第九届“泰迪杯”数据挖掘挑战赛

基于 LDA 主题模型和 LightGBM 分类模型的在线旅游评论挖掘及分析

基于 LDA 主题模型和 LightGBM 分类模型的在线旅游评论挖掘及分析

摘要

近年来，随着网络技术的快速发展和旅游业信息的高度密集，在线旅游（Online Travel Agent，简称 OAT）已经成为用户获取信息、表达观点、相互交流的重要途径，大量以评论和游记等为形式的非结构化数据不断涌现，借助自然语言处理技术来提取旅游者对目的地的真实印象，能够掌握影响游客满意度的重要因素，有针对性地提高游客满意度、提升目的地美誉度，不仅能够保证客源稳定，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。

针对任务一，由于原始数据集含有大量噪声，本文首先对原始数据进行预处理，包括文本去重、分词、词性标注和去停用词。由于与景点和酒店相关的印象词语具有一定的特殊性，保留句子中词性为名词、名动词和形容词的词汇。对比了基于词频、TF-IDF 和 textrank 三种算法进行文本关键词提取的结果，最终，保留 TF-IDF 算法提取的关键词和对应权重，分别制作 50 个景区和 50 家酒店的印象词云表。

针对任务二，首先，对评论数据进行文本预处理，保留分词结果中词性标注为名词的词语，基于 LDA 模型进行主题识别，依据主题聚类结果和相关文献，分别构建不同评价维度下景区及酒店评论的主题词词典。其次，使用哈工大开源 LTP 分词库，通过将标点符号替换为换行符，实现对评论的分句。再者，基于主题词典，实现分句筛选和分类，并对分句进行情感分析，进行情感得分规范化。采用以好评数比例扩大好评影响力的方法，修正情感得分规范化时，最终得分偏低的情况。最后，基于用户对各评价维度的关注度，计算每个景区或酒店总得分时的权重。最终结果中，对景区在服务、位置、设施、卫生和性价比上预测评分的均方误差分别为 0.24，0.37，0.21，0.10 和 0.10；对酒店在服务、位置、设施、卫生和性价比上预测评分的均方误差分别为 0.11，0.24，0.10，0.03 和 0.24。

针对任务三，查阅相关文献，本文分别从内容相关性、内容有用性和内容简单重复三个方面建立了评论有效性评价体系，8 个二级指标分别为是否主题词数、是否包含情感词、是否包含广告词、是否包含违禁词、包含文本字符数、语义丰富度、情感是否极端、内容是否重复。随后，分别抽取去重后的景区评论 5862 条，去重后的酒店评论 2014 条，进行人工标注，并对标注结果进行一致性检验，两标注者对景区和酒店的标注结果 Kappa 值分别为 0.627、0.912。最后，对由人工和机器共同标注的数据集，共 9060 条带标签的评论数据，划分训练集和测试集，构建 LightGBM 模型，进行模型训练和参数调优，最终，得到模型在测试集上的分类准确率为 94%，对有效评论的识别精度为 94%，F1 值 84%，对无效评论的识别精度为 93%，F1 值为 96%。

针对任务四，基于任务二中的综合评价结果，分别筛选了高中低三个层次中各三个景区和酒店。首先，基于主题词典，识别用户对不同层级中景区和酒店关注点的分布差异。其次，基于评论的情感分类结果，分别对正负向评论进行主题识别。最终结果中，对景区而言，服务、设施和性价比是正向评论中的热点，门票价格是用户负面评价中普遍关注的问题；对酒店而言，服务、位置和设施是用户正面评价中关注的热点，设施和卫生为用户负面评价中关注的热点。最后，基于不同层级中景区和酒店的特征词频，绘制词云图，分别对各景区和酒店进行特色分析。

关键词：TF-IDF 模型；LDA 模型；LightGBM；主题识别；有效性评价

目　录

图 录

1 引言

1.1 挖掘背景及意义

旅游形象是旅游者对某一旅游目的地的总体感知和全部印象的总和，是目的地竞争优势的核心来源，同时也是评价旅游地发展的一项重要指标[1]。提升景区及酒店等旅游目的地美誉度是各地文旅主管部门和旅游相关企业非常重视和关注的工作，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。近年来，随着网络技术的快速发展和旅游业信息的高度密集，在线旅游（Online Travel Agent，简称 OAT）已经成为用户获取信息、表达观点、相互交流的重要途径[2]。在线旅游评论是一个公共资源，不仅是当前顾客的发表看法和提出意见的场地，还能够为潜在顾客的决策提供参考[3]。

然而，旅游者通过在线评论平台畅所欲言，大量以评论和游记等为形式的非结构化数据不断涌现，除了包含旅行者真实意见的文本语料，不乏无用的、重复的垃圾评论。传统的调研方式已无法满足如今动辄上万的数据挖掘需求，通过借助非结构化文本挖掘技术和自动化技术来提取旅游者对目的地的真实印象，能够掌握影响游客满意度的重要因素，有针对性地提高游客满意度、提升目的地美誉度，不仅能够保证客源稳定，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。

基于此，本文采用文本挖掘技术，自动识别旅游景点及酒店评论文本中关键词，构建景点及酒店的综合评价模型，使用有监督学习的方法分析评论文本的有效性，最终，基于评论文本主题识别的结果，分析不同景点及酒店的特色。

1.2 问题描述

问题一分析：由于网络评论文本数据量庞大，需要借助机器自动提取文本中的关键词，并计算词频。观察实验语料可知，存在部分内容完全重复的评论，可能会影响统计结果，因此，需要依据内容对评论进行去重处理。并且，评论语料中常包含大量标点符号和无意义的词汇，因此，在进行词频统计时，需要对语料进行预处理。由于与景点和酒店相关的印象词语具有一定的特殊性，词性通常为名词、名动词和形容词，如“长老峰”、“爬山”、“秀丽”等，考虑基于词性标注的结果进一步筛选关键词，绘制印象词云表。

问题二分析：对景区及酒店的综合评价涉及服务、位置、设施、卫生和性价比五个维度，与新闻文本分类问题不同，评论文本篇幅通常较短并且可能会对景区或酒店多个方面的进行评价，因此，需要构建合适的评价体系。对于任务二，由于一条评论中，评论者对目的地不同方面的评价所蕴含的情感可能存在差异，因此，需要采用基于属性的情感分析，进一步对属性所属维度的评分进行计算。并且，由于评论者对于不同维度的关注度存在差异，可以采用基于用户关注度的权重确立，计算景点或酒店的综合得分。再将最终结果与实际数据进行比较，计算均方误差。

问题三分析：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。对评论文本进行有效性分析，自动识别无效评论，对提升评论质量，从而帮助平台识别用户真正需求，改善服务质量有重要意义。对于任务三，由于评论文本并没有事先给定的标签，为保证机器识别的严谨性，采用人工标注的方式，对随机抽取的评论语料进行有效性标注。再基于有效性划分的方法，使用机器对评论进行特征提取。对有标签数据划分训练集和测试集，使用机器学习或深度学习的模型，分析模型分类的准确度。

问题四分析：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。基于任务二中综合评价的结果，分别对属于不同层次酒店和景点评论语料进行主题识别，提取目的地关键词，分析酒店及景区的特色。

2 景区及酒店印象分析

2.1 数据描述

文中有两个评论数据集，分别是 59106 条针对景区和 25525 条针对酒店的评论。针对景区的评论涉及 50 个景点，统计各景点的评论数量，由于景点数较多，仅展示评论量前 15 位（评论数大于 1000 条）景点评论量分布，如图 1 所示。分析图 1 可知，所有景点中，评论量排名前 15 位的景点累积评论量近总评论量的 70%，据此，可推断这 15 个景点为旅游热门景点，其中，评论量最多的为 A01 景点，评论人数超过 4500 条。

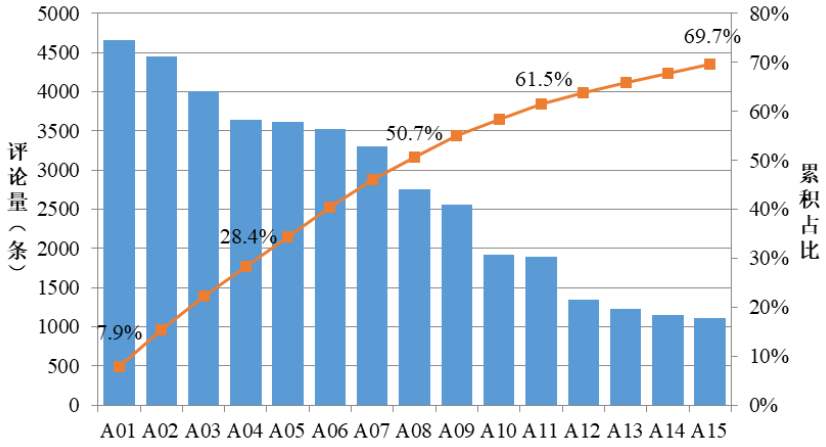

图 1：TOP15 景点评论量分布

针对酒店的评论涉及 50 家酒店，统计各酒店的评论数量，仅展示评论量前 15 位（评论数大于 550 条）酒店的评论量分布，如图 2 所示。分析图 2 可知，与景点评论量相比，酒店评论量较少，大于 1000 条评论的仅 H01 一家酒店。评论量排名前 15 位的酒店累积评论量不到总评论量的 50%，评论量总体分布较为均匀。

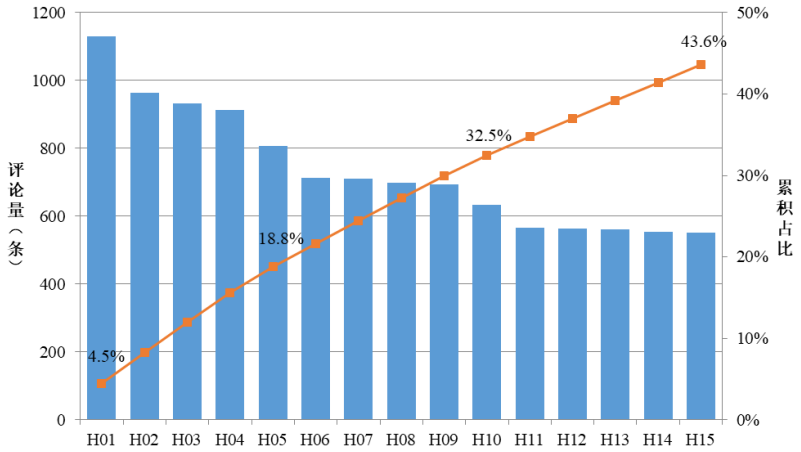

图 2：TOP15 酒店评论量分布

2.2 数据预处理

2.2.1 文本去重

2.2.1.1 去除原因

原始评论文本中，存在内容完全一致的评论，可能是发布者通过简单复制粘贴产生，本文假定最早发布的评论具有参考价值。而对于内容相似评论，由于围绕某景点和酒店描述的关键词通常具有相似性，因此保留内容不完全一致，但描述相似的评论。表 1 中给出了内容完全一致和内容相似评论的示例。

表 1：重复评论及相似评论示例

类型	评论实体	评论时间	评论内容
完全重复	A26	2020-07-02	除了贵除了晒，一切都好！请个讲解小姐姐更好，对历史有更深入的了解。环境很不错，房子都保护得很好！值得一去！
		2020-07-02	除了贵除了晒，一切都好！请个讲解小姐姐更好，对历史有更深入的了解。环境很不错，房子都保护得很好！值得一去！
	H01	2020-06-17	酒店很好不错的
		2020-06-23	酒店很好不错的
相似评论	A31	2020-06-24	**，房屋保持原生态，表演好看
		2020-08-02	**，房屋保持原生态，寨民热情待客 篝火晚会热闹，值得一去
	H01	2020-01-05	酒店服务挺不错的。环境也还可以。
		2020-06-23	很好的酒店服务不错

2.2.1.2 处理方法

对评论进行按时间排序，使用 python 程序判断评论是否完全重复，保留发布时间最早评论。经过上述处理，分别得到每个景区及酒店评论中重复文本的数量占比，如图 3、图 4 所示。分析图 3 可知，每个景区完全重复的评论占比相对较低，比例均低于 3%。分析图 4 可知，除酒店 H16 及 H48，每个酒店完全重复的评论占比均低于 8%。

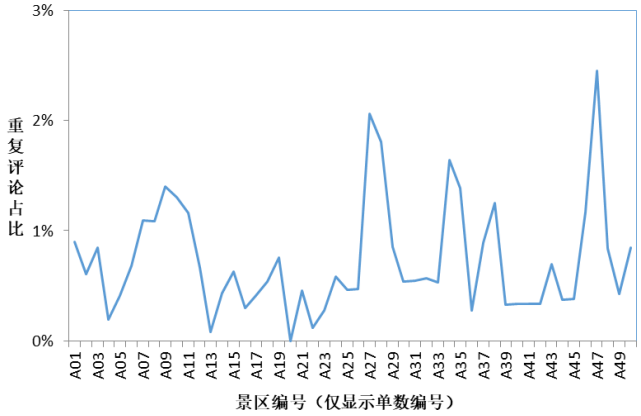

图 3：景区重复评论分布

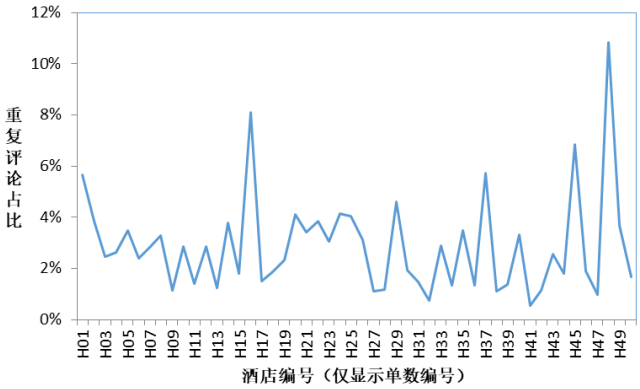

图 4：酒店重复评论分布

2.2.2 文本分词

2.2.2.1 分词工具

（1）Jieba 分词

Jieba 分词是一款中文开源分词包，具有高性能、高准确率、可扩展性等特点，包含精确模式、全模式、搜索引擎等分词模式。其分词原理是使用基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG）；采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。并且，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。此外，还支持自定义词典及停用词词典，提供词性标注的功能。

（2）NLPIR 分词系统

NLPIR 是中科院计算所研制的基于多层隐马尔科夫模型的中文词法分析系统。该系统提供了中文分词、词频统计、词性标注、命名实体识别、新词识别等功能。

（3）pyltp

语言技术平台(LTP)由哈工大社会计算与信息检索研究中心研发和推广，pyltp 是 LTP 的 Python 封装。它提供的功能包括中文分词、词性标注、命名实体识别、依存句法分析、语义角色标注等。

本章节采用 Jieba 分词工具进行评论文本分词，由于 Jieba 分词支持在原有词库的基础上添加未登录词词典，并且，根据分词结果不断修正分词词库。

2.2.2.2 词性标注

（1）标注原因

由于与景点和酒店相关的印象词语具有一定的特殊性，词性通常为名词、动名词和形容词，如“长老峰”、“爬山”、“秀丽”等，为保证后续印象词识别的准确性，本文基于词性标注的结果进一步筛选关键词。

（2）词性标记含义

本章节采用 jieba 对分词的词性进行标注，部分常见词性标记如下：

表 2：Jieba 常见词性标记释义

词性	符号	释义
形容词	a	取形容词 adjective 的第 1 个字母
名形词	an	具有名词功能的形容词。
副形词	ad	直接作状语的形容词。
副词	d	取副词 adverb 的第 2 个字母。
代词	r	代词 pronoun 的第 2 个字母。
介词	p	介词 prepositional 的第 1 个字母。
名词	n	取名词 noun 的第 1 个字母。
动词	v	取 verb 的第一个字母。
副动词	vd	直接作状语的动词。
名动词	vn	指具有名词功能的动词。

（3）处理方法

导入 Jieba 分词中词性标注函数 jieba.posseg，保留分词词性为名词、动词、名动词、形容词的词语。针对一条景区评论的分词示例如下图 3 所示：

```
import jieba.posseg as pseg

sentence =pseg.cut("A03非常值得一去，可以在里面玩一天，不去世界各地也能观赏到埃菲尔铁塔，凯撒宫，等等的辉煌")
for w in sentence:
    print (w.word+'/'+w.flag,end="")
```

```
A03/eng非常/d值得/v一/m去/v，/x可以/c在/p里面/f玩/v一天/m，/x不/d去/v世界各地/l也/d能/v观赏/v到/v埃菲尔铁塔/nz，/x凯撒/nrt宫/nr，/x等等/u的/uj辉煌/a
```

图 5：景区评论文本分词示例

2.2.3 过滤停用词

由于评论文本中包含大量无用标点符号及停用词，本文采用哈工大停用词表，并结合分词结果，不断修正停用词表。

2.3 评论热点识别

2.3.1 文本关键词提取算法

识别景区及酒店评论热点的主要任务是提取评论文本中关键词，并输出关键词所占重要性权重。本文采用了三种方法进行关键词提取，以 A01 景区和 H01 酒店评论为例，对比三种提取结果，分析关键词提取方法的适用性，基于关键词提取结果，对 A01 景区及 H01 酒店的进行印象分析。

2.3.1.1 基于词频的关键词提取

词频（term frequency，简称 TF）是指词或短语在给定文档中出现的频率，通常认为词频越高，其在文档中的重要度越高，成为关键词的可能性就越大[4]。

本章节对文本预处理后的分词结果，进行词频统计，分别保留各景区及酒店词频最高的前 20 个关键词，作为目的地热门词。

2.3.1.2 基于 TF-IDF 的关键词提取

词频-逆文档频率（TF-IDF）[5]结合词频和逆文档频率来衡量候选关键词的重要度，TF-IDF 的主要思想为：如果某个词在一篇文档中出现的频率越高，即 TF 越高；并且在语料库中其他文档中很少出现，即文档频率（DF）越低，逆文档频率（IDF）越高，则认为该词具有较好的区分能力。本章节采用 jieba.analyse 库中 extract_tags()方法，对各景区及酒店评论文本预处理后的结果，重要性权重排名前 20 位的关键词作为目的地热门词。

2.3.1.3 基于 textrank 的关键词提取

textrank 方法[6]的核心思想是将文本中的词语当作图中的节点，通过边相互连接，不同的节点会有不同的权重，权重高的节点可以作为关键字。本章节采用 jieba.analyse 库中的 textrank()方法，选择重要性权重排名前 20 位的关键词节点作为目的地热门词。

2.3.2 结果对比及印象分析

分析三种方法的关键词提取结果，基于词频的提取方法中，虽然较多的无用词，如“看”、“取票”等词被识别出来，而 TF-IDF 和 textrank 的结果相近，并且提取效果较好。下图 6、8 分别展示基于词频的景区和酒店 TOP15 关键词词频分布，图 7、图 9 分别展示基于 TF-IDF 景区和酒店 TOP15 关键词权重分布。分析图 6、图 7 可知，A01 景区有马戏、动物园、过山车等项目，适合小朋友游玩，并且价格实惠。分析图 8、图 9 可知，H01 酒店交通便利，离车站较近，客房服务和设施配备都给旅游者留下了不错的印象。

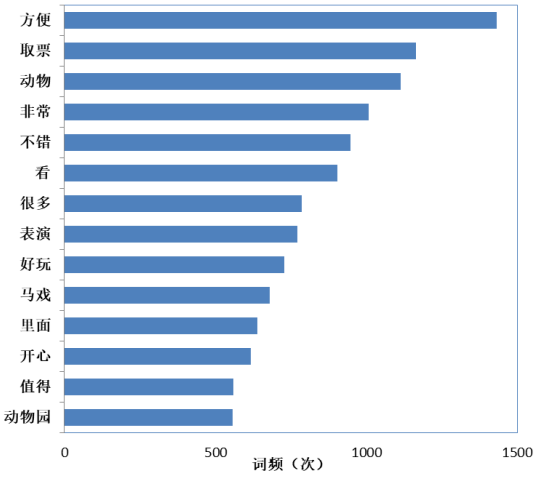

图 6：基于词频的 A01 景区 TOP15 关键词

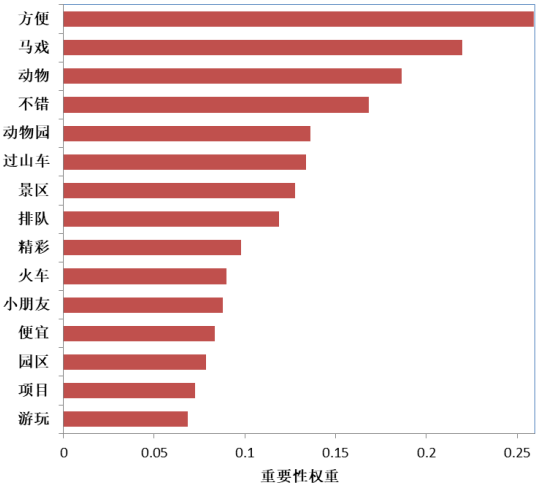

图 7：基于 TF-IDF 的 A01 景区 TOP15 关键词

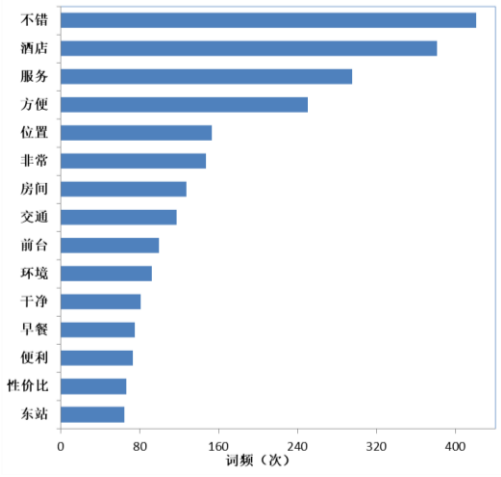

图 8：基于词频的 H01 酒店 TOP15 关键词

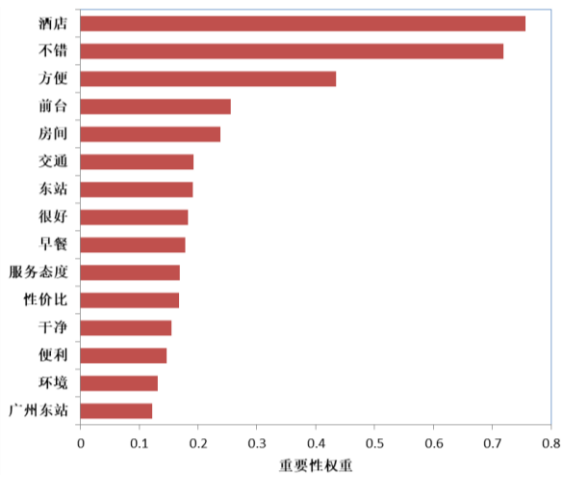

图 9：基于 TF-IDF 的 H01 酒店 TOP15 关键词

3 景区及酒店综合评价

由于针对景区或酒店的评论主题性较强，主要论点通常通过名词的形式进行表达[3]，并且涉及评价对象的多个方面，图 10 和图 11 分别展示了景区及酒店评论语料中名词性短语的词云图。可见，对景区和酒店进行评价时，首先，需要识别观点句中包含的实体属性，再通过基于属性的情感分析，分别计算景区及酒店在服务、位置、设施、卫生、性价比五个方面的情感综合得分，以用户对各评价维度的关注度作为权重，计算各景区及酒店的总得分，以实现对目的地的综合评价。

图 10：景区评论词云图

图 11：酒店评论词云图

3.1 基于 LDA 模型的主题词典构建

3.1.1 LDA 主题提取算法

LDA（Latent dirichlet allocation）[7]是有 Blei 于 2003 年提出的三层贝叶斯主题模型，通过无监督的学习方法发现文本中隐含的主题信息，目的是要以无指导学习的方法从文本中发现隐含的语义维度。它是一种无监督的文档主题生成模型，认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程，这些主题被集合中的所有文档所共享，每个文档有一个特定的主题比例。对应结构如下图所示：M 表示评价数，N 表示不同的主题数，文档词频 $W_{t,n}$ 是一个已知的统计量，它依赖于对这个话题的指派 $Z_{t,n}$ 以及话题所对应的词频β_k；同时，话题指派 $Z_{t,n}$ 依赖于话题分布θ，θ 依赖于 Dirichlet 分布参数α，话题的词频则依赖于参数η，大矩形表示从狄利克雷分布中为每个文档 d 中反复抽取主体分布θ_d，小矩形表示从主体分布中迭代产生文档 d 的词{w_1，$w_{1,}$ $w_{1.}$…，w_n}。

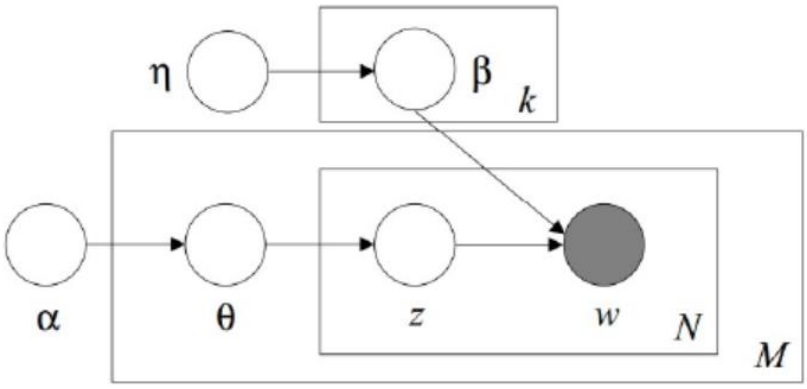

图 12：LDA 主题提取流程图

当给定一个评论集合 D，包含 M 条评价和 N 个不同的主题词，每条评论 d 包含一个词

序列{w_1,w_2,w_3⋯,w_N}，在评论集合 D 对应的 LDA 模型中，假设主题数目固定为 k，则一个文档 d 的产生可以表示为以下两个步骤：

（1）从 Dirichlet 分布 p(θ|α)中随机选择一个 k 维的向量θ_d，表示文档 d 中的主题混合比例。

（2）根据主题比例对文档 d 中的每个词均进行反复抽样，得到 p(w_n|θ_d,β)，其中参数α是一个 k 维的 Dirichlet 的一个参数，如公式 3.1 所示。

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdot\cdot\,\theta_k^{\alpha_k-1} \qquad \text{（式 3.1）}$$

其中，Γ(.)是 Gamma 函数，α是模型参数，β是一个 K×N 的矩阵，具体的写法为：$\beta_{ij} = p(w_j = 1|z_i = 1), i = 1,2,\cdots K; j = 1,2,\cdots N$。

3.1.2 流程设计

基于 LDA 主题提取的原理，本模块利用词性标注技术，找出评论中用户意见对应的名词短语，然后通过将名词进行聚类，结合聚类结果和领域相关文献，构建景区和酒店主题词典。具体流程设计如下图 13。

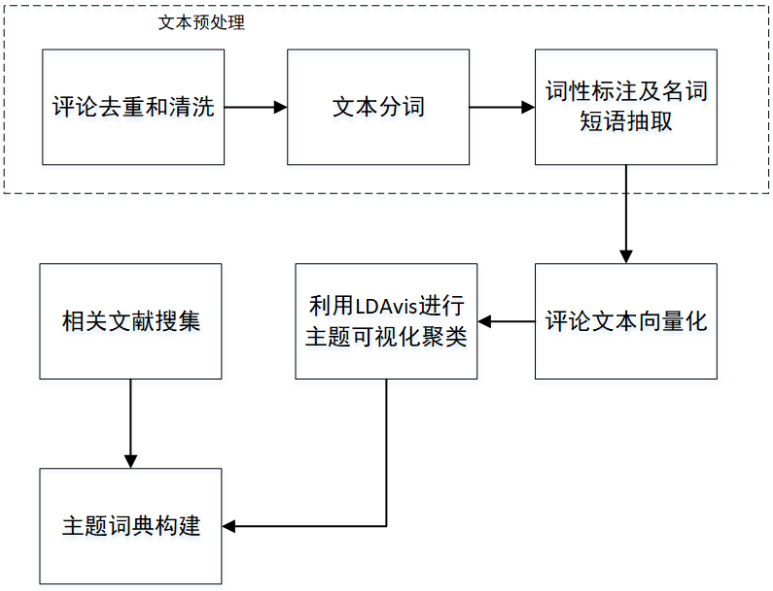

图 13：基于 LDA 模型的主题词典构建流程

3.1.3 主题建模及词典构建

基于第一章中热点识别的结果可知，基于词频的关键词提取会提取出较多无实际意义或不具备代表性的词语，而使用逆文档频率可以弥补词频排序带来的缺点，使更多合理的词排在前面，有助于挖掘景区或酒店的隐含属性。因此，本模块使用 pyltp 对评论文本进行分词和词性标注，使用 TF-IDF 的算法对去停用词并保留名词性短语的评论进行文本向量化处理。利用 LDA 主题模型对主题词进行聚类可视化，根据可视化结果，不断调整聚类主题。

本模块使用 pyLDAvis 对评论进行可视化的主题聚类，左侧的圆圈代表了不同的主题，圆圈之间的距离是每个主题之间的相似度。在选定某一个主题后，右侧面板会相应地显示出

该主题下相关性最高的 30 个词汇，通过总结这些词汇表达的意义，归纳出该主题的意思。参考五维度的评价标准，利用实验通过以 K=5 为基准，依次升高 K 值的方法选择各主题下的属性词。

对于景区评论文本，多次调整 K 值后发现，当主题数 K=5 时各主题交叉较少，分布均匀，效果最好。如图 12 所示，选中主题 2，该主题内部包含的主题词有“摩天轮”、“大峡谷”、“广州”、“东西”、“美人鱼”、“深圳”、“温泉水”等主题词，归纳该主题主要与景区设施和景区地理位置相关。

对于酒店评论文本，多次调整 K 值后发现，当主题数 K=6 时各主题交叉较少，分布均匀，效果最好。如图 13 所示，选中主题 4，该主题内部包含的主题词有“广州”、“深圳”、“东西”、“美”、“珠江”、“交通”、“性价比”等主题词，归纳该主题主要与酒店所处地理位置和性价比相关。

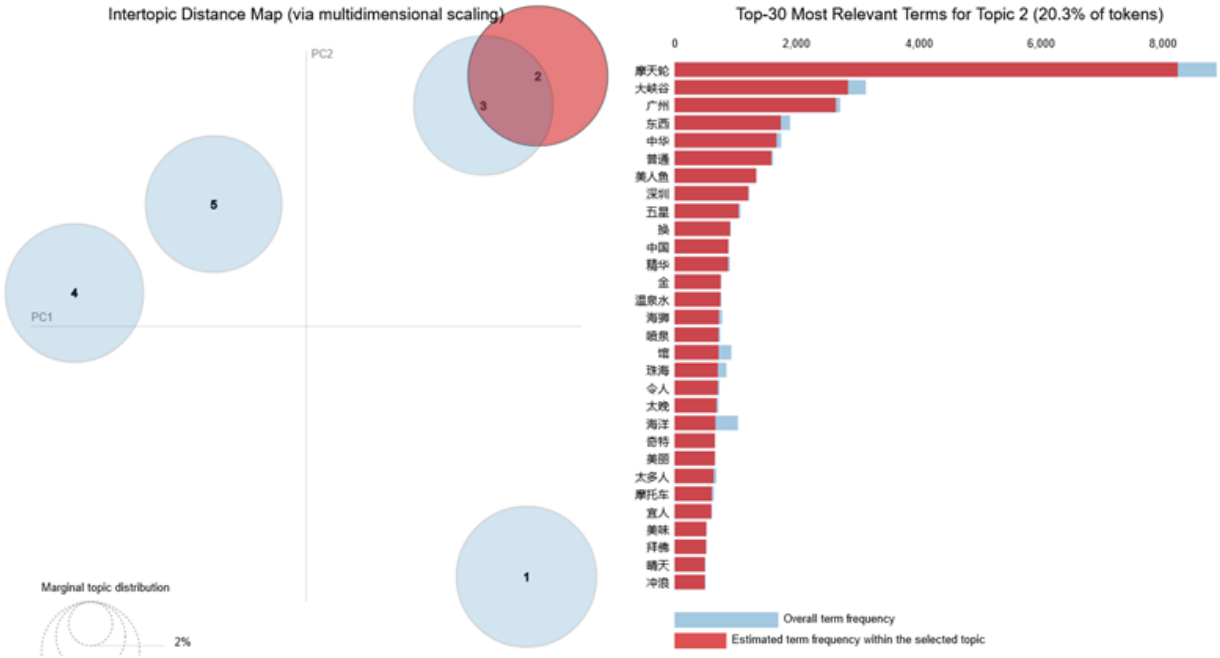

图 14：景区评论 pyLDAvis 主题可视化（Topic2）

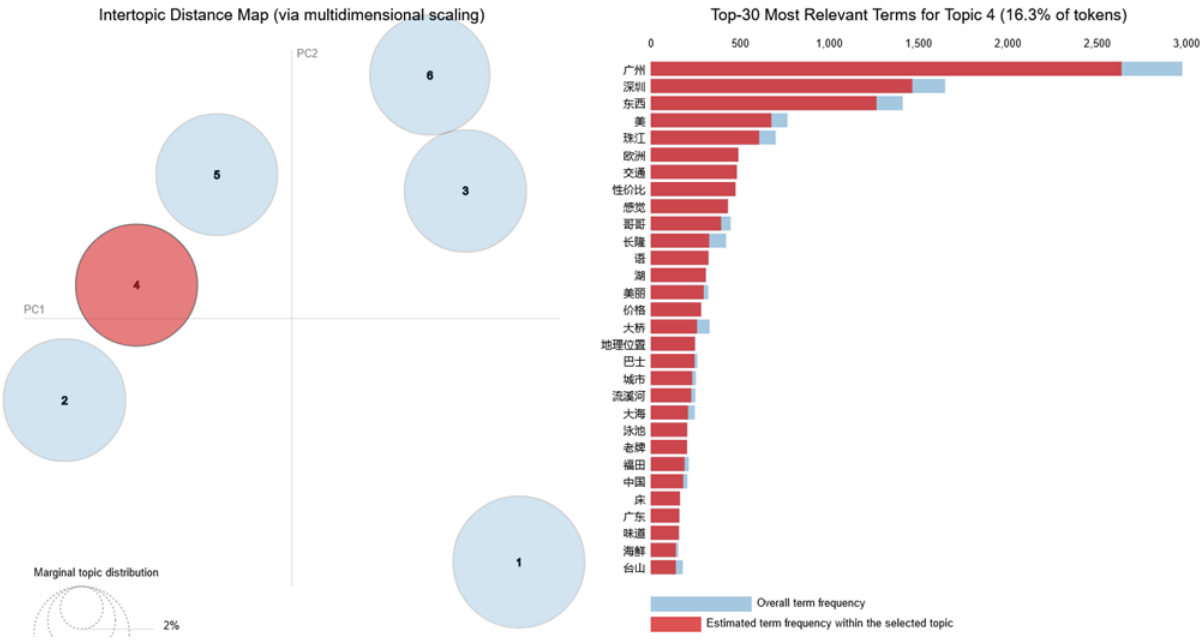

图 15：酒店 pyLDAvis 主题可视化（Topic 4）

借助陈天琪等[2]对景区评价和缪章伟等[8]对酒店评价进行主题归纳的结果，以及使用 pyLDAvis 进行可视化聚类，对景区及酒店评价的相关主题词进行了扩充和丰富，分别归纳景区和酒店的评价维度和具体分类下的主题属性词。景区和酒店的主题词典示例分别如表 3、表 4 所示。

表 3：景区主题词典示例

评价维度	具体分类	主题词
服务	餐饮	服务（服務）、早餐、美味、美食、菜品、口味、好吃、味道、品种、种类、自助餐、自助、牛腩、餐厅、餐馆、菜品、口感、菜品
	服务人员	服务态度、售票员、态度、前台、保安、热情、小姐、和善、经理、服务员、热情、服务态度、贴心、亲切、主动、敬业、体贴、笑容
	服务细节	售票、订票、拿票、取票、管理、速度、接待、行李、效率、细节、周到、到位、方便、服务水平、刷卡、素质、文明
位置	地理位置	地理位置、广州东站、北京路、花城、地标
	周边环境	机场、车站、出口、入口、附近、靠近、周围、购物、周边、地段、世界之窗、地铁口、湖边、商场、边上、广场、旁边、公园、大桥
	交通	交通、巴士、拥堵、地铁、自驾、车、打车、步行、出门、方便、公交车、公交、出行、出租车、车程
设施	游玩设施	设施、设备、景点、观光车、游览车、升级、摩天轮、大峡谷、庄园、馆、摩托车、冲浪、索道、安全、规范、危险、农家乐、游船
	景区配置	大门、动物、猛兽、海狮、孔雀、老虎、狮子、熊猫、鲤鱼、白鲸、喷泉、瀑布、海洋馆、温泉、温泉水、沙滩、微缩、阳元石、阴元石
	公共设施	取票机、休息室、卫生间、洗手间、厕所、马桶、指示牌、寄存处、路标、大厅、大堂、环境、停车、停车场、豪华、颜值、富丽堂皇、
卫生	卫生条件	空气、空气质量、气味、清新、新鲜、卫生（衛生）、干净、防疫、疫情、整洁、灰尘、霉味、乱、脏、差、臭、舒适、舒服

性价比	价格	性价比、免费、优惠、消费、价格、经济、自费、价位、便宜、门票、票价、值得、划算、贵、涨价、不值、购物、价格不菲、公道、金额、
	风景品质	风景、品质、景色、景观、天堂、谷、峰、湖、海、美、云霄、钟灵毓秀、山清水秀、亮点、特色、景色宜人、雪山、奇特、美丽、夜景

表 4：酒店主题词典示例

评价维度	具体分类	主题词
服务	餐饮	服务（服務）、早餐、美味、正宗、普通、口味、好吃、味道、品种、种类、丰富、自助餐、自助、酒廊、牛腩、餐厅、菜品、口感
	人员	服务态度、态度、前台、迎宾、保安、热情、经理、服务员、礼貌、温馨、热情、贴心、亲切、主动、敬业、体贴、笑容、员工
	服务细节	速度、及时、接待、行李、效率、细节、周到、到位、方便、服务水平、刷卡、素质
位置	地理位置	地理位置、广州、广州东站、广东、北京路、深圳、成都、珠海、珠江、长隆、流溪河、台山、中国、深圳市、东莞、花城、小蛮、地标
	周边环境	机场、出口、附近、靠近、周围、购物、风景、边上、广场、旁边、公园、大桥、沙滩、大海、景观（景觀）、酒吧、地段、商圈、夜景
	交通	交通、巴士、地铁站、打车、步行、出门、方便、公交车、公交、出行、出租车
设施	客房设施	床、桌子、灯、枕头、电视、拖鞋、空调、电脑、上网、陈旧、温馨、升级、房间设施、卫生间、浴缸、浴室、一次性、用品、视野、视野
	隔音效果	隔音、声音、效果、安静
	酒店设施	酒店、电梯、设施、硬件、楼层、风格、装修、商务、装潢、繁华、度假、大厅、环境、游泳池、温泉、停车场、滑梯
卫生	卫生条件	房间（房間）、房、卫生（衛生）、干净、防疫、疫情、整洁、灰尘、霉味、乱、脏、差
性价比	价格	性价比、免费、优惠、消费、价格、经济、价位、便宜、值得、划算、贵、涨价、不值、价格不菲、公道、金额、额外收费、价钱
	品质	老牌、品质、星级、五星、四星、三星、二星、一星、特色、档次、舒适、舒服

3.2 基于属性的细粒度情感分析

由于依据评论中可能出现针对多主题的观点，本模块通过利用 Python 编程对接哈工大开源 LTP 分词库，对句子进行分词处理，并通过将词性标注为标点符号（主要包括逗号、问号、句号、分号）的元素替换为换行符，解决一条评论句中出现多主题的问题。如对于评论“酒店服务挺不错的。环境也还可以。”即可拆分为“酒店服务挺不错的”和“环境也还可以”两条评论。利用扩充后的主题词典，通过属性词匹配的方式，找出包含主题词的评论分句，统计各评价维度对应的评论分句数量。假定一条分句中，仅包含评论者对评论实体某一主题的评价，将分句情感得分的结果转换为对评价主体属性的得分。

3.2.1 流程设计

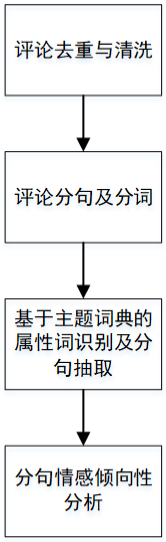

图 16：基于评论分句的情感分析流程

3.2.2 中文情感分析工具

3.2.2.1 SnowNLP

SnowNLP 是一个 python 写的类库，可以方便地处理中文文本内容，是受到了 TextBlob 的启发而写的，针对中文文本进行情感分析，和 TextBlob 不同的是，SnowNLP 没有用 NLTK，所有的算法都是由创作者自己实现。SnowNLP 支持的中文自然语言操作包括中文分词、词性标注、情感分析、文本分类等。

3.2.2.2 Senta

Senta 为百度开发的一款一键式情感分析预测工具，使用了百度研究团队提出的基于情感知识增强的情感预训练算法（SKEP），该算法采用无监督方法自动挖掘情感知识，然后利用情感知识构建预训练目标，从而让机器学会理解情感语义。在三个典型情感分析任务，即句子级情感分类（Sentence-level Sentiment Classification），评价对象级情感分类（Aspect-level Sentiment Classification）、观点抽取（Opinion Role Labeling），共计 14 个中英文数据上进一步验证了情感预训练模型 SKEP 的效果。实验表明，以通用预训练模型 ERNIE（内部版本）作为初始化，SKEP 相比 ERNIE 平均提升约 1.2%，并且较原 SOTA 平均提升约 2%。

本模块对比了 SnowNLP 和 Senta 对于一些典型观点句的情感打分效果，对消极评论情感得分计算示例如图 15 所示。分析图 15 可知，使用两种工具对两条评论“这里一点都不好玩”和“这里卫生真差劲”进行打分，Senta 将两个句子均归为消极评论（sentiment_label 为 0）；SnowNLP 的评分的大小代表句子积极情感倾向的强弱，通过对结果得分的判断可知，SnowNLP 将第一条分句划分为积极评论（得分大于 0.5）。因此，为保证情感倾向计算的准确性，本模块选择使用 Senta 作为计算分句情感得分的工具。

```
input_dict = {"text": ['这里一点都不好玩','这里卫生真差劲']}
result = senta.sentiment_classify(data=input_dict)
print('Senta情感分析结果：'+str(result))
```

```
Senta情感分析结果：[{'text': '这里一点都不好玩', 'sentiment_label': 0, 'sentiment_key': 'negative', 'positive_probs': 0.103, 'negative_probs': 0.897}, {'tex
t': '这里卫生真差劲', 'sentiment_label': 0, 'sentiment_key': 'negative', 'positive_probs': 0.0127, 'negative_probs': 0.9873}]
```

```
from snownlp import SnowNLP
text1='这里一点都不好玩'
text2='这里卫生真差劲'
s1 = SnowNLP(text1)
s2=SnowNLP(text2)
print('Snownlp情感分析结果：'+'\n'
      +'text1情感得分：'+str(s1.sentiments)+'\n'+
      'text2情感得分：'+str(s2.sentiments))#
```

```
Snownlp情感分析结果：
text1情感得分：0.6315125408371073
text2情感得分：0.043397022562916776
```

图 17：Senta 和 SnowNLP 情感得分结果对比

3.3 综合评价模型

3.3.1 流程设计

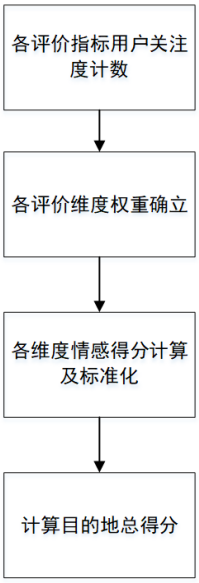

图 18：基于情感得分的综合评价流程

3.3.2 模型构建

（1）情感得分规范化

本模块以分句的积极情感倾向性得分作为最终得分的计算依据，由于倾向性得分输出结果范围在 0~1 之间，使用如下公式（式 3.2）将评分规范化到 1~5 之间。

$$\text{score}_i = 1 + 4 * pos\_probs_i \quad \text{（式 3.2）}$$

（式 3.2）中，pos_probs_i表示第 i 条评论的积极情感倾向性得分，score_i表示规范化后的得分。

本模块尝试了两种计算最终得分的方法，第一种，直接计算各景区或酒店每个维度下分句情感得分的均值，即为该景区或酒店该维度下的最终得分。

第二种，以好评数比例扩大好评影响力，修正第一种计算方式中，进行情感得分规范化时，最终得分偏低的情况。具体计算方式如下公式（式 3.3）。

$$\text{dimention\_score}_d = posp_d * \sum_{i=1}^{M} \frac{score_i}{M} + (1 - posp_d) * \sum_{i=1}^{N} \frac{score_i}{N} \quad \text{（式 3.3）}$$

（式 3.3）中，dimention_score_d表示景区或酒店第 d 个维度下的最终得分，$posp_d$表示

第 d 个维度下的分句语料中好评数所占比例，M 表示好评数量，N 表示非好评的数量（包括差评和中评）。

（2）基于用户关注度的权重确立

在景区评论文本（如图 19）中，顾客意见中对服务、设施、位置、性价比和卫生的关注度依次减弱，顾客的意见主要集中在的景区的服务上，总占比达到了 42%，其中针对卫生和性价比的评论数较少。在酒店评论文本（如图 20）中，顾客意见中对性价比、设施、服务、位置和卫生的关注度依次减弱，顾客的意见主要集中在的酒店的性价比上，总占比达到了 33%，其中针对卫生的评论数较少。

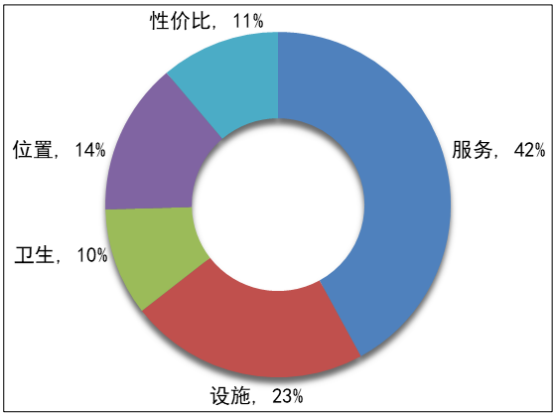

图 19：景区用户关注度分布

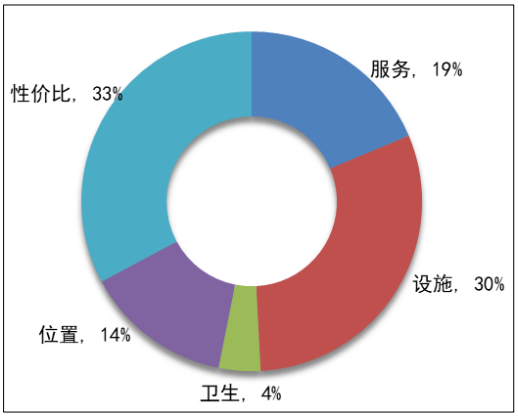

图 20：酒店用户关注度分布

用户关注度一定程度上反映了用户更加注重该方面的需求，因此，本模块将用户对各维度的关注度占比作为计算每个景区或酒店总得分时的权重。具体计算公式（式 3.4）如下：

$$\text{final score} = \sum_{d}^{D} attention_d * \text{dimention\_score}_d \quad \text{（式 3.4）}$$

（式 3.4）中，final score表示景区或酒店的最终得分，$attention_d$表示景区或酒店评论文本中，各维度分句数所占比例，D 表示评价维度集合{服务，位置，设施，卫生，性价比}。

3.3.3 模型评价

本章节依据均方误差（Mean Squared Error, MSE）进行模型评价，该统计参数是预测数据和原始数据对应点误差的平方和的均值，具体计算公式为：

$$\text{MSE} = \frac{1}{50}\sum_{c=1}^{50} (\text{final score}_c - true\ score_c)^2 \quad \text{（式 3.5）}$$

（式 3.5）中，final score_c表示第 C 个景区或酒店的预测评分，$true\ score_c$表示第 C 个景区或酒店的真实评分。

使用第一种情感得分规划化方法，对景区的评分预测中，在服务、位置、设施、卫生和性价比上得分的均方误差分别为 0.12，0.79，0.63，0.41 和 0.39；对酒店的评分预测中，在服务、位置、设施、卫生和性价比上得分的均方误差分别为 0.45，0.41，0.72，0.25，0.09。基于上述结果进一步调整了主题词典，并且观察到预测结果评分普遍偏低。考虑到可能是由于数据规范化时，扩大了数据之间的差距，导致与真实值的误差进一步扩大，因此采用第二种方法。依据评论文本中好评数量多于差评及中评数的情况，扩大好评的影响力，最终得到

的结果中，对景区的评分预测中，在服务、位置、设施、卫生和性价比上得分的均方误差分别为 0.24，0.37，0.21，0.10，0.10；对酒店的评分预测中，在服务、位置、设施、卫生和性价比上得分的均方误差分别为 0.11，0.24，0.10，0.03，0.24，除对于景区位置评分的均方误差大于 0.3，其余维度评分的均方误差均低于 0.25，并且在景区卫生、景区性价比、酒店服务、酒店设施和酒店卫生方面的均方误差值为 0.1 左右。

3.3.4 景区及酒店综合评价

依据（式 3.4）计算各景区及酒店的总评分，按总评分从大到小排序，将 50 个酒店及景区划分至高中低三个层次，不同层次景区及酒店的划分结果如下表 5：

表 5：景区及酒店综合评价结果汇总

评价层级	评分详情	景区名称	酒店名称
高	final score>4.4	A12、A28、A13	H20、H16、H03、H29、H01、H35、H10、H12、H30、H17、H48、H37、H46
中	final score=4.4	A07、A03、A11、A09、A27、A33、A01、A19、A35、A42、A37、A23、A20、A45、A10、A02、A14、A04、A05、A06、A32、A15、A21、A25、A40、A24	H25、H15、H05、H21、H04、H22、H08、H14、H49、H26、H23、H31、H02、H33、H42、H24、H18、H41、H11、H07、H09、H45、H43、H19、H50、H34、H40、H13、H06、H44、H47、H39
低	final score<4.4	A08、A16、A50、A26、A46、A38、A49、A22、A36、A17、A43、A47、A48、A29、A31、A41、A39、A44、A18、A34、A30	H32、H28、H36、H38、H27

4 网评文本有效性分析

4.1 构建有效性评价体系

针对网络评论文本经常出现的内容不相关、简单复制修改和无有效内容的问题，本模块分别从内容相关性、内容有效性和内容简单重复三个方面，结合相关文献，构建评论文本有效性评价体系如下表 6 所示。

表 6：评论文本有效性评价指标体系

一级指标	二级指标	指标说明
内容相关性	是否包含主题词	主题词描述景区或酒店的核心词，一般表现为名词。
	是否包含评价词	评价词为与旅游相关的情感词，一般表现为形容词或副词。
	是否包含广告词	基于常用广告词典，或当评论中涉及电话号码、QQ 账号等，即认为包含广告词
	是否包含违禁词	违禁词即含有恶性攻击的词，如 tm、脑残、烂货等。
内容有用性	包含文本字符数	计算剔除标点符号后的文本字符数量
	语义丰富度	识别评论中包含名词性实体的数量
	情感是否极端	识别一条评论情感极性的强弱
内容简单重复	内容是否重复	识别一条评论是否为内容完全重复

4.1.1 内容相关性

在线旅游产品是一种特殊的电商产品，类比电商产品垃圾评论的概念[9]，旅游评论文本针对的是某个具体的景区或酒店，包含产品对各方面的特征和使用体验的真实陈述才是有用的评论，其他的评论都可以视为垃圾评论。因此，探究评论文本是否有价值，可以从评论内容与描述主题是否相关，即内容相关性入手。

4.1.1.1 主题词

产品主题词是描述产品的核心词，也是产品评论的核心词，一般是与产品相关的核心名词。高质量的、有用的、好的评论定义为：能具体描述商品的特征、性能等信息，辅助潜在用户做出适当决策的评论[10]。因此，一条正常评论中一般应该包括与评价主体相关的属性名词。如第二章中，通过对景区及酒店的评论文本进行主题识别，发现用户对于景区主要关注点为景区设施和地理位置，对酒店的关注点集中在酒店地理位置和性价比。

4.1.1.2 评价词

一条正常的评论除了要包括与评价对象相关的属性名词外，一般还应包含有针对该属性名词的评价词。通过计算一条评论中所包含的评价词比例，对于识别无关评论有重要意义[11]。本章节通过对评论语料进行词性标注，识别一条评论语料中与属性相关的形容词或副词的数量，计算评价词数。

4.1.1.3 广告词

对于广告类的垃圾评论一般都包含一些比较明显的关键词，例如一些广告类的垃圾评论，“开业酬宾，想了解更多信息+V 138****3322”，通常包含与商业广告非常相关的关键词，如“开业酬宾”、特定符号组合和手机号。识别这些关键词对评论相关性评价非常重要，本模块整理使用的广告词典包含 56 个常见广告词，部分样例如表:7 所示，进行评论文本中广告词的识别。另外，通过构造正则表达式，如“[1-9]([0-9]{5,11})”判断评论中是否存在 QQ 号，使用 cocoNLP 判断评论中是否存在手机号。

表 7：广告词示例

类别	示例
广告词	火热、特惠、免费、留言、回复、火爆、招商、抢购、积分、充值、淘宝、加微信、加 QQ、咨询电话、客户、联系电话、+V

4.1.1.4 违禁词

针对部分涉及人身攻击的垃圾评论通常具有比较显著的特征，即含有若干网络低俗用语，例如“谁他妈再买这个，谁就是傻逼”中的“他妈”、“傻逼”。针对这类垃圾评论，本文使用的敏感词典包含 14600 个违禁词，部分样例如表 8 所示

表 8： 违禁词示例

类别	示例
敏感词	他妈的、特么、tmd、傻逼、sb、傻叉、脑残、装逼、草蛋、尼玛、妈的

4.1.2 内容有用性

4.1.2.1 评论文本长度

在评论有用性的研究中，Racherla 等认为评论长度代表信息量和评论者的严谨性，是影响顾客行为意图的重要因素，对评论有用性有积极影响[12]。Mudambi 等认为长评论可能包含商品的细节信息，如商品如何被购买和使用，有助于提高评论有用性[13]。基于上述研究，本模块做出如下假设：评论中的剔除标点符号后的文本字符数越多，其内容有用性就越大。

4.1.2.2 语义丰富度

评论的内容和风格是吸引顾客的重要影响因素，现有研究主要以评论所含信息量[14]、评论内容深度[12]等来衡量语义，因此本研究以评论所包含主题词的数量表征语义丰富度，假设一条评论中所包含主题词数越多，其内容有用性就越大。

4.1.2.3 评论情感分析

现有研究多以评分来量化评论者对商品的情感倾向和体验效果，Ghose 等以评分衡量情感发现，温和评论（中立情感评论）的有用性更低[15]。Forman 等也发现，评论情感的极端程度会影响评论有用性[16]。基于此，本模块以评论文本情感倾向性得分（正向或负向得分的最大值）作为评价情感是否极端的依据，假设评论情感倾向性得分越高，该评论有用性就越高。

4.1.3 内容简单重复

有些无效评论直接来自于前一条垃圾评论的复制粘贴，尤其是某些广告类评论为增加曝光率，通常被重复发表。故本文检验数据集中每条评论是否存在重复，如果数据集中存在相同评论，则该属性取值 1，否则为 0。

4.2 基于 LightGBM 的有效评论识别

由于实验所用数据集为无标签文本，本文分别抽取 5862 条去重后的景区评论和 2014 条去重后的酒店评论（重复评论由机器自动标注），进行人工标注，为后续使用机器学习方法自动识别有效评论文本构造有标签数据集。基于评论有效性评价体系，从内容相关性、内容有用性和内容重复性三个方面，将评论文本的定性特征转化为定量指标，使用集成学习算法在有标签数据集上进行训练，并将结果与传统机器学习算法进行对比。

4.2.1 流程设计

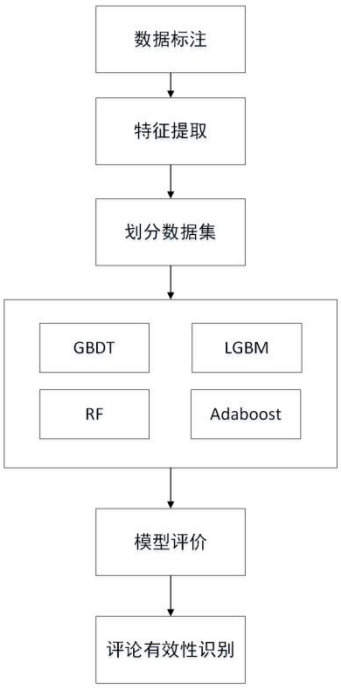

图 21：评论有效性识别流程设计

4.2.2 数据标注

两名标注者分别依据评论有效性评价体系进行标注，为保证标注结果的一致性，事先抽取去重后的 100 条景区评论和 100 条酒店评论，两名标注者进行共同标注，并对标注结果进行一致性检验，使用 Kappa 系数衡量一致性程度，Kappa<0.2 则说明一致性程度较差；0.2~0.4 之间说明一致性程度一般；0.4~0.6 之间说明一致性程度中等；0.6~0.8 之间说明一致性程度较强；0.8~1.0 之间说明一致性程度很强。最终得到的结果中，景区评论标注结果的 Kappa 值为 0.627（p 值<0.01），酒店评论标注结果的 Kappa 值为 0.912（p 值<0.01），表明两个标注者标注结果的一致性较强。

4.2.3 模型简介

4.2.3.1 GBDT 算法

GBDT（Gradient Boosting Decision Tree）又称为梯度提升决策树。GBDT 内部集成了多棵决策树，而最终结果是多棵决策树结果的累加，有效地避免了单棵决策树的过拟合问题，使得算法能够很好地推广，并广泛应用于各个领域。

GBDT 算法的主要过程为：GBDT 的第一棵决策树会从原始数据样本进行处理，产生一个结果和残差，然后第二棵树对第一棵树的结果和残差进行学习，后面的树不断重复这一过程。所谓残差指的是真实值与预测值之间的差值。残差在此过程中会衰减，直至小于规定阈值。GBDT 的核心思想[17]在于通过不断地对前一棵树的残差进行学习进而降低损失函数。GBDT 每次都会选择残差减少幅度大的点作为分裂点，减少的越多，分裂点越好。GBDT 会利用梯度迭代的方法使得每次迭代产生的损失函数最小，通过梯度下降的方法让损失函数越

来越小，使模型变得越来越精。

4.2.3.2 改进的 GBDT——LightGBM

LightGBM（Light Gradient Boosting Machine Method，以下简称 LGBM）是轻量级梯度提升器，是微软在 2017 年提出的一个基于决策树的梯度提升算法框架，是在 GBDT 基础上的改进，以解决高维度大样本数据运行耗时及可拓展性差的问题。LGBM 的实质是一种将弱学习器提升为强学习器的集成学习算法，具体是将许多准确率较低的树模型组合起来，经过不断迭代并采用梯度下降的方法，在每次迭代时通过向损失函数的负梯度方向移动来使得损失函数越来越小，最终得到一棵较优的树，并以此作为预测模型。该算法对于 GBDT 模型的改进主要表现在以下 5 个方面：

（1）Histogram 算法

LGBM 利用直方图算法加速数据处理，提升算法性能。主要过程为：将特征值的浮点型离散为 K 个整数型，并构造宽度 K 的直方图。把训练样本数据导入之后，直方图算法会遍历数据，在遍历数据的时候，算法会将数据离散化后所得到的值作为索引，然后在直方图中累计统计量。最后可以从离散值中找出最佳分裂点。

使用直方图不仅能降低计算内存，离散后的算法只需要计算 K 次信息增益，降低了计算的复杂度。但由于将特征值离散成了 K 个值，牺牲了数据的精准性，所以直方图算法也会有一定的缺点。

（2）带深度限制的 Leaf-wise 的叶子生长策略

常用的 GBDT 的叶子是按层生长的，可以优化多线程优化，控制复杂度且不易过拟合。但同时也造成了许多多余的计算负担。LGBM 则采用了更为高效的按叶子生长(leaf-wise)算法，基本思路在于从某一层叶子中寻找分裂增益最大进行分裂并不断重复，可以在同样的分裂次数下获得更优的精度，缺点在于树会分裂较深，易发生过拟合，因此需要对决策树设置深度限制。

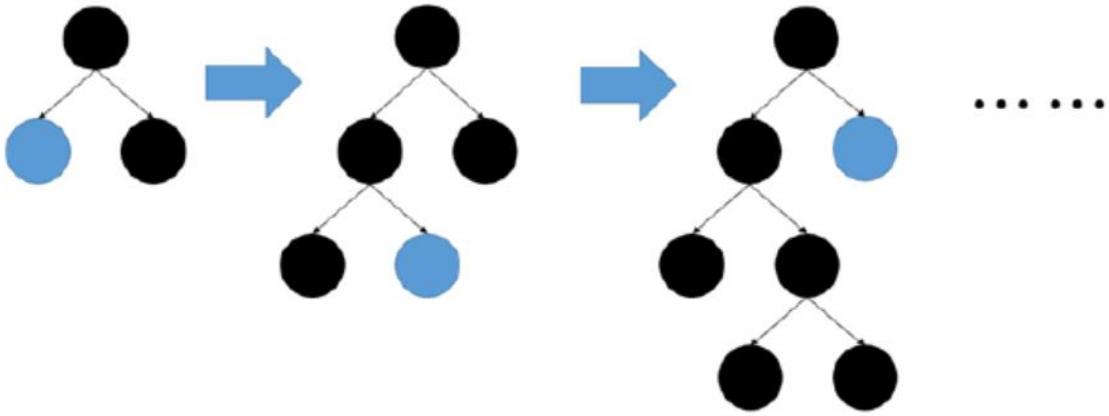

图 22：LightGBM 决策树生长策略[18]

（3）GOSS 抽样方法

GOSS(Gradient-based One-Side Sampling)是一种新颖的 GBDT 抽样方法，可以在减少数据实例数量和保持决策树学习的准确性之间取得良好的平衡。GOSS 的基本原理为：保存下全部大梯度的数据实例，并对小梯度的数据实例随机抽样。然后用小梯度的实例乘以一个常数以放大抽样样本，这样就可以在不改变原始数据分布的情况下，提升 LGBM 的性能，并减小抽样对数据分布的影响。

（4）独立特征捆绑

独立特征捆绑（Exclusive Feature Bundling，EFB）是一种有效减少特征数量的新方法。高维数据一般很稀疏。其稀疏性提供了使用算法减少特征维度的机会，且这些特征一般互斥，可以安全地将独立的特征捆绑到一个特征（称之为独立特征束）。运用特征扫描算法，可以将这些“束”构建与单个特征一样的特征直方图。此时复杂度大大降低，GBDT 的训练速度加快且精度不受影响。

如何选择捆绑的特征以及如何构建捆绑，是两个主要需要解决的问题。关于问题一，LGBM 首先构造一个有权重边的图，权重对应于特征之间的总冲突。其次，按照下降的顺序排列。最后，查看有序列表中的每个特征，并将其分配给一个具有小冲突的现有特征束，或者创建一个新的特征束。关于问题二，该问题的重点在于从特征束中识别原始特征。直方图算法的存储值为离散型，故可以将各箱中的独立特征实现特征捆绑。这可以通过将偏移量添加到特征原始值来完成。例如，假设一个特征束捆绑了两个特征。最初，特征 A[0,10)和特征 B[10,20)。对特征 B 的值进行 10 的修正，修正后为[10,30]。之后，合并特征 A 和 B 并使用范围为[0,30]的特征束来替换它们[18]。

（5）直方图差加速

LightGBM 运用了直方图做差加速的方法实现速度的提升。其基本原理为利用父节点与兄弟节点的直方图相减得到所求节点的直方图，只需遍历直方图的 K 个箱即可。运算速度因此得以提升。

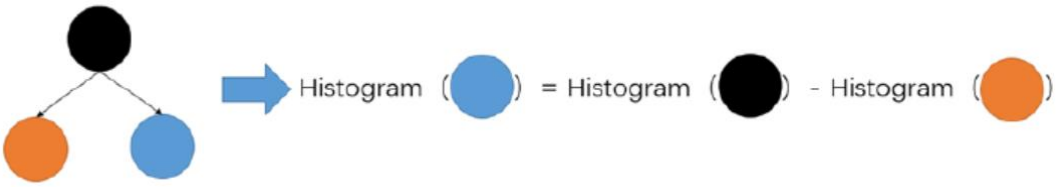

图 23：LightGBM 直方图加速原理

4.3 模型评价及对比

4.3.1 评价指标

4.3.1.1 准确率

准确率是分类模型中最常见的性能评价指标，准确率越高，分类器效果越好。对于给定的测试集数据，准确率是被模型正确预测的样本数的累加与总样本数之比。具体计算公式如

下：

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{（式 4-1）}$$

4.3.1.2 F1 值

F1 值是查全率（Recall，简记为 R）和查准率（Precision，简记 P）的调和平均值。F1 值综合考虑了 P 和 R 的结果，F1 越高，则说明模型在数据集上越有效。具体计算公式如下：

$$\text{P} = \frac{TP}{TP+FP} \quad \text{（式 4-2）}$$

$$\text{R} = \frac{TP}{TP+FN} \quad \text{（式 4-3）}$$

$$\text{F1} = \frac{2*P*R}{P+R} \quad \text{（式 4-4）}$$

公式（4-1、4-2、4-3、4-4）中，TP 表示为正确地被识别为有效评论的数量，TN 表示为正确地被识别为无效评论的数量，FP 表示错误地被识别为有效评论的数量，FN 表示为错误地识别为无效评论的数量。

4.3.2 模型构建及对比

数据集由人工和机器共同标注，共 9060 条带标签的评论数据，其中，6300 条为景区评论，2760 条为酒店评论。首先，使用 python 程序，按照有效性评价体系进行特征提取。其次，按 3：1 的比例划分训练集和测试集。再者，本模块分别使用了 python lightgbm 包和 sklearn.ensemble 中的 GradientBoostingClassifier 方法，分别构建 LGBM 和 GBDT 模型，输入训练集数据，进行模型训练。最后，通过输入测试集数据验证模型的准确性，通过交叉验证的方式调整模型参数。此外，在确定最优参数后，本模块还分别构建了决策树、SVM、BP 神经网络、随机森林和 Adaboost 模型，将模型预测的结果与 LGBM 模型的结果进行对比。最终，汇总所有模型在测试集上的对有效评论分类预测的准确率、不同标签上的分类精度及 F1 值如下表 9 所示。

表 9：不同模型分类准确度和 F1 值

名称	**模型名称**													
	LGBM		**GBDT**		**决策树**		**SVM**		**BP**		**RF**		**Adaboost**	
分类准确度	0.94		0.93		0.89		0.90		0.91		0.92		0.92	
指标名称	**P**	**F1**	**P**	**F1**	**P**	**F1**	**P**	**F1**	**P**	**F1**	**P**	**F1**	**P**	**F1**
1:有效评论	**0.94**	**0.84**	0.93	0.83	0.74	0.75	0.92	0.75	0.93	0.75	0.87	0.82	0.98	0.79
0:无效评论	**0.93**	**0.96**	0.94	0.96	0.93	0..93	0.90	0.94	0.90	0.94	0.94	0.95	0.91	0.95

分析表 9 可知，LGBM 模型的分类准确度最高，为 0.94，其次是 GBDT 模型，分类准确度为 0.93。Adaboost 算法对有效评论的识别精度最高，但召回率较低，仅 0.79。

4.4 文本有效性预测分析

基于训练好的 LGBM 模型，预测得到景区评论的总体有效性为 89.4%，酒店评论的总体有效性为 82.1%。分别针对 50 个景区及 50 个酒店的评论自动进行有效性识别，最终得到

的结果如图 24、图 25 所示。分析图 24 可知，50 个景区的有效文本数均在 80%以上，各景区重复评论数所占比例较少，A04 景区无效评论数占比最多，A49 景区有效评论数占比最多。分析图 25 可知，除 H35、H45 和 H48 酒店外，剩余酒店的有效评论数占比均位于 80%左右，酒店的重复评论数占比较少，H45 酒店的无效评论数占比最多，H41 酒店的有效评论数占比最多。

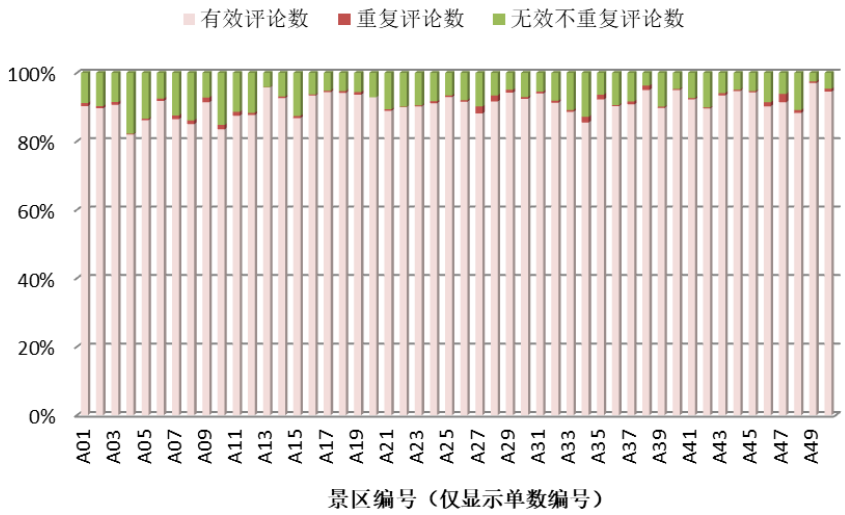

图 24：各景区评论有效性分布

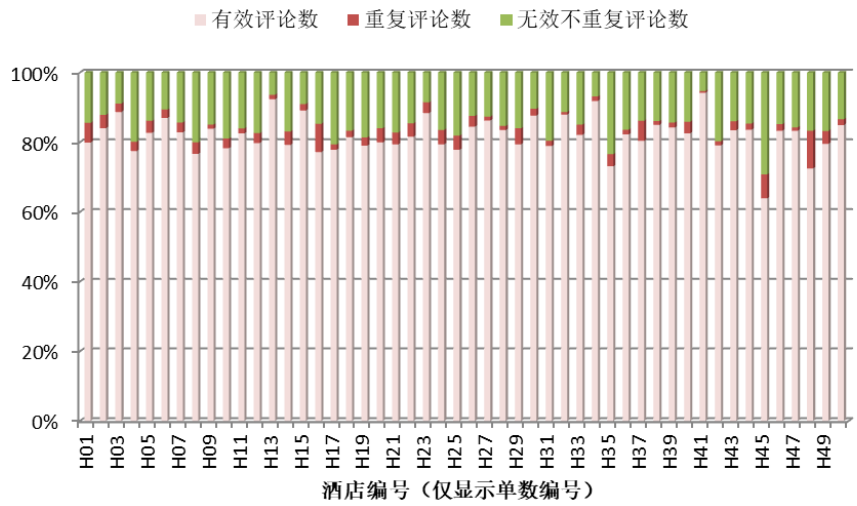

图 25：各酒店评论有效性分布

5 景区及酒店特色分析

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。基于第 3 章中综合评价模型的结果，分别选取综合排名高中低三个层次中的各三家景点和三家酒店，九个景点及九家酒店的综合得分结果如表 10 所示，分别依据提及维度、正负向评论主题识别及特征词分析，分析用户对各景区及酒店的关注点，挖掘不同层次中景区及酒店各自的特色。

表 10：不同层级九个景区及九家酒店综合评价结果

评价层级	名称	总得分	服务得分	位置得分	设施得分	卫生得分	性价比得分
高	A12	**4.5**	4.5	4.4	4.3	4.6	4.5
	A28	**4.5**	4.4	4.6	4.3	4.7	4.3
	A13	**4.5**	4.4	4.4	4.4	4.6	4.5
中	A19	**4.4**	4.5	4.4	4.2	4.6	4.3
	A35	**4.4**	4.4	4.4	4.1	4.7	4.4
	A42	**4.4**	4.3	4.3	4.3	4.7	4.4
低	A18	**4.3**	4.4	4.0	4.1	4.6	4.3
	A34	**4.3**	4.4	4.1	4.1	4.3	4.3
	A30	**4.2**	4.3	4.2	4.1	4.3	4.3
高	H20	**4.6**	4.5	4.5	4.6	4.8	4.7
	H16	**4.5**	4.4	4.5	4.5	4.8	4.7
	H03	**4.5**	4.5	4.3	4.5	4.7	4.6
中	H41	**4.4**	4.4	4.2	4.4	4.7	4.6
	H11	**4.4**	4.4	4.2	4.4	4.7	4.5
	H07	**4.4**	4.4	4.3	4.3	4.6	4.4
低	H36	**4.3**	4.2	4.2	4.3	4.6	4.5
	H38	**4.3**	4.2	4.1	4.3	4.5	4.4
	H27	**4.3**	4.2	4.1	4.3	4.5	4.4

5.1 数据描述

使用基于主题词典的方法，分别识别用户对各景区及酒店在服务、位置、设施、卫生和性价比方面具体的关注点差异，对高中低三个层级中各三家景点的用户关注度的统计分布如下图 26 所示，对三家酒店用户关注度的统计分布如下图 27 所示。

基于图 26，分析高层级的三家景区可知，对 A12 景区，用户提及量最多的依次为景区的票价、风景品质和服务细节，涉及服务人员的评论较少；对 A28 景区，用户提及量最多的为景区的服务人员，其次是景区票价和餐饮；对 A13 景区，景区服务人员的被提及量最高，其次是景区的公共设施和餐饮服务。

在中层级的三家景区中，对 A19 景区，用户最关注景区的服务人员、票价和公共设施；对 A35 景区，用户最关注景区的票价、风景品质和景区配置；对 A42 景区，用户提及量较高的依次为服务人员、票价和风景品质。

分析低层级的三家景区可知，对 A18 景区，用户提及景区的票价、服务人员以及风景品质较多；对 A34 景区，用户最关注票价，其次是风景品质和公共设施；对 A30 景区，用户最关注景区的配置、票价和交通条件。

综上，用户对不同景区的关注点各有侧重。对于高层级的景区，服务（服务细节、服务人员、餐饮服务）一直是用户关注的热点，A12 和 A28 景区的票价被用户提及较多，A13 景区公共设施方面被用户提及较多；对于中层次的景区，票价一直是用户关注的热点，A19

景区及 A42 景区的服务人员的被提及量较高，A35 和 A42 景区的风景品质受到用户较多关注；对于低层次的景区，票价同样是用户评论的热点话题，A18 景区和 A34 景区的风景品质受到用户较多关注，A30 景区的交通条件被用户提及较多。

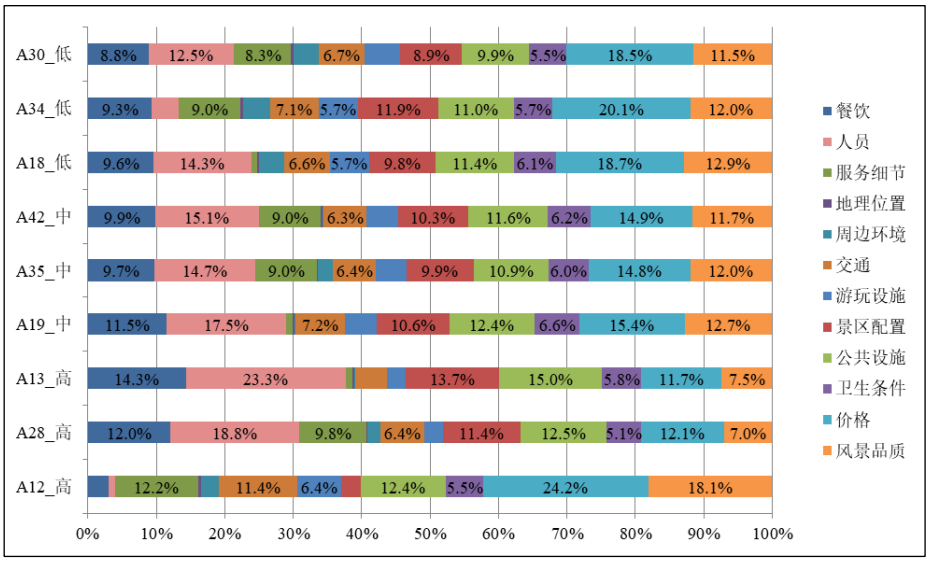

图 26：不同层级景区用户关注度分布

基于图 27，分析高层级的三家酒店可知，对 H03 酒店，服务人员和餐饮是用户的主要关注点，其次是服务的细节，这三者均与酒店的服务有关；对 H16 酒店，用户最关注酒店的服务人员、酒店设施以及餐饮条件；对 H20 酒店，用户提及酒店设施、服务人员和卫生条件较多；

分析中层级的三家酒店可知，对 H07 酒店，餐饮和酒店设施是用户关注的热点；对 H11 酒店，用户最关注酒店的餐饮、服务人员和酒店设施；对 H41 酒店，用户提及餐饮服务、人员和酒店设施较多。

分析低层级的三家酒店可知，对 H27 酒店，用户最关注餐饮、酒店设施和服务人员；对 H36 酒店，餐饮、服务人员和设施是用户评论的热点；对 H38 酒店，用户最关注的三个方面依然是酒店的餐饮、服务人员和酒店设施。

综上，服务和设施是用户主要关注的热点，即使处于不同层次的酒店，用户评论中最关注的基本围绕酒店服务和设施展开。对于高层次的酒店，H03 酒店服务备受关注，H16 和 H20 酒店的设施受到用户较多关注；对于中层级酒店，H27 酒店的餐饮和设施受到较多关注，H41 酒店和 H27 酒店最受用户关注的方面占比相差不大，但 H41 酒店的服务细节所受关注度高于 H27 酒店；对于低层次酒店，三个酒店最受用户关注的同样为餐饮、设施和人员，但用户对 H36 和 H38 酒店服务细节的关注度占比高于 H27 酒店。

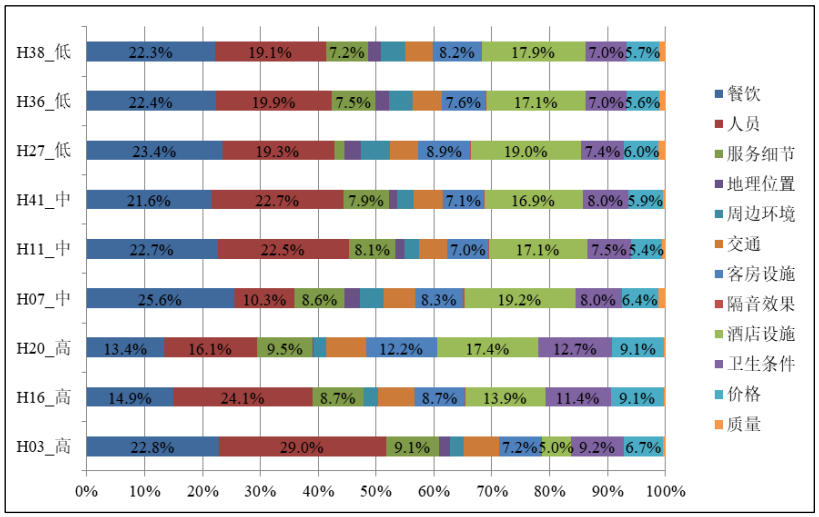

图 27：不同层级酒店用户关注度分布

5.2 正负向评论的主题识别

本模块依据 Senta 对景区及酒店的全部评论进行情感分析，依据评论标签 1 或 0 划分正负向评论，标签为 1 表示正向评论，标签为 0 表示负向评论。通过分别对正负向评论进行主题挖掘，在发现景区及酒店特色的同时，了解各景区及酒店的不足之处，为提升服务质量提供参考。首先，对各景区及酒店的评论进行文本预处理（保留分词词性标注为名词及形容词的结果），依据 LDA 模型主题识别的结果，选取聚类主题数，每个主题数由若干词来概括，每个词都有对应的权重，权重反映该词对于该主题的概括程度，即生成概率，选取每个主题生成概率前 10 的词汇，表征评论主题。

以 A12 景区为例，正负面主题识别结果示例分别如下图 28 和图 29 所示。为保证主题识别的全面性，依据主要评价维度选择主题聚类数为 5 个，主题词数为 10 个。对于 A12 景区，正面评论中主要关注景区的服务和性价比，具体表现为对景区中方便快捷取票服务表示满意，对景区中优美的环境表示认可；负面评论主要集中在性价比和设施上，主要关于门票价格太高和景区风景配置单一。

```
corpus/景区/A12_pos_cut.txt正面主题分析
0.020*"不错" + 0.013*"鱼" + 0.013*"玩" + 0.012*"方便" + 0.012*"适合" + 0.011*"地方" + 0.010*"值得" + 0.010*"票" + 0.009*"里面" + 0.009*"环境"
0.023*"不错" + 0.019*"鱼" + 0.015*"值得" + 0.014*"方便" + 0.013*"里面" + 0.012*"票" + 0.012*"适合" + 0.012*"玩" + 0.012*"地方" + 0.010*"景色"
0.030*"不错" + 0.016*"玩" + 0.015*"方便" + 0.014*"鱼" + 0.013*"适合" + 0.012*"里面" + 0.011*"值得" + 0.011*"地方" + 0.011*"环境" + 0.009*"景色"
0.030*"不错" + 0.016*"方便" + 0.015*"玩" + 0.015*"鱼" + 0.013*"环境" + 0.013*"里面" + 0.012*"适合" + 0.011*"挺" + 0.011*"票" + 0.011*"地方"
0.029*"不错" + 0.023*"鱼" + 0.017*"值得" + 0.016*"里面" + 0.015*"环境" + 0.014*"方便" + 0.013*"玩" + 0.012*"票" + 0.011*"适合" + 0.011*"地方"
```

图 28：A12 景区正面评论潜在主题

```
corpus/景区/A12_neg_cut.txt负面主题分析
0.023*"买" + 0.022*"鱼" + 0.018*"太" + 0.013*"玩" + 0.013*"票" + 0.012*"贵" + 0.011*"感觉" + 0.011*"门票" + 0.010*"里面" + 0.009*"逛"
0.025*"鱼" + 0.022*"买" + 0.017*"太" + 0.014*"票" + 0.013*"玩" + 0.012*"里面" + 0.012*"门票" + 0.011*"贵" + 0.010*"感觉" + 0.010*"好玩"
0.024*"鱼" + 0.016*"里面" + 0.016*"太" + 0.016*"买" + 0.015*"玩" + 0.012*"票" + 0.012*"门票" + 0.011*"贵" + 0.010*"好多" + 0.009*"没什么"
0.021*"鱼" + 0.016*"太" + 0.016*"票" + 0.016*"买" + 0.015*"里面" + 0.013*"玩" + 0.011*"门票" + 0.009*"贵" + 0.009*"感觉" + 0.009*"好多"
0.022*"鱼" + 0.017*"买" + 0.016*"太" + 0.013*"票" + 0.011*"门票" + 0.010*"里面" + 0.009*"感觉" + 0.008*"贵" + 0.008*"玩" + 0.008*"门口"
0.018*"鱼" + 0.012*"太" + 0.011*"买" + 0.010*"票" + 0.009*"门票" + 0.009*"里面" + 0.008*"玩" + 0.007*"逛" + 0.007*"好玩" + 0.007*"好多"
```

图 29：A12 景区负面评论潜在主题

使用同样的方法，分别对不同层级景区及酒店的正负面评论识别结果进行对比总结（所有主题识别具体结果见附录），使用表格形式展示，括号中为部分相关主题词的具体含义。

5.2.1 景区正负面评论主题分析

5.2.1.1 高层次景区

表 11：高层次景区评论主题识别

名称	正面评论主题	负面评论主题
A12	服务（取票快捷）、性价比（环境优美）	性价比（票价高、风景单一）
A28	服务（取票快捷）、设施（沙滩）	性价比（票价高）、服务（排队）
A13	设施（温泉）、服务（服务态度热情）	性价比（收费服务）、卫生（水质）

分析表 11 可知，对于 A12 景区，正面评论中主要关注景区的服务和性价比，具体表现为对景区中方便快捷取票服务表示满意，对景区中优美的环境表示认可；负面评论主要关注景区的票价以及风景品质。对于 A28 景区，正面评论主要围绕快捷的取票服务、便宜的票价以及沙滩设施，负面评论主要围绕门票价格和排队情况。对 A13 景区，景区的温泉设施和热情的服务态度给顾客留下了深刻的印象，负面评论与收费服务和水质相关。

5.2.1.2 中层次景区

表 12：中层次景区评论主题识别

名称	正面评论主题	负面评论主题
A19	性价比（风景优美）、卫生（空气清新）	性价比（票价高）、设施（游船）
A35	性价比（风景壮观）、服务（取票方便）	性价比（票价高）
A42	服务（服务不错）、设施（温泉）	服务（温泉没开）

分析表 12 可知，对 A19 景区，正面评论主要关注优美的风景、清新的空气，负面评论主要关于票价和游船设施。对于 A35 景区，正面评论主要围绕壮观的风景和便捷的取票服务，负面评论与票价相关。对于 A42 景区，正面评论主要与服务和温泉体验相关，而负面评论主要与温泉服务未开放相关。

5.2.1.3 低层次景区

表 13：低层次景区评论主题识别

名称	正面评论主题	负面评论主题
A18	性价比（门票免费）、设施（建筑景观）	性价比（风景单一）
A34	性价比（好玩）、设施（项目）	性价比（票价高）
A30	设施（动物、海洋馆）、性价比（值得游玩）	性价比（票价高）、设施（设施陈旧）

分析表 13 可知，对 A18 景区，正面评论主要关注景区的门票和设施，负面评论与单一的风景品质相关。对 A34 景区，正面评论主要围绕景区的游玩体验和设施项目，负面评论与票价相关。对 A30 景区，正面评论中，用户主要关注景区设施和服务，具体表现为对动物园配置和取票服务的认可；负面评价主要围绕性价比和设施，用户主要反映门票价格太贵，设施有待升级。

5.2.1.4 总结分析

综合来看，服务、设施和性价比是正向评论中的热点，需要景区管理者继续维持上述特征，进一步扩大景区的影响力。门票价格是用户负面评价中普遍关注的问题，对于景区的负面评价，商家也需要及时关注，做出相应的改进措施，以提高用户的满意度。

5.2.2 酒店正负面评论主题分析

5.2.2.1 高层次酒店

表 14：高层次酒店评论主题识别

名称	正面评论主题	负面评论主题
H03	性价比（五星级）、卫生（房间环境）	服务（催退房）
H16	服务（前台服务人员）、卫生（房间环境）	设施（隔音效果不好）
H20	卫生（房间干净）、设施（设施齐全）	设施（硬件设施陈旧）、服务（餐饮）

分析表 14 可知，对 H03 酒店，用户正面评价主要围绕酒店的性价比和房间环境，负面评价集中在退房服务上。对 H16 酒店，正面评价围绕服务人员和房间环境，而负面评价与房间的隔音效果相关。对 H20 酒店，用户主要关注酒店卫生和设施，说明用户对酒店干净整洁的客房布置和齐全的设施表示满意；负面评价集中于酒店的设施和服务方面，主要表现为对电视、网络等硬件设施以及前台服务和餐饮服务的不满。

5.2.2.2 中层次酒店

表 15：中层次酒店评论主题识别

名称	正面评论主题	负面评论主题
H07	服务（早餐）、位置（江景）	卫生（房间卫生）
H11	服务（前台、早餐）、设施（沙滩、泳池）	设施（电梯）
H41	服务（早餐、方便）、设施（温泉）	服务（前台服务人员）

分析表 15 可知，对 H07 酒店，用户主要关注酒店的早餐和位置，主要表现为对早餐及周边环境的满意，负面评价主要与房间卫生相关。对于 H11 酒店，用户主要关注酒店的服务和设施，说明用户对酒店的服务人员和餐饮品质表示满意，对泳池、沙滩等配置表示认可；负面评价主要关于酒店卫生和电梯设施，说明酒店的卫生条件有待改善，电梯等设施需要升级。对 H41 酒店，正面评价主要关注酒店的服务和设施，具体表现为对酒店的早餐服务、便捷入住和温泉设施的满意；负面评价主要与前台服务人员相关。

5.2.2.3 低层次酒店

表 16：低层次酒店评论主题识别

名称	正面评论主题	负面评论主题
H27	服务（早餐）、位置（小镇）	设施（设施陈旧）
H36	服务（早餐）、位置（七星岩）	卫生（房间卫生）
H38	设施（泳池、沙滩）、服务（早餐）	设施（设施陈旧）

分析表 16 可知，对 H27 酒店，用户主要关注酒店的服务和地理位置，具体表现为对餐饮服务和周边环境的认可；负面评价主要围绕酒店的餐饮服务和设施，说明该酒店餐饮服务的质量有待提高，设施有待更新。对 H36 酒店，正面评价主要关于早餐服务和周边环境，负面评价主要与房间卫生相关。对 H38 景区，正面评价主要关于酒店设施和服务，具体表现为对沙滩、泳池等设施的满意，对早餐服务的认可，而负面评价主要与陈旧的设施相关。

5.2.2.4 总结分析

综上，服务、位置和设施是用户正面评价中关注的热点，需要酒店运营者继续维持上述特征，以维系酒店的口碑。设施和卫生为用户负面评价中关注的热点问题，酒店管理者需要重点关注，及时更新酒店设施，提升用户体验，提高用户的满意度。

5.3 基于特征词的特色分析

通过 LDA 主题识别的结果可知，用户对于不同景区或不同酒店的关注点各有侧重，关注的细节各异。为进一步挖掘各景区及酒店的特色，本模块通过分析正面评论中具体的特征词完善各景区和酒店的特色分析。

5.3.1 各景区特色分析

5.3.1.1 高层次景区（A12、A28、A13）词云对比

图 30：高层次景区正向词云对比图（从左到右依次为 A12、A28 及 A13）

分析高层次景区正向词云图可知，A12 景区以园林景观为特色，风景优美；景区中有锦鲤池塘，适合带小朋友参观游玩；支持网上订票，取票方便。A28 景区以海洋沙滩为主要景观，餐饮以海鲜为特色；订票快捷、取票方便，且价格优惠。A13 景区以温泉服务为特色，景区服务人员热心负责，取票方便，温泉水干净卫生，游玩体验舒适。

5.3.1.2 中层次景区（A19、A35、A42）词云对比

图 31：中层次景区词云图对比（从左到右依次为 A19、A35 及 A42）

分析图 31 可知，A19 景区以湖上风光为特色，主要活动为游船，湖水清澈，环境优美，空气清新，适合晴天游玩。从 A35 景区的正向评论词云图分析可知，该景区以自然景观为特色，有壮观的瀑布群和清新的空气，是夏日踏青出游的好去处。A42 以温泉服务为特色，环境卫生干净，价格实惠，顾客游玩体验舒适。

5.3.1.3 低层次景区（A18、A34、A30）词云对比

图 32：低层次景区特色分析（从左往右依次为 A18、A34 和 A30）

分析低层次景区词云图可知，A18 景区为一个特色建筑群，以各式建筑为主要景观，部门景点的门票免费，景色优美。A34 景区是一个游乐园，取票方便，以丰富刺激的娱乐设施为特色，适合带小朋友游玩。A30 景区为一个大型的动物园，大熊猫、长颈鹿、老虎以及海洋馆中的各类动物给游客们留下了深刻的印象；由于景区中有各类动物，还有动物表演，非常适合带小朋友参观游玩；景区交通便利，距离地铁站近。

5.3.2 各酒店特色分析

5.3.2.1 高层次酒店（H03、H16、H20）词云对比

图 33：高层次酒店词云图分析（从左往右依次为 H03、H16 和 H20）

分析高层次酒店的词云图可知，H03 酒店以贴心的服务和干净的卫生为特色，提供早餐，部分房间可以看到江景。H16 酒店以热情的服务人员和整洁的卫生为特色，酒店的性价比较高。H20 酒店同样以干净的卫生条件和热情的服务为特色，同时酒店设施齐全，还提供有健身房，交通比较便利。

5.3.2.2 中层次酒店（H07、H11、H41）词云对比

图 34：中层次酒店词云图分析（从左往右依次为 H07、H11 和 H41）

分析中层次酒店的词云图可知，H07 酒店为五星级的老牌酒店，提供特色的餐饮服务，

周边环境优美，提供江景房，交通便利。H11 酒店以热情周到的服务和室外设施为特色，贴心的服务人员和沙滩泳池是用户评论的热点。H41 酒店为一处度假酒店，距离园区较近，提供温泉服务和早餐服务，环境干净卫生。

5.3.2.3 低层次酒店（H27、H36、H38）词云对比

图 35：低层次酒店特色分析（从左往右依次为 H27、H36 和 H38）

分析低层次酒店的词云图可知，H21 酒店可能位于一个小镇上，周边风景优美，适合旅游度假居住；酒店提供早餐，还设置有游泳池。H36 酒店可能为一个五星级酒店，位于市中心，交通便利，距离七星岩景区较近；房间舒适宽敞，提供早餐服务。H38 酒店为一处度假酒店，以沙滩泳池为特色，提供海景房，设施齐全，环境安静优美。

6 总结

随着网络技术的快速发展和旅游业信息的高度密集，在线评论平台上出现了大量以评论和游记等为形式的非结构化数据，传统的调研方式已无法满足如今动辄上万的数据挖掘需求，本文主要通过 TF-IDF 算法、LDA 主题模型、文本情感分析和 LightGBM 模型解决了景区及酒店的印象分析、综合评价、网评文本有效性识别、特色分析四个任务。

为制作任务一的印象词云表，本文在对原始数据进行预处理后，保留词性标注为名词和形容词的词语，对比了基于词频、TF-IDF 和 textrank 三种关键词提取算法的计算结果，结果显示 TF-IDF 算法的提取结果最优。

为解决任务二中的综合评价问题，本文首先通过保留评论文本中的名词性短语，基于 LDA 主题模型进行主题识别，并结合相关文献构建的主题词词典。随后，通过对评论文本进行分句处理，解决一条评论句中出现多主题的问题。再基于主题词典进行分句筛选和分类，使用 Senta 对分句进行情感分析，并使用特定公式规范和修正情感得分。最后，基于用户对各评价维度的关注度，计算每个景区或酒店总得分时的权重，进行综合评价。通过计算模型在对各景区和酒店预测评分的均方误差，评价模型的有效性。

为解决任务三中网评文本有效性分析问题，文本查阅相关文献，分别从文本相关性、文本有用性、文本简单重复三个方面构建了评论文本有效性评价体系。采用监督学习的方式，构建 LightGBM 模型并训练，并对比了该模型与其他模型的分类准确度，结果表明，本文采用的模型在识别有效性文本时的准确度最高，并使用该模型对所有景区及酒店的评论文本进行有效性预测。

为解决任务四中的特色分析问题，基于任务二中的综合评价结果，分别筛选了高中低三个层次中各三个景区和酒店。对比了对不同层级中各个景区和酒店的用户具体关注点的差异，分析正负向评论中，用户关注的热点问题，了解不同景区及酒店的亮点和不足；通过分析不同层级中景区和酒店的正向评论的特征词云图，进一步挖掘各景区及酒店的特色。

在对赛题进行充分研究后，本文对四个任务提出了合适的解决方案，基本实现了赛题设立的目标。

参考文献

[1] Crompton .John L. Motivations for Pleasure Vacation[J]. Annals of Tourism Research, 1979,6(4):408-424.

[2] 陈天琪,张建春.基于文本挖掘的景区旅游形象感知研究——以杭州西溪国家湿地公园为例[J/OL].资源开发与市场:1-9[2021-04-24].http://kns.cnki.net/kcms/detail/51.1448.n.20210412.1629.012.html.

[3] 张振. 基于字符级卷积神经网络的民宿顾客意见挖掘[D].重庆师范大学,2019.

[4] 常耀成,张宇翔,王红,万怀宇,肖春景.特征驱动的关键词提取算法综述[J].软件学报,2018,29(07):2046-2070.

[5] Salton G, Buckley C. Term-Weighting approaches in automatic text retrieval. Informati-on Processing & Management, 1988,24(5): 513–523.

[6] Mihalcea R, Tarau P. TextRank: Bringing order into texts. In: Proc. of the EMNLP. Stroudsburg: ACL, 2004. 404–411.

[7] David M. Blei, AndrewY. Ng, Michael I. Jordan, LatentDirichlet Allocation, Journal of Machine Learning Research 3, p993-1022,2003

[8] 缪章伟.酒店顾客满意度评价体系研究[D].浙江工商大学,2019.

[9] 唐子豪.基于改进 LDA 的在线商城垃圾评论识别研究[D].西安理工大学,2020.

[10] 林煜明,王晓玲，朱涛，等.用户评论的质量检测与控制研究综述[J].软件学报,2014,25(3):506-527.

[11] 昝红英,毕银龙,石金铭.基于 Adaboost 算法与规则匹配的垃圾评论识别[J].郑州大学学报(理学版),2017,49(01):24-28.

[12] Racherla, P.,Friske, W..Perceived “Usefulness” of Online Consumer Reviews: An Exp-loratory Investigation across Three Services Categories. Electronic Commerce Research & Applications, 2012, 11(6): 548-559.

[13] Mudambi, S. M.,Schuff, D.. What Makes a Helpful Online Review? A Study of Cus-tomer Reviews on Amazon. com. MIS Quarterly, 2010, 34(1): 185-200.

[14] Mckinney, V.,Yoon, K.,Zahedi, F. M.. The Measurement of Web-customer Satisfaction: An Expectation and Disconfirmation Approach. Information Systems Research, 2002, 13(3): 296-315.

[15] Ghose, A.,Ipeirotis, P. G.. Estimating the Helpfulness and Economic Impact of Produ-ct Reviews: Mining Text and Reviewer Characteristics. Social Science Electronic Publishing, 2011, 23(10): 1498-1512.

[16] Forman, C.,Ghose, A.,Wiesenfeld, B.. Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. Information Systems Research, 2008, 19(3): 291-313.

[17] 王天华.基于改进的 GBDT 算法的乘客出行预测研究[D].大连理工大学, 2016.

[18] 范德祥. 基于 LightGBM 的居民出行方式选择模型及应用研究[D].华中科技大学,2018.

附录

1 景区正负面评论 LDA 主题识别结果

（1）高层次景区（A12、A28、A13）

```
../corpus/景区/A12_pos_cut.txt正面主题分析
0.028*"不错"·+·0.013*"方便"·+·0.012*"值得"·+·0.012*"适合"·+·0.011*"鱼"·+·0.011*"玩"·+·0.011*"里面"·+·0.010*"环境"·+·0.009*"小朋友"·+·0.009*"好玩"
0.028*"不错"·+·0.016*"地方"·+·0.015*"方便"·+·0.014*"值得"·+·0.013*"适合"·+·0.012*"里面"·+·0.012*"环境"·+·0.010*"玩"·+·0.009*"景色"·+·0.008*"小朋友"
0.030*"不错"·+·0.017*"里面"·+·0.015*"值得"·+·0.012*"方便"·+·0.012*"地方"·+·0.009*"小朋友"·+·0.009*"玩"·+·0.009*"景色"·+·0.009*"适合"·+·0.008*"开心"
0.024*"不错"·+·0.014*"里面"·+·0.013*"方便"·+·0.012*"玩"·+·0.012*"值得"·+·0.011*"适合"·+·0.011*"环境"·+·0.010*"地方"·+·0.010*"小孩"·+·0.009*"景色"
0.025*"不错"·+·0.015*"里面"·+·0.015*"方便"·+·0.013*"适合"·+·0.012*"值得"·+·0.011*"小朋友"·+·0.011*"玩"·+·0.011*"地方"·+·0.010*"小孩"·+·0.009*"环境"

../corpus/景区/A28_pos_cut.txt正面主题分析
0.034*"方便"·+·0.020*"便宜"·+·0.019*"不错"·+·0.018*"取票"·+·0.015*"排队"·+·0.011*"套票"·+·0.009*"比较"·+·0.009*"沙滩"·+·0.008*"挺"·+·0.007*"不用"
0.042*"方便"·+·0.028*"取票"·+·0.025*"不错"·+·0.022*"便宜"·+·0.016*"排队"·+·0.011*"比较"·+·0.010*"套票"·+·0.009*"沙滩"·+·0.009*"买"·+·0.008*"好玩"
0.034*"方便"·+·0.022*"取票"·+·0.020*"不错"·+·0.015*"排队"·+·0.014*"便宜"·+·0.009*"套票"·+·0.008*"玩"·+·0.008*"沙滩"·+·0.008*"窗口"·+·0.007*"不用"
0.047*"方便"·+·0.028*"取票"·+·0.023*"不错"·+·0.017*"排队"·+·0.015*"便宜"·+·0.012*"套票"·+·0.012*"沙滩"·+·0.009*"挺"·+·0.009*"不用"·+·0.009*"买"
0.025*"方便"·+·0.018*"不错"·+·0.011*"取票"·+·0.011*"便宜"·+·0.008*"套票"·+·0.008*"排队"·+·0.008*"不用"·+·0.007*"沙滩"·+·0.006*"比较"·+·0.006*"旅游"

../corpus/景区/A13_pos_cut.txt正面主题分析
0.045*"温泉"·+·0.023*"服务"·+·0.021*"前台"·+·0.018*"不错"·+·0.017*"环境"·+·0.015*"服务态度"·+·0.012*"池子"·+·0.012*"服务员"·+·0.010*"值得"·+·
0.009*"干净"
0.037*"温泉"·+·0.020*"环境"·+·0.019*"服务"·+·0.019*"前台"·+·0.017*"不错"·+·0.017*"服务态度"·+·0.011*"服务员"·+·0.011*"池子"·+·0.010*"热情"·+·
0.010*"态度"
0.047*"温泉"·+·0.034*"服务"·+·0.024*"前台"·+·0.021*"不错"·+·0.019*"服务态度"·+·0.019*"环境"·+·0.011*"池子"·+·0.010*"值得"·+·0.010*"热情"·+·
0.010*"服务员"
0.062*"温泉"·+·0.036*"服务"·+·0.034*"不错"·+·0.030*"环境"·+·0.029*"前台"·+·0.018*"服务态度"·+·0.012*"下次"·+·0.012*"池子"·+·0.012*"挺"·+·0.012*"喜欢"
0.050*"温泉"·+·0.030*"服务"·+·0.029*"前台"·+·0.023*"不错"·+·0.019*"环境"·+·0.015*"服务态度"·+·0.013*"池子"·+·0.013*"热情"·+·0.011*"服务员"·+·
0.010*"推荐"

../corpus/景区/A12_neg_cut.txt负面主题分析
0.017*"里面"·+·0.012*"鱼"·+·0.012*"玩"·+·0.011*"门票"·+·0.010*"好多"·+·0.009*"有点"·+·0.009*"感觉"·+·0.009*"南粤"·+·0.008*"买"·+·0.008*"票价"
0.017*"买"·+·0.013*"门票"·+·0.013*"鱼"·+·0.013*"玩"·+·0.011*"好玩"·+·0.010*"感觉"·+·0.010*"有点"·+·0.010*"没什么"·+·0.010*"里面"·+·0.008*"门口"
0.018*"买"·+·0.016*"鱼"·+·0.015*"里面"·+·0.013*"玩"·+·0.012*"门票"·+·0.011*"有点"·+·0.010*"好玩"·+·0.010*"感觉"·+·0.010*"苑"·+·0.009*"票价"
0.014*"里面"·+·0.012*"门票"·+·0.011*"买"·+·0.011*"鱼"·+·0.010*"感觉"·+·0.010*"门口"·+·0.010*"好玩"·+·0.010*"玩"·+·0.008*"票"·+·0.008*"有点"
0.017*"买"·+·0.014*"有点"·+·0.011*"好多"·+·0.011*"玩"·+·0.010*"鱼"·+·0.010*"门票"·+·0.010*"票价"·+·0.009*"里面"·+·0.009*"门口"·+·0.009*"适合"

../corpus/景区/A28_neg_cut.txt负面主题分析
0.017*"排队"·+·0.014*"贵"·+·0.013*"取票"·+·0.009*"买"·+·0.009*"海鲜"·+·0.008*"钱"·+·0.008*"吃饭"·+·0.008*"方便"·+·0.008*"坐船"·+·0.008*"还要"
0.016*"贵"·+·0.012*"排队"·+·0.011*"买"·+·0.010*"吃饭"·+·0.010*"二楼"·+·0.010*"窗口"·+·0.010*"取票"·+·0.009*"码头"·+·0.009*"方便"·+·0.008*"比较"
0.015*"贵"·+·0.013*"取票"·+·0.011*"窗口"·+·0.011*"排队"·+·0.011*"钱"·+·0.009*"吃饭"·+·0.009*"买"·+·0.009*"方便"·+·0.009*"海边"·+·0.008*"海鲜"
0.014*"贵"·+·0.012*"排队"·+·0.011*"取票"·+·0.010*"吃饭"·+·0.009*"海边"·+·0.009*"景色"·+·0.009*"窗口"·+·0.008*"二楼"·+·0.008*"海鲜"·+·0.008*"方便"
0.016*"贵"·+·0.015*"排队"·+·0.012*"取票"·+·0.009*"景色"·+·0.009*"码头"·+·0.009*"窗口"·+·0.009*"海鲜"·+·0.008*"海边"·+·0.008*"吃饭"·+·0.008*"二楼"

../corpus/景区/A13_neg_cut.txt负面主题分析
0.019*"温泉"·+·0.011*"感觉"·+·0.010*"服务"·+·0.008*"里面"·+·0.007*"前台"·+·0.007*"水果"·+·0.007*"比较"·+·0.007*"池"·+·0.007*"服务态度"·+·0.007*"丶"
0.023*"温泉"·+·0.009*"感觉"·+·0.008*"里面"·+·0.008*"丶"·+·0.008*"服务"·+·0.007*"前台"·+·0.006*"水果"·+·0.006*"比较"·+·0.006*"服务员"·+·0.005*"收费"
0.038*"温泉"·+·0.017*"感觉"·+·0.014*"服务"·+·0.009*"前台"·+·0.009*"池"·+·0.009*"里面"·+·0.008*"丶"·+·0.008*"比较"·+·0.008*"水果"·+·0.008*"池子"
0.025*"温泉"·+·0.013*"服务"·+·0.012*"感觉"·+·0.010*"里面"·+·0.009*"池"·+·0.009*"水果"·+·0.009*"比较"·+·0.008*"丶"·+·0.007*"特色"·+·0.007*"池子"
0.021*"温泉"·+·0.012*"服务"·+·0.011*"感觉"·+·0.011*"池"·+·0.008*"比较"·+·0.008*"里面"·+·0.008*"丶"·+·0.008*"水果"·+·0.007*"水质"·+·0.007*"前台"
```

（2）中层次景区（A19、A35、A42）

```
../corpus/景区/A19_pos_cut.txt正面主题分析
0.022*"不错"·+·0.015*"湖"·+·0.011*"方便"·+·0.011*"风景"·+·0.009*"景色"·+·0.009*"绿"·+·0.008*"值得"·+·0.008*"游船"·+·0.008*"水"·+·0.008*"取票"
0.023*"不错"·+·0.017*"湖"·+·0.013*"值得"·+·0.013*"风景"·+·0.012*"景色"·+·0.012*"方便"·+·0.011*"湖水"·+·0.009*"取票"·+·0.009*"岛"·+·0.007*"游船"
0.023*"不错"·+·0.022*"湖"·+·0.012*"方便"·+·0.011*"值得"·+·0.010*"取票"·+·0.010*"湖水"·+·0.010*"景色"·+·0.009*"风景"·+·0.009*"景点"·+·0.009*"绿"
0.021*"不错"·+·0.017*"湖"·+·0.012*"风景"·+·0.009*"方便"·+·0.009*"值得"·+·0.009*"湖水"·+·0.008*"景点"·+·0.008*"取票"·+·0.007*"景色"·+·0.007*"空气"
0.021*"不错"·+·0.015*"湖"·+·0.013*"值得"·+·0.012*"风景"·+·0.011*"湖水"·+·0.010*"方便"·+·0.008*"绿"·+·0.008*"景色"·+·0.008*"游船"·+·0.007*"水"

../corpus/景区/A35_pos_cut.txt正面主题分析
0.039*"不错"·+·0.026*"瀑布"·+·0.020*"景色"·+·0.016*"值得"·+·0.011*"景点"·+·0.009*"壮观"·+·0.009*"空气"·+·0.009*"景区"·+·0.009*"瀑布群"·+·0.008*"方便"
0.026*"瀑布"·+·0.020*"不错"·+·0.016*"景色"·+·0.011*"值得"·+·0.010*"壮观"·+·0.010*"景区"·+·0.008*"瀑布群"·+·0.007*"方便"·+·0.007*"景点"·+·0.006*"门票"
0.040*"瀑布"·+·0.034*"不错"·+·0.022*"景色"·+·0.018*"值得"·+·0.013*"景区"·+·0.012*"景点"·+·0.010*"门票"·+·0.010*"壮观"·+·0.008*"方便"·+·0.008*"挺"
0.030*"瀑布"·+·0.029*"不错"·+·0.021*"值得"·+·0.019*"景色"·+·0.015*"景区"·+·0.010*"景点"·+·0.009*"壮观"·+·0.009*"方便"·+·0.009*"环境"·+·0.008*"挺"
0.033*"不错"·+·0.024*"瀑布"·+·0.019*"景色"·+·0.014*"壮观"·+·0.012*"值得"·+·0.011*"方便"·+·0.010*"门票"·+·0.010*"景区"·+·0.008*"景点"·+·0.008*"空气"

../corpus/景区/A42_pos_cut.txt正面主题分析
0.044*"不错"·+·0.037*"温泉"·+·0.024*"环境"·+·0.015*"挺"·+·0.012*"玩"·+·0.011*"服务"·+·0.010*"舒服"·+·0.009*"比较"·+·0.009*"方便"·+·0.009*"下次"
0.047*"不错"·+·0.036*"温泉"·+·0.024*"环境"·+·0.014*"挺"·+·0.011*"舒服"·+·0.011*"值得"·+·0.010*"方便"·+·0.009*"下次"·+·0.009*"泡"·+·0.009*"服务"
0.039*"环境"·+·0.035*"温泉"·+·0.033*"不错"·+·0.013*"舒服"·+·0.012*"服务"·+·0.012*"池"·+·0.011*"下次"·+·0.011*"地方"·+·0.010*"值得"·+·0.010*"挺"
0.034*"温泉"·+·0.032*"不错"·+·0.028*"环境"·+·0.015*"挺"·+·0.010*"值得"·+·0.010*"舒服"·+·0.009*"方便"·+·0.008*"服务"·+·0.008*"开心"·+·0.008*"下次"
0.024*"不错"·+·0.019*"温泉"·+·0.018*"环境"·+·0.009*"舒服"·+·0.008*"下次"·+·0.007*"挺"·+·0.007*"地方"·+·0.006*"服务"·+·0.006*"玩"·+·0.006*"好好"


../corpus/景区/A19_neg_cut.txt负面主题分析
0.013*"景点"·+·0.012*"没什么"·+·0.012*"岛"·+·0.011*"景色"·+·0.011*"湖"·+·0.009*"三个"·+·0.008*"湖水"·+·0.008*"有点"·+·0.008*"不错"·+·0.008*"感觉"
0.011*"没什么"·+·0.010*"景色"·+·0.009*"岛"·+·0.009*"湖"·+·0.009*"景点"·+·0.007*"游船"·+·0.007*"有点"·+·0.007*"水"·+·0.007*"不错"·+·0.007*"风景"
0.016*"没什么"·+·0.014*"景点"·+·0.012*"湖"·+·0.012*"景色"·+·0.011*"不错"·+·0.010*"岛"·+·0.009*"三个"·+·0.009*"感觉"·+·0.009*"不值"·+·0.009*"风景"
0.013*"没什么"·+·0.012*"岛"·+·0.011*"不错"·+·0.010*"景色"·+·0.010*"三个"·+·0.010*"湖"·+·0.009*"感觉"·+·0.009*"坐船"·+·0.009*"游船"·+·0.008*"开发"
0.011*"岛"·+·0.010*"景点"·+·0.010*"湖"·+·0.010*"没什么"·+·0.009*"景色"·+·0.009*"风景"·+·0.008*"不错"·+·0.007*"湖水"·+·0.007*"门票"·+·0.007*"有点"

../corpus/景区/A35_neg_cut.txt负面主题分析
0.020*"门票"·+·0.013*"景点"·+·0.012*"景区"·+·0.011*"瀑布"·+·0.011*"不值"·+·0.009*"元"·+·0.009*"小时"·+·0.009*"贵"·+·0.008*"取票"·+·0.007*"景色"
0.029*"门票"·+·0.019*"景区"·+·0.018*"景点"·+·0.017*"元"·+·0.015*"瀑布"·+·0.011*"贵"·+·0.010*"取票"·+·0.010*"便宜"·+·0.008*"太贵"·+·0.008*"价格"
0.022*"门票"·+·0.017*"景区"·+·0.017*"瀑布"·+·0.016*"景点"·+·0.015*"元"·+·0.010*"不值"·+·0.010*"贵"·+·0.009*"取票"·+·0.009*"小时"·+·0.008*"便宜"
0.022*"景区"·+·0.021*"门票"·+·0.019*"元"·+·0.016*"瀑布"·+·0.016*"景点"·+·0.011*"不值"·+·0.010*"贵"·+·0.009*"取票"·+·0.009*"小时"·+·0.009*"洗手间"
0.020*"门票"·+·0.018*"景点"·+·0.017*"景区"·+·0.016*"瀑布"·+·0.013*"元"·+·0.011*"取票"·+·0.009*"不值"·+·0.008*"价格"·+·0.008*"停车费"·+·0.008*"贵"

../corpus/景区/A42_neg_cut.txt负面主题分析
0.018*"温泉"·+·0.016*"服务"·+·0.011*"地方"·+·0.009*"不错"·+·0.008*"服务态度"·+·0.007*"前台"·+·0.007*"池"·+·0.006*"提供"·+·0.006*"设施"·+·0.006*"元"
0.018*"温泉"·+·0.016*"服务"·+·0.014*"不错"·+·0.013*"地方"·+·0.010*"设施"·+·0.009*"池"·+·0.008*"服务态度"·+·0.008*"提供"·+·0.008*"没开"·+·0.008*"房间"
0.017*"服务"·+·0.011*"温泉"·+·0.009*"不错"·+·0.008*"地方"·+·0.008*"设施"·+·0.007*"池"·+·0.006*"服务态度"·+·0.006*"没开"·+·0.006*"前台"·+·0.005*"办理"
0.011*"服务"·+·0.009*"温泉"·+·0.008*"地方"·+·0.007*"不错"·+·0.007*"服务态度"·+·0.007*"提供"·+·0.006*"池"·+·0.006*"感觉"·+·0.005*"知道"·+·0.005*"前台"
0.014*"温泉"·+·0.012*"服务"·+·0.011*"地方"·+·0.010*"不错"·+·0.009*"池"·+·0.009*"服务态度"·+·0.008*"设施"·+·0.007*"房间"·+·0.007*"没开"·+·0.006*"拖鞋"
```

（3）低层次景区（A18、A34、A30）

../corpus/景区/A18_pos_cut.txt正面主题分析
0.014*"不错"+0.011*"珠海"+0.010*"门票"+0.009*"免费"+0.009*"北京"+0.009*"里面"+0.007*"值得"+0.007*"表演"+0.007*"景色"+0.006*"建筑"
0.019*"不错"+0.015*"免费"+0.014*"里面"+0.012*"门票"+0.011*"值得"+0.010*"北京"+0.009*"地方"+0.009*"珠海"+0.008*"景点"+0.008*"景色"
0.016*"免费"+0.014*"不错"+0.012*"门票"+0.011*"珠海"+0.011*"景点"+0.010*"里面"+0.010*"北京"+0.009*"值得"+0.008*"建筑"+0.007*"景色"
0.014*"门票"+0.012*"不错"+0.011*"珠海"+0.010*"值得"+0.008*"免费"+0.008*"里面"+0.008*"表演"+0.007*"北京"+0.007*"建筑"+0.006*"地方"
0.020*"不错"+0.014*"免费"+0.014*"珠海"+0.012*"门票"+0.009*"值得"+0.008*"景点"+0.007*"里面"+0.007*"景色"+0.007*"公园"+0.007*"挺"

../corpus/景区/A34_pos_cut.txt正面主题分析
0.033*"玩"+0.027*"不错"+0.021*"项目"+0.020*"好玩"+0.011*"比较"+0.011*"方便"+0.009*"小孩"+0.009*"值得"+0.008*"地方"+0.008*"开心"
0.033*"不错"+0.033*"玩"+0.019*"好玩"+0.018*"项目"+0.014*"开心"+0.012*"刺激"+0.009*"方便"+0.009*"比较"+0.008*"小孩"+0.007*"孩子"
0.040*"玩"+0.033*"不错"+0.022*"好玩"+0.017*"项目"+0.011*"方便"+0.010*"开心"+0.009*"值得"+0.009*"比较"+0.009*"地方"+0.008*"设施"
0.029*"玩"+0.021*"不错"+0.018*"好玩"+0.015*"项目"+0.011*"方便"+0.009*"开心"+0.009*"设施"+0.008*"比较"+0.008*"孩子"+0.008*"里面"
0.024*"玩"+0.017*"不错"+0.016*"好玩"+0.015*"项目"+0.010*"方便"+0.009*"比较"+0.008*"开心"+0.007*"孩子"+0.006*"地方"+0.006*"里面"

../corpus/景区/A30_pos_cut.txt正面主题分析
0.027*"动物"+0.020*"动物园"+0.012*"不错"+0.011*"挺"+0.011*"门票"+0.009*"方便"+0.009*"里面"+0.009*"值得"+0.007*"地方"+0.007*"喜欢"
0.042*"动物"+0.020*"动物园"+0.013*"不错"+0.010*"挺"+0.010*"门票"+0.009*"方便"+0.009*"里面"+0.009*"值得"+0.008*"喜欢"+0.008*"海洋馆"
0.055*"动物"+0.018*"动物园"+0.016*"不错"+0.011*"里面"+0.011*"玩"+0.011*"小朋友"+0.010*"地方"+0.009*"方便"+0.009*"海洋馆"+
0.009*"门票"
0.036*"动物"+0.026*"动物园"+0.014*"挺"+0.014*"不错"+0.012*"里面"+0.010*"小朋友"+0.009*"门票"+0.009*"值得"+0.009*"海洋馆"+
0.007*"方便"
0.045*"动物"+0.024*"动物园"+0.015*"不错"+0.012*"门票"+0.012*"挺"+0.011*"方便"+0.011*"海洋馆"+0.011*"小朋友"+0.010*"里面"+0.008*"玩"

../corpus/景区/A18_neg_cut.txt负面主题分析
0.010*"公园"+0.009*"有点"+0.008*"感觉"+0.008*"门票"+0.008*"建筑"+0.008*"免费"+0.007*"北京"+0.007*"景点"+0.007*"珠海"+0.007*"表演"
0.014*"免费"+0.013*"公园"+0.010*"表演"+0.010*"感觉"+0.009*"门票"+0.009*"景点"+0.009*"珠海"+0.009*"建筑"+0.009*"玩"+0.008*"里面"
0.013*"免费"+0.012*"公园"+0.011*"门票"+0.011*"景点"+0.011*"感觉"+0.011*"表演"+0.010*"珠海"+0.009*"建筑"+0.009*"地方"+0.008*"有点"
0.011*"感觉"+0.010*"公园"+0.010*"建筑"+0.010*"免费"+0.009*"门票"+0.009*"珠海"+0.008*"景点"+0.008*"表演"+0.008*"没什么"+0.007*"地方"
0.010*"免费"+0.010*"门票"+0.009*"景点"+0.009*"感觉"+0.008*"建筑"+0.007*"玩"+0.007*"北京"+0.007*"地方"+0.007*"公园"+0.007*"表演"

../corpus/景区/A34_neg_cut.txt负面主题分析
0.028*"玩"+0.018*"项目"+0.013*"门票"+0.008*"设施"+0.008*"贵"+0.007*"比较"+0.007*"东西"+0.007*"里面"+0.007*"太"+0.007*"孩子"
0.032*"玩"+0.021*"项目"+0.009*"太"+0.009*"东西"+0.009*"门票"+0.008*"孩子"+0.007*"景区"+0.007*"比较"+0.007*"设施"+0.007*"里面"
0.026*"玩"+0.023*"项目"+0.012*"门票"+0.010*"比较"+0.009*"里面"+0.009*"东西"+0.008*"设施"+0.008*"贵"+0.008*"孩子"+0.007*"太"
0.021*"项目"+0.019*"玩"+0.009*"门票"+0.007*"设施"+0.007*"比较"+0.007*"孩子"+0.007*"里面"+0.007*"东西"+0.006*"元"+0.006*"贵"
0.036*"玩"+0.034*"项目"+0.011*"设施"+0.010*"东西"+0.010*"比较"+0.009*"里面"+0.009*"门票"+0.008*"地方"+0.008*"贵"+0.007*"太"

../corpus/景区/A30_neg_cut.txt负面主题分析
0.043*"动物"+0.031*"动物园"+0.012*"门票"+0.011*"海洋馆"+0.010*"里面"+0.009*"小孩"+0.009*"地方"+0.009*"景区"+0.009*"设施"+
0.008*"小时"
0.051*"动物"+0.046*"动物园"+0.009*"海洋馆"+0.009*"地方"+0.009*"设施"+0.008*"门票"+0.007*"小时"+0.007*"玩"+0.007*"贵"+0.007*"小孩"
0.039*"动物"+0.029*"动物园"+0.018*"门票"+0.009*"海洋馆"+0.008*"设施"+0.008*"里面"+0.008*"玩"+0.007*"地方"+0.007*"比较"+
0.007*"小孩子"
0.026*"动物"+0.022*"动物园"+0.008*"里面"+0.007*"门票"+0.006*"海洋馆"+0.006*"景区"+0.006*"比较"+0.005*"看到"+0.005*"地方"+
0.005*"设施"
0.042*"动物"+0.025*"动物园"+0.012*"门票"+0.008*"看到"+0.008*"比较"+0.008*"里面"+0.008*"海洋馆"+0.007*"小时"+0.007*"设施"+0.006*"元"

2 酒店正负向评论主题识别结果

（1）高层次酒店（H03、H16、H20）

../corpus/酒店/H03_pos_cut.txt正面主题分析
0.035*"服务"+0.031*"酒店"+0.025*"不错"+0.022*"前台"+0.012*"环境"+0.010*"房间"+0.010*"大堂"+0.009*"服务态度"+0.008*"热情"+
0.007*"入住"
0.047*"酒店"+0.045*"服务"+0.030*"不错"+0.022*"前台"+0.015*"环境"+0.014*"大堂"+0.014*"服务态度"+0.012*"房间"+0.009*"早餐"+
0.009*"经理"
0.026*"酒店"+0.025*"服务"+0.018*"不错"+0.016*"前台"+0.010*"房间"+0.009*"经理"+0.008*"服务态度"+0.008*"大堂"+0.007*"热情"+
0.006*"环境"
0.054*"酒店"+0.050*"服务"+0.023*"不错"+0.020*"前台"+0.014*"房间"+0.013*"服务态度"+0.013*"环境"+0.012*"热情"+0.011*"大堂"+
0.009*"经理"
0.057*"服务"+0.056*"酒店"+0.021*"不错"+0.018*"环境"+0.017*"房间"+0.016*"前台"+0.014*"热情"+0.013*"大堂"+0.012*"经理"+
0.011*"服务态度"

../corpus/酒店/H16_pos_cut.txt正面主题分析
0.061*"服务"+0.060*"不错"+0.051*"前台"+0.027*"小姐姐"+0.026*"酒店"+0.026*"房间"+0.024*"推荐"+0.024*"服务态度"+0.024*"环境"+
0.021*"入住"
0.057*"不错"+0.047*"前台"+0.039*"服务"+0.023*"酒店"+0.022*"服务态度"+0.021*"房间"+0.020*"小姐姐"+0.020*"环境"+0.017*"干净"+
0.015*"推荐"
0.093*"不错"+0.062*"前台"+0.050*"服务"+0.039*"酒店"+0.029*"服务态度"+0.024*"小姐姐"+0.023*"房间"+0.023*"推荐"+0.022*"环境"+
0.022*"入住"
0.067*"前台"+0.050*"不错"+0.044*"服务"+0.031*"环境"+0.031*"酒店"+0.029*"房间"+0.025*"推荐"+0.025*"服务态度"+0.019*"小姐姐"+
0.017*"方便"
0.077*"不错"+0.060*"前台"+0.055*"服务"+0.036*"酒店"+0.031*"房间"+0.029*"服务态度"+0.027*"推荐"+0.025*"环境"+0.021*"热情"+
0.021*"入住"

../corpus/酒店/H20_pos_cut.txt正面主题分析
0.029*"干净"+0.027*"酒店"+0.025*"服务"+0.023*"前台"+0.023*"房间"+0.022*"不错"+0.020*"方便"+0.019*"环境"+0.018*"设施"+0.013*"卫生"
0.042*"干净"+0.038*"房间"+0.031*"服务"+0.029*"不错"+0.027*"酒店"+0.026*"前台"+0.021*"方便"+0.020*"设施"+0.018*"环境"+0.014*"特别"
0.040*"干净"+0.037*"房间"+0.035*"服务"+0.030*"方便"+0.027*"酒店"+0.025*"前台"+0.024*"不错"+0.023*"环境"+0.022*"推荐"+0.019*"设施"
0.047*"干净"+0.030*"房间"+0.029*"服务"+0.028*"酒店"+0.027*"不错"+0.025*"环境"+0.025*"前台"+0.021*"方便"+0.020*"推荐"+0.016*"舒服"
0.040*"干净"+0.038*"酒店"+0.037*"服务"+0.034*"房间"+0.028*"环境"+0.026*"前台"+0.025*"不错"+0.024*"方便"+0.020*"设施"+0.017*"卫生"

../corpus/酒店/H03_neg_cut.txt负面主题分析
0.026*"酒店"+0.019*"服务"+0.013*"入住"+0.011*"房间"+0.010*"早餐"+0.010*"不错"+0.008*"前台"+0.006*"退房"+0.005*"经理"+0.005*"催"
0.034*"酒店"+0.030*"服务"+0.014*"房间"+0.014*"入住"+0.012*"前台"+0.009*"不错"+0.008*"催"+0.008*"服务态度"+0.008*"退房"+0.008*"五星"
0.027*"服务"+0.026*"酒店"+0.015*"入住"+0.011*"不错"+0.009*"前台"+0.008*"房间"+0.007*"退房"+0.007*"催"+0.007*"五星"+0.007*"冯家铭"
0.048*"酒店"+0.030*"服务"+0.017*"入住"+0.014*"房间"+0.013*"前台"+0.011*"不错"+0.009*"催"+0.009*"退房"+0.008*"早餐"+0.008*"一下"
0.028*"酒店"+0.024*"服务"+0.012*"入住"+0.010*"房间"+0.009*"早餐"+0.008*"不错"+0.007*"前台"+0.007*"住"+0.006*"一下"+0.006*"服务态度"

../corpus/酒店/H16_neg_cut.txt负面主题分析
0.025*"好好"+0.022*"酒店"+0.021*"房间"+0.015*"服务"+0.015*"住"+0.014*"位置"+0.014*"前台"+0.012*"体验"+0.011*"隔音"+0.009*"态度"
0.029*"好好"+0.027*"位置"+0.024*"酒店"+0.022*"房间"+0.015*"体验"+0.015*"前台"+0.012*"服务"+0.012*"隔音"+0.011*"比较"+0.010*"睡觉"
0.021*"酒店"+0.020*"好好"+0.018*"房间"+0.015*"服务"+0.014*"前台"+0.013*"体验"+0.011*"位置"+0.011*"隔音"+0.009*"住"+0.009*"早餐"
0.023*"房间"+0.021*"酒店"+0.016*"位置"+0.014*"好好"+0.013*"前台"+0.013*"服务"+0.013*"体验"+0.011*"住"+0.010*"隔音"+0.009*"真的"
0.023*"好好"+0.022*"房间"+0.016*"位置"+0.014*"酒店"+0.014*"服务"+0.013*"体验"+0.013*"隔音"+0.012*"前台"+0.012*"住"+0.010*"态度"

../corpus/酒店/H20_neg_cut.txt负面主题分析
0.033*"房间"+0.028*"酒店"+0.021*"住"+0.017*"打电话"+0.016*"前台"+0.011*"挺"+0.010*"卫生"+0.010*"现在"+0.010*"知道"+0.010*"电视"
0.044*"房间"+0.034*"酒店"+0.021*"住"+0.016*"前台"+0.016*"打电话"+0.016*"电视机"+0.014*"早餐"+0.014*"网络"+0.012*"差评"+0.012*"挺"
0.033*"房间"+0.027*"酒店"+0.025*"住"+0.019*"前台"+0.015*"差评"+0.014*"电视"+0.013*"打电话"+0.012*"挺"+0.012*"早餐"+0.011*"环境"
0.036*"房间"+0.031*"酒店"+0.025*"前台"+0.024*"住"+0.015*"打电话"+0.012*"挺"+0.012*"电视"+0.011*"网络"+0.011*"差评"+0.010*"凌晨"
0.032*"房间"+0.022*"酒店"+0.020*"住"+0.018*"前台"+0.016*"打电话"+0.013*"网络"+0.011*"挺"+0.011*"差评"+0.010*"电视"+0.010*"早餐"

（2）中层次酒店（H07、H11、H41）

../corpus/酒店/H07_pos_cut.txt正面主题分析
0.035*"酒店"+0.021*"服务"+0.017*"房间"+0.013*"沙面"+0.013*"不错"+0.013*"环境"+0.010*"住"+0.009*"广州"+0.009*"入住"+0.008*"位置"
0.050*"酒店"+0.018*"服务"+0.014*"房间"+0.014*"环境"+0.013*"不错"+0.011*"沙面"+0.011*"入住"+0.010*"住"+0.009*"早茶"+0.009*"江景"
0.037*"酒店"+0.021*"房间"+0.018*"不错"+0.018*"服务"+0.015*"沙面"+0.011*"环境"+0.010*"广州"+0.010*"入住"+0.008*"位置"+0.008*"住"
0.047*"酒店"+0.023*"房间"+0.021*"不错"+0.016*"服务"+0.015*"环境"+0.010*"广州"+0.010*"江景"+0.010*"沙面"+0.008*"住"+0.008*"方便"
0.047*"酒店"+0.016*"房间"+0.016*"不错"+0.015*"服务"+0.012*"环境"+0.011*"广州"+0.010*"沙面"+0.010*"住"+0.009*"位置"+0.008*"早餐"

../corpus/酒店/H11_pos_cut.txt正面主题分析
0.039*"酒店"+0.032*"服务"+0.020*"不错"+0.013*"入住"+0.012*"前台"+0.011*"房间"+0.011*"泳池"+0.011*"孩子"+0.011*"特别"+0.011*"早餐"
0.046*"酒店"+0.033*"服务"+0.015*"早餐"+0.014*"房间"+0.014*"不错"+0.013*"前台"+0.012*"环境"+0.012*"入住"+0.011*"下次"+0.011*"特别"
0.054*"酒店"+0.029*"服务"+0.015*"不错"+0.014*"房间"+0.014*"孩子"+0.014*"早餐"+0.012*"前台"+0.011*"沙滩"+0.010*"入住"+0.010*"环境"
0.036*"酒店"+0.034*"服务"+0.018*"不错"+0.015*"前台"+0.014*"早餐"+0.013*"房间"+0.011*"孩子"+0.010*"入住"+0.010*"沙滩"+0.009*"环境"
0.032*"酒店"+0.031*"服务"+0.018*"不错"+0.011*"房间"+0.011*"前台"+0.010*"入住"+0.009*"沙滩"+0.009*"早餐"+0.008*"环境"+0.008*"特别"

../corpus/酒店/H41_pos_cut.txt正面主题分析
0.039*"酒店"+0.024*"不错"+0.021*"温泉"+0.021*"房间"+0.016*"服务"+0.011*"环境"+0.010*"早餐"+0.008*"挺"+0.007*"干净"+0.007*"感觉"
0.035*"酒店"+0.025*"不错"+0.023*"温泉"+0.020*"房间"+0.016*"服务"+0.014*"环境"+0.010*"早餐"+0.009*"挺"+0.007*"方便"+0.007*"前台"
0.030*"酒店"+0.020*"不错"+0.018*"温泉"+0.016*"环境"+0.014*"房间"+0.014*"服务"+0.010*"早餐"+0.009*"入住"+0.008*"感觉"+0.007*"特别"
0.035*"酒店"+0.020*"温泉"+0.019*"不错"+0.016*"房间"+0.015*"环境"+0.014*"服务"+0.011*"早餐"+0.009*"挺"+0.008*"感觉"+0.007*"舒适"
0.029*"酒店"+0.024*"温泉"+0.021*"不错"+0.016*"房间"+0.015*"服务"+0.012*"早餐"+0.012*"环境"+0.008*"干净"+0.008*"方便"+0.008*"舒服"

../corpus/酒店/H07_neg_cut.txt负面主题分析
0.021*"酒店"+0.017*"房间"+0.014*"早餐"+0.013*"住"+0.011*"服务"+0.011*"房"+0.008*"江景"+0.007*"环境"+0.006*"不错"+0.005*"感觉"
0.028*"房间"+0.023*"早餐"+0.022*"酒店"+0.012*"住"+0.010*"服务"+0.010*"房"+0.008*"不错"+0.008*"环境"+0.008*"感觉"+0.007*"江景"
0.023*"房间"+0.021*"酒店"+0.021*"早餐"+0.015*"住"+0.012*"房"+0.010*"服务"+0.007*"不错"+0.006*"环境"+0.006*"江景"+0.006*"应该"
0.021*"早餐"+0.020*"酒店"+0.016*"房间"+0.011*"服务"+0.010*"房"+0.009*"住"+0.007*"设施"+0.006*"应该"+0.006*"感觉"+0.006*"不错"
0.019*"酒店"+0.015*"房间"+0.014*"早餐"+0.012*"住"+0.011*"服务"+0.010*"房"+0.008*"江景"+0.007*"环境"+0.007*"不错"+0.006*"设施"

../corpus/酒店/H11_neg_cut.txt负面主题分析
0.025*"酒店"+0.015*"服务"+0.010*"孩子"+0.008*"不错"+0.008*"早餐"+0.008*"房间"+0.008*"电梯"+0.006*"入住"+0.006*"前台"+0.006*"洲际"
0.031*"酒店"+0.015*"服务"+0.011*"孩子"+0.011*"房间"+0.010*"电梯"+0.010*"早餐"+0.009*"前台"+0.008*"不错"+0.007*"洲际"+0.007*"入住"
0.029*"酒店"+0.018*"服务"+0.013*"孩子"+0.011*"早餐"+0.011*"房间"+0.009*"电梯"+0.009*"不错"+0.009*"前台"+0.007*"意识"+0.006*"入住"
0.022*"酒店"+0.015*"服务"+0.013*"房间"+0.011*"早餐"+0.010*"不错"+0.009*"电梯"+0.008*"孩子"+0.008*"洲际"+0.007*"入住"+0.007*"前台"
0.030*"酒店"+0.020*"服务"+0.011*"早餐"+0.010*"不错"+0.010*"电梯"+0.009*"孩子"+0.009*"房间"+0.008*"前台"+0.006*"洲际"+0.006*"被子"

../corpus/酒店/H41_neg_cut.txt负面主题分析
0.013*"温泉"+0.011*"酒店"+0.010*"早餐"+0.009*"房间"+0.009*"前台"+0.008*"住"+0.008*"入住"+0.007*"服务"+0.005*"不错"+0.005*"别墅"
0.015*"温泉"+0.014*"前台"+0.014*"早餐"+0.013*"酒店"+0.011*"房间"+0.011*"服务"+0.010*"住"+0.009*"入住"+0.007*"电话"+0.006*"想"
0.019*"温泉"+0.016*"酒店"+0.014*"服务"+0.013*"房间"+0.011*"住"+0.011*"早餐"+0.008*"前台"+0.008*"入住"+0.007*"找"+0.006*"想"
0.014*"酒店"+0.014*"温泉"+0.013*"房间"+0.013*"服务"+0.012*"早餐"+0.010*"前台"+0.009*"入住"+0.008*"住"+0.006*"找"+0.006*"不错"
0.017*"温泉"+0.014*"前台"+0.012*"服务"+0.012*"房间"+0.012*"酒店"+0.012*"早餐"+0.008*"住"+0.007*"找"+0.007*"入住"+0.006*"想"

（3）低层次酒店（H27、H36、H38）

../corpus/酒店/H27_pos_cut.txt正面主题分析
0.039*"酒店"+0.022*"不错"+0.022*"房间"+0.017*"环境"+0.014*"早餐"+0.014*"适合"+0.013*"小镇"+0.012*"服务"+0.011*"玩"+0.010*"欧洲"
0.022*"房间"+0.021*"酒店"+0.018*"不错"+0.015*"小镇"+0.014*"环境"+0.014*"早餐"+0.010*"适合"+0.009*"服务"+0.009*"欧洲"+0.007*"玩"
0.035*"酒店"+0.024*"不错"+0.022*"房间"+0.019*"服务"+0.016*"环境"+0.015*"小镇"+0.010*"欧洲"+0.009*"适合"+0.009*"升级"+0.009*"早餐"
0.022*"不错"+0.020*"房间"+0.018*"酒店"+0.012*"服务"+0.011*"小镇"+0.011*"早餐"+0.010*"环境"+0.010*"适合"+0.008*"欧洲"+0.008*"升级"
0.024*"酒店"+0.022*"房间"+0.017*"不错"+0.017*"服务"+0.016*"小镇"+0.014*"环境"+0.011*"早餐"+0.010*"欧洲"+0.009*"适合"+0.009*"玩"

../corpus/酒店/H36_pos_cut.txt正面主题分析
0.032*"酒店"+0.029*"不错"+0.020*"早餐"+0.018*"房间"+0.013*"位置"+0.012*"环境"+0.011*"方便"+0.010*"七星岩"+0.008*"比较"+0.008*"服务"
0.048*"酒店"+0.027*"不错"+0.019*"房间"+0.018*"位置"+0.017*"早餐"+0.017*"环境"+0.016*"方便"+0.009*"比较"+0.009*"有点"+0.009*"住"
0.036*"酒店"+0.023*"不错"+0.019*"早餐"+0.017*"位置"+0.017*"环境"+0.016*"方便"+0.015*"房间"+0.011*"住"+0.010*"比较"+0.010*"服务"
0.040*"酒店"+0.027*"不错"+0.019*"位置"+0.018*"早餐"+0.017*"环境"+0.016*"房间"+0.016*"方便"+0.010*"七星岩"+0.010*"比较"+0.009*"设施"
0.042*"酒店"+0.026*"不错"+0.022*"早餐"+0.022*"位置"+0.018*"房间"+0.015*"方便"+0.011*"服务"+0.010*"环境"+0.009*"比较"+0.009*"好好"

../corpus/酒店/H38_pos_cut.txt正面主题分析
0.050*"酒店"+0.033*"不错"+0.020*"早餐"+0.015*"泳池"+0.015*"环境"+0.015*"设施"+0.014*"沙滩"+0.012*"有点"+0.011*"适合"+0.011*"房间"
0.045*"酒店"+0.025*"不错"+0.020*"环境"+0.020*"早餐"+0.018*"沙滩"+0.015*"设施"+0.012*"比较"+0.011*"适合"+0.010*"房间"+0.009*"泳池"
0.051*"酒店"+0.023*"不错"+0.022*"早餐"+0.018*"环境"+0.016*"设施"+0.013*"沙滩"+0.011*"比较"+0.010*"服务"+0.010*"泳池"+0.009*"挺"
0.056*"酒店"+0.029*"不错"+0.026*"早餐"+0.014*"房间"+0.013*"环境"+0.012*"比较"+0.012*"有点"+0.011*"海滩"+0.011*"设施"+0.011*"沙滩"
0.041*"酒店"+0.028*"不错"+0.024*"早餐"+0.015*"沙滩"+0.014*"环境"+0.013*"设施"+0.013*"服务"+0.011*"泳池"+0.010*"房间"+0.010*"比较"

../corpus/酒店/H27_neg_cut.txt负面主题分析
0.023*"酒店"+0.016*"房间"+0.012*"早餐"+0.010*"设施"+0.010*"小镇"+0.009*"服务"+0.008*"入住"+0.006*"比较"+0.006*"有点"+0.006*"吃"
0.018*"酒店"+0.015*"房间"+0.014*"早餐"+0.010*"设施"+0.009*"服务"+0.008*"感觉"+0.008*"比较"+0.008*"入住"+0.007*"吃"+0.007*"小镇"
0.020*"酒店"+0.016*"房间"+0.010*"早餐"+0.009*"服务"+0.009*"设施"+0.009*"小镇"+0.008*"玩"+0.007*"比较"+0.007*"入住"+0.006*"吃"
0.033*"酒店"+0.021*"房间"+0.011*"早餐"+0.010*"设施"+0.009*"感觉"+0.009*"有点"+0.009*"入住"+0.007*"小镇"+0.007*"味道"+0.007*"吃"
0.029*"酒店"+0.023*"房间"+0.014*"早餐"+0.009*"感觉"+0.009*"设施"+0.008*"小镇"+0.008*"比较"+0.008*"服务"+0.008*"入住"+0.008*"吃"

../corpus/酒店/H36_neg_cut.txt负面主题分析
0.043*"酒店"+0.016*"早餐"+0.011*"服务"+0.010*"房间"+0.010*"前台"+0.009*"有点"+0.008*"设施"+0.007*"入住"+0.007*"位置"+0.007*"不错"
0.038*"酒店"+0.013*"早餐"+0.011*"服务"+0.011*"前台"+0.009*"房间"+0.007*"住"+0.007*"比较"+0.007*"餐厅"+0.007*"入住"+0.007*"设施"
0.036*"酒店"+0.017*"早餐"+0.013*"服务"+0.013*"房间"+0.010*"前台"+0.010*"有点"+0.007*"住"+0.007*"设施"+0.006*"餐厅"+0.006*"不错"
0.026*"酒店"+0.012*"早餐"+0.009*"住"+0.008*"服务"+0.008*"设施"+0.007*"前台"+0.007*"房间"+0.007*"餐厅"+0.007*"位置"+0.007*"入住"
0.028*"酒店"+0.022*"早餐"+0.012*"房间"+0.011*"服务"+0.010*"前台"+0.009*"餐厅"+0.009*"住"+0.008*"位置"+0.007*"有点"+0.006*"设施"

../corpus/酒店/H38_neg_cut.txt负面主题分析
0.034*"酒店"+0.023*"房间"+0.020*"设施"+0.013*"沙滩"+0.013*"入住"+0.012*"早餐"+0.009*"服务"+0.008*"环境"+0.007*"住"+0.006*"比较"
0.030*"酒店"+0.016*"设施"+0.015*"房间"+0.013*"早餐"+0.009*"沙滩"+0.008*"服务"+0.008*"入住"+0.007*"老旧"+0.007*"真的"+0.007*"比较"
0.038*"酒店"+0.020*"房间"+0.010*"早餐"+0.010*"服务"+0.010*"沙滩"+0.010*"设施"+0.009*"入住"+0.008*"比较"+0.007*"住"+0.006*"不错"
0.028*"酒店"+0.019*"房间"+0.018*"设施"+0.013*"入住"+0.011*"早餐"+0.009*"服务"+0.009*"沙滩"+0.006*"陈旧"+0.006*"真的"+0.006*"比较"
0.035*"酒店"+0.020*"房间"+0.015*"设施"+0.015*"早餐"+0.011*"入住"+0.011*"沙滩"+0.009*"服务"+0.007*"住"+0.007*"比较"+0.006*"环境"