第九届“泰迪杯”数据挖掘挑战赛

基于文本挖掘的旅游目的地印象分析

摘要

近来，网络评论在旅游生态中的地位显著提升，很大程度上直接反映了游客的消费体验感受、关注焦点与情感诉求，从而勾勒出游客对旅游目的地的整体印象。本文选取景区和酒店的评论文本，挖掘游客的关注焦点，同时为景区及酒店等相关经营者、文旅部门做出更优决策提供理论支撑和数据支持。

针对任务一，本文首先对 50 家景区和 50 家酒店共计 84333 条评论数据，进行热词挖掘并计算其热度。先对原始数据集进行数据清洗，避免因存在较大的噪声造成研究误差。再利用 Text Rank 算法对去噪后的评论进行关键词提取，并结合语义网络分析，获取印象热词，并设计了一套科学的热度评价指标体系，计算其热度值。

针对任务二，本文以游客满意度理论为基础，以拆分后的短句为研究对象，设计了两类评价模型。其一是基于情感分析的游客满意度评价模型，选取 K-means、GMM、HAC、AP 中聚类效果的算法获取游客满意度影响因素。利用 NB、SVM、Text CNN 算法与情感分析技术获得游客情感分类及极性得分，结合情感分析获取游客满意度总得分及分项得分。其二是构建基于模糊综合评价方法的游客满意度评价模型。以专家打分作为标准，利用均方误差（MSE）评估上述两个模型的合理性，结果显示基于情感分析的游客满意度评价模型更为科学合理。最后依据情感模型得分，将游客目的地划分为高、中、低不同层次。

针对任务三，本文以信息质量理论为基础，从评论内容质量和表达形式质量两个维度，综合时效性、情感性、相关性、完整性、可理解性五个指标，利用随机森林（Random Forest）构建基于信息质量视角的文本有效性评价体系，筛选出景区及酒店的高效评论，以减少信息搜寻成本，提高评论质量，增强评论生态的活力。此外，本文创新性地基于主题建模和词向量相似度，构建高效评论排序模型，精简冗余以减少游客的检索时间。

针对任务四，本文依据前三题的研究结果，将情感分类和 LDA 主题挖掘进行融合，提取出景区及酒店的积极高效评论文本集。以各家景区及酒店积极且高效的评论数占其总评论数的比重为指标，筛选出高、中、低不同层次的各三家景区及酒店，建立 LDA 主题挖掘模型，并将其可视化，挖掘各自的特色和亮点，以吸引游客提升竞争优势。

关键词：文本挖掘；情感分析；有效性评价；LDA 主题挖掘；竞争优势

目录

表录

图录

1 引言

1.1背景及意义

2020 年 11 月我国文化和旅游部、发展改革委、教育部、工业和信息化部等十部门联合发布《关于深化“互联网+旅游”推动旅游业高质量发展的意见》。5G、大数据、云计算等信息技术成果应用普及，加快了旅游领域数字化、网络化、智能化转型。网络评论或点评的规范化、数据化、平台化升级，一方面帮助游客准确了解交通、景区与酒店等场所的基本信息与服务信息，根据相应的评论内容，做出合理的旅行消费选择；另一方面，景区与酒店基于大量的评论反馈进行更有针对性与实效性的质量管理。网络评论对于监测旅游舆情、改善旅游生态、助力旅游业高质量、高效率、高水平发展起着关键作用。

在实践管理中，网络评论具有数量规模大、内容覆盖广、标准差异化等特点，这对于景区及酒店筛选有效评价信息、科学分析评价内容、切实提高服务质量等工作极具挑战。再者，推进旅游业现代化建设发展，增加旅游业服务能力，我国文旅主管部门和旅游业相关企业需要满足游客高质量的消费需求。其中，获取游客的真实服务评价体验是重要前提。

自然语言处理（NLP）技术是一门融合多学科的技术，是计算机科学领域与人工智能领域的重要发展方向，近年来其在各领域逐步深度地展开相关应用。基于自然语言处理技术的游客目的地印象分析，能解决网络评论中语义理解问题，为提高目的地的美誉度明确具体的发展方向。

本文针对景区及酒店的游客评论数据，科学构建游客目的地满意度综合评价指标体系，为景区和酒店提高游客满意度提供决策依据。游客满意度越高，目的地美誉度就越大，两者有着密不可分的关系，这涉及到如何稳定客源、取得竞争优势、吸引游客到访消费等重要事项。因此，印象分析能够促进旅游业持续健康发展，对景区及酒店具有参考价值与借鉴意义，而且对于旅游企业科学监管、资源优化配置以及市场持续开拓具有长远而积极的作用。

1.2文本挖掘目标

“游客目的地”文本挖掘的目标主要包含 4 个部分：

1.景区及酒店印象分析

许多旅游软件，例如携程、Airbnb 等每天都会接收大量的游客评论。在海量的评论中，提取存在着许多反映共同问题、表达共同诉求的评论，对它们进行针对性地处理，对提高景区和酒店的服务有重要的意义。因此，利用自然语言处理中的文本分类（Text classification)技术力图实现评论话题自动分类。简言之，通过自然语言处理中的话题检测与跟踪范畴，从大量评论中，自动发现评论中反映的热点问题，并度量其热度以增强印象分析的真实性与整体性。

2.景区及酒店的综合评价

景区与酒店基于大量的评论反馈可以进行综合性分析，同时也可以对服务、位置、卫生、环境和性价比等方面进行针对性分析。因此，优化满意度评价指标体系可以全面反映游客的需求和情感导向，实现景区和酒店的级别评定与划分，为其工作发展提供具体引导。

3.网评文本的有效性分析

由于评论文本自身存在差异，可能会出现内容不相关、简单复制修改和无有效内容等现象，这会妨碍游客从网络评价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。探其根本，自动地评价文本有效性，有助于提升景区和酒店的服务质量，将游客的需求落到实处，同时有利于旅游平台的在线运营。因此，本文融合五项指标，从相关性、完整性、可解释性、情感性、时效性等角度，自动地评价文本的有效性，并创新性地进行了高效评论排序。

4.景区及酒店的特色分析

旅游运营模式雷同化与可复制化会给游客的选择带来一定困难，尤其是面对评分接近的景区或酒店，如何仅根据评分进行取舍成为当下多数游客的困扰。因此，在网评文本中挖掘出景区或酒店各自的特色和亮点，以其个性化吸引游客提升竞争优势极其重要，实现从“人无我有”到“人有我优”的展示。因此，利用自然语言处理中的主题挖掘，实现对景区及酒店的特色化与个性化分析。

1.3 研究思路

在现有研究成果基础上，本文研究思路如下：利用景区及酒店的游客评论文本，在海量信息的时代背景下，运用信息处理技术手段评估游客对目的地的整体印象、评论文本的有效性等。为解决这些问题，提出在游客满意度理论、信息质量理论体系基础上结合文本挖掘、情感分析、特征选择、机器学习等技术手段构建相关评价体系，并进行模型运用。最后，基

于评估结果，分析并总结提升景区及酒店竞争力的相应策略。研究思路如下图 1-1 所示。

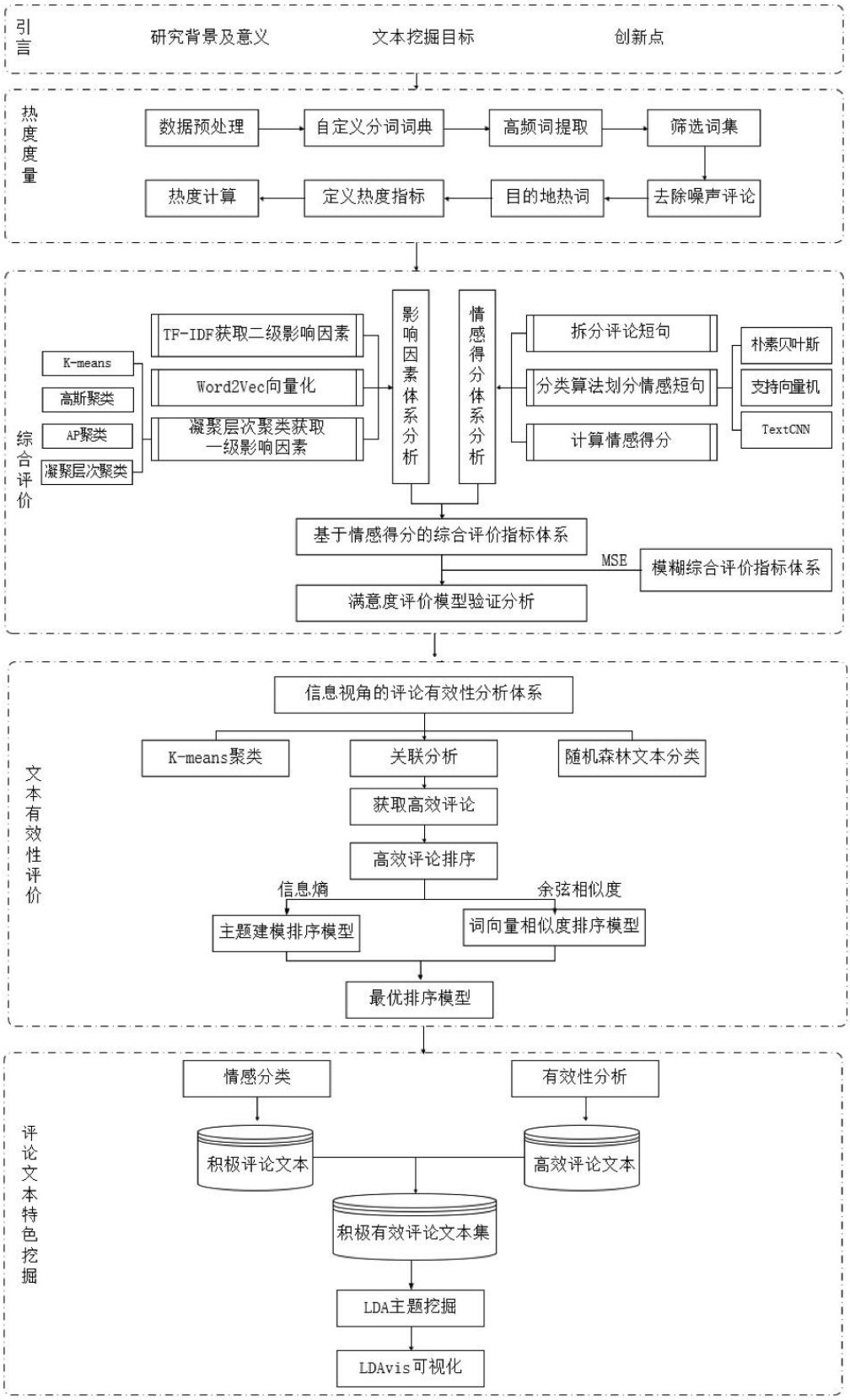

图 1-1　本文研究思路图

1.4创新点

1.采用 pkuseg 分词构建自定义分词词典

目前，我国的文本研究主要基于 Jieba 分词工具，而本文采用一套全新的中文分词工具包——pkuseg。在 Linux 测试环境下，使用 Jieba 分词和 pkuseg 对新闻数据（MSRA）和混合型文本（CTB8）数据进行分词的准确率测试。参照第二届国际汉语分词测评发布的国际中文分词测评标准，Jieba 分词对这两个文本的误差率高达 18.55%和 20.42%，而 pkuseg 的误差率为 3.25%与 4.32%。

各领域专业词汇不尽相同，因此为了得到更好的分词结果，本文结合景区及酒店评论语料、知网情感词典构建贴合景区、酒店评论的自定义分词词典，提高 pkuseg 工具的分词效果。

2.利用语义网络分析修正热度词

关键词提取结果中，存在例如“不错”、“非常”等指代对象不明且与游客印象无关的特征词，这需要借助人工识别和剔除，才能使结果更为精准。评论文本多表达游客的情感，表达口语化，为此需要借助语义网络分析来进一步挖掘隐藏的信息。

3. 科学对比多种聚类算法

传统的 K-means 聚类算法在无监督任务时，需预设聚类数量。但许多时候，会存在不知如何预设聚类数量的问题。因此本文选取了无需预设聚类数量的 AP（AffinityPropagation）聚类和凝聚层次（HAC）聚类法，以及须预设数目的高斯混合（GMM）模型与 K-means 进行效果对比，选取效果最佳的算法进行聚类。

4.引入模糊综合评价满意度模型

在现有综合评价领域，模糊和专家社会网络是研究的热点。因此本文在综合评价时，引入模糊综合评价法，将结果与专家得分和基于情感分析的满意度得分相比较，并利用均方误差（MSE）进行效果评价，增强结果可信度。

5. 构建基于信息质量视角的有效性评价体系，并提出高效评论排序模型

以信息质量理论为基础，结合网评文本的信息特征，从信息内容质量和信息表达形式质量两个独立的维度，融合非结构化的文本特征，构建了由分类指标、排序指标组成的高效评论挖掘指标体系。主要目的包括：对评论进行整理与分类，挖掘出高效的评论，剔除参考价值低的文本；为了更高效获取信息，减少信息查找成本，本文将高效评论进行排序，将最好

的评论置于最前的位置，方便游客发现对决策有用的评论。

6.构建基于情感分类的 LDA 主题挖掘模型

现有的研究中，将情感分类和 LDA 主题模型结合的很少见，多数研究如何提升情感分类准确度。本文在研究思路上结合了情感分类和 LDA 主题模型，并利用 LDA 可视化系统对主题挖掘深入分析，以分析各景区、酒店在发展过程中的竞争优势，为文旅主管部门和旅游相关企业提出更有针对性的建议。

2 相关算法介绍

2.1 TF-IDF 算法

TF-IDF 由词频（Term Frequency）和逆文档频率（(Inverse Document Frequency）组成，是文本相似度量中最常用的方法。在使用此算法之前，文本通常表示为多个单词的向量，一个词出现在文本中的频率称为词频（TF）。词频与该词在文本中的重要性呈正相关，而该词在整个语料库中出现的频率与该词在文本中的重要性呈负相关。因此这部分频率称为逆文档频率（IDF）。TF-IDF 的计算方法为两部分相乘，具体计算公式如下：

$$TF-IDF(w_i)=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\log\frac{|D|}{|\{j:w\_i\in d\_j\}|} \quad (2\text{-}1)$$

每个参数的含义如表 2-1 所示。

表 2-1　TF-IDF 算法参数含义

参数	含义
d_j	文本
$n_{i,j}$	文本d_j中该词出现的次数
$\sum_k n_{k,j}$	文本d_j中该词出现的次数之和
$\|D\|$	语料库中的文本总数
$\|\{j:w_i\in d_j\}\|$	包含该词的文本数

2.2 聚类方法简介

聚类方法是一个无监督学习技术，不需要先验知识，根据数据对象之间的相似性对数据

进行分区，应用范围极其广泛。聚类分析可以处理各种类型的大型数据，以发现隐藏的模式、未知的相关关系以及其他潜在有用的信息。然而，Naldietal.指出，不同的聚类方法可能会产生不同的分区。因此，如何选择一个好的聚类方法仍然是一个重要的且具有挑战性的难题。目前聚类方法有近百种，如 K-means 聚类方法、CLARANS 聚类方法、AP（Affinity Propagation）聚类、DBSCAN 聚类方法、BIRCH 聚类方法、GMM 聚类方法等。下文将 K-means、AP、GMM、HAC 聚类方法进行效果比较，选取最佳算法进行分析。

2.2.1 K-means 聚类

K-means（K 均值聚类算法）是通过迭代方法求解的分析算法，作用机理是把一堆数据点分成若干类。在指定的 K 个簇中，聚类效果与簇内的数据样本相似度成正比。基于以上思想，该聚类的目标函数如公式（2-2）所示:

$$J(C_1, C_2, ..., C_k) = \sum_{j=1}^{k} \sum_{i}^{n_j} (x_i - c_j)^2 \qquad (2\text{-}2)$$

其中公式中各符号含义如表 2-2 所示：

表 2-2 K-means 聚类目标函数公式符号含义

符号	解释
c_j	第 j 个簇的簇中心
x_i	第 j 个簇的样本 i
n_j	第 j 个簇的样本总量

对于该目标函数而言，c_j 为未知的参数。只有事先知道 c_j 的值，才能求得目标函数的最小值。为达到目标，K-means 的步骤阐述如下：

步骤一：确定聚类数据 K，在特征空间中选择 K 个点作为初始的聚类中心点。

步骤二：计算每个特征到聚类中心点的距离，并分配给距离最短的中心点。

步骤三：得到每个聚类中心点对应的簇，从而完成第一轮聚类。

步骤四：每轮结束后，需结合各个簇所对应聚类点的特征坐标，根据设定方法更新聚类中心，以新的中心重新聚类，选出新的簇。不断重复上述过程，知道所有的簇聚成一类为止。

2.2.2 AP（Affinity Propagation）聚类

2007 年 Frey 等人在著名科学杂志 science 上提出了 AP（Affinity Propagation）聚类算法。AP 算法根据 N 个数据点之间的相似度进行聚类，不需要事先指定聚类数目，而是将所有的数据点都作为潜在的聚类中心。

N 个数据点之间的相似度，就组成一个 $N \times N$ 的相似度矩阵 S，并以对角线上的值 $\mathrm{S}(i,i)$ 即参考度 p（preference）作为第 i 个数据点能否成为聚类中心 k 的评判依据，该值越大，表明该数据点成为聚类中心的可能性也就越大。

表 2-3 AP 算法中传递两种类型的信息表

信息	表达形式	表示内容	反映内容
吸引度	r(i,k)	从点 i 发送到聚类中心 k 的数值信息	k 点作为 i 点聚类中心的合适程度
归属度	a(i,k)	从聚类中心 k 发送到 i 的数值信息	i 点选择 k 点作为聚类中心的合适程度

可看出，吸引度和归属度越强，k 点作为 i 点聚类中心的可能性越大。AP 算法就是通过多次迭代，更新每一个点的吸引度和归属度信息。当已经历最大迭代次数，或数值收敛，则对于任意点 i，计算它与所有样本的吸引度与之和，那么 i 点的聚类中心 k 点如下式选择：

$$\mathrm{k} = \mathrm{argmax}\left(\mathrm{a}(\mathrm{i},\mathrm{k}) + r(i,k)\right) \tag{2-3}$$

AP 聚类具有如下优势：

（1）不需要事先指定聚类的数量。聚类的数量，由参考度(preference)S(i,i)的初始值与数据的分布共同决定；

（2）聚类的结果不会多次运行而随机变化。这比通用的 k-means 聚类更加稳定；

（3）适用于非对称与稀疏的相似性矩阵。

2.2.3 高斯混合模型（GMM）

高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型，这 K 个子模型是混合模型的隐变量（Hidden variable），K 的取值需要事先确定，具体的形式化定义如下：

$$\mathrm{P}\left(y \mid \theta\right) = \sum_{k=1}^{K} \partial_k \phi\left(y \mid \theta_k\right) \tag{2-4}$$

其中 ∂_k 是样本集合中 k 类被选中的概率：$\partial_k = P\left(z = k \mid \theta\right)$，其中 z=k 指的是样本属

于 k 类，那么$\phi(y|\theta_k)$可以表示为$\phi(y|\theta_k)=P(y|z=k,\theta)$，很显然$\partial_k \geq 0, \sum_{k=1}^{K}\partial_k = 1$，y 是观测数据。

首先可以先假设聚成 k 类，然后选择参数的初始值θ_0（总共 2K 个变量），这里需要引进一个变量γ_{jk}，表示的是第 j 个观测来自第 k 个 component 的概率，即数据 j 由第 k 个 component 生成的概率，根据后验概率计算得到：

$$\gamma_{jk} = P(z=k|y_j,\theta) = \frac{\partial_k \phi(y_j|\theta_k)}{\sum_{k=1}^{K}\partial_k \phi(y_j|\theta_k)} \tag{2-5}$$

注：这个与∂_k的区别，∂_k指的是第 k 个 component 被选中的概率，需要γ_{jk}对所有的数据进行累加。

上面是根据数据 j 计算各个 component 的生成概率，而现在根据每个 component 生成了 1,2,…,N 点数据，每个 component 又是一个高斯分布，那么根据∂,μ,σ^2的定义又可以直观地得出如下式子：

$$\partial_k = \frac{\sum_{j=1}^{N}\gamma_{jk}}{N} \tag{2-6}$$

$$\mu_k = \frac{\sum_{j=1}^{N}\gamma_{jk}y_j}{\sum_{j=1}^{N}\gamma_{jk}} \tag{2-7}$$

$$\sigma_k^2 = \frac{\sum_{j=1}^{N}\gamma_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\gamma_{jk}} \tag{2-8}$$

这样其实只是把原本样本一定属于某一类改成了样本属于某类的概率，而 k 类样本数量N_k变成了概率相加，$N_k = \sum_{j=1}^{N}\gamma_{jk}$，就可以直接得出上述公式，进行相互迭代，直到收敛，高斯混合模型就聚类完成。

2.2.4 凝聚层次聚类（HAC）

层次聚类算法属于无监督学习的一种聚类算法，分为 2 种聚类方式：凝聚式层次聚类（Hierarchical Agglomerative Clustering），简称 HAC 和分裂式层次聚类（DHC）。

凝聚式层次聚类的思想是：首先将每个数据单独成簇，之后按照相似性度量标准将相似

性最高的数据先进行合并，依照数据相似度从高到低的顺序依次合并成簇，簇间的相似度随着簇的合并而降低，直到达到给定的相似度阈值才会停止。本文采用凝聚式层次聚类法来对调频广播文本数据进行聚类，具体方法如下。

首先将原始样本（$F_1, F_2, \ldots, F_n$）中每个样本自成一类，原始的类中心为（$C_1, C_2, \ldots, C_n$），然后根据相似性度量标准将类中心最近的 2 个文本合并，根据聚类收敛的条件不断重复这一过程直至所有可合并的类合并完成。本文采用的相似性度量标准为欧式距离法，为

$$d(F_i, F_j)=\sqrt{\sum_{m=1}^{n}(w_{m,i}-w_{m,j})^2} \quad (2\text{-}9)$$

式中：F_i、F_j 分别表示文本集合中第 i 个和第 j 个文本；$w_{m,i}$、$w_{m,j}$ 分别表示文本 F_i 和 F_j 的第 m 个特征项的权重；n 表示文本特征项的数目。通过式（2-9）可计算出在数据集中文本 F_i 和文本 F_j 的距离，并将距离最近的 2 个文本合并为一个类。

该方法的优点在于：K-Means 等方法需要事先设定初始聚类中心和聚类个数，导致不同的聚类中心和聚类种类产生不同的聚类结果不尽相同，且产生局部最优解的可能性较大。对于 AHC 而言，因这种方法不需要事先设置数据的聚类中心，而是以整个输入数据为聚类中心，不断进行融合凝聚，最后达到聚类效果，不会陷入局部最优解，因此该方法能在一定程度上增加聚类的有效性。

2.3 分类算法简介

文本分类作为机器学习的一个分支，是通过使用机器学习模型对带有类别标签的文本数据集进行训练学习，进而使用这个模型对测试集进行预测。许多经典的分类器模型都被用于文本分类领域，其中比较常用的有朴素贝叶斯（NB），支持向量机（SVM），神经网络，随机森林、Text CNN 卷积神经网络等。下文选取最具代表性的 NB、SVM 和目前较为流行的随机森林、Text CNN 卷积神经网络进行效果比较，选取最合适的算法进行分析。

2.3.1 朴素贝叶斯（NB）

朴素贝叶斯（NB）算法是统计学的一种方法，该方法能运用到大型数据库中，且使用

方法简单，分类准确率较高[1]。算法原理为：假设 $P(\mathrm{X},\mathrm{Y})$ 独立分布，则得到朴素贝叶斯理论的表达形式如公式（2-10）所示：

$$y=f(\mathrm{x})=\arg\max_{c_k}\frac{P(\mathrm{Y}=\mathrm{c}_k)\Pi_j P(\mathrm{X}^{(\mathrm{j})}=\mathrm{x}^{(\mathrm{j})}\mid \mathrm{Y}=\mathrm{c}_k)}{\sum_k P(\mathrm{Y}=\mathrm{c}_k)\Pi_j P(\mathrm{X}^{(\mathrm{j})}=\mathrm{x}^{(\mathrm{j})}\mid \mathrm{Y}=\mathrm{c}_k)} \tag{2-10}$$

其中，输入 $X\subseteq R_n$ 为 n 维向量的集合，输出为类标记集合 $Y=\{c_1,c_2,...,c_{\mathrm{k}}\}$，输入为特征向量，输出为类的标记，$T=\{(\mathrm{x}_1,\mathrm{y}_1),(\mathrm{x}_2,\mathrm{y}_2)...,(\mathrm{x}_N,\mathrm{y}_N)\}$ 为训练集合。

2.3.2 支持向量机（SVM）

在支持向量机（SVM）算法中，训练模型的过程实际上是对每个数据点对于数据分类决定边界的重要性进行判断[2]。也就是说支持向量机的最终目的是在特征空间中寻找一个尽可能将数据集合进行分开的超级平面，SVM 的目标函数如公式（2-11）所示:

$$J(\mathrm{w},\mathrm{b},\mathrm{i})=\arg_{w,b}\max\min(d_i) \tag{2-11}$$

式中 d_i 表示样本点 i 到某条固定分割面的距离；$\min(d_i)$ 表示所有样本点与某分割面之间的最小距离值；$\arg_{w,b}\max\min(d_i)$ 表示从所有分割面中寻找“分隔带”最宽的“超平面”，其中 w 和 b 代表线性分割面的参数，进而达到分类效果。对于线性不可分的情况，则需要引入核函数，如高四核函数、多项核函数等。

2.3.3 Text CNN 卷积神经网络

Text CNN 卷积神经网络由 Yoon Kim（2014 年）提出，其将 CNN 运用到文本分类任务中，并且为了更好地寻找不同的局部信息，Text CNN 利用多个大小不同的卷积核（kernel）对文本语句进行关键词的提取（类似于多窗口大小的 n-gram），从而能够更好地捕捉局部相关性。与传统图像的 CNN 网络相比，Text CNN 卷积神经网络在网络结构上没有任何变化，甚至更加简单了，因此在文本分类、推荐等 NLP 领域应用广泛。从图 2.9 可以看出 Text CNN 卷积神经网络其实只有一层卷积，一层 max-pooling,最后将输出外接 soft max 来进行 n 分类。其主要结构如图 2-1 所示。

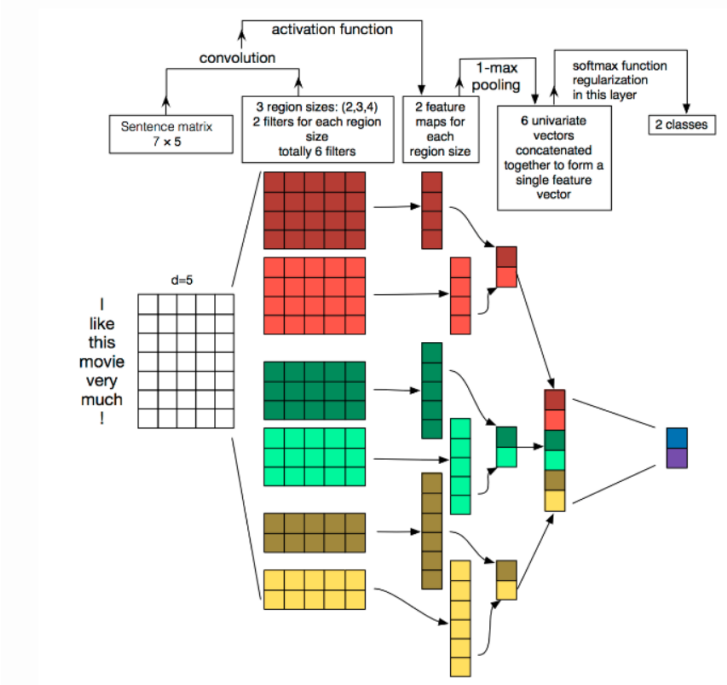

图 2-1　Text CNN 卷积神经网络结构示意图

Text CNN 卷积神经网络中有一层 embedding layer,该层主要导入预先训练好的词向量。将数据集里所有的词用一个向量的形式进行表示，得到一个嵌入矩阵 MM,其中每一行都是词向量。这个矩阵可以是固定不变的，静态(static)的；也可以根据反向传播进行更新，非静态(non-static)的。

Text CNN 详细过程：

1.wordembedding 的维度是 5，对于句子 I like this movie very much，转换成矩阵 AϵR7×5；

2.有 6 个卷积核，尺寸为(2×5),(3×5),(4×5)，每种尺寸各 2 个，A 分别与以上卷积核进行卷积操作（这里的 Stride Size 相当于等于高度 h）；

3.再用激活函数激活，每个卷积核得到了特征向量(feature maps)；

4.使用 1-maxpooling 提取出每个 feature map 的最大值；

5.然后在级联得到最终的特征表达；

6.将特征输入至 soft max layer 进行分类,在这层可以进行正则化操作(l2-regulariation)。

Text CNN 和一般 CNN 的主要区分在于 Text CNN 在一维卷积的前提下，使用更多不同高度的卷积核，得到特征表达也更为丰富。Text CNN 卷积神经网络中包括很多不同窗口大小的卷积核，常用的卷积核规模为{3,4,5},每个卷积核的 feature maps=100o 这里的特征图就是不同的 k 元语法。不同的卷积核在大小不同的范围进行操作，当不同卷积核生效的范围发

生重叠时，模型可以学习到不同的特征，使得最终的学习结果更为优化。

2.3.4 随机森林

随机森林是一种有监督的集成学习方法，由于其不存在过拟合问题并具有良好的分类性能，因此被广泛应用于各种分类问题中。采用 CART[3]作为随机森林元分类器，用 Bagging 方法[4]生成每棵决策树的随机训练样本集。当构造一棵树时，随机选取训练样本中的特征来确定决策树的节点分裂。Bagging 方法和 CART 的结合，并且随机选择特征进行属性分割，使得 RF 能较好的容忍噪声，并具有较好的分类性能。

随机森林是由多个决策树{h(x, θ k),k=1,2,⋯,n}组成的组合分类器，其中{ θ k}是一个独立同分布的随机向量，通过对所有决策树结果的投票产生输出结果。一个随机森林由 N 棵决策树构成，所有决策树（如决策树 T1,T2,…,TN）是一个分类器，随机森林的决策结果由所有决策树分类结果的组合策略得出。

本文构建的基于随机森林算法的文本分类模型的过程如图 2-2 所示，分为四个步骤。

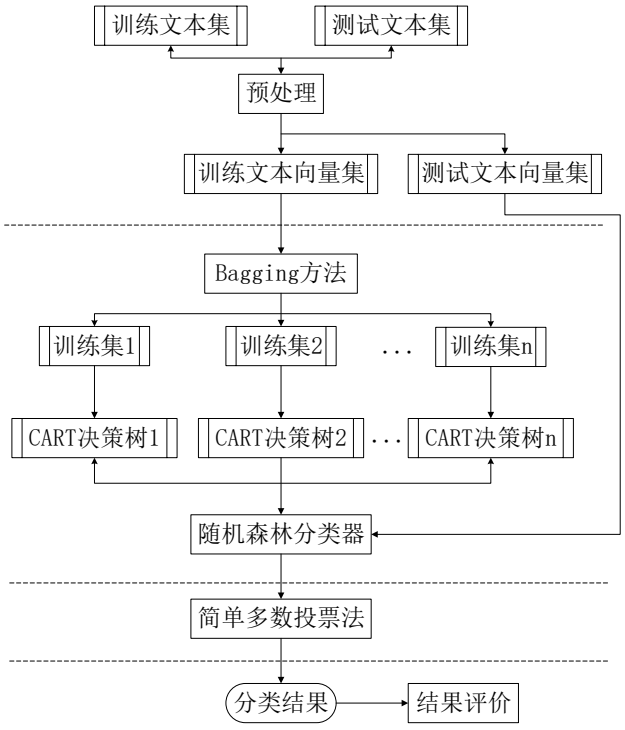

图 2-2 基于随机森林算法的文本分类模型流程图

步骤 1：建立文本向量集

对文本集进行预处理，形成一个可由随机森林算法识别的文本向量集，然后按一定的方式划分为训练集和测试集。

步骤 2：构建随机森林文本分类器

① 原始训练集 S 包含是所有评论的集合，从酒店和景区中各抽取 80%的样本作为建立决策树的训练集。

② Create Tree 分为两步进行

a.从样本 S 的 m 维特征中随机地选取 mtry 维特征，参考参数优选进行 mtry 的取值，候选的分裂特征变量从各个决策树的各个节点处随机选取 mtry 个变量。

b.根据属性将结点分为两个分支，递归调用 Create Tree 构造各分支，直到决策树训练集分类精度较高，或所有属性均被遍历。

③不对建立的决策树进行剪枝。

③ 重复②③④直到森林中有 k 棵决策树。

步骤 3 运用分类器进行分类

在对未知样本进行分类时，采用简单多数投票法确定随机森林的输出类别。

步骤 4：分类结果评价

本文使用的测试文本集是已知类别的样本。通过将分类结果与已知类别进行比较，评价分类器的分类效果。

对于分类预测实验的评价标准，本文采用统计学领域常见的精度（Accuracy）、查准率（Precision）、查全率（Recall）、F 值（F-measure）作为综合评价指标。其中，精度表示分类正确的样本占样本总数的比例，反应分类器对整个样本的判定能力；查准率表示所有预测为有效的样本数中真正有效的样本数；查全率是相对查准率而言，即所有有效的样本中有多少被预测正确；F 值是查准率和查全率相结合的综合评价指标。上述指标公式如下：

$$Accuracy=\frac{预测正确的样本数}{总样本数}=\frac{TP+TN}{TP+TN+FP} \tag{2-12}$$

$$Precision=\frac{预测为1且正确预测的样本数}{所有预测为1的样本数}=\frac{TP}{TP+FP} \tag{2-13}$$

$$Recall=\frac{预测为1且正确预测样本数}{真实情况下所有为1的样本数}=\frac{TP}{TP+FN} \tag{2-14}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{2-15}$$

TP 是预测有效且正确预测的评论数，TN 表示预测无效且正确预测的评论数，FP 表示无效评论被错误识别的评论数，FN 表示有效评论被错误识别的评论数。

2.4模糊综合评价方法

模糊综合评价分析方法来源于我们学习对的模糊数学，这方法就是把需要考察的模糊对象和反映模糊对象的那些模糊概念当成一定的模糊集合，然后建立适当的模糊隶属函数，并通过模糊集合论的相关运算与变换，对这个模糊对象进行定量分析[5]。

模糊综合评价方法的特点概括起来，主要有一下几点[6]：

(1)模糊综合评价方法可以进行多层次评价，并且其评价过程是可以不断循环的。前面过程的综合评价结果，可以作为后面过程综合评价的投入数据依据。即，对于一个比较复杂的评价对象我们可以进行单级模糊综合评价与多级模糊综合评价。

(2)评价指标的权重处理。模糊综合评价中的指标权重系数向量，是人为的估价权，是一个模糊向量，而不是模糊综合评价过程中所伴随生成的。

(3)模糊综合评价方法本身的性质也决定了其评价结果只能是一个向量集。所以其评价结果只是一个模糊向量集，并不是一个具体的值，并且其评价结果对于被评价对象也具有唯一性。

(4)评价等级论域的设立。在模糊综合评价方法里，总会设有一个评语等级的论域，且各个等级的含义必须是明确的。

模糊综合评价流程图如下所示：

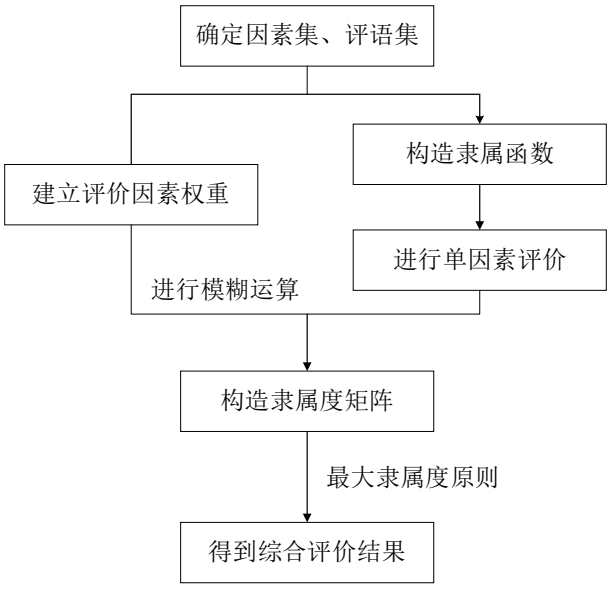

图 2-3　模糊综合评价流程图

2.5 LDA 主题模型

2003 年，Blei 提出了著名的 LDA(Latent Dirichlet Allocation)主题模型，这是一个文档生成概率模型。它主要在概率潜在语义分析模型(probabilistic Latent Semantic Analysis, pLSA)中添加了贝叶斯框架层，主要目标是使用无监督学习方法来提取大量文档中的隐藏主题信息，以帮助游客或读者快速理解文档中的信息[7]。

LDA 主题模型本质是利用文本特征词的共线特征来提取文本主题，具有三个清晰的层次，分别为文档层、主题层和特征词层。

LDA 的主要思想如下：文档集基于主题的概率分布，主题又基于特征词的概率分布。因此，存在一个式（2-16）的概率公式，表示文档M_m中词W_n出现概率：

$$p(w_n|M_m) = \sum_{k\in K} p(w_n|K_k)p(K_k|M_m) \quad (2\text{-}16)$$

上式中，N 为特征词总数，M 为文档数目，K 为主题总数，且$n \in N$，$m \in M$，$k \in K$。

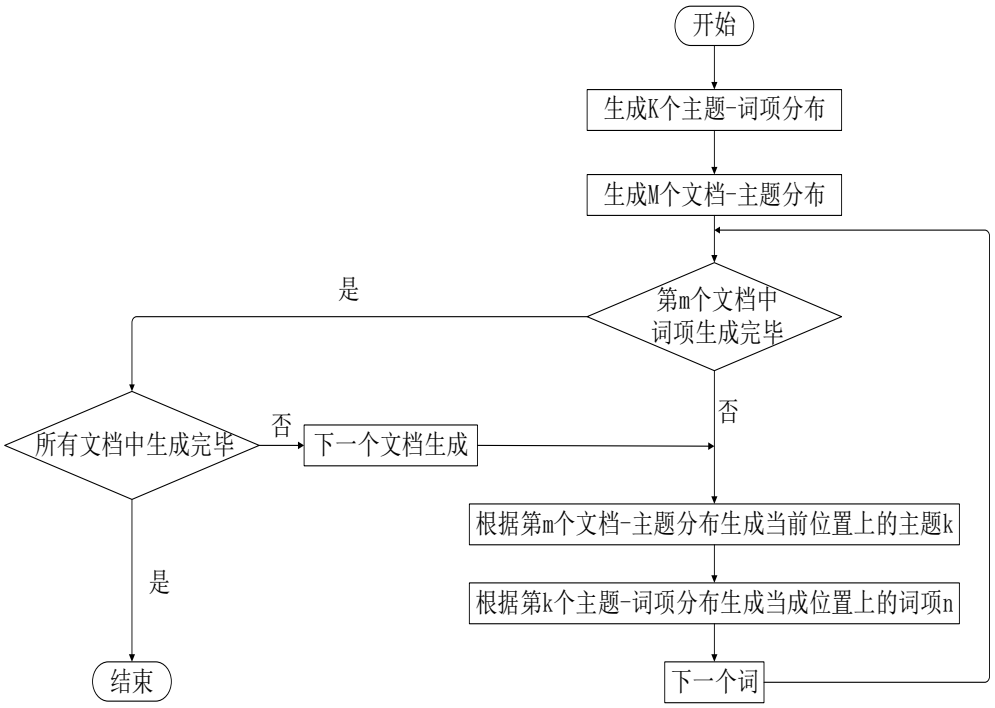

图 2-4 LDA 主题文档生成流程图

LDA 主题文档生成过程如图 xxx 所示。对于“文档-主题”多项式分布过程，该分布必须服从参数为α的 Dirichlet 先验分布（Dirichlet 分布是多维 Beta 分布），式（2-17）为 Beta 分布的密度函数。

$$\mathrm{f(p;\alpha,\beta)} = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad (2\text{-}17)$$

其中$B(\alpha,\beta)$表示参数为(α,β)的 Beta 分布，p表示事件发生的概率。K 维的 Dirichlet 分

布式如式（2-18）所示。

$$\text{Dirichlet}(\vec{p}|\vec{\alpha}) = \frac{\tau(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\tau(\alpha_k)}\prod_{k=1}^{K}P_k^{\alpha_k-1} \quad (2\text{-}18)$$

由此可见，Dirichlet 分布在二维状态下的特殊形式就是 Beta 分布，对参数(α,β)进行确定就是 LDA 模型所要做的，确定过程如图 2-5 所示。

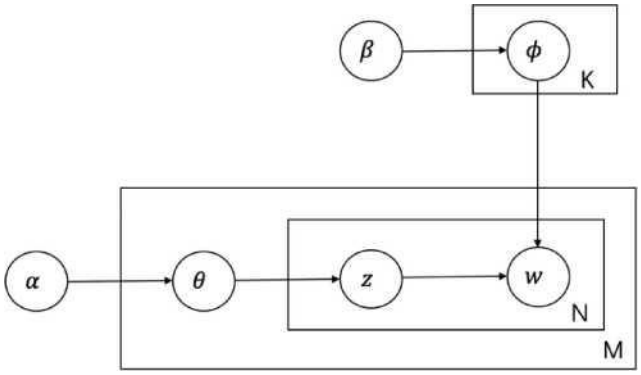

图 2-5　LDA 主题文档生成流程图（参数）

模型图中每个参数如表 2-4 所示：

表 2-4　LDA 模型公式中符号含义

符号	含义
$\emptyset$	词分布，满足$\vec{\emptyset}$~$Dirichlet(\vec{\alpha})$
θ	主题分布，满足$\vec{\theta}$~$Dirichlet(\vec{\beta})$
α	主题分布θ的先验分布(即 Dirichlet 分布)参数
β	词分布$\emptyset$的先验分布参数
N	文档的特征词总数，n∈N
M	文档的总数，m∈M
K	主题总数，k∈K

LDA 主题模型的训练过程，主要是对参数α和β进行训练，其典型代表是 EM 估计和 Gibbs 抽样。

2.6LDA 可视化系统

2004 年，Cason Sievert 和 Kenneth E.Shirley 提出基于 Web 的交互式可视化主题工具—LDA vis。它提供了一个主题挖掘的整体视图，可用于探讨主题及其差异，同时还可以加深与主题高度相关的关键词的理解深度。LDA vis 的关键创新在于根据主题中关键词的相关性

进行排名，并通过描述游客研究的结果来学习相关性计算中的最佳参数。

在 LDA vis 中，通过关键词在主题内的相关性对其进行排名，并且通过描述一项游客研究的结果，以学习相关性计算中的最佳调整参数。

首先令$\emptyset_{kw}$表示主题$k \in \{1,\cdots,K\}$中项$w \in \{1,\cdots,V\}$的概率，其中V表示语料中的词语的个数，同时p_w用于表示词项 w 的边际概率。通常使用变分贝叶斯模型或 Collapsed Gibbs 抽样对 LDA 中的$\emptyset$进行估算，并根据语料库的经验分布估算p_w。给定权重参数$\lambda(0 \le \lambda \le 1)$, 将词项 w 与主题 k 的相关性如式（2-19）计算得到：

$$r(\mathrm{w},\ell \mid \lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\log(\frac{\phi_{kw}}{P_w}) \qquad (2\text{-}19)$$

式（2-19）中确定了主题 k 下项 w 的概率相对于其提升的权重(均在对数刻度上进行测量)。设置$\lambda = 1$会导致关键词按照其主题特定概率的降序排列，而设置$\lambda = 0$则仅根据其升序对关键词进行排名。LDA vis 研究可以为主题解释学习一个“最优”的λ值。

LDA vis 作为一个交互式的可视化系统，试图反映主题模型的关键问题：(1)每个主题的含义是什么?这个依靠对每个主题中的关键词进行展示，通过关键词了解主题想表达的信息。(2)每个主题的流行程度如何？这个问题 LDA vis 中可以展示每个主题中的关键词占所有语料的比例，以此反映该主题在所有语料中的流行度。(3)这些主题之间如何相互关联？首先 LDA vis 可以直观展示每个主题之间的距离，以此判断主题之间的相关性，另外通过主题中的关键词可以了解到主题之间内容的重叠程度。LDA vis 中不同的视觉组件回答了这些问题，其中一些是原始的 LDA 模型中已经包含的信息，而另一些是从 LDA vis 展示中获得的。

3 数据处理

3.1 数据准备

本文共有四个数据集，其中附件一包含景区和酒店评论两个数据集，附件二为专家对景区和酒店的评分。景区评论集共有 59106 条数据，涉及三个指标，包括“景区名称”、“评论日期”与“评论详情”。酒店评论集共有 25225 条数据，涉及四个指标，包括“酒店名称”、“评论日期”、“评论内容”、“入住房型”。专家打分数据包括总得分、服务得分、位置得分、设施得分、卫生得分和性价比得分。

通过随机抽样，将数据集按 8：1：1 分为训练集(train set)，验证集(validation set)和测试

集(test set)，以寻求模型最优参数。其中训练集用来估计模型，验证集用来确定网络结构或者控制模型复杂程度的参数，测试集用来检验最终选择最优的模型的性能如何。

3.2 文本预处理

3.2.1 数据清洗

数据清洗的步骤，主要包括文本去重、压缩去词、短句删除等。但考虑到相似留言会影响后续热度值的计算，口语性重复描述可以度量情绪强烈程度，所以本文仅去除完全相同的数据，且未压缩去词。处理的原则是：利用 Python 程序判断语料库中是否存在完全重复的文本，若存在，则保留一条完全重复文本。

中文表达中，会有几个词意思相同的情况或者是评论文本中出现繁体字。为了解决上述问题，本文使用同义词词库和繁转简词库对数据集进行处理。

经过数据清洗后，景区评论集共去除 305 条完全重复评论，剩余 58801 条有效评论文本。酒店评论集共去除 248 条完全重复评论，剩余 24977 条有效评论文本。

3.2.2 中文分词

1. pkuseg 分词

在人机自然语言交互中，成熟的中文分词算法能够达到更好的自然语言处理效果，帮助计算机理解复杂的中文语言。目前，成熟的分词工具有：Jieba 分词、NLPIR 分词系统、HanLp 分词、Snow NLP、北京大学 PK Use、哈工大 LTP、p useg、Baidu Lac、等。可根据个人不同的需求去选择不同的分词工具。

本文使用北京大学语言计算与机器学习研究组研制推出的一套全新的中文分词 pkuseg 工具包。pkuseg 简单易用，支持细分领域分词，有效提升了分词准确度。照第二届国际汉语分词测评发布的国际中文分词测评标准，Jieba 分词对这两个文本的误差率分别高达 18.55% 和 20.42%，而 pkuseg 的误差率只有 3.25%与 4.32%。随后通过分词效果对比，使用 pkuseg 分词工具所得效果更好。

表 3-1　pkuseg 分词结果展示

评论编号	评论语料
1	(‘酒店’，‘n’),(’在’，‘p’),(‘市区’，‘s’),(‘交通’，‘n’),(‘便利’，‘a’)(‘,’，‘w’),(‘周边’，‘n’),(‘环境’，‘n’),(‘优美’，‘a’),(‘,’，‘w’)……
2	(‘城站’，‘n’),(‘地铁站’，‘n’),(‘D 口’，‘v’),(‘出来’，‘v’),(‘走路’，‘v’),(‘差不多’，‘d’),(‘公里’，‘q’),(‘就’，‘d’),(‘到’，‘v’)……

2. 自定义分词词典—专用词基准库建立

关于虚词的处理，现有的分词方法并不适用于情感分类的研究，因为不带情感的虚词与实词的组合单元也可以表达情感倾向，分词将这组合单元拆分开会造成这部分情感信息的损失。并且不同领域中包含的词汇也不尽相同，例如在酒店和服装评论中，文本的描述就不一样，酒店评论中“方便”、“干净”等词用来表达正面情感的频率较高，而服装评论中用来表达正面情感的词可能是“好看”、“衬肤色”等词。

针对上述问题，本文以“pkuseg”为基本分词工具，借鉴 N-gram 语言模型的特点，以知网 HowNet 情感词典为基础，从景区及酒店领域的评论语料中训练抽取常见的组合单元来构建自定义分词词典。HowNet 词典相比 NTUSD 词典更为丰富，其不仅包含积极情感词和消极情感词，还囊括了各种程度级别词语，将上述词语合并后，得到积极情感词 4366 个和消极情感词 4370 个。褒贬义词典中包含积极情感词 728 个和消极情感词 933 个。

为获取每一个词的 TF-IDF 值，本文采用 Gensim 模块的 Corpora 函数以及 Model 函数进行处理。以 8：1：1 将评论语料分为训练集、测试集和验证集，然后加载 TF-IDF 算法对语料数据进行训练，提取 TF-IDF 值靠前的 100 个词汇作为自定义词典。基于自定义词典重新对酒店及景区的评论文本进行分词，得到更精确的结果。

3.2.3　去停用词

在自然语言处理过程中，文本分析中有大量无用的数词、量词、连词、助词等，例如“啊”，“的”、“和”等，对语义作用很小。因此，过滤停用词能有效提高文本的检索效率和效果，使分类结果更加精确，同时过滤标点符号、常规英文字母与数字以及长度为 1 的词。

本文通过实践发现先进行分词处理，再去除标点、常规英文字母和数字的效果优于先去除后分词。因此分词结束后，选用“哈工大停用词表”去除停用词和标点符号，得到完成初步清洗的景区及酒店数据集。对比表 3-1，得到的最终数据如表 3-2 所示。

表 3-2　去除停用词结果展示

评论编号	评论语料
1	(‘酒店’, ‘n’), (‘在’, ‘p’), (‘市区’, ‘s’), (‘交通’, ‘n’), (‘便利’, ‘a’), (‘周边’, ‘n’, (‘环境’, ‘n’), (‘优美’, ‘a’)……
2	(‘城站’, ‘n’), (‘地铁站’, ‘n’), (‘D 口’, ‘v’), (‘出来’, ‘v’), (‘走路’, ‘v’)), (‘差不多’, ‘q’), (‘公里’, ‘q’), (‘到’, ‘v’)……

3.3 目的地热词提取

3.3.1　高频词提取

1.基于词云图的词频统计

本文使用基于 Python 语言的 World cloud 模块输出酒店及景区评论的词云图。词云图由无数个词语构成，借用艺术化的方式来呈现词语在文本中的重要程度。Word cloud 把词云当作对象，以每个词语的频率为参照依据进行绘制，通过字号大小、颜色、形状、位置属性等参数来表达某一词语在文本的重要性。绘制一个词云图要先配置相关属性的参数，包括背景、字体、颜色等，加载文本，设定存储格式，最后生成词云图保存到本地。

提取出所有评论主题的关键词后，汇总并统计各个关键词的词频，设定阈值为 4，将词频大于阈值的关键词设为筛选词集，绘制为筛选词集词云图。以 A01、H01 为例，词云图中可以看出，酒店评论中“不错”、“服务”、“位置”、“方便”等词具有较高的词频。景区评论中，“取票”、“好玩”，“景区”，“值得”等词具有较高的词频。

图 3-1　景区词云图

图 3-2　酒店词云图

2.基于 TF-IDF 算法的关键词提取

TF-IDF(词频-逆文档频次算法)是一种基于统计的计算方法，常用来评估在一个文档中一个词对某文档的重要程度，这显然很符合常理。一个词对文档越重要，那么这个词就越可能是文档的关键词。因此选取每个景区及酒店中 TF-IDF 值靠前的词作为备选词库，后续再进行筛选。

表 3-3　景区及酒店评论中的高频词关键词（示例）

A01 为例					H01 为例			
序号	词语	词频	TF-IDF 值		序号	词语	词频	TF-IDF 值
1	动物	1765	0.081117531		1	不错	432	0.450846647
2	方便	1503	0.147752913		2	服务	385	0.373511117
3	取票	1294	0.207523507		3	方便	250	0.287520708
4	不错	992	0.094562323		4	位置	204	0.226718841
5	表演	902	0.076867148		5	房间	129	0.162022547
6	好玩	829	0.115895736		6	交通	117	0.136230817
7	马戏	794	0.116810486		7	环境	107	0.108592931
8	值得	688	0.065974746		8	东站	102	0.20198487
9	开心	651	0.087848857		9	前台	99	0.172266743
10	孩子	595	0.037499678		10	干净	82	0.112833171

3.基于 Text Rank 算法的关键词提取

Text Rank 算法是一种用于文本的基于图的排序算法，其基本思想来源于 PageRank 算法。通过词之间的相邻关系构建网络，然后用 PageRank 迭代计算每个节点的 rank 值，排序

rank 值即可得到关键词。PageRank 本来是用来解决网页排名的问题，网页之间的链接关系即为图的边，迭代计算公式如下：

$$PR(\mathrm{Vi}) = (1-\mathrm{d}) + \mathrm{d} * \sum_{\mathrm{j} \in In(\mathrm{V_i})} \frac{1}{|\mathrm{Out}(\mathrm{V_j})|} PR(\mathrm{V_j})$$

其中，$PR(\mathrm{Vi})$ 表示结点 Vi 的 rank 值，$In(\mathrm{V_i})$ 表示结点 Vi 的前驱结点集合，$\mathrm{Out}(\mathrm{V_j})$ 表示结点 $\mathrm{V_j}$ 的后继结点集合，d 为 damping factor 用于做平滑。

评论详情体现游客的情感倾向，但语句冗长，语言多样，难以把握中心思想，因此本文利用 Text Rank 算法对数据集每条评论提取关键词。下表为提取出的评论关键词示例。

表 3-4　景区评论详情关键词（示例）

景区名称	评论详情	关键词
A01	1.小火车很值得坐，缆车不太值得玩。2.动物园里的项目越来越多了。3.从北门进，从南门出，这样正好坐免费穿梭车去大马戏，非常方便。4.大马戏表演很精彩，很多精彩表演，超级赞！	马戏精彩表演值得缆车穿梭北门动物园南门超级火车免费正好方便项目很多
A02	非常不错，圆满，小孩玩的很开心，就是一定得排队，整体环境很棒，尤其是烟火，服务人员，都很有耐心，和认真，期待下次再去玩。	不错圆满小孩开心环境烟火耐心认真期待

表 3-5　酒店评论详情关键词（示例）

酒店名称	评论详情	关键词
H01	服务态度也非常棒！而且酒店管理在不断进步中！下次还会入住	入住 进步 态度 服务 不断
H02	房间卫生间设施是差了点，特别像快捷酒店。房间装修还算舒适，也不觉得特别老旧，但是霉味太重了，从走出电梯就闻到，进入房间特别不适应。个人觉得，服务还行吧，笑脸相迎，早餐也不错，中式西式都有几样，对于早餐很可以了。	房间 早餐 特别 霉味 笑脸 相迎 西式 老旧 中式 卫生 间 快捷 电梯 装修 舒适 设施 适应 不错 个人 服务

3.3.2 相关度度量

由于词频统计分析仅分析众多关键词在评论文本中出现的次数，无法反映这些出关键词之间的关联性。共词分析能统计出一对词语同时出现的频率，若一起出现的频率越高，则这对词语之间的联系越紧密。共词分析的工作原理为:在词频统计的基础上，进行聚类分析，从而挖掘出该文章的主题结构。该方法能挖掘出一个主题词语与另一个主题词语之间在同一领域中的关联性。

通过研究，本次构建共现矩阵的基本思路为:首先构建一个单位矩阵;然后通过算法统计出目标对象的共现情况；最后在对应的矩阵元素中输入结果。对于这种矩阵来说，矩阵的索引、列名使用对应的目标对象的名称会比较方便，因而本文使用 Python 中 Pandas 库的 DataFrame 来构建矩阵。下面以 A01、H01 为例展示评论的共词矩阵。

	取票	方便	马戏	好玩	满意	开心	动物	表演	排队	精彩	过山车	便宜	小朋友	刺激	火车	游玩	门票	订票	乐园	垂直
取票	1083	897	157	142	27	146	260	175	164	110	96	197	70	99	104	65	62	103	54	81
方便	897	1372	189	169	34	194	320	195	182	138	93	263	79	105	114	101	93	130	69	76
马戏	157	189	621	51	12	62	242	199	46	216	28	46	45	55	54	42	34	23	42	15
好玩	142	169	51	752	9	87	120	75	96	36	115	51	34	161	33	24	29	20	59	92
满意	27	34	12	9	81	11	26	19	11	6	8	3	11	6	9	7	9	5	5	3
开心	146	194	62	87	11	612	147	79	55	42	38	43	74	59	59	42	26	27	30	29
动物	260	320	242	120	26	147	1080	307	79	160	15	90	115	46	231	94	74	54	67	10
表演	175	195	199	75	19	79	307	656	75	226	58	37	70	87	114	55	41	33	35	42
排队	164	182	46	96	11	55	79	75	447	28	87	47	27	99	52	36	35	31	37	79
精彩	110	138	216	36	6	42	160	226	28	448	19	29	36	50	41	27	19	17	13	13
过山车	96	93	28	115	8	38	15	58	87	19	335	29	10	186	5	10	17	9	32	250
便宜	197	263	46	51	3	43	90	37	47	29	29	426	15	33	24	18	44	45	28	28
小朋友	70	79	45	34	11	74	115	70	27	36	10	15	278	20	38	22	16	12	12	7
刺激	99	105	55	161	6	59	46	87	99	50	186	33	20	460	7	19	23	12	45	150
火车	104	114	54	33	9	59	231	114	52	41	5	24	38	7	322	39	32	16	12	4
游玩	65	101	42	24	7	42	94	55	36	27	10	18	22	19	39	227	23	15	23	7
门票	62	93	34	29	9	26	74	41	35	19	17	44	16	23	32	23	224	13	20	17
订票	103	130	23	20	5	27	54	33	31	17	9	45	12	12	16	15	13	165	7	9
乐园	54	69	42	59	5	30	67	35	37	13	32	28	12	45	12	23	20	7	241	27
垂直	81	76	15	92	3	29	10	42	79	13	250	28	7	150	4	7	17	9	27	272

图 3-3　景区评论中的共词矩阵

	满意	服务	方便	位置	东站	前台	房间	交通	早餐	干净	环境	性价比	便利	卫生	设施	态度	地理	出差	地铁	出行	五星级	舒适	周边	热情
满意	30	10	7	9	1	3	8	5	0	8	4	3	5	5	0	3	1	1	1	3	0	1	0	2
服务	10	362	57	54	17	75	44	35	20	32	34	10	16	30	18	62	15	10	8	6	3	9	3	20
方便	7	57	233	91	63	13	33	73	25	21	18	26	19	18	21	4	16	15	30	31	7	8	20	9
位置	9	54	91	204	33	17	29	18	17	15	16	17	36	9	20	7	51	9	27	16	4	8	8	7
东站	1	17	63	33	96	6	19	20	20	6	2	11	10	6	14	3	6	12	17	10	5	2	12	4
前台	3	75	13	17	6	96	16	5	6	8	4	3	6	8	6	15	3	1	3	1	3	2	2	14
房间	8	44	33	29	19	16	##	14	24	29	7	7	4	20	18	4	10	5	6	5	9	7	5	13
交通	5	35	73	18	20	5	14	117	12	16	10	12	48	15	12	2	3	4	3	3	1	2	6	7
早餐	0	20	25	17	20	6	24	12	70	4	6	8	5	4	10	4	4	7	8	3	5	3	4	3
干净	8	32	21	15	6	8	29	16	4	82	9	5	7	40	6	5	6	4	4	3	5	5	4	8
环境	4	34	18	16	2	4	7	10	6	9	106	7	4	14	7	7	7	4	6	2	1	9	4	2
性价比	3	10	26	17	11	3	7	12	8	5	7	66	6	6	5	2	3	3	6	6	5	1	6	1
便利	5	16	19	36	10	6	4	48	5	7	4	6	75	8	5	1	2	2	8	2	1	1	1	5
卫生	5	30	18	9	6	8	20	15	4	40	14	6	8	75	4	3	6	2	2	2	7	2	3	8
设施	0	18	21	20	14	6	18	12	10	6	7	5	5	4	59	2	6	2	6	3	3	3	7	5
态度	3	62	4	7	3	15	4	2	4	5	7	2	1	3	2	67	4	1	2	1	0	5	0	0
地理	1	15	16	51	6	3	10	3	4	6	7	3	2	6	6	4	52	2	4	1	0	1	2	1
出差	1	10	15	9	12	1	5	4	7	4	4	3	2	2	2	1	2	39	3	0	0	1	2	0
地铁	1	8	30	27	17	3	6	3	8	4	6	6	8	2	6	2	4	3	52	7	2	0	5	1
出行	3	6	31	16	10	1	5	3	3	3	2	6	2	2	3	1	1	0	7	34	0	0	3	4
五星级	0	3	7	4	5	3	9	1	5	5	1	5	1	7	3	0	0	0	2	0	26	1	3	0
舒适	1	9	8	8	2	2	7	2	3	5	9	1	1	2	3	5	1	1	0	0	1	31	1	0
周边	0	3	20	8	12	2	5	6	4	4	4	6	1	3	7	0	2	2	5	3	3	1	28	0
热情	2	20	9	7	4	14	13	7	3	8	2	1	5	8	5	0	1	0	1	4	0	0	0	28

图 3-4　酒店评论中的共词矩阵

3.3.3 语义网络分析

基于语义关系的可视化利用句法结构分析、语义关系分析以及自语言理解等技术方法在词频分析的基础上，挖掘词语间的内在语义联系后以可视化形式呈现。常用的可视化形式，如树状图、网络图等，典型的应用有 Word Tree，Phrase Net，NLP-Win 等。

1968 年美国著名心理学家 R. Quilian 第一次提出，语义网络是一种以网络格式表达某种语言的概念及关系的方法。随后 1972 年，美国人工智能领域的专家 R.E Simmons 和 J. SIo 率先将其运用到自然语言处理工作中。与向量空间模型不同，利用语义网络分析可以挖掘出词项之间的语义关联，在一定程度上可以将由分词所导致的凌乱的文本结构关系重新整合，从而还原出单独词项无法表达出的部分原始文本信息。

语义网络由众多的语义单元构成，语义单元又叫语义基元。语义基元可借助三元组来描述，如图 5-1 所示，结点 A、弧 R 和结点 B 构成了一个三元组，弧 R 代表结点 A,B 之间的语义关系，语义网络由此生成。语义网络通常以图的形式表现，称之为语义网络图，语义网络图实际上是用于描述词项之间关系的有向图，通常由多个代表词项的节点和代表语义关联的有向弧所构成。

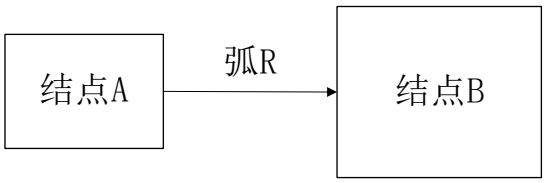

图 3-5 语义基元的基本结构

在研究景区及酒店评论的文本数据时，语义网络图常被用于提取游客选择目的地的关注点。由于中文分词会打乱原来的语句结构、语义关系，通过语义网络能够很好的重建语义之间的联系，从中得出潜藏的信息。例如:当“酒店”与“干净”被切分开，两者间的语义关系就被打乱，单个的“酒店”或“干净”存在很多理解方式，通过其中的一个词语无法厘清两者之间的内在联系。只有当“酒店”和“干净”通过语义网络连接，人们能够清楚的知道该酒店住宿环境很干净。

本文使用 Networkx 生成所需语义网络图。Networkx 作为一个图论与复杂网络建模工具，基于 Python 语言，内部配置了标准的图与复杂网络的分析算法，能够胜任仿真建模、复杂网络数据分析等工作。此外，Networkx 也包含了常用的图论算法，节点可设置成任意数据，随意调节边值维度，功能强大且操作简单易上手。借助 Networkx 可实现分析网络的结构、

构建网络的模型、设计新的网络算法、绘制网络等功能。在存储方式上，既能存储标准化的数据格式，也能存储非标准化的数据格式。下面以 H01 为例展示景区及酒店的语义网络图。

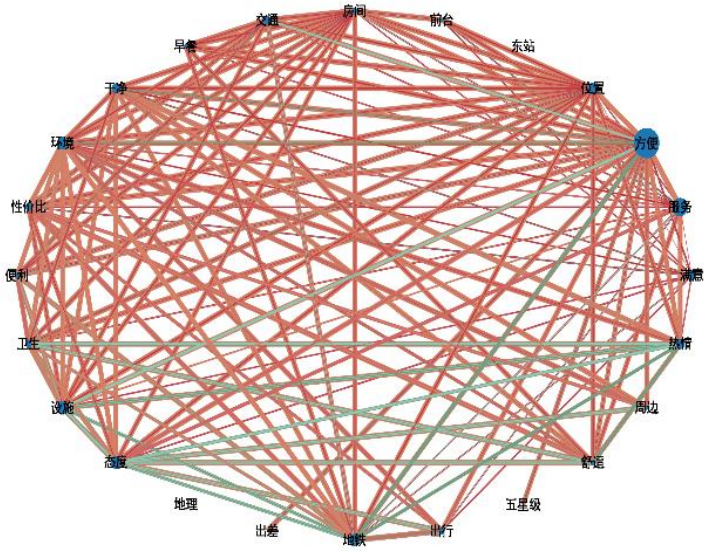

图 3-6　酒店语义网络图

3.4结果分析

本文将“噪声评论”定义为：“评论主题”的关键词在全部评论中出现频数小、较少人关注的评论。因为噪声评论不满足成为热词的条件，所以在原始数据集中剔除噪声评论不仅能减少相似矩阵的维度，还能提高后续热点抽取的效果。具体步骤为：对每个景区和酒店的每条评论提取关键词，提取前 30 个关键词汇总成筛选词集；然后利用 Python 程序对评论主题的关键词进行遍历，其中不包含筛选词集内关键词的评论文本，视为噪声评论，将其删除。

经过以上操作，景区评论集共去除噪声评论 473 条，剩余评论 58633 条；酒店评论集共去除噪声评论 1514 条，剩余评论 23711 条。

表 3-6　噪声评论（示例）

景区名称	噪声评论		酒店名称	噪声评论
A01	哈哈哈哈哈哈过哈哈哈哈哈哈哈哈		H02	还哈哈哈哈哈哈
A03	还吃哈哈哈哈哈哈哈哈哈		H03	一般般吧

4 游客满意度综合评价模型

本章研究包括 Word2Vec、基于监督学习的特征分类算法、情感分析及模糊综合评价，并结合游客满意度相关理论构建基于情感分析和模糊综合评价的满意度评价模型，得到各特征维度得分。通过得分能够清晰地分析出游客的属性特征偏好及其情感倾向对满意度的影响，从而帮助经营者调整其经营模式，提供更好的服务。

4.1 热词挖掘

为找到目标景区和酒店 TOP20 的热门词，本文在提取关键词并进行语义分析后，设计了一套热度评价指标体系，进行热度度量，取排名前 20 的词作为目的地印象词。

4.1.1 定义热度指标

热词，即热门词汇，作为一种词汇现象，反映了特定人群在某段时间普遍关注的问题和事物。由此可见，在定义热度评价指标时，需要综合考虑时间段、相关评论数量、以及某一问题的情感倾向性。通过查阅相关文献后，本文选取以下指标进行热度评价，针对每一个景区及酒店，分别计算其热度指标：

（1）该目的地的**评论数量 n**（x_1），评论数量是热度的重要表现；

（2）该目的地最早评论日期与最晚评论日期的**间隔天数 m**（x_2），热点问题往往在较短时间内集中产生；

（3）该目的地评论的**正向情感次数 a**（x3）与**负向情感次数 b**（x4），评论中情感词数越多，也反映出更多的关注度。

则目的地的热度公式如下式（4-1）：

$$L=\sqrt{\frac{\sum_{i=1}^{n}(10+\frac{a+b}{5})}{4[1+\log_2(m+5)]}} \tag{4-1}$$

4.1.2 热度计算

基于以上提出的热度评价指标体系，计算出各景区、酒店的热度值，如下表 4-1 和表 4-2。

表 4-1　景区热度值

景区名称	热度	景区名称	热度	景区名称	热度
A01	35.36996	A18	15.12066	A35	9.36119
A02	34.21964	A19	15.00475	A36	9.27769
A03	31.61870	A20	14.73611	A37	8.81371
A04	29.49110	A21	14.58783	A38	8.96962
A05	29.82034	A22	14.21054	A39	8.42665
A06	29.66963	A23	13.03808	A40	8.81400
A07	28.14819	A24	12.86179	A41	8.59658
A08	25.48302	A25	12.65464	A42	8.38079
A09	25.39642	A26	12.62610	A43	8.49943
A10	21.65671	A27	12.34641	A44	6.24474
A11	21.25444	A28	12.01280	A45	8.06958
A12	18.02373	A29	12.29612	A46	7.97013
A13	17.15827	A30	11.61656	A47	7.92809
A14	16.66240	A31	11.72009	A48	7.90538
A15	16.62648	A32	11.09668	A49	7.82983
A16	15.65846	A33	10.90965	A50	7.98515
A17	15.51925	A34	9.45457		

表 4-2　酒店热度值

酒店名称	热度	酒店名称	热度	酒店名称	热度
H01	17.97897	H18	12.87909	H35	10.41664
H02	17.41939	H19	12.14572	H36	10.70115
H03	17.59162	H20	12.8579	H37	10.15603
H04	17.38139	H21	12.39846	H38	10.87500
H05	16.13271	H22	12.10528	H39	10.44567
H06	15.98453	H23	12.17335	H40	10.98087

H07	15.23584	H24	12.48979	H41	12.76710
H08	14.29081	H25	11.68050	H42	10.44808
H09	15.20004	H26	11.85800	H43	10.35190
H10	13.78907	H27	12.44204	H44	9.96976
H11	13.44399	H28	11.26544	H45	9.05950
H12	13.05850	H29	12.18820	H46	9.70177
H13	14.91516	H30	11.29400	H47	9.78690
H14	12.98774	H31	11.04485	H48	9.02822
H15	14.29242	H32	11.40995	H49	9.47315
H16	12.66893	H33	10.80398	H50	9.62064
H17	12.73461	H34	11.50054		

再利用公式L_1=L ×(词的TF-IDF值)计算每个词的热度值，最后得到每个景区及酒店中热门词的热度。

表 4-3　景区印象词云表（A01 为例）

评论热词	TF-IDF 值	热度	评论热词	TF-IDF 值	热度
取票	0.207523507	7.340098208	排队	0.063600972	2.249563863
方便	0.147752913	5.226014674	精彩	0.061688151	2.181907446
马戏	0.116810486	4.131582263	景区	0.057819066	2.045058057
好玩	0.115895736	4.099227575	过山车	0.053615401	1.896374598
不错	0.094562323	3.344665614	便宜	0.051608314	1.825383999
开心	0.087848857	3.10721059	小朋友	0.047878333	1.693454739
动物	0.081117531	2.869123831	刺激	0.040427761	1.429928289
表演	0.076867148	2.718787977	火车	0.039800405	1.407738726
动物园	0.06876864	2.432344074	园区	0.039456773	1.395584504
值得	0.065974746	2.333524132	游玩	0.036259533	1.282498248

表 4-4　酒店印象词云表（H01 为例）

评论热词	TF-IDF 值	热度	评论热词	TF-IDF 值	热度
不错	0.450846647	8.105756148	环境	0.108592931	1.952388525
服务	0.373511117	6.715343347	性价比	0.1066281	1.917062898
方便	0.287520708	5.169324788	便利	0.103995118	1.869724602
位置	0.226718841	4.076170134	卫生	0.098285124	1.767064819
东站	0.20198487	3.631478931	入住	0.095614321	1.719046544
前台	0.172266743	3.097177765	设施	0.07918073	1.42358759
房间	0.162022547	2.912997725	态度	0.078494012	1.411241096
交通	0.136230817	2.449289102	地理	0.068854857	1.237939075
早餐	0.117864491	2.119081577	出差	0.06663858	1.198092704
干净	0.112833171	2.028623638	地铁	0.052997988	0.952848971

4.2游客满意度因素分析

获取景区及酒店游客满意度影响因素的过程是在满意度理论和相关研究的基础上进行的。首先利用 TF-IDF 算法获得满意度二级影响因素，然后再借助 Word2Vec 模型对二级影响因素词汇进行向量化处理，最后采用凝聚层次聚类获取词向量的聚类结果。根据聚类结果定义满意度一级影响因素。具体流程如图 4-1 所示。

图 4-1　确定景区及酒店游客满意度影响因素流程图

4.2.1　特征提取

机器学习无法直接将文本作为特征信息进行处理，而计算机只能处理数值型的数据，因此在进行一级影响因素的提取时，需先将文本转化为可表示的向量。本文提取景区及酒店 TF-IDF 值靠前的词汇作为游客满意度二级影响因素，并采用 Word2Vec 对上文中得到的二级特征词汇进行向量化处理，以得到每一个特征词汇的向量表示。

Word2Vec 是一种利用深度学习思想将词表征为实数值向量的高效算法模型，把对文本内容的处理简化为 K 维向量空间中的向量运算，同时文本语义上的相似度被转化为空间向量上的相似度[8]。其核心思想是利用训练将每个词映射成 K 维实数向量，通过词之间的距离来判断它们之间的语义相似度。Word2Vec 可通过 CBOW 模型和 Skip-gram 模型两种方法实现，它们分别基于 HierarchicalSoftmax 和 NegativeSampling 进行设计，两个模型都包含三层：输入层、投影层和输出层，如图 4-2 所示。

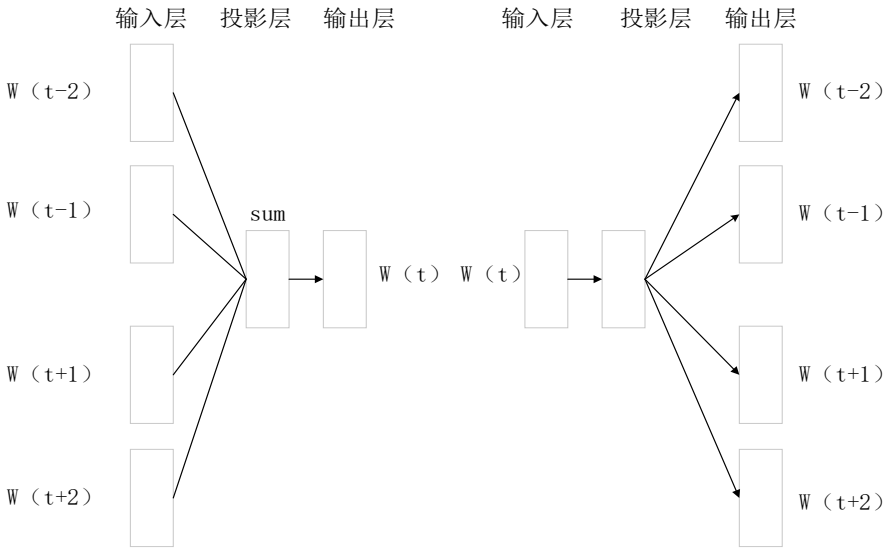

图 4-2　Word2Vec 模型展示（a）CBOW 模型（b）Skip-gram 模型

本文利用 Word2Vec 模型对景区及酒店游客满意度二级影响因素词汇进行向量化处理。Word2Vec 模型可以很好地利用词的上下文信息及句子的内部结构信息，基于神经网络模型将词映射成一个低维、稠密的实数向量。

4.2.2　文本聚类法提取影响因素

本文选用 K-means 聚类、AP 聚类、高斯混合聚类（GMM）、凝聚层次四种不同的方法对各景区及酒店进行二级影响因素聚类。采用轮廓系数评估聚类质量。轮廓系数取值在[-1,1]，越接近于 1，所包含的簇越紧凑，并且远离其他簇，这是可取的情况。反之，当轮廓系数为负值时，这种情况不可取。

四种方法的聚类结果如下表显示，凝聚层次聚类的效果最优，并结合各聚类算法的特点，下文采用凝聚层次法进行聚类。下面以 A01 为例，展示聚类效果评估。

表 4-5　四种聚类算法效果评估

聚类方法	K-means	AP	GMM	HAC
轮廓系数	0.238	0.231	0.194	0.307

基于 Rust 和 Oliver 的服务质量三重属性计量模型，对提炼出的 50 个满意度评价影响因素进行深入分析，从服务、位置、设施、卫生、性价比五个维度对指标进行初步筛选。

4.2.3　内容分析法提取影响因素

内容分析是大众传播研究的内容和方法之一，通过对大众传播内容量和质的分析，能够认识和判断某一时期的传播重点和某些问题的倾向、态度、立场，以及传播内容在某一时期的变化规律等。前文对景区及酒店的游客观点进行综合聚类，以 H01 酒店为例，共提炼出 50 个满意度评价影响因素。接下来，利用内容分析法中的定性分析，来解读、判断和挖掘信息中所蕴涵的本质内容，从而对提取出的二级指标影响因素进行调整。

表 4-6　评论文本分析样例

原始评论	语义片段	涉及因素
酒店地理位置非常好！地铁口出来 100	“地理位置非常好”	地理位置
米。服务态度也非常棒！而且酒店管理	“服务态度也非常棒”	服务态度
在不断进步中！	“酒店管理在不断进步”	服务质量

本文综合机器学习聚类与内容分析结果，最终将景区、酒店游客满意度影响因素分为 5 类，分别是服务、位置、设施、卫生、性价比。具体综合评价指标体系如下表所示如下表 4-7 和表 4-8 所示。

表 4-7　景区游客满意度影响因素评价体系

一级影响因素	二级影响因素
服务	取票、态度、导游、服务、讲解、订票、方便、指引
位置	轻轨、景色、山、沙滩、小岛、海边、环境、地铁
设施	缆车、索道、滑道、过山车、温泉、餐厅、停车场、酒店
卫生	干净、清洁、卫生间、垃圾桶、
性价比	门票、餐饮、票价、实惠、特惠票

表 4-8　酒店游客满意度影响因素评价体系

一级影响因素	二级影响因素
服务	前台、态度、服务、热情、餐厅、周到、早餐
位置	交通、位置、便利、地理、东站、机场、环境、周边、出行、地铁
设施	房间、温泉、大堂、停车、游泳池、装修、风景、设施
卫生	干净、舒适、卫生、整洁
性价比	性价比、价格、实惠、五星

4.3游客满意度评价模型

4.3.1　基于情感分析的游客满意度评价模型

1.拆分评论短句

每条评论可能包含不止一个影响因素，可能既包含服务又包含卫生，为方便后续处理，本文将评论划分为短句，示例结果如表 4-9 所示：

表 4-9　评论短句拆分结果示例

处理前评论	拆分后评论	去除不包含特征因素的短句
“服务态度挺好，尤其前台 c2022 服务员，取票很快，小朋友很喜欢，下次再来”	“服务态度挺好/尤其前台 c2022 服务员/取票很快/小朋友很喜欢/下次再来”	“服务态度挺好/尤其前台 c2022 服务员/取票很快/小朋友很喜欢/下次再来”

2.情感得分计算

在测试过程中，本文应用情感倾向分析接口对包含主观信息的文本进行情感倾向性类别(积极、消极和中性)的判断。例如评论：“挺好，真的现在还在这里准备出发去珠海，这是第一站，熊猫酒店的各种服务好的没的说，感觉很舒服，一楼大厅还时不时有活动，有玩的地方，离得马戏团很近，就是有点热。”该评论的计算处理结果如表 4-10 所示。

表 4-10　情感倾向性分析示例

属性	positive_prob	negative_prob	confidence	sentiment
数值	0.96857	0.03143	0.98424	2

其中，positive_prob 代表此条评论的积极类别概率为，negative_prob 代表此条评论的消极类别概率为，confidence 表示此条评论的置信度为。sentiment 为代表此条评论的情感倾向为正向，sentiment 有三种取值， 0 表示负向，1 表示中性，2 表示正向。

满意度模型各指标构建如下所示：

步骤一：计算二级影响因素满意度得分s_j，s_j为第 j 个二级影响因素的满意度得分，由于量化后的满意度影响因素情感极性值在 0-1 之间，为了统一性，本文满意度按照 5 分制满分的评价标准进行计算，公式如下所示：

$$s_j=5\times\left(\overline{\sum_{k=1}^{p}\left(positive\_prob\right)_{jk}}-\overline{\sum_{k=1}^{p}\left(negative\_prob\right)_{jk}}\right) \tag{4-2}$$

其中$\overline{\sum_{k=1}^{p}\left(positive_prob\right)_{jk}}$为第 j 个二级影响因素所包含评价单元的积极概率和的平均值，$\overline{\sum_{k=1}^{p}\left(negative_prob\right)_{jk}}$为消极概率和的平均值。

步骤二：计算二级影响因素权重ω_{ij}，ω_{ij}为第i个一级影响因素下第 j 个二级影响因素测评指标的权重，公式如下所示：

$$\omega_{ij}=\frac{\left(TF\_IDF\right)_{ij}}{\sum_{j=1}^{m}\left(TF\_IDF\right)_{ij}} \tag{4-3}$$

其中$\sum_{j=1}^{m}\left(TF_IDF\right)_{ij}$为第$i$个一级影响因素下第 j 个二级影响因素的TF_IDF值的和，$\left(TF_IDF\right)_{ij}$为第i个一级影响因素下第 j 个二级影响因素的TF_IDF值。

步骤三：计算一级影响因素权重ω_i，ω_i为一级游客满意度影响因素中第i个测评指标权重，公式如下所示：

$$\omega_i=\frac{\overline{\sum_{j=1}^{m}\left(TF\_IDF\right)_{ij}}}{\sum_{i=1}^{n}\overline{\sum_{j=1}^{m}\left(TF\_IDF\right)_{ij}}} \tag{4-4}$$

其中，$\overline{\sum_{j=1}^{m}(TF\_IDF)_{ij}}$ 为所有一级影响因素的 TF_IDF 值的和，$\overline{\sum_{j=1}^{m}(TF\_IDF)_{ij}}$ 为第 i 个一级影响因素的 TF_IDF 值。

步骤四：构建基于情感分析的游客满意度评估模型，如公式（）所示，用于计算整体满意度得分 $HCSI$。

$$HCSI=\sum_{i=1}^{n}\left(\omega_{ij}\sum_{j=1}^{m}\omega_{ij}s_j\right) \tag{4-5}$$

3.满意度得分

综合以上结果并结合游客满意度相关理论及研究，确定评价指标体系。以情感极性值为基础来确定二级影响因素的满意度得分，采用预处理中 TF-IDF 算法得到的结果作为基础，计算两级影响因素的权重。根据二级指标的权重得到一级指标得分，最后根据一级指标权重计算最后得分。最终所得分数与游客满意度成正比，分数越高，游客满意度越高。

运用基于情感分析的酒店游客满意度评估模型，对本文的数据进行满意度评估计算，得到景区及酒店游客满意度一级、二级影响因素得分和整体游客满意度得分。具体结果如表所示，其中二级得分保留 2 位小数，其他指标保留 4 位小数。下表 4-11 是以 H01 为例，计算基于情感分析的满意度得分。

表 4-11　各影响因素权值及满意度得分

一级指标	一级权重	二级指标	二级权重	二级得分	$\omega_{ij}s_j$	$\sum\omega_{ij}s_j$	$\omega_i\sum\omega_{ij}s_j$	HCSI
服务	0.2873	前台	0.0392	4.7932	0.1878	4.8441	1.3919	4.4759
		态度	0.1062	4.9618	0.5269			
		服务	0.0647	4.7834	0.3097			
		热情	0.0858	4.8668	0.4174			
		餐厅	0.1977	4.8983	0.9686			
		周到	0.3242	4.7250	1.5318			
		早餐	0.1822	4.9505	0.9019			
位置	0.2506	交通	0.1316	4.1856	0.5508	4.8895	1.2254	
		位置	0.1004	4.2443	0.4263			
		便利	0.2190	4.5191	0.9896			

		地理	0.0665	4.4895	0.2986			
		东站	0.1951	4.1979	0.8190			
		机场	0.1139	4.4963	0.5119			
		环境	0.0334	4.1875	0.1401			
		周边	0.0864	3.7428	0.3235			
		出行	0.1473	4.1981	0.6185			
		地铁	0.0512	4.1254	0.2112			
设施	0.2130	房间	0.0493	4.7526	0.2342	4.7424	1.0103	
		温泉	0.2856	4.6000	1.3139			
		大堂	0.2688	4.7768	1.2842			
		停车	0.0647	4.8722	0.3153			
		游泳池	0.0474	4.7217	0.2240			
		装修	0.1417	4.8050	0.6809			
		风景	0.0524	4.7589	0.2495			
		设施	0.0900	4.8958	0.4405			
卫生	0.1026	干净	0.1481	4.8509	0.7185	4.7788	0.4901	
		舒适	0.3596	4.7262	1.6996			
		卫生	0.1442	4.8246	0.6959			
		整洁	0.3480	4.7835	1.6648			
性价比	0.0859	性价比	0.5120	3.3805	1.7308	4.1689	0.3582	
		价格	0.1910	3.0830	0.5889			
		实惠	0.4628	3.2000	1.4810			
		五星	0.1159	3.1772	0.3682			

分析表 4-11 数据可得，游客在该期间入住酒店的总体满意度得分为 4.48 分，得分率为 0.8952，表明游客景区或酒店服务质量满意度比较高，但仍存在一定的提升空间。表中各影响因素权重的大小代表了游客对该影响因素的特征偏好程度，即酒店经营者应该给予关注度的大小。对于酒店管理者而言，应结合权重值以及满意度得分值对各级影响因素进行具体维度的分析，改善相应的因素以减少游客不满情绪。酒店游客满意度一级影响因素满意度得分

由高到低依次为位置、服务、设施、性价比、卫生，具体得分如图 4-3 所示。

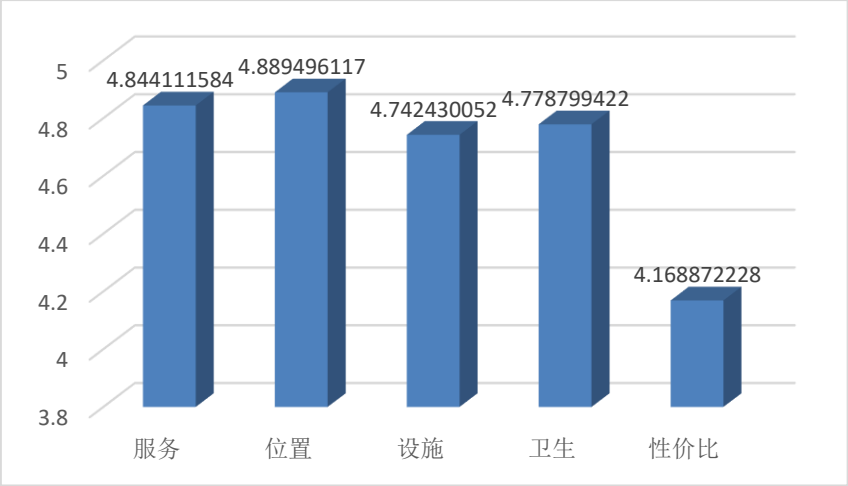

图 4-3　一级影响因素满意度得分

针对酒店游客满意度二级影响因素某一细分特征的满意度得分进行分析，发现交通出行以及酒店价格带来的游客体验较差。有针对性地改善酒店存在的问题能帮助经营者以较小投入获得更高的游客满意度，以游客体验较差的排名后十位的游客满意度影响因素为例绘制条形图，如图 4-4 所示，

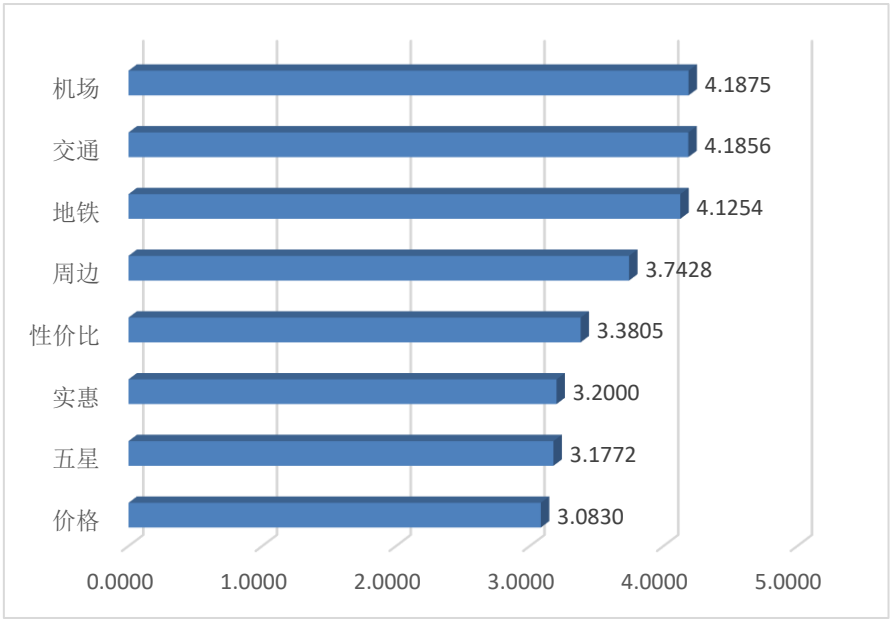

图 4-4　二级影响因素满意度得分

4.满意度划分标准

根据以上结果，景区及酒店的综合得分基本都高于 4 分，因此制定了相关满意度等级，划分如下表 4-12：

表 4-12　得分与满意度等级划分

酒店得分	游客满意度	景区得分
4.0-4.5	低	4.0-4.4
4.5-4.75	中	4.4-4.65
4.75-5.0	高	4.65-5.0

游客满意度为高，则该景区或酒店对应的综合评价为高层次；游客满意度为中，则综合评价等级为中；游客满意度为低，则对应的景区或酒店的综合评价等级为低。

4.3.2　基于模糊综合评价方法的游客满意度评价模型

模糊综合评价以模糊数学为基础，主要是为了将因素实现定量化，进而实现综合评价。通常事物的状况都是与多种因素相关，不能仅依据单一因素进行判断，应当综合考虑可能产生的各种相关指标和因素。本文将运用模糊综合评价方法对各家景区及酒店的游客满意度进行综合评价，计算过程分为三个步骤：1.从上述得出的各体系指标与收集的某一家景区或酒店的数据中确定评价指标集以及评语集。2.通过所收集的数据得到权重向量矩阵 A 以及权重判断矩阵 R 等。3.利用权重向量矩阵和权重判断矩阵对一家景区或酒店的游客满意度体系进行多层模糊计算以得出决策评价。

（一）确定指标权重

指标权重是一个相对的概念，是某一指标在整体评价体系中的相对重要性。目前较为常见的权重确定方法有专家估计法、加权统计法、层次分析法等。本文的游客满意度指标体系中的指标权重是根据各指标在所收集评论中的频次来确定，即收集的所有评论中被提及的频次越高，所占得权重越高。以酒店评论文本集 H01 为例，计算得出所属权重如表所示：

表 4-13　酒店模糊综合评价体系指标权重

一级指标	权重系数	二级指标	权重系数
服务	0.3151	前台	0.1924
		态度	0.0605
		服务	0.2297
		热情	0.1352
		餐厅	0.0357

		周到	0.0539
		早餐	0.1692
位置	0.2724	交通	0.1281
		位置	0.2232
		便利	0.0842
		地理	0.0569
		东站	0.1116
		机场	0.0492
		环境	0.2199
		周边	0.0306
		出行	0.0383
		地铁	0.0580
设施	0.2832	房间	0.1505
		温泉	0.4842
		大堂	0.1463
		停车	0.0484
		游泳池	0.0347
		装修	0.0432
		风景	0.0295
		设施	0.0632
卫生	0.0602	干净	0.3960
		舒适	0.1436
		卫生	0.3416
		整洁	0.1188
性价比	0.0692	性价比	0.3147
		价格	0.2371
		实惠	0.1164
		五星	0.3319

（二）评价体系计算过程

模糊综合评价一般分为一级模糊综合评价和多级模糊综合评价。其中多层次模糊综合评价常用于较为复杂的体系系统[9]。

步骤一：

根据之前收集的 H01 酒店的评价集和各指标所占权重，分别对应数据集。

一级指标评价指标集：U={服务，位置，设施，卫生，性价比}

二级指标评价指标集：

$U_{1=}${前台，态度，服务，热情，餐厅，周到，早餐}

$U_{2=}${交通，位置，便利，地理，东站，机场，环境，周边，出行，地铁}

U_3={房间，温泉，大堂，停车，游泳池，装修，风景，设施}

U_4={干净，舒适，卫生，整洁}

U_5={性价比，价格，实惠，五星}

对应评语集：V={积极，消极}

对应各评价指标集的评价指标权重向量:A=[0.3151,0.2724,0.2832,0.0602,0.0692]

A_1={0.1924,0.0605,0.2297,0.1352,0.0357,0.0539,0.1692}

A_2={0.1281,0.2232,0.0842,0.0569,0.1116,0.0492,0.2199,0.0306,0.0383,0.0580}

A_3={0.1505,0.4842,0.1463,0.0484,0.0347,0.0432,0.0295,0.0632}

A_4={0.3960,0.1436,0.3416,0.1188}

A_5={0.3147,0.2371,0.1164,0.3319}

步骤二：

根据 H01 酒店评论集的数据与对应评语集 V，通过以下公式计算，确定指标的权重判断矩阵 R 。

$$R_i = U_i\text{因素的}V_i\text{评论集} \Big/ U\text{因素的总评价集} \quad (4\text{-}6)$$

表 4-14　二级指标隶属度矩阵

一级指标	二级指标	积极	消极
服务	前台	0.958639175	0.041360825
	态度	0.992357143	0.007642857
	服务	0.956671429	0.043328571
	热情	0.973354839	0.026645161

	餐厅	0.979666667	0.020333333
	周到	0.945	0.055
	早餐	0.990101266	0.009898734
位置	交通	0.837115044	0.162884956
	位置	0.848854839	0.151145161
	便利	0.903818182	0.096181818
	地理	0.897909091	0.102090909
	东站	0.839585586	0.160414414
	机场	0.89925	0.10075
	环境	0.8375	0.1625
	周边	0.7485625	0.2514375
	出行	0.839625	0.160375
	地铁	0.825085714	0.174914286
设施	房间	0.950514019	0.049485981
	温泉	0.92	0.08
	大堂	0.955363636	0.044636364
	停车	0.974444444	0.025555556
	游泳池	0.944333333	0.055666667
	装修	0.961	0.039
	风景	0.95177	0.04823
	设施	0.979153846	0.020846154
卫生	干净	0.97018	0.02982
	舒适	0.94524	0.05476
	卫生	0.964910448	0.035089552
	整洁	0.9567	0.0433
性价比	性价比	0.676098361	0.323901639
	价格	0.616608696	0.383391304
	实惠	0.64	0.36
	五星	0.635434783	0.364565217

根据综合模糊评价法，结合各因素权重及对对应的评价集，计算如下：

根据上表得出二级指标“卫生”的评价矩阵R_4

$$\mathrm{R}_4=\begin{bmatrix} 0.97018 & 0.02982 \\ 0.94524 & 0.05476 \\ 0.96491045 & 0.03509 \\ 0.9567 & 0.0433 \end{bmatrix}$$

通过公式：$U_i = A_i * R_i$ （4-7）

得出“卫生”=[0.963197113　　0.036802887]

采用上述方法和步骤，对各一级指标下的二级指标U_i对应权重系数A_i，计算得出：

“服务”隶属度向量=[0.84925952　　0.02734048];

“位置”隶属度向量=[0.848917078　0.151082922];

“设施”隶属度向量=[0.939692476　0.060307524];

“卫生”隶属度向量=[0.963197113　0.036802887];

“性价比”隶属度向量=[0.64436288 0.35573712]。

步骤三：

通过对应满意状态的得分向量进行计算得到一级指标满意度指数如下：

服务 S1=4.2804732，位置 S2=4.433439043,设施 S3=4.773846784，卫生 S4=4.861989174，性价比 S5=3.666485801。

通过模糊综合评价法计算得出整个 H01 酒店的隶属度向量，计算公式为：

一级指标的权重向量 A • 一级指标隶属度向量=酒店整体隶属度向量 （4-8）

根据满意度得分的计算公式：S=4.454811875。

H01 酒店通过计算得出酒店整体游客满意度指数为 4.45，其中“服务”的游客满意度指数为 4.28，“位置”游客满意度指数为 4.43，“设施”的游客满意度指数为 4.77，“卫生”游客满意度指数为 4.86，“性价比”游客满意度指数为 3.50。

运用上述模型方法及步骤，同理可计算所有景区及酒店的游客满意度指数。

4.4游客满意度评价模型验证分析

为了验证本文构建模型的准确性和合理性，将基于情感分析的游客满意度得分及基于模

糊综合评价方法的游客满意度得分与专家打分进行比较，并运用 MSE 评估模型的合理性。以 H01、A01 为例进行结果展示，如表 4-15 和表 4-16。

表 4-15　H01 酒店得分数据对比表

体系 / 得分	总得分	服务	位置	设施	卫生	性价比	MSE
专家评价	4.8	4.8	4.8	4.7	4.8	4.0	---
情感评价	4.47	4.84	4.88	4.74	4.77	4.16	0.1555
模糊评价	4.45	4.28	4.43	4.77	4.86	3.67	0.3276

表 4-16　A01 景区得分数据对比表

体系 / 得分	总得分	服务	位置	设施	卫生	性价比	MSE
专家评价	4.4	3.8	4.9	4.9	4.5	4.5	---
情感评价	4.53	3.98	4.81	4.87	4.54	4.59	0.1086
模糊评价	4.63	4.21	4.87	4.85	4.62	4.67	0.2131

为了更加直观地展示对比情况，绘制图 4-5 和图 4-6。可以看到，根据均方误差 MSE 说明基于情感分析建立的模型比模糊评价综合评价相比更为合理。

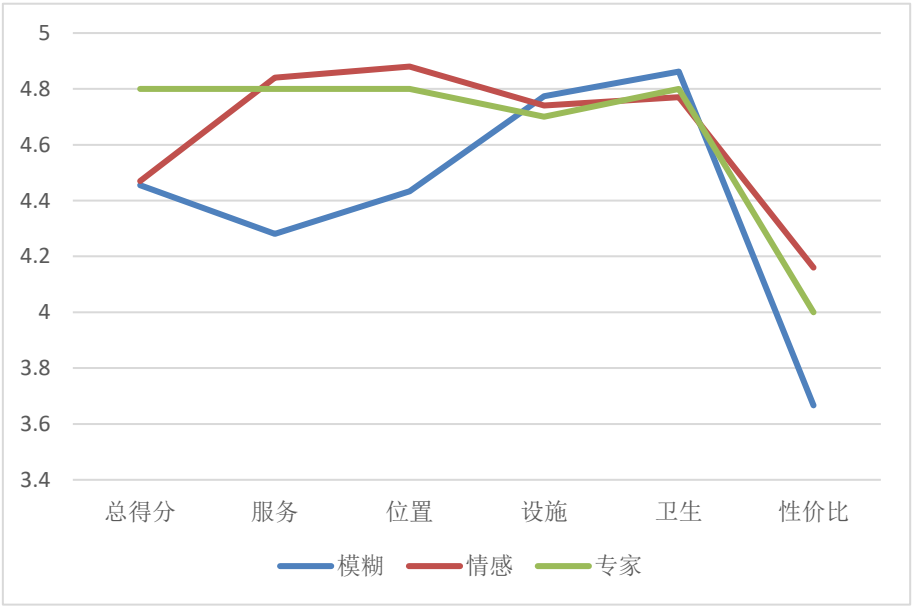

图 4-5　H01 酒店模型得分对比

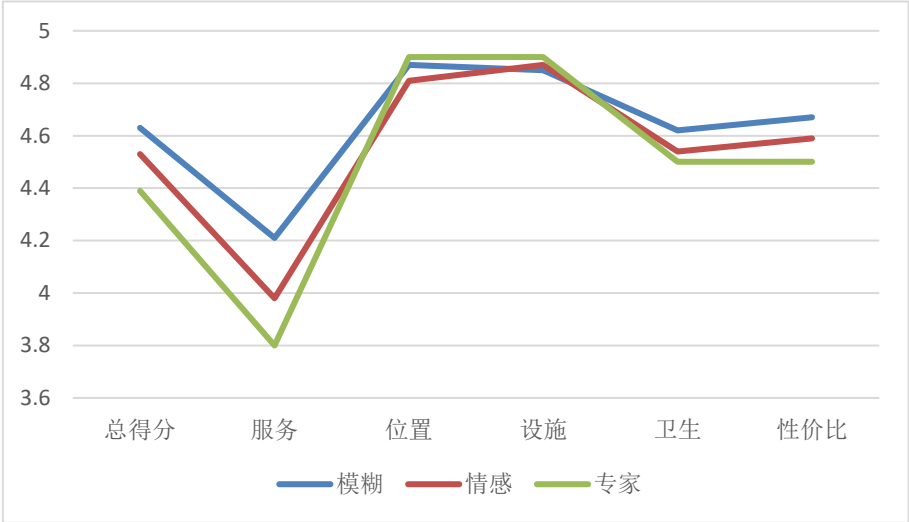

图 4-6　A01 景区模型得分对比

基于情感分析的游客满意度评价模型得到的满意度评分与专家打分趋势基本一致，验证了本文构建模型的合理性和科学性。

依据基于情感分析的游客满意度评价模型，得到 50 家景区和 50 家酒店的情感得分，并按标准将景区和酒店划分为不同层次，结果如下表 4-17。

表 4-17　50 家酒店的情感得分及层次划分结果

低层次酒店		中层次酒店		高层次酒店	
酒店名称	综合情感得分	酒店名称	综合情感得分	酒店名称	综合情感得分
				H03	4.86
		H05	4.62	H04	4.81
		H09	4.53	H06	4.84
H01	4.47	H14	4,61	H07	4.91
H02	4.38	H16	4.71	H10	4.78
H08	4.49	H17	4.58	H11	4.91
H18	4.39	H19	4.55	H12	4.83
H26	4.49	H21	4.72	H13	4.93
H29	4.47	H24	4.62	H15	4.88
H33	4.45	H25	4.68	H20	4.79
H36	4.43	H27	4.69	H22	4.92
H38	4.15	H28	4.59	H23	4.77
H39	4.36	H31	4.63	H30	4.87
H43	4.46	H35	4.71	H32	4.91
H47	4.46	H41	4.74	H34	4.84
H48	4.47	H42	4.59	H37	4.89
H50	4.41	H46	4.65	H40	4.88
		H49	4.57	H44	4.86
				H45	4.79

表 4-18　50 家景区的情感得分及层次划分结果

低层次景区		中层次景区		高层次景区	
景区	综合情感得分	景区	综合情感得分	景区	综合情感得分
				A06	4.66
		A01	4.57		
				A14	4.74
A02	4.19	A07	4.47		
				A13	4.73
A03	4.28	A10	4.42		
				A16	4.68
A04	4.17	A12	4.59		
				A17	4.77
A05	4.15	A20	4.61		
				A21	4.83
A08	4.16	A22	4.63		
				A23	4.89
A09	4.32	A25	4.53		
				A35	4,82
A11	4.07	A26	4.52		
				A36	4.92
A15	4.38	A28	4.49		
				A37	4.81
A18	4.35	A29	4.49		
				A38	4.94
A19	4.26	A30	4.44		
				A39	4.96
A24	4.27	A32	4.52		
				A40	4.79
A27	4.25	A33	4.48		
				A44	4.93
A31	4.33	A34	4.63		
				A45	4.88
A41	4.36	A42	4.44		
				A48	4.79
A43	4.27	A46	4.57		
				A49	4.87
		A47	4.53		
				A50	4.84

5 基于随机森林的信息质量有效性分析

近年来，随着大数据时代的到来，人们不再满足于简单地从互联网上获取信息，而是更加关注互联网能否带来更加准确、纯粹、高效的体验。在旅游业，一方面，海量的旅游评论为游客提供了全面的旅游决策信息；另一方面，海量评论也导致信息过滤的低效率。本章旨在从信息内容质量和信息表达形式质量两个方面的五个评价指标来评论的有效性。每个指标的目的是从不同的深度和广度分析旅游评论的信息质量。此外，如何挖掘高效评论，降低游客的信息搜索成本，快速获取有价值的信息也是本章的重点。

本题所使用的数据集为任务一数据预处理后得到的数据集，包括文本分词、去除特殊字符、去除停用词、去除噪声评论等。基于上述内容，本章流程图如下图 5-1 所示：

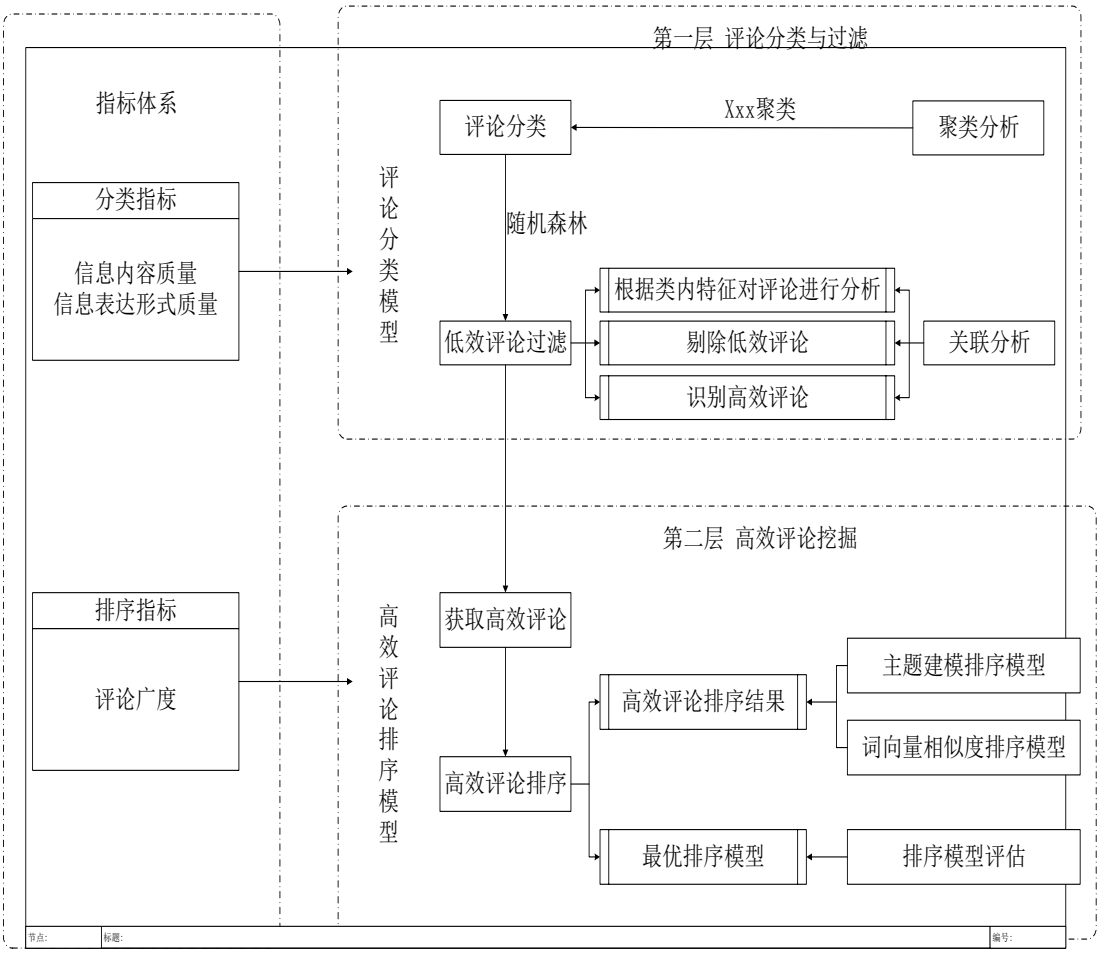

图 5-1 文本有效性研究思路图

5.1文本有效性评价指标体系

信息质量（InformationQuality）从研究数据质量开始。游客在进行旅游目的地选择时，大量低质量的信息容易对其形成误导，因此高质量的评论对于快速有效做出决策发挥着至关重要的作用。

早在 1988 年，RichardWang 和 S.Madnick[10]提出了一个衡量标准，通过由 17 个信息质量维度组成的指标，这尽可能全面地衡量了信息质量。Rogova[11]提出的信息质量本体模型（如图 5-2 所示）很好的填补了信息质量研究领域的缺失，从信息融合设计者的角度阐明了不同维度的信息质量是相互关联的，并从信息表达形式质量、信息内容质量、信息源质量这三个维度进行了详细的信息质量指标研究。

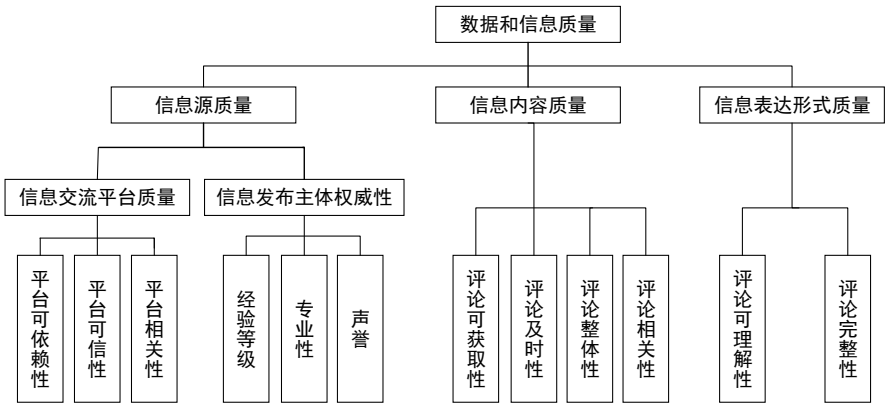

图 5-2　信息质量本体论模型

由于不同信息交流平台的产品评论可能来自消费者个人体验、专家推荐、明星代言、商业营销等途径，因此不同平台的信息可信度衡量指标不同。同时，信息的质量也受信息发布者的专业性和美誉度影响。由于本文研究问题的数据不涉及信息来源，故而不考虑这一指标。因此，根据旅游评论信息的特点，本文构建文本有效性评价指标体系，如下图 5-3 所示。

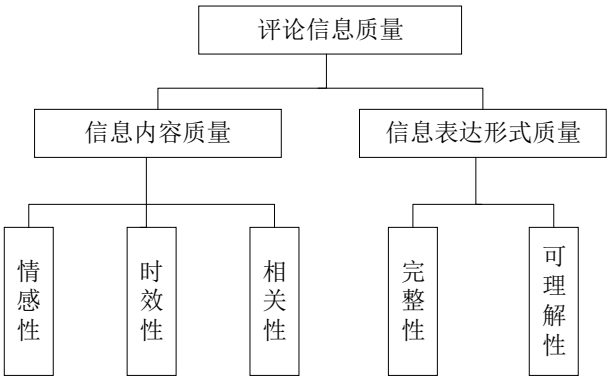

图 5-3　评论信息质量评价指标体系

5.1.1 信息内容质量与评论有效性

1.时效性

（1）指标含义

文本内容的时效性是指消费者一定时间内感受到内容价值的属性[12]，信息内容的时效性表明时间能够决定信息的效用，或者说信息内容的效用往往具有一定的时间期限[13]。Jin[14]等人发现，消费者的时间偏好、最近的评论对及时的购买决策更有吸引力，而长期评论对游客的长期购买决策影响更大。

（2）指标量化

发表日期和阅读时间之间的差异被用来量化评论的时效性。然而，由于评论发布时间的离散分布，直接将评论阅读时间和评论发布时间的差异作为评论时效性的量化值，可能造成较大的偏差和数据波动。本文将设置参数σ，评论及时性（X_1）的取值范围控制在[0-10]，计算公式为：X_1=[T_max-(T_w-T_r)]/σ。其中 T_max 是评论发布时间和评论阅读时间之间的最大差值，T_w 是评论发布时间，T_r 是评论阅读时间，σ取值为 T_max/10。特别注意的是，评论阅读时间定为 2021 年 5 月 1 日。

2.情感性

（1）指标含义

评论文本中含有的“非常好、满意、差评、失望”等情感词语，可以直接表达出游客对目的地的态度，也可以指导其他游客的旅游选择。在绝大多数目的地评论中，正面评论多于负面评论。当负面评论出现在首页或显眼位置时，更容易引起消费者的负面情绪反应，因为人们倾向于相信坏的而不是好的。

（2）指标量化

用情感词的频率很难表达评论中表述的积极或消极情绪的强度。本文基于 Hownet 情感词典，根据情感词典中对应词的相似度对采集到的文本进行评分，并考虑程度副词和否定词对情感值计算的影响，并加权计算出句子情感强度，输出评论文本的情感极性和强度。

3.相关性

（1）指标含义

评论内容的相关性考察评论者的评论内容与产品的主要特征之间的相似性，并通过评论的受众群体、评价对象、评论与产品词之间的相似性来评价信息质量，即评论中产品特征的

丰富性。在评论中引入一些能够客观地解释商品或服务的属性特征和价值的关键词。那些包含更多维度的关键词的评论与游客的信息需求更符合，对帮助游客了解目的地的服务和环境具有较高的参考价值。

（2）指标量化

评论相关性由评论中包含的目的属性特征词来量化。属性特征词通常是名词或形容词，是目的地本身及其相关服务的客观表达。经过对评论文本内容的细节提取，并进行分词、词性标注等处理，得到目的地评论词集合 C_1={c_1,c_2,⋯,c_s }。根据 TF-IDF 词频的统计，对名词和形容词进行要素量化，并人工归纳得到属性特征词集 C_2={c_1,c_2,⋯,ctf-idf}。构建空间向量模型，T_i 表示 i 个评论文本，匹配评论词集 C_1 和属性特征词集 C_2，统计评论中每个特征词的出现频率。例如，景区、酒店的特征词人工分为“服务”、“位置”、“设施”、“卫生”、“性价比”五大类。

5.1.2 信息表达形式质量与评论有效性

1.完整性

（1）指标含义

评论文本用来表达评论员对某种商品或服务的观点。评论文本中的关键词和字符越多，评论中所包含的商品描述信息就越多，这将丰富论据的信息来源，为潜在消费者提供全方位的参考信息，从而提高消费者的决策效率，也就是说评论表现出来的“说服性论证的力量”更强[15]。

（2）指标量化

网评文本的深度主要是指对目的地特征的详细描述。较长的评论通常包含更丰富的内容，并对目的地的服务、位置、性价比和其他信息有更详细的描述。本文统计了每条评论的字数，统计所得字符数分布较为不均，对字符数进行取对数，Length=Ln(Nr)，其中 Nr 为评论字数。

2.可理解性

（1）指标含义

一般而言，图文结合能给游客提供更高的参考价值，并产生更好表达效果。但由于在纯文本挖掘领域没有相应的图片评论信息，我们只能选择从文本自身的角度去解析表达形式的可理解性。

（2）指标量化

评论的可理解性，主要是指文本前后的关联性。本文利用语义文本向量来度量评论的可理解性。向量语义的概念是将一个单词表示为多维语义空间的一个点。表示单词的向量通常称为嵌入，因为单词嵌入到特定的向量空间中。

表 5-1　指标对应变量描述

文本有效性维度	信息质量指标	指标说明	变量对应	编码
信息内容质量	时效性	评论发表与评论阅读的时间间隔时长	时间间隔	X1
	情感性	评论表达出的情感倾向	情感倾向值	X2
	相关性	评论中描述目的地属性特征词数	关键词数	X3
信息表达形式质量	完整性	评论的长度	评论长度	X4
	可理解性	语义文本向量	余弦相似度	X5

将评论量化后的数据进行描述性统计分析，显示全部文本评论的数据集字段平均值没有显著差异，方差均在 10 以内，说明指标量化效果良好，数据波动性较小，对于模型分析具有较好的拟合效果。利用随机森林进行有效性分类时无须再进一步处理这些变量。

5.1.3　关联分析

(1)K-means 聚类

聚类是一种简单有效的数据挖掘技术。聚类是按照一定的标准将一些事物划分为若干类别的过程。相似的被聚为一类，不相似的被聚为不同的类。聚类算法种类繁多，比较常见的有：K-means 聚类、密度聚类、期望最大化聚类等。

文本所采用的是 K-means 聚类算法。K-means 聚类的模型构建思路如图 5-4 所示：

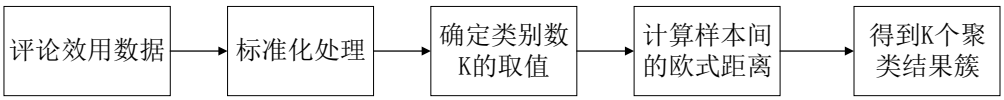

图 5-4　K-means 聚类的模型构建思路图

(2) Apriori 算法

Apriori 算法是一种基本的发现频繁项集的算法。Apriori 算法由连接和剪枝两个步骤组

成。连接是了找到L_k，通过$L_{(k-1)}$与自己连接产生候选 k 项集的集合C_k；剪枝是通过计算每个 k 项集的支持度来得到L_k，为减少计算量，可利用到该算法的性质即如果一个 k 项集的(k-1)项子集不在$L_{(k-1)}$中，则该候选也不是频繁的，可以直接从C_k中删除。其中支持度、置信度、提升度是用来衡量关联性强弱的三个核心指标。

关联规则的模型构建思路如图 5-5 所示：

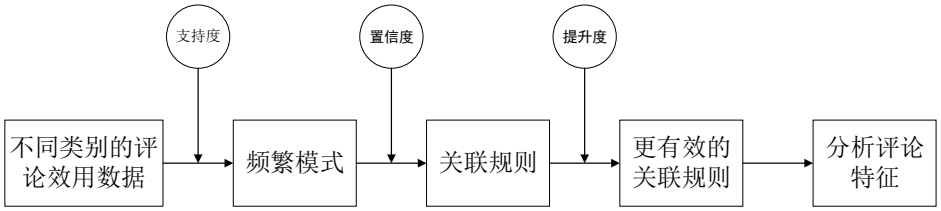

图 5-5　关联规则模型构建思路图

（3）关键技术路线

评论文本数据需要进行文本分析，研究技术路线也比较复杂，涉及分词、去停用词、词数统计等步骤。总体的核心技术步骤如下：

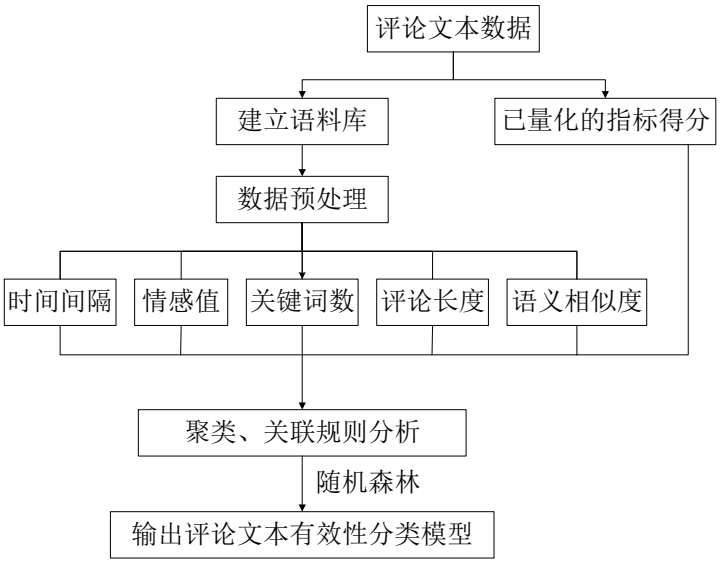

图 5-6　评论分类模型关键技术线路图

5.2随机森林分类模型应用

文本分类作为处理大量文本数据的关键技术，可以在较大程度上解决“信息爆炸”的问题。Breiman 提出的随机森林算法具有泛化性强、稳健性、对噪声不敏感、能处理连续属性等特点，非常适合于建立文本分类模型[16]。因此在文本有效性评估时，选用随机森林进行分类。下列模型以 A01 为例。

5.2.1 随机森林分类模型结果

本文利用得到的测试集对基于随机森林的文本分类模型效果进行评价。

（1）参数选择

随机森林模型有 3 个重要的可调参数：

表 5-2 随机森林模型可调参数

参数	含义
nodesize	包含样本的叶节点数
ntree	森林中树的数目
mtry	每个节点的候选特征数

一般而言，节点大小为 1 表示分类，5 表示回归。因此，本文取 nodesize=1。

研究表明，mtry 是影响随机森林模型性能最明显的参数，要求 mtry<<M，它表示每个分段中随机选择的候选变量数。在分类中，mtry 的建议值是整个变量个数的均方根；在回归中，mtry 的建议值是变量总数的 1/3。通过实证，发现 mtry 对模型的分类性能影响不大，因此本文取 mtry=7。

ntree 的设置相对简单，只要随机森林的总体误差率趋于稳定。Breiman 定义了随机森林的间隔函数，并根据大数定律证明了 RF 的泛化误差随着林中树数的增加趋于有限上界。因此只要 ntree 值足够大，就可以保证 RF 收敛。实证表明，当 ntree=59 时，分类效果最好，准确率达到 0.8235。

因此，随机森林的最优模型参数如下，达到了最好的分类效果：

表 5-3 随机森林最优模型参数

参数	含义
nodesize	1
ntree	59
mtry	7

（2）随机森林分类模型效果评价

对于有效性分类预测预实验的评价标准，本文采用统计学领域常见的精度（Accuracy）、查准率（Precision）、查全率（Recall）、F 值（F-measure）作为综合评价指标。以 A01 为例。

表 5-4　随机森林分类模型效果评价

准确率（Accuracy）	精确率（Precision）	召回率（Recall）	F1 值（F-measure）
0.8235	0.8077	0.9845	0.7802

结果显示准确率为 82.35%，即所有被预测的样本，预测正确的概率为 82.35%；准确率为 80.77%，说明分类器识别样本的能力较好；召回率为 98.45%，表示被预测的所有正样本，能够被正确预测的占比；F1 值是一个综合的评价指标。

5.2.2　文本分类结果与分析

利用随机森林分类器，本文将去除噪声后的 58633 条景区评论和 23711 条酒店数据集分为四类。景区高质量评论共 46441 条，占 77.5%，景区低质量评论共 13192 条，占 22.5%；酒店高质量评论 14179 条，占 59.8%，酒店低质量评论 9532 条，占 40.2%。因此，对于游客来说，至少有 60%左右的评论不需要阅读，而分类模型的目的是减少阅读低效评论的时间，只需关注高价值评论和潜力评论。

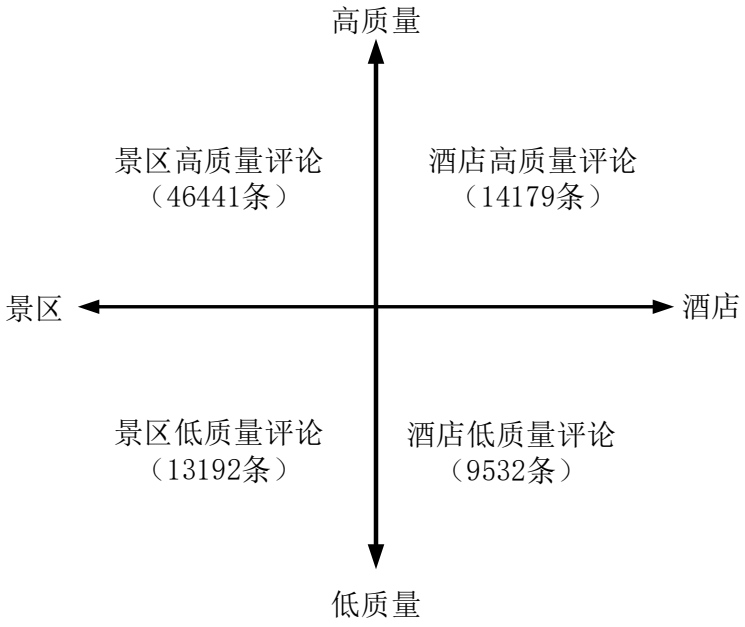

图 5-7　景区及酒店高质量评论和低质量评论分布

表 5-5 景区高质量评论示例

景区名称	评论时间	评论详情
A01	2020-06-23	水的项目:激流勇进好玩，海盗漂流一般，海战那个做好挨淋、挨喷、挨晒的准备，少人玩不是没理由。过山车类型很多，垂直过山车、火箭过山车挑战你的肾上腺。四维电影体验一下也不错。情侣票虽然挺实惠，但进到里面额外花的钱真不少:雨衣 15，雪糕两个 25，两根香肠加王老吉加水就 50 了，还有最重要的是，想要精彩瞬间的相片吗，至少 79。最后的建议:难得去一次的话，尽量选人气淡的时候，某些项目就不要早早去挤着排队了，如垂直过山车中午时要排 70 分钟，下午 6 点多就只用排 5 分钟，需合理安排时间。
A05	2021-03-25	非常愉快的一天，几个景点都很好玩，景色也很美，没有任何强制消费，导游照相技术不错哦，给我们拍了好多美美的照片，那是相当 happy 滴

表 5-6 酒店高质量评论示例

酒店名称	评论日期	评论详情
H09	2020-08-19	出差广州定酒店还是选择交通比较方便的，这家店地理位置不错，出门就是地铁站，对面就是广州东站，交通超方便。距离我办事的地方也很近。因为新冠肺炎入住酒店的人还是比较少的，来的时候保安主动打招呼测提问让我感觉很安心。酒店大堂很宽敞，在前台就有消毒酒精和预防知识，前台的酒店工作人员很细心的指引我到电梯，电梯要刷房卡才能按电梯，相比之下会比楼上的公寓安全得多。酒店干净卫生，客房服务也很热情，感觉超赞的。可惜只待了一天，听说附近有东方宝泰商场，逛街吃饭很方便，等疫情结束，下次再来体验。
H02	2020-12-23	前台服务很好，很热情，房间比较大住着舒服，卫生环境好

从以上景区及酒店高质量评价示例中，可以看到高质量评论基本上能包含“服务”、“卫生”、“环境”、“设施”、“性价比”等综合信息，涉及游客对目的地的真实评价，并且时效性较高、相关度较强。

表 5-7　景区低质量评论示例

景区名称	评论时间	评论详情
A09	2015-05-02	还不错，就是太累了！
A11	2015-04-22	第二次购买了，很不错

表 5-8　酒店低质量评论示例

酒店名称	评论时间	评论详情
H43	2020-08-27	挺好的挺好的
A11	2020-07-06	小如好评多多

从示例中可以看到，低质量评论文本会存在评论长度短、评论时间久远、评论内容不具有针对性等问题，参考性较弱。

5.3 高效评论排序模型

分类模型存在的问题是，虽然游客不需要阅读低效评论，在一定程度上降低了游客获取信息的时间成本，但高效评论的数量仍然众多。如何向游客优先展示最高效的评论是一个重要的问题。因此，本文提出基于高效评论的排序模型。关键技术路线如图 3-7 所示。重点是建立两种排序模型：基于 LDA 的排序模型和基于词向量相似度的排序模型。主要过程如下：

（1）数据预处理。对评论分词，与停用词表进行匹配，去除无用词，减少评论的噪音。

（2）构建高效评论-词频矩阵。采用向量空间模型构造文本向量，用 TF-IDF 函数变换。

（3）LDA 建模。主要使用的是建模后得到的评论——主题分布表，得到每个主题词在每条评论中的概率，为排序模型做准备。

（4）关键词向量计算。提取所有评论关键词，构造最优关键词向量，取值为 TF-IDF。

（5）基于 LDA 的排序模型构建。根据评论——主题表，计算评论主题信息的熵值，并基于该值对评论进行排序。

（6）基于词向量相似度的排序模型构建。计算每条评论的关键词向量与最优关键词向量之间的距离，并根据距离对评论进行排序。

（7）模型对比评估。提取两个排序模型的 top-k 条评论，比较 k 条评论在有效性评价的二级指标均值，选出最佳排序模型。

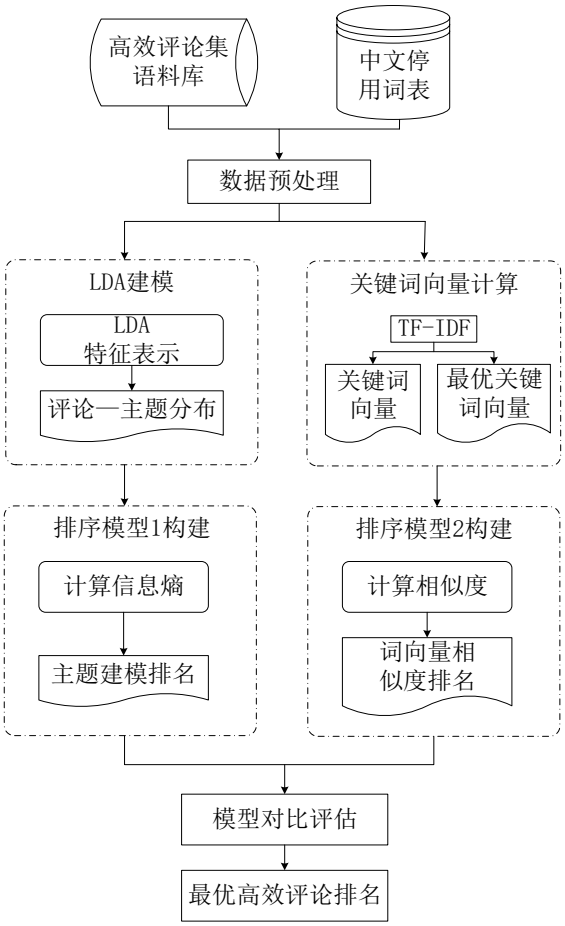

图 5-8　高效评论排序模型关键技术线路

5.3.1　指标选择

在高效评论排序模型的研究中，我们选择了评论广度作为排序指标，即评论信息的详尽程度。Shelat 和 Egger（2002）在研究中指出，评论广度对游客的购买决策有着重要的影响，是衡量评论效用的一个重要指标[17]。董丹凤（2013）也得出结论，评论广度对评论效用有着重要影响，可以增加游客信任感[18]。因此，本文选择评论广度作为高效评论排序模型的指标。评论广度越大，评论的效用就越高，这类评论应该被优先展示。

评论广度的度量方法有：通过文本主题发现模型来挖掘评论潜在的主题，主题能够显示评论的具体内容，是剔除冗余信息后最具价值信息的反映[19]；或提取评论关键词，形成一个词向量，通过关键词的丰富度来衡量评论所包含的信息量。

5.3.2 基于主题建模的排序模型结果

（1）信息熵

计算信息熵进行主题建模排名。信息量是一个很难描述的概念，直到 1948 年信息论之父 Shannon 提出信息熵的概念，信息度量问题才得以解决[20]。信息熵的概念是指热力学中热熵的概念，用来衡量信息的不确定性、稳定性和信息量。信息熵越大，信息越无序，信息熵越小，信息越有序。由于信息中必然存在冗余，因此冗余的大小与信息组成要素出现的概率有关。组成元素出现的单词越多，概率越大，信息的不确定性越小，信息量越大，反之亦然[21]。因此，信息熵是剔除冗余后的平均信息量，具体数学公式如下：

$$H(x) = -\sum_{i=1}^{m} p(x_i) \log p(x_i) \quad (5\text{-}1)$$

近年来，信息熵越来越多地应用于评论话题的研究。根据信息熵的定义，可以得出如下结论：评论主题的信息熵越大，评论主题越分散，评论内容的范围越大，评论广度越大；反之亦然。因此，评论广度信息熵的计算公式如下：

$$H(z_{ij}) = -\sum_{i=1}^{m} p(z_{ij}) \log p(z_{ij}) \quad (5\text{-}2)$$

其中z_{ij}表示某个主题在某条评论中出现的概率。

（2）LDA 建模结果与分析

下列以 H01A 为例，展示排序模型结果。

在使用 LDA 进行主题建模时,需要预先确定主题的个数，通常采用困惑度（$Perplex$ity）来评价模型生成主题的泛化能力，困惑度越低，模型的效果越好。本文选取了 10—30 个主题数，确定最佳主题数为 12。

根据信息熵计算公式(5-2)计算每条评论的信息熵。信息熵越大，评论包含的信息越高，排名越高。表 4-5 列举了信息熵排名前 5 的评论及其原始排名。由此可见，这些评论的原始排名位置相对较低，因此游客很难按照原始排名来搜索这些评论。从这 5 条评论的具体内容来看，每条评论基本上都包含了“服务”、“卫生”、“环境”、“设施”、“性价比”等综合信息。此外评论中有正面的评价，同时也有负面的评价，可以让游客对目的地有更全面客观的了解。

表 5-9　基于主题建模的评论排序结果

排名	原排序	信息熵	评论详情
1	478	3.882	非常推荐的老牌五星酒店，就在东站门口，对面就是 1 号地铁站和公交总站。对面就是东方宝泰，吃穿住行非常方便。永旺也方便购物，星巴克、肯德基、麦当劳也全都有。广东老牌陶陶居也在这里，出酒店转弯就是点都德，还有阿强家酸菜鱼。屋内房间非常大，服务特别好，卫生也非常到位！这个价位已经是相当相当的超值了。酒店自己的餐饮也值得推荐
2	692	3.879	帮同事一起订的，已经点评了。该酒店地理位置很好，靠近广州东站。老五星，所以设施确实陈旧了一点，一开始感觉有味道的，刚进房间嗓子干咳，适应一段时间好了。周边吃饭地方很多，非常方便。选择住在这里主要还是因为性价比高，促销 400 元一晚不到，很合算了，疫情期间大多数酒店打折，这个价位肯定会每一次都选择入住，不用多说了。
3	700	3.863	该酒店地理位置很好，靠近广州东站。老五星，所以设施确实陈旧了一点，一开始感觉有味道的，刚进房间嗓子干咳，适应一段时间好了。周边吃饭地方很多，非常方便。选择住在这里主要还是因为性价比高，促销 400 元一晚不到，很合算了，疫情期间大多数酒店打折，这个价位肯定会每一次都选择入住，不用多说了。
4	62	3.860	与火车东站一路之隔，去坐火车非常方便；连着东站的东方宝泰广场里有很多吃东西的地方，里面也有陶陶居。酒店稍微老了一些，床头没有插座，手机充电不方便。早餐很丰富，中西式、韩日的都有，还有现滚的各种粥，挺不错。服务态度也不错，我的戒指落在床头柜上，服务员发现后酒店立即联系我，十分感谢。
5	683	3.849	老牌五星级酒店，装修有点老，但胜在酒店位置方便，地铁火车东站 F 出口向左就到了。周边餐饮吃的也很多。前台帮升级到了行政豪华房型，600 多住这种性价比很高了。房间空调有点冷，出风口声音很大。早餐品类算比较丰富的。

5.3.3 基于词向量相似度的排序模型结果

（1）余弦相似度

常用的文本相似度计算方法有基于欧氏距离的相似度计算、皮尔逊相关系数计算和余弦相似度计算。由于余弦相似度计算方法能够更好地描述两个向量之间的角度差，故本文选择余弦相似度作为文本相似度的度量方法，并基于相似度进行评论排序。

在文本相似度计算中，一般将文本表示为由词组成的向量，具体值为上述 TF-IDF 权重。因此，两个文本的相似度由两个向量的相似度来表示，通常用余弦相似度来计算。夹角越小，越相似。具体公式见公式（5-3）。余弦值越接近 1，角度就越接近 0，也就是说，两个向量越相似。

$$\cos\theta = \frac{\sum_{i=1}^{n} X_i \times Y_j}{\sqrt{\sum_{i=1}^{n} {X_i}^2} \times \sqrt{\sum_{i=1}^{n} {Y_i}^2}} \tag{5-3}$$

（2）词向量相似度排序结果与分析

在实验中，首先将评论中的所有关键词聚集在一起，使用 TF-IDF 进行转换。此外，根据公式（3-10）计算每条评论的关键词向量与最优关键词向量之间的余弦相似度，并根据相似度从高到低进行排序。表 5-10 列出了前 5 条评论的具体内容和原始排名。同样，这些评论的原始排名不高，因此用户很难快速浏览这些评论。考虑到五条评论的内容，较为全面地描述了“服务”、“卫生”、“设施”、“位置”、“性价比”等信息，但单条评论所包含的信息量比基于主题建模的排名模型要少。通过对前 30 条评论的观察，我们也发现了这样的现象：虽然有些评论只是在“服务”上进行了详细的描述和评价，但相似度仍然很高，说明排名模型容易找到游客比较关注的关键话题，但相对会忽略话题的丰富性。

表 5-10 基于词向量相似度的评论排序结果

排名	原排序	相似度	评论内容
1	478	0.193	非常推荐的老牌五星酒店，就在东站门口，对面就是 1 号地铁站和公交总站。对面就是东方宝泰，吃穿住行非常方便。永旺也方便购物，星巴克、肯德基、麦当劳也全都有。广东老牌陶陶居也在这里，出酒店转弯就是点都德，还有阿强家酸菜鱼。屋内房间非常大，服务特别好，卫生也非常到位！这个价位已经是相当相当的超值了。酒店自己的餐饮也值得推荐
2	692	0.191	帮同事一起订的，已经点评了。该酒店地理位置很好，靠近广州东站。

			老五星，所以设施确实陈旧了一点，一开始感觉有味道的，刚进房间嗓子干咳，适应一段时间好了。周边吃饭地方很多，非常方便。选择住在这里主要还是因为性价比高，促销 400 元一晚不到，很合算了，疫情期间大多数酒店打折，这个价位肯定会每一次都选择入住，不用多说了。
3	533	0.187	酒店位置不错，出行方便，就在东站门口，对面就是 1 号地铁和公交总站。周边有东方宝泰，有超市，比较方便。性价比较高。
4	683	0.183	老牌五星级酒店，装修有点老，但胜在酒店位置方便，地铁火车东站 F 出口向左就到了。周边餐饮吃的也很多。前台帮升级到了行政豪华房型，600 多住这种性价比很高了。房间空调有点冷，出风口声音很大。早餐品类算比较丰富的。
5	788	0.175	老牌酒店，设施略旧，卫生比较干净，毗邻广州东站，交通方便，楼下很多便利店，不到 400 在广州住五星级酒店性价比很高。

5.3.4 排序模型总结

本章的最终目标是挖掘最有效的评论，而用户有耐心浏览的往往是前几条评论。因此，与评论的最终排名相比，衡量哪种排名算法对排名前 k 位的评论更有效是最重要的问题，而且 k 条评论的排名顺序存在一定的偶然性，不能作为评价的标准。田韶存（2014）在对书评排名的研究中提出，按照排名算法得到的前 6 条评论已经可以覆盖大部分评论信息[22]。因此，本文选取两个排序模型得出的前 10 名、前 20 名和前 30 名评论进行评价，计算 5 个指标的平均值，选择效果较好的模型作为最优排序模型。

排序模型的结论是：基于主题建模的排序算法能够得到广泛的评论和众多的主题，但它局限于主题模型的缺点，而且一些模型学习到的主题很难理解。基于词向量相似度的排序算法只能得到相对较少的主题，但可以挖掘出重要的关键词，供用户通过关联获取主题。由于基于主题建模的排序模型在前 10、前 20 和前 30 条评论指数得分上具有明显优势，因此该模型被选为最佳排序模型（见表 5-11）。本文提出的排序模型可以在分类模型的基础上进一步简化评论，去除冗余，保证第一次显示的评论包含的信息量最大，减少信息搜索的时间和成本，大大提升用户体验。

表 5-11　排序模型对比及总结

排序模型	优点	缺点	占优势的评估指标数量（去除得分相同）		
			前 10	前 20	前 30
基于主题建模的排序模型	评论内容覆盖面广，包含的主题多	部分主题难以理解	3	3	3
基于词向量相似度的排序模型	通过重要关键词联想得到主题，便于理解	评论内容主题较少、多个关键词可能表达同一个意思	2	1	1

6　基于 Text CNN 的 LDA 主题特色挖掘

6.1 基于 Text CNN 的主题挖掘模型构建

6.1.1　数据集描述

田韶存[22]在对图书评论的排序研究中提出，根据排序算法得到的 TOP6 条评论已经可以覆盖大部分评论的信息。为了从评论文本中挖掘出景区和酒店的特色和亮点，以吸引游客提升竞争优势。本题基于前两题的情感分类和有效性评价，得到“积极且高效的评论文本集”，其中景区数据集共 42627 条；酒店数据集共 18193 条。

6.1.2　Text CNN 评论文本情感分类

目前常用的分类方法包括朴素贝叶斯方法(NB)、支持向量机(SVM)、随机森林、Text CNN 卷积神经网络等。

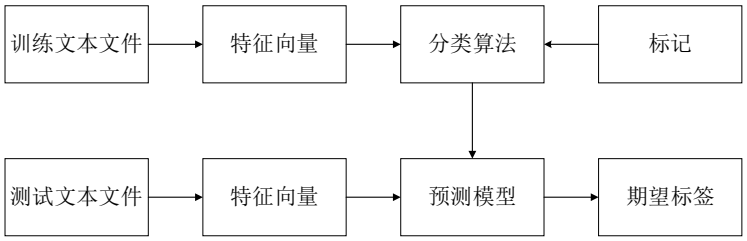

图 6-1　分类算法流程图

为进一步验证本章模型适合哪种分类器，故对常用的四种分类器进行验证测试。分类处理流程如图 6-1 所示。本文首先将训练集和测试集按照酒店顾客满意度二级影响因素进行人工标记分类，以“.txt”的文本形式分别存到对应的类别特征文件夹下进行操作。利用 TF-IDF 发现特征词，分别构建训练集和测试集的词袋模型来反映文档主题的特征，然后利用三种分类算法对生成的模型进行训练，如表 6-1 文本分类结果所示，结果保留 4 位小数。

表 6-1 文本分类结果

分类器	朴素贝叶斯 NB	支持向量机 SVM	TextCNN 卷积神经网络
准确率	0.8684	0.8947	0.9153
精度	0.8919	0.8947	0.9072
召回率	0.9706	0.9845	0.9946
F1 得分	0.9296	0.9444	0.9517

通过表 6-1 可以看出，三种分类算法采用 tf-idf 模式对实验结果的参数进行验证。综合准确率、精度、召回率及 F1 得分这四个评价指标，TextCNN 卷积神经网络算法对本研究实验而言效果更好。因此，本章采用 TextCNN 卷积神经网络分类器对本研究的文本数据集进行分类，将各情感单元分类到各二级影响因素之下，为主题挖掘打好基础。

6.1.3 LDA 模型构建及评价

（1）LDA 参数设定

由于以上分类结果无法直观地发现景区和酒店的特点，为进一步发掘其各自的优势和特色，本文对各景区和景点的积极且高效评论文本集进行主题挖掘。

主题挖掘思路为：一是对使用卷积神经网络分类好的文本进行分词；二是使用 Counter Vectorizer（python 中 scikit-learn 矢量化工具）对文档集合进行向量化；三是在 scikit-learn 工具箱中调用 Latent Dirichlet Allocation 函数，在参数调整和可视化结合下，选定主题数为 4；四是经过最多 40 次迭代，初步识别出主题。

下

表 6-2 显示了 Counter Vectorizer 和 Latent Dirichlet-Allocation 的参数设置。

表 6-2 Counter Vectorizer 和 Latent Dirichlet-Allocation 的模型参数

	可调参数	参数值	参数解释
Counter Vectorizer	encodeing	utf-8	分析器默认解码方式为 utf-8
	max_df	0.8	构造语料库关键词集时，频率大于 80%的词不是关键词
	min_df	0.1	构造语料库关键词集时，频率小于 10%的词不是关键词
	max_feature	5000	对所有关键词的出现频率降序排序，只取前 5000 个作为关键词集
Latent Dirichlet Allocation	n_topics	4	模型选取 4 个主题数
	Learning_method	batch	模型选取的算法为 batch,即变分推断 EM 算法
	Max_iter	40	模型中 EM 算法的最大迭代次数为 40 次

（2）LDA 主题挖掘结果示例

LDA 是无监督学习的一种，所以主题的名称标签不能包含在结果中，仅使用一系列高频关键词来描述该主题。

表 6-3 H01 酒店积极高效文本各主题关键词

主题	关键词
主题 1	前台、态度、热情、舒适、耐心、效率、礼貌、专业、齐全、感受
主题 2	东站、机场、出行、地铁、飞机、火车站、公交车、方便、快捷、停车场
主题 3	风景、周边、环境、便利、地理、附近、商圈、购物、距离、吃饭
主题 4	干净、舒适、浴缸、整洁、舒服、整齐、洗漱、消毒、崭新、霉味

以 H01 酒店为例，利用其积极且高效的评论文本数据集进行 LDA 主题挖掘，每个主题输出 10 个关键词，得到如表 6-3 的结果。

主题 1 的内容主要表达该酒店服务方面的描述，结果显示该酒店服务态度热情周到；主题 2 表达的主要内容大致和交通相关，主要表达了该酒店可选择交通方式多，且靠近交通运输枢纽，出行十分便利；主题 3 主要表达该酒店价格地理位置十分优越，周围遍布商圈，出门游玩吃饭都很方便，并且环境很好；主题 4 主要表达该酒店的卫生状况良好，但仍有些许

改进空间。我们发现 LDA 主题模型得到了较好的主题划分结果，但主题 2 和 3 之间的界限较为模糊。前文也提到了，LDA 的主题分类结果受参数调整影响较大，而 LDA 模型的超参数调整又是一个比较难解决的问题，因此本文在下文引入 LDAvis 工具，进行 LDA 可视化，帮助验证 LDA 的分类结果，并对各个主题的核心内容进行更好的区分。

6.1.4 基于 LDA vis 的主题模型可视化

本文利用 py.LDAvis 软件包进行交互式可视化检验。LDA 分析的结果用作输入变量传输到 pyLDAvis，同时支持更加直观的参数调整，检查上一步 LDA 主题挖掘的结果，并使得各主题的区分度更显著。

图 6-2 显示，LDA 主题挖掘结果与主题分类相一致。pyLDAvis 工具通过编号来区分不同的主题，从而在二维向量空间中直观地表示了不同对象之间的距离。每个数字所在的圆圈的大小说明主题中的语料库中包含多少个词。评论的四个主题在二维向量空间中有显著差异，彼此之间没有重叠，分布相对分散，主题之间分类较好。在这些主题中，主题 1 的圆圈最大，主题 1 的内容最受关注，在评论文本中出现的次数最多。反之，主题 4 包含的内容出现次数最少，关注也较少。

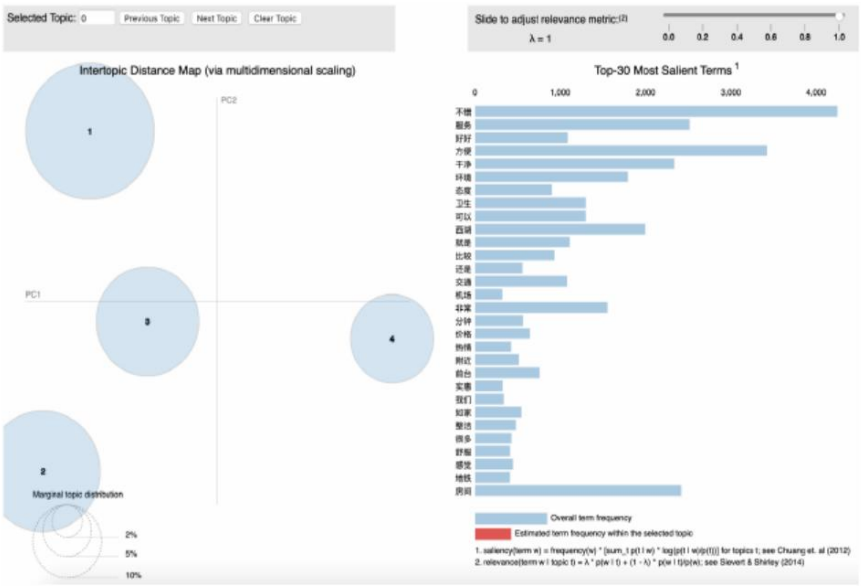

图 6-2 LDA 结果可视化

右上角有一个用于调整参数λ的拖曳条。当λ=1 时，显示结果与 LDA 分类结果的高频词相同，关键词排名仅取决于词频；当λ=0 时，具有主题之间区分度的关键词位置排序会显著提升。选择λ没有规则，通过调整参数λ，获取最具区分度的主题关键词，寻找潜在的主题规

律。

以 H01 酒店为例，如图 6-3 所示，当λ调整为 0.4 左右时，词语排序出现显著变化。根据各个主题的关键词列表，可以推断出各个主题对应的主要内容。

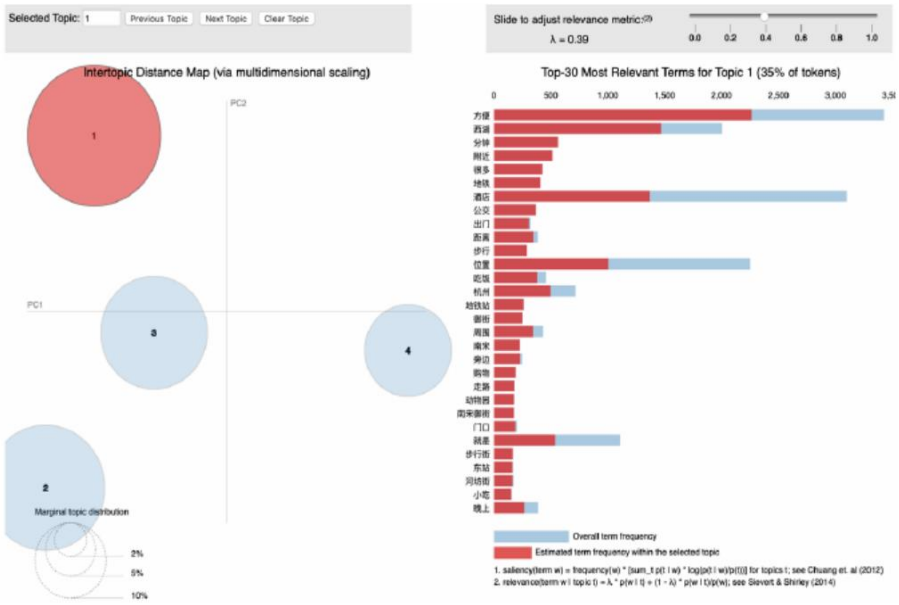

图 6-3　H01 酒店积极且高效评论主题 1 关键词展示

同理，依次分析主题 2、主题 3、主题 4。该酒店给客户的整体良好印象主要集中在以下四点：1.酒店客服人员服务热情周到，设施齐全，入住体验好。2.酒店靠近交通枢纽，出行方便。3.地理位置优越，酒店周边商圈遍布。4.酒店卫生状况良好。

6.2 主题挖掘结果分析

本章融合前述的分析结果，提取景区及酒店的积极且高效的评论文本集。以该景区或酒店的积极且高效的评论数占其总评论数的比例为指标，分别在评分为高、中、低三个层次的景区及酒店中选取前三名，共 18 家，进行主题挖掘。同时，有针对性的进行 LDA 主题挖掘建模和可视化分析，归纳出景区和酒店最具特色的部分，为游客选择提供有效参考，并为营业者提升竞争力。结果如下所示：

图 6-4 酒店特色挖掘词云图

图 6-4 从左到右，从上到下，分别为 H01、H39、H43、H09、H27、H41、H07、H32、H40，融合 LDA 主题挖掘结果和可视化图形得到如下表 6-4 的酒店特色。

表 6-4 酒店特色分析

层次	酒店	特色			
低层次	H01	商圈遍布	靠近交通枢纽	卫生不错	服务周到
	H39	位置繁华	出行便利	房间温馨	性价比高
	H43	繁华江景	靠近景区	怀旧情怀	视野开阔
中层次	H09	海边漫步	亲子出行	私密性强	设施丰富
	H27	欧洲情调	小镇风情	度假体验	入住便捷
	H41	温泉民宿	乡土风情	空气清新	休闲娱乐
高层次	H07	江畔美居	文艺小资	文化底蕴	美食环绕
	H32	靠近机场	免费接送	专车服务	入住方便
	H40	小资浪漫	美味早茶	体验性强	温泉别墅

酒店评论中，消费者会表达对于酒店服务的多种不同服务要素的意见，对于同一种服务要素，又会关注不同的服务特征，不同的消费者对服务细节有不同的要求和看法，细化服务的程度也不相同。通过对酒店评论进行主题挖掘并结合可视化与词云图分析酒店特色，所选择的九家酒店竞争优势如上表所示。

酒店服务有别于一般的电商实体产品，它是有地域性的，是一种典型的异地消费服务型产品，周边一般会因为交通、商圈或、景点或特色等吸引要素有密集流动的人群。酒店以这些吸引要素为基点向周边辐射，面向某一或某几种特定的人群提供特定的服务。确定目标受众，并针对性的展开运营策略，对酒店长远发展极具优势。

图 6-5　景区特色挖掘词云图

图 6-5 从左到右，从上到下，分别为 A05、A31、A43、A01、A20、A26、A14、A23、A50。融合 LDA 主题挖掘结果和可视化图形得到如下表 6-5 的景区特色。

表 6-5 景区特色分析

层次	景区	特色			
低层次	A05	世界之窗	奢华夜景	精彩表演	特色建筑
	A31	民族特色	瑶族服装	古寨风情	歌舞演出
	A43	海岛沙滩	休闲度假	渔家海鲜	红树林
中层次	A01	动物园	游乐园	亲子出游	精彩马戏
	A20	空气清新	风景优美	索道缆车	攀登顶峰
	A26	特色建筑	文化之旅	世界遗产	影视基地
高层次	A14	地下长河	自然景观	溶洞风采	游船赏景
	A23	苏州园林	江南烟雨	古色古香	岭南风光
	A50	避暑胜地	瀑布群	森林公园	自驾出行

旅游景点的不同带给游客的情感体验亦有差异，海岛风情、游乐园、山清水秀的优势各不相同。不同类型的景区所面向的群体不同，竞争优势也不相同。因此以可视化手段将结果进行呈现，并总结归纳各景区的竞争优势，使景区经营者能够清晰地针对游客的需求优化弱势，宣传特色，以增加客流量。

7 结论与建议

7.1 结论

本文的核心是通过构建合理的指标选择和严谨的模型设计，以景区及酒店的评论文本集作为数据源，利用自然语言处理技术，为景区及酒店战略管理与发展提供辅助。得出如下结论：

（1）将数据预处理后，首先根据构建的热度计算指标体系，得到各家景区及酒店的热度值；其次将提取的关键词作为二级影响因素，并对比 K-means、AP、GMM、HAC 四种聚类算法，选用结果最好的凝聚层次聚类法并结合内容分析法得到一级影响因素；再次对比朴素贝叶斯、支持向量机、Text CNN 卷积神经网络情感分类结果，得出情感得分，由此构建基于情感分析的游客满意度评价模型；另构建基于模糊综合评价方法的游客满意度评价模型；最后，以专家打分为标准得分，将以上两个模型与专家打分进行对比，利用均方误差

（MSE）进行模型评估。

（2）首先根据信息质量理论构建文本有效性评价指标体系，依据信息内容质量的三个指标和信息表达形式的两个指标，利用随机森林进行有效性分类，并对分类模型进行效果评价；其次构建基于主题挖掘的排序模型和基于词向量相似度的排序模型，选择最优排序模型；最后根据分类和排序结果进行分析。

（3）基于评论短文本情感倾向性分析结果和有效性分析结果，得到一个积极且有效的评论文本集，并进行 LDA 主题模型训练与构建。通过景区及酒店个性化特色挖掘，最后根据主题模型挖掘结果及其可视化展示分析结果。

7.2 建议

7.2.1 提高酒店服务水平

游客评论提及最多的是酒店服务水平。酒店的服务水平在很大程度上会影响游客满意度。贴心细致的服务能增加游客的满意度，这其中又包括了酒店前台接待服务、酒店清洁人员的服务水平、酒店餐厅的用餐服务等。因此，酒店要提升前台人员的接待水平，定时对员工进行统一的培训，做到服务标准化、规范化，做到服务热情、贴心、及时、细致，使得游客能够在服务中体会到宾至如归的温暖。提升酒店清洁员工的服务水平，要增加其责任感，同时应当建立一套规范化的卫生水平评判标准。对于表现优秀的员工，应当给予适当的奖励，以此激起员工工作的积极性；对于工作质量欠佳的员工也要适当的批评，责令其尽快改进。

文旅部门应定期对当地酒店进行检查以及动态管理，参考网络上的评论数据，建立健全酒店服务考评体系，重视服务质量和生产管理，主动开展自查自纠和隐患排查、线索摸排工作，制定质量提升行动方案。

7.2.2 优化酒店硬件设施

酒店的硬件设施，包括装修风格、卫生间、停车场等，是影响游客入住体验的重要因素。对于酒店来说，硬件设施的资金投入很大程度上决定了酒店房间的层次，设备设施作为固定资产对酒店的运营成本产生了重要的影响。在保证性价比的前提下，酒店管理者应重视不断完善酒店的硬件设施。例如，增强酒店房间的隔音效果。

此外，注重提升硬件设施水平的同时绝不能轻视安全性，文旅相关部门应当定期对酒店进行安全生产检查，可以通过实地查看、翻阅资料、现场询问座谈等方式，针对问题逐项排查。重点就食品安全管理情况、消防设施、安全通道、电气设备、电梯等进行详细检查。做到“谁经营，谁负责”，把安全生产各项工作落到实处，杜绝安全事故的发生。

7.2.3 提升景区旅游层次

旅游既是大产业，又是大民生，我们已经进入了大众旅游时代。旅游已经成为人民幸福生活的必需品。要提升旅游景区的层次，不断发掘景区的特色。从评论数据的分析结果来看，可以知道游客最在意的是一个景区是否“好玩”。而“好玩”是一个比较宽泛的词语，对于类似游乐园这样的景区来说，游客体验到的是感官上的刺激，因此游乐园要“好玩”，就要做到各种游乐项目、表演、建筑风格等方面的推陈出新。而对于各种历史文化景区来说，“好玩”就要体现在景区浓厚的人文气息，为此要重视对景区包含的历史文化价值的介绍，在寓教于乐的同时使游客感受到全新的地方特色文化。比如西安著名的景点大唐不夜城，以盛唐文化为背景，以唐风元素为主线，运用盛唐文明的元素与现代展示手法相结合的方式突出在现代商业中的地位，创造了巨大的经济效益的同时，又宣传了盛唐文化，体现了实用性与科学性、艺术性相结合。各地政府应重视与文旅相关部门等充分合作，结合地方具体情况，为发掘本土的特色旅游共同努力。

参考文献

[1] 涂铭,刘祥,刘树春.Python 自然语言处理实战核心技术与算法[M].北京:机械工业出版社,2018.

[2] 刘顺祥.从零开始学 Python 数据分析与挖掘[M].北京:清华大学出版社,2018.

[3] BreimanL,FriedmanJ, Olshen R,alet.ClassificationandRegressionTrees[M].NeYork:Chapman&Hall,1984.

[4] BreimanL.Bagging Preditors [J]. Machine Learning,1996,24(2):123-140.

[5] 黄厚波.基于 AHP-模糊综合评价法的长沙市公交乘客满意度评价研究[D].中南林业科技大学,2014.]

[6] 张丽娜.模糊综合评价法在生态工业园区评价中的应用大连：大连理工大学，2006:66-67

[7] 程海琪.基于情感分类的酒店评论短文本主题挖掘[D].浙江工商大学,2020.

[8] 谢宗彦,黎嶸,周纯洁.基于 word2vec 的酒店评论情感分类研究[J].北京联合大学学报,2018,32(04):34-39.

[9] 李柏年.模糊数学及其应用[M].合肥工业大学出版社,2007.

[10]WangRY,StrongDMBeyondAccuracy:WhatDataQualityMeanstoDataConsumers[J].JournalofManagementInformationSystems,1996,12(4):5-33.

[11] RogovaGL,BosseE. Information quality in information fusion[C]//InformationFusion.IEEE,2010.

[12] 张宁,袁勤俭.游客视角下的学术社交网络信息质量影响因素研究——基于扎根理论方法[J].图书情报工作,2018(5):105-113.

[13] 江彦,娄策群,江秀,毕达宇.评论者对在线商品评论信息质量的影响及提升策略研究[J].图书馆学研究,2019(03):78-83+94.

[14] JinL, HuB,HeY. The Recent versus The Out-Dated: An Experimental Examination of the Time Variant Effects of Online Consumer Reviews[J]. Journal of Retailing,2014,90(4):552-566.

[15] 石文华,高羽,胡英雨.基于情感倾向和观察学习的在线评论有用性影响因素研究[J].北京邮电大学学报(社会科学版),2015,17(05):32-39.

[16] 罗新.基于随机森林的文本分类模型研究[J].农业图书情报学刊,2016,28(11):50-54.

[17] ShelatB, EggerFN. What makes people trust online gambling sites? [A]. CHI'02 Extended Abstracts on Human Factors in Computing Systems[C]. New York:ACM,2002:852-853.

[18] 董丹凤.点评社区评论有用性影响因素研究[D].华东师范大学,2013.

[19]阮光册.基于 LDA 的网络评论主题发现研究[J].情报杂志,2014,(03):161-164.

[20] Shann once. A mathematical theory of communication[J]BellSystemTech,1948,27(3):379-423.

[21] 王莉亚,张志强.基于信息熵的信息整合主题演化研究[J].图书情报工作,2012,(06):102-106.

[22] 田韶存.在线社区游客评论有用性研究[D].山东大学,2014.