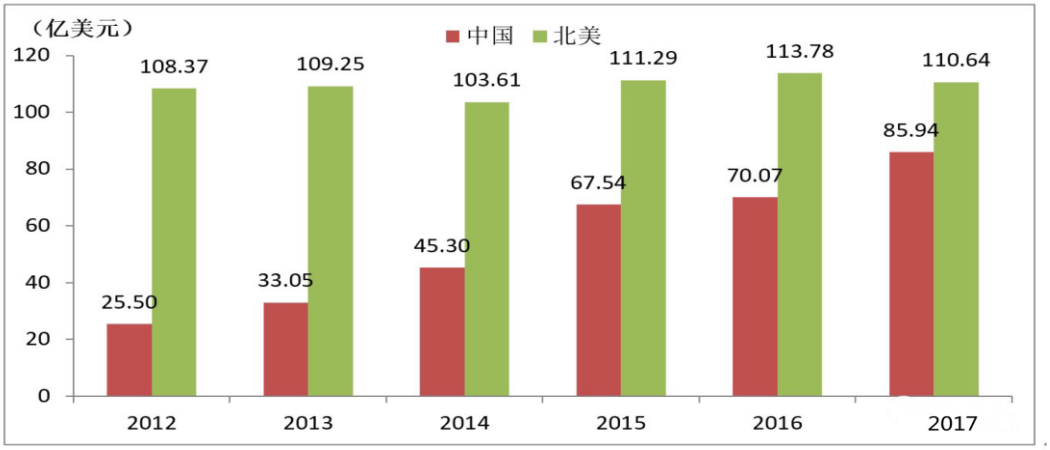
探究影响电影评论数及相似度推荐的变量

摘要：本文通过用 18 项指标建立的 XGBoost 的模型，探索影响评论数的因素。用电影的剧情梗概内容来定义电影，通过输入电影的剧情信息，进行分词统计词频，进而定义词向量，采用 TF-IDF 算法对文本内容进行相似度的计算；对电影类型、电影关键词、电影导演和电影前三名主演的姓名标签，用 sklearn 里的 CountVectorizer 统计词频，选用 sklearn 里的 cosine_similarity 计算相似度。将基于电影内容做一个相似电影的推荐，力图深度挖掘电影的文本信息。

一、背景介绍

在全球新兴电影市场特别是中国等亚洲电影市场的带动下，全球电影市场近年来保持着稳定增长的基本态势。根据统计数据显示，2017 年全球电影票房达 406 亿美元，同比 2016 年增长约 5%，创下历史新高。近十年来全球电影票房持续保持增长态势，2008-2017 年全球电影票房复合增长率为 4.3%。作为全球第二大电影市场——中国正以 30%左右的平均增长速度领跑全球市场，2017 年票房为 559.11 亿元人民币，票房同比增长 13.45%。



中国电影市场的高速发展不仅仅源于电影产业技术的日益提高，也受益于中国新一代消费群体对电影等娱乐产业的需求日益提高。而广大消费者对电影娱乐的需求也不仅仅体现在电影票房收入方面，更多地人们的观影需求会体现在互联网视频网站等网络资源分享网站上。因而，基于不同类型消费群体的不同观影需求，我们开展了关于影响电影评论数因素以及相似电影推荐的研究。

二、数据说明

本报告使用的是 TMDB5000 电影数据集，共 4803 个样本。在数据预处理阶段，上映年份只有一个缺失值，电影时长有两个缺失值，均可通过网络查询真实值进行填充。提取原始数据中上映日期的年份作为新变量，命名为：上映年份；选取演职人员前三位作为新的指标，命名为：主演。采用同义词替换方法处理关键词信息，并过滤掉出现次数小于 5 次或出现在不同电影的电影个数少于三个的关键词。最后将电影类型、导演、关键词和主演这四个文本变量转化为 one-hot 向量用于相似度的度量。本文选取 7 个因变量分析电影评论数影响因素，具体的变量说明如表 2-1 所示。

表 2-1：数据变量说明表

变量类型		变量名	详细说明	取值范围
因变量		评论数	连续变量	[0,13752]
自变量	外部因素	上映年份	连续变量	[1929,2017]
		热度	连续变量	[0,74.82]
		收入	连续变量	[0,253625427]
		评分	连续变量	[0,10]
	内部因素	预算	连续变量	[0,380000000]
		电影时长	连续变量	[0,276]
		类型	分类变量	冒险、动作等 20 种

三、描述分析

在对电影评论数的影响因素进行模型探究之前，首先对各变量进行描述性分析，为后续研究做铺垫。

（一）因变量：评论数

在本案例中，评论数最大值为 13752，对应的电影是《盗梦空间》，《黑暗骑士》、《复仇者联盟》、《阿凡达》紧随其后，分别有着 12002、11776 和 11800 个评论。有 63 部电影的评论数为 0，如《黑水船运公司》。

通过电影评论数分布直方图（图 3.1）可以看到，各部电影评论数差异较大。评论数的均值为 690，中位数为 235，中位数与均值差异较大，符合基本认知，即存在少数优质电影评论数较高，拉高了整体的平均值。

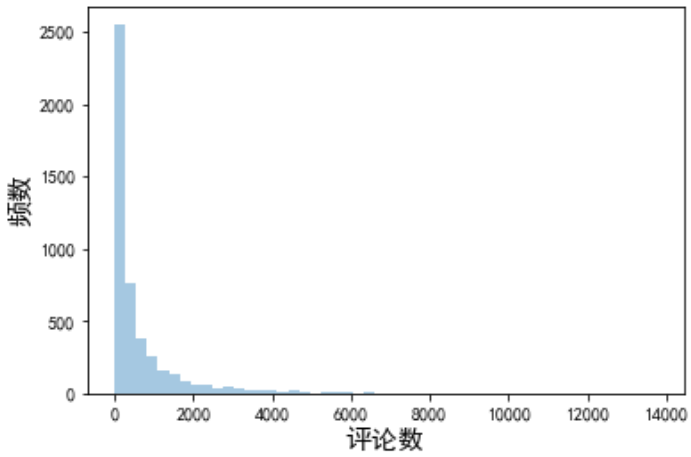


图 3.1 电影评论数分布直方图

（二）自变量：外部因素

外部因素包括上映年份、热度、收入和评分四个变量。

上映年份被分为 1940 以前、1940-1960、1960-1980、1980-2000 和 2000 以后，各个时间段电影数量逐渐增多，评论数均值逐渐增大但增幅不大，实际研究意义不大故忽略。

通过电影评论数与热度和收入的散点图（图 3.2）可以看到，热度越高，电影评论数越高，同理收入越高，电影评论数越高。

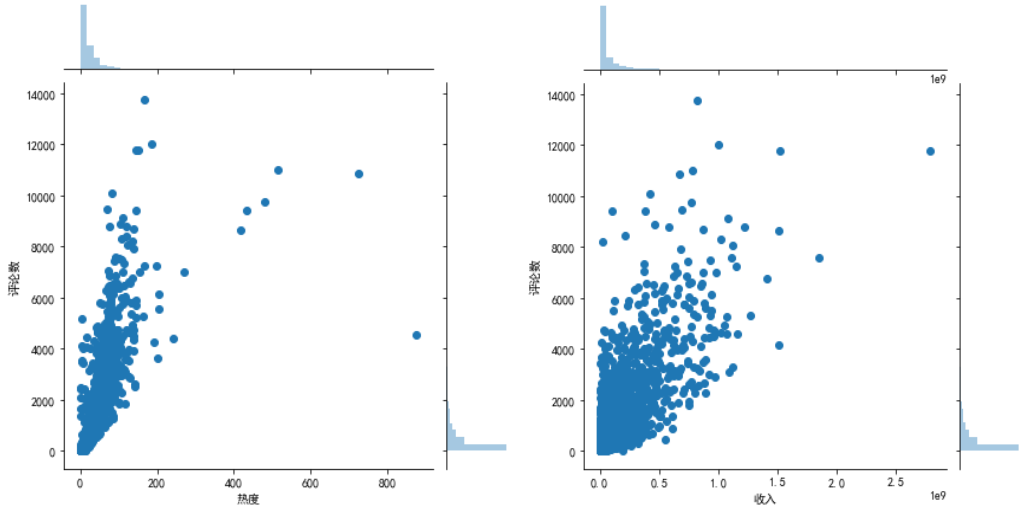


图 3.2 电影评论数与热度和收入的散点图

通过电影评论数与评分的散点图（图 3.3）可以看到，高评论数的电影，评分不会低，集中于 6 分至 8 分。

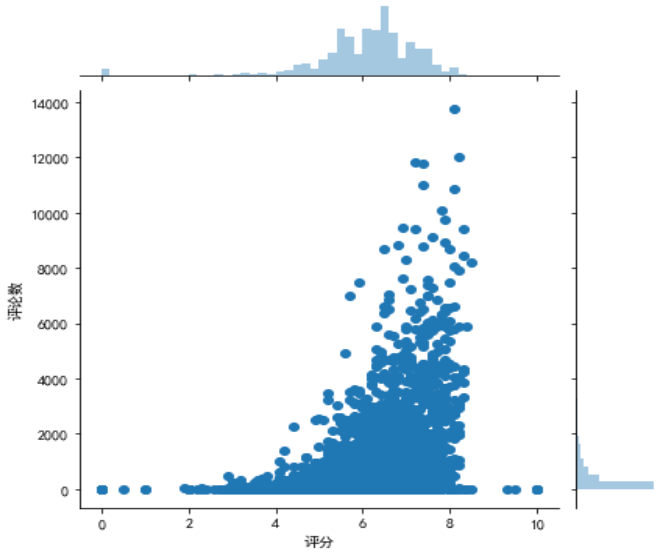


图 3.3 电影评论数与评分的散点图

（三）自变量：内部因素

内部因素预算、电影时长和类型三个变量。

通过散点图发现预算和电影评论数的关系不是太明显，就先不进一步探究它了。

通过电影评论数与电影时长的散点图（图 3.4）可以看到，电影时长小于 75 分钟或大于 200 分钟时，电影评论数均较小，电影时长适中更容易吸引观影者评论，具体评论数多少还要受其他因素影响。

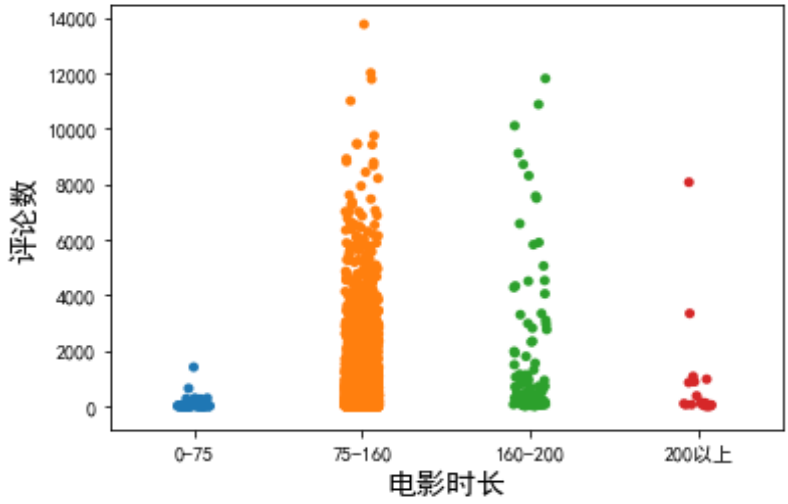


图 3.4 评论数与电影时长的散点图

通过电影评论数与电影类型的箱线图（图 3.5）可以看到，在 20 个电影类型中，按照均值大小划分，动画类、虚幻类、冒险类和科幻类评论数最高，而纪录片类、外国类和电视电影类评论数最低。宽度代表电影类型的样本个数，戏剧类、戏剧类、惊悚类和动作类样本个数最多，其中戏剧类样本个数为 2297，而纪录片类、西方类、外国类和电视电影类样本个数最少，电视电影只包含 8 个样本。

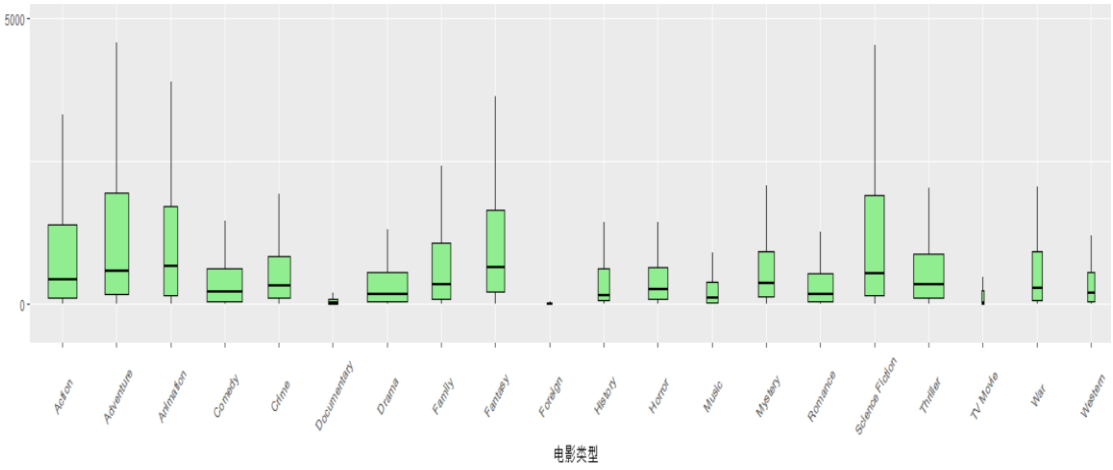


图 3.5 电影评论数与电影类型的箱线图

（四）自变量：导演与演员

众多观影者会关注一部电影的导演，导演的代表作越多，其新执导的电影更容易有超高的票房。执导电影数最多的导演为 Steven Spielberg（史蒂文·斯皮尔伯格），导演了 26 部电影，代表作为《辛德勒的名单》、《拯救大兵瑞恩》等，获得过美国电影电视金球奖终身成就奖和 2018 帝国电影奖终身成就奖。

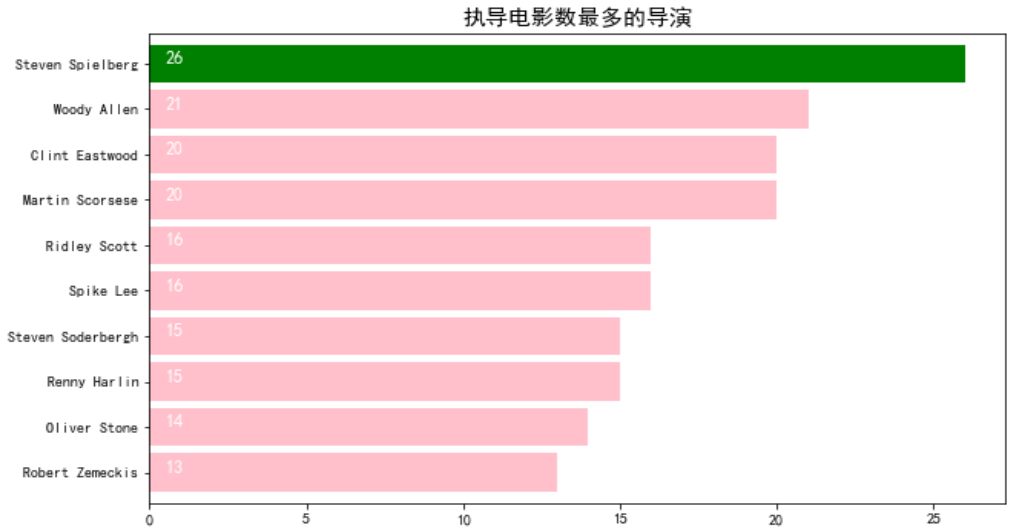


图 3.6 执导电影数排名前 10 的导演

演员同样也是一部电影的重要组成部分，Robert De Niro（罗伯特·德尼罗）是出演电影次数最多的演员，共出演 46 次，代表作品有《教父 2》、《出租车司机》、《美国往事》、《愤怒的公牛》等。

四、模型建立

（一）XGBOOST 探究影响因素

该数据集中的可选变量有预算、热度、年份、收入、时长、评分六个数值型变量，考虑到电影类型、演员和导演信息也很有可能是评论数的影响因素，一些忠实影迷通常会热衷于他们所喜爱的演员所拍的戏，也会积极地进行评论。不同类型的影片数量有者较大的差别，在这里我们选取了影片数量最多的前十大电影类型，分别是 drama(2292 部)、comedy(1721 部)、thriller(1273 部)、action(1153 部)、romance(892 部)、adventure(788 部)、crime(695 部)、science fiction(534 部)、horror(519 部)、family(512 部)，将其作为 10 个哑变量，同时基于演员出演电影次数的分布情况和导演执导电影的情况，找出了出演次数最多的前五位演员和执导电影数排名前十一的导演，将影片中演员是否为前五大演员和导演是否为前十一大导演分别设置成哑变量，加上前面的 6 个数值型变量的 10 个电影类型哑变量一共 18 个变量作为模型的 feature，以电影评论数作为 respond。由于评论数的跨度范围很大并且严重拖尾，尾部极为稀疏，在尝试用 lasso，pcr 等模型来找出影响显著的 feature 时发现误差都非常大，故最后决定使用 xgboost 进行建模分析。

首先随机取出数据集的 80%作为训练集，其余的作为测试集。默认每棵树的最大树深为 6，同时为了减小过拟合，将 min_child_weight 设置为 2，并将 eta(学习率)降为 0.2。为了估计误差，对于训练集，我们采用 leave-one-out 的方式，将其随机分为 10 份，其中的 9 份用于训练，其余的一份用做估计，得到如下训练误差和估计误差的变化情况：

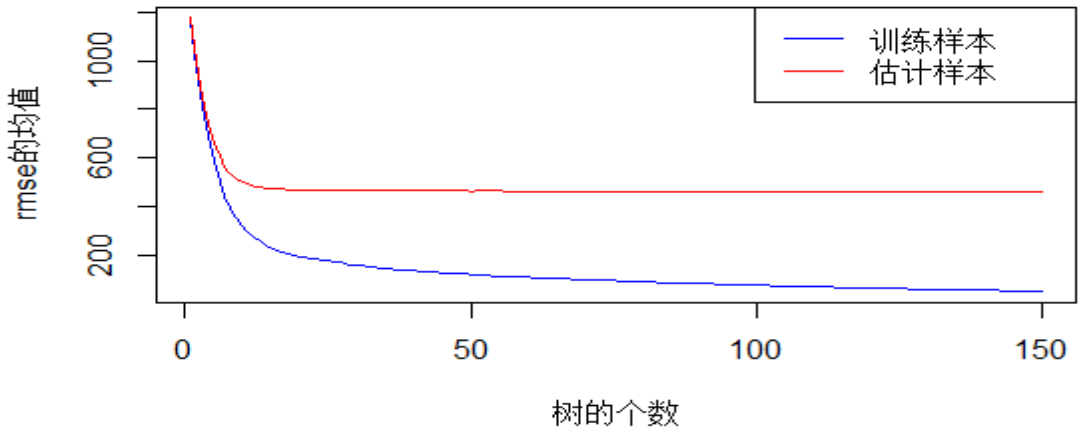


图 4.1 RMSE 均值随树个数的变化

可以看到，随着 tree 的数量不断增多，RMSE 也在不断下降，并且训练样本的 RMSE 的均值和标准差可以随着 tree 的不断增多而一直减小，当 tree 的数量为 150 时，train 上的平均 RMSE 大致为 50 左右，但估计的 RMSE 的均值最终只会稳定在 450 左右，标准差最终也是稳定在 40 左右。

最终，基于主要考虑测试集误差的原则，只选取了 30 棵树（此时估计误差的波动已经比较稳定），其余参数和上面一致，以整个训练集来拟合模型，得到测试集上的 RMSE 为 443（这和训练集上的误差估计也比较接近），整个训练集上的 RMSE 为 167（已经比 SVR 有了较多的改善）。根据变量的重要性矩阵，我们得到如下变量重要性信息：

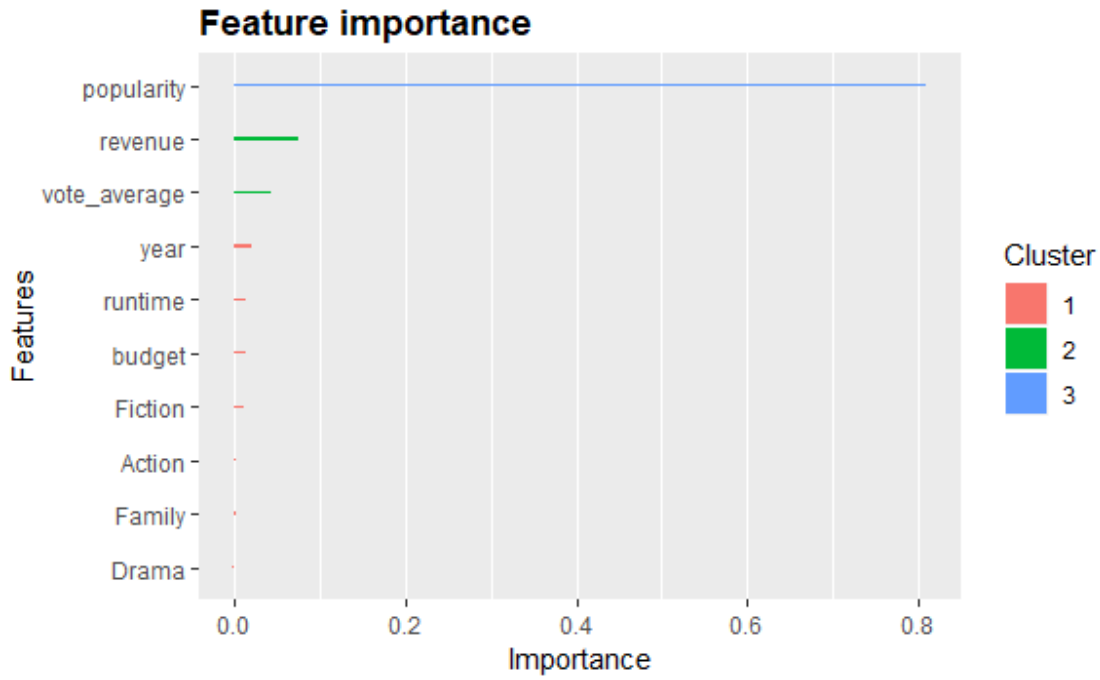


图 4.2 特征重要性

从图 3 中可以看到，对估计电影的评论数最重要的变量是 popularity，这与我们的认识也是比较符合的，即一部电影的受欢迎程度越高，评论数也会越多，两者之间有者较强的关联性。因此，作为制片人，如果从评论数的角度出发，可以考虑在电影中多添加一些当下流行的元素，选择一些比较热门的题材等等。次重要的变量为 revenue 和 vote_average，而 year、runtime 和 budget 对评论数的影响则比较微弱。在电影类型中，我们可以看到小说类的电影评论数于其它类型的电影的评论数差别最大，戏剧类的电影虽然数量最多，但在评论数上和其它类型的电影差别并不大。导演的信息虽然也是 tree 生成过程中使用的变量，但在该模型中对评论数的影响并不大，而演员的信息甚至没有出现在 tree 生成的过程中，这可能是因为在处理这两个变量时没有充分提取到信息（演员只提取了出演次数最多的前五个演员，导演只提取了执演次数最多的前 11 位）。

（二）深度挖掘数据集中文本信息

通过初步观察和筛选文本数据，我们发现每部电影对应的文字内容包括电影剧情梗概和电影基本信息这两方面，变量“overview”展示的信息为电影的剧情梗概；变量“genres”为电影类型标签，且具有多层分类特征，变量“keywords”为电影关键词标签，变量“actor”为电影前三主演的姓名标签，变量“director” 为电影导演的姓名标签。结合自身对浏览电影信息网站的行为习惯，我们认为电影剧情梗概的相似度越高，将其推荐给相应阅读者能够投其所好的概率可能越大；电影基本信息，即电影所属类型、电影关键词、电影主要演职人员的相似度越高，将其推荐给相应阅读者能够符合其个性化需求的概率也很可能越大。因此，基于这样的推测，我们给出两种方案来对数据集中的电影进行相似推荐。

第一种方案为用电影的剧情梗概内容来定义电影，通过输入电影的剧情信息，进行分词统计词频，进而定义词向量，进行相似度的计算。第二种方案是用电影的类型、关键词、主演、导演等已经提取出的文字标签来定义电影，统计词频组成词向量，然后进行相似度的计算。本研究通过把文本转换成词向量，比较电影之间的相似度就转换成了度量词向量之间的距离，本研究选用 cosine 余弦来度量词向量之间夹角的大小。最后，输入数据集中的某一或某几个电影的名称，根据两种方案分别获取相似电影的名称，并根据两种方案给出的相似电影推荐结果，大体判断哪一种相似电影推荐的效果更好。

具体方法如下：

首先，对于变量“overview”中电影剧情梗概的内容，我们采用 TF-IDF 算法对方案一中的文本内容进行处理。TF-IDF 这种统计词频的方法可以突出输入文本中独特的高频词，而对于剧情信息中普遍常用的词的权重会被消弱。这个方法对于我们数据集中电影剧情梗概的内容特性，较为适宜。通过使用 sklearn 里的 TfidfVectorizer 把原始文本转化为 TF-IDF 的特征矩阵，再通过 sklearn 里的 linear_kernel 方求 TF·IDF 矩阵和其本身转置矩阵的点积，所得的结果就是 cosine 余弦值，即我们所要度量的相似度。

对于方案二中的电影基本信息标签，我们选取已经提取好的电影类型、电影关键词、电影导演和电影前三名主演的姓名标签，此外为了提高算法最总得出的相似度，我们额外提取到每条电影信息里的制片和编剧的姓名标签。随后，将上述标签作为元信息，整合成一个长文本统计词频组成的词向量。在对标签信息进行相似度计算时我们没用到 TF-IDF 对常用词过滤，原因在于对标签信息我们认为都是重要的，且不存在普遍常用词，因此直接选用 sklearn 里的 CountVectorizer 统计词频，选用 sklearn 里的 cosine_similarity 计算相似度。

在完成相似度的计算之后，我们需要定义获取相似电影的方法。在这里我们采用输入电影名称，获取 id 号的方法，从相似度矩阵中提取输入名字的电影对应的列，进而获取相似度最高的前十部电影（除自身外）并返回前十部电影的名称，分别按照两种方案进行相似电影的查找。

从下面表中展示的相似电影推荐结果中可以得到，无论是超级英雄系列电影、经典剧情类电影还是动画类电影，根据标签信息，即根据电影类型、关键词、主要演职人员等信息得到的相似推荐效果更好，且系列电影的推荐效果最好，这样的结果也与我们预想的相符，系列电影由于演职人员较为固定，电影关键词类似而得到较好的推荐效果，而剧情类电影的推荐结果相对差一些，原因可能是由于剧情类电影除了电影类型，关键词的标签可能相似度高之外，其他标签信息的相似度不够高。而根据剧情梗概的相似推荐得到的结果不尽如人意，原因也可能主要在于对剧情文本的实质性内容提取不够，导致相似度的计算不能反映电影之间的真实相关性。

表 4.1 根据剧情梗概的相似推荐结果

根据剧情梗概的相似推荐			
输入电影名称	The Avengers	The Shawshank Redemption	Frozen
输入电影类型	系列电影	剧情类电影	动画片电影
推荐相似电影			
1	Avengers: Age of Ultron	Civil Brand	Stardust
2	Plastic	Prison	Ida
3	Timecop	Escape Plan	Leap Year
4	This Thing of Ours	Fortress	The Promise
5	Thank You for Smoking	Penitentiary	Splash
6	The Corruptor	The 40 Year Old Virgin	Two Girls and a Guy
7	Wall Street: Money Never Sleeps	Fatal Attraction	The Prince of Tides
8	Team America: World Police	A Christmas Story	Forrest Gump
9	The Fountain	The Longest Yard	Royal Kill
10	Snowpiercer	Toy Story 3	Black Snake Moan

表 4.2 根据标签信息的相似推荐结果

根据标签信息的相似推荐			
输入电影名称	The Avengers	The Shawshank Redemption	Frozen
输入电影类型	系列电影	剧情类电影	动画片电影
推荐相似电影			
1	Avengers: Age of Ultron	Amidst the Devil's Wings	Valiant
2	Captain America: Civil War	Escape from Alcatraz	Return to Never Land
3	Captain America: The Winter Soldier	Brooklyn's Finest	Pocahontas
4	Iron Man 2	Witness	Tangled
5	Iron Man	The Green Mile	Atlantis: The Lost Empire
6	Ant-Man	Cradle Will Rock	Alpha and Omega: The Legend of the Saw Tooth Cave
7	Captain America: The First Avenger	Dark Blue	Aladdin
8	Thor: The Dark World	Dead Man Walking	Enchanted
9	Iron Man 3	The Bad Lieutenant: Port of Call - New Orleans	The Smurfs 2
10	X-Men Origins: Wolverine	The Yards	Why I Did (Not) Eat My Father

五、结论与建议

通过上述一系列的分析，我们得到了以下结论：

1、影响电影评论数的主要因素有：(1)外部因素：热度、收入和评分均对电影评论数有影响；(2)内部因素：电影的预算、电影时长的适中更倾向于有较高的评论数。(3)还有电影本身的演员导演等因素。电影的导演越知名执导的电影越多，越有多的评论数；演员越知名，评论数也越多，这也可能是未来电影发展的一个趋势。

2、影响电影推荐的主要因素有：电影类型、关键词、主要演职人员等信息得到的相似推荐效果更好，且系列电影的推荐效果最好。

电影厂商可以根据以上影响因素，制作比较优秀的电影，并进行有效的推荐。

第三次分工：

数据预处理：王嬰、张媛媛

描述分析：张云华

模型建立：柳素问，张继丹

报告整合：栗书敬

报告修改：王嬰