

实验三、网络爬虫 (25")

- 1、爬取所有豆瓣电影评分 Top250 的电影的信息 (10")
 - a) 正文链接
 - b) 英文名 (如有), 中文名
 - c) 其他信息

从首页进行信息爬取, 包括以下几部分:



其中, 使用 `parse1` 的 `xpath` 以及 `re` 正则表达式来提取, 遇到一些问题需要异常处理——一句话短评以及电影英文名称处有些电影是不具备的!

爬取后得到的 `excel` 表格如下:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
227	阿飞正传	阿飛正傳	王家卫	张国荣	1990	中国香港	犯罪 剧情	8.5	390192人	王家卫	https://movie.douban.com/subject/1305690/				
228	谍影重重2	The Bourne	保罗·格林	马特·达蒙	2004	美国 德国	动作 悬疑	8.7	281380人	谁说王	https://movie.douban.com/subject/1308767/				
229	地球上的星星	Taare Zam	阿米尔·汗	达席尔·萨	2007	印度	剧情 儿童	8.9	164161人	天使保护	https://movie.douban.com/subject/2363506/				
230	完美陌生人	Perfetti scc	保罗·格诺	马可·贾利	2016	意大利	剧情 喜剧	8.5	432249人	来啊，互相	https://movie.douban.com/subject/26614893/				
231	彗星来的那	Coherence	詹姆斯·沃	艾米丽·芭	2013	美国 英国	科幻 悬疑	8.5	403048人	小成本大	https://movie.douban.com/subject/25807345/				
232	战争之王	Lord of W	安德鲁·尼	尼古拉斯·	2005	美国 法国	剧情 犯罪	8.7	279563人	做一颗让	https://movie.douban.com/subject/1419936/				
233	香水	Perfume: T	汤姆·提克	本·卫肖	2006	德国 法国	剧情 犯罪	8.5	448272人	一个单凭	https://movie.douban.com/subject/1706622/				
234	谍影重重	The Bourne	道格拉斯·	马特·达蒙	2002	美国 德国	动作 悬疑	8.6	336048人	哗啦啦啦	https://movie.douban.com/subject/1304102/				
235	朗读者	The Reade	史蒂芬·戴	凯特·温丝	2008	美国 德国	剧情 爱情	8.6	387055人	当爱情跨	https://movie.douban.com/subject/2213597/				
236	东京物语	東京物語	小津安二郎	笠智众 Ch	1953	日本	剧情 家庭	9.2	102089人	东京那么	https://movie.douban.com/subject/1291568/				
237	猜火车	Trainspot	丹尼·博伊	伊万·麦克	1996	英国	剧情 犯罪	8.5	353591人	不可猜的	https://movie.douban.com/subject/1292528/				
238	再次出发之	Begin Aga	约翰·卡尼	凯拉·奈特	2013	美国	喜剧 爱情	8.6	332857人	爱我就给	https://movie.douban.com/subject/6874403/				
239	千钧一发	Gattaca	安德鲁·尼	伊桑·霍克	1997	美国	剧情 科幻	8.8	199923人	一部能引	https://movie.douban.com/subject/1300117/				
240	浪潮	Die Welle	丹尼尔·甘	伊尔·霍夫	2008	德国	剧情 惊悚	8.7	217580人	世界孤独	https://movie.douban.com/subject/2297265/				
241	驴得水	无对应英	周申 Shen	任素汐 Su	2016	中国大陆	剧情 喜剧	8.3	724650人	过去的如	https://movie.douban.com/subject/25921812/				
242	黑帝帝国2	The Matrix	安德尼·沃	基努·里维	2003	美国 澳大	动作 科幻	8.6	291263人	一个精彩	https://movie.douban.com/subject/1304141/				
243	聚焦	Spotlight	托马斯·麦	马克·鲁弗	2015	美国	剧情 传记	8.8	224648人	新闻人的	https://movie.douban.com/subject/25954475/				
244	东京教父	東京ゴッ	今敏 Satoshi	江守彻 To	2003	日本	剧情 喜剧	9	132039人	你要相信	https://movie.douban.com/subject/1301077/				
245	我爱你你	그대를 사	秋敏熙 Ch	宋在河 Ja	2011	韩国	剧情 爱情	9.1	124365人	宝莱坞的	https://movie.douban.com/subject/5908478/				
246	小萝莉的虎	Bairangi B	卡比尔·汗	萨尔曼·汗	2015	印度	剧情 喜剧	8.4	390905人	诺兰的牛	https://movie.douban.com/subject/26393561/				
247	追随	Following	克里斯托	杰里米·西	1998	英国	犯罪 悬疑	8.9	145147人	还原真实	https://movie.douban.com/subject/1397546/				
248	无间道2	無間道II	刘伟强 / 麦	陈冠希 / 梁	2003	中国香港	动作 犯罪	8.6	315071人	只有有信	https://movie.douban.com/subject/1307106/				
249	黑鹰坠落	Black Haw	雷德利·斯	乔什·哈奈	2001	美国 英国	动作 历史	8.7	231248人	无	https://movie.douban.com/subject/1291824/				
250	一次别离	جدایی نادر از	阿斯哈·法	佩曼·莫阿	2011	伊朗 法国	剧情 家庭	8.7	210469人	无	https://movie.douban.com/subject/5964718/				
251	网络追踪	Searching	阿尼什·查	约翰·达	2018	美国 俄罗	剧情 犯罪	8.6	415106人	无	https://movie.douban.com/subject/27615441/				
252															

肖申克的救赎 (豆瓣)

× +

← → ↺ ⌂ ↶ ☆

https://movie.douban.com/subject/1292052/

想看

看过

评价: ☆☆☆☆☆

写短评

写影评

分享到

推荐

肖申克的救赎的剧情简介 ·····

一场谋杀案使银行家安迪（蒂姆·罗宾斯 Tim Robbins 饰）蒙冤入狱，谋杀妻子及其情人的指控将囚禁他终生。在肖申克监狱的首次现身就让监狱“大哥”瑞德（摩根·弗里曼 Morgan Freeman 饰）对他另眼相看。瑞德帮助他搞到一把石锤和一幅女明星海报，两人渐成患难之交。很快，安迪在监狱里大显其才，担当监狱图书管理员，并利用自己的金融知识帮助监狱官避税，引起了典狱长的注意，被招致麾下帮助典狱长洗黑钱。偶然一次，他得知一名新入狱的小偷能够作证帮他洗脱谋杀罪。燃起一丝希望的安迪找到了典狱长，希望他能帮自己翻案。阴险伪善的狱长假装答应安迪，背后却派人杀死小偷，让他唯一能合法出狱的希望泯灭。沮丧的安迪并没有绝望，在一个电闪雷鸣的风雨夜，一场暗藏几十年的越狱计划让他自我救赎，重获自由！老朋友瑞德在他的鼓舞和帮助下，也勇敢地奔向自由。

本片获得1995年奥... (展开全部) @豆瓣

肖申克的救赎的演职员 ····· (全部 35)



弗兰克·德拉...
导演



蒂姆·罗宾斯
饰 安迪·杜佛...



摩根·弗里曼
饰 艾利斯·波...



鲍勃·冈顿
饰 监狱长山姆...



威廉姆·赛德勒
饰 海伍德 He...



克兰西·布朗
饰 上尉哈德利...

肖申克的救赎的视频和图片 ····· (预告片2 | 视频评论3 · 添加 | 图片723 · 添加)







单线程爬的过程中加了中断处理：

- 3、加分项：
 - a) 不限于豆瓣的简介，影评
 - b) 是否分析了演员与电影类型的关联程度
 - c) 是否分析了演员与演员的关系？
 - d) 是否对简介和影评进行词云分析？
 - e) 其他信息

其中a在第一问解决了，现在绘制词云图：

```

38     wCloud.to_file(path + '/' + '{}.jpg'.format(name))
39
40     # 需要将生成的词云图片放到对应的子目录下
41     path = '爬取的数据'
42     f = open(path + '/' + "豆瓣250电影信息.csv", mode='r', encoding='utf-8')
43     csv_reader = csv.reader(f)
44     rows = [i for i in csv_reader]
45     # 去除第一行后，每行的最后一列就是评论内容
46     comments = [row[-1] for row in rows[1:]]
47     instruction = [row[-2] for row in rows[1:]]
48     make_cloud(path, comments, "短评")
49     make_cloud(path, instruction, "简介")
50

```

Run: task3_绘制词云图 ×

```

D:\python\python.exe D:/系统默认/桌面/机器学习/codes/task3_绘制词云图.py
Building prefix dict from the default dictionary ...
Dumping model to file cache D:\系统缓存\jieba.cache
Loading model cost 1.425 seconds.
Prefix dict has been built succesfully.

进程已结束，退出代码 0

```

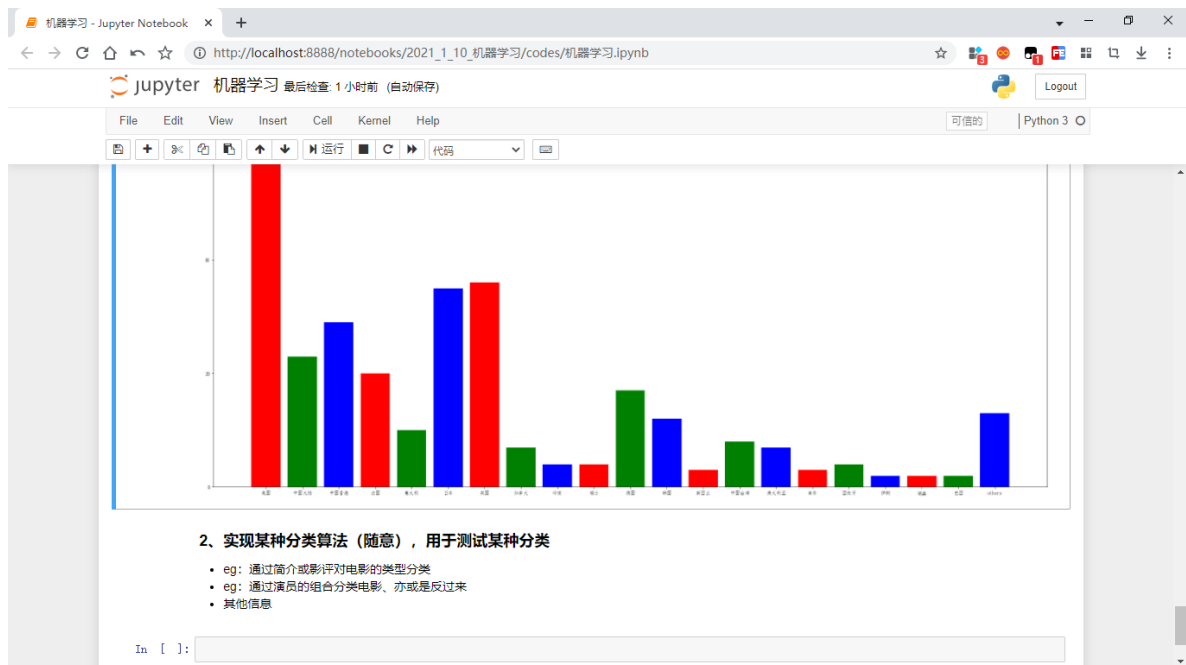
得到的词云效果图如下：



实验四、机器学习 (25")

- 1、实现对获取的电影数据的统计分析 (10")
 - a) 可以考虑类型、语言、地区或演员的特征维度
 - b) 可以考虑对简介、影评进行语义分析出来的结果进行统计
 - c) 绘制相关图形

首先选择两列数据：`首映地区和类型`，进行数据处理后，使用 `matplotlib` 绘制柱状图！



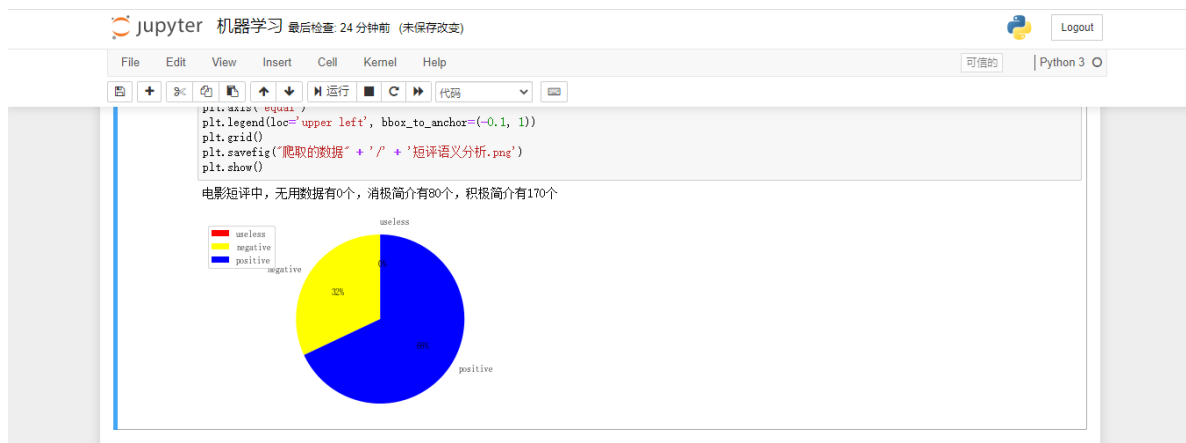
安装 SnowNLP 库，以方便中文句子打分：

```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [版本 10.0.18363.1256]
(c) 2019 Microsoft Corporation. 保留所有权利。

C:\Users\HP>pip install snownlp
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting snownlp
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/3d/b3/3756768662100d3bce62d3b0f2adec18ab4b9ff2b61abd7a61c39343cld/snownlp-0.12.3.tar.gz (37.6MB)
    | 37.6MB 6.8MB/s
Building wheels for collected packages: snownlp
  Building wheel for snownlp (setup.py) ... done
  Stored in directory: C:\Users\HP\AppData\Local\pip\Cache\wheels\05\9d\33\3cb257f3b8fef0093cfe5a2bac10a8ab7e9ce18f8538f7e5
Successfully built snownlp
Installing collected packages: snownlp
Successfully installed snownlp-0.12.3

C:\Users\HP>
```

使用该库对短评和简介数据进行打分：



使用随机森林根据上映时间、评分以及评价人数和最后一句话短评的正负性得分进行训练和验证模型

机器学习 - Jupyter Notebook x +

http://localhost:8888/notebooks/2021_1_10_机器学习/codes/机器学习.ipynb

jupyter 机器学习 最后检查 37 分钟前 (未保存改变)

File Edit View Insert Cell Kernel Help 可信的 Python 3

```
# 训练: 测试=7: 3
X_train, X_test, y_train, y_test=train_test_split(X, y, test_size=0.3, random_state=1)
rf = RandomForestClassifier(
    criterion='entropy',
    n_estimators=30,
    max_depth=5,
    min_samples_split=10, # 定义至少多少个样本的情况下才继续分叉
    min_samples_leaf=4,
    min_weight_fraction_leaf=0.05 # 定义叶子节点最少需要包含多少个样本(使用百分比表达), 防止过拟合
)

# 训练模型
rf.fit(X_train, y_train)
# 做预测
y_pred = rf.predict(X_test)
# 模型的准确率
print('i=number_of_trees=:4, accuracy=', metrics.accuracy_score(y_test, y_pred))
# 保存model
joblib.dump(rf, 'rf.pkl')

i=number_of_trees=: 4 , accuracy= 0.958904109589041

Out[103]: ['rf.pkl']

In [ ]:
```