销全球的《荆棘鸟》，然而《恺撒大传》被评价为“故事性差”、“读不下去”、“翻译太差”，不过确实因为作品有失水准而上榜的她，比起紧随其后的费尔南多佩索阿也倒不算冤枉。

费尔南多佩索阿是葡萄牙诗人，《惶然录》评分 9.0，然而 2013 年在中国出版的《我的心略大于整个宇宙》却被 43.5%的用户打了一星，而另外有近 30%的用户打了五星。评分两极分化来源于一场抄袭风波，闵雪飞指出该书的译者韦白抄袭了自己和杨铁军先前对费尔南多作品的翻译，尽管韦白否认抄袭，然而出版该书的世纪文景出版社发表声明，“《我的心略大于整个宇宙：佩索阿诗选》（韦白译）涉嫌对杨铁军、闵雪飞译作《斜雨》组诗等内容存在著作权侵权行为。作为该书出版方，世纪文景决定即日起对该出版物停止发货”[4]，并表示对抄袭行为零容忍。而这场抄袭风波直接导致了大量用户恶意差评以表达对韦白抄袭的不满。无辜的费尔南多莫名躺枪，若他泉下有知，不知道作何感想？

4.6 各个地区最受欢迎和评分最高的作家

国家	最受欢迎作者	平均每部作品评价人次	平均作品得分	评分最高作者	平均每部作品的评价人次	平均作品得分
英国	J·K·罗琳	71617.67	8.88	阿兰·摩尔	1023.4	9.22
美国	卡勒德·胡赛尼	120554.33	8.7	侯世达	4505	9.4
加拿大	亦舒	18614	7.83	露西·莫德·蒙哥马利	275.5	8.52
德国	赫尔曼·黑塞	5134.17	8.92	伊曼努尔·康德	993.67	9.08
俄罗斯	列夫·托尔斯泰	4901.17	8.55	费奥多尔·陀思妥耶夫斯基	4208.17	8.9
日本	东野圭吾	133872.5	8.35	尾田荣一郎	4576.67	9.4
中国台湾	龙应台	31605.17	8.67	袁哲生	166.17	8.93
中国香港	梁文道	14348.83	7.78	李碧华	7880.5	8.28

德国队的代表选手赫尔曼·黑塞（46 岁入瑞士籍）一生曾获多种文学荣誉，比较重要的有：冯泰纳奖、诺贝尔奖、歌德奖。1946 年获诺贝尔文学奖。作品多以小市民生活为题材，表现对过去时代的留恋，也反映了同时期人们的一些绝望心情。主要作品有《彼得·卡门青》、《荒原狼》、《东方之旅》、《玻璃球游戏》等。

美国队最受欢迎的同样是诺贝尔文学奖的获得者，卡勒德胡赛尼描述阿富汗人民生活的三部曲《追风筝的人》、《灿烂千阳》、《群山回唱》保持了一贯的高水准，并且体现出他日益高超的叙事技巧，尤其是第三部《群山回唱》，描述了几个家庭三代人的悲欢离合，地点跨越美国、阿富汗和法国，叙事宏大，线索巧妙，可以看出作者很有野心。凭借处女作《追风筝的人》一举获得诺贝尔奖的卡勒德胡赛尼并没有止步不前，反而在之后的两部作品中持续发力，令人惊喜。

5. 技术难点

5.1 提高爬取速度

用 python 自带的 threading 库，采用多线程爬虫提高速度。

5.2 反反爬虫

Request 一定要传 header 参数，从我自己的经验来看，每请求一些网页就 time.sleep()几秒并不能逃脱豆瓣的反爬。爬一些所谓免费高匿的代理作为参数传入 request（）函数，豆瓣还是能识别出你真正的 IP，并且封掉你的 IP，解决方案是 PPTP 动态 IP 技术，可以让你的电脑自动的每隔几十秒更换一次 IP，这样你的本机 IP 真的会显示成全国各地，而且用这个方法，除了传 header 参数以

外，并不需要别的反反爬虫操作，暴力爬就是了，反正打一枪换一个地方。当然我还是在程序了设置了每循环一次就 sleep 一下，毕竟不能太过分。在从一个 IP 切换到下一个 IP 之间会有几秒钟我真正的本机 IP 暴露出来的时段，由于我的 IP 已经被豆瓣封了，所以这个时段本该被爬取的数据都没有爬到，这也是为什么预计爬取 800 本书的代码运行起来平均只能爬到 600 本书的原因。

5.3 出错处理

爬虫基本上是编码测试能用了之后，run 一下就可以干别的去了，让程序自己慢慢爬数据，如果爬到一半突然一个你之前没有预料到的错误导致你程序报错，之前爬的数据也白费了，非常浪费时间。这种没有预料到的错误包括但不限于，教学区突然断网，网站返回一个“假网页”，网站拒绝服务，正则表达式读取的串和你预期的格式不一样.....所以一定要在所有可能出错的地方加上 try 和 except 语句，保证即是出错你的程序还能继续运行。

6. 参考文献

[1] https://developers.douban.com/wiki/?title=api_v2
[2] http://www.tubangzhu.com
[3] https://book.douban.com/review/8923354/
[4] https://book.douban.com/review/6145992/