豆瓣电影数据分析

一、实现步骤：

1. 数据采集：Python
2. 数据存储：Python + 数据库
3. 数据清洗：Python
4. 数据分析：Python + R
5. 数据可视化：R + Web

二、编程语言与环境

1. C++/Java
2. Python
3. R
4. Web:PHP/HTML/CSS/Javascript
5. 环境：Pycharm/Rstudio/Sublime/WAMP

三、实验结果

1. 一次性爬取豆瓣所有电影的概要信息：

使用 urllib2.Request 函数获取 url 返回值，使用 urllib2.urlopen 函数获取 response,并使用 json.loads 函数将返回值获取为 tag 形式，按格式输出如下：

```
id;title;url;cover;rate
25746375;我是路人甲;http://movie.douban.com/subject/25746375/;http://img3.douban.com/view/movie_
5446197;铁拳;http://movie.douban.com/subject/5446197/;http://img3.douban.com/view/movie_poster_
25885212;我们梦中见;http://movie.douban.com/subject/25885212/;http://img4.douban.com/view/movie_
25728581;少年透明人;http://movie.douban.com/subject/25728581/;http://img4.douban.com/view/movie_
5156558;撒迦利亚;http://movie.douban.com/subject/5156558/;http://img3.douban.com/view/movie_pos
24325815;非我;http://movie.douban.com/subject/24325815/;http://img3.douban.com/view/movie_poste
25922902;唇上之歌;http://movie.douban.com/subject/25922902/;http://img4.douban.com/view/movie_p
25738406;如晴天，似雨天;http://movie.douban.com/subject/25738406/;http://img4.douban.com/view/mov
25767747;故事的故事;http://movie.douban.com/subject/25767747/;http://img3.douban.com/view/movie_
25823840;奸臣;http://movie.douban.com/subject/25823840/;http://img3.douban.com/view/movie_poste
10533913;头脑特工队;http://movie.douban.com/subject/10533913/;http://img4.douban.com/view/movie_
25814707;小森林 冬春篇;http://movie.douban.com/subject/25814707/;http://img4.douban.com/view/mov
10773239;小男孩;http://movie.douban.com/subject/10773239/;http://img4.douban.com/view/movie_pos
25870236;可爱的你;http://movie.douban.com/subject/25870236/;http://img4.douban.com/view/movie_p
26147706;花与爱丽丝杀人事件;http://movie.douban.com/subject/26147706/;http://img3.douban.com/view/
25761178;百元之恋;http://movie.douban.com/subject/25761178/;http://img4.douban.com/view/movie_p
26219652;少年班;http://movie.douban.com/subject/26219652/;http://img3.douban.com/view/movie_pos
25786077;末日崩塌;http://movie.douban.com/subject/25786077/;http://img3.douban.com/view/movie_p
25958787;深夜食堂 电影版;http://movie.douban.com/subject/25958787/;http://img3.douban.com/view/m
3608742;冲出康普顿;http://movie.douban.com/subject/3608742/;http://img3.douban.com/view/photo/ph
11541282;魔力麦克2;http://movie.douban.com/subject/11541282/;http://img3.douban.com/view/movie_
25837175;谜城;http://movie.douban.com/subject/25837175/;http://img3.douban.com/view/movie_poste
10463953;模仿游戏;http://movie.douban.com/subject/10463953/;http://img3.douban.com/view/movie_p
24397586;小羊肖恩;http://movie.douban.com/subject/24397586/;http://img3.douban.com/view/movie_p
6866928;进击的巨人真人版：前篇;http://movie.douban.com/subject/6866928/;http://img3.douban.com/vie
```

2. 根据电影 ID 获取详细信息：

首先使用 urllib2.Request 函数返回 url 的 request 值，使用 urlopen 函数返回 response 值，使用 read 函数读取 html 形式，并使用 BeautifulSoup 库转换为可提取信息的 html 文件形式；然后使用 html.select 函数，html.get_text 等函数根据豆瓣 ID 提取字段；最后使用 html.find_all 函数将电影简介写入 description 变量，并写入数据，输出如下：

```
id^title^url^cover^rate^director^composer^actor^category^district^language^showtime^length^oth
25746375^我是路人甲^http://movie.douban.com/subject/25746375/^http://img3.douban.com/view/movie_
5446197^铁拳^http://movie.douban.com/subject/5446197/^http://img3.douban.com/view/movie_poster_
25885212^我们梦中见^http://movie.douban.com/subject/25885212/^http://img4.douban.com/view/movie_
25728581^少年透明人^http://movie.douban.com/subject/25728581/^http://img4.douban.com/view/movie_
5156558^撒迦利亚^http://movie.douban.com/subject/5156558/^http://img3.douban.com/view/movie_pos
24325815^非我^http://movie.douban.com/subject/24325815/^http://img3.douban.com/view/movie_poste
25922902^唇上之歌^http://movie.douban.com/subject/25922902/^http://img4.douban.com/view/movie_p
25738406^如晴天，似雨天^http://movie.douban.com/subject/25738406/^http://img4.douban.com/view/mov
25767747^故事的故事^http://movie.douban.com/subject/25767747/^http://img3.douban.com/view/movie_
25823840^奸臣^http://movie.douban.com/subject/25823840/^http://img3.douban.com/view/movie_post
10533913^头脑特工队^http://movie.douban.com/subject/10533913/^http://img4.douban.com/view/movie_
25814707^小森林 冬春篇^http://movie.douban.com/subject/25814707/^http://img4.douban.com/view/mov
10773239^小男孩^http://movie.douban.com/subject/10773239/^http://img4.douban.com/view/movie_pos
25870236^可爱的你^http://movie.douban.com/subject/25870236/^http://img4.douban.com/view/movie_p
26147706^花与爱丽丝杀人事件^http://movie.douban.com/subject/26147706/^http://img3.douban.com/view/
25761178^百元之恋^http://movie.douban.com/subject/25761178/^http://img4.douban.com/view/movie_p
26219652^少年班^http://movie.douban.com/subject/26219652/^http://img3.douban.com/view/movie_pos
25786077^末日崩塌^http://movie.douban.com/subject/25786077/^http://img3.douban.com/view/movie_p
25958787^深夜食堂 电影版^http://movie.douban.com/subject/25958787/^http://img3.douban.com/view/mo
3608742^冲出康普顿^http://movie.douban.com/subject/3608742/^http://img3.douban.com/view/photo/ph
11541282^魔力麦克2^http://movie.douban.com/subject/11541282/^http://img3.douban.com/view/movie_
25837175^谜城^http://movie.douban.com/subject/25837175/^http://img3.douban.com/view/movie_poste
10463953^模仿游戏^http://movie.douban.com/subject/10463953/^http://img3.douban.com/view/movie_p
24397586^小羊肖恩^http://movie.douban.com/subject/24397586/^http://img3.douban.com/view/movie_p
6866928^进击的巨人真人版：前篇^http://movie.douban.com/subject/6866928/^http://img3.douban.com/vie
```

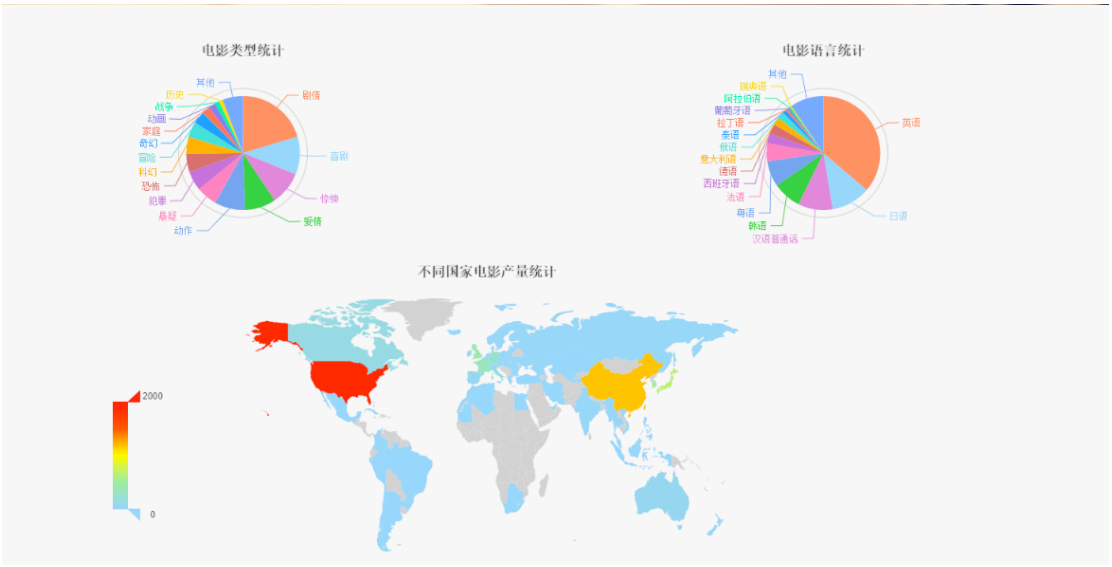
3. 对已爬取的 4589 部电影进行清洗：

定义 nameMap 数组并将英文与中文输入数组一一对应，对豆瓣电影数据进行固定格式清洗，输出如下：

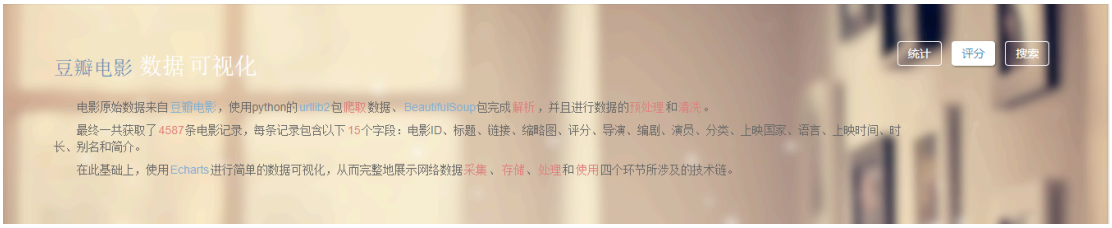
```
id^title^url^cover^rate^director^composer^actor^category^district^language^showtime^length^ot
25746375^我是路人甲^http://movie.douban.com/subject/25746375/^http://img3.douban.com/view/movie_
5446197^铁拳^http://movie.douban.com/subject/5446197/^http://img3.douban.com/view/movie_poster_
25885212^我们梦中见^http://movie.douban.com/subject/25885212/^http://img4.douban.com/view/movie_
25728581^少年透明人^http://movie.douban.com/subject/25728581/^http://img4.douban.com/view/movie_
5156558^撒迦利亚^http://movie.douban.com/subject/5156558/^http://img3.douban.com/view/movie_pos
24325815^非我^http://movie.douban.com/subject/24325815/^http://img3.douban.com/view/movie_poste
25922902^唇上之歌^http://movie.douban.com/subject/25922902/^http://img4.douban.com/view/movie_p
25738406^如晴天，似雨天^http://movie.douban.com/subject/25738406/^http://img4.douban.com/view/mo
25767747^故事的故事^http://movie.douban.com/subject/25767747/^http://img3.douban.com/view/movie_
25823840^奸臣^http://movie.douban.com/subject/25823840/^http://img3.douban.com/view/movie_post
10533913^头脑特工队^http://movie.douban.com/subject/10533913/^http://img4.douban.com/view/movie_
25814707^小森林 冬春篇^http://movie.douban.com/subject/25814707/^http://img4.douban.com/view/mov
10773239^小男孩^http://movie.douban.com/subject/10773239/^http://img4.douban.com/view/movie_pos
25870236^可爱的你^http://movie.douban.com/subject/25870236/^http://img4.douban.com/view/movie_p
26147706^花与爱丽丝杀人事件^http://movie.douban.com/subject/26147706/^http://img3.douban.com/view/
25761178^百元之恋^http://movie.douban.com/subject/25761178/^http://img4.douban.com/view/movie_p
26219652^少年班^http://movie.douban.com/subject/26219652/^http://img3.douban.com/view/movie_pos
25786077^末日崩塌^http://movie.douban.com/subject/25786077/^http://img3.douban.com/view/movie_p
25958787^深夜食堂 电影版^http://movie.douban.com/subject/25958787/^http://img3.douban.com/view/mo
3608742^冲出康普顿^http://movie.douban.com/subject/3608742/^http://img3.douban.com/view/photo/ph
11541282^魔力麦克2^http://movie.douban.com/subject/11541282/^http://img3.douban.com/view/movie_
25837175^谜城^http://movie.douban.com/subject/25837175/^http://img3.douban.com/view/movie_poste
10463953^模仿游戏^http://movie.douban.com/subject/10463953/^http://img3.douban.com/view/movie_p
24397586^小羊肖恩^http://movie.douban.com/subject/24397586/^http://img3.douban.com/view/movie_p
6866928^进击的巨人真人版：前篇^http://movie.douban.com/subject/6866928/^http://img3.douban.com/vie
```

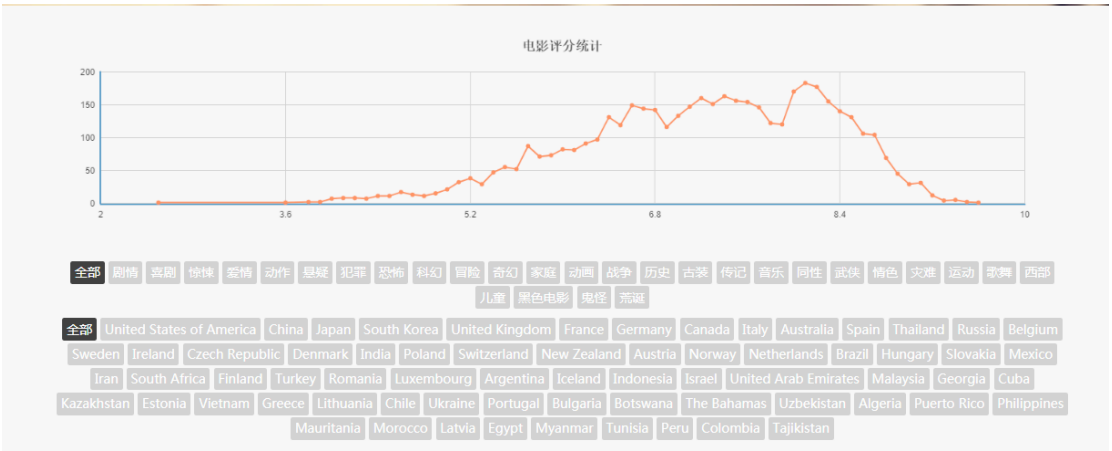
4. 将清洗的数据插入数据库：

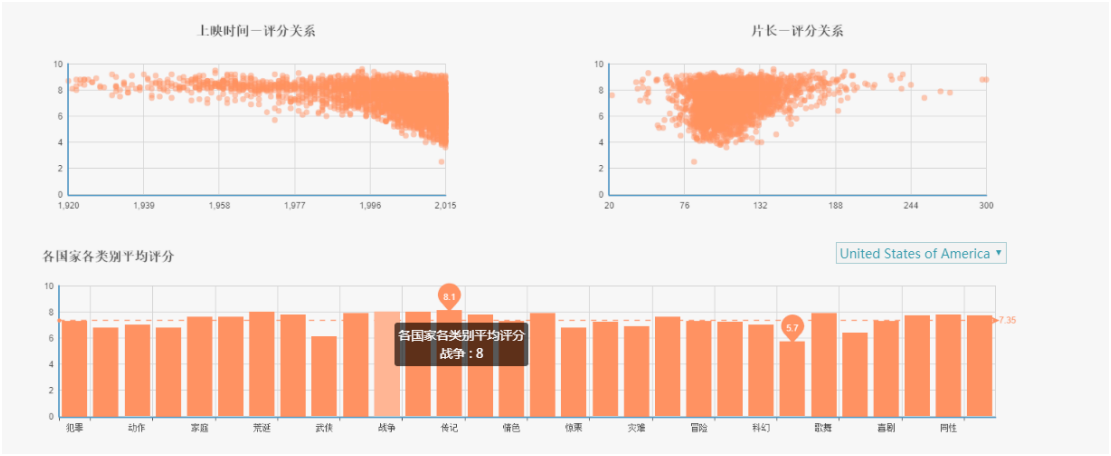
使用 mysql 存储电影信息,并对清洗后的电影数据进行基本统计，然后使用 WAMP 搭建 Web 环境：Apahce,Nginx 处理 Web 请求，MySQL 存储管理数据，PHP 后端操作。使用 HTML/CSS/Javasrcipt/Echart,实现数据前端动态可视化，最后进行结果展示：统计/评分/搜索功能

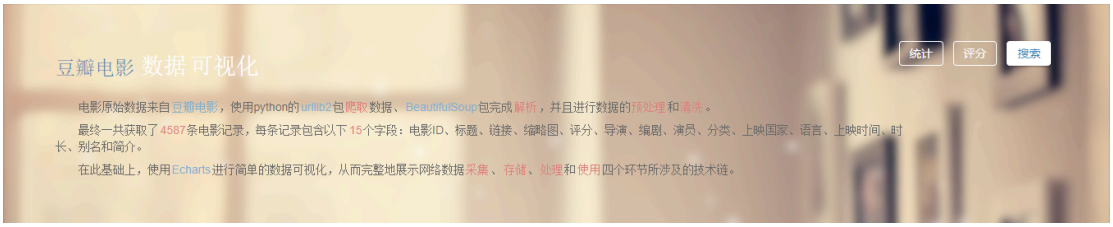

电影类型统计
其他
历史
战争
动画
家庭
奇幻
冒险
科幻
恐怖
犯罪
悬疑
动作
剧情
喜剧
惊悚
爱情
电影语言统计
其他
瑞典语
阿拉伯语
葡萄牙语
拉丁语
泰语
俄语
意大利语
德语
西班牙语
法语
粤语
韩语
汉语普通话
英语
日语
不同国家电影产量统计
2000
0


历年电影产量统计
500
400
300
200
100
0
1,915
1,936
1,957
1,978
1,999
2,020
电影时长统计
150
125
100
75
50
25
0
0
60
120
180
240
300


豆瓣电影 数据 可视化
统计
评分
搜索
电影原始数据来自豆瓣电影，使用python的urllib2包爬取数据、BeautifulSoup包完成解析，并且进行数据的预处理和清洗。
最终一共获取了4587条电影记录，每条记录包含以下15个字段：电影ID、标题、链接、缩略图、评分、导演、编剧、演员、分类、上映国家、语言、上映时间、时长、别名和简介。
在此基础上，使用Echarts进行简单的数据可视化，从而完整地展示网络数据采集、存储、处理和使用四个环节所涉及的技术链。


电影评分统计
200
150
100
50
0
2
3.6
5.2
6.8
8.4
10
全部 剧情 喜剧 惊悚 爱情 动作 悬疑 犯罪 恐怖 科幻 冒险 奇幻 家庭 动画 战争 历史 古装 传记 音乐 同性 武侠 情色 灾难 运动 歌舞 西部
儿童 黑色电影 鬼怪 荒诞
全部 United States of America China Japan South Korea United Kingdom France Germany Canada Italy Australia Spain Thailand Russia Belgium
Sweden Ireland Czech Republic Denmark India Poland Switzerland New Zealand Austria Norway Netherlands Brazil Hungary Slovakia Mexico
Iran South Africa Finland Turkey Romania Luxembourg Argentina Iceland Indonesia Israel United Arab Emirates Malaysia Georgia Cuba
Kazakhstan Estonia Vietnam Greece Lithuania Chile Ukraine Portugal Bulgaria Botswana The Bahamas Uzbekistan Algeria Puerto Rico Philippines
Mauritania Morocco Latvia Egypt Myanmar Tunisia Peru Colombia Tajikistan


上映时间－评分关系
片长－评分关系
各国家各类别平均评分
United States of America
各国家各类别平均评分
战争：8
犯罪
动作
家庭
荒诞
武侠
战争
传记
情色
惊悚
灾难
冒险
科幻
歌舞
喜剧
同性


豆瓣电影 数据可视化
统计
评分
搜索
电影原始数据来自豆瓣电影，使用python的urllib2包爬取数据、BeautifulSoup包完成解析，并且进行数据的预处理和清洗。
最终一共获取了4587条电影记录，每条记录包含以下15个字段：电影ID、标题、链接、缩略图、评分、导演、编剧、演员、分类、上映国家、语言、上映时间、时长、别名和简介。
在此基础上，使用Echarts进行简单的数据可视化，从而完整地展示网络数据采集、存储、处理和使用四个环节所涉及的技术链。


肖申克的救赎
共 1 条标题含 肖申克的救赎 的电影
肖申克的救赎 1994年 9.6分
导演 弗兰克·德拉邦特
编剧 弗兰克·德拉邦特 斯蒂芬·金
主演 蒂姆·罗宾斯 摩根·弗里曼 鲍勃·冈顿 威廉姆·赛德勒 克兰西·布朗 吉尔·贝罗斯 马克·罗斯顿 詹姆斯·惠特摩 杰弗里·德曼 拉里·布兰登伯格 尼尔·吉恩托利 布赖恩·利比 大卫·普罗瓦尔 约瑟夫·劳格诺 祖德·塞克利拉
类型 剧情 犯罪
产地 美国
语言 英语
片长 142分钟
又名 月黑高飞(港) 刺激1995(台) 地狱诺言 铁窗岁月 消香克的救赎
剧情简介
20世纪40年代末，小有成就的青年银行家安迪（蒂姆·罗宾斯 Tim Robbins 饰）因涉嫌杀害妻子及她的情人而锒铛入狱。在这座名为肖申克的监狱内，希望似乎虚无缥缈，终身监禁的惩罚无疑注定了安迪接下来灰暗绝望的人生。未过多久，安迪尝试接近囚犯中颇有声望的瑞德（摩根·弗里曼 Morgan Freeman 饰），请求对方帮自己搞来小锤子。以此为契机，二人逐渐熟稔，安迪也仿佛在鱼龙混杂、罪恶横生、黑白混淆的牢狱中找到属于自己的求生之道。他利用自身的专业知识，帮助监狱管理层逃税、洗黑钱，同时凭借与瑞德的交往在犯人中间也渐渐受到礼遇。表面看来，他已如瑞德那样对那堵高墙从憎恨转变为处之泰然，但是对自由的渴望仍促使他朝着心中的希望和目标前进。而关于其罪行的真相，似乎更使这一切朝前推进了一步……
本片根据著名作家斯蒂芬·金（Stephen Edwin King）的...