

python 期中大作业（豆瓣电影数据集）报告

徐艺峰 2100013125

05/03/2023

目录

| | |
|--|-----------|
| 1 Task1: 数据预处理与特征向量化 | 2 |
| 1.1 标签 GENRES、类型 TAGS 数据的分词、清理、向量化及降维 | 2 |
| 1.2 简介 STORYLINE 字段的清洗、分词和 word2vec 向量化 | 3 |
| 2 Task2: 特征融合和降维 | 4 |
| 2.1 归一化数值特征、拼接特征向量 | 4 |
| 2.2 PCA 降维 | 5 |
| 3 Task3: K-means 聚类 | 5 |
| 3.1 肘部法确定最佳簇数量 | 5 |
| 3.2 K-means 聚类 | 6 |
| 3.3 分析聚类结果，进行可视化展示 | 6 |
| 4 Task4: 为导演和演员生成 Embedding，进行无监督分类，并分析 2-3 个属于不同类别导演和演员的特点 | 11 |
| 4.1 对导演和演员生成特征向量、再分别进行层次聚类 | 11 |
| 4.2 导演和演员聚类结果的简单分析 | 12 |
| 4.3 可视化分析 2 个不同类型导演的特点 | 14 |
| 4.4 可视化分析 2 个不同类型演员的特点 | 15 |

1 Task1：数据预处理与特征向量化

1.1 标签 GENRES、类型 TAGS 数据的分词、清理、向量化及降维

查看原始数据集发现，标签和类型字段都类似“剧情/爱情”这种格式，因此只需要以“/”符号为界分词即可。在此处发现，有一部分标签和类型字段的词是完全一致的，但有简中、繁中和英文三个版本（比如“喜剧”“喜劇”和“Comedy”），在分词前进行了统一，全部替换为简中表示的标签。

```
# 类型和标签数据中有一些标签有简、繁、英三种表示，在此统一将繁体字和英文同义替换成对应的简体字，同时处理掉缺失数据
dict_mapping = {"": ["NaN"], "动作": ["動作", "Action"], "喜剧": ["喜劇", "Comedy"], "爱情": ["愛情", "Romance"], "儿童": ["兒童", "Kids"], \
    "纪录片": ["紀錄片", "Documentary"], "音乐": ["音樂", "Music"], "动画": ["動畫", "Animation"], "惊悚": ["驚悚", "Thriller"], \
    "剧情": ["劇情", "Adult", "Drama"], "真人秀": ["Reality-TV"], "国": ["國"], "电": ["電"], "湾": ["灣"], "韩": ["韓"]}

# 类型和标签数据按“/”符号为界进行分词即可
def clean_and_cut(text):
    for key, value in dict_mapping.items():
        # 将需要替换的单词用“|”连接，组成正则表达式
        pattern = "|".join(value)
        # 使用正则表达式将单词替换为对应的value
        text = re.sub(pattern, key, str(text))
    return re.sub('/', ' ', text)
```

图 1: 分词和统一标签

```
GENRES字段分词后的结果:
0      剧情 爱情
1      动作 爱情
2      剧情
3      爱情
4      剧情 历史
...
38184    惊悚 恐怖
38185    喜剧 爱情
38186    剧情 动作 犯罪
38187    悬疑 惊悚 恐怖 犯罪
38188    剧情
Name: GENRES, Length: 38189, dtype: object

TAGS字段分词后的结果:
0      甘肃 临夏 伊斯兰 中国 2016 中国大陆 烂片 宣传伊斯兰教的电影
1      穿越 华语
2
3      小波 王小波 爱情 小说改编 文学改编 剧情 中国 2017
4
...
38184    血腥 恐怖 美国 惊悚 暴力 德洲电锯杀人狂前传 美国电影 2006
38185    爱情 美国 美国电影 麻雀变王妃 喜剧 浪漫喜剧 童话 电影
38186    韩国 动作 黑帮 暴力 韩国电影 犯罪 2006 柳承完
38187    惊悚 美国 悬疑 恐怖 我看过的英语电影 我看过的电影 我看过的恐怖电影 美国电影
38188    法国 法国电影 冒险 旅行探索 动物 俄罗斯 2006 Serko
Name: TAGS, Length: 38189, dtype: object
```

图 2: GENRES 和 TAGS 字段的分词结果

随后，对 GENRES 字段和 TAGS 字段分别使用 `TfidfVectorizer().fit_transform` 进行向量化，生成 `tfidf` 矩阵。在上一步统一了简繁体之后，GENRES 字段只剩下 35 个词，而 TAGS 字段的词数仍然较多，所以此处只保留了出现频率最高的前 500 个词作为词表，将 `tdidif` 向量限制在 500 维。

```
# 对类型 (GENRES) 字段, 将全部的35个词作为词表
tfidf_GENRES = TfidfVectorizer()
weight_GENRES = tfidf_GENRES.fit_transform(data['GENRES']).toarray()
word_GENRES = tfidf_GENRES.get_feature_names()

# 对标签 (TAGS) 字段, 由于分词后不同的词太多, 只取出现频率最高的前500个词作为词表
tfidf_TAGS = TfidfVectorizer(max_features = 500)
weight_TAGS = tfidf_TAGS.fit_transform(data['TAGS']).toarray()
word_TAGS = tfidf_TAGS.get_feature_names()
```

图 3: tfidf

GENRES字段tf-idf矩阵的大小:
(38189, 35)

TAGS字段tf-idf矩阵的大小:
(38189, 500)

GENRES字段的词表 (35词) :

```
[('news', 0), ('传记', 1), ('儿童', 2), ('冒险', 3), ('剧情', 4), ('动作', 5), ('动画', 6), ('历史', 7), ('古装', 8), ('同性', 9), ('喜剧', 10), ('奇幻', 11), ('家庭', 12), ('恐怖', 13), ('悬疑', 14), ('情色', 15), ('惊悚', 16), ('戏曲', 17), ('战争', 18), ('歌舞', 19), ('武侠', 20), ('灾难', 21), ('爱情', 22), ('犯罪', 23), ('真人秀', 24), ('短片', 25), ('科幻', 26), ('纪录片', 27), ('脱口秀', 28), ('舞台艺术', 29), ('西部', 30), ('运动', 31), ('音乐', 32), ('鬼怪', 33), ('黑色电影', 34)]
```

TAGS字段的词表 (500词) :

```
[('1080p', 0), ('11', 1), ('12', 2), ('123', 3), ('2000s', 4), ('2006', 5), ('2007', 6), ('2008', 7), ('2009', 8), ('2010', 9), ('2010s', 10), ('2011', 11), ('2012', 12), ('2013', 13), ('2014', 14), ('2015', 15), ('2016', 16), ('2017', 17), ('2018', 18), ('2019', 19), ('2020', 20), ('3d', 21), ('3m稀影基地', 22), ('3m稀影基地论坛', 23), ('anime', 24), ('asfun', 25), ('bbc', 26), ('biff', 27), ('bl', 28), ('brigadier general', 29), ('b级片', 30), ('cg', 31), ('cctv6', 32), ('china', 33), ('cult', 34), ('снг', 35), ('davidtenant', 36), ('dc', 37), ('disney', 38), ('doctorwho', 39), ('doxtv', 40), ('dvd', 41), ('espana', 42), ('france', 43), ('français', 44), ('gay', 45), ('hallmark', 46), ('hbo', 47), ('horror', 48), ('ireland', 49), ('isfvf', 50), ('itv', 51), ('les', 52), ('lgbt', 53), ('lgbtq', 54), ('live', 55), ('marvel', 56), ('movie', 57), ('nbc', 58), ('netflix', 59), ('ntlive', 60), ('oad', 61), ('opera', 62), ('ova', 63), ('production', 64), ('r15', 65), ('rapid', 66), ('short', 67), ('shortlist', 68), ('siff', 69), ('sp', 70), ('stand', 71), ('sweden', 72), ('tba', 73), ('tv', 74), ('ugc', 75), ('uk', 76), ('up', 77), ('us', 78), ('usa', 79), ('wwe', 80), ('p и', 81), ('с с с п ф', 82), ('アニメ', 83), ('一流', 84), ('上海国际电影节', 85), ('上海电影节', 86), ('不看', 87), ('丧尸', 88), ('丧尸片', 89), ('中国', 90), ('中国动画', 91), ('中国大陆', 92), ('中国电影', 93), ('丹麦', 94), ('丹麦电影', 95), ('旋律', 96), ('乡村', 97), ('二战', 98), ('二次元', 99), ('京阿尼', 100), ('亲情', 101), ('人性', 102), ('人生', 103), ('以色列', 104), ('饭面ライダー', 105), ('伊朗', 106), ('伊朗电影', 107), ('传记', 108), ('伦理', 109), ('伦理片', 110), ('伪纪录片', 111), ('低成本', 112), ('体育', 113), ('侦探', 114), ('俄罗斯', 115), ('俄罗斯电影', 116), ('信仰', 117), ('假面骑士', 118), ('僵尸', 119), ('儿童', 120), ('克里斯蒂', 121), ('公视人生剧展', 122), ('公路', 123), ('内地', 124), ('内地电影', 125), ('冒险', 126), ('军事', 127), ('农村', 128), ('冷门', 129), ('剧场版', 130), ('剧情', 131), ('功夫', 132), ('加拿大', 133), ('加拿大电影', 134), ('动作', 135), ('动漫', 136), ('动物', 137), ('动画', 138), ('动画片', 139), ('动画电影', 140), ('动画短片', 141), ('励志', 142), ('匈牙利', 143), ('北欧', 144), ('北欧五国', 145), ('华语', 146), ('华语电影', 147), ('南印电影', 148), ('卡通', 149), ('印度', 150), ('印度电影', 151), ('印影', 152), ('历史', 153), ('友情', 154), ('反乌托邦', 155), ('反转', 156), ('变态', 157), ('古天乐', 158), ('古装', 159), ('台湾', 160), ('台湾电影', 161), ('史诗', 162), ('同志', 163), ('同志电影', 164), ('同性', 165), ('同性恋', 166), ('名著改编', 167), ('后宫', 168), ('吸血鬼', 169)]
```

图 4: 向量化的结果, 以及 GENRES 字段和 TAGS 字段的词表 (一部分)

随后, 使用 TruncatedSVD 对向量化后的数据进行降维, 将 GENRES 和 TAGS 对应的特征向量分别降到 20 和 30 维。

```
# 使用TruncatedSVD对向量化后的数据进行降维 (用时较久, 约2分钟, 初始化TruncatedSVD时设置random_state参数)
svd_GENRES = TruncatedSVD(n_components=20, random_state=100)
svd_RESULT_GENRES = svd_GENRES.fit_transform(weight_GENRES)
svd_TAGS = TruncatedSVD(n_components=30, random_state=100)
svd_RESULT_TAGS = svd_TAGS.fit_transform(weight_TAGS)
print(svd_RESULT_GENRES.shape)
print(svd_RESULT_TAGS.shape)
```

(38189, 20)
(38189, 30)

图 5: TruncaSVD 降维

1.2 简介 STORYLINE 字段的清洗、分词和 word2vec 向量化

对 STORYLINE 字段判断中英文, 分别进行分词 (中文使用 jieba 库)。

STORYLINE字段分词后的结果:

```

0    电影 情定 夏天 使然 讲述 临夏 新一代 青年人 发奋图强 借助 国家 一带 一路 战略 ...
1    桀骜不驯 如龙 武功 高强 一场 比赛 打成 重伤 被诊 今生 不能 再用 功夫 女友 荆兰...
2    平民 女孩 李莉 只身 初入 曼哈顿 求学 历经 迷失 困惑 之后 凭借 努力 善良 收获 ...
3    王小波 经典 中篇小说 绿毛 水怪 改编 电影 绿毛 水怪 王小波 早期 手稿 作品 天马行...
4    1932 上海 虹口 爆炸案 韩国 国父 金九在 褚辅成 朱爱宝 这些 普通群众 帮助 逃到...
      ...
38184  德州 屠宰场 肥胖 女工 正在 案板 切肉 突然 感到 腹中 剧痛 原来 怀有孕 羊水 破产...
38185  皇家 婚禮 2003 浪漫 喜劇片 麻雀 王妃 繢集 集中 婦愛上 愛得華 不知 如假 包換...
38186  首爾 重案 刑警 郑泰秀 郑斗洪饰 接到 好友 吴汪才 死讯 于是 赶回 十年 未曾 亲近 ...
38187  a pair of siamese twins are separated and one ...
38188  1889 为了 修建 欧亚 铁路 沙皇 属下 四处 征用 马匹 期间 阿穆尔河 大公 为了 ...
Name: STORYLINE, Length: 38189, dtype: object

```

图 6: STORYLINE 字段的分词结果

随后使用 Word2Vec 对分词后的“简介”进行向量化。Word2Vec 模型参数中，将 vector_size 设置为 50 维，词的最小出现次数 min_count 设置为 10 次。得到每个词的特征向量以后，再使用文本中出现的所有词的特征向量相加求平均值的方式，由词的特征向量得到电影 STORYLINE 字段的特征向量。

```

# 词向量累加得到STORYLINE文本的特征向量
vectors = []
for sentence in sentences:
    vector = np.zeros(50)
    count = 0
    for word in sentence:
        try:
            vector += model.wv[word]
            count += 1
        except:
            continue
        if count != 0:
            vectors.append(vector / count)
        else:
            vectors.append(vector)

print("STORYLINE特征向量矩阵的大小: ")
print(np.array(vectors).shape)

```

STORYLINE分词后符合要求的总词数:
18069
STORYLINE特征向量矩阵的大小:
(38189, 50)

图 7: 由词的特征向量累加成 STORYLINE 的特征向量

2 Task2：特征融合和降维

2.1 归一化数值特征、拼接特征向量

将 GENRES 和 TAGS 的 tfidf、word2vec 后的 STORYLINE 向量、电影年份 YEAR、豆瓣评分 DOUBAN_SCORE 横向拼接起来，归一化之后作为电影的 embedding 向量。

对特征向量矩阵的每一列（即向量的每一维）做 min-max 归一化。

```

# 归一化数值特征
def min_max_normalize(x):
    x_min = np.min(x, axis=0, keepdims=True)
    x_max = np.max(x, axis=0, keepdims=True)
    x_normalized = (x - x_min) / (x_max - x_min)
    return x_normalized

#拼接 (GENRES和TAGS的tfidf、word2vec后的STORYLINE向量、年份、豆瓣评分)
embedding = np.hstack((svd_RESULT_GENRES, svd_RESULT_TAGS, np.array(vectors), np.array(data['YEAR']).reshape(-1, 1) \
                      , np.array(data['DOUBAN_SCORE']).reshape(-1, 1)))

# 对每个特征向量按列进行归一化到[0, 1]范围内
embedding_normalized = min_max_normalize(embedding)

print("embedding特征向量矩阵的大小 (20+30+50+1+1=102) : ")
print(embedding_normalized.shape)
print(embedding_normalized)

embedding特征向量矩阵的大小 (20+30+50+1+1=102) :
(38189, 102)

```

图 8: 特征向量拼接和归一化

2.2 PCA 降维

接下来对电影的特征向量进行 pca 降维，设定保留 85% 以上的主成分，输出结果发现降到了 26 维。

```

# 计算累计方差贡献率
def get_cumulative_variance_ratio(pca):
    cumulative_variance_ratio = np.cumsum(pca.explained_variance_ratio_)
    return cumulative_variance_ratio

# PCA设定保留85%以上的主成分
pca = PCA(n_components=0.85, random_state=0)
embedding_pca = pca.fit_transform(embedding_normalized)

print("降维后的embedding特征向量矩阵的大小 (降成了26维) : ")
print(embedding_pca.shape, '\n')

print("累计方差贡献率: ")
print(get_cumulative_variance_ratio(pca)[embedding_pca.shape[1] - 1])

print(embedding_pca[0])

```

降维后的embedding特征向量矩阵的大小 (降成了26维) :

(38189, 26)

图 9: PCA 降维，输出保留的主成分

3 Task3: K-means 聚类

3.1 肘部法确定最佳簇数量

借助 kmeans.inertia_ 变量，进行 cluster-inertia 可视化。

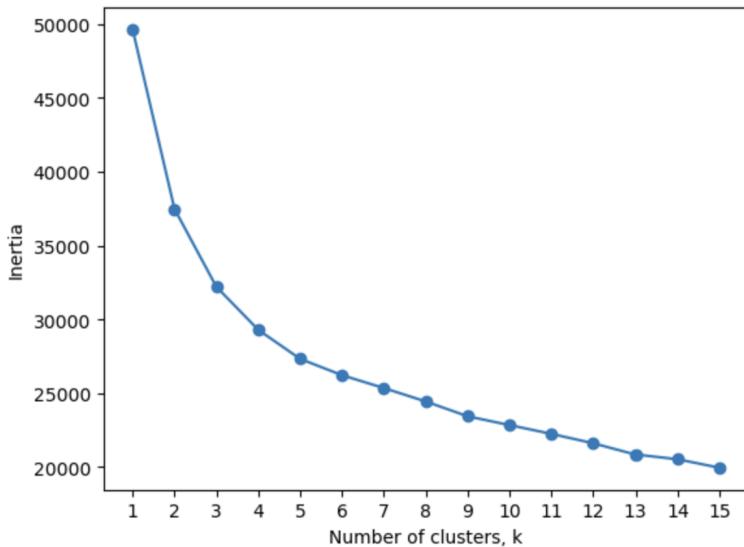


图 10: cluster-inertia 可视化

3.2 K-means 聚类

根据上图，肘部应该在 $k=4$ 或 5 的位置，但此处根据实际情况选择聚为 7 类更加合适。

```
# 使用KMeans进行聚类, 按上图k=4, 5时为肘部, 最终根据实际情况选择聚成7类
kmeans = KMeans(n_clusters=7, random_state=0)
kmeans.fit(embedding_pca)
labels = kmeans.labels_

print("聚类结果 (查看前100部电影的分类标签): ")
print(labels[:100])

聚类结果 (查看前100部电影的分类标签):
[3 3 0 3 3 3 3 3 3 3 3 3 3 1 3 3 3 5 3 0 0 3 3 2 3 2 2 3 2 5 3 3 3 5 3
3 3 3 1 5 1 3 3 3 1 3 3 3 3 3 0 2 3 1 3 3 2 1 3 3 3 3 0 3 3 2 3 1 0 3 3 3
1 3 2 5 3 1 3 3 3 2 1 3 3 3 3 3 1 3 3 3 3 6 2 2 3 1]
```

图 11: 聚类结果

3.3 分析聚类结果，进行可视化展示

对于这 7 类，我对每个簇都输出了簇的代表电影及其类型（离簇中心最近的电影）、簇中观看人数最多的前 10 部电影及其信息、簇中电影的总数、簇中电影的评分分布情况（分布柱状图、平均值、标准差）、簇中评分最高的三部电影、以及簇中电影的类型 GENRES 分布柱状图。

根据这些信息，我对每个簇的电影特点进行了解读。

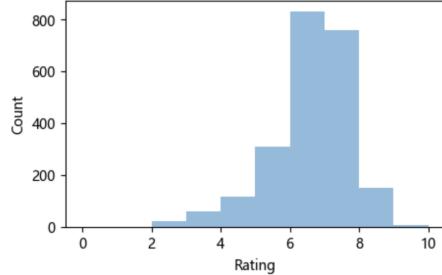
(此处统计簇内电影信息的代码较长，报告中不粘贴代码，只展示结果)

| Cluster 0 : | | | | | |
|---|----------|--------------|--------------|-----|--|
| 纯粹的剧情片，设定都是基于现实世界的，内容平实、没有架空元素，这类电影质量相对较高 | | | | | |
| 代表电影及其标签： | | | | | |
| 漂洋过海来爱你 剧情 | | | | | |
| 观看人数最多的前10部电影： | | | | | |
| NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | | |
| 19282 | 小森林 夏秋篇 | 剧情 | 231551 | 9.0 | |
| 20421 | 十二公民 | 剧情 | 208473 | 8.3 | |
| 15812 | 小森林 冬春篇 | 剧情 | 200877 | 9.0 | |
| 21485 | 百鸟朝凤 | 剧情 | 168071 | 8.1 | |
| 5004 | 过春天 | 剧情 | 139716 | 7.7 | |
| 37514 | 追风筝的人 | 剧情 | 119346 | 8.2 | |
| 29394 | 海洋天堂 | 剧情 | 112413 | 7.9 | |
| 31845 | 成长教育 | 剧情 | 105483 | 7.7 | |
| 28008 | 再见我们的幼儿园 | 剧情 | 101162 | 8.7 | |
| 34135 | 七磅 | 剧情 | 100276 | 8.1 | |

簇中电影的数量：5394

豆瓣评分分布情况：

Rating Distribution in Cluster 0



评分平均值（忽略0分电影）：6.57

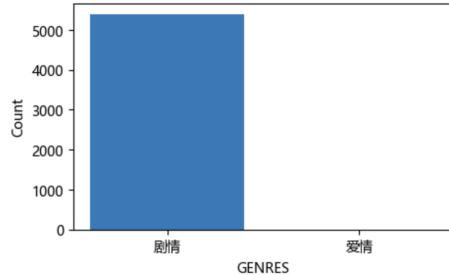
评分标准差（忽略0分电影）：1.13

Top 3 Movies:

1. 红色 (9.3)
2. 谁害怕弗吉尼亚·伍尔夫? (9.1)
3. 小森林 冬春篇 (9.0)

GENRES电影类型分布情况：

GENRES Distribution in Cluster 0

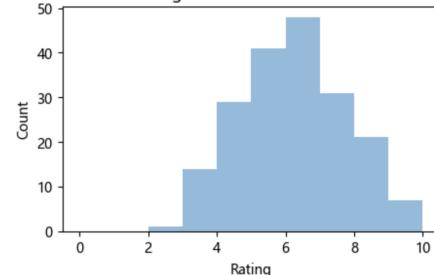


| Cluster 1 : | | | | | |
|--|-------------|----------------|--------------|-----|--|
| 类别比较杂，以剧情、喜剧、动作为主，大部分为娱乐商业片，可以认为是没什么思考量的‘爽片’ | | | | | |
| 代表电影及其标签： | | | | | |
| Paul Chowdhry: What's Happening White People? 喜剧 脱口秀 | | | | | |
| 观看人数最多的前10部电影： | | | | | |
| NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | | |
| 29744 | 狄仁杰之通天帝国 | 动作 悬疑 古装 | 199365 | 6.5 | |
| 29035 | 魔法师的学徒 | 剧情 喜剧 动作 奇幻 冒险 | 53981 | 6.1 | |
| 34921 | 东京！ | 剧情 喜剧 奇幻 | 13264 | 7.8 | |
| 31287 | 李小龙我的兄弟 | 剧情 动作 爱情 传记 | 7481 | 6.3 | |
| 34762 | 郎在远方 | 爱情 战争 | 1654 | 7.4 | |
| 23790 | 侯门之险 | 剧情 悬疑 | 1237 | 5.6 | |
| 21186 | 迷人的保姆 | 爱情 情色 | 1132 | 3.2 | |
| 13587 | 扯蛋圣诞节 | 喜剧 | 1083 | 8.1 | |
| 23539 | 西蒙·阿姆斯特朗：麻木 | 喜剧 脱口秀 | 952 | 9.3 | |
| 23805 | 无条件的爱 | 剧情 惊悚 同性 | 946 | 6.7 | |

簇中电影的数量：5251

豆瓣评分分布情况：

Rating Distribution in Cluster 1



评分平均值（忽略0分电影）：6.19

评分标准差（忽略0分电影）：1.54

Top 3 Movies:

1. 黄子华栋笃笑之金盆洗手 (9.4)
2. 杀手乐团：皇家艾伯特音乐厅演唱会 (9.4)
3. 迪兰·莫兰：脱身 (9.3)

GENRES电影类型分布情况：

GENRES Distribution in Cluster 1

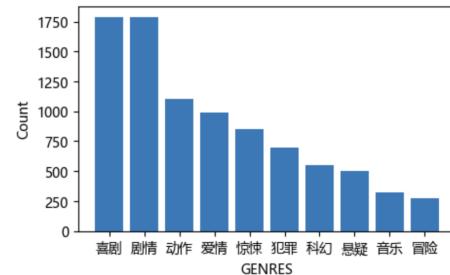


图 12: cluster 0

图 13: cluster 1

cluster 0：纯粹的剧情片，设定都是基于现实世界的，内容平实、没有架空元素，这类电影质量相对较高。

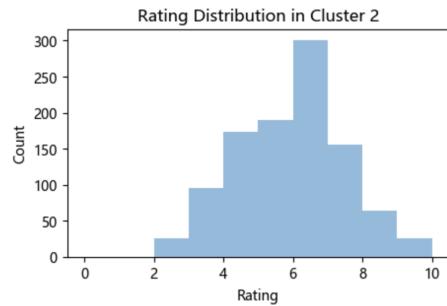
cluster 1：类别比较杂，以剧情、喜剧、动作为主，大部分为娱乐商业片，可以认为是没什么思考量的‘爽片’。

| Cluster 2 : | | | | | | | Cluster 3 : | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|-------------------|--------------|--------------|--|--|--|-------------------|--------|--------------|--------------|-----------------|-------|-------|-----|-----------------|----------|------|-----|------------|-------|------|-----|------------------|----|------|-----|------------|-------------|------|-----|----------------|----------|------|-----|--------------|----------|------|-----|------------------|-------|------|-----|----------------|----------|------|-----|--------------------------|-------|------|-----|---|------|--------|--------------|--------------|----------------|-------|-------|-----|------------------|-------|-------|-----|-----------|----------|-------|-----|------------|----------|-------|-----|-----------|----------|-------|-----|----------|-------------------|-------|-----|------------|----------|-------|-----|-----------|-------|-------|-----|-----------|----------|-------|-----|--------|----|-------|-----|
| 外国电影，这一类电影内容基本上都与西方的历史/文化/艺术有关，国内观众可能对此不太感冒。 | | | | | | | 烂片，豆瓣评分集中在2-4分之间。 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 代表电影及其标签： | | | | | | | 代表电影及其标签： | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 偷心加油站 剧情 西部 | | | | | | | 我未成年 剧情 儿童 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 观看人数最多的前10部电影： | | | | | | | 观看人数最多的前10部电影： | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <thead> <tr> <th>NAME</th><th>GENRES</th><th>DOUBAN_VOTES</th><th>DOUBAN_SCORE</th> </tr> </thead> <tbody> <tr><td>13879 极品基老伴：完结篇</td><td>喜剧 同性</td><td>17474</td><td>9.3</td></tr> <tr><td>30598 神秘博士：圣诞颂歌</td><td>科幻 奇幻 冒险</td><td>4315</td><td>8.6</td></tr> <tr><td>28306 铁甲公敌</td><td>动作 冒险</td><td>3394</td><td>6.6</td></tr> <tr><td>31094 蒂塔·万·提斯疯马秀</td><td>歌舞</td><td>3097</td><td>8.4</td></tr> <tr><td>32630 黄昏公主</td><td>剧情 爱情 历史 战争</td><td>3048</td><td>8.4</td></tr> <tr><td>34575 爱因斯坦与爱丁顿</td><td>剧情 传记 历史</td><td>3023</td><td>7.7</td></tr> <tr><td>14179 亡命徒与天使</td><td>剧情 犯罪 西部</td><td>2973</td><td>7.5</td></tr> <tr><td>20109 神秘博士：最后的圣诞</td><td>剧情 科幻</td><td>2798</td><td>8.4</td></tr> <tr><td>30021 性、毒品和摇滚乐</td><td>剧情 音乐 传记</td><td>2424</td><td>7.4</td></tr> <tr><td>13395 乐高DC超级英雄：正义联盟之宇宙冲击</td><td>动作 动画</td><td>2272</td><td>7.5</td></tr> </tbody> </table> | | | | | | | NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | 13879 极品基老伴：完结篇 | 喜剧 同性 | 17474 | 9.3 | 30598 神秘博士：圣诞颂歌 | 科幻 奇幻 冒险 | 4315 | 8.6 | 28306 铁甲公敌 | 动作 冒险 | 3394 | 6.6 | 31094 蒂塔·万·提斯疯马秀 | 歌舞 | 3097 | 8.4 | 32630 黄昏公主 | 剧情 爱情 历史 战争 | 3048 | 8.4 | 34575 爱因斯坦与爱丁顿 | 剧情 传记 历史 | 3023 | 7.7 | 14179 亡命徒与天使 | 剧情 犯罪 西部 | 2973 | 7.5 | 20109 神秘博士：最后的圣诞 | 剧情 科幻 | 2798 | 8.4 | 30021 性、毒品和摇滚乐 | 剧情 音乐 传记 | 2424 | 7.4 | 13395 乐高DC超级英雄：正义联盟之宇宙冲击 | 动作 动画 | 2272 | 7.5 | <table border="1"> <thead> <tr><td>NAME</td><td>GENRES</td><td>DOUBAN_VOTES</td><td>DOUBAN_SCORE</td></tr> </thead> <tbody> <tr><td>22937 天机·富春山居图</td><td>动作 冒险</td><td>98654</td><td>2.9</td></tr> <tr><td>15964 纯洁心灵·逐梦演艺圈</td><td>剧情 喜剧</td><td>84842</td><td>2.2</td></tr> <tr><td>6034 欧洲攻略</td><td>喜剧 动作 爱情</td><td>52218</td><td>3.5</td></tr> <tr><td>13968 封神传奇</td><td>剧情 动作 奇幻</td><td>50268</td><td>2.9</td></tr> <tr><td>25361 血滴子</td><td>动作 武侠 古装</td><td>49288</td><td>4.6</td></tr> <tr><td>28447 战国</td><td>剧情 动作 爱情 悬疑 战争 古装</td><td>35299</td><td>3.9</td></tr> <tr><td>20624 大话天仙</td><td>喜剧 奇幻 古装</td><td>27247</td><td>3.1</td></tr> <tr><td>17915 放手爱</td><td>喜剧 爱情</td><td>24965</td><td>2.3</td></tr> <tr><td>6616 武林怪兽</td><td>喜剧 奇幻 武侠</td><td>21835</td><td>3.5</td></tr> <tr><td>下一任：前任</td><td>爱情</td><td>20398</td><td>2.8</td></tr> </tbody> </table> | NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | 22937 天机·富春山居图 | 动作 冒险 | 98654 | 2.9 | 15964 纯洁心灵·逐梦演艺圈 | 剧情 喜剧 | 84842 | 2.2 | 6034 欧洲攻略 | 喜剧 动作 爱情 | 52218 | 3.5 | 13968 封神传奇 | 剧情 动作 奇幻 | 50268 | 2.9 | 25361 血滴子 | 动作 武侠 古装 | 49288 | 4.6 | 28447 战国 | 剧情 动作 爱情 悬疑 战争 古装 | 35299 | 3.9 | 20624 大话天仙 | 喜剧 奇幻 古装 | 27247 | 3.1 | 17915 放手爱 | 喜剧 爱情 | 24965 | 2.3 | 6616 武林怪兽 | 喜剧 奇幻 武侠 | 21835 | 3.5 | 下一任：前任 | 爱情 | 20398 | 2.8 |
| NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13879 极品基老伴：完结篇 | 喜剧 同性 | 17474 | 9.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 30598 神秘博士：圣诞颂歌 | 科幻 奇幻 冒险 | 4315 | 8.6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 28306 铁甲公敌 | 动作 冒险 | 3394 | 6.6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 31094 蒂塔·万·提斯疯马秀 | 歌舞 | 3097 | 8.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 32630 黄昏公主 | 剧情 爱情 历史 战争 | 3048 | 8.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 34575 爱因斯坦与爱丁顿 | 剧情 传记 历史 | 3023 | 7.7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 14179 亡命徒与天使 | 剧情 犯罪 西部 | 2973 | 7.5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 20109 神秘博士：最后的圣诞 | 剧情 科幻 | 2798 | 8.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 30021 性、毒品和摇滚乐 | 剧情 音乐 传记 | 2424 | 7.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13395 乐高DC超级英雄：正义联盟之宇宙冲击 | 动作 动画 | 2272 | 7.5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 22937 天机·富春山居图 | 动作 冒险 | 98654 | 2.9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 15964 纯洁心灵·逐梦演艺圈 | 剧情 喜剧 | 84842 | 2.2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6034 欧洲攻略 | 喜剧 动作 爱情 | 52218 | 3.5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13968 封神传奇 | 剧情 动作 奇幻 | 50268 | 2.9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 25361 血滴子 | 动作 武侠 古装 | 49288 | 4.6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 28447 战国 | 剧情 动作 爱情 悬疑 战争 古装 | 35299 | 3.9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 20624 大话天仙 | 喜剧 奇幻 古装 | 27247 | 3.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 17915 放手爱 | 喜剧 爱情 | 24965 | 2.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6616 武林怪兽 | 喜剧 奇幻 武侠 | 21835 | 3.5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 下一任：前任 | 爱情 | 20398 | 2.8 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

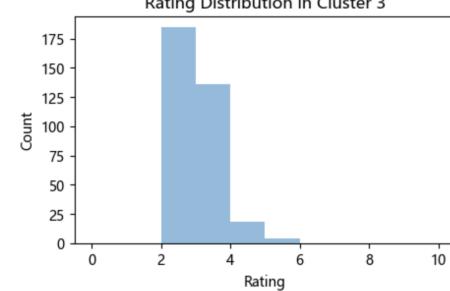
簇中电影的数量: 6958

簇中电影的数量: 7510

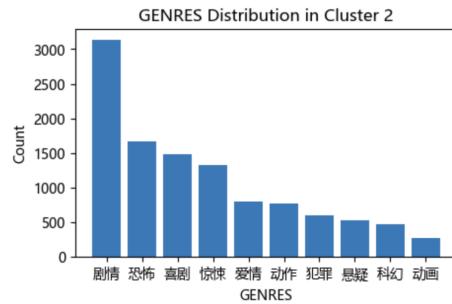
豆瓣评分分布情况:



豆瓣评分分布情况:

评分平均值（忽略0分电影）: 5.90
评分标准差（忽略0分电影）: 1.52评分平均值（忽略0分电影）: 3.00
评分标准差（忽略0分电影）: 0.56Top 3 Movies:
1. Coldplay - Ghost Stories Live 2014 (9.5)
2. 跳出我天地音乐剧 (9.4)
3. 五位(还嫌少)博士重启 (9.4)Top 3 Movies:
1. 非你莫属 (5.3)
2. 油车殿下 (5.1)
3. 一吻定情电影版3: 求婚篇 (5.0)

GENRES电影类型分布情况:



GENRES电影类型分布情况:

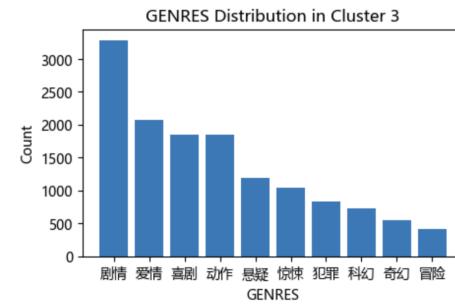


图 14: cluster 2

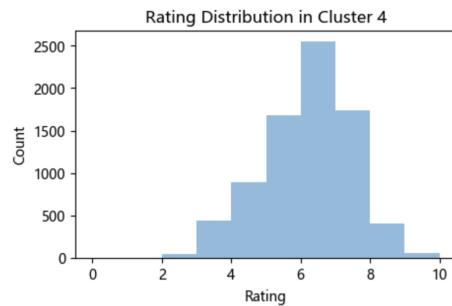
图 15: cluster 3

cluster 2: 外国电影，这一类电影内容基本上都与西方的历史/文化/艺术有关，国内观众可能对此不太感冒，所以在豆瓣平台上明显观影人数少。

cluster 3: 烂片，豆瓣评分集中在 2-4 分之间。

| Cluster 4 : | | | | | Cluster 5 : | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|--------------|--------------|--------------|--|------------------------------------|--------|--------------|--------------|------|---------|-------------|------------|-------|--------|-------------|------------|------|------|----------|------------|-------|-------|-------------|------------|-------|-------|----------|------------|------|------|-------|------------|------|------|-------|------------|-------|--------------|----------|------------|-------|--------|----------|------------|------|----|-------------|------------|--|------|--------|--------------|--------------|-------|------|----------|------------|-------|-------|----------|------------|-------|--------|----------|------------|-------|---------|----------|------------|-------|------|----------|------------|-------|-----|-------|------------|-------|-----------|-------|------------|------|-----|----------|------------|-------|------|-------|------------|-------|--------|-------|------------|
| 人气电影，观影人数明显比其他类多，整体上评分也相对稍好一些 | | | | | 动画/动漫电影，这类电影整体上评分较高，说明动画类电影的好片占比较高 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 代表电影及其标签： 周六夜现场 剧情 喜剧 舞台艺术 | | | | | 代表电影及其标签： 潜艇总动员3：彩虹宝藏 动画 儿童 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 观看人数最多的前10部电影： | | | | | 观看人数最多的前10部电影： | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <thead> <tr> <th>NAME</th><th>GENRES</th><th>DOUBAN_VOTES</th><th>DOUBAN_SCORE</th></tr> </thead> <tbody> <tr><td>3498</td><td>哪咤之魔童降世</td><td>剧情 喜剧 动画 奇幻</td><td>889431 8.6</td></tr> <tr><td>12661</td><td>摔跤吧！爸爸</td><td>剧情 家庭 传记 运动</td><td>870905 9.0</td></tr> <tr><td>6273</td><td>头号玩家</td><td>动作 科幻 冒险</td><td>841114 8.7</td></tr> <tr><td>32201</td><td>飞屋环游记</td><td>剧情 喜剧 动画 冒险</td><td>789525 9.0</td></tr> <tr><td>13969</td><td>你的名字。</td><td>剧情 爱情 动画</td><td>775533 8.4</td></tr> <tr><td>6203</td><td>无名之辈</td><td>剧情 喜剧</td><td>699120 8.1</td></tr> <tr><td>4607</td><td>一出好戏</td><td>剧情 喜剧</td><td>665824 7.1</td></tr> <tr><td>26268</td><td>那些年，我们一起追的女孩</td><td>剧情 喜剧 爱情</td><td>653722 8.1</td></tr> <tr><td>13868</td><td>看不见的客人</td><td>悬疑 惊悚 犯罪</td><td>650397 8.8</td></tr> <tr><td>6646</td><td>无双</td><td>剧情 动作 悬疑 犯罪</td><td>647077 8.1</td></tr> </tbody> </table> | | | | | NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | 3498 | 哪咤之魔童降世 | 剧情 喜剧 动画 奇幻 | 889431 8.6 | 12661 | 摔跤吧！爸爸 | 剧情 家庭 传记 运动 | 870905 9.0 | 6273 | 头号玩家 | 动作 科幻 冒险 | 841114 8.7 | 32201 | 飞屋环游记 | 剧情 喜剧 动画 冒险 | 789525 9.0 | 13969 | 你的名字。 | 剧情 爱情 动画 | 775533 8.4 | 6203 | 无名之辈 | 剧情 喜剧 | 699120 8.1 | 4607 | 一出好戏 | 剧情 喜剧 | 665824 7.1 | 26268 | 那些年，我们一起追的女孩 | 剧情 喜剧 爱情 | 653722 8.1 | 13868 | 看不见的客人 | 悬疑 惊悚 犯罪 | 650397 8.8 | 6646 | 无双 | 剧情 动作 悬疑 犯罪 | 647077 8.1 | <table border="1"> <thead> <tr> <th>NAME</th><th>GENRES</th><th>DOUBAN_VOTES</th><th>DOUBAN_SCORE</th></tr> </thead> <tbody> <tr><td>30921</td><td>驯龙高手</td><td>动画 奇幻 冒险</td><td>446545 8.7</td></tr> <tr><td>36470</td><td>秒速5厘米</td><td>剧情 爱情 动画</td><td>419194 8.3</td></tr> <tr><td>32273</td><td>玛丽和马克思</td><td>剧情 喜剧 动画</td><td>293141 8.9</td></tr> <tr><td>34408</td><td>悬崖上的金鱼姬</td><td>动画 奇幻 冒险</td><td>265528 8.4</td></tr> <tr><td>21393</td><td>怪兽大学</td><td>喜剧 动画 冒险</td><td>248447 8.1</td></tr> <tr><td>16654</td><td>小王子</td><td>动画 奇幻</td><td>200509 8.1</td></tr> <tr><td>19055</td><td>哆啦A梦：伴我同行</td><td>剧情 动画</td><td>199372 8.0</td></tr> <tr><td>2960</td><td>狮子王</td><td>剧情 动画 冒险</td><td>190326 7.4</td></tr> <tr><td>22459</td><td>言叶之庭</td><td>爱情 动画</td><td>187584 8.2</td></tr> <tr><td>17666</td><td>小黄人大眼萌</td><td>喜剧 动画</td><td>185214 7.5</td></tr> </tbody> </table> | NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | 30921 | 驯龙高手 | 动画 奇幻 冒险 | 446545 8.7 | 36470 | 秒速5厘米 | 剧情 爱情 动画 | 419194 8.3 | 32273 | 玛丽和马克思 | 剧情 喜剧 动画 | 293141 8.9 | 34408 | 悬崖上的金鱼姬 | 动画 奇幻 冒险 | 265528 8.4 | 21393 | 怪兽大学 | 喜剧 动画 冒险 | 248447 8.1 | 16654 | 小王子 | 动画 奇幻 | 200509 8.1 | 19055 | 哆啦A梦：伴我同行 | 剧情 动画 | 199372 8.0 | 2960 | 狮子王 | 剧情 动画 冒险 | 190326 7.4 | 22459 | 言叶之庭 | 爱情 动画 | 187584 8.2 | 17666 | 小黄人大眼萌 | 喜剧 动画 | 185214 7.5 |
| NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3498 | 哪咤之魔童降世 | 剧情 喜剧 动画 奇幻 | 889431 8.6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 12661 | 摔跤吧！爸爸 | 剧情 家庭 传记 运动 | 870905 9.0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6273 | 头号玩家 | 动作 科幻 冒险 | 841114 8.7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 32201 | 飞屋环游记 | 剧情 喜剧 动画 冒险 | 789525 9.0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13969 | 你的名字。 | 剧情 爱情 动画 | 775533 8.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6203 | 无名之辈 | 剧情 喜剧 | 699120 8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4607 | 一出好戏 | 剧情 喜剧 | 665824 7.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 26268 | 那些年，我们一起追的女孩 | 剧情 喜剧 爱情 | 653722 8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13868 | 看不见的客人 | 悬疑 惊悚 犯罪 | 650397 8.8 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6646 | 无双 | 剧情 动作 悬疑 犯罪 | 647077 8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 30921 | 驯龙高手 | 动画 奇幻 冒险 | 446545 8.7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 36470 | 秒速5厘米 | 剧情 爱情 动画 | 419194 8.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 32273 | 玛丽和马克思 | 剧情 喜剧 动画 | 293141 8.9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 34408 | 悬崖上的金鱼姬 | 动画 奇幻 冒险 | 265528 8.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 21393 | 怪兽大学 | 喜剧 动画 冒险 | 248447 8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 16654 | 小王子 | 动画 奇幻 | 200509 8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 19055 | 哆啦A梦：伴我同行 | 剧情 动画 | 199372 8.0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2960 | 狮子王 | 剧情 动画 冒险 | 190326 7.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 22459 | 言叶之庭 | 爱情 动画 | 187584 8.2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 17666 | 小黄人大眼萌 | 喜剧 动画 | 185214 7.5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 簇中电影的数量：7805 | | | | | 簇中电影的数量：2058 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

豆瓣评分分布情况：

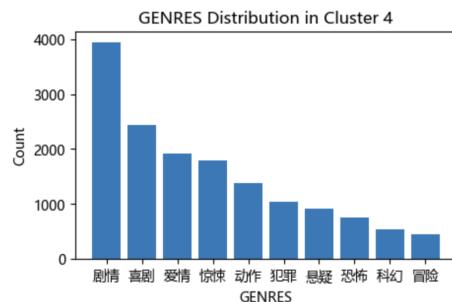


评分平均值（忽略0分电影）：6.17
评分标准差（忽略0分电影）：1.26

Top 3 Movies:

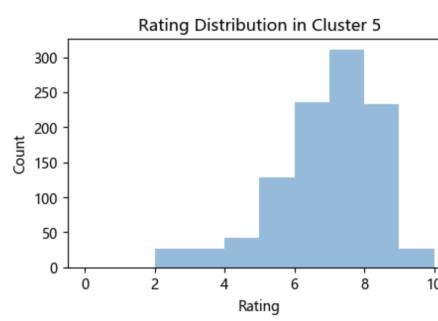
1. 悲惨世界：25周年纪念演唱会 (9.6)
2. 神秘博士：DT的视频日志：最后的日子 (9.6)
3. 第十二夜 (9.5)

GENRES电影类型分布情况：



| Cluster 5 : | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|-----------|--------------|--------------|--|------|--------|--------------|--------------|-------|------|----------|------------|-------|-------|----------|------------|-------|--------|----------|------------|-------|---------|----------|------------|-------|------|----------|------------|-------|-----|-------|------------|-------|-----------|-------|------------|------|-----|----------|------------|-------|------|-------|------------|-------|--------|-------|------------|
| 动画/动漫电影，这类电影整体上评分较高，说明动画类电影的好片占比较高 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 代表电影及其标签： 潜艇总动员3：彩虹宝藏 动画 儿童 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 观看人数最多的前10部电影： | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <thead> <tr> <th>NAME</th><th>GENRES</th><th>DOUBAN_VOTES</th><th>DOUBAN_SCORE</th></tr> </thead> <tbody> <tr><td>30921</td><td>驯龙高手</td><td>动画 奇幻 冒险</td><td>446545 8.7</td></tr> <tr><td>36470</td><td>秒速5厘米</td><td>剧情 爱情 动画</td><td>419194 8.3</td></tr> <tr><td>32273</td><td>玛丽和马克思</td><td>剧情 喜剧 动画</td><td>293141 8.9</td></tr> <tr><td>34408</td><td>悬崖上的金鱼姬</td><td>动画 奇幻 冒险</td><td>265528 8.4</td></tr> <tr><td>21393</td><td>怪兽大学</td><td>喜剧 动画 冒险</td><td>248447 8.1</td></tr> <tr><td>16654</td><td>小王子</td><td>动画 奇幻</td><td>200509 8.1</td></tr> <tr><td>19055</td><td>哆啦A梦：伴我同行</td><td>剧情 动画</td><td>199372 8.0</td></tr> <tr><td>2960</td><td>狮子王</td><td>剧情 动画 冒险</td><td>190326 7.4</td></tr> <tr><td>22459</td><td>言叶之庭</td><td>爱情 动画</td><td>187584 8.2</td></tr> <tr><td>17666</td><td>小黄人大眼萌</td><td>喜剧 动画</td><td>185214 7.5</td></tr> </tbody> </table> | | | | | NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | 30921 | 驯龙高手 | 动画 奇幻 冒险 | 446545 8.7 | 36470 | 秒速5厘米 | 剧情 爱情 动画 | 419194 8.3 | 32273 | 玛丽和马克思 | 剧情 喜剧 动画 | 293141 8.9 | 34408 | 悬崖上的金鱼姬 | 动画 奇幻 冒险 | 265528 8.4 | 21393 | 怪兽大学 | 喜剧 动画 冒险 | 248447 8.1 | 16654 | 小王子 | 动画 奇幻 | 200509 8.1 | 19055 | 哆啦A梦：伴我同行 | 剧情 动画 | 199372 8.0 | 2960 | 狮子王 | 剧情 动画 冒险 | 190326 7.4 | 22459 | 言叶之庭 | 爱情 动画 | 187584 8.2 | 17666 | 小黄人大眼萌 | 喜剧 动画 | 185214 7.5 |
| NAME | GENRES | DOUBAN_VOTES | DOUBAN_SCORE | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 30921 | 驯龙高手 | 动画 奇幻 冒险 | 446545 8.7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 36470 | 秒速5厘米 | 剧情 爱情 动画 | 419194 8.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 32273 | 玛丽和马克思 | 剧情 喜剧 动画 | 293141 8.9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 34408 | 悬崖上的金鱼姬 | 动画 奇幻 冒险 | 265528 8.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 21393 | 怪兽大学 | 喜剧 动画 冒险 | 248447 8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 16654 | 小王子 | 动画 奇幻 | 200509 8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 19055 | 哆啦A梦：伴我同行 | 剧情 动画 | 199372 8.0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2960 | 狮子王 | 剧情 动画 冒险 | 190326 7.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 22459 | 言叶之庭 | 爱情 动画 | 187584 8.2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 17666 | 小黄人大眼萌 | 喜剧 动画 | 185214 7.5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 簇中电影的数量：2058 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

豆瓣评分分布情况：



评分平均值（忽略0分电影）：6.92
评分标准差（忽略0分电影）：1.45

Top 3 Movies:

1. 夏目友人帐 第六季 特别篇 铃响的残株 (9.6)
2. 夏目友人帐 五 特别篇：一夜酒杯 (9.5)
3. 狐妖小红娘剧场版：月红篇 (9.3)

GENRES电影类型分布情况：

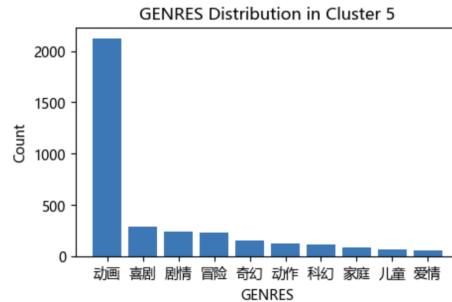


图 16: cluster 4

图 17: cluster 5

cluster 4：人气电影，观影人数明显比其他类多，整体上评分也相对稍好一些。

cluster 5：动画/动漫电影，这类电影整体上评分较高，说明动画类电影的好片占比较高。

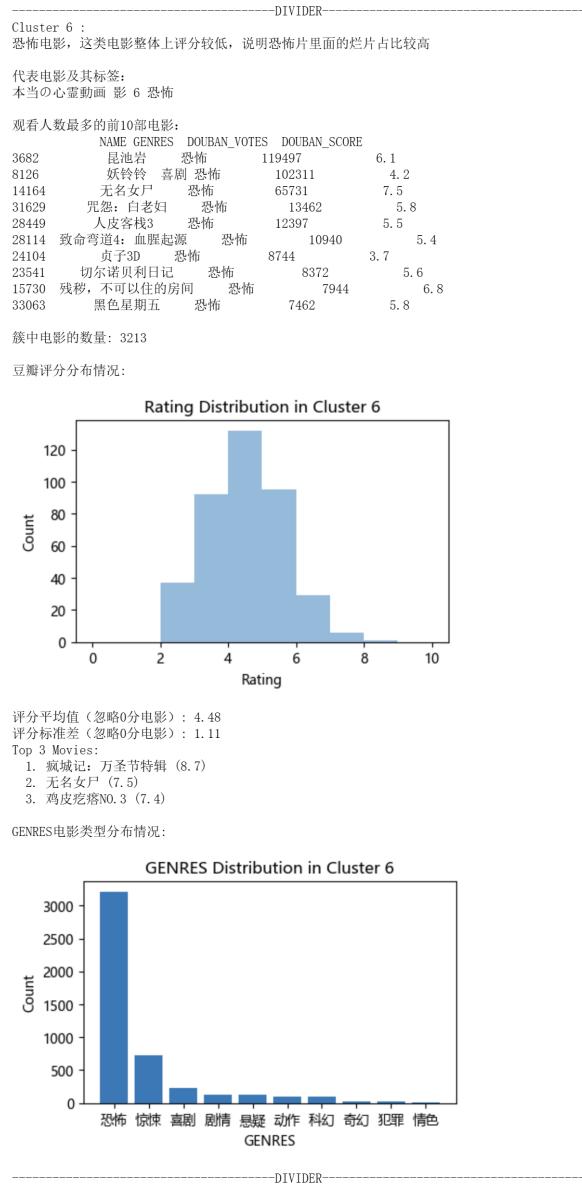


图 18: cluster 6

cluster 6：恐怖电影，这类电影整体上评分较低，说明恐怖片里面的烂片占比较高。

最后我以簇为横坐标、以评分均值和标准差为纵坐标，输出了两张柱状图，能更好地横向对比每类电影的评分情况，印证上面对每个簇电影共同特点的分析。

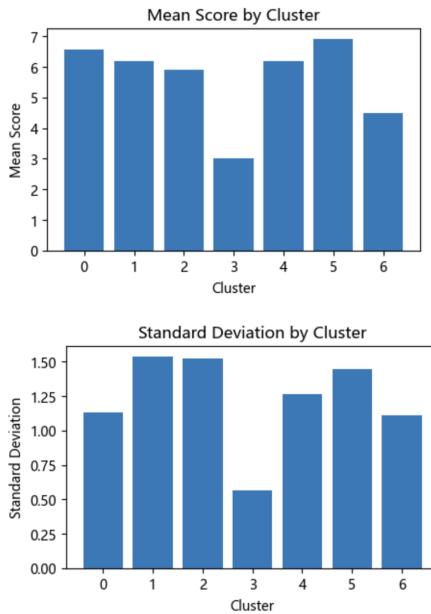


图 19: 不同簇的评分均值与标准差

4 Task4：为导演和演员生成 Embedding，进行无监督分类，并分析 2-3 个属于不同类别导演和演员的特点

4.1 对导演和演员生成特征向量、再分别进行层次聚类

对导演和演员字段进行分词，然后统计所有导演和演员各自参与过的电影，以演员名:[所出演的电影编号的列表]这样一个字典的形式储存（对导演同理）。统计之后发现数据集中共涉及 50663 名演员，12839 名导演。

为了缩小规模，我再进行了一次筛选，只保留那些出演了 8 部以上电影的演员和执导了 3 部电影以上的导演，并且参与过的电影至少有 2 部有评分。筛选完成后剩余 2710 名演员和 1278 名导演。

```

数据集中总共涉及的演员人数:  
50663  
数据集中总共涉及的导演人数:  
12839  
  
再进行一次筛选，只保留那些出演了8部以上电影的演员和执导了3部电影以上的导演，且电影不能全部无评分。  
筛选后数据集中总共涉及的演员人数:  
2710  
筛选后数据集中总共涉及的导演人数:  
1278  
  
查看其中10位演员的信息（演员名:[所出演的电影编号的列表]）：  
['王博': [0, 3364, 6876, 6902, 7908, 9055, 9645, 11330, 14378, 14549, 15451, 27658, 32255], '王婧': [0, 737, 1994, 2516, 7338, 17315, 18552, 19006, 22641, 24103, 24145, 25058, 32529, 33352, 35259, 37088], '沈丹萍': [0, 6382, 15665, 18688, 19358, 25487, 28085, 31039, 31545, 33881], '吴孟达': [1, 994, 1159, 3493, 4159, 6963, 9804, 13948, 14928, 17808, 19662, 24151, 25058, 28362, 35214], '曾志伟': [1, 73, 994, 2106, 2181, 2247, 4473, 5466, 6102, 6963, 7532, 9119, 9271, 11295, 12129, 13582, 14968, 15206, 16009, 17373, 19065, 19662, 20374, 21477, 21533, 21925, 2227, 22771, 24049, 24229, 26351, 27087, 29419, 31118, 31344, 32072, 33260, 33402, 34335, 35037, 35214, 35385, 35439, 36102, 36935, 37487], '林海音': [1, 73, 4600, 9484, 9969, 10582, 11535, 11816, 12243, 12782, 12908, 16826, 17477, 17810, 20073, 20367, 22058, 24766, 24851, 27377, 31212, 31534, 32675], '奥莉薇·瑟尔比': [15, 4305, 10658, 13449, 17128, 24130, 24871, 25565, 28319, 28484, 32079, 33764, 34791, 36699, 37486], '西格妮·韦弗': [15, 38, 60, 72, 1007, 1274, 11314, 12889, 13977, 23888, 24094, 35122, 35530], '埃文·蕾切尔·伍德': [15, 312, 2215, 860, 15369, 17843, 19287, 20596, 22979, 28156, 32018, 35388, 37075, 37076], '王丽坤': [22, 3818, 9304, 14935, 15516, 19849, 22795, 34666]}  
  
查看其中10位导演的信息（导演名:[所执导的电影编号的列表]）：  
['比利·奥古斯特': [25, 9010, 17926, 23841], '速达': [31, 8609, 24920], '莱恩·约翰逊': [39, 75, 23877, 34828], '汤姆·摩尔': [41, 18717, 20506, 33666], '陈力': [47, 25162, 26615], '周星驰': [54, 1820, 21900], '拉斯·霍尔斯道姆': [59, 23328, 28492, 29138], '韩在林': [66, 21558, 36664], '黄真真': [71, 593, 12277, 18481, 26559, 30071, 36092], '罗兰·艾默里奇': [86, 611, 4988, 10178, 14764, 21403, 28446, 34008, 35243, 36586]}

```

图 20: 统计和筛选导演、演员的数据

然后，对这些导演和演员，使用对应的电影 embedding 向量累加求平均值的方式，生成导演和演员的 embedding 向量。

```
# 计算导演和演员的embedding
actors_names = [] # 初始化演员姓名列表
directors_names = [] # 初始化导演姓名列表
# 为导演生成embedding
directors_embeddings = np.zeros((len(director_dict), embedding_pca.shape[1])) # 初始化导演embedding矩阵
for i, (director_name, movie_ids) in enumerate(director_dict.items()):
    director_embeddings = np.zeros((len(movie_ids), embedding_pca.shape[1])) # 初始化该导演的embedding矩阵
    for j, movie_id in enumerate(movie_ids):
        director_embeddings[j] = embedding_pca[movie_id] # 获取该电影的embedding
    mean_embeddings = np.mean(director_embeddings, axis=0) # 对该导演的所有电影embedding取平均值
    directors_embeddings[i] = mean_embeddings
    directors_names.append(director_name)

# 为演员生成embedding
actors_embeddings = np.zeros((len(actor_dict), embedding_pca.shape[1])) # 初始化演员embedding矩阵
for i, (actor_name, movie_ids) in enumerate(actor_dict.items()):
    actor_embeddings = np.zeros((len(movie_ids), embedding_pca.shape[1])) # 初始化该演员的embedding矩阵
    for j, movie_id in enumerate(movie_ids):
        actor_embeddings[j] = embedding_pca[movie_id] # 获取该电影的embedding
    mean_embeddings = np.mean(actor_embeddings, axis=0) # 对该演员的所有电影embedding取平均值
    actors_embeddings[i] = mean_embeddings
    actors_names.append(actor_name)

print("导演数据的embedding矩阵大小: ", directors_embeddings.shape)
print("演员数据的embedding矩阵大小: ", actors_embeddings.shape)

导演数据的embedding矩阵大小: (1278, 26)
演员数据的embedding矩阵大小: (2710, 26)
```

图 21: 求导演和演员的 embedding 向量

最后使用 sklearn 库中的 AgglomerativeClustering() 进行层次聚类，导演和演员各聚为 5 类，设置参数为 linkage='ward' 采用方差和最小的方式聚类。

```
# 对导演、演员分别进行无监督分类（此处使用层次聚类）
from sklearn.cluster import AgglomerativeClustering
# 对导演进行层次聚类
n_clusters_directors = 5
model_directors = AgglomerativeClustering(n_clusters=n_clusters_directors, linkage='ward')
labels_directors = model_directors.fit_predict(directors_embeddings)
# 对演员进行层次聚类
n_clusters_actors = 5
model_actors = AgglomerativeClustering(n_clusters=n_clusters_actors, linkage='ward')
labels_actors = model_actors.fit_predict(actors_embeddings)

print("导演的分类结果（查看前100位导演的分类标签）：")
print(labels_directors[:100])
print("演员的分类结果（查看前100位演员的分类标签）：")
print(labels_actors[:100])

导演的分类结果（查看前100位导演的分类标签）：
[1 3 0 3 2 0 2 0 0 0 1 0 0 0 0 3 0 0 0 2 0 0 0 0 1 2 2 2 0 0 0 0 1 0 0
 0 0 1 0 2 1 0 1 0 0 3 0 0 0 0 0 3 0 0 0 0 0 0 0 2 3 0 1 3 0 3 1 2 2
 0 0 1 0 3 0 2 0 0 4 2 0 3 0 2 0 0 3 2 0 4 3 0 0 0 3]
演员的分类结果（查看前100位演员的分类标签）：
[4 4 4 4 4 4 1 1 1 4 4 4 4 4 1 4 1 1 1 4 4 1 1 1 1 1 1 1 1 1 4 1 1 1 1 0 1
 1 4 1 1 1 1 4 4 1 1 1 1 1 0 4 4 4 4 1 1 1 1 1 4 1 4 1 1 1 4 4 1 1 1 1]
```

图 22: 导演和演员的聚类结果

4.2 导演和演员聚类结果的简单分析

对导演的聚类结果，对每一簇，我输出了簇内的导演数、代表导演（离簇中心最近）、簇中执导电影数目最多的 4 位导演以及他们分别的 3 部代表作。演员同理进行统计。

（此处统计簇内信息的代码较长，报告中不粘贴代码，只展示结果）

| Cluster 0 : 导演 | | | | | | |
|-------------------------------|--------------------------------------|--------------|--------------|-----|--|--|
| 簇内导演数: 516 代表导演: 刘伟强 | | | | | | |
| 簇中执导电影数最多的4位导演, 以及每人的三部作品: | | | | | | |
| 刘伟强 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 16991 | 九纹龙史记之陈胜吴广 | 剧情 动作 武侠 | 105 | 6.4 | | |
| 12541 | 九纹龙史记之戚继明 | 剧情 动作 | 66 | 6.1 | | |
| 18672 | 顾大嫂与孙新 喜剧 爱情 国际 武侠 古装 | | 77 | 5.9 | | |
| 邱礼涛 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 35144 | 性工作者2: 我们不卖身, 我们爱 | 剧情 | 739 | 6.9 | | |
| 21833 | 性工作者3: 我们不再沉默 | 剧情 动作 惊悚 | 20348 | 6.2 | | |
| 27076 | 变女郎 海芋恋 动作 悬疑 犯罪 | | 27393 | 6.1 | | |
| 王晶 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 871 | 追龙 动作 犯罪 | 23068 | 7.2 | | | |
| 32598 | 金枝玉叶 犯罪 | 42686 | 6.4 | | | |
| 2733 | 大话西游 铜雀台 刑警 | | 351 | 6.3 | | |
| 钟玲 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 16130 | 福星高照 动画 | 208 | 7.5 | | | |
| 27252 | 简单的美丽 剧情 | 114 | 7.3 | | | |
| 31308 | 一屋五班 剧情 | | 76 | 5.7 | | |
| Cluster 1 : 导演 | | | | | | |
| 簇内导演数: 145 代表导演: 陈嘉上 · 阿来 | | | | | | |
| 簇中执导电影数最多的4位导演, 以及每人的三部作品: | | | | | | |
| 晋文安 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 20246 | 人妖江湖014 喜剧 | 157 | 6.7 | | | |
| 22999 | 这高中也有爱 喜剧 恐怖 | 4517 | 6.0 | | | |
| 16663 | 那个年代 喜剧 喜剧 爱情 | | 73 | 5.7 | | |
| 马克·阿特拉斯 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 19531 | 机器侠警 动作 科幻 | 75 | 4.0 | | | |
| 8085 | 蓝调夜馆 动作 科幻 | 150 | 3.5 | | | |
| 12273 | 蓝色星球 动作 科幻 恐怖 灾难 | | 319 | 3.0 | | |
| 弗雷雷·欧··雷 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 11320 | 狙击手: 特别行动 剧情 动作 惊悚 | | 579 | 4.1 | | |
| 32301 | 变种蛇蝎 动作 喜剧 恐怖 | | 187 | 3.7 | | |
| 24458 | 揪机 剧情 动作 科幻 恐怖 暴殓 | | 103 | 3.5 | | |
| 克里斯蒂安·佩纳尔多 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 24649 | 巴比伦塔 喜剧 | 4847 | 7.4 | | | |
| 34203 | 黑金三角 喜剧 | 682 | 7.1 | | | |
| 36396 | 邵氏 喜剧 爱情 恐怖 | | 442 | 7.0 | | |
| Cluster 2 : 导演 | | | | | | |
| 簇内导演数: 329 代表导演: 张艺谋 | | | | | | |
| 簇中执导电影数最多的4位导演, 以及每人的三部作品: | | | | | | |
| 田少波 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 25740 | 秦时明月 剧情 动画 | 165 | 7.3 | | | |
| 29597 | 战地诱惑 战争 情感 | 285 | 7.3 | | | |
| 27043 | 女王 剧情 | | 117 | 6.9 | | |
| 周伟 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 36890 | 离别也是爱 喜剧 | 517 | 7.1 | | | |
| 37418 | 棋王和他的儿子 剧情 儿童 儿童 | | 1198 | 6.9 | | |
| 35317 | 升生乐章 喜剧 | | 652 | 6.7 | | |
| 邓衍成 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 33985 | 扶墙道之突厥死亡 剧情 动作 | 3619 | 7.7 | | | |
| 34013 | 火柴盒凶之惊魂魔女 剧情 喜剧 | | 3600 | 7.6 | | |
| 36443 | 陆小凤传奇之大金鹏王 动作 武侠 古装 | | 10715 | 7.5 | | |
| 高峰 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 16914 | 土地公 土地公 喜剧 | 315 | 8.4 | | | |
| 34778 | 剧场版 黑子 喜剧 | 132 | 7.9 | | | |
| 23772 | 蝶变者 动漫 惊悚 监禁 | | 2900 | 7.8 | | |
| Cluster 3 : 导演 | | | | | | |
| 簇内导演数: 110 代表导演: 日高政光 | | | | | | |
| 簇中执导电影数最多的4位导演, 以及每人的三部作品: | | | | | | |
| 水鸟努 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 26686 | 棋圣 高清国棋 剧情 动画 | 669 | 8.9 | | | |
| 19144 | 母鸡变凤凰 喜剧 爱情 | 1782 | 8.7 | | | |
| 22917 | 剧场版2代目 Q&D 动画 | | 228 | 8.2 | | |
| 石原正也 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 15155 | 吹响吧! TV未放送前 喜剧 爱情 动画 音乐 | | 1160 | 9.0 | | |
| 31869 | 团子家族第一季番外篇: 一年前的事 剧情 动画 | | 715 | 8.8 | | |
| 32186 | 团子家族第二季番外篇: 在蘑菇的树下 剧情 动画 | | 593 | 8.7 | | |
| 羽原直久 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 8897 | 宇宙舰队大号2203: 传说的拉杜 第一季 喜剧 科幻 | | 141 | 7.9 | | |
| 27846 | 剧场版 黑之介 第五章 黑之介之谜 喜剧 | | 1016 | 7.9 | | |
| 30700 | 剧场版 黑之介 第四章 黑之介之谜 喜剧 | | 1007 | 7.9 | | |
| 新井裕树 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 22524 | 剧场版魔法少女小圆 魔法 少女 喜剧 | | 9206 | 9.0 | | |
| 24174 | 剧场版 魔法少女小圆 魔法 少女 喜剧 | | 3552 | 8.8 | | |
| 20395 | 怦·再见: 魔法少女! (Magic Girl Box 全纪念版) 动画 | | 132 | 8.7 | | |
| Cluster 4 : 导演 | | | | | | |
| 簇内导演数: 178 代表导演: 赵卓彦 | | | | | | |
| 簇中执导电影数最多的4位导演, 以及每人的三部作品: | | | | | | |
| 吴京 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 16914 | 战狼2 喜剧 动作 | 621 | 9.5 | | | |
| 32813 | SPEC 亂 剧情 悬念 | 27367 | 9.5 | | | |
| 19298 | 攀登者 喜剧 战争 | | 4797 | 8.0 | | |
| 成龙秀才 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 32539 | 红灯里的女孩儿 喜剧 | | 87 | 7.2 | | |
| 19353 | 美人鱼白书 喜剧 情色 | | 184 | 6.2 | | |
| 32931 | 18岁 喜剧 喜剧 | | 128 | 5.2 | | |
| 三浦友和 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 2429 | 初恋 喜剧 动情 | | 72 | 7.9 | | |
| 32964 | 热血高校2 热血 高校 动作 | | 61273 | 7.9 | | |
| 29714 | 十三朝花木 动作 古装 | | 13446 | 7.7 | | |
| DIVIDER | | | | | | |
| Cluster 0 : 演员 | | | | | | |
| 簇内的演员数: 385 代表演员: 热拉尔·德帕迪约 | | | | | | |
| 簇中参演电影数目最多的4位演员, 以及每人的三部作品: | | | | | | |
| 埃里克·侯纳 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 21517 | 八人华尔街 剧情 惊悚 | | 10613 | 6.5 | | |
| 27959 | 惊天大逆转 喜剧 动作 科幻 惊悚 歌舞 | | 5572 | 6.5 | | |
| 21374 | 技惊四座 喜剧 传记 | | 9981 | 6.3 | | |
| 热拉尔·德帕迪约 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 30156 | 与玛格丽特的午后 喜剧 | | 23372 | 8.8 | | |
| 32779 | 汕头 喜剧 | | 236 | 7.6 | | |
| 29739 | 俄狄的世界 喜剧 | | 152 | 7.5 | | |
| 汤姆·莱塞尼尔 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 31664 | 21 and a Wake-up 喜剧 战争 | | 54 | 6.3 | | |
| 13914 | 杜兰朵 色情 黑色 传记 暗恋 惊悚 | | 151 | 6.1 | | |
| 17026 | 6种死亡方式 动作 惊悚 犯罪 | | 365 | 5.6 | | |
| 马修·阿马立克 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 23757 | 杀熟比尔整个世界 喜剧 动作 犯罪 | | 14059 | 8.5 | | |
| 15011 | 八人入局 喜剧 惊悚 动作 | | 14556 | 8.4 | | |
| 31778 | 不朽者 喜剧 惊悚 动作 | | 741 | 6.8 | | |
| 丹尼·特雷雷 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 13868 | 逃亡骑士 挑战大飞 喜剧 动画 音乐 | | 27460 | 7.5 | | |
| 20889 | 人见是亲 喜剧 | | 803 | 6.6 | | |
| 29717 | 新铁血战士 动作 科幻 惊悚 剧情 | | 23361 | 6.3 | | |
| 尼古拉斯·卡索 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 6594 | 少年泰坦电影版 喜剧 动作 惊悚 | | 3874 | 8.1 | | |
| 37281 | 海扁王 喜剧 动作 犯罪 | | 24437 | 7.8 | | |
| 29498 | 海扁王 喜剧 动作 犯罪 | | 187722 | 7.5 | | |
| Cluster 1 : 演员 | | | | | | |
| 簇内的演员数: 509 代表演员: 邱彦宏 | | | | | | |
| 簇中参演电影数目最多的4位演员, 以及每人的三部作品: | | | | | | |
| 竹中直人 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 2261 | 海贼王: 狂狂行动 动画 冒险 | | 1346 | 8.9 | | |
| 30491 | 安魂夜梦 梦境 乐章 最终乐章 后篇 喜剧 | | 32114 | 8.8 | | |
| 32368 | 安魂夜梦 梦境 乐章 后篇 喜剧 音乐 | | 30599 | 8.5 | | |
| 光石研 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 18797 | 哪吒闹海神上村 喜剧 惊悚 音乐 | | 11109 | 8.5 | | |
| 2118 | 今天是见鬼的日子 喜剧 动作 | | 5547 | 8.3 | | |
| 18446 | 暗金鸟岛2 喜剧 喜剧 犯罪 | | 5718 | 7.9 | | |
| 小日向文世 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 1510 | 行骗天下 剧场版 动画 喜剧 暗恋 | | 1941 | 8.5 | | |
| 3336 | 行骗天下2 行骗篇 动画 喜剧 | | 18939 | 8.3 | | |
| 34511 | 魔幻乐章 喜剧 | | 23856 | 8.3 | | |
| 吳谷鶴 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 30082 | 无头骑士异闻录: 天国与我 动画 | | 1429 | 8.8 | | |
| 21162 | 命运之门 剧场版: 水滑梯 喜剧 动画 | | 11614 | 8.7 | | |
| 33885 | 新宿事件 喜剧 惊悚 动画 音乐 | | 1558 | 8.6 | | |
| 袴谷浩史 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 9866 | 相思多少 暗恋 钟鸣林 剧情 暗恋 音乐 | | 477 | 8.9 | | |
| 10357 | 排球少年! 剧场版: 才能与感觉 喜剧 动画 | | 438 | 8.9 | | |
| 7310 | 我的英雄学院MHO 死亡地带 喜剧 动画 | | 1952 | 8.8 | | |
| 花泽香菜 | | | | | | |
| NAME | GENES | DOUBAN_VOTES | DOUBAN_SCORE | | | |
| 30082 | 无头骑士异闻录: 天国与我 动画 | | 1429 | 8.8 | | |
| 21162</td | | | | | | |

视化分析，在此就先不写了。

4.3 可视化分析 2 个不同类型导演的特点

此处对导演聚类的 cluster 3 和 cluster 1 进行了分析。对每一类，我首先统计出该类导演执导过的全部电影，由此可以获取该类导演的类型 GENRES 偏好（将全部电影的标签统计起来，取频率最高的 6 个标签绘制雷达图）。随后对该类导演所执导电影的 DOUBAN_SCORE 分布（将全部电影的分数统计起来，忽略评分为 0 的电影）也进行了统计和绘制雷达图。此外，还输出了该类导演的完整名单作为参考。（此处统计簇内信息的代码较长，报告中不粘贴代码，只展示结果）

首先分析导演的 Cluster 3：

Cluster 3 中的所有导演名单：

「速达」，「汤姆·摩尔」，「宫崎骏」，「汤浅政明」，「史蒂夫·马蒂诺」，「森本晃司」，「黄伟明」，「汤姆·麦格拉思」，「莱因哈德·克洛斯」，「克里斯·威廉姆斯」，「水岛努」，「王云飞」，「拜伦·霍华德」，「王川」，「诺拉·托梅」，「宋岳峰」，「曾宪林」，「新海诚」，「汤继业」，「丁亮」，「刘山姆」，「羽原信义」，「金泽洪充」，「山本宽」，「长井龙雪」，「殷玉麒」，「沃尔特·拜克」，「原惠一」，「河浪荣作」，「郑成峰」，「菊地康仁」，「李蒙凌」，「上村泰」，「石原立也」，「米谷良知」，「马特·皮特斯」，「阿部记之」，「安藤真裕」，「多田俊介」，「新房昭之」，「河森正治」，「巴里·库克」，「伊桑·斯波尔丁」，「妮娜·佩利」，「后信治」，「村田和也」，「长崎健司」，「溝仲勳」，「安立奎·高德」，「王昕」，「大槻敦史」，「大森贵弘」，「岸诚二」，「王章俊」，「高桥敦史」，「川口敬一郎」，「陆锦明」，「古桥一浩」，「合田浩章」，「曾利文彦」，「史派克·布兰特」，「神保昌登」，「太田雅彦」，「大地丙太郎」，「尾石达也」，「杨广福」，「周彬」，「安藤裕章」，「濑下宽之」，「柳泽哲也」，「托尼·塞沃恩」，「米歇尔·欧斯洛」，「日高政光」，「锅岛修」，「神山健治」，「黄军」，「汤山邦彦」，「夏目真悟」，「草川启造」，「迫井政行」，「葛谷直行」，「川崎逸朗」，「宫繁之」，「善晓一郎」，「岩崎良明」，「武本康弘」，「市村徹夫」，「佐藤顺一」，「施屹」，「黄瀬和哉」，「古田丈司」，「马可·A·Z·迪普」，「邓东明」，「比尔·普莱顿」，「杰伊·欧力瓦」，「工藤进」，「松江名俊」，「贺梦凡」，「石平信司」，「出渊裕」，「山本秀世」，「平尾隆之」，「青木荣」，「入江泰浩」，「山本泰一郎」，「赵崇邦」，「吉原正行」，「山田尚子」，「网野哲朗」，「劳伦·蒙哥马利」】

Cluster 3 中的电影总数： 464

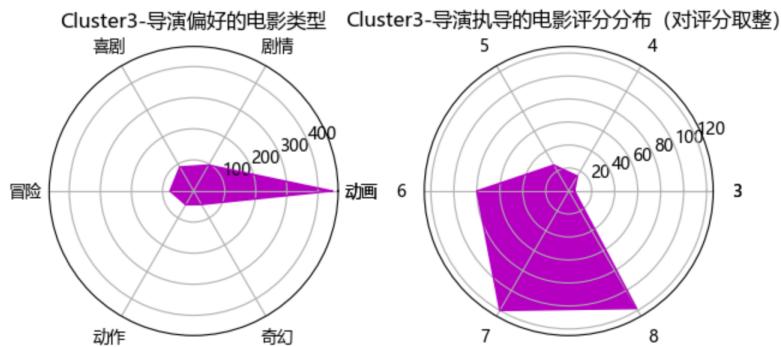


图 25: 导演 cluster 3

可以看出，Cluster 3 中的 464 部电影几乎全部是动画电影，其他频率较高的类型还有剧情、喜剧、冒险、奇幻等，也与动画相关联。

对照导演名单可以看出，这一类导演混杂中国与外国、东西方的导演，其中日本导演最多，且都是以制作动漫电影出名的导演，与电影特征相一致。

这部分电影的评分集中在 6-8 分，特别是 7-8 分之间，与平均水平相比有明显提高，说明（至少在这个数据集中）动画类电影比较出色。

下面对导演的Cluster 1进行分析:

Cluster 1中的所有导演名单:

「比利·奥古斯特」,「亚历山大·阿嘉」,「拉加·高斯内尔」,「泰勒·海克福德」,「亚当·温加德」,「马丁·斯科塞斯」,「尼古拉斯·斯托勒」,「王子逸」,「斯科特·德瑞克森」,「西娅·夏罗克」,「肖恩·麦克纳马拉」,「泰伦斯·马力克」,「塞德里克·康」,「李·杜兰·克里格」,「安德烈·艾弗道夫」,「汤姆·哈伯」,「阿尔伯特·塞拉」,「威尔·古勒」,「彭浩翔」,「安迪·坦纳特」,「艾德亚多·桑奇兹」,「丹尼·博伊尔」,「克里斯蒂安·佩措尔德」,「马库斯·邓斯坦」,「瑞恩·墨菲」,「亚历桑德拉泰雷斯·肯宁」,「文森佐·纳塔利」,「诺亚·鲍姆巴赫」,「马库斯·尼斯佩尔」,「亚当梅森」,「戴维·布莱尔」,「尤金·格林」,「小理查德·贝茨」,「马克·佩灵顿」,「S·S·拉贾穆里」,「迈克尔·霍夫曼」,「伯纳德·罗斯」,「阿扬·慕克吉」,「马克·杨」,「莱·拉索扬」,「乔安娜·霍格」,「德尼·科泰」,「李洙成」,「马克·阿特金斯」,「瑞卡多·米拉尼」,「史蒂芬·戴德利」,「斯科特·斯皮尔」,「罗宁·提姆」,「尼克·里昂」,「詹姆斯·肯特」,「伊沃·冯·霍夫」,「托德·海因斯」,「法比耶娜·贝尔托」,「彼得·休伊特」,「刘杰」,「罗曼·卡兹」,「大卫·米奇欧德」,「艾瑞克·英格兰德」,「史蒂文·R·蒙若尔」,「陈俊彦」,「肯尼思·布拉纳」,「马库斯·戈勒」,「塞巴斯蒂安·古提耶雷兹」,「曼斯·马林德」,「比约恩·史坦」,「安东尼·C·费兰特」,「弗雷德·谢皮西」,「米夏埃尔·艾斯」,「埃里克·布罗斯」,「提莫·贝克曼贝托夫」,「凯文·史密斯」,「帕尼勒·费舍尔·克里斯藤森」,「亚历克斯·罗斯·派瑞」,「伦尼·阿布拉罕森」,「吉姆·洛奇」,「让弗朗索瓦·波略特」,「弗雷德·欧伦·雷」,「普安农」,「蔡明亮」,「北村龙平」,「全秀一」,「蒂波尔·塔卡克斯」,「马克·罗特蒙德」,「扎克·希尓迪奇」,「杰森·伯格」,「杰弗里·沃克」,「山姆·贾巴尔斯基」,「莎拉·哈丁」,「拉里·克拉克」,「斯蒂芬·柯尼」,「布莱恩·莱温特」,「加比·德拉尔」,「全圭煥」,「罗伯特·文斯」,「理查德·隆克瑞恩」,「萨杜斯·奧沙利文」,「詹妮弗·林奇」,「格里夫·弗斯特」,「保罗·杰诺维塞」,「杰森·康纳利」,「艾德·加斯·多内利」,「泽维尔·多兰」,「迈克·门德兹」,「罗伯·威廉姆斯」,「普利亚当沙」,「罗伯·莱纳」,「詹姆斯·弗兰科」,「卡斯帕·安德瑞斯」,「艾拉·萨克斯」,「史蒂芬·C·米勒」,「彭力·云旦拿域安」,「沃夫冈·格罗斯」,「奥利弗·希施比格尔」,「麦克·鲍力施」,「吉姆·温诺斯基」,「叶天伦」,「申渊植」,「普拉部·地伐」,「迈克尔·费法」,「李宗弼」,「奉万大」,「科林·塞斯」,「米凯莱·普拉奇多」,「托德·维罗」,「尼基尔·阿德瓦尼」,「朴庭凡」,「金赵光寿」,「萨维里奥·科斯坦佐」,「尼古拉斯·雷顿」,「乔·斯万博格」,「丹·里德」,「蒂姆·费威尔」,「伊维斯·西蒙尼奥」,「豪尔赫·托雷格罗萨」,「查理·帕尔默」,「汉内斯·赫尔姆」,「曼努埃尔·德·奥利维拉」,「金曲」,「金宣」,「詹姆斯·哈维斯」,「罗伯特·扬」,「小丹尼尔」,「比尔·安德森」]

Cluster 1中的电影总数: 585

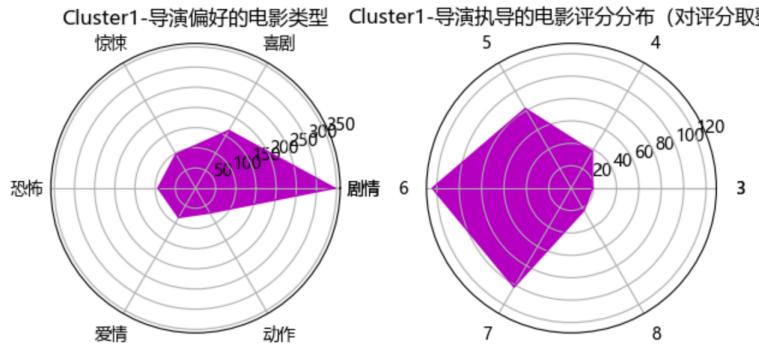


图 26: 导演 cluster 1

Cluster 1 中的导演绝大部分是西方导演，此外还有极少数的日本、韩国等亚洲导演。

从类型分布雷达图中，可以得知以西方人为主导的这部分导演的偏好，585 部电影中有超过一半是剧情片，此外喜剧、惊悚、恐怖等题材也比较受这些导演的欢迎。

这部电影的评分集中在 5-7 之间，比之前的 Cluster 3 要差一些，算是平均水平，原因可能是这一类导演只是以地区做了主要的共同特征，并没有按照特定的题材偏好或者评分水平进行分类，所以导演水平自然良莠不齐，最终的结果也比较平均。

4.4 可视化分析 2 个不同类型演员的特点

此处选择对演员的 cluster 2 和 cluster 4 进行可视化，对演员的数据分析方式与对导演的方式几乎一致。

首先分析演员的Cluster 2:

Cluster 2中的前200位演员名单:

[寺岛忍', '浅野忠信', '渡辺谦', '饭丰万里江', '忍成修吾', '谷村美月', '染谷将太', '绫瀬遥', '西島秀俊', '成田凌', '山田凉介', '芳根京子', '莲佛美沙子', '妻夫木聰', '菊地凛子', '大泽隆夫', '贺来贤人', '广瀬爱丽丝', '岩田刚典', '高岛政宏', '芦名星', '玉城蒂娜', '余贵美子', '三浦友和', '中村优子', '池田纯矢', '市道真央', '夏帆', '苍井优', '古关宽广', '泷正则', '大仓孝二', '佐藤健', '柳乐优弥', '桥本爱', '岩城滉一', '千叶雄大', '滨边美波', '中井贵一', '佐佐木藏之介', '广末凉子', '广瀬铃', '堤真一', '吉泽亮', '高畑充希', '山崎贤人', '斋藤工', '松隆子', '福山雅治', '神木隆之介', '菅田将晖', '小松菜奈', '荣仓奈奈', '山本美月', '二阶堂富美', '高杉真宙', '松重丰', '前野朋哉', '竹中直人', '柄本佑', '村上虹郎', '森山未来', '前田吟', '吉冈秀隆', '桥爪功', '小林稔侍', '长泽雅美', '东出昌大', '小日向文世', '竹内结子', '三浦春马', '江口洋介', '小栗旬', '植木玲弥', '大杉涟', '浅川梨奈', '秋山莉奈', '樱田通', '市原隼人', '中村友理', '津田宽治', '五十嵐信次郎', '池松壮亮', '村上淳', '西田尚美', '佐野史郎', '永濑正敏', '门胁麦', '松浦祐也', '筱原友希子', '横田雄司', '吹越満', '奥田瑛二', '福田麻由子', '本田翼', '平泉成', '松山研一', '堺雅人', '田中圭', '吉田钢太郎', '林遣都', '内田理央', '真岛秀和', '大塚宁宁', '志尊淳', '泽村一树', '前田敦子', '仲野太贺', '户田惠梨香', '佐藤浩市', '北村有起哉', '中村伦也', '矢本悠马', '塙地武雅', '池田铁洋', '古川雄辉', '柄本时生', '松坂桃李', '阿部宽', '安田显', '和田聰宏', '藤冈靛', '井浦新', '清水寻也', '长谷川京子', '寺胁康文', '伊武雅刀', '香里奈', '阿部纯子', '有村架纯', '山下莉绪', '田中丽奈', '瑛太', '绪形直人', '大地康雄', '中村苍', '高桥和也', '石桥静河', '中村贺津雄', '丰川悦司', '和久井映见', '光石研', '恒松祐里', '岸井雪乃', '峯田和伸', '中条彩未', '田边诚一', '青木崇高', '长谷川博己', '深水元基', '筒井真理子', '高良健吾', '吉冈里帆', '石田百合子', '小池栄子', '木村佳乃', '役所广司', '池田依来沙', '塙本高史', '桐山涟', '小宫有纱', '多部未华子', '萩原利久', '滨田麻里', '藤原季节', '貫地谷栞', '原田泰造', '山谷花纯', '木村了', '绫野刚', '杉咲花', '柄本明', '片冈礼子', '黒泽明日香', '根岸季衣', '松平健', '黒島結菜', '井上真央', '音尾琢真', '稻垣吾郎', '田中泯', '川荣李奈', '国村隼', '安藤政信', '伊势谷友介', '麻生久美子', '磨赤儿', '久保田悠来', '渡辺哲', '岸部一徳', '大东骏介', '渡辺真起子', '宇野祥平', '涩川清彦', '奥野瑛太', '板谷由夏', '浅田美代子', '树木希林', '山崎努', '小市慢太郎']

Cluster 2中的电影总数: 6956

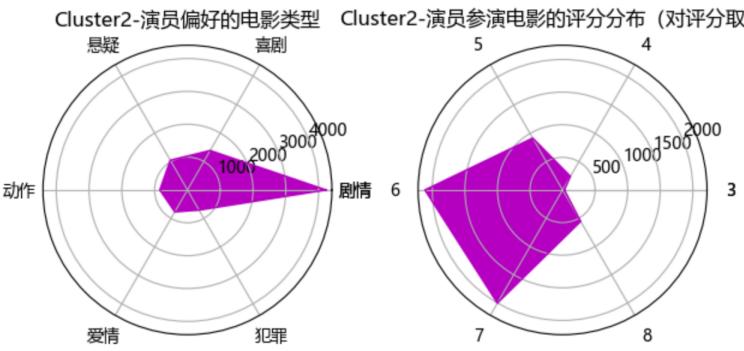


图 27: 演员 cluster 2

可以看出, Cluster 2 中的演员全部为日本演员, 偏好参演的电影类型有剧情、喜剧、悬疑、动作等。

这部电影的评分集中在 6-7 分, 稍好于平均水平。

下面对演员的Cluster 4进行分析：

Cluster 4中的前200位演员名单：

「王博」,「王姬」,「沈丹萍」,「吴孟达」,「曾志伟」,「杜海涛」,「王丽坤」,「舒淇」,「佟大为」,「姜武」,「夏雨」,「刘德华」,「潘粤明」,「托尼·贾」,「徐峰」,「张静初」,「林依晨」,「汤姆·赫兰德」,「阿萨·巴特菲尔德」,「高圆圆」,「姚星彤」,「陈观泰」,「章子怡」,「胡明」,「吴京」,「刘涛」,「张嘉译」,「邹兆龙」,「黄一飞」,「张亮」,「李易祥」,「李菁」,「王德顺」,「刘仪伟」,「赵英俊」,「赵薇」,「陈冲」,「谢霆锋」,「吴彦祖」,「李灿森」,「古天乐」,「刘烨」,「成龙」,「范冰冰」,「梁朝伟」,「萨姆·科罗纳多」,「汤唯」,「张涵予」,「王婷」,「宋洋」,「葛优」,「王学兵」,「张榕容」,「罗翔」,「郑伊健」,「王祖蓝」,「凡妮莎·柯比」,「沈腾」,「董立范」,「王宝强」,「张子枫」,「高捷」,「景甜」,「刘嘉玲」,「包贝尔」,「李成敏」,「许君聪」,「廖蔚蔚」,「吴亦凡」,「陈伟霆」,「郭采洁」,「山姆·克拉弗林」,「郑恺」,「李晨」,「张钧甯」,「李梦」,「李光洁」,「金世佳」,「刘陆」,「李勤勤」,「唐文龙」,「苑琼丹」,「魏小欢」,「余少群」,「多布杰」,「赵毅」,「周浩东」,「吴君如」,「周杰伦」,「王珞丹」,「欧豪」,「甄子丹」,「李连杰」,「郑佩佩」,「卡琳娜·卡普尔」,「矢野浩二」,「黄海冰」,「曾江」,「颜卓灵」,「蔡瀚亿」,「陈小春」,「谢天华」,「保剑锋」,「余男」,「黄奕」,「闫妮」,「斯琴高娃」,「陈威旭」,「何超仪」,「黄渤」,「郑秀文」,「王双宝」,「杨子姗」,「任达华」,「孙红雷」,「谢依霖」,「马丽」,「黄小蕾」,「黄晓明」,「张学友」,「王学圻」,「吴毅将」,「刘洋」,「张震」,「春夏」,「马思纯」,「彭于晏」,「范伟」,「秦沛」,「白冰」,「陈之辉」,「戴立忍」,「黄健玮」,「何育骏」,「房祖名」,「王凯」,「钱嘉乐」,「蒋雯丽」,「徐帆」,「陈奕迅」,「林雨申」,「加布里埃尔·伯恩」,「王景春」,「曾国祥」,「王敏奕」,「杨洋」,「唐嫣」,「赵文卓」,「洪金宝」,「石兆琪」,「于荣光」,「丁海峰」,「刘青云」,「谢君豪」,「姜皓文」,「洪天明」,「黄德斌」,「梁家仁」,「郑人硕」,「连凯」,「姜潮」,「吴刚」,「唐宁」,「邓家佳」,「黄觉」,「潘斌龙」,「陈意涵」,「程媛媛」,「金巧巧」,「果靖霖」,「汤镇业」,「王璐瑶」,「谢孟伟」,「郭富城」,「付赫安琪」,「王千源」,「谭卓」,「黄轩」,「王子文」,「王砚辉」,「李璐兵」,「倪慕斯」,「乔乔」,「刘青」,「林雪」,「白鹿杨」,「郝劭文」,「谢苗」,「石班瑜」,「惠英红」,「陈静」,「张继聪」,「余安安」,「张艺兴」,「蔡卓妍」,「周柏豪」,「卫诗雅」,「马里奥·毛瑞尔」,「何佩瑜」]

Cluster 4中的电影总数： 9030

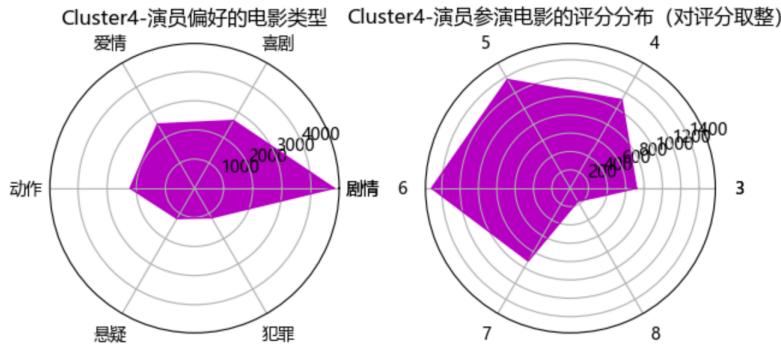


图 28: 演员 cluster 4

可以看出，Cluster 4 中的演员以我国演员为主，包括了内地和港澳台的各知名演员。

有趣的是，我国演员参演电影的类型偏好和上面 Cluster 2 中的日本演员高度相似，都是剧情、喜剧、动作、爱情等，然而评分却集中在 4-6 分之间，明显要低于 Cluster 2，这说明我国电影在同样题材的情况下，和日本电影的水平相比还有一定的差距。