基于文本数据挖掘的周边游需求图谱分析

摘要

随着网络技术的快速发展和旅游业信息的高度密集，文本形式的在线旅游（OTA）和游客的用户生成内容（UGC）数据成为了解旅游市场现状的重要信息来源。然而 OTA 和 UGC 数据具有碎片化和非结构化的特点，内容较为分散，要使用它们对某一特定旅游目的地进行研究时，需要一种能够从文本中抽取相关的旅游要素，并挖掘要素之间的相关性和隐含的高层概念的可视化分析工具。本文基于文本数据挖掘，使用 LDA 主题分类模型、情感分析模型和 Apriori 算法进行了周边游需求图谱分析。

针对任务一，本文首先对微信公众号文章原始数据进行预处理，包括文本去重、去除文本缺失值与基于 jieba 的文本分词、词性标注和去停用词。由于微信公众号文章主题词语具有一定的特殊性，保留文章中词性为名词、名动词的词汇。TF、TF-IDF 和 textrank 三种算法进行文本关键词提取的结果，最终选择 TF-IDF 算法提取的关键词。利用 sklearn 库中的 LDA 模型进行主题分类，根据模型困惑度确定主题数，并标记文章主题。再根据 Jaccrad 相似度判断分类主题是否与文旅主题相关。最终通过文章主题标记得出分类结果。

针对任务二，首先对评论数据进行文本预处理，将同一品牌的各连锁店归为同一产品。从景区评论、酒店评论和餐饮评论的名称中提取旅游产品，并编号（产品的依托语料为产品第一次出现的语料）。在产品热度分析时，利用 textrank4zh 库对评论进行分句，并通过属性词匹配的方式，计算有效评论的情感得分。然后设置该年产品评论数与产品评论情感得分的权重计算该年产品的产品热度。最终按照不同年份分别以产品热度排序。

针对任务三，先将产品名称通过人工词典处理为产品简称，以游记攻略、微信公众号文章、景区评论、酒店评论和餐饮评论为语料库，利用 Apriori 算法计算其关联度，找到以景区、酒店、餐饮等为核心的强关联规则。凭借得到的强关联规则进行构建本地旅游图谱。

针对任务四，通过比较新冠疫情前后本地旅游图谱和旅游产品热度变化，书写一封旅游行业发展的政策建议信。

关键词：TF-IDF 模型、LDA 模型、情感分析、Apriori 算法、知识图谱

目录

图录

一、 引言

1.1 挖掘背景及意义

随着互联网和自媒体的繁荣，文本形式的在线旅游（Online Travel Agency，简称 OTA ）和游客的用户生成内容（User Generated Content，简称 UGC）数据成为了解旅游市场现状的重要信息来源。由于 OTA 和 UGC 数据的内容较为分散和碎片化，要使用它们对某一特定旅游目的地进行研究时，迫切需要一种能够从文本中抽取相关的旅游要素，并挖掘要素之间的相关性和隐含的高层概念的可视化分析工具。

知识图谱是由 Google 公司在 2012 年提出的新概念。用信息可视化技术将知识以图的形式表示，图由节点和边构成，节点对应知识图谱的实体，自然界中的每个对象都可以称之为一个实体。本质上，知识图谱旨在描述真实世界中存在的各种实体或概念及其关系，其构成一张巨大的语义网络图，节点表示实体或概念，边则由属性或关系构成。

本地旅游知识图谱是在通用知识图谱的基础上加入了更多针对旅游行业的需求。本地旅游图谱采用图的形式直观全面地展示特定旅游目的地“吃住行娱购游”等旅游要素，以及它们之间的关联。旅游要素分为多个等级，需要从文本中挖掘出面对不同要素游客所关注的下一级要素。旅游要素之间会存在关联关系，在本地旅游图谱中使用连接两个节点的一条边来表示。

在近年来新冠疫情常态化防控的背景下，我国游客的旅游消费方式已经发生明显的转变。在出境游停滞，跨省游时常因为零散疫情的影响被叫停的情况下，中长程旅游受到非常大的冲击，游客更多选择短程旅游，本地周边游规模暴涨迎来了风口。疫情防控常态化背景下研究分析游客消费需求行为的变化，对于旅游企业产品供给、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 问题描述

问题一分析：由于微信公众号文章数据量庞大，需要借助机器自动提取文章中的关键词，并计算词频。观察实验语料可知，存在完全重复文本内容或文本内容具有缺少，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和去缺失值处理。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。

问题二分析：由于景区、酒店和餐饮评论里，旅游产品直接表现在名称中，故可以直接提取。由于有同一平牌的连锁店，需要借助机器提取不同连锁店的同一品牌名作为产品。由于旅游产品评论的情感倾向对热度具有影响，故将同一产品的所有评论的积极情感倾向得分之和与该产品的出现频次按一定权重计算产品热度。

问题三分析：将游记攻略、景区评论、酒店评论、餐饮评论和与文旅相关的微信公众号文章作为语料库，利用关联算法计算问题二提取出的旅游产品关联度，提取强关联

规则，在此基础上构建旅游图谱。

问题四分析：根据疫情前后的本地旅游图谱和前三问提取出的数据分布进行可视化分析，得到新冠疫情前后茂名市旅游产品的变化，并撰写一封茂名旅游行业发展的政策建议信。

二、 微信公众号文章分类

2.1 流程设计

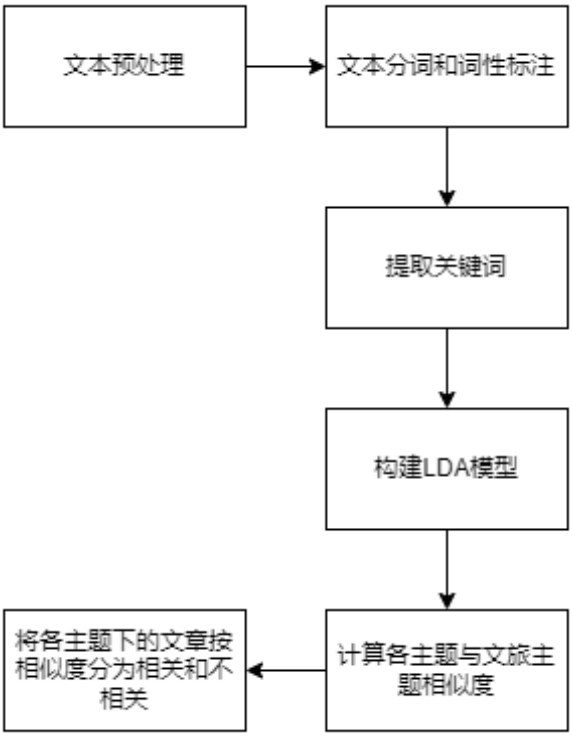

图 1　微信公众号文章分类流程图

2.2 数据预处理

2.2.1 文本去重

原始微信公众号的推送文章，存在内容完全一致的文章，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公众号文章具有参考价值。图 2 中给出了公众号文章内容完全一致的示例。

文章ID	公众号标题	发布时间	正文
1194	【1元购票】鼎龙湾德萨斯水世界9月钜惠点进来！	2018-08-31 16:21	1元可购票？？？确有此事德萨斯水上乐园9月特惠大酬宾...化州分部地址：化州河东汽车总站旁
1195	【1元购票】鼎龙湾德萨斯水世界9月钜惠点进来！	2018-08-31 17:10	1元可购票？？？确有此事德萨斯水上乐园9月特惠大酬宾...化州分部地址：化州河东汽车总站旁

图 2　重复文本

本文对原始微信公众号文章进行按时间排序，使用 python 程序判断内容是否完全重复，保留发布时间最早的评论。

2.2.2 去除文本缺少值

原始微信公众号的推送文章，存在内容为不到一句话，甚至内容为空的文章。图 3 中给出了公众号文章内容少和空的示例。

文章ID	公众号标题	发布时间	正文
1005	关于粤K27618号大客车排气管“喷火”事件的情况说明	2018-01-05 16:57	
1079	一骑红尘妃子笑 原是茂名荔枝来	2018-06-15 14:00	点击上方蓝字，关注我们

图 3　缺少值

本文通过 Python 程序判断内容是否满足文章条件，进行删除操作。

2.2.3 文本分词

（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。

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。故本文采用 Jieba 分词工具进行文本分词。

词性标注　由于与文旅相关的词语具有一定的特殊性，词性通常为名词、动名词等，如“旅游”、“爬山”、“景点”等，为保证后续主题词识别的准确性，本文基于词性标注的结果进一步筛选关键词。

本文采用 jieba 对分词的词性进行标注，部分常见词性标记如下：

n	名词	取英语名词 noun的第1个字母。
nr	人名	名词代码 n和“人(ren)”的声母并在一起。
ns	地名	名词代码 n和处所词代码s并在一起。
nt	机构团体	“团”的声母为 t，名词代码n和t并在一起。
nz	其他专名	“专”的声母的第 1个字母为z，名词代码n和z并在一起。
o	拟声词	取英语拟声词 onomatopoeia的第1个字母。
p	介词	取英语介词 prepositional的第1个字母。
q	量词	取英语 quantity的第1个字母。
r	代词	取英语代词 pronoun的第2个字母,因p已用于介词。
s	处所词	取英语 space的第1个字母。
tg	时语素	时间词性语素。时间词代码为 t,在语素的代码g前面置以T。
t	时间词	取英语 time的第1个字母。
u	助词	取英语助词 auxiliary
vg	动语素	动词性语素。动词代码为 v。在语素的代码g前面置以V。
v	动词	取英语动词 verb的第一个字母。
vd	副动词	直接作状语的动词。动词和副词的代码并在一起。
vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起。

图 4　结巴部分词性标注

导入 Jieba 分词中词性标注函数 jieba.posseg，保留分词词性为名词、人名、名动词、地名的词语。

```
In [8]: import jieba.posseg as psg
   ...: sentence=psg.cut("2017的旅程已经结束2018的未来拉开了帷幕新的一年里，请对自己好一点,从现在开始，享受自己的人生")
   ...: for w in sentence:
   ...:     print(w.word+'/'+w.flag,end='')
   ...:
2017/m的/uj旅程/n已经/d结束/v2018/m的/uj未来/t拉开/v了/ul帷幕/n新/a的/uj一年/m里/f，/x请/v对/p自己/r好/a一点/m,/x从/p现在/t开始/v，/x享受/v自己/r的/uj人生/n
```

图 5　文本分词示例

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，本文采用百度和哈工大停用词表的结合，并根据分词结果，不断修正停用词表。

2.3 文本关键词提取算法

识别公众号文章内容是否与文旅相关，主要任务之一是提取文章内容关键字 [1]。

基于 TF 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

基于 TF-IDF 的关键词提取 词频-逆文档频率（TF-IDF）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。

基于 TextRank 的关键词提取 textrank 方法 TextRank 算法是由 Google 搜索的核心网页排序算法 PageRank 改编而来，利用图模型来提取文章中的关键词，其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。

分析三种方法的关键词提取结果，基于词频的提取方法中，有较多的无用词，如“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。故本文采用了 TF-IDF 算法提取关键词。

2.4 基于 LDA 模型的主题词典构建

2.4.1 LDA 主题提取算法

LDA（Latent dirichlet allocation）是有 Blei 于 2003 年提出的三层贝叶斯主题模型，[2] 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：

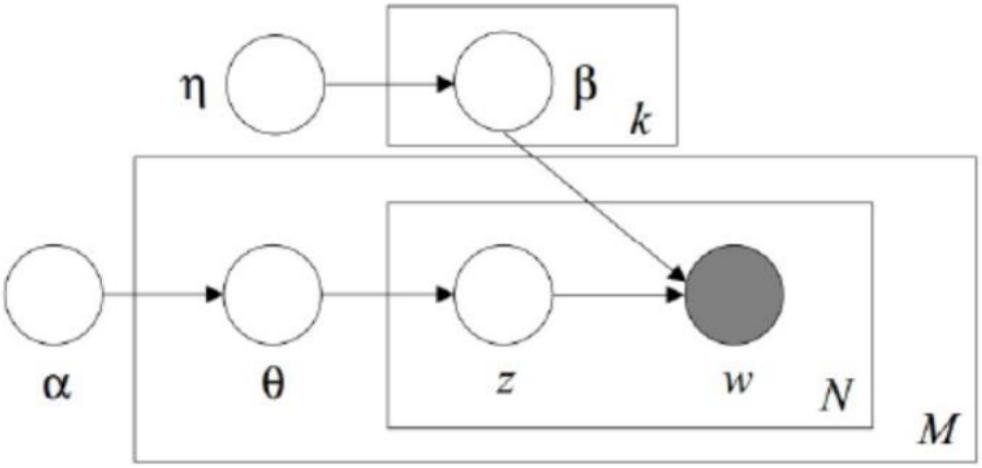

图 6　LDA 主题提取流程图

其中 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，依赖于 Dirichlet 分布参数 α，话题的词频则依赖于参数 λ，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 w1, w 1, w1...,wn。

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 包含一个序列 w1,w2,w3,„wN，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 $p(\theta|\alpha)$ 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 $p(w_n|\theta_d,\beta)$，其中参数 α 是一个 k 维的 Dirichlet 的一个参数，如公式 1 所示。

$$p(\theta|\alpha)=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\Pi_{i=1}^{k}\Gamma(\alpha_i)}\theta_i^{\alpha_1-1}\cdot\cdot\theta_k^{\alpha_k-1} \tag{1}$$

2.4.2 主题建模

基于 LDA 主题提取的原理 [3]，使用基于 Python 语言的机器学习工具 Sklearn (全称 Scikit-Learn) 进行实现，并基于模型提取出的每篇微信公众号文章内容的主题分布，选取概率最大的主题对文章进行标记。根据分类主题数 1-100 的模型主题困惑度曲线，选定使模型主题困惑度尽量小且周围曲线光滑的分类主题数。

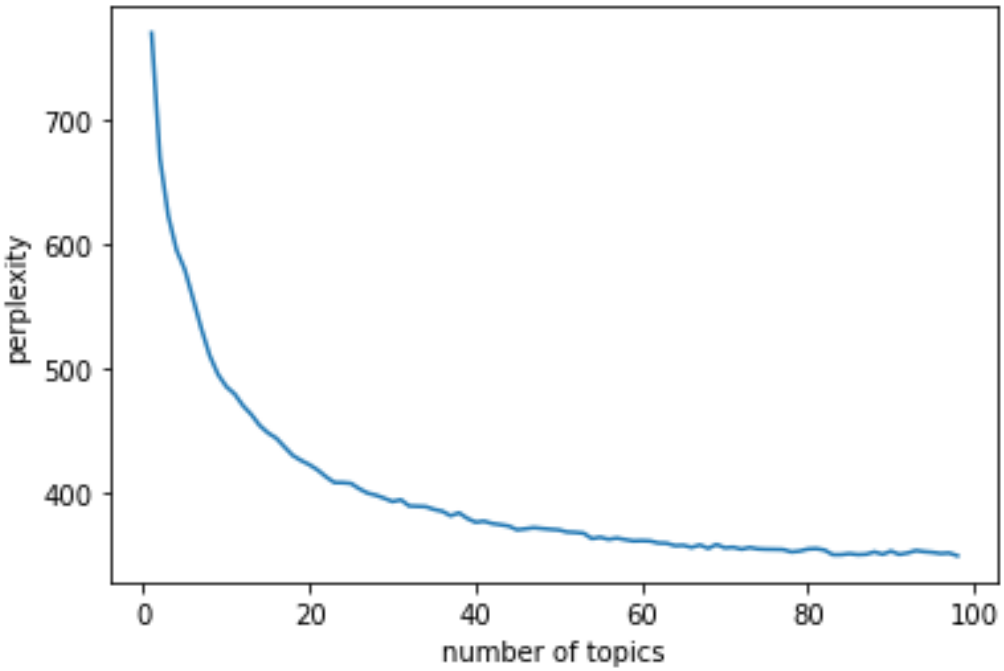

图 7　1-100 主题数的困惑度曲线

本模块使用 pyLDAvis 对内容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该主题下相关性最高的 30 个词汇。

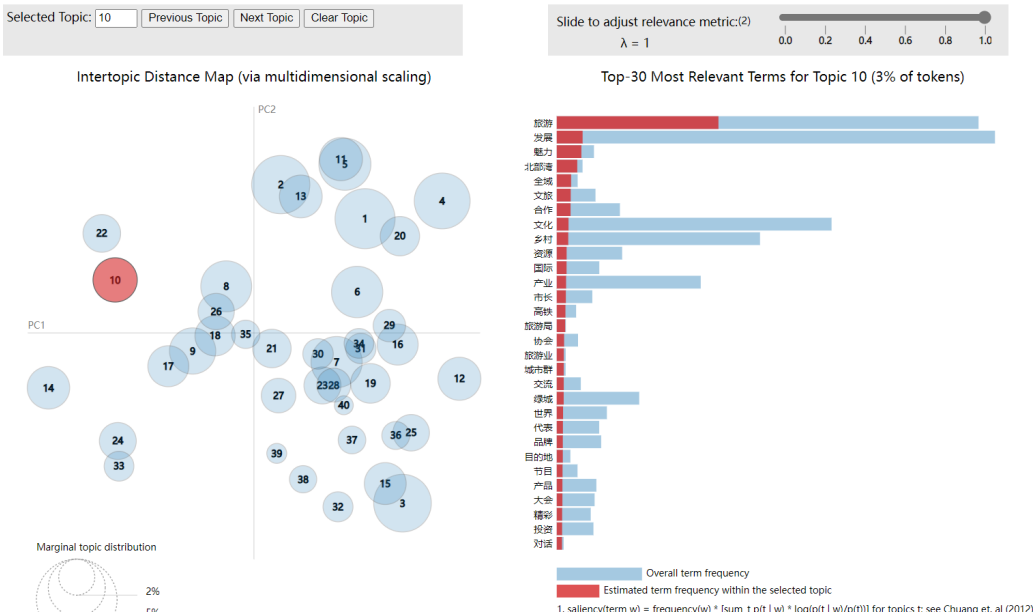

图 8　微信公众号文章 pyLDAvis 主题分类可视化（Topic10）

2.4.3 标签分类

Jaccard 相似度 jaccard index 又称为 jaccard similarity coefficient 用于比较有限样本集之间的相似性和差异性。主要应用场景：

（1）比较文本的相似度，用于文本的查重与去重；

（2）计算对象间的距离，用于数据聚类。

用于衡量有限样本集之间的相似程度：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \tag{2}$$

jaccard 距离公式：

$$d_j(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}=\frac{A\Delta B}{|A\cup B|} \tag{3}$$

jaccard 系数取值范围 [0,1]：

当 A==B 时，jaccard 系数为 1；当 A 与 B 不相交，jaccard 系数为 0。

jaccard 距离表示样本或集合的不相似程度，jaccard 距离越大，样本相似度越低。故 jaccard 距离用于描述不相似度，缺点是只适用于二元数据的集合。

基于上面的主题分类模型的分类主题关键词，我们将每个分类主题和人工给定的文旅相关主题进行 Jaccard 相似度计算，给定合适相似度阈值，将所有分类主题标记为文旅相关和不相关。再以此为基础，对每篇已经标记过主题的微信公众号文章分为相关和不相关两类。

文章 ID	分类标签
1001	相关
1002	相关
1003	相关
1004	相关
1005	不相关
1006	相关
1007	不相关
1008	不相关
1009	相关
1010	相关
1011	相关
1012	相关

图 9　微信公众号文章部分分类结果

三、 周边游产品热度分析

3.1 流程设计

图 10　周边游产品热度分析流程图

3.2 提取旅游产品

在观察附件提供的 OTA、UGC 数据结构与内容时，发现景区评论，酒店评论和餐饮评论的产品名称可以直接通过其名称进行提取。

景区评论ID	城市	景区名称	评论日期	评论内容
1366	茂名	白水瀑布	2019-06-16	白水瀑布位于阳江阳春市八甲镇10多公里外的崇山峻岭中。瀑布垂直落差…
1489	茂名	白水瀑布	2019-11-05	只能到山脚！水电站封路还能进！
1224	茂名	保利·南海 1 号展览馆	2018-12-14	非常值得参观。海上丝绸之路的实物见证，既体现了800多年前海船的前世…
1255	茂名	保利·南海 1 号展览馆	2019-02-03	就是看沉船讲历史，很好的沙滩。
1151	茂名	保利银滩浴场	2018-08-02	今年6月中旬时去过，当时以为有沙雕展才进去，结果沙雕展还没开始，目…
1469	茂名	保利银滩浴场	2019-10-12	感觉沙子确实挺细软的人也不是特别多感觉还是挺不错的
1372	茂名	潮州风吹岭石刻群	2019-06-24	潮州柘林风吹岭是个值得去游玩的地方啊！风吹岭，岭高面海，一年四季，…
1022	茂名	寸金动物园	2018-01-26	那里很大，有各部分景点，有游乐设施，有动物园，也有博物馆，景色很美
1248	茂名	寸金动物园	2019-01-17	寸金公园全名为“寸金桥公园”，坐落在湛江市赤坎区西侧，寸金桥公园因…

图 11　景区评论

酒店评论ID	城市	酒店名称	评论日期	评论内容	入住日期	入住房型
1001	茂名	茂名君悦商务酒店	2018-12-02	干净卫生服务好	2018-12-02	标双
1002	茂名	维也纳国际酒店(茂名电白店)	2018-12-03	环境可以，干净！	2018-12-03	豪华双床房
1003	茂名	茂名永利之家	2018-12-04	环境不错，房间卫生都很好，生活也很方便，就是隔音效果不理想，有时太吵。我定的优惠价，性价比很高的，是一个不错选择。	2018-12-04	标准单人间
1004	茂名	茂名诚荟酒店	2018-12-05	很好.......舒服态度不错	2018-12-05	豪华大床房
1005	茂名	茂名华景商务酒店	2018-12-06	#卫生# #设计风格# #酒店餐饮#	2018-12-06	特惠单人间(无窗)
1006	茂名	茂名荔晶大酒店	2018-12-07	交通方便，离父母又比较近些，所以还是选择这里	2018-12-07	商务房
1007	茂名	好莱登商务宾馆(信宜绍秀体育馆店)	2018-12-08	酒店设施不错，热水很快，中央空调效果很好，服务态度很好，每次出差都是入住这个酒店，值得大家前往入住！	2018-12-08	三人房
1008	茂名	茂名海豚酒店	2018-12-09	酒店隔音效果很差，杯子里有茶垢，窗子只能打开一点点没啥感觉，去听听他吝吝吐吐吝吝吐吐吝吝吐吐	2018-12-09	标准单人房
1009	茂名	茂名荔晶大酒店	2018-12-10	晚上11：00以后，还是很安静的，但早上5:00以后，傍边的河东市场就喧闹起来了	2018-12-10	商务房
1010	茂名	茂名海豚酒店	2018-12-11	没有早餐，有停车场，酒店离海有点点远，很干净，不能加床，我们是国庆去的，平时去住就是120元的，国庆320元	2018-12-11	标准双人房
1011	茂名	茂名南越印象岭南文化主题酒店火车站店	2018-12-12	很有格调哦，房间空间比较大，床也舒服	2018-12-12	精品大床房

图 12　酒店评论

餐饮评论ID	城市	餐饮名称	评论日期	评论内容	标题
1001	茂名	盛香烧鹅（东方市场店）	2018-01-01	很好吃推荐！	主食3选1，有赠品
1002	茂名	清香面包店（车田街店）	2018-01-01	超级好吃 老板又好　量又多	水果忌廉夹心蛋糕（二层）1个，约4磅，圆
1003	茂名	功夫鸡排（光华北路店）	2018-01-01	好吃，不得不说比门口正对面那家好吃。服务态度也好。炸的皮松软，不会炸得很硬。	小吃6选1，提供免费WiFi
1004	茂名	茂名浪漫海岸温德姆酒店望海餐厅	2018-01-02	品种少，冷冻食品多。 7点就不再出食物，难以接受。	自助晚餐（浪漫海岸跨年音乐节）
1005	茂名	清香面包店（车田街店）	2018-01-03	味道不错，至少吃起来不腻，并且全家7个人吃还是挺多的了，就比较实惠咯	水果忌廉夹心蛋糕（二层）1个，约4磅，圆
1006	茂名	优之品西点	2018-01-04	蛋糕香啡，奶油嫩滑，水果新鲜！超赞！到店买的抹茶慕斯蛋糕，榴莲蛋糕更是唇齿留香、回味无穷！	蛋糕2选1，提供免费WiFi
1007	茂名	优之品西点	2018-01-04	芒果好酸好生啊，看起来黄黄的应该挺甜的，实际好酸。。。	蛋糕2选1，提供免费WiFi
1008	茂名	清香面包店（车田街店）	2018-01-04	蛋糕很漂亮也非常好吃，奶油很厚不腻	水果忌廉夹心蛋糕（一层）1个，约2.5磅，圆
1009	茂名	旺角亭（为民店）	2018-01-05	很好，一直都有在吃他家的东西，紫菜卷很好吃	休闲单人套餐，提供免费WiFi
1010	茂名	盛香烧鹅（东方市场店）	2018-01-06	来了好多次了，饭菜都适合我的口味，不错的快餐，值得推荐！还会再来	主食3选1，有赠品
1011	茂名	鲜滋味蛋糕·季念日生日蛋糕（茂南区店）	2018-01-06	榴莲千层没什么榴莲味，而且不新鲜，那点榴莲肉有杂物黑色的	榴莲千层/榴莲雪媚娘2选1

图 13　餐饮评论

在提取过程中，我们将同一品牌的却不同位置的连锁店归为同一类产品，并将品牌名作为产品名称。在进行产品编号时，我们将第一次出现同一产品的语料 ID 作为该产品 ID 所依托的语料 ID。

语料 ID	产品 ID	产品名称
酒店评论-1001	ID1	茂名君悦商务酒店
酒店评论-1002	ID2	维也纳国际酒店
酒店评论-1003	ID3	茂名永利之家
酒店评论-1004	ID4	茂名诚荟酒店
酒店评论-1005	ID5	茂名华景商务酒店
酒店评论-1006	ID6	茂名荔晶大酒店
酒店评论-1007	ID7	好莱登商务宾馆
酒店评论-1008	ID8	茂名海豚酒店
酒店评论-1011	ID9	茂名南越印象岭南文化主题酒店火车站店
酒店评论-1013	ID10	茂名卓钰精品酒店
酒店评论-1014	ID11	化州汇嘉大酒店
酒店评论-1016	ID12	茂名海云雁酒店
酒店评论-1017	ID13	茂名威利国际酒店

图 14　部分旅游产品提取表

3.3 多维度热度评价模型

由于产品热度与评论频次和评论的情感倾向有关，本模块通过利用 python 中的 textrank4zh 库对评论进行分句，并通过属性词匹配的方式，计算有效分句情感得分。再通过求均值的方式，计算该条评论所有分句的情感平均得分。再设置产品评论数与产品评论情感得分权重计算得产品的产品热度。

3.4 中文情感分析工具

3.4.1 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。

3.4.2 Senta

Senta 为百度开发的一款一键式情感分析预测工具 [4]，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%。

图 15　Senta 和 SnowNLP 情感得分结果对比

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所示。分析图可知，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（sentimentlabel 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。

3.5 热度分析

本模块根据评论的积极情感倾向性平均得分作为产品热度得分的计算依据，将产品的评论数和产品每条评论的积极情感倾向性平均得分之和按照 2：1 权重计算得产品热度。

$$Score_i = 2 * fre_i + \sum_{j=1}^{k} posprobs_{ij} \tag{4}$$

score 为热度得分，fre 为产品出现频次，posprobs 为积极情感倾向得分。

将评论按不同年份分开，对每一年的旅游产品计算热度，并各年度按热度排名。

产品 ID	产品类型	产品名称	产品热度	年份
ID201	特色餐饮	果之度	269.929	2018
ID191	特色餐饮	清香面包店	251.498	2018
ID194	特色餐饮	优之品西点	199.219	2018
ID202	特色餐饮	小乔紫菜卷	151.235	2018
ID200	特色餐饮	甜在心扉千层蛋糕	145.447	2018
ID190	特色餐饮	盛香烧鹅	135.977	2018
ID209	特色餐饮	优乐堡	111.273	2018
ID195	特色餐饮	旺角亭	91.6224	2018
ID203	特色餐饮	爱可咖啡馆	91.1229	2018
ID207	特色餐饮	渔乐码头烤鱼牛蛙晚饭餐厅	78.6099	2018
ID196	特色餐饮	鲜滋味蛋糕·季念日生日蛋糕	57.6911	2018
ID212	特色餐饮	CAKE情迷黑森林	57.2282	2018

图 16　2018 旅游产品热度部分排名

产品 ID	产品类型	产品名称	产品热度	年份
ID212	特色餐饮	CAKE情迷黑森林	215.934	2019
ID2	酒店	维也纳国际酒店	162.138	2019
ID201	特色餐饮	果之度	134.89	2019
ID211	特色餐饮	相聚时光	128.474	2019
ID15	酒店	城市便捷酒店	120.564	2019
ID213	特色餐饮	元晟坊蛋糕	114.011	2019
ID191	特色餐饮	清香面包店	112.587	2019
ID17	酒店	茂名国际大酒店	99.9548	2019
ID8	酒店	茂名海豚酒店	95.0926	2019
ID209	特色餐饮	优乐堡	80.2821	2019
ID214	特色餐饮	优悦西餐厅	77.7572	2019
ID194	特色餐饮	优之品西点	73.5478	2019
ID98	景区	广垦国家热带农业公园	64.3184	2019

图 17　2019 旅游产品热度部分排名

产品 ID	产品类型	产品名称	产品热度	年份
ID223	特色餐饮	Hello炸鸡	415.689	2020
ID218	特色餐饮	菠斯蒂蛋糕	317.732	2020
ID201	特色餐饮	果之度	261.718	2020
ID222	特色餐饮	螺香汇·柳州螺蛳粉	253.63	2020
ID211	特色餐饮	相聚时光	228.771	2020
ID212	特色餐饮	CAKE情迷黑森林	227.185	2020
ID220	特色餐饮	十七门重庆老火锅	207.058	2020
ID214	特色餐饮	优悦西餐厅	202.852	2020
ID217	特色餐饮	爵士厚牛排自助餐馆	198.715	2020
ID209	特色餐饮	优乐堡	175.561	2020
ID213	特色餐饮	元晟坊蛋糕	135.104	2020

图 18　2020 旅游产品热度部分排名

四、 本地旅游图谱构建与分析

4.1 Apriori 算法

Apriori 算法 [5] 用于解决大规模数据集的关联分析问题。关联分析（association analysis）或关联规则学习（association rule learning）是从大规模数据集中寻找物品间的隐含关系。但是，寻找物品的不同组合是一项十分耗时的任务，计算代价高，蛮力搜索并不能解决问题，所以需要更智能的方法在合理时间范围内找到频繁项集。Apriori 算法就是解决这个问题的。

产品 ID	产品类型	产品名称	产品热度	年份
ID228	特色餐饮	椒王火锅	1152.8	2021
ID211	特色餐饮	相聚时光	725.893	2021
ID226	特色餐饮	法兰度航空主题餐厅	584.353	2021
ID243	特色餐饮	草原郎烤羊	578.021	2021
ID227	特色餐饮	友情有意音乐餐厅	548.95	2021
ID225	特色餐饮	贵族自助牛排餐厅	538.125	2021
ID234	特色餐饮	付记·招牌酸菜鱼	525.308	2021
ID248	特色餐饮	麦壳西点	461.655	2021
ID255	特色餐饮	围炉海鲜烤肉自助	381.245	2021
ID220	特色餐饮	十七门重庆老火锅	333.294	2021
ID217	特色餐饮	爵士厚牛排自助餐馆	328.287	2021
ID235	特色餐饮	麦当劳	296.122	2021
ID215	特色餐饮	顺德火焰醉鹅坊	295.616	2021

图 19　2021 旅游产品热度部分排名

4.1.1 基本概念

（1）项与项集　数据库中不可分割的最小单位信息称为项 (或项目)，用符号 i 表示，项的集合称为项集。设集合 I=i_1,i_2,$\cdots$,i_k 是项集，I 中项目的个数为 k，则集合 I 称为 k-项集。

(2) 事务　设 I=i_1,i_2,$\cdots$,i_k 是由数据库中所有项目构成的集合，事务数据库 T=t_1,t_2,$\cdots$,t_n 是由一系列具有唯一标识的事务组成的。每一个事务 t_i(i=1,2,$\cdots$,n) 包含的项集都是 I 的子集。例如，顾客在商场里同一次购买多种商品，这些购物信息在数据库中有一个唯一的标识，用以表示这些商品是同一顾客同一次购买的，称该用户的本次购物活动对应一个数据库事务。

（3）项集的频数 (支持度计数)　包括项集的事务数称为项集的频数 (支持度计数)。

（4）关联规则　关联规则是形如 $X \Rightarrow Y$ 的蕴涵式，其中 X,Y 分别是 I 的真子集，并且 $X \cap Y = \phi$。X 称为规则的前提，Y 称为规则的结果。关联规则反映 X 中的项目出现时，Y 中的项目也跟着出现的规律。

（5）关联规则的支持度 (support)　关联规则的支持度是交易集中同时包含 X 和 Y 的交易数与所有交易数之比，它反映了 X 和 Y 中所含的项在事务集中同时出现的频率，记为 support($X \Rightarrow Y$)，即

$$support(X \Rightarrow Y) = support(X \cup Y) = P(XY) \tag{5}$$

（6）关联规则的置信度 (confidence) 关联规则的置信度是交易集中包含 X 和了的交易数与所有交易数与包含 x 的交易数之比，记为 confidenee($X \Rightarrow Y$)，置信度反映了包含 X 的事务中出现 Y 的条件概率。即

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} = P(Y|X) \tag{6}$$

（7）最小支持度与最小置信度 通常用户为了达到一定的要求，需要指定规则必须满足的支持度和置信度阈值，此两个值称为最小支持度阈值 (min_stp) 和最小置信度阈值 (min_conf)。其中，min_sup 描述了关联规则的最低重要程度，min_conf 规定了关联规则必须满足的最低可靠性。

（8）强关联规则 support($X \Rightarrow Y$)>min_sup 且 confidence($X \Rightarrow Y$)≥ min_conf，称关联规则 $X \Rightarrow Y$ 为强关联规则，否则称 $X \Rightarrow Y$ 为弱关联规则。通常所说的关联规则一般是指强关联规则。

（9）频繁项集 设 $U \subseteq I$，项目集 U 在数据集 T 上的支持度是包含 U 的事务在 T 中所占的百分比，即

$$support(U) = \frac{||t \in T|U \subseteq t||}{||T||} \tag{7}$$

式中，$\|\cdot\|$ 表示集合中的元素数目。对项目集 I，在事务数据库 T 中所有满足用户指定的最小支持度的项目集，即不小于 min_sup 的 I 的非空子集，称为频繁项目集或大项目集。

（10）项目集空间理论 Agrawal 等人建立了用于事务数据库挖掘的项目集空间理论。理论的核心为：频繁项目集的子集仍是频繁项目集；非频繁项目集的超集是非频繁项目集。

4.1.2 算法原理

最著名的关联规则发现方法是 R.Agrawal 提出的 Apriori 算法 [6]。Apriori 算法的基本思想是通过对数据库的多次扫描来计算项集的支持度，发现所有的频繁项集从而生成

关联规则。Apriori 算法对数据集进行多次扫描。第一次扫描得到频繁 1-项集的集合 L_1，第 k(k>1) 次扫描首先利用第 (k-1) 次扫描的结果 L_{k-1}。来产生候选 k-项集的集合 C_k，然后在扫描的过程中确定 C_k 中元素的支持度，最后在每次扫描结束时计算频繁 k-项集的集合 L_k，算法在当候选 k-项集的集合 C_k 为空时结束。

Apriori 原理可以避免项集数目的指数增长，从而在合理时间内计算出频繁项集。

4.2 产品简称

在观察游记攻略、景区评论、酒店评论、餐饮评论和微信公众号文章内容语料的内容时发现，其中提到的旅游产品大部分是以简称的形式出现。例如“中国第一滩旅游度假区”在语料中一般以“中国第一滩”出现

为了方便寻找某种旅游产品是否在语料中出现，本模块创建了一个词典包含“茂名”，“高州”等地名和“生态旅游区”，“旅游度假区”等词性的词语，删除产品名称中在词典中的词语，从而得到产品简称

4.3 旅游产品关联

在游记攻略、景区评论、酒店评论、餐饮评论和微信公众号文章语料的文章内容中计算每一个旅游产品的支持度，对满足最小支持度的旅游产品再组合成 2-项集进行计算支持度。对满足最小支持度的 2-项集计算置信度，满足最小置信度等于 0.2 的 2-项集具有强关联规则。

4.4 本地旅游图谱

根据得到的强关联规则，使用 python 中的 networkx 库绘制知识图谱 [7]:

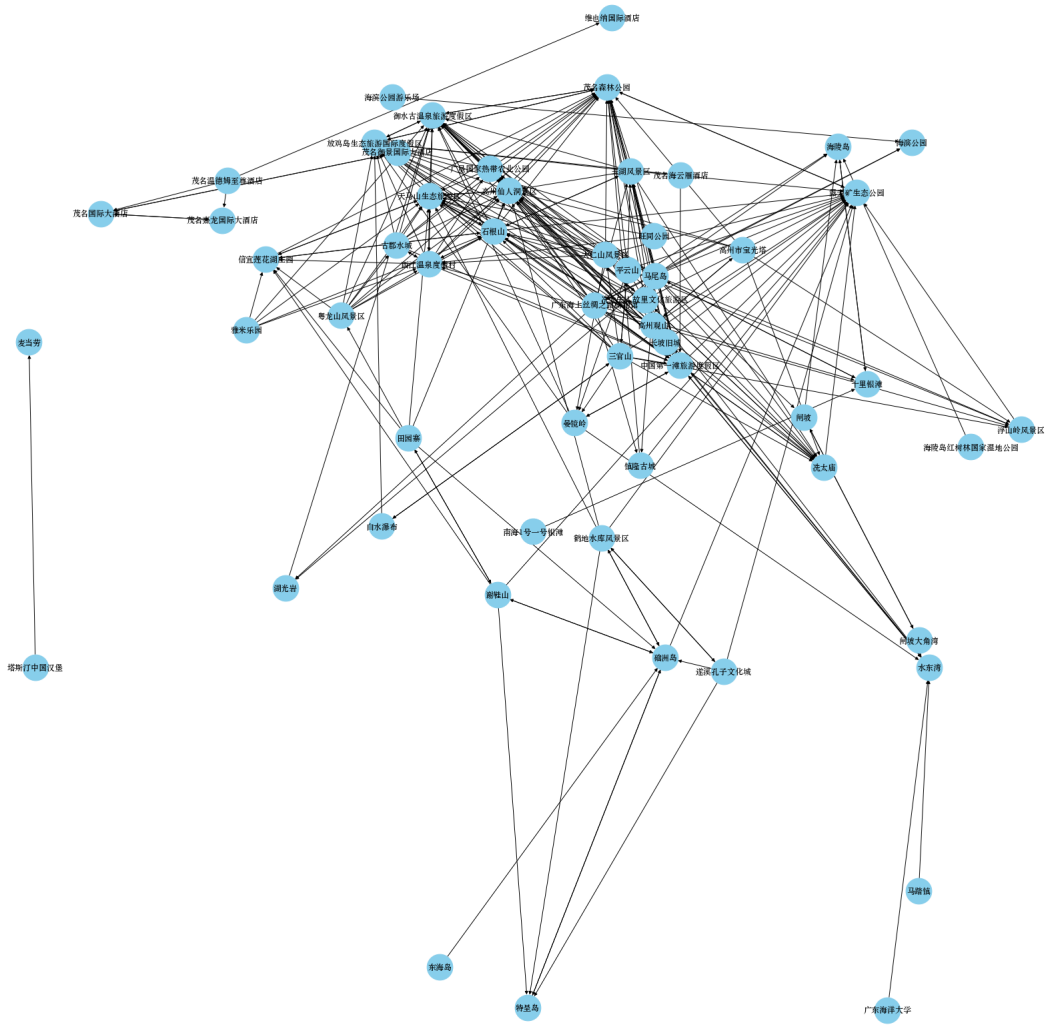

图 20　2018-2021 本地旅游图谱

五、 疫情前后旅游产品需求的变化分析

5.1 本地旅游图谱分析

通过第三问模型，分别得到 2018-2019，2020-2021 旅游产品间的强关联规则，并绘制本地旅游图谱：

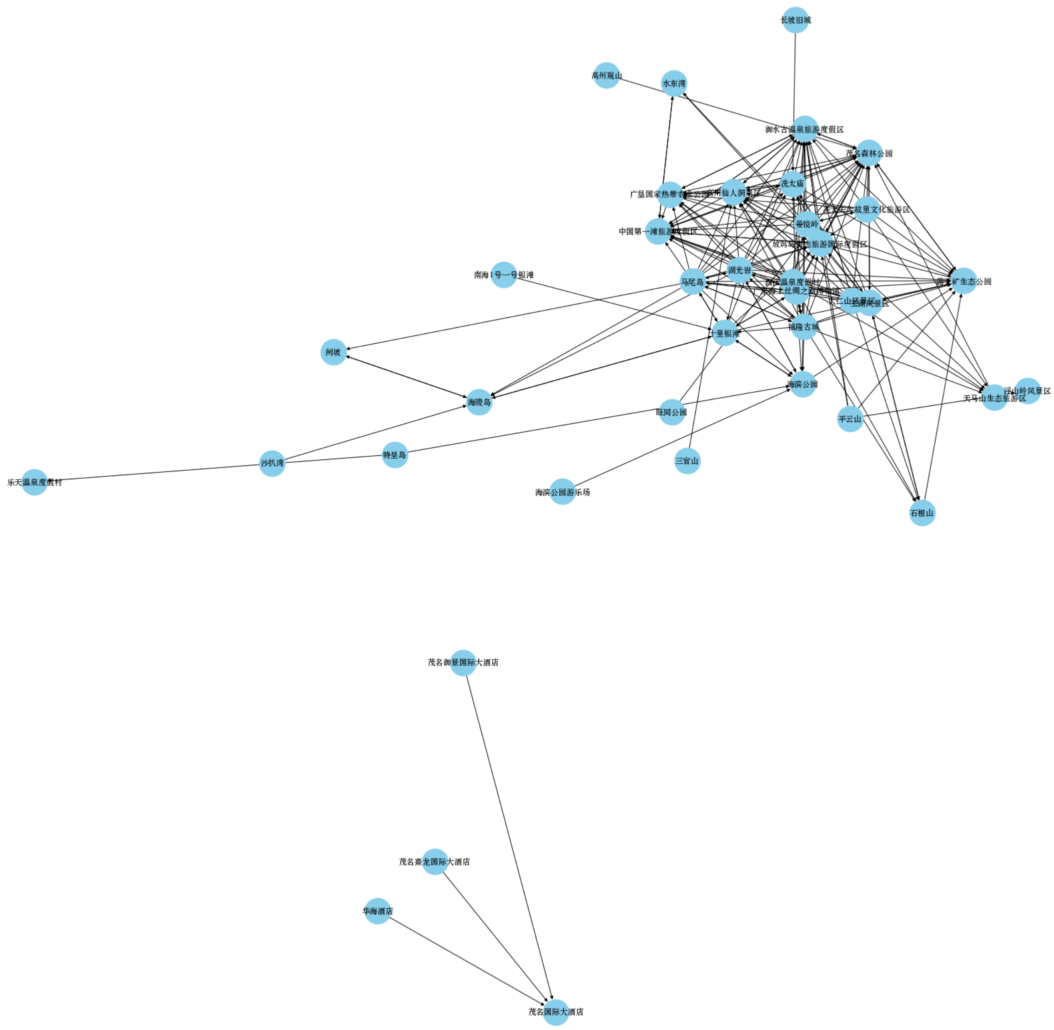

图 21　2018-2019 本地旅游图谱

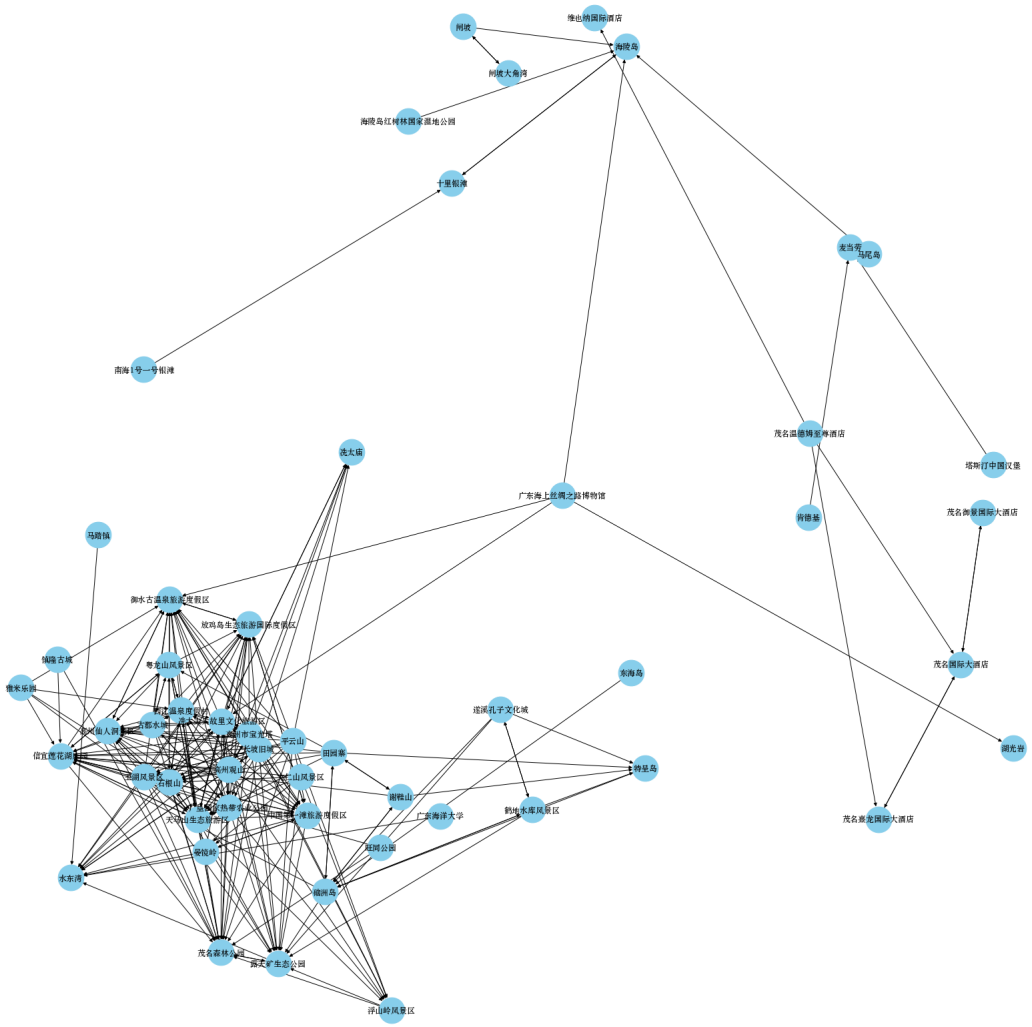

图 22　2020-2021 本地旅游图谱

5.2 旅游产品热度分析

根据问题一得到的各年度微信公众号文章的主题与文旅主题的相关性，绘制比例图：

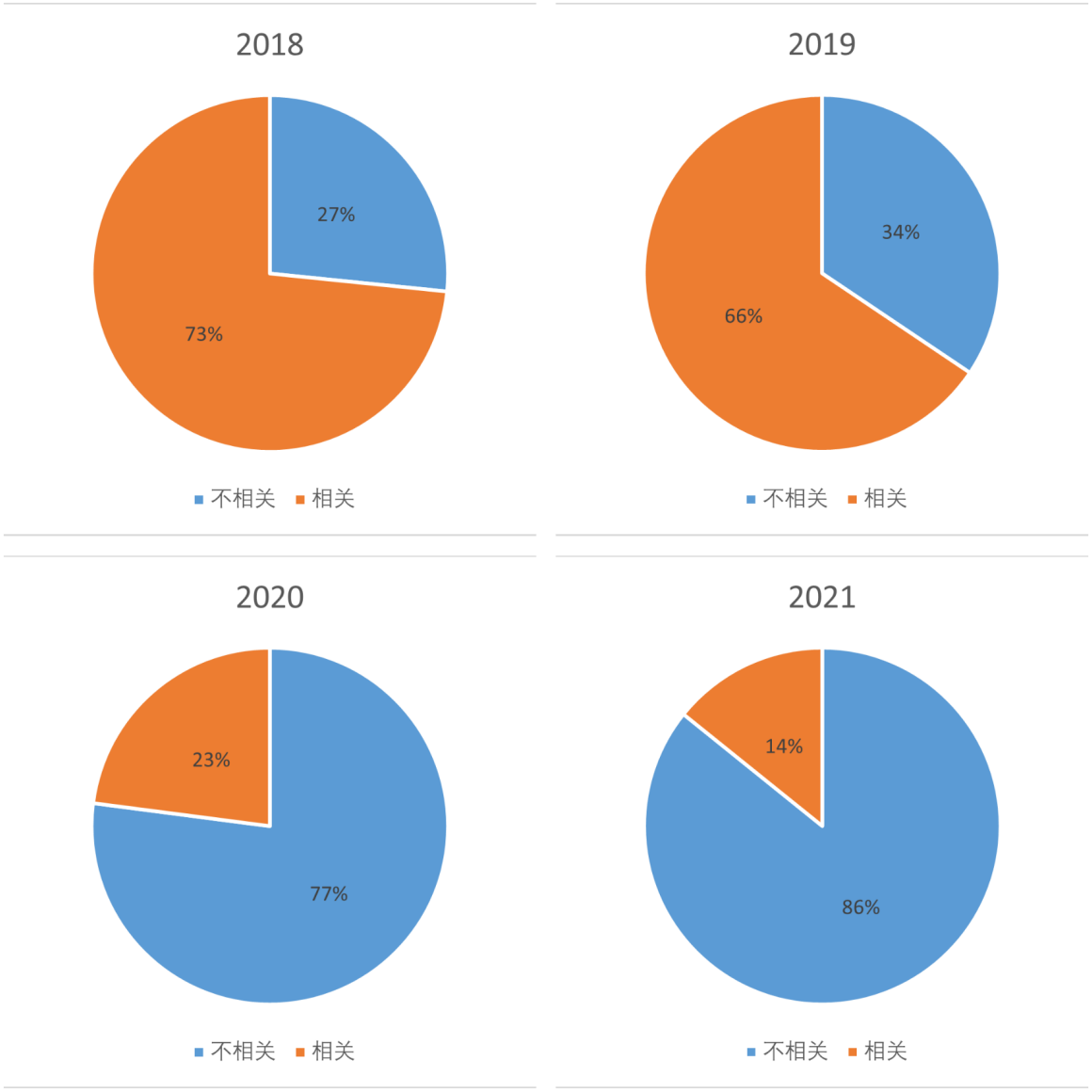

图 23　各年度微信公众号文旅相关比例

可以看出，在新冠疫情前（2018-2019 年），微信公众号文章中与文旅主题有关的占比很大；而在新冠疫情爆发后（2020-2021 年），由于疫情防控政策的宣传，这一占比一再减小。从这一情况可以间接说明旅游业的发展处于下滑状态，也许要等到疫情结束后才会再次呈爆发式增长。

根据问题二得到的景区、酒店和特色餐饮的年度热度，绘制各年度这三个产品类别占总热度的比例图：

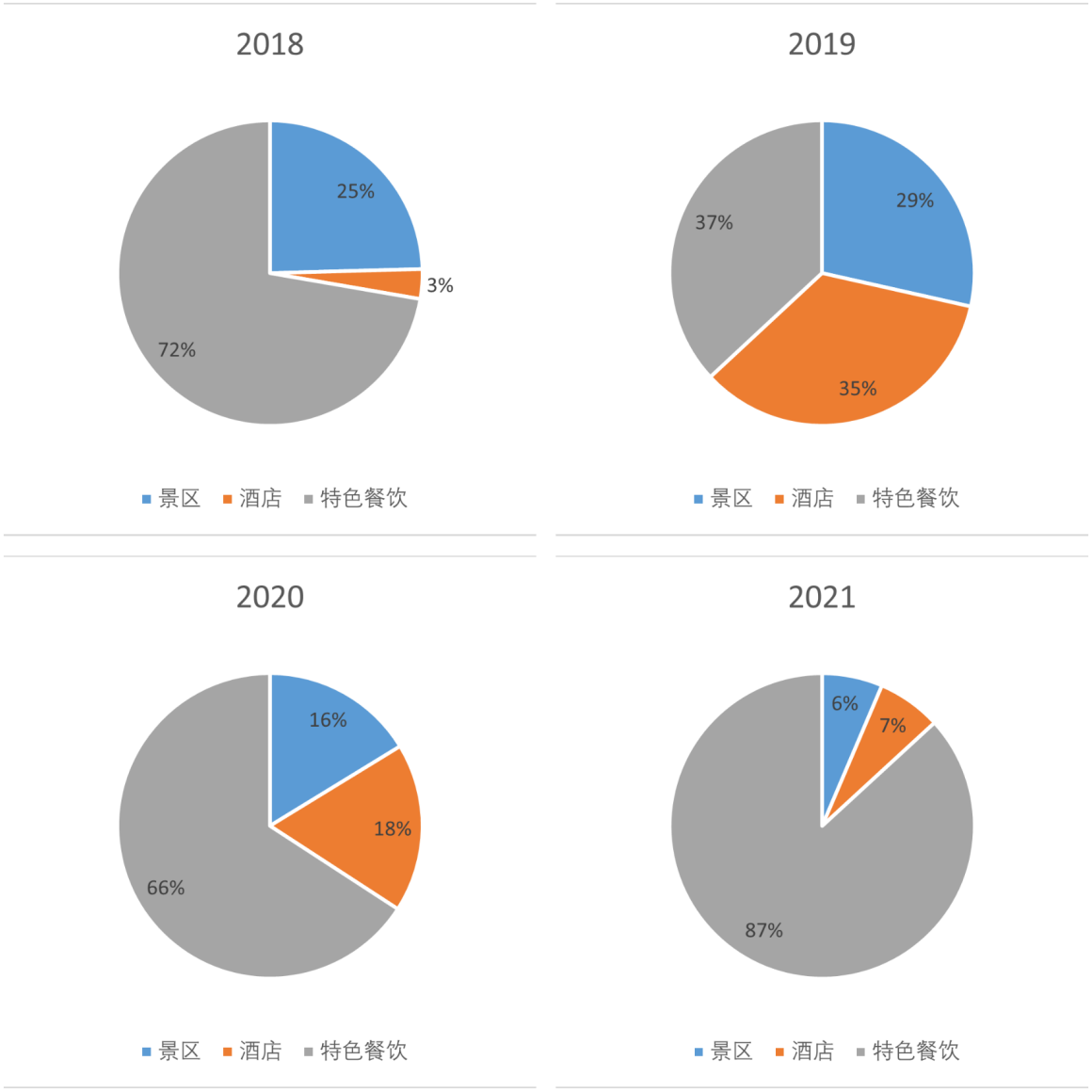

图 24　各年度产品类型总热度占比

反观旅游产品的热度占比，可以发现随着时间的推移，景区的热度占比逐年下滑，而酒店的热度占比处于上下波动状态，可能是由于疫情时的酒店隔离政策影响。而特色餐饮的热度占比一直处于较高水平，受疫情的影响最小，可作为日后发展的重点对象。

5.3 基于旅游图谱分析的建议信

关于茂名市旅游行业发展的建议信

尊敬的茂名市旅游主管部门领导们：

您们好！

在近年来新冠疫情常态化防控的背景下，我国游客的旅游消费方式已经发生了明显转变。在出境游停滞，跨省游时常因为零散疫情的影响被叫停的情况下，中长程旅游受到非常大的冲击，游客更多选择短程旅游，本地周边游规模暴涨迎来了风口。

在通过本地旅游图谱分析了新冠疫情前后茂名市旅游产品的变化后，可以发现：

由于疫情的影响，景区和酒店的热度逐年锐减，然而“民以食为天”，特色餐饮的相关产品一直保持着较高热度。同时基于本地旅游图谱可以发现，游客们的旅游路线与景区的地理位置、酒店的居住环境和特色餐饮的热度具有很大的关联，可以构建旅游产品空间化、一体化，打造“食住行”全方位的旅游体验项目，从而让游客更加便捷地享受旅游和饮食的乐趣。

对此，我们向您们提出以下建议：

一、加快推进旅游数字化。

充分挖掘旅游大数据区域中心优势，做好旅游数据加工及产品研发工作，为全省旅游提供信息支持，为本市居民与域外游客有针对性推送特色旅游产品与服务。加快重点景区管理信息化设施建设，为管理部门和游客提供直观准确景点景区人流动态信息。

二、稳步推进文旅融合发展。

我市文化资源比较丰富，具有典型的地方特征，建议加快推进“旅游 +”和“+ 旅游”战略，将历史文化、民俗风情、特色剧种、艺术教育等资源与旅游深度融合，提升长春旅游礼品文化韵味，培育新的增长点。

三、着力解决细分新旅游消费群体的需求，重点打造新型文化旅游目的地。

打造旅游与文创、美丽、健康等跨界融合产品和服务。充分释放文旅“打卡经济”活力，尝试利用美食、网红景点、热播电视剧、电影的取景地等与文旅进行联动。着力打造“音乐节 + 旅游”“电竞 + 旅游”等新业态的旅游产品，形成时尚旅游消费目的地。

四、补齐重点新旅游业态短板，释放消费潜力。

建立健全乡村旅游管理服务网络，提升乡村旅游品质，提高旅游安全突发事件的应急能力。加大财政支持力度，用于乡村源污染治理和资源保护工作。大力提升乡村特色民宿、健康食品、人文风俗等对旅游者的吸引力。

五、降低门票价格，发挥门票价格的牵引作用。

在旅游业中，经济效益的更多来源并非来自景点门票本身。门票收入在其中只是很小的一部分，但有着强大的牵引拉动力量。地方经济应该以旅游景点为核心，拓展相关资源形成产业链，包括住宿业、餐饮业、交通运输业、游览娱乐业、旅游用品和纪念品销售行业等等相关行业，重点是把整个产业链做大做强。景区门票降价在一定程度上减轻游客旅游负担，吸引游客数量增长，使周边游需求增长，同时促进二次旅游消费，从

而对激发对景区营收贡献值提升。

希望这场疫情可以尽快得到控制，也祝旅游行业发展得越来越好。茂名加油! 中国加油!

热心市民

2022 年 4 月 29 日

六、 总结

本文主要通过 TF-IDF 算法、LDA 主题模型、Jaccard 相似度、文本情感分析和 Apriori 算法解决了微信公众号文章分类、周边游产品热度分析、本地旅游图谱构建与分析、疫情前后旅游产品需求的变化分析四个任务。

为解决任务一中微信公众号文章分类问题，本文首先对原始数据进行预处理后，保留词性标注为名词和名动词的词语，对比了基于 TF、TF-IDF 和 textrank 三种关键词提取算法的计算结果，结果显示 TF-IDF 算法的提取最优。使用 sklearn 库中的 LDA 主题模型进行主题分类，根据模型困惑度确定主题数，并标记文章主题。然后根据 Jaccrad 相似度判断分类主题是否与文旅主题相关。最终通过文章主题标记得出分类结果。

为解决任务二中周边游产品热度分析问题，本文首先对各评论中的产品名称进行提取品牌名作为产品名称，随后通过 textrank4zh 库对评论文本进行分句处理，使用 Senta 对分句进行情感分析，最后根据产品出现频次和评论情感倾向得分按照一定的权重计算热度得分。

为构建任务三的本地旅游图谱，本文先将产品名称通过人工词典处理为产品简称，以游记攻略、微信公众号文章、景区评论、酒店评论和餐饮评论为语料库，利用 Apriori 算法计算其关联度，找到以景区、酒店、餐饮等为核心的强关联规则。根据得到的强关联规则构建本地旅游图谱进行可视化分析并挖掘出旅游产品间隐含的关联模式。

为进行任务四中疫情前后旅游产品需求的变化分析，本文基于前三问构造的模型，通过比较新冠疫情前后本地旅游图谱、微信公众号文旅相关文章占比和各类旅游产品热度分布，得出新冠疫情前后茂名市旅游业的变化情况，并撰写一封关于茂名旅游行业发展的政策建议给旅游主管部门。

参考文献

[1] 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

[2] 崔宁, 赵宗良, 吴瑞雪. 基于 LDA 主题模型和偏序集的在线商品评论研究 [J]. 情报探索.2021 年 12 期

[3] 谢宗彦, 张公鹏, 郝志成, 向征. 基于 LDA 的游客网络评论主题分类: 以故宫为例 [J].《情报工程》.2017 年第 3 期:55-63

[4] 张小艳, 白瑜. 基于加权融合字词向量的中文在线评论情感分析 [J].《计算机应用研究》.2022 年第 1 期:31-36

[5] 王媛媛, 胡学钢. 关联规则挖掘研究 [C]. 全国第 16 届计算机科学与技术应用 (CACIS) 学术会议论文集.2004:808-812

[6] 王振武. 数据挖掘算法原理与实现 [M]. 北京：清华大学出版社.2017 年 1 月第 2 版:35-37

[7] 户文月. 全域旅游知识图谱研究 [J].《青岛职业技术学院学报》2021 年第 1 期:67-73