摘要

随着线上线下融合的旅游模式的蓬勃发展，在线查看评论已经成为游客出游前的必要操作，消费者的反馈也在一定程度上影响着商家经营策略和方向。然而目前基于消费的评价和打分模式却只能作为参考而并非直接依据，究其原因，一方面是网上有不少系统自动默认的好评、用户为了领券而凑字数的评价，这些并无实际参考意义，甚至有恶意抹黑竞争对手的虚假评论，如何筛选出有价值的评论进而消费对于用户来说是个难题；另一方面，用户对景区及酒店的综合评分直接基于用户评分计算，其中就有不少用户来不及评分而被默认满分，导致该计算结果明显虚高，无法给用户提供良好的参考，也不能给景区及酒店提供决策支持。此外，如何利用用户评论凸显酒店景区的特色，从而提供多样化服务，满足用户个性化需求对于未来旅游业的发展具有重大意义，但多数旅游和生活类 App 均未涉及。

为此，本文拟对以上三个问题进行研究。1）提出了一个基于统计学和机器学习算法的评分系统。将游客评论进行五维度拆分，再基于中文词汇的正反感情权值进行评分计算，创新性地使用了多项式朴素贝叶斯算法以及岭回归算法，实验结果证明了本系统评分具有更好的参考价值。2）通过对用户评论数据进行深度挖掘和分析，建立了一个保持用户原意的有效性评价分析模型。首先采用 Viterbi 中文分词算法进行数据预处理，然后对有效评论的热词进行了加权分析，最后筛选出用户的有效评价，较之传统的词袋法，其精确度更高。3）最后通过有效评论的相似度计算突出各酒店和景区的特色，满足用户个性化需求。从不同层级的酒店及景区中挑选出 3 家进行特色分析实验可知，本文提出的方法能彰显出酒店景区的特色。

关键词：用户评论，多项式朴素贝叶斯，岭回归，Viterbi 算法，游客满意度

Abstract

With the booming development of online and offline tourism mode, online review has become a necessary operation for tourists before traveling, and consumer feedback also affects business strategy and direction to a certain extent. However, the current evaluation and scoring mode based on consumption can only be used as a reference rather than a direct basis. The reason is that, on the one hand, there are many online systems that automatically default to positive comments and users' evaluation of the number of words in order to get coupons, which have no practical reference significance and even have false comments that maliciously smear competitors. How to filter out valuable reviews and then consume them is a difficult problem for users; On the other hand, users' comprehensive ratings of scenic spots and hotels are directly calculated based on user ratings, among which many users are not able to score and are given full marks by default. As a result, the calculation results are obviously inflated, which cannot provide a good reference for users and cannot provide decision support for scenic spots and hotels. In addition, it is of great significance for the future development of tourism industry to make use of user reviews to highlight the characteristics of hotel scenic spots, provide diversified services and meet the personalized needs of users, but most tourism and lifestyle APPs do not cover it.

Therefore, this paper intends to study the above three problems. 1) A scoring system based on statistics and machine learning algorithms is proposed. The tourist comments are divided into five dimensions, and then the rating is calculated based on the positive and negative sentiment weights of Chinese words. The polynomial Naive Bayes algorithm and ridge regression algorithm are innovatively used. The experimental results prove that this system has a better reference value. 2) Through in-depth mining and analysis of user review data, a validity evaluation and analysis model is established to maintain the original intention of users. First, the Viterbi Chinese word segmentation algorithm is used for data preprocessing, and then the weighted analysis is carried out on the hot words of valid comments. Finally, the effective comments of users are screened out. Compared with the traditional word bag

method, the accuracy is higher. 3) Finally, through the similarity calculation of effective reviews, the characteristics of each hotel and scenic spot are highlighted to meet the personalized needs of users. Three hotels and scenic spots at different levels were selected to carry out characteristic analysis experiments, and the method proposed in this paper can highlight the characteristics of hotel scenic spots.

Keywords: User comments, Polynomial Naive Bayes, Ridge regression, Viterbi algorithm, Tourist satisfaction

目录

一、绪论

1.1 背景

随着信息技术和旅游经济的发展，关于旅游景区及酒店的网评文本数量不断增长，但部分商家为了吸引游客，便开始进行刷好评或者给竞争对手恶意差评等操作，例如 2016 年央视 315 晚会曝光了网络刷单行为在短时间内获得好评、快速提升店铺等级的现象[1]，还有一些用户因为某次住宿或游玩感受被冷落，而自己带领家人朋友恶意诋毁商家。一方面，各旅游相关企业希望有效挖掘出用户真正的需求，增强用户黏度。另一方面，在用户决策过程中，越来越多的游客习惯于出游前关注社交网络上评论文本，在线评论也以 77.5%的比例成为消费者决策时最为关注的因素[2]。为此，筛选出用户有效评论成为一个极具实用价值的研究问题。

此外，由于用户在进行消费决策时主要参考商家的评分情况，但是在实际情况中，商家的评分差距较小，用户无法快速找到心仪的酒店或景区。并且由于刷分等现象的存在，用户对于差评的信任度明显高于好评，游客在旅游目的地的消费和体验过程中也会存在网络负面口碑，因此，结合网评文本及评分进行综合评价分析使得评论酒店及景区更为客观，对用户来说可信度更高。有学者指出，当可选择项超载时不但会导致消费者延迟消费决策甚至不做消费决策从而直接影响销售业绩，而且会使消费者失去信心，消费满意度降低，并产生不良消费体验[3]。因此，综合特色分析成为提升景区及酒店等地竞争优势的关键任务，对于游客的消费决策而言具有极高的应用性。

本文结合了酒店及景区的网评文本有效性分析、综合评价分析以及特色分析等进行整合、筛选、分析，为游客及各网络平台提供了有效建议和数据支撑。

1.2 研究内容和研究意义

（1）提出了一个基于统计学和机器学习算法的评分系统，精准把控游客满意度，针对性改正不足，突出优势，提升景区和酒店竞争力；

（2）建立了一个保持用户原意的有效性评价分析模型，给各地文旅主管部门提供决策支持，提升景区及酒店的美誉度；

（3）通过有效评论的相似度计算突出各酒店和景区的特色，对旅游企业科学监管、资源优化配置及市场可持续开拓具有长远而积极的作用。

二、相关研究

用户评论的有效性是指用户对商家提供的产品或服务质量综合评价的真实可靠性，有效性分析不仅能帮助商家明确产品和服务的改进方向，而且能极大降低做出正确决策的成本。因此，用户评论的有效性分析具有重要的现实意义。用户的有效性研究最早由 Kumar 等[4] 提出，用于研究推荐系统和用户评论对消费者的影响，对可能影响实验结果的混杂因素(如产品推荐、服务条款等)进行了有效过滤，但是仍然存在刷好评等评论无法有效过滤的问题。Deng 等[5]进行了自动化识别虚假评论的尝试，识别欺骗评论的精度达到了 80%。Liu 等[6]学者提出了基于用户评论标签摘要的有效性研究，通过商品的标签摘要判断用户评论的有效性，对于不具备标签摘要的服务型商品具有一定的局限性。

另一些学者对游客的满意度计算进行了分析研究，为游客提供数据参考。Liu[7]等构建 SEM 模型，借助 Amos 等统计软件对景区游客满意度的影响因子进行分析对比。Wu 等[8]基于 Word2Vec 的机器学习方法对构建计量经济模型分析酒店特征评价与用户满意度的关系，但是存在样本选择不够全面的问题。Jin[9]等学者基于 word2vec 算法和主题情感的分析方法计算用户评论的情感极性和强度，融合情感强度计算每一个商家的综合评分，克服了传统的计算方法难以对同一类型评分进行更细粒度区分的缺点。

还有一些学者进行了景区及酒店的特色分析研究，Zhang[10]等学者基于 K-Means 聚类算法对景区数据分析，通过数据结果更直观地体现了景区的特色，但是其准确度不高。Yuan[11]提出基于特征加权与密度聚类的景区信息挖掘系统的研究与实现，通过引进空间密度聚类算法和特征加权算法帮助用户更好的挑选个性化旅游景点，但是仍然存在文本类别偏斜数据集的处理效果不佳的问题。

三、数据预处理

本文采用的数据预处理方法主要包括三个步骤：数据处理，文本分词、停用词过滤，具体如图 3-1 所示。

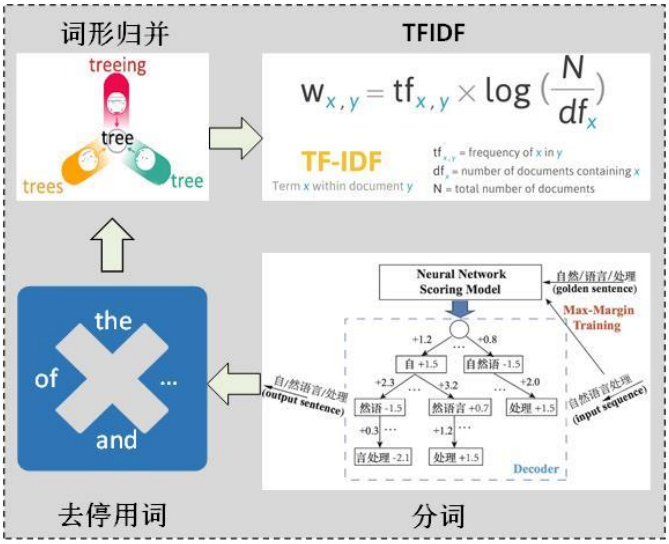

图 3-1 数据预处理图

3.1 数据处理

由于附件一的酒店与景区网评数据量较大，直接取出来的数据为脏数据，需要经过处理提取出可供模型使用的数据才能精准提取出景区及酒店目的地的热门词，并做下一步的模型分析。本文的数据预处理主要分为以下步骤：

1）中文繁体转简体。由于网友的评论存在一定数量的繁体评论，需要对繁体评论文本转换为简体，本文采用 Python 提供的开源常用库 OpenCC 进行中文繁简体转换。

2）重复值和缺失值处理。对于评论数据而言，出现重复数据一定为数据出错，对于文本数据而言，多特征的数据对模型的建立毫无意义，故对此数据进行删除。

3）机械语料压缩。由于游客、住户的评论数据中可能存在没有意义的评论，为了减少相同词组出现的评率，提高热门词提取的有效性，我们需要对单一重复与过分强调的数据进行机械预料压缩。因此，本文自定义了类似于 HashMap 的数据结构将语句进行压缩。

3.2 文本分词

由于中文词与词之间没有像英文那样固有的间隔符（空格），因此，进行中文文本挖掘时，首先应对文本进行分词，即将连续的字序列按照一定的规范重新组合成词序列的过程，分词结果的准确性对后续文本挖掘算法有着不可忽视的影响。

首先，对一个待分词的字串 S,按照从左到右的顺序取出全部候选词w_1，w_2, …,…, w_n。查找每个候选词的概率值 P(w_i)，并记下每个候选词的全部左邻词，按照式 (1)(2) 计算每个候选词的累积概率，同时比较每个候选词的最佳左邻词。如果当前词w_n是字串 S 的尾词，且累计概率 P; (w_n) 最大，则w_n就是 S 的终点词。从w_n开始，按照从右到左顺序，依次将每个词的最佳左邻词输出，即为 S 的分词结果。

$$P(w) = P(w_1, w_2, \dots, w_i) \approx P(w_1) \times P(w_2) \times \dots \times P(w_n) \quad (1)$$

$$P(w_i) = \frac{w_i \text{在语料库中的出现次数 } n}{\text{语料库中总词数 } N} \quad (2)$$

在酒店及景区网评数据处理中不难发现其存在交集型歧义，组合型歧义，混合型歧义。网评中甚至还出现了未登录词，即在字典中都没有收录过，但又确实能称为词的那些词。为了将酒店及景区网评数据按照词组的形式分开，我们通过以下两种方法进行解决。

第一步：对汉字进行标注，即由字构词(组词)；不仅考虑了文字词语出现的频率信息，同时考虑上下文语境；

第二步：采用 jieba 分词器以及 HMM 模型（隐式 Markov 模型）中的 Viterbi 中文分词算法进行分词。jieba 分词有功能丰富、提供多种编程语言实现、使用简单等优点，分词处理后的部分数据如表 3-1 所示:

表 3-1 分词处理后的部分数据

序号	名称	评论
0	H01	适合 家庭 出行
1	H01	升级 房间 延迟 退房 很赞
2	H01	几年 广州 每次 住 位置 性价比 不错 免费 升级 房间
…	…	…
25221	H50	老字号 宾馆 装修 设施 陈旧 房间 空调 卫生间 墙壁 剥落
25222	H50	还好 价格 偏高 高档 服务 设施 偏老 很大 喜欢 楼下 早茶
25223	H50	隔音 极差 睡眠 崩溃
25224	H50	位置 好找 停车

3.3 停用词过滤

进行评论文本分词后，文本中仍然含有对评论无意义的词语，比如文本中的副词、介词、语气词以及其它一些无实际意义的词语，应尽量规避此类词语对文本挖掘工作的不良影响。针对以上问题，本文主要通过以下步骤进行停用词的过滤：

①根据评论分词结果进行选取拟定合适的停用词表，具体停用词表见stopword.txt。

②通过运用 pandas 的 apply 函数检测在上一步的词语中是否存在停用词表中的词，若存在则进行删除。

酒店及景区数据经过去重，机械预料压缩，短句删除，文本分词，停用过滤词处理后的部分结果如表 3-2 所示。

表 3-2 数据处理后部分数据

序号	评论
0	住店 客人 预约 早茶 不需要 排位 早茶 价格 要贵 一倍 份量…
1	酒店 位置 不错 沙面 珠江 游玩 江景 订 豪华 江景 房 设施…
2	早餐 小贵 建议 在外 吃 早茶 周边环境 太 友好 超 喜欢 建筑…
…	…
25221	买票 提前 价格 会贵 一点 晚 几小时 价格 涨 十几块 吃饭…
25222	划 重点 图片 比例 顺序 来图 过山车 爽 翻转 挑战 里边 刺激…
25223	路上 交通 跨市 倒 公交 不建议 方式 出发 路上 交通 堵塞…
25224	好玩 好多 剧场 表演 真的 太棒了

四、问题解决思路及流程

4.1 问题一

题目：依据附件 1 中景区及酒店网评文本，按表 1 格式计算出目的地 TOP20 热门词，并保存为文件“印象词云表.xls”。

4.1.1 问题分析及解决思路

（1）问题分析

该题意从用户评论中分析出具有参考意义的热词并就出现次数进行排序，得到酒店及景区的最终词云表。该题的重难点在于中文的准确分词和统计实意词。

（2）解决思路

1）附件 1 的数据存在一定数量的繁体评论和重复评论，需要将繁体评论转换成简体以及去重；

2）为了降低词与词组合之间的耦合性，采用 jieba 分词器以及 Viterbi 中文分词算法进行分词。

3）最后采用 TF 计算词的词频，得出景区及酒店的 TOP20 热词。

具体解决思路如图 4-1 所示：

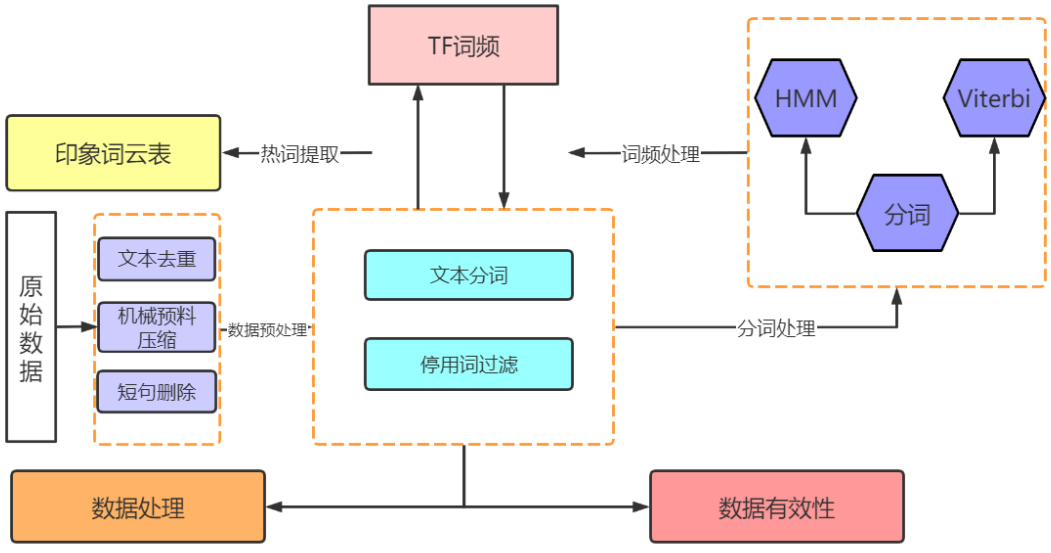

图 4-1 词云热度计算流程图

4.1.2 关键技术及应用

维特比算法（Viterbi）是一个特殊但应用最广的动态规划算法，它是针对篱

笆网络的有向图（Lattice）的最短路径问题而提出的。凡是使用隐含马尔可夫模型描述的问题都可以用维特比算法来解码，包括今天的分词、数字通信、语音识别、机器翻译、拼音转汉字等，具体原理如图 4-2 所示。

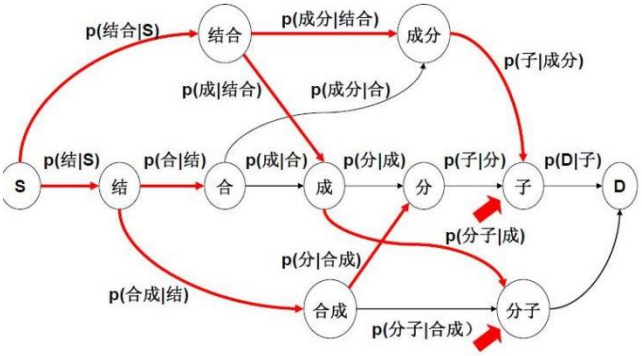

图 4-2 维特比算法原理图

首先对于输入的词序列，根据语法词典，列出每个词可能的词性候选，构成词网格，即状态空间。其次采用 Viterbi 算法搜索词网格，搜索最佳路径。最后计算相关概率时，取-log 对数形式，目的是将乘法运算变成加法运算。

jieba 分词第一步先调用函数 cut(sentence)，cut 函数会将输入句子进行解码，然后调用_cut 函数进行处理。cut 函数就是 jieba 分词中实现 HMM 模型分词的主函数。cut 函数再调用 viterbi 算法，求出输入句子的隐藏状态，然后基于隐藏状态进行分词。

4.1.3 目的地印象分析

TF 即词频，是对词数(term count)的归一化，以防止它偏向长的文件。将数据中的词语再文本中的出现次数除以文本的中词数，计算出词频的权重，具体计算如公式(3)所示。

$$TF_w = \frac{\text{在某一类中词条 w 出现的次数}}{\text{该类中所有的词条数目}} \tag{3}$$

对于某一文档d_j里的词语t_i来说，t_i的词频可以表示如公式(4)所示：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

向量化 TF 文本中各个词出现频率统计，并作为文本的特征，利用这一特性，可以很巧妙的提取出热门印象词。分词结束后对关键词进行抽取，并列出词频排列在前 20 位的词语，将这 20 个词以词云的形式展现出来，具体如图 4-3 所示。

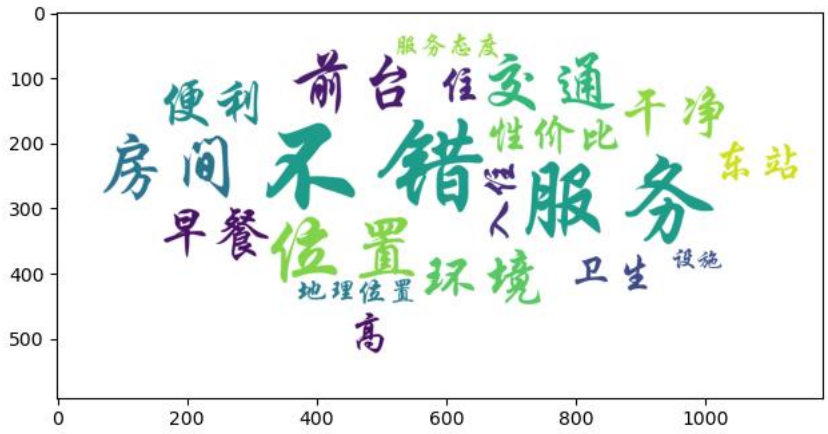

图 4-3 酒店印象词云图

4.2 问题二

题目：根据附件 1 景区及酒店网评文本及附件 2 景区及酒店得分建立合理的数学模型及相应算法，按满分为 5 分对景区及酒店的服务、位置、设施、卫生、性价比五个方面进行评分，并按照均方误差（Mean Squared Error, MSE）进行模型评价。

4.2.1 问题分析及解决思路

（1）问题分析

该题意通过评论文本数值化并结合景区及酒店的得分计算出 5 项指标的得分及总得分，进行目的地的综合评价分析。该题的重难点在于挑选出更合适的模型进行分类与提高该模型的准确性。

（2）解题思路

1）首先对附件一的数据预处理步骤，使用 TF-IDF 转化为词向量将文本数值化；

2）通过对比四种分类方法的均方误差值采用多项式朴素贝叶斯进行分类，文本分类方法的性能比较如表 4-1 所示。

表 4-1 文本分类方法结果比较

模型	性能比较	MSE
KNN	对稀有类别的预测准确率低	0.022058823529411766
极限树	最佳分叉属性随机选择	0.014946524064171123
随机森林	模型容易过拟合，训练时间长	0.013885360962566844
Multinomial NB	分类决策存在一定的偏差	0.013688168449197862

由表可知，文本分类的最佳模型为多项式朴素贝叶斯。

3）最后使用 Ridge 岭回归算法将五项得分综合得出各景区及酒店的总得分。具体解题思路如图 4-4 所示。

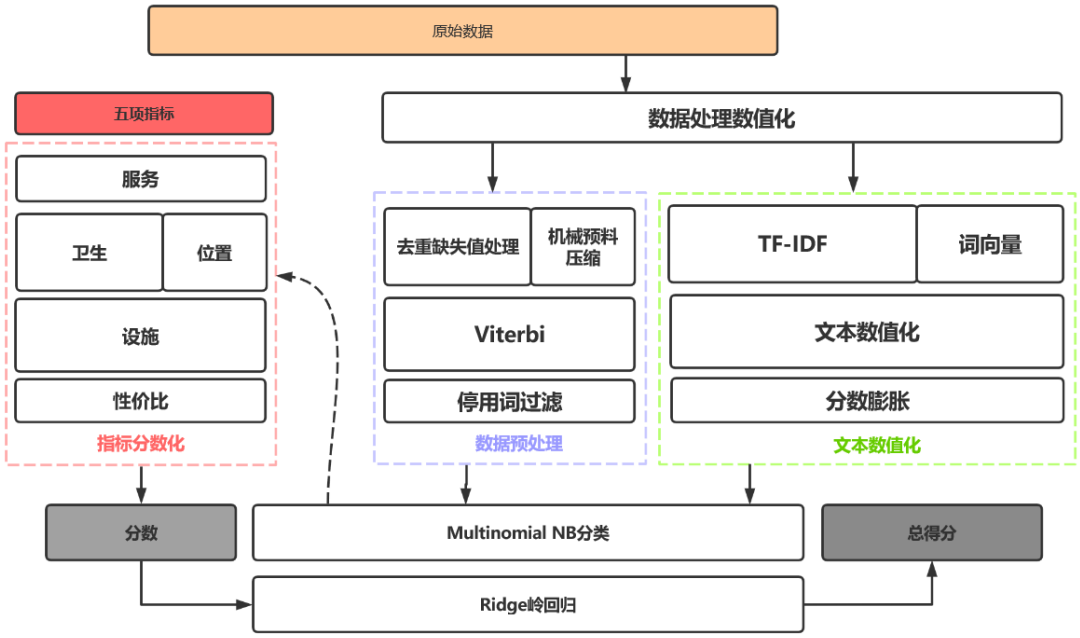

图 4-4 评分系统图

4.2.2 文本词向量

将文本进行预处理和分词后，需要将数据转换成数值矩阵，可采用的方法有 TF-IDF、One-Hot 等方法，由于 One-hot 编码未考虑词与词之间的顺序关系和语义关系（而这两点因素对于文本分类等工作具有重要影响），因此不宜采用 One-Hot 方法，故本文采用的是 TF-IDF 模型。

采用 TF-IDF 模型返回的结果是一个矩阵。矩阵每行存储的是每个词语在一个文本中的权值数据，排列顺序与列表中词语的排列顺序一致，若某个词并未出现在某个文本中，则权值为 0。TF-IDF 从词出现在文本中的频率和在文本集中的分布情况两方面很好地衡量了词的重要性。本文将文本词向量分成以下三个步骤：

1）首先计算数据词频 TF，TF 的矩阵结果如表 4-2 所示。

表 4-2 TF 矩阵结果

([[0.51973,	0.64912,	0.55545,	...	0.44654,	0.28978]
[0.52060,	0.54018,	0.39218,	...	0.36279,	0.23536]
[0.45788,	0.38008,	0.40225,	...	0.45719,	0.28683]
[0.56166,	0.43575,	0.61241,	...	0.43731,	0.35022]

2）其次，计算逆向文件频率 IDF。某一特定词语的 IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到，如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力，IDF 的输出矩阵如表 4-3 所示。

表 4-3 IDF 矩阵结果

([[16512, 7294, 4285, ... 4654, 8978]
[6514, 9051, 5587, ... 3685, 8547]
[8012, 3021, 4855, ... 6324, 11785]
[66138, 3075, 6731, ... 3629, 3321]

最后计算 TF-IDF 值，某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。采用 TF-IDF 过滤掉常见的词语，保留重要的词语，得到最终的稀疏矩阵如表 4-4 所示。

表 4-4 TF-IDF 稀疏矩阵

([[0., 0 ., 0. , ..., 0 ., 0. , 0 .],
[0. , 0. , 0. , ..., 0. , 0 ., 0 .],
[0. , 0 ., 0 ., ..., 0 ., 0 ., 0 .],
…,
[0. , 0. , 0. , ..., 0. , 0. , 0.],
[0. , 0. , 0. , ..., 0. , 0. , 0.],
[0. , 0. , 0. , ..., 0. , 0. , 0.]])

4.2.3 Multinomial NB 分类器

文本分类的目的是让分类模型能把文本映射到已存在的多个类别中的某一类，使查询速度更快、准确率更高。主要的分类方法有：朴素贝叶斯分类算法、k-近邻分类算法、Random Forest(随机森林)决策树 Design Tree 分类算法以及 Extra Tree 极限树分类算法。经过对以上四种分类方法的性能评估，朴素贝叶斯分类算法中的 Multinomial NB 多项式朴素贝叶斯分类性能更好，具体结果如图 4-5 所示。

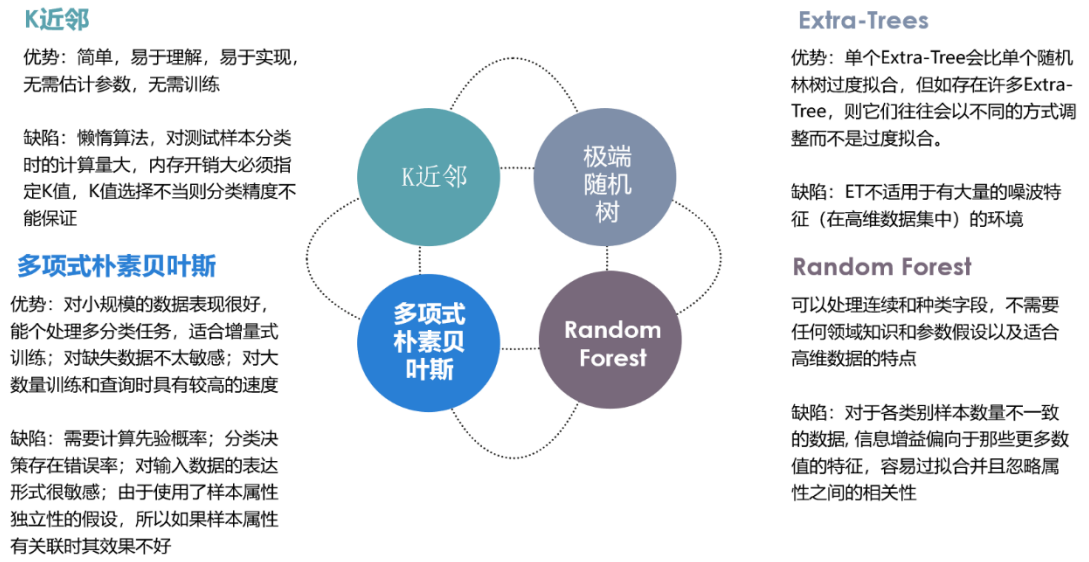

图 4-5 各分类模型性能比

朴素贝叶斯分类模型由于其简单性、高效性和有效性被广泛由于解决分类问题，基本的朴素贝叶斯常常用于解决名词类型的分类问题[12]。为了处理属性取值为连续值的文本分类问题，需要扩展朴素贝叶斯，由此产生了朴素贝叶斯文本分类模型。朴素贝叶斯文本分类模型是构建在前面介绍的向量空间模型上的，它的基本假设是给定一篇文档的类别变量的情况下，单词之间相互独立。朴素贝叶斯文本分类模型包括基于二项分布的伯努利朴素贝叶斯以及基于多项分布的多项式朴素贝叶斯、补集朴素贝叶斯和两者的结合模型。

多项式模型（Multinomial Naive Bayes，简称 MNB）将文档看作一个词袋模型，认为单词在一篇文档中出现的频率对文档类别的预测有影响。因此在计算条件概率的时候，MNB 需要统计单词出现的频率，这一点与 BNB 有显著的不同。在 MNB 模型中，一篇文本可以表示为一个向量 $d=\{w_1, w_2, w_3,\ldots, w_m\}$,$w_i \in N$，其中$w_i$表示单词在文档 d 中出现的频率。同 BNB 模型一样，MNB 也作了一个条件独立假设，这样不同条件概率估计互不影响。在待测文档 d 给定情况下，MNB 模型使用公式(5)对文档 d 进行预测：

$$c(d) = \underset{c \in C}{argmax}[log_2 p(c) + \sum_{i=1}^{m} f_i log_2 p(w_i|c)] \quad (5)$$

其中先验概率 P(c)可以通过公式(6)计算，该式同样采用了拉普拉斯估计，1 表示类别属性 C 的属性值个数。

$$p(c) = \frac{\sum_{j=1}^{n} \delta(c_j, c)+1}{n+l} \quad (6)$$

条件概率$P(w_i|c)$表示属于 c 类样本中单词w_i出现的概率，条件概率$P(w_i|c)$采用公式(7)计算：

$$p(w_i|c)=\frac{\sum_{j=1}^{n} f_{ji}\delta(c_j,c)+1}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ji}\delta(c_j,c)+m} \tag{7}$$

其中f_{ji}表示第 j 篇文档中第 i 个单词的频率，c_j表示第 j 篇文档的类标记，m 表示属性个数。

使用 Multinomial NB 进行分类，可能会损失结构化信息，而利用拉普拉斯平滑可以很巧妙的解决这一问题，Multinomial NB 分类结果如下表所示。

表 4-5 指标得分 MSE 计算结果

数据项	服务	位置	设施	卫生	性价比
酒店	0.014	0.014	0.025	0.016	0.008
景区	0.098	0.147	0.196	0.109	0.107

4.2.3 Ridge 岭回归

使用 Multinomial NB 进行分类得到五项评分后，最后采用 Ridge 岭回归输入景区及酒店的服务、位置、设施、卫生、性价比得分可得总得分。

岭回归(Ridge Regression ,RR) 方法是一种利用正则化的最小二乘法方法，它主要包括以下步骤: ①生成训练样本点的多变量标签矩阵; ②学习线性分类器，即投影矩阵; ③对新样本进行分类识别。岭回归方法使用正则单形顶点(regular simplex vertices)作为训练样本的多变量标签，将高维特征空间映射到低维特征空间，并使样本投影到这些正则单形顶点的周围。圆柱与抛物面的交点对应的 β1、β2 值，即为满足约束项条件下的能取得的最小的 β1 和 β2.从 β1β2 平面理解，即为抛物面等高线在水平面的投影和圆的交点。如图 4-6 所示。

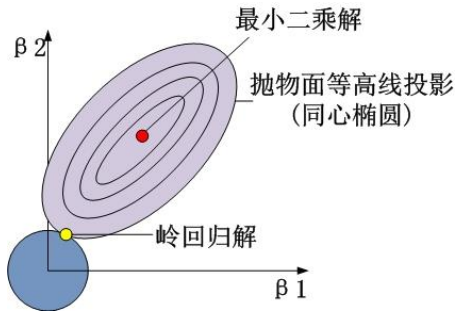

图 4-6 岭回归原理图

岭回归用于控制模型系数的大小来防止过度拟合。岭回归通过在成本函数中

加入模型参数的正则项以平衡数据的拟合和系数的大小。在求解岭回归的函数时，本文采用梯度下降与学习率的思想。

梯度下降是迭代法的一种,可以用于求解最小二乘问题(线性和非线性都可以)。在求解机器学习算法的模型参数，即无约束优化问题时，梯度下降（Gradient Descent）是最常采用的方法之一，另一种常用的方法是最小二乘法。在求解损失函数的最小值时，可以通过梯度下降法来一步步的迭代求解，得到最小化的损失函数和模型参数值。

梯度下降法(gradient descent)是一个最优化算法，常用于机器学习和人工智能当中用来递归性地逼近最小偏差模型。一般情况下，梯度向量为 0 的话说明是到了一个极值点，此时梯度的幅值也为 0。而采用梯度下降算法进行最优化求解时，算法迭代的终止条件是梯度向量的幅值接近 0 即可，可以设置个非常小的常数阈值。

学习率(Learning rate)作为监督学习以及深度学习中重要的超参，其决定着目标函数能否收敛到局部最小值以及何时收敛到最小值。合适的学习率能够使目标函数在合适的时间内收敛到局部最小值。

固定学习率时，当到达收敛状态时，会在最优值附近一个较大的区域内摆动；而当随着迭代轮次的增加而减小学习率，会使得在收敛时，在最优值附近一个更小的区域内摆动。

4.2.4 综合评价分析

通过结果-目标值的不断训练可以得出岭回归权重方程的均方误差值，结果-目标值训练图如图 4-7 所示，均方误差值的具体结果如表 4-6 所示。

表 4-6 岭回归权重与总得分 MSE 计算结果

数据项	服务 W	位置 W	设施 W	卫生 W	性价比 W	MSE
酒店	0.25	0.09	0.15	0.24	0.12	0.017
景区	0.18	0.15	0.22	0.17	0.008	0.015

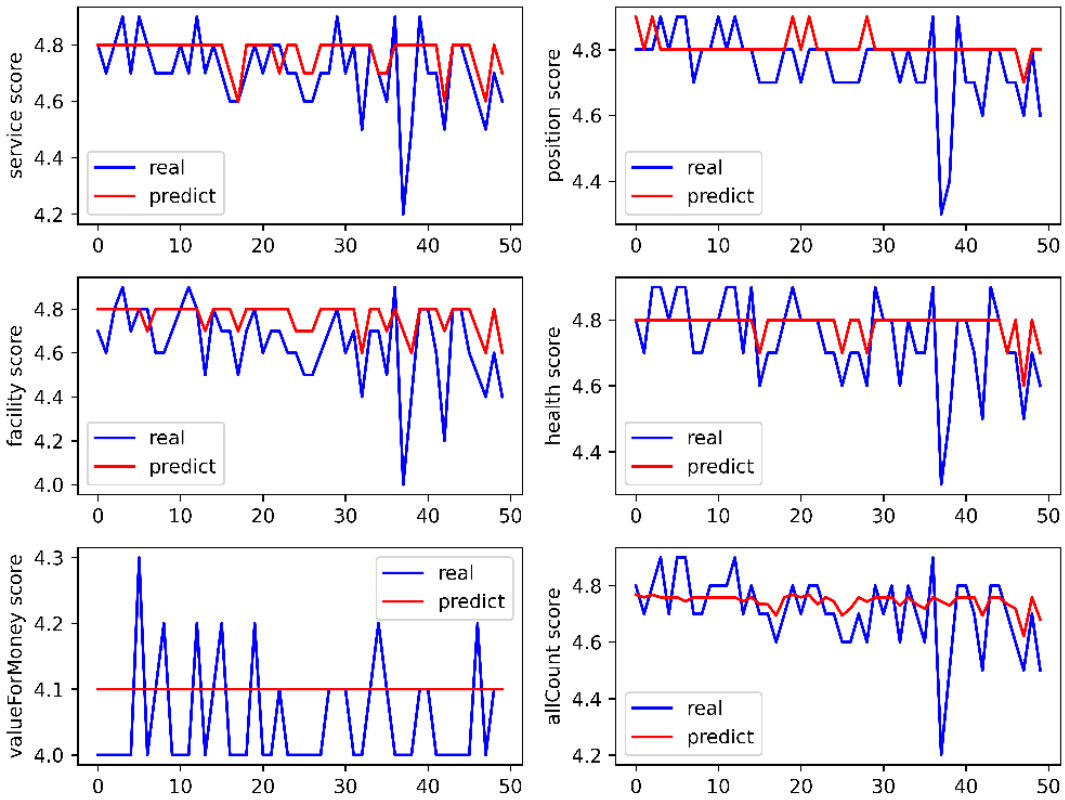

图 4-7 结果-目标值训练图

由表可知，酒店的主要影响因子为服务与卫生，总得分 MSE 为 0.017。根据上文对评论文本中关键词的提取及满意度的评估可以看出，酒店对游客提供的服务与卫生是对游客总体满意度影响较大的因素，其次是设施与性价比。因此，提高游客满意度，酒店提升自身美誉度要着重注意服务质量与卫生情况。贴心细致的服务与干净卫生环境可以增加消费者对酒店的依赖性与忠诚度。

游客对酒店设施的关注度仅次于服务水平与卫生环境，健全的设施是酒店满足消费者物质需求的保证。酒店要建立设备设施保养维护体系，定期对设备设施进行检查，对设备的使用损耗进行及时检修，保障其工作效率。

而景区的主要影响因子为设施，卫生与服务，总得分 MSE 为 0.015。从表中可看出景区综合评价的情况，游客选择景点会较注重其设施，其次是服务，卫生与位置。因此，提高游客满意度，景区在提升自身美誉度要着重注意设施情况；景区基础配套设施、卫生安全设施、休闲游憩设施、资源维护设施都是游客的关注点。景区工作人员作为景区服务主要提供群体直接与游客进行对接，其服务态度、服务热情以及面对顾客需求的及时性与专业性都对顾客满意度有重要影响。游客进人旅游景区首先感受到的是景区的卫生状况，并且卫生状况自始至终都影

响着游客的整个游览过程，卫生状况对景区吸引力有着重要的影响。景区区位的选择直接影响到其客源的丰富度，游客出行交通的便利性是影响游客满意度的重要特质。

4.3 问题三

题目：出于各种原因，网络评论常常出现内容不相关、简单复制修改和无有效内容等现象，妨碍了游客从网络评论中获得有价值的信息，也为各网络平台的运营工作带来了挑战。请从文本分析的角度，建立合理的模型，对附件 1 景区及酒店网络评论的有效性进行分析。

4.3.1 问题分析及解决思路

（1）问题分析

该题意从用户评论分析出具有参考意义的关键词，文本将其转化为词向量得出词的相似度最终解决无效评论问题。该题的重难点在于文本词向量的转换与关键词提取的准确性。

（2）解决思路

针对评论归类而言，需要将评论数据进行无监督的文本聚类。文本聚类主要是依据著名的聚类假设：同类的文档相似度较大，而不同类的文档相似度较小。那么便可根据 Word2Vec 算法对留言数据进行量化，利用基于语义的词向量模型比基于词频统计的算法模型效果更加，从而进一步提高聚类效果，流程图如图 4-8 所示。

图 4-8 有效性评价分析模型图

4.3.2 文本聚类

（1）词向量获取

根据官方提供的数据统计，附件一共有 84333 条数据，数据量较大，在词向量的获取方面上考虑 Word2Vec 模型进行构建，利用 Jieba 分词器对附件一的评论进行分词后采用 Word2Vec 对评论的主题进行构建模型。

（2）Word2Vec 模型

无论是传统方法的机器学习，还是现在较流行的深度学习，在学习数据输入方面，必须要输入计算机能识别和认知的数据。而在自然语言处理（NLP）这个信息工程的子领域上，其所获得的数据皆为文本数据，并不是计算机可直接识别使用的数据。我们需要将这些数据转换为数值的形式后再进行数据的使用。这种嵌入方式称为为词嵌入(Word Embedding)，本文使用 Word2Vec 进行实现。

Word2Vec 本质是一种简单的神经网络模型，通过训练将文本内容处理简化为 K 维向量空间的向量，向量的相似性可用于表示文本的语义相似性[13]。Word2Vec 模型基本思想如图 4-9 所示。

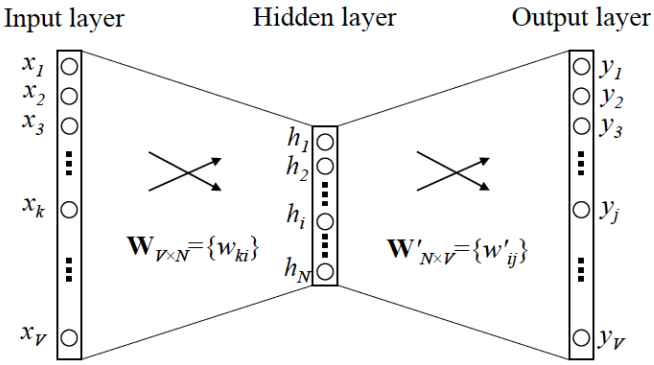

图 4-9 Word2Vec 模型基本思想

Word2Vec 主要使用 CBOW 和 Skip-gram 模型进行训练，具体内容如下：

1）Skip-gram 模型

Skip-Gram 模型又称作为连续跳过语法架构，与 CBOW 模型恰恰相反，通过利用一个单词来预估上下文的，相对 CBOW 模型较慢，但是对不常用的单词来说表现更好，其原理模型图如图 4-10 所示。

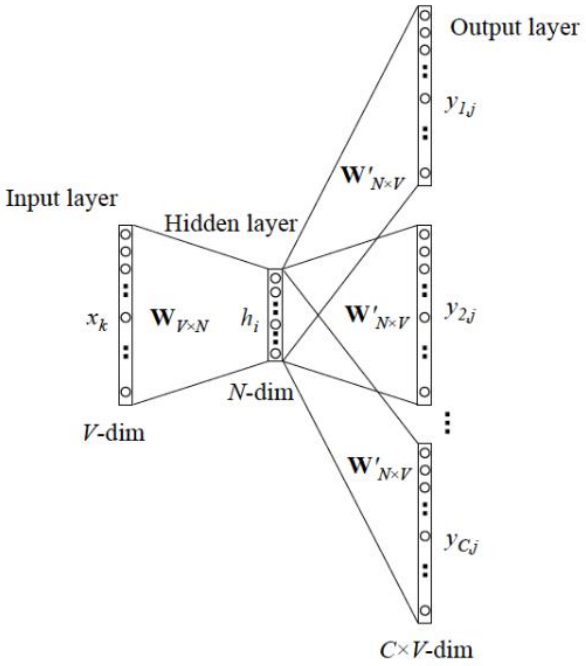

图 4-10 skip-gram 模型结构图

2）CBOW 模型

CBOW(Continuous Bag Of Words)又称作为连续词袋模型，是一个三层神经网络，在连续词袋体系结构中，该模型从周围上下文词的窗口中实现预测当前词。输出对当前单词的预测，其原理模型图如图 4-11 所示：

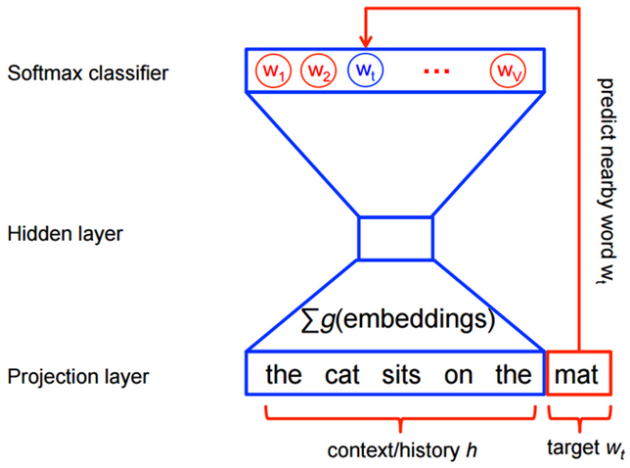

图 4-11 CBOW 模型结构图

（3）操作流程

1) 利用分词软件对用户的评价进行中文分词，然后通过开源 Word2Vec 模型

中基于预测的 CBOW 算法训练得到词向量模型。

2) 将景区及酒店的评论类别分布的统计结果转化为向量的形式，构建出“文本-类别分布”矩阵,最终文本表示形式如公式(8)所示:

$$D = \begin{bmatrix} \omega'_{11}, \omega'_{12}, \dots, \omega'_{1i}, \dots, \omega'_{1N} \\ \vdots, \quad \vdots \quad, \dots \quad, \vdots, \dots \quad, \vdots \\ \omega'_{j1}, \omega'_{j2}, \dots, \omega'_{ji}, \dots, \omega'_{jN} \\ \vdots, \quad \vdots \quad, \dots \quad, \vdots, \dots \quad, \vdots \\ \omega'_{M1}, \omega'_{M2}, \dots, \omega'_{Mi}, \dots, \omega'_{MN} \end{bmatrix} \tag{8}$$

其中, N 代表类别数量, M 代表文本数。

通过 Word2Vec 模型得出网评文本的相似度后进行查找相似的关键词做聚类，即可得出酒店和景区的聚类关键词。

4.3.3 文本有效性分析

原始数据经过数据去重、去除短句、机械语料压缩等数据预处理操作后初步过滤了网评文本中简单复制修改和无有效内容的评论，关于内容不相关的文本通过聚类出网评文本的关键词，并生成网评文本的词频向量，根据该词频向量与关键词的词频向量的余弦值来判断是否相关，余弦值越大，表示二者越相似，相关性也越大。具体解决步骤如下：

（1）文本聚类提取关键词

通过文本聚类得出酒店关键词为服务、位置、设施、卫生、性价比，景区的关键词为环境和交通。部分数据聚类效果图如下：

1）酒店聚类关键词: 服务，位置，设施，卫生，性价比。部分结果如表 4-7 所示。

表 4-7 酒店关键词聚类部分数据

名称	服务关键词	位置关键词	设施关键词	卫生关键词	性价比关键词
H01	前台 热情 玩耍 高兴	广州东站 交通 必备	机器人 隔音 热水	干净 舒服 环境 卫生	合理 消费 必选 纸质
H02	礼宾 不错 阿姨 特别	地理位置 很近 居家	浴缸 设备 配套 标准	体验 简单 房间 非常	入门 商场 实惠 费用
H03	态度 大堂 干净 愉快	不错 优越 出差 首选	完备 安全 智能 装修	安静 预想 整洁 保洁	车站 便宜 价格 实在
H04	早餐 舒适 房间 行政	紧邻 好评 便利 附近	硬件 陈旧 商务 WIFI	头发 清洁 阿姨 态度	无语 免费 餐费 实惠
H05	免费 服务一流 周到	方便 地铁 出行 便捷	淋浴房 水池 电视 空调	脏乱 小刘 床 天花板	偏僻 堵车 出票 房费
H06	专业 接待 服务员	地段 出租车 两三分钟	超市 老旧 小礼品 漏水	整齐 香薰 清新 很好	水果 钱 昂贵 小费
H07	礼貌 优质 很棒 延迟	靠近 饮食 办事 附近	停车场 小孩 窗帘 被套	灰尘 洗漱 异味 积水	小时 晚间 推荐 太贵

2）景区聚类关键词：环境，交通。部分结果如表 4-8 所示。

表 4-8 景区关键词聚类部分数据

名称	环境关键词	交通关键词
A04	景区 愉快 玩耍 高兴	学生票 拥挤 国内 纸质
A05	小孩 不错 独特 安静	入门 商场 实惠 很快
A06	美美 适合 震撼 愉快	车站 便宜 方便快捷 近
A07	徒步 漂亮 周末 游玩	国内 地铁 无语 停车
A08	风景 景色宜人 出游 好看	偏僻 堵车 出票 五一
A09	宜居 生态 自然 壮观	夜景 美丽 仓促 保养
A10	风光 喜欢 亲子 娱乐	小时 晚间 公交 邮轮

（2）计算关键词与网评文本的余弦值

通过提取出网评文本中的关键词后，将网评文本与关键词进行相似度比较，在此过程中采用了 Word2Vec 的余弦相似度原理，该原理是用来判断两个文章或者句子相似程度的一个算法。根据向量坐标，将词向量绘制在空间中，求得夹角的 Cos 值，Cos 值越接近 1，则说明夹角越小，即两向量相似。具体计算如公式(9)所示。

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{9}$$

通过使用 word2vec 进行训练，得到词语的向量模型，在模型中直接计算与关键词最接近的词，这里以设施、环境关键词所对应的前 5 个相似度最高的词为例，具体如图 4-12 所示。酒店五项指标相似度的值，以酒店 H01 为例，如表 4-9 所示：

```
=== 词库中距离“设施”的前5个最近的词是： ===  | Elapsed Time: 0:00:00  ETA:
0:00:00 125.52  B/s
('前台', 0.88415687505314)
('早餐', 0.81910075209882)
('健身房', 0.80458239787651)
('设备', 0.78596645821417)
('服务', 0.74378964724687)
=== 词库中距离“环境”的前5个最近的词是： ===  | Elapsed Time: 0:00:00  ETA:
0:00:00 735.82  B/s
('风景', 0.95702295401120)
('好看', 0.90287547901455)
('视野', 0.85708756631753)
('壮观', 0.80551893843741)
('游玩', 0.79620894546319)
('优美', 0.73259884715093)
```

图 4-12 景区相似度计算

表 4-9 酒店相似度值比较

服务相似度	位置相似度	设施相似度	卫生相似度	性价比相似度
0.999671	0.999583	0.999821	0.999733	0.999660
0.999647	0.999550	0.999797	0.999694	0.999560
0.999639	0.999449	0.999781	0.999661	0.999483
0.999627	0.999369	0.999763	0.999623	0.999473
0.999611	0.999365	0.999730	0.999620	0.999437

4.4 问题四

题目：旅游业繁荣发展给游客带来了选择困难的问题，评分接近的景区或酒店很难根据评分进行取舍。请建立合理的模型和算法，从景区及酒店的网评文本中挖掘他们各自的特色和亮点，以吸引游客提升竞争优势。请选择综合评价高、中、低三个层次的各 3 家景点和 3 家酒店，结合模型的结果，分析他们各自的特色。

4.4.1 问题分析及解决思路

（1）问题分析

问题四旨在根据网评文本及评分，分析高、中、低三个层次的景区和酒店的特色，帮助用户在评分接近的景区和酒店中进行取舍，该题的重难点在于通过网评文本挖掘出目的地真正的特色以及特色分析的准确性和实用性。

（2）解决思路

1）通过问题二综合了网评文本及评分情况，得出了服务、位置、设施、卫生、性价比这五个方面的最终评分及总得分，并根据总得分分出高、中、低三个层次；

2）由于酒店及景区的最后总得分分数十分接近，用户难以进行消费决策，例如，酒店的最终总得分分别为 4.5、4.6、4.7（满分为 5 分），根据最终得分用户无法较为全面地分析出酒店的特色与优势，因此，本文基于高、中、低三个层次用户期望值，从中分别挑选出 3 家距离本层期望最近的景区及酒店；

3）基于用户评论的相似度计算特色，原理是用户评论中出现的具有特色特征的词汇，这些词汇可以反映酒店和景区的特色。

整体解决思路如图 4-13 所示。

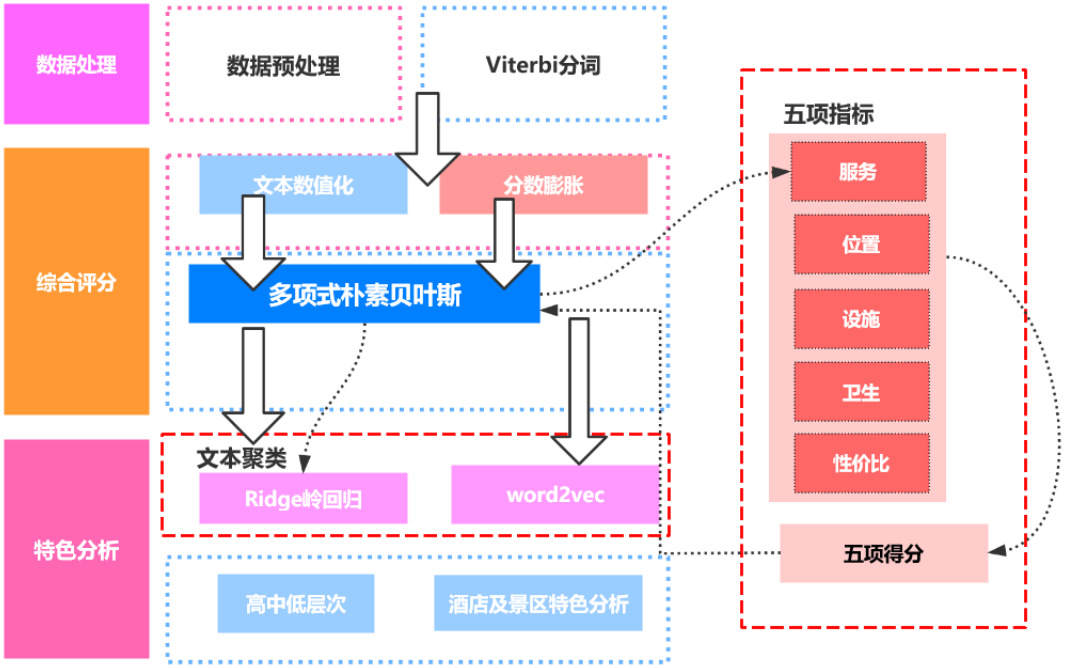

图 4-13 酒店及景区特色计算流程图

4.4.2 特色分析

综合网评文本和打分情况后，高、中、低三个层次的酒店和景区结果分别如图 4-14 和图 4-15 所示。

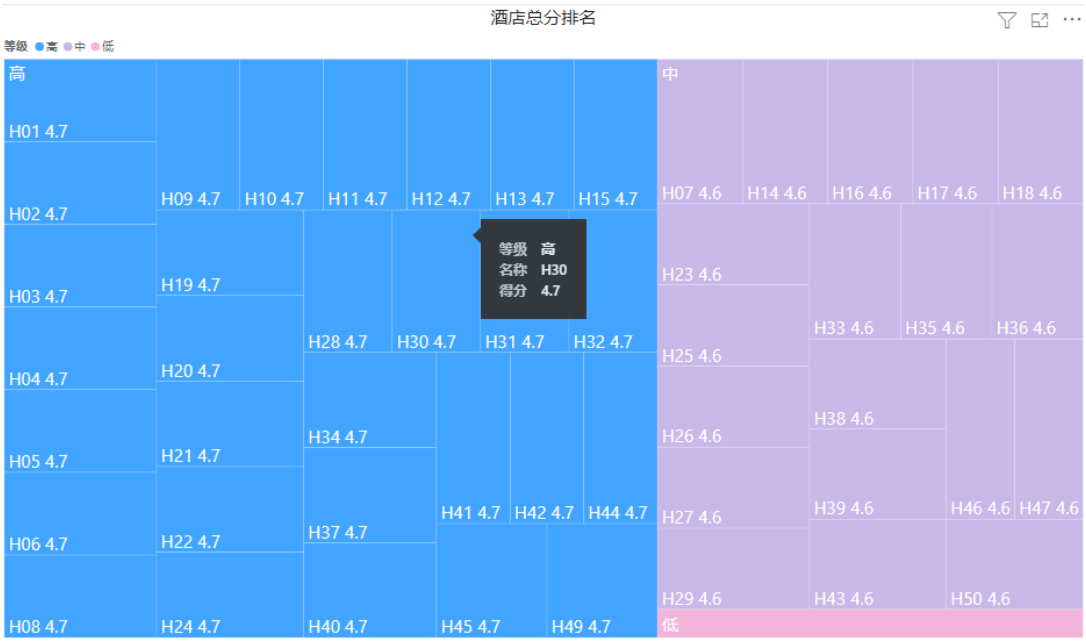

图 4-14 酒店排名结果

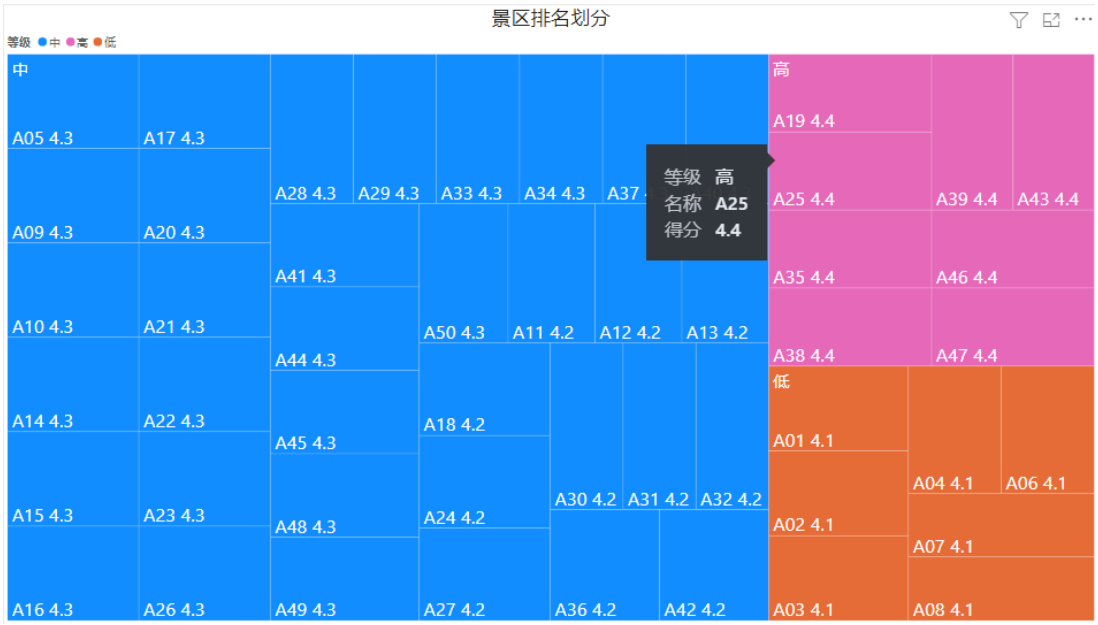

图 4-15 景区排名结果

本文从距离三个层次的期望值最近入手，选择出高层次酒店 H15，中层次酒店 H35，低层次酒店 H48，高层次景区 A46，中层次景区 A13，低层次景区 A07。

1）通过文本相似度计算，只有该酒店的评论中出现了：晚会表演、装潢高档，基于此判断这是一家具有特色表演的高端酒店，因此建议消费人群为：高端旅游人群、出差任务并不繁重的高端商务人士。

2）中层次 H35 酒店的评论中则出现了：河源特色，由此推测 H35 酒店为一家具有当地特色的、含有人文环境的酒店，建议消费人群为：追求自由轻松、喜欢乡土气息的用户；

3）低层次 H48 酒店的评论中出现了：年代久远，由此推测 H48 酒店具有一定的年代感、底蕴丰厚的酒店，建议消费人群为：推崇老品牌、有怀旧感的用户。

4）通过文本相似度计算，只有在高层次 A46 景区的评论文本中出现了：营地遗址、地貌丰富的字样，可以推断出该景区兼具自然美景与人文历史，适合喜欢领略自然风光及文化古韵的游客；

5）中层次景区 A13 的网评文本中则出现了：海洋温泉、温泉节，因此判断该景区主打海洋养生主题，适合追求健康以及想要放松身心的游客；

6）最后，在低层次景区 A07 的网评文本中出现了：全中国的微缩景观，由此推断该景区的民族特色强烈，历史文化浓厚，适合喜欢民俗风情、想要了解中国文化的游客。

五、总结及展望

5.1 总结

本文首先对酒店及景区的用户评论计算 TOP20 热词，基于此，提出了一个基于统计学和机器学习算法的评分系统，经过训练，得到最终模型的 MSE 为 0.01368，充分证明该系统的评分比较有参考价值，且有助于商家和景区精准把控游客满意度，针对性改正不足，突出优势，提升竞争力；

通过对用户评论数据进行深度挖掘和分析，建立了一个保持用户原意的有效性评价分析模型。实验结果证明该模型能够去除无效和恶意评论，保留高质量评论文本，可以给各地文旅主管部门提供决策支持，提升景区及酒店的美誉度；

最后对景区的及酒店的特色进行分析，基于用户评论的相似度计算得到酒店景区的特色评价，并基于该特色给出了适合居住和旅行人群的建议。该方法对旅游企业科学监管、资源优化配置及市场可持续开拓具有长远而积极的作用。

5.2 展望

虽然本文的方法对于精准解读用户评论、量化酒店和景区评分、凸显各酒店景区特色具有参考价值，仍然存在几点不足需要在进一步的研究中进行改进：

（1）本文利用传统机器学习的方法，利用多项式朴素贝叶斯算法进行分类，虽然均方误差值达到了 0.01368，但是仍存在局限性，无法继续精确，且是在特定的语料库上建立的分类，可以进行适当调整。

（2）本文所提出的研究方法只在旅游相关领域进行了实验验证，未来研究可以拓展至其他领域，进一步验证该研究方法的通用性。

参考文献

[1] 潘伟芬.政企合作模式下网络刷单规制现状分析[J].上海商业,2021(02):194-196.

[2] 刘莹,李宝库.负面评论对消费者购买意愿的影响机制研究——基于双系统模型[J].财经论丛,2021(03):93-102.

[3] Kumar N,Benbasat I.Research note:The influence of recommendations and consumer reviews on evaluations of websites[J].Information Systems Research,2006,17(4):425-439.

[4] 邓莎莎,张朋柱,张晓燕,李欣苗.基于欺骗语言线索的虚假评论识别[J].系统管理学报,2014,23(02):263-270.

[5] 刘景方,李嘉,张朋柱,刘璇.用户评论标签摘要系统的有效性研究[J].系统管理学报,2016,25(04):613-623.

[6] 刘志成,钱怡伶.基于 SEM 模型武陵源生态旅游景区游客满意度研究[J].湖南社会科学,2019(03):121-127.

[7] 吴维芳,高宝俊,杨海霞,孙含琳.评论文本对酒店满意度的影响:基于情感分析的方法[J].数据分析与知识发现,2017,1(03):62-71.

[8] 金家华,吴浪涛,张婷婷,闫相斌.基于情感分析的外卖商家评分研究[J].管理学刊,2020,33(02):66-75.

[9] 张亚成,朱涤尘,夏换.基于 K-Means 聚类算法的景区数据分析——以黄果树景区为例[J].信息技术与信息化,2021(02):208-212.

[10] 袁娜. 基于特征加权与密度聚类的景区信息挖掘系统研究与实现[D].武汉理工大学,2019.

[11] 梁胜,成卫青.基于组合型中文分词技术的改进[J].南京邮电大学学报(自然科学版),2013,33(06):112-117.

[12] 张伦干. 多项式朴素贝叶斯文本分类算法改进研究[D].中国地质大学,2018.

[13] 陈丹华,王艳娜,周子力,赵晓函,李天宇,王凯莉.基于 Word2Vec 的 WordNet 词语相似度计算研究[J/OL].计算机工程与应用:1-11.