电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

《综合课程设计》报告

2019-2020-2

BACHELOR THESIS

论文 题目 基于在线旅游网站评论数据的乡村文化旅游多元需求分析

学 院 公共管理学院

专 业 信息管理与信息系统

时 间 2020/7/9

小组成员及分工表

学号	姓名	主要工作
2017120101007	贾坤泽	数据爬取，主题建模，情感分析，论文撰写
2017120101003	潘春辉	查询相关文献，数据清洗，加油鼓劲
2017120101027	李泓岩	数据爬取，描述性统计，论文撰写，分词
2017120101028	谭旭航	数据爬取，论文撰写，共现分析

摘　要

近年来，乡村文旅逐渐成为现在的一个焦点。本文以乡村文旅为研究对象，以携程网评论数据为基础，根据所挑选的旅游景点抓取 2014 年至今的部分评论数据。从不同的角度分析旅客对乡村旅游景区的不同需求，并对游客的情感进行研究，以及结合价格相关关键词寻找影响旅客对价格看法的因素。本研究立足于乡村旅游的运行实际情况，使用携程大数据在乡村旅游领域进行分析，是对利用大数据对乡村旅游景点进行定向规划的尝试，大数据的分析方法和结果能够让乡村旅游景区和管理部门更好的理解旅客的需求，了解乡村旅游的发展趋势，从而更好的制定相应的政策或进行景区进一步的管理。

关键词：乡村文旅；多元需求；在线评论； LDA 主题建模

ABSTRACT

In recent years, the village culture brigade has gradually become a focus of the present. Based on Ctrip's review data, this paper captures some of the comment data from 2014 to the present based on the selected tourist attractions. From different angles to analyze the different needs of tourists in rural tourist attractions, and to study the feelings of tourists, and combined with price-related keywords to find factors affecting the passengers' view of prices. Based on the actual situation of rural tourism, the use of Ctrip big data in the field of rural tourism analysis, is the use of big data to carry out targeted planning of rural tourist attractions, big data analysis methods and results can make rural tourist attractions and management departments better understand the needs of tourists, understand the development trend of rural tourism, so as to better formulate the corresponding policies or further management of scenic spots.

Keywords: Country Tourism； Multiple needs; Online reviews; LDA Theme Modeling

目录

引言

乡村旅游是以旅游度假为宗旨，以村庄野外为空间，以人文无干扰、生态无破坏、以游居和野行为特色的村野旅游形式。近几年来，随着我国居民收入不断增长，人们的旅游需求也在日益攀升，十一黄金周，五一黄金周的出行人数不断上升，这也造成了热门景区人满为患的情况出现，严重影响了人们的出行体验。而作为一种新的旅行模式，乡村旅行具有很多特色，包括未被开发、纯天然无污染的环境，在城市中难以触及的生活方式，以及各种乡土气息浓厚的文化项目。同时，乡村旅游也可以促进社会资源和文明成果在城乡之间的共享以及财富重新分配的实现，并为地区间经济发展差异和城乡差别的逐步缩小、产业结构优化等做出很大贡献，推动欠发达、开发不足的乡村地区经济、社会、环境和文化的可持续发展，对于加快实现社会主义新农村建设及城乡统筹发展具有重要意义。由此，乡村旅游迅速发展，2015 年中央一号文件提出，要积极开发农业多种功能，挖掘乡村生态休闲、旅游观光、文化教育价值，2016 年中央一号文件强调，大力发展休闲农业和乡村旅游。强化规划引导，采取以奖代补、先建后补、财政贴息、设立产业投资基金等方式扶持休闲农业与乡村旅游业发展。2017 年《国家发改委“十三五”时期文化旅游提升工程实施方案》的颁布，标志着我国 对“旅游是载体，文化是灵魂。” 这个核心理念的认可。2018 年中央一号文件中，乡村振兴战略被高度重视。将文化脉络融入乡村肌理，通过文化提升乡村旅游品位，能够有效解决乡村旅游同质化、传统文化特色不足等问题又可以进一步接近乡村振兴的目标。

近年来，基于互联网产生的大数据已经成为重要的数据来源，这些大数据产生于搜索引擎、社交媒体等不同的互联网平台，数据量大、数据类型丰富、生成速度快且具有较高的价值.互联网大数据在学术界研究和业界应用中均发挥显著的作用，互联网大数据为经济、金融、能源、旅游等行业的精准预测提供新的数据支持，并能显著提高预测精度在旅游活动中，游客在景点查询、门票及酒店预订、餐饮选择、交通、服务评价等各环节中产生的数据具有重要的研究价值。已有研究表明，这些互联网大数据能够表征游客特征及偏好，其中，携程网拥有最为广阔的受众面，较高的使用率和景区占有率，基于携程网的旅游大数据研究较为普遍。

国内外在旅游需求的研究方面较为成熟，旅游需求影响因素的衡量指标多以传统的统计调研数据为主、以计量经济模型和人工智能模型为主要建模工具，数据频度多为月度和年度近年来，随着大数据的发展，研究者开始尝试从大数据中

提取能够影响旅游需求的因素，提高旅游需求预测精度。在线旅游数据也是一种重要的大数据来源，携程网作为国内较大的在线旅游机构’所产生的游客数据类型丰富，包括旅游产品的订单数据以及游客对于旅游产品。及服务的评论数据等，因此，携程网的游客数据在学术研究中具有重要的应用价值。

已有研究通过利用携程网中用户对于酒店的评论数据，研究了互联网的评论对于酒店销量的影响，通过利用游客在携程网中订购旅游目的地的机票、酒店和门票等数据可以分析　游客旅游行为和偏好．通过采集游客对旅游产品、酒店服务等的评价，能够反映游客对旅游目的地的态度和满意度，从而更好地进行旅游目的地的管理。

第一章 文献综述

目前，国内有不少学者对挖掘网络在线评论开展了研究，其中包含了政治，商业，旅游，舆情等多个方面。孙宗锋[1]运用大数据分析方法对青岛市市长信箱里公民诉求情感进行研究。发现公民诉求表达过程中公民积极和消极情感比例相当，且波动性较小。在不同领域，公民诉求情感态度差异化明显；金吉琼利用在线评论探究了电子烟市场的消费热点[2]，通过建立主题模型利用分类评论主题对消费者的消费行为进行剖析。其结果可为烟草企业设计和优化电子烟产品提供支持；陈雪琳利用共词分析探究了我国双创领域的研究热点[3]以期对双创参与主体及政策制定主体提供一定的参考，进一步推动我国创新创业的高质量发展。在旅游板块，张尧政，张若愚等研究利用在线评论数据来分析用户对迪士尼，江西五A级景区等景点的情感倾向[4-5]，既能很好的为游客提供决策信息，又能为旅游经营者提供改进信息；程翠琼,徐健构建的模型能够有效地反映旅游地的游客情感随时间变化的波动,进而为旅游管理者、潜在旅游者信息获取提供新的信息参考渠道[6]；刘思叶,田原等基于机器学习方法,开展游客微博主题情感分析方法比较研究[7]：针对饮食、娱乐、购物、景观、交通和住宿6个旅游主题，基于机器学习方法，开展游客微博主题情感分析方法比较研究；王少兵,吴升构建了能够反映游客关注度和情感分析的方法，为游客决策提供参考依据[8]；李圆圆利用 LDA 主题发现模型对兵马俑在线评论进行分析[9]，结果表明，携程网和途牛网的游客更关注历史文化，去哪儿网的游客更关注门票服务；王新宇将基于词典和机器学习相结合，对旅游网络评论进行情感分析研究[10]，提出一种基于旅游情感词典和机器学习相结合的方法；倪海燕也找出了使用情感分析探究旅客对莫干山民宿态度的方法[11] 。

第二章 研究方法

2.1 jieba 分词

2.1.1 jieba 分词概述

分词是本文在线评论数据处理的第一步，也是比较关键的一步。所谓分词，就是将由字符序列构成的句子按照一定的规则重新组合成词的集合，中文分词就是指将句子中汉字序列切分成词集合。相对于英文而言，中文分词要复杂得多。中文分词时如何界定 “词”、如何消除歧义、如何识别未登录词？这个是我们面临的问题。中文分词算法主要基于字符串匹配算法、基于理解算法和基于统计分词算法。 Jieba 分词主要提供了一下三种模式：

精确模式：根据相应的算法，可以将句子或者文本内的句子精准切分开来，在文本分析中得到广泛的应用。

全模式：将本文或者句子内可以构成成词的词语全部快速扫描出来，但是在中间切分产生的词语可能会产生歧义的问题。

搜索引擎模式：在精确模式产生的长词的基础上，将长词再次切分，这样做可以有效提高召回率，在搜索引擎分词中得到广泛的应用。

2.1.2 jieba 分词的基本原理

（1）jieba 分词采用了基于 Trie 树结构的算法，jieba 分词利用该算法高效实现了词图扫描，并且利用词图扫描将得到句子中汉字所有的成词可能，并且将这些所有成词可能的 情况构成有向无环图（DAG），为下一步打下基础。通过分词，jieba 分词的源码可以发现，jieba 分词本身就包含了一个有 2 万多词条的词典。基于 Trie 结构的扫描就是将这些词条放到 Trie 的树结构之中，一旦扫描的词条中和这些放在 Trie 结构中的词条有着相同的前缀，那么就实现了快速查找。

（2）jieba 分词中最大概率路径的实现采用了动态规划查找的方法，动态规划查找以找出根据词频的最大切分组合。分析 jieba 分词的源码可以发现 jieba 分词不仅仅将字典生成 Trie 树，同时，将把每个词的出现次数转换为了频率。

（3）对于在 jieba 分词自带的词典中未出现的词，jieba 分词采用了 Viterbi 算法，并且将用于汉字成词的 HMM 模型应用其中。

2.2 LDA 主题模型

LDA 是基于三层贝叶斯概率的主题生成模型，用来识别大规模文档的潜在主题信息，由词、主题和文档三层结构。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“这样一个关系得到。其三者关系如下：

$$P(词语|文档)=\sum p(词语|主题)\times p(主题|文档)$$

文档主题分布 P(主题 | 文档)指不同主题在同一个文档中所占比重。文档词语分布 P(词语|主题)指每个主题中不同词语出现的概率。文档词语分布 P(词语|文档)指每个文档中不同词语出现的概率。则 LDA 生成过程为：

(1)对每一篇文章，从主题分布中抽取一个主题；

(2)从上述被抽到的主题所对应的单词分布中抽取一个单词；

(3)重复上述的过程直至遍历文档中的每一个单词。

由图 1 可知，要先按照一定的先验概率$p(d_i)$选择一篇文档，然后通过狄利克雷分布 α 中取出生成文档的主题分布 θ i，之后取出文档d_i的第j个词的主题z_{ij}，从狄利克雷分布 β 中取样生成主题对应的词语分布 φ，最后从词语多项式分布中采样最终生成词语w_{ij}，这就是 LDA 生成文档的过程。

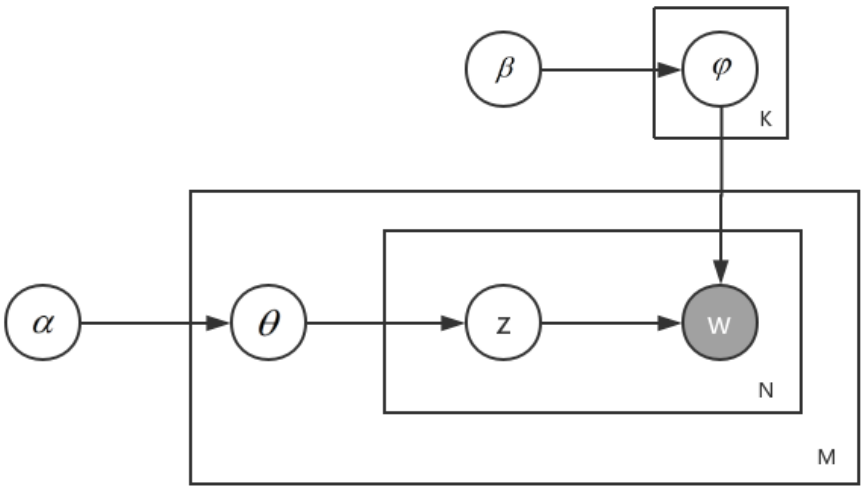


图 1 LDA 工作原理

2.3 情感分析

情感分析又称意见挖掘、倾向性分析等。简单而言,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。本文使用自然语言处理库中的 snowNLP 库，对 LDA 建模后得到的每个主题下包含评论的情感分析,得到游客各个主题的情感态度。

snowNLP 是自然语言处理工具。可以输出文本的情感倾向。情感倾向以 0 到 1 的数据表示。0 表示完全消极的情感。1 表示完全积极的情感。0.5 表示中性的情感。越接近于 1,，则表明文本的情感越积极；接近于 0，则表明文本的情感越消极。如：“很好，很有意思，非常值得一去。”最终输出的值为 0.9976，表示为情感积极的文本；“我的心情很糟糕，不建议去” 最终输出的值为 0.1919，表示为情感消极的文本。

2.4 共词分析

共词分析法最早于 20 世纪 70 年代由法国文献计量学家提出，它是将各种信息载体中的共现信息定量化的分析方法，以揭示信息的内容关联和特征项所隐含的寓意。一般而言，同时包含两个关键词的文献数量越多，则表示这两个关键词之间的联系越密切、“距离”越近。利用因子分析、聚类分析、多维尺度分析等多元统计的方法，根据“距离”的远近对关键词进行分类，从而形成文献集的研究关注热点和内部结构。

第三章 实证研究

本文的实证研究的思路如图 2 所示。具体包括五个环节：数据采集、数据清洗、LDA 主题建模、情感分析以及共词分析。一般情况下，每一条旅客评论都是旅客对旅游地的相关评价，但是这些数据往往不止包含一个方面，可能是多维度的描述，且这种描述往往也带有不同的感情要素，例如某条数据中可能同时存在对价格过高的不满和对景区风景的赞美，再对这些文本构成的非结构化数据进行处理和情感分析的的时候，首先要处理成结构化数据，然后抽取主题以及相应的评论观点。通过情感分析方法以及可视化工具来评价出旅客对于各个主题的情感状况，在涉及价格相关时，则可以通过筛选出与价格相关的评论，并进行共现词分析，通过确定经常和价格同时出现的名词所对应的主题来反映出导致用户价格敏感的要素。

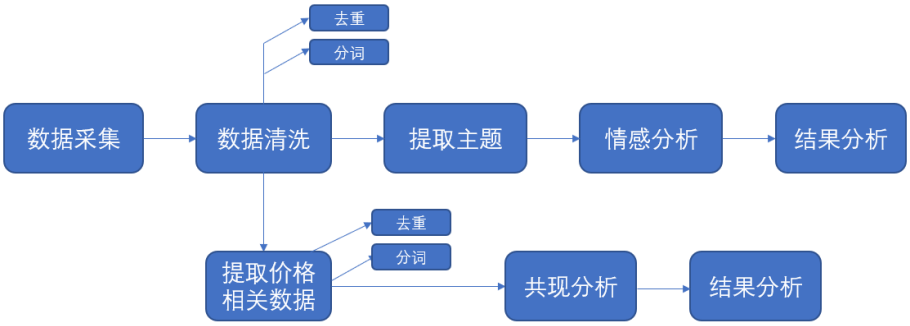


图 2 研究思路

3.1 数据采集

本文利用网络爬虫工具对携程网旅游平台中的用户评论进行抓取，首先确定所抓取的乡村旅游景点名单。《生态体育》提出了十种乡村农业旅游的主要类型，本文基于该分类并整合收集到的在线评论数据，最终将乡村景区分为三种类型，分别为民俗古镇、乡村庄园以及现代科技农园。民俗古镇指具有当地特色文化以及悠久历史的乡村古镇。乡村庄园指集生态农业、乡村旅游、养生度假、休闲体验等功能为一体，实现经济价值、社会价值和生态价值的新型农业旅游产业综合体。现代科技农园指立足农业优势产业，探索现代农业发展新路径，突出科技引领和示范带动，引进科技化和智能化项目，发展高科技农业的高科技农园。在携程网站中以这三部分为关键词进行搜索，标定评论数较多的景点作为我们的研究对象。每条数据包括以下属性，用户姓名、性别、关注数、粉丝数、评论文本以

及有用数。最终获取的数据包括全国 48 个乡村景点，10049 条在线评论，形成

表 1 初始评论数据集

	name	gender	关注	粉丝	time	rating	text	useful
1	_CFT01****3482347	female	0	0	2016/10/11	width:100%	就是景色，既然是庄园，怎么没有园林呢，常家的后花园也是风景很美，	有用（45）
2	ROMAO	female	0	0	2015/12/15	width:100%	车一小时直达，庄园很大很美丽，多部影视剧的取景地，亭台水榭，鸟语	有用（26）
3	keava	male	14	103	2014/6/1	width:100%	乔家大院、王家大院、常家庄园相比，感觉常家庄园最好，王家大院次之，乔家大院相比之下根本不值一提，	有用（14）
4	西风品味	female	0	9	2017/2/7	width:40%	讲解器到处进院，景区实际复完开放可参观的部分不多，像瓜式的讲解	有用（13）
5	_WeCh****38190	male	0	0	2016/9/12	width:100%	有思想，后花园也不错，景色宜人，流连忘返啊，是许多电影电视剧的取	有用（10）
6	ZY张张	female	2	0	2017/2/17	width:100%	大的就是常家庄园，文化底蕴深刻，一年四季不同景，给人印象极好，去	有用（9）
7	_WeCh****50188	male	0	0	2016/10/2	width:100%	无大，值得，值得一去，导游小姑娘讲解的也不错，门口常家家宴价位也	有用（8）
8	泰江孤影1965	male	7	63	2015/10/4	width:80%	积超级大，院落里有很多木雕，砖雕的艺术品，最值得去的地方，还是	有用（8）
9	M24****7846	female	0	0	2018/2/26	width:100%	常家的家庭教育，体味晋商的深厚文化底蕴，建筑的砖雕，木雕，彩绘	有用（7）
10	_WeCh****80463	male	0	0	2017/3/20	width:80%	得大家去看一看，可想当时主人的财富！可以想而知当时的晋商，在社	有用（7）
11	e33****50	female	0	0	2018/2/19	width:100%	比刚去过的乔家大院还要大很多啊，除了古建筑，还有这么大的园林，可	有用（6）
12	1361130****	male	19	352	2018/1/13	width:100%	宅院4公顷，园林8公顷，附属房屋3公顷，庄墙12公里，形成一山，一园	有用（6）
13	海棠花	female	0	0	2017/10/7	width:100%	特色：而前面的院落，则是典型的北方建筑，深宅大院，庄严肃穆，最	有用（6）
14	M44****335	female	0	0	2016/8/31	width:100%	悠久的历史和灿烂的文化，建筑宏伟，景色美丽，集建筑美学与自然风	有用（6）
15	M81****86	female	0	0	2015/10/30	width:100%	第一家，气势恢弘叹为观止！精美的砖雕木雕令人赞叹，私家花园绿草茵	有用（6）

初始评论数据集。

表 2 景区信息

乡村庄园	蓝调庄园	张裕爱斐堡国际酒庄	常家庄园	伏尔加庄园	金卉庄园	…
现代科技农园	中国科技农业展示园	太仓现代农业园	北大荒现代农业园	兰陵国家农业公园	都江堰市高科技农业旅游园	…
民俗古镇	西江千户苗寨	凤凰古城	平遥古镇	束河古镇	川民俗园	…

3.2 数据清洗

为进行后续的主题挖掘及情感分析，首先对评论文本进行去噪处理，去除文本空格、链接，并将繁体转换为简体、小写转换为大写。剔除评论数少于 100 的景点数据，一是因为其不具有代表性，说服能力比较差；二是因为其可能对于整体的游客满意度分析带来较大的偏差。其次，剔除重复评论数据以及评论文本字段小于 5 的数据。最终得到 46 个景区共计 9882 条在线评论数据。为提高分词效果和准确度，根据评论文本和初次分词结果构建用户词典，并整合停用词库以及自建的文本停用词表，构建较为全面的停用词表。加载用户词典，采用 jieba 进行分词，同时调用停用词表，去除停用词。进行词频统计，并去除掉无实意的高频词和低频词后重新进行分词。最终的分词结果代入 LDA 中进行主题建模。

3.3 LDA 主题建模

本文从两个维度进行 LDA 主题建模，分别为景区类型维度和满意度维度。对于景区类型维度，从满意度维度讲，本文将在线评论数据分为两类，好评类数据和差评类数据。其中好评类数据是指用户对景点评分大于三星的数据。差评类数据是指用户对景点评分是小于三星的数据。之后对两个维度各类数据分别进行 LDA 主题建模。本文采用困惑度和一致度来确定 LDA 主题建模的最优主题个数，

基于最优主题个数来进行 LDA 主题建模最终得到每类评论数据的主题信息。确定之后要人工归并主题，进而分别统计每个主题下性别比例。以及高影响力游客（本文设定粉丝数目大于 500 的评论者为高影响力的评论者）关注的主题分布。从中分析不同的游客是否存在不同的需求。

图 3 LDA 建模过程

3.4 情感分析

本文对各个主题下的评论数据进行情感分析，得到游客对于各个主题下的情感分布，从而分析不同的游客在不同主题下的情感需求。

3.5 共现词分析

在本次研究中，我们选择 “价格，钱，值得，亏”等关键词进行筛选，最终取得了 1789 条评论数据，生成共现矩阵后，纵轴保留价格相关关键词，横轴统计其他词出现的频率，删除共现总数小于 200 的词汇得到共现矩阵。

表 1 共现矩阵

z'x	不错	方便	小时	感觉	景色	建筑	时间	庄园	博物馆	需要	特别	游客	历史	特色	城堡	看到	演出	网上
门票	1014	546	545	520	417	468	333	400	276	353	278	253	237	186	146	144	124	146
值得	340	329	261	184	260	167	189	148	100	99	105	77	71	0	77	43	28	118
价格	514	144	246	216	289	282	106	269	0	86	43	80	60	63	60	71	58	116
性价比	283	122	252	264	118	71	75	0	61	0	43	0	27	0	0	0	0	88
便宜	236	41	71	111	106	106	30	35	41	41	19	0	0	76	19	0	0	70
收费	105	38	0	47	0	0	0	24	38	24	0	0	0	0	0	0	20	0
订票	19	0	27	59	0	0	59	0	59	0	19	0	0	0	19	0	0	0
购票	32	44	32	0	0	32	0	0	44	44	0	32	76	0	0	0	0	0
票价	34	62	34	34	0	28	0	28	34	28	0	0	0	0	0	0	0	0
不值	47	0	0	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0
块钱	30	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	30
性价比	0	0	0	28	28	0	28	28	0	0	0	28	0	0	0	0	28	0
不贵	27	0	0	0	0	0	0	27	0	27	0	27	0	0	0	0	0	0
购买	0	0	0	25	0	0	0	0	25	0	0	0	0	25	0	25	0	0
值得	25	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2706	1326	1493	1535	1218	1154	850	959	678	702	507	497	471	350	321	283	258	568
	旅行体验	旅行体验	旅行体验	旅行体验	自然风景	人文景观	旅行体验	人文景观	人文景观	旅行体验	特色内容	人文景观	人文景观	特色内容	人文景观	旅行体验	特色内容	旅行体验

第四章 研究结果

4.1 描述性统计

首先，本文对收集的数据进行统计描述性统计，从性别的分布来看，这些评论者的性别接近 1:1。

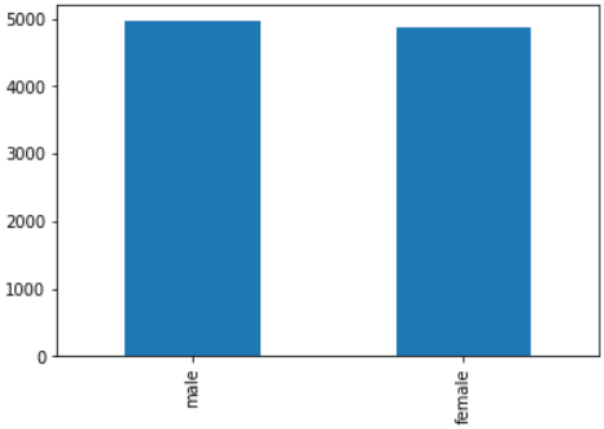


图 4 性别分布

对收集的评论时间按照年份与月份做进一步的分析，从年份来看本文收集的大部分数据来源于 14 年至今的数据，其中 17 年的数据最多，说明本文收集的数据具有时效性的特点。对于月份来说，评论数目随着月份的变化处于波动状态，其中在 10 月份数据最多。

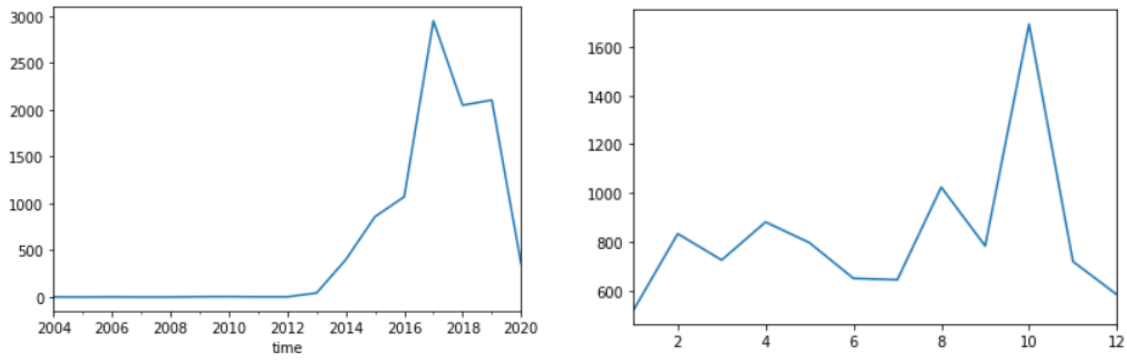
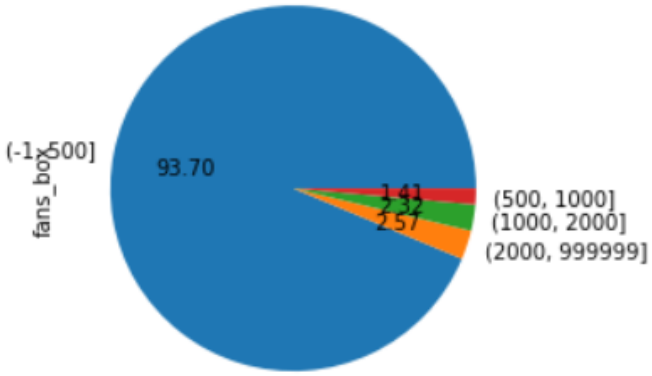


图 5 评论时间分布（左：按年份统计；右：按月份统计）

对于收集的评论者的粉丝数目进行分析，发现大多数的评论者的粉丝数目少于 500，粉丝数大于 500 小于等于 1000 占据总评论者的 1.41%，粉丝数大于 1000 小于等于 2000 占据总评论者的 2.32%，粉丝数大于 2000 占据总评论者的 2.57%。拥有较高粉丝的评论者是影响力比较高的一类人群，具有较高研究价值。因此本



文将着重分析粉丝数大于 500 这一类群体。

对评分进行分析，有 62.66%的评论者对景点评分为 5 星，22.32%的评论者对景点评分为 4 星，6.74%的评论者对景点评分为 3 星。由此可见，大部分评论者对于乡村文旅持满意态度，仅有少量游客对于乡村文旅持不满意的态度。

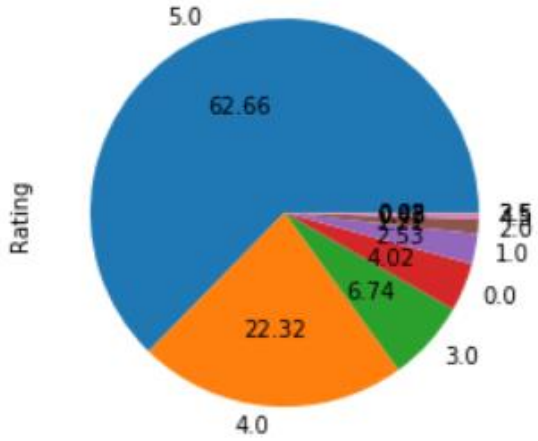


图 7 评分分布

4.2 LDA 主题建模结果

本文从两个维度进行 LDA 主题建模，分别为景区类型维度和满意度维度，得到最终各类评论数据的主题信息如下:

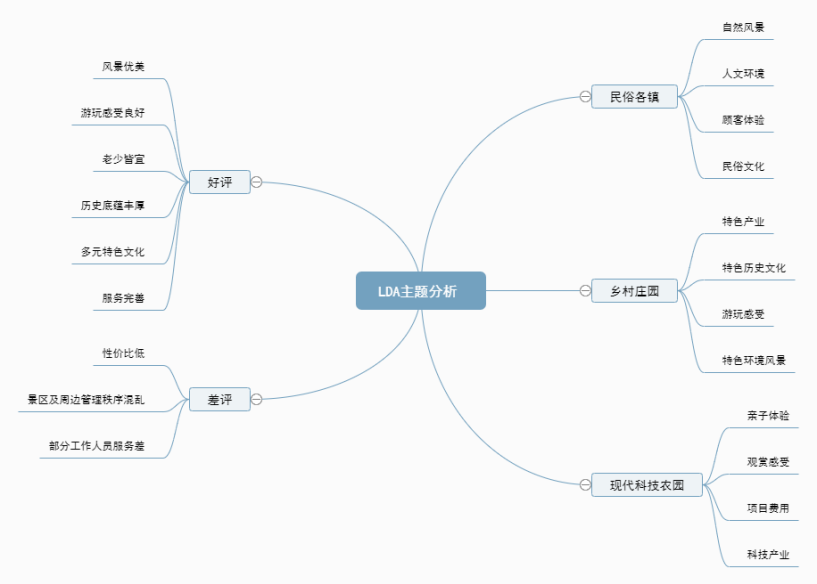


图 8 LDA 主题分析

接下来对每个主题下面的评论数据进行性别分析，得到每个主题下面的性别分布。以此来研究不同性别在不同主题下是否有特定的需求。同时我们选择具有较高影响力的评论者（即粉丝数目较多的评论者）来研究此类群体在不同主题下是否有特定的需求。

具体的研究结果如下所示：

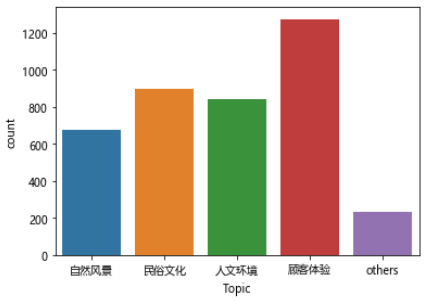


图 9 民俗古镇类主题分布

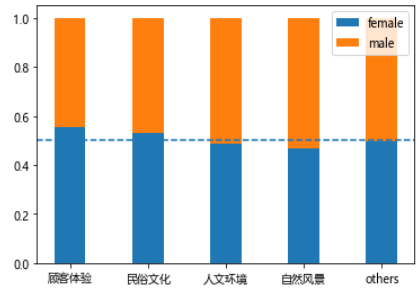


图 10 民俗古镇类性别分布

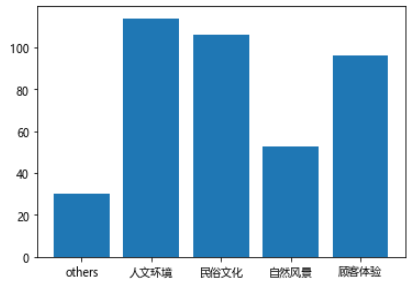


图 11 民俗古镇类高影响力游客关注的主题分布

(1). 本文对民俗古镇的评论文本内容进行主题建模，得到了四个主题，分别为自然风景，民俗文化，人文环境以及顾客体验。自然风景主题指的是景区自然风光一类词语的集合，主要包括“风景”“景色”“湖面”等词语；民俗文化主题指的是景区所蕴含的历史文化及特色的民族文化一类词语的集合，主要包括“文化”“民族”“历史等词语；人文环境主题指的是景区文物古迹、现代艺术及科学活动场以及地区和民族的特殊人文景观一类词语的集合，主要包括“建筑”“石板”“砖雕”等词语。顾客体验主题指的是游客参观景点产生的最直接的主观感受的一类词语的集合，包括“性价比”“值得”“心情”等词语。Others 类主题中

的词语过于杂乱，无法判断出其归属于某类特定的主题，因此本文不考虑此类主题。对各类主题的统计分析可以看出游客更关心自己的主观感受，对良好的游玩体验需求更高，而对自然风景，民俗文化，人文环境主题的需求差别不大。对性别比例统计发现，男性游客对优美风景的需求更高，而女性游客对良好体验的需求更高。此外，高影响力的游客对优美风景的需求相对较低

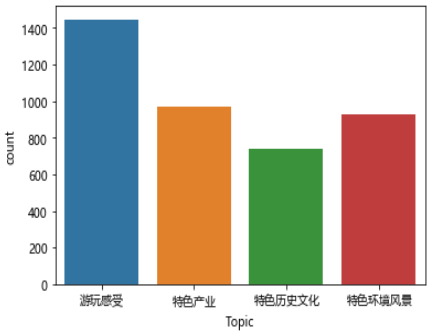


图 12 乡村庄园类主题分布

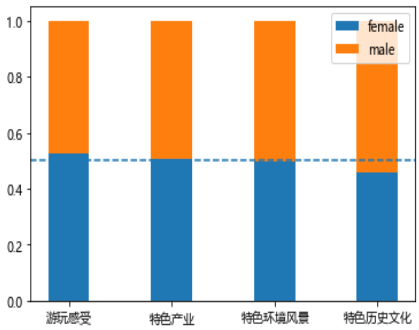


图 13 乡村庄园类性别分布

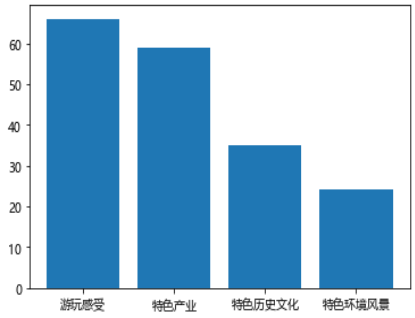


图 14 乡村庄园高影响力游客关注的主题分布

(2). 对乡村庄园的评论文本内容进行主题建模，也得到了四个主题，分别为游玩感受，特色产业，特色历史文化以及特色环境风景。游玩感受主题指的是游客游玩景点产生的最直接的主观感受一类词语的集合，主要包括“游玩”“性价比”等词语；特色产业主题指的是景区所包括的主导经营产业一类词语的集合，主要包括“葡萄酒”“酒庄”等词语；特色历史文化主题指的是景区丰富的历史底蕴和特色的庄园文化一类词语的集合，主要包括“历史”“文化”“时期”等词语。特色环境风景主题指的是庄园独特景色一类词语的集合，包括“气派”“宅院”等词语。对各类主题的统计分析可以看出游客更关心自己的游玩感受，对此需求更高，而对特色产业，特色历史文化，特色环境风景主题的需求差别不大。对性别比例统计发现，男性游客对特色历史文化的需求更高，男女游客对于其他类主题的需求大致相当。此外，高影响力的游客对庄园特色的环境风景的需求相对较低

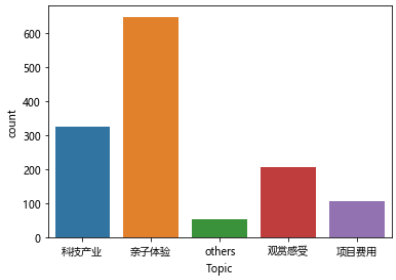


图 15 现代科技农园类主题分布

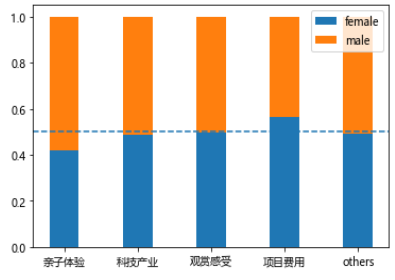


图 16 现代科技农园类性别分布

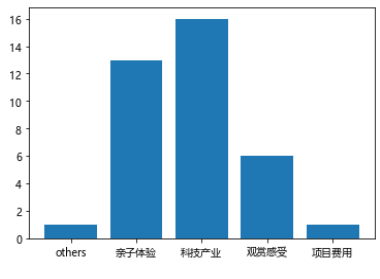


图 17 现代科技农园高影响力游客关注的主题分布

(3). 对现代科技农园的评论文本内容进行主题建模，得到了四个主题，分别为亲子体验，科技产业，观赏感受以及项目费用。亲子体验主题指的是父母同子女之间互动体验一类词语的集合主要包括“孩子”“亲子”“小朋友”等词语；科技产

业主题指的是景区科技化和智能化农业项目一类词语的集合，主要包括“科技”“品种”“花卉“等词语；观赏感受主题指的是的是游客对科技农园的产业及周边环境产生的最直接的主观感受一类词语的集合，主要包括“菊花”“环境”“园区”等词语。项目费用主题指的是科技农园门票及其他项目费用的一类词语的集合，包括“门票”“套票”“项目”等词语。Others 类主题中的词语过于杂乱，无法判断出其归属于某类特定的主题，因此本文不考虑此类主题。对各类主题的统计分析可以看出游客更关心自己与家人之间的互动体验，对良好的亲子游玩体验需求更高，而对性价比的需求最低。从性别比例统计发现，男性游客对亲子体验的需求更高，而女性游客对项目费用的需求更高，男女游客对于其他类主题的需求大致相当。此外，高影响力的游客对科技产业的需求最高同时对性价比的需求最低。

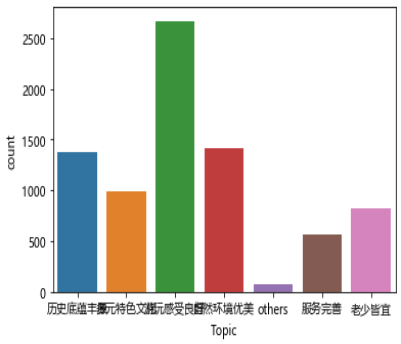


图 18 好评类主题分布

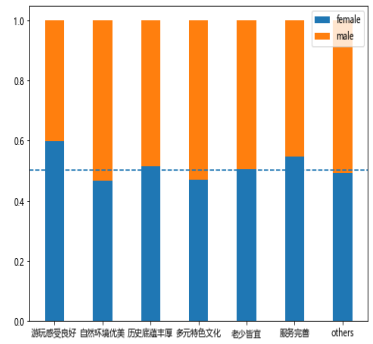


图 19 好评类主题性别分布

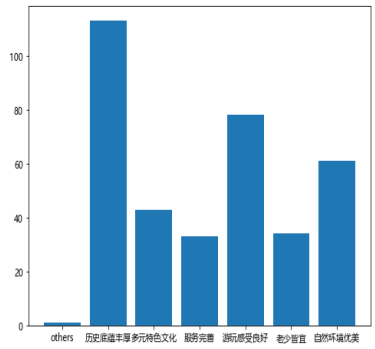


图 20 好评类高影响力游客关注的主题分布

(4).对好评类的评论文本内容进行主题建模，得到了六个主题，分别为自然环境优美、游玩感受良好、老少皆宜、历史底蕴丰厚、多元特色文化以及服务完善。自然环境优美主题指的是景区良好的自然风光一类词语的集合，主要包括“很漂亮”“景色”“观光”等词语；老少皆宜主题指的是适宜各个年龄段人群一类词语的集合，主要包括“老人”“孩子”“家人“等词语；历史底蕴丰厚主题指的是景区丰富的历史底蕴和特色的乡村文化一类词语的集合，主要包括“古代”“建筑”“时期”等词语。多元特色文化主题指的是景区所蕴含的特色民族文化和庄园文化一类词语的集合，主要包括“文化”“民族”“欧式”等词语；服务完善主题指的是景区及周边完善的服务设施一类词语的集合，主要包括“导游”“住宿”“周边”。Others 类主题中的词语过于杂乱，无法判断出其归属于某类特定的主题，因此本文不考虑此类主题。从性别比例我们可以看出，女性更关注自己的主观感受，如游玩感受良好、服务完善。男性更关注一些客观的事物，如自然环境和特

色文化。此外，高影响力的游客对丰厚的历史底蕴、优美的自然环境以及良好的游玩感受的需求最高，对其他主题的需求相差不多。

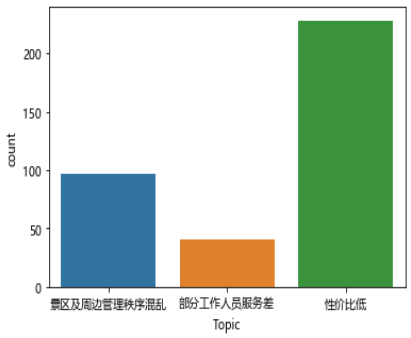


图 21 差评类主题分布

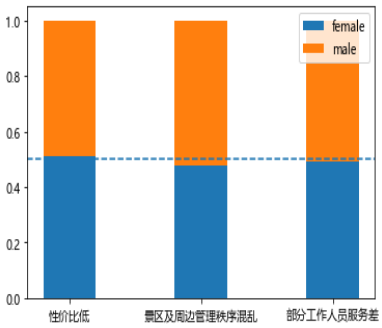


图 22 差评类主题性别分布

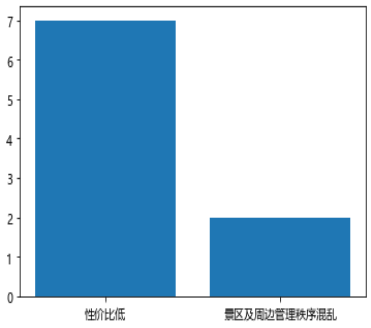


图 23 差评类高影响力游客关注的主题分布

(5).对差评类的评论文本内容进行主题建模，得到了三个主题，分别为性价比低、景区及周边管理秩序混乱以及部分工作人员服务差。性价比低主题主要包括“门票”“买票”“没什么”等词语；景区及周边管理秩序混乱主题主要包括“很漂亮”“景色”“观光”等词语；部分工作人服务差主题主要包括“老工作人员”“管理”“窗口“等词语；不同性别对差评类主题没有明显的区别。此外，由于此类数据中高影响力的游客仅发表 9 条评论，不具有参考价值，因此无法分析出该类群体在差评主题类的分布。

(6).从民俗古镇、现代科技农园以及乡村庄园的数据可以看出，三种情况下对主观感受类主题的关注度均较多，这也和好评类数据中游玩感受良好占比最多相一致。同时，高影响力的游客对优美风景的需求低于大众对优美风景的需求。

4.3 情感分析结果

从情感分析中可以看出，不论是民俗古镇、乡村庄园还是现代科技农园。其游客的情感倾向均大于 0.8。只有少部分游客会存在一些消极负面的情感。可见游客对乡村文旅景点大多持满意的态度。结合当下乡村文旅的发展现状来看，乡村文旅仍具有很大发展空间。

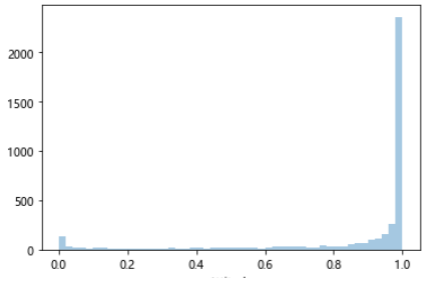


图 24 乡村庄园评论情感分布

在满意度维度下进行情感分析，首先观察所有好评数据的情感分析，发现大多数游客情感倾向均符合自己的评分。只有少部分游客的情感倾向与自己的评分不符。我们具体分析，得到了两个原因。其一是来自游客自身的原因，游客发表了负面消极的评论，但是却在评分打了高分。其二是来自内部原因，当前的情感分析工具，对自然语言情感分析的识别结果仍存在偏差。因此最终得到了如图 25 分布。

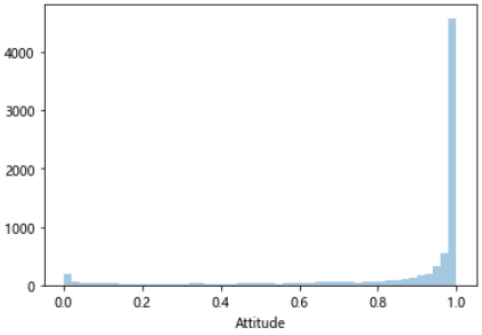


图 25 好评数据情感分布

同理，观察所有差评数据的情感分析，和上述分析类似，发现大多数游客的情感倾向均符合自己的评分。只有少部分游客的情感倾向与自己的评分不符。具体分析得到了两个原因。其一是来自游客自身的原因，游客发表了正面积极的评论，但是却在评分打了低分。其二也依旧是来自内部原因，当前的情感分析工具，对自然语言情感分析的识别仍存在偏差。因此最终得到了如图 26 分布。

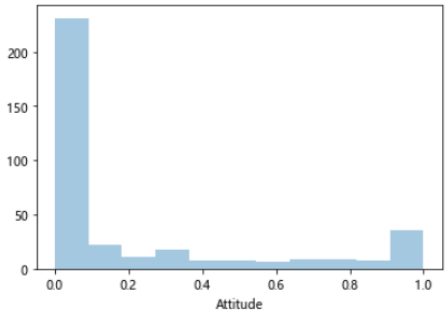


图 26 差评数据情感分布

4.4 共现词分析结果

在共现词分析中,根据 LDA 分析的结果，我们将上述关键词分为四个主要主题，分别是旅行体验，自然风景，人文景观，特色内容。

本文对共现主题的分布进行统计

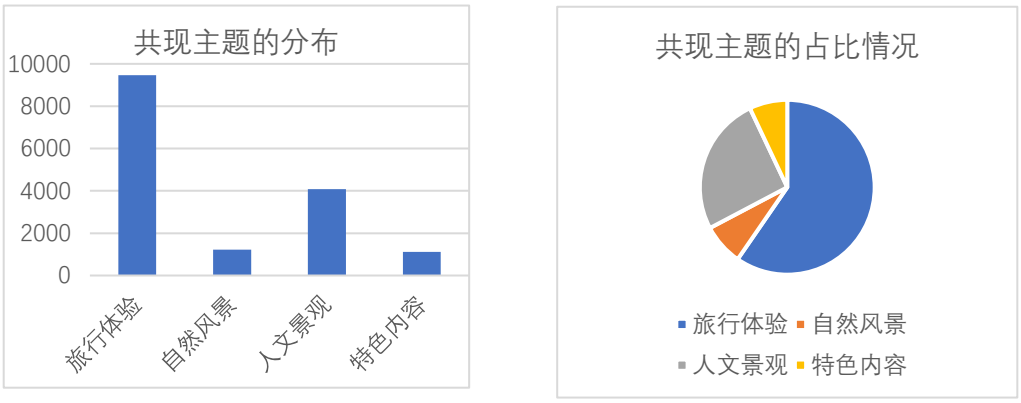


图 27 共现主题分布

在结果分析中，旅行体验是最能影响人们对价格态度的，人文景观也是关于价格讨论的重点，自然风景和特色内容则排在最后，这种结果可能来自于不是每个景区都包括这两种要素，旅行体验是最为重要的，根据对涉及相关关键词的进一步解读，在旅行的时间中，能不能给予充分的体验内容往往非常重要，如果景区的内容较为单调，或者由于拥堵等原因导致旅客并没有充分体验，就会影响旅客对价格值不值得的判断，这种影响往往比特色内容是否精彩，自然风光是否美丽更能对旅客产生影响。

4.5 小结

随着旅游平台的快速发展，人们越来越倾向于在线安排旅游，并在之后进行在线评论。分析景点旅客的客观评价，不仅为其他旅客决策提供信息，同时也为景点运营者提供改进信息。本文根据在线评论文本情感分析以及共现分析，总结出以下结论。

1.针对不同的景区而言，旅客所关注的侧重点不同，针对民俗古镇，旅客更看重良好的游玩体验,而对自然风景,民俗文化,人文环境主题的需求差别不大。对乡村庄园而言，对各类主题的统计分析可以看出游客更关心自己的游玩感受，对此需求更高，而对特色产业，特色历史文化，特色环境风景主题的需求差别不大。对于科技农园而言，游客更关心自己与家人之间的互动体验，对良好的亲子游玩体验需求更高，而对性价比的需求最低。

2.针对男女性别差异而言，从性别比例我们可以看出，女性更关注自己的主观感受，如游玩感受良好、服务完善。男性更关注一些客观的事物，如自然环境和特色文化。男女游客对于其他类主题的需求大致相当。针对这一点，景区可以进行区分化发展，尽量更多的满足双方面需求。

3.针对游客的影响力而言，高影响力的游客对优美风景，庄园特色的环境风景的需求相对较低，对科技产业的需求最高同时对性价比的需求最低。如果景区希望在高影响力用户中获得更多好评，可以从增强特色这一步入手。

4.针对差评和好评而言，影响差评的主要因素包括游玩体验差和内容少，这也同时是影响旅客对价格的态度，为了争取更多的回头客和更少的差评，景区应该加强对员工的管理和园内的安置，并对内容加以丰富。

第五章 不足与展望

1．在数据采集阶段中，本文爬虫采集数据的效率低下，最终只获得了 1 万余条数据。在后续的研究中，可以增加数据量，以保证研究结论的准确性。

2.乡村文旅作为近几年来新兴话题。部分乡村景区的建设仍处于初级阶段，许多乡村景区并未入驻携程，这也是最终数据量比较少的原因之一。

3．本文采用的情感分析方法，并不能准确的判断出评论的情感倾向。对部分评论文本情感的判断会有偏差，在后续研究可以采取依存句法等其他方法来判断情感倾向。

4．共现分析仅能得到价格与其他因素之间存在相关关系，并不能判断游客对价格需求的影响因素。

参考文献

[1]孙宗锋,赵兴华.网络情境下地方政府政民互动研究——基于青岛市市长信箱的大数据分析[J].电子政务,2019(05):12-26.
[2]金吉琼,刘鸿,郑赛晶.基于在线评论文本挖掘技术的电子烟市场消费热点分析[J].烟草科技,2019,52(12):106-114.
[3]陈雪琳,鲁若愚.基于共词分析的我国双创政策关注热点研究[J].电子科技大学学报(社科版),2019,21(02):9-17.
[4]张尧政,邓少灵.上海迪士尼度假区游客在线评论的情感倾向分析[J].现代商业,2019(05):42-43.
[5]张若愚. 基于文本情感分析的江西省 5A 级景区网络口碑综合评价[D].华东交通大学,2017.
[6]程翠琼,徐健.面向网络游记时间特征的情感分析模型[J].数据分析与知识发现,2017,1(02):87-95.
[7]刘思叶,田原,冯雨宁,庄育龙.游客微博主题情感分析方法比较研究[J].北京大学学报(自然科学版),2018,54(04):687-692.
[8]王少兵,吴升.采用在线评论的景点个性化推荐[J].华侨大学学报(自然科学版),2018,39(03):467-472.
[9]李圆圆.基于 LD A 的游客在线评论主题分类——以秦始皇兵马俑博物馆为例[J].河北企业,2020(04):43-44.
[10]王新宇.基于情感词典与机器学习的旅游网络评价情感分析研究[J].计算机与数字工程,2016,44(04):578-582+766.
[11]倪海燕. 基于网络评论的莫干山网红民宿情感认同研究[D].华东师范大学,2019.