第十届“泰迪杯”
数据挖掘挑战赛

优秀作品

作品名称：疫情背景下的周边游需求图谱分析

荣获奖项：一等并网宿创新奖

作品单位：中国地质大学（武汉）

作品成员：叶浩翔 徐晓虎 李雯凤

指导老师：朱镇

封面为后期添加，原作品没有此页。

“泰迪杯”挑战赛优秀作品

疫情背景下的周边游需求图谱分析

摘 要

新冠疫情给旅游业带来了巨大的冲击。后疫情时代，如何在疫情防控常态化的情况下逐步复苏旅游业、使得旅游业经济保持稳中向好的局面，是各级政府、相关企业以及游客群体共同关心的问题，而促进旅游业的复苏，需要能够迅速识别市场需求。本文依托于旅游市场的 UGC（用户生成内容）和 OTA（在线旅游）的语料数据，采用以深度学习为代表的多种自然语言处理技术，挖掘语料中的有用信息，建立知识图谱并寻找其中的隐含关联。最终给出切实可行的政策建议。

针对问题一，本文首先对 2018-2021 年间的 6286 篇微信公众号文章进行人工标注，判断文章是否与文旅相关；之后通过两个机器学习算法 SVM 与 XGBoost 以及两个深度学习算法 Bi-LSTM 与 GRU 对预处理后的公众号文章语料进行分类。综合考虑准确率、召回率和F_1值，选择最优的 Bi-LSTM 算法对剩余的微信公众号文章进行预测，最后识别出 2217 篇与文旅相关的微信公众号文章。

针对问题二，为了获取评论、游记攻略以及公众号文章中的相关旅游产品实例，本文采取命名实体识别的方法，构建 BERT+Bi-LSTM+CRF 模型从上述三类语料中提取旅游产品的实体，经过实体对齐之后，构建以频次为基础的实体热度计算公式计算旅游产品实体每年的热度值。

针对制作茂名旅游产品图谱的问题三。本文的主要工作是利用 Apriori 算法计算问题二中提取到实体之间的关联度，在此基础上制作知识图谱；此外，本文使用依存句法分析获取了酒店、餐饮、景区三类实体评论中的评价方面-评价二元组，将评价方面作为关系名，评价作为节点名绘入知识图谱当中，制作更细粒度的知识图谱。

针对问题四，是前三部分所得到的结果的分析和针对性的政策建议。为了保证政策建议的科学性、合理性与针对性，我们结合问题二中茂名旅游产品实体热度的结果，使用网络爬虫获取了美食店铺、酒店、旅游景点、旅游景区四类产品实体的经纬度坐标，结合热度值在 ArcGIS 软件中以年度为单位绘制产品热度分布地图，分析旅游地点热度在疫情前后的变化规律。

关键词：知识图谱；旅游需求识别；自然语言处理；深度学习

目 录

1 引言.....1
1.1 研究背景及意义.....1
1.2 研究内容及安排.....1
1.3 主要贡献.....3
2 相关技术.....4
2.1 文本分类.....4
2.1.1LDA.....4
2.1.2 支持向量机（SVM）.....5
2.1.3XGBoost.....5
2.1.4Bi-LSTM.....6
2.1.5GRU.....6
2.2 命名实体识别.....7
2.3 知识图谱与关联算法.....8
2.3.1 知识图谱.....8
2.3.2 Apriori 算法.....9
3 微信公众号文章分类.....10
3.1 问题分析.....10
3.2 数据预处理.....11
3.2.1 人工标注.....11
3.2.2 数据清洗.....12
3.2.3 中文分词.....13
3.3 实验设置与结果分析.....14
3.3.1 实验配置.....14
3.3.2 实验步骤与评价指标.....14
3.3.3 实验结果及分析.....15
3.4 小结.....16
4 周边游产品热度分析.....16
4.1 问题分析.....16

“泰迪杯”挑战赛优秀作品

4.2 实验配置.....17
4.3 旅游产品识别.....17
4.3.1 数据预处理.....17
4.3.2 实验步骤.....18
4.3.3 实验结果与分析.....19
4.4 产品热度分析.....22
4.4.1 定义热度指标.....22
4.4.2 热度计算.....23
5 旅游产品关联分析与旅游产品图谱构建.....24
5.1 问题分析.....24
5.2 关联分析.....25
5.3 旅游产品评价属性提取.....27
5.4 旅游产品图谱构建.....28
6 旅游产品需求变化分析及政策建议.....30
6.1 需求变化分析.....30
6.2 政策建议.....33
7 总结与展望.....35
7.1 论文主要工作总结.....35
7.2 今后研究工作展望.....36
参考文献.....37

“泰迪杯”挑战赛优秀作品

1 引言

1.1 研究背景及意义

新冠疫情对各行各业都带来了巨大的冲击，旅游行业首当其冲，遭受重创。在疫情防控常态化的情况下，如何促使国内旅游行业逐渐复苏，已成为了游客、相关企业和各级政府共同关心的问题[1]。因此，如果能快速准确地识别出疫情前后旅游行业的差别以及消费者新的需求，就能针对性的设定政策，达到促进旅游业逐渐复苏的目的。

消费者发布的 UGC 内容中蕴含了丰富的潜在旅游需求信息，这些内容给分析如何设计消费者满意的旅游产品提供了宝贵的有效信息。然而 UGC 内容具有非结构化特征，质量普遍较低，对研究有用的信息往往隐藏在大量无关信息中，如何有效地、低成本地从海量的 UGC 内容中快速准确地提取有用信息已成为学术界和商业界共同关心的问题。

随着近些年深度学习的迅猛发展，这类技术被学术界和商业界广泛地应用于分析大数据文本，即自然语言处理任务，主要包括文本分类、命名实体识别等[2]。使用机器自动判断文本是否属于某一特定的类别以及从文本中自动抽取出一些预定义的实体类别，用于后续的研究分析。目前，以谷歌的 BERT 为代表的预训练模型在各类自然语言处理任务中都取得了较为优秀的结果，成为了大家解决大数据文本的常用模型。

基于此，本文采用以深度学习为代表的各类自然语言处理处理技术，针对赛题提供的各类 OTA 及 UGC 文本数据，挖掘其中的有用信息并构建知识图谱，尝试找出他们之间的隐含信息。除此之外，根据不同问题的实际情况使用 Python、ArcGIS 等软件对挖掘出的信息进行可视化，并结合各类信息给出促进旅游市场逐步回暖的政策建议。

1.2 研究内容及安排

本文分为以下七个章节：

第一章为引言。首先介绍了旅游行业复苏的迫切性以及消费者发布的 UGC 内容对分析如何设计消费者满意的旅游产品的重要性。随后介绍了目前学术界和

商业界解决这个问题的方法以及本文的研究内容、技术路线及创新点。

第二章为相关技术介绍。该章节主要介绍了本文使用到的一些技术方法，包括文本分类的相关模型、命名实体识别问题的研究综述以及知识图谱技术的简单介绍。

第三章为研究问题一，微信公众号文章分类。首先对问题进行了分析，明确了对判断是否“相关”的定义以及使用的模型。然后介绍了预处理的过程，包括人工标注、数据清洗和中文分词。接着介绍了实验的设置以及实验结果。最后对整个问题进行小结。

第三章为研究问题二，周边游产品热度分析。该章节包含两个研究问题，即旅游产品实例的提取以及对旅游产品按年度进行热度分析。首先对问题进行分析，然后介绍了实验的配置以及旅游产品识别的步骤。随后定义热度指标，对提取出的旅游产品进行热度计算。

第五章为研究问题三，旅游产品关联分析与旅游产品图谱构建。包括问题分析，产品关联度定义与计算，利用依存句法分析进行细粒度的属性提取以及最后的知识图谱构建。

第六章为研究问题四，疫情前后旅游产品需求的变化分析及政策建议。本章使用前文提取出的各种信息，针对不同的情况灵活使用多种工具，对数据进行可视化分析并针对性地给出相应的政策建议。

第七章为总结和展望。首先总结了论文的主要工作，然后针对本研究的一些不足指出了在未来的研究中可以完善的地方。

本文的研究流程图如图 1.1 所示。

“泰迪杯”挑战赛优秀作品

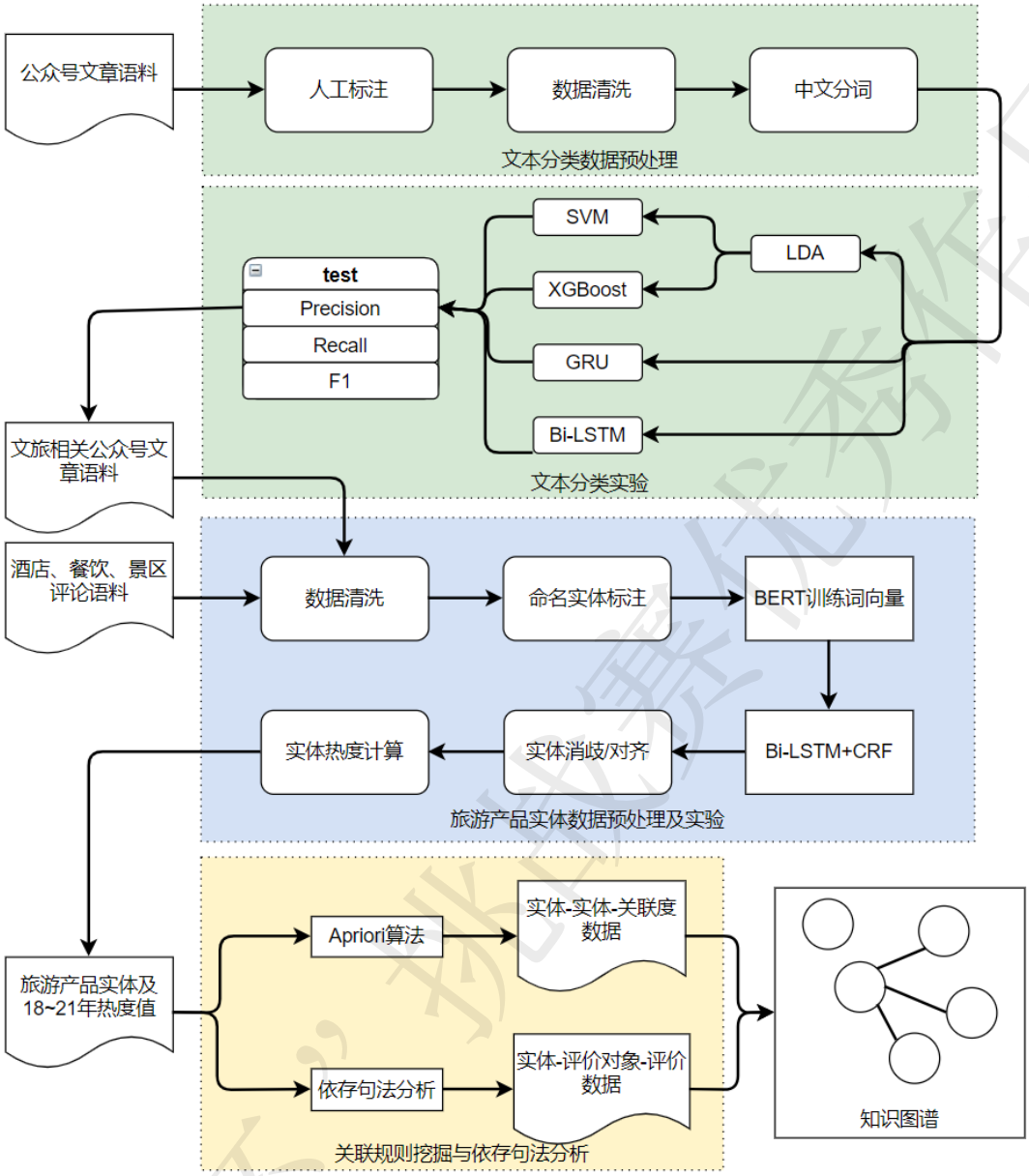

图 1.1 研究流程图

1.3 主要贡献

（1）基于依存句法分析的细粒度属性抽取

在构建知识图谱的过程中，本文使用依存句法分析，提取出旅游产品更细粒度的属性（如服务好、性价比一般等），并将其与旅游产品关联加入到知识图谱中。这些细粒度的属性不但可以丰富知识图谱，方便我们更容易地获取各种隐含关联，发现新的知识，还可以为后续的情感分析等研究提供数据基础。

（2）引用外部知识并灵活运用可视化工具进行展示

根据从数据中提取的旅游产品及其热度，通过网络爬虫爬取对应产品的经纬度坐标作为外部知识，将爬取到的产品经纬度坐标与产品热度值结合，使用 ArcGIS 软件进行可视化。从图中直观的观察旅游热点地区的变化趋势，识别疫情对茂名旅游需求的影响，从而针对性地提出建议，达到促进旅游业逐步恢复的目的。

2 相关技术

2.1 文本分类

文本分类（Text Categorization，简称 TC）技术是信息检索和文本挖掘的重要基础，其主要任务是在预先给定的类别标记（label）集合下，根据文本内容判定它的类别。文本分类在自然语言处理与理解、信息组织与管理、内容信息过滤等领域都有着广泛的应用[3]。目前已有许多针对文本分类问题的研究，主要包括无监督模型、传统的机器学习模型以及深度学习模型等[4]。下文将介绍本研究所使用到的文本分类算法。

2.1.1LDA

LDA 是一种文档主题生成模型，也称为三层贝叶斯概率模型，包含词、主题和文档三层结构。其核心思想是文档可以表示为一系列潜在主题的随机混合，其中每一个主题都代表了在所对应的文档集中全部词的概率分布，与潜在主题相关的词的概率分布较高[5]。

在使用数据集训练 LDA 模型时，为了得到最佳的主题模型，需要确定最佳的主题数 k。Blei 等在其研究中详细阐述了使用困惑度（Perplexity）作为确定主题数的标准，通过绘制 Perplexity-topicnumber 曲线，寻找最佳主题数。

困惑度是一种信息理论的测量方法，可以理解为，对于一篇文档 d，我们的模型对文档 d 属于哪个 topic 有多不确定，这个不确定程度就是 Perplexity。其他条件固定的情况下，topic 越多，则 Perplexity 越小，但是容易过拟合。

吴江等使用 LDA 模型对在线医疗社区文本进行分类，取得了较好的效果[6]。宁宁提出一种基于 LDA 和 Doc2Vec 的改进句向量模型，有效提升了文本分类的效果[7]。除了将 LDA 与句向量进行融合之外，还有一些研究将其与深度学习模

型进行融合来提高文本分类的效果，比如付静等将 LDA 与 BERT 模型结合用于新闻短文本的分类[8]；胡吉明等在此基础上进一步融合了注意力机制[9]；康宸等将 LDA 主题模型和二维卷积进行融合用于处理短文本分类问题[10]，均取得了不错的效果。

2.1.2 支持向量机（SVM）

支持向量机（Support Vector Machine，SVM）方法在很多领域都有应用。它的基本模型是在特征空间上找到最佳的分离超平面使得训练集上正负样本间隔最大[11]。SVM 是用来解决二分类问题的有监督学习算法，在引入了该方法之后 SVM 也可以用来解决非线性问题。同时支持向量机在训练样本较少的情况下的表现要优于其他分类算法。在分类时要使超平面和支持向量之间的间隔尽可能的大，这样超平面才可以将两类样本准确的分开，而保证间隔尽可能的大就是保证我们的分类器误差尽可能地小，尽可能地健壮[12]。

在文本分类领域，SVM 被广泛应用。早在 1999 年，卢增祥等人就将 SVM 与主动学习相结合以解决在有监督学习问题中，训练样本量较少，样本数据维数较大的问题[13]。此后，不断有学者对 SVM 方法进行改进，拓展其使用场景并提高准确率[14, 15]。例如巩知乐等为提高支持向量机算法的性能,提出了一种采用免疫算法对支持向量机参数进行优化的文本分类算法（IA-SVM）算法减少了对支持向量机参数选择的盲目性，提高了 SVM 的预测精度[16]。如今，SVM 已成为解决文本分类问题的标杆算法。

2.1.3XGBoost

XGBoost 是对 Boosting 算法的一种优化，它集成弱分类器为一个强分类器。XGBoost 算法通过不断迭代，生成一棵新树拟合前一棵树的残差，随着迭代次数的增多，精度不断提高[17]。XGBoost 分类器的分类是由学习模型、参数调节和最优化目标函数组成。学习模型是基函数和权重组合的结果，参数调节是建立模型和优化模型的必要过程，目标函数的优化程度决定了最终模型学习的准确率。目标函数优化的越好，模型泛化能力也就越好。即通过最小化损失函数和增加模型复杂度的惩罚项来实现对目标函数的优化，实现模型复杂度和误差综合最优[18]。

在文本分类任务中，XGBoost 通常与其他模型配合使用。如龚维印等为解决

文本分类问题中词袋模型带来的特征维数高、数据稀疏以及文本分类精度低等问题。提出了一种基于卷积神经网络和 XGBoost 的文本分类模型 CNNs-XGB。首先利用 word2vec 对预处理后的数据进行词向量表示，其次利用多尺寸卷积核卷积神经网络进行数据特征提取，最后利用 XGBoost 对深度提取的特征进行分类处理，在清华大学自然语言研究室提供的新闻数据上进行实验，证明了其效果好于卷积神经网络与支持向量机（SVM）、最近邻分类（KNN）、贝叶斯分类（NB）、随机森林分类（RFC）组合的分类模型[19]。

2.1.4Bi-LSTM

近年来，因深度学习具有强大的特征选择和特征抽取能力，能够自动捕获文本更高层次、更抽象和更全面的语义信息，所以其在文本分类研究中得到了广泛应用。其中，循环神经网络（Recurrent Neural Network，RNN）在时间序列数据处理上具有优势，其通过自身网络结构的循环机制，对文本序列整体进行建模和特征提取，可以提高文本的分类效果[9]。作为循环神经网络中的一种——长短期记忆网络（Long Short-Term Memory，LSTM）在 RNN 的基础上增加了神经元中控制信息出入、有记忆功能的处理器——门（gate）机制，解决长短距离依赖的问题，适合处理信息传播过程中间隔和延迟相对较长的信息，提取重要特征。

在文本分类任务中，每个特征词存在序列关系且其表达受上下文不同距离特征词的影响。双向长短时记忆时间递归神经网络（Bi-LSTM）中的每个序列向前和向后分别是两个 LSTM 层，弥补了 LSTM 缺乏下文语义信息的不足[20]。

Bi-LSTM 凭借其在分类任务上优秀的效果，已成为学者们研究分类问题的首选模型。如范昊等使用 Bi-LSTM 和词嵌入对新闻进行分类，取得了较之前研究更好的效果[21]。李超凡等则针对中文电子病历文本分类的高维稀疏性、算法模型收敛速度较慢、分类效果不佳等问题，提出了一种将 BiLSTM 模型与注意力机制结合的病历文本分类模型，有效地提升了文本分类的效果[22]。

2.1.5GRU

随着 LSTM 在自然语言处理特别是文本分类任务的广泛应用，人们逐渐发现 LSTM 具有训练时间长、参数较多、内部计算复杂的缺点。Cho 等人在 2014 年进一步提出了更加简单的、将 LSTM 的单元状态和隐层状态进行合并的、还有

一些其他变动的 GRU 模型。GRU 模型是一种保持了 LSTM 效果、具有更加简单的结构、更少的参数、更好的收敛性的模型。GRU 模型由更新门和重置门两个门组成。前一个时刻的输出隐层对当前隐层的影响程度由更新门控制，更新门的值越大说明前一时刻的隐层输出对当前隐层的影响越大；前一时刻的隐层信息被忽略的程度由重置门控制，重置门的值越小说明忽略得越多[23]。

在最近的研究中，GRU 也越来越多的受到研究者的关注。冼广铭等针对文本中的特征信息不足以及传统模型上下文依赖关系方面不足的问题，提出经过 TF-IDF 加权的词向量和 LDA 主题模型相融合，利用双向门控循环神经网络层（BiGRU）充分提取文本深度信息特征的分类方法，在天池比赛新闻文本分类数据集上取得了优秀的效果[24]。

2.2 命名实体识别

命名实体识别（Named Entity Recognition，NER）是信息抽取和信息检索中一项重要的任务，其目的是识别出文本中表示命名实体的成分，并对其进行分类。汉语的命名实体识别较之英语要更为复杂困难，这主要表现在汉语文本中没有表示词语边界的分隔符号，命名实体的识别效果很大程度受自动分词结果影响，而汉语自动分词的效果往往也受制于命名实体的识别[25]。

目前，NER 任务的主要方法包括：基于规则的方法、基于统计机器学习的方法以及基于深度学习的方法。尤其是近年来，由于使用词向量来表示词语的方法一方面解决了高纬度向量空间带来的数据稀疏问题，另一方面词向量本身也比人工选择的特征包含更多的语义信息。使得深度学习方法成为了 NER 任务的主流方法，主要包括 BERT 预训练语言模型、Bi-LSTM 等。由于 Bi-LSTM 模型在上文已经介绍过，因此下面主要介绍本文将使用的 BERT 预训练语言模型。

BERT 是一种基于微调的多层双向 Transformer 编码器,它的特征表示在所有层中共同依赖于左右两侧的上下文[26]。它能够计算词语之间的相互关系，并利用所计算的关系调节权重提取文本中的重要特征，利用自注意力机制的结构来进行预训练，基于所有层融合左右两侧语境来预训练深度双向表征，比起以往的预训练模型，它捕捉到的是真正意义上的上下文信息，并能够学习到连续文本片段之间的关系，能够解决一词多义问题。该模型融合了其他模型的优点，并摒弃了它

们的缺点，在诸多自然语言处理的后续特定任务上取得了良好的效果[27]。

BERT 模型凭借其在实体识别任务中优秀的结果，已成为研究者使用的最广泛的模型之一。在旅游领域，对 BERT 模型的使用也十分广泛。赵平等为解决旅游文本在特征表示时的一词多义问题，针对旅游游记文本景点实体识别中景点别名的问题，研究了一种融合语言模型的中文景点实体识别模型——首先使用 BERT 语言模型进行文本特征提取获取字粒度向量矩阵，再利用 BiLSTM 与 CRF 提取全局最优序列，最终得到景点命名实体[28]。陈赟等参考 BERT 模型，通过共享 BERT 编码层，提出了一种基于双仿射注意力机制的实体关系联合抽取模型 BAMRel 模型，用于旅游领域的实体和关系联合抽取[29]。

2.3 知识图谱与关联算法

2.3.1 知识图谱

知识图谱是一种结构化的语义知识库，用于以符号的形式描述客观世界中的概念及其相互关系[30]。其以三元组（头实体、关系、尾实体）和实体及其相关属性-值对的形式表现实体及其之间的关系，实体之间以有向的关系相互联结，构成网络状的结构。本质上，知识图谱是一种揭示实体之间关系的语义网络，可以对现实世界的事物及其相互关系进行形式化地描述[31]。

依赖于传统关系型数据库的信息检索方式随着互联网信息总量的指数式增长而渐显乏力。知识图谱最早由谷歌公司提出，其试图借助知识图谱的非关系型数据库结构来构建下一代智能化搜索引擎，解决传统关系型数据库所带来的信息检索困难问题，这是知识图谱最早、最主流的应用领域[32]。随着智能信息服务应用的不断发展，知识图谱也被广泛应用于智能问答、个性化推荐、语义分析、命名实体消歧、信息提取等方面[33]。

现在的知识图谱已被用来泛指各种大规模的知识库，从覆盖范围上来看，知识图谱可以分为通用知识图谱和行业知识图谱。通用知识图谱注重内容的广度，较行业知识图谱来说准确度略低，主要注重于融合更多的实体，应用于智能搜索等领域；行业知识图谱通常需要依靠特定行业的数据来构建，具有特定的行业意义。一般来说，行业知识图谱的实体属性与数据模式相对比较丰富，考虑到不同的业务场景和使用人员[31]。

知识图谱的构建分为自顶向下与自底向上两种模式，如图 2.1。自顶向下是指先定义好知识图谱的本体和数据模式，再将实体加入到知识库中，该构建方式需要利用一些现有的结构化知识库作为基础；自底向上的构建模式应用较多，它是指从一些开放链接数据中提取出实体加入到知识库，再构建顶层本体的模式。本文使用自顶向下的模式构建茂名旅游产品图谱。

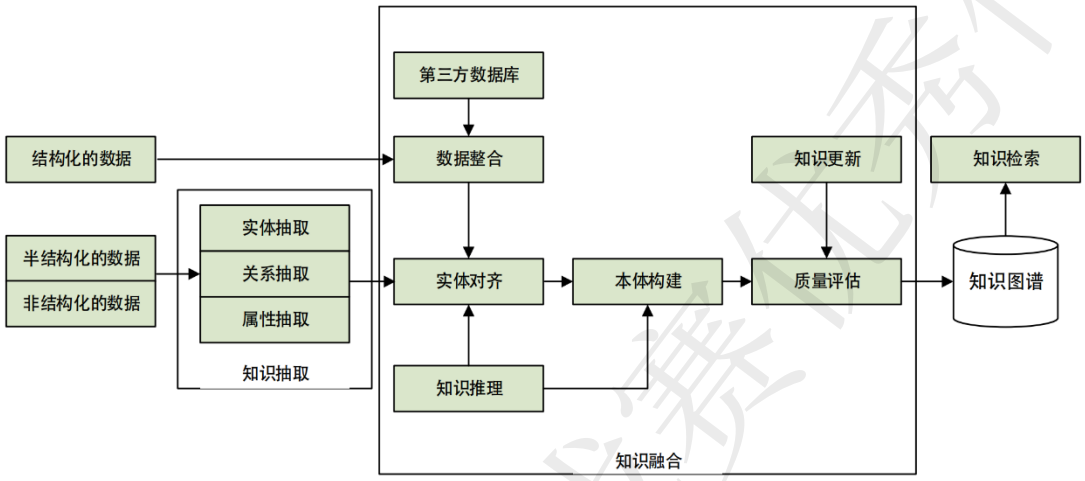

图 2.1 知识图谱的构建模式[33]

2.3.2 Apriori 算法

Apriori 关联规则挖掘算法基于两个原理:频繁项集的子集是频繁项集;非频繁项集的超集是非频繁项集[34]。它利用逐层搜索的迭代方法找出数据库中项集的关系，以形成规则，其过程由连接（类矩阵运算）与剪枝（去掉那些没必要的中间结果）组成。该算法中项集（Itemset）的概念即为项的集合。包含 K 个项的集合为 k 项集。项集出现的频率是包含项集的事务数，称为项集的频率。如果某项集满足最小支持度，则称它为频繁项集[35]。

Apriori 算法的操作步骤如下：

（1）设定最小支持度 s 和最小置信度 c。

（2）Apriori 算法使用候选项集。首先产生出候选的项的集合，即候选项集，若候选项集的支持度大于或等于最小支持度，则该候选项集为频繁项集。

（3）在 Apriori 算法的过程中，首先从数据库读入所有的事务，每个项都被看作候选 1-项集，得出各项的支持度，再使用频繁 1-项集集合来产生候选 2-项集集合，因为先验原理保证所有非频繁的 1-项集的超集都是非频繁的。

（4）再扫描数据库，得出候选 2-项集集合，再找出频繁 2-项集，并利用这些频繁 2-项集集合来产生候选 3-项集。

（5）重复扫描数据库，与最小支持度比较，产生更高层次的频繁项集，再从该集合里产生下一级候选项集，直到不再产生新的候选项集为止。

在此算法中要不断地重复两个步骤：连接和剪枝。具体内容如下：

（1）连接。为找 F_k，通过 F_{k-1} 与自己连接产生候选 k-项集。该候选项集的集合记做 L_k。设 F1 和 F2 是 F_{k-1} 中的项集。执行连接 $F_{k-1}\infty F_{k-1}$，其中 F_{k-1} 的元素 F_1 和 F_2 是可以连接的。

（2）剪枝。L_k 的成员不一定都是频繁的，所有的频繁 k-项集都包含在 L_k 中。扫描数据库，确定 L_k 中每个候选集计数，并利用 F_{k-1} 剪掉 L_k 中的非频繁项，从而确定 F_k。

3 微信公众号文章分类

3.1 问题分析

微信在中国社交媒体市场占据着绝对主流的地位，据 2021 年腾讯财务报告显示，微信月活跃账户数量已经达到 12.682 亿[1]。微信公众号基于微信自身巨大的用户基数，已成为各类机构进行宣传推广的重要平台。微信公众号凭借其多媒体图文推送、互动方便快捷等优势备受用户青睐，为政府、企业、组织、媒体及个人提供了一种新的信息传播方式。

旅游业往往需要广泛的宣传才能为大众所知，吸引游客前来观光消费，而微信公众号依托于其高效便捷的信息传播模式，已成为旅游行业宣传的重要阵地，每天都有大量与文旅相关的内容产出。由于公众号的内容五花八门，对于消费者、旅游相关部门以及各类旅游线路供应商来说，得到与文旅相关的信息是其进行旅游线路设计、消费线路选择的第一步。然而人工分类效率低下，成本高昂，难以达到理想的效果。所以利用自然语言处理的方式对公众号文章进行自动分类，找出与文旅相关的文章很有现实意义。

[1] 数据来源：https://new.qq.com/omn/20220420/20220420A074W700.html

在针对本研究进行文本分类的过程中，我们发现对于判断是否“相关”的概念定义较为模糊，可分为整篇文章主旨为文旅的“强相关”与文章内容部分包含文旅的“弱相关”。例如，在公众号语料库中存在一篇题为《茂名启动台风Ⅲ级应急响应！市旅游局发布防御台风安全提示》的文章，虽然其主题是防范台风，但其中涉及到了大量对游客的游玩建议、景区的旅游安全管理等内容。

综合考虑本项目的四个研究问题之后，我们认为赛题中的四个题目并非独立存在，而是相互关联的，可以将抽取出的与文旅相关的公众号文章加入至游记、各类评论等其他语料中，丰富语料库，并进一步利用这些语料作为旅游产品图谱构建的基础。因此，本团队将问题一中的“相关”概念进行拓展，既包括强相关关系，也包括弱相关关系，以便在后续知识图谱的分析中挖掘产品间更丰富的隐含关联。

操作层面上，将语料库中的大量文章分为“相关”或“不相关”是机器学习中典型的二分类问题。在自然语言处理领域，已有较多成熟的研究模型被用来解决二分类问题，主要可以分为机器学习与深度学习方法。本文选取在大量数据挖掘比赛中被广泛运用、效果强大的模型来解决本研究的问题，包括传统机器学习中的SVM（Support Vector Machines）模型以及XGBoost（Extreme Gradient Boosting）模型；深度学习中善于处理较长文本、能够捕捉较长距离依赖关系的BiLSTM（Bi-directional Long Short-Term Memory）模型以及GRU（Gate Recurrent Unit）模型，分别进行实验，最后对实验结果进行评价比对，采用在本研究问题中表现最优的模型进行后续预测分析。

3.2 数据预处理

公众号文章的表现形式多样、结构复杂，掺杂着大量的图文混合内容，而且语料库数据格式不够规范，无法直接使用。因此，为了能让计算机简单有效地处理文本信息，需要将原始的数据集变为有规则、有结构、有组织的数据，需要进行数据预处理。

3.2.1 人工标注

由于机器学习与深度学习都属于有监督学习，需要有标注的数据来训练。因此，如前所述，本研究在将“相关”的概念进行拓展之后，对公众号文章进行人

工标注。在人工标注过程中，我们先随机抽取了 300 篇公众号，团队三名成员都对这 300 篇公众号进行人工标注，将与文旅相关的公众号文章标注为 1，不相关文章标注为 0，对于三位同学标注的不一致的文章进行讨论投票确定其是否与文旅相关，并制定一致的标注规则，形成文档。最后，大家按照制定的标注规则各自再随机抽取 200 篇公众号进行标注，之后计算得到的 Kappa 系数达到 0.935，证明我们的标注结果较为科学合理。最终共标注 900 篇公众号文章，414 篇相关，486 篇不相关，数据结果相对平衡。标注结果如图 3.1 所示。

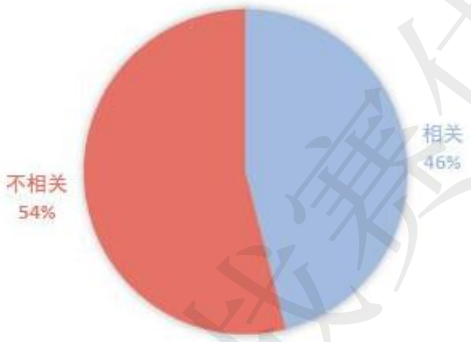

图 3.1 标注结果

3.2.2 数据清洗

在标注公众号的过程中，我们发现公众号文章的标题往往是文章内容的精华提取，包含内容的主要信息。而且公众号文章中存在一些噪声数据，如某些公众号文章内容包含与文旅相关的专有名词，但其主题实际上与文旅无关。

如某些公众号文章内容中存在“市文化广电旅游体育局”一词，但其内容主旨是该单位内部的歌唱活动，与文旅无关，然而，由于文章中包含“旅游”这个关键词，会对模型的预测产生错误的影响。此外，我们发现某些公众号末尾会出现类似“往期精彩”的词，用于推荐其他公众号。这些被推荐的公众号会导致一些与文旅无关的文章中会出现那些与文旅相关文章的标题，给分类带来错误影响。

所以本文进行了以下数据清理工作：

（1）纳入公众号标题。将公众号文章的标题纳入到公众号文章正文中，一起用于识别公众号文章与文旅的相关与否；

（2）清理公众号文章中的噪声文本。本文通过人工阅读大量公众号文章，

识别出在公众号文章中常见的、对后续模型处理影响较大的噪声文本，统一进行清理工作。表 3.1 列举了部分需处理的文本内容。

表 3.1 数据清洗内容示例

文本内容	处理
市文化广电旅游体育局	删除文本
茂名市旅游服务中心	删除文本
茂名市旅游集散中心	删除文本
往期精彩	删除文本及文本后续的内容
茂小布推荐	删除文本及文本后续的内容
咨询电话：	删除文本及文本后续的内容

3.2.3 中文分词

在自然语言处理中，词是能够独立活动的最小语言单位。分词就是通过一定的规则，将原始一段文字序列划分为以词或词语为基本单位的词序列的过程[36]。中文词相较于英文来说更加灵活，且存在各种同音字，比英文分词更加困难。

目前，学术界与企业界常用的中文分词工具包括 jieba 分词、SnowNLP、THULAC 和 NLPIR 等。本文参照曹丹阳等在构建旅游景点本体的过程中使用了 jieba 分词工具并取得良好的分词效果，在本研究中也采用 jieba 分词工具。

jieba 分词工具基于隐马尔科夫模型，相对其他分词工具而言，能够获得更好的分词结果。它的基本思想如下:

（1）基于前缀词典实现词图扫描，根据句子中所有可能成词的情况生成有向无环图；

（2）用动态规划根据词频查找最大概率路径，由此找出基于词频的最大切分组合；

（3）对于未登录词，在 HMM 模型的基础上使用 Viterbi 算法进行计算。

由于可以在 jieba 分词结果中输出词语对应的词性。考虑到本研究的具体问题是判断公众号文章是否与“文旅”相关，而与“文旅”相关与否的判断依据主要是文章中是否存在景区、餐饮、酒店等名词、动名词。因此，为了减少模型输入的复杂度，提高运行效率，本文只提取分词后的名词、动名词输入到后续模型

中，用于模型预测。

3.3 实验设置与结果分析

3.3.1 实验配置

本实验软硬件配置如表 3.2 所示。

表 3.2 实验软件硬件设置

类别	配置
硬件	**CPU:**Intel Xeon Platinum 8260M **操作系统：**Ubuntu 16.04 Linux64 位 **内存：**60G
软件	Python:3.6 Tensorflow:2.7.0

3.3.2 实验步骤与评价指标

本研究采用四种模型来解决研究问题，四种模型的处理过程如下。

（1）LDA+SVM 与 LDA+XGBoost

首先，将上一步预处理过后的名词和动名词词语输入 LDA 模型中，用于提取特征。然后，使用 LDA 构建词模型，包括构建模型、拟合数据以及根据拟合模型转换数据。最后，把 LDA 构建好的词模型分别输入到 SVM 与 XGBoost 分类器中进行分类，针对不同的分类器模型分别使用网格搜索寻找其最优参数。

（2）GRU 与 Bi-LSTM

首先，将分词的结果输入 word2vec 模型构建词向量。Word2Vec 可以通过大量的语料库信息来训练语料库中的文本来获取词向量信息。Word2Vec 包含两种模型，CBOW 与 Skip-gram 模型。CBOW 模型是通过固定窗口内的所有随机初始词向量累加来推断中心词，而 Skip-gram 是通过中心词推断窗口内其他词。

其次，由于不同的词构建的词向量长度不同，为了统一进行运算，本文将词向量中最大的词向量长度作为统一的词向量长度，对于不足的词向量在后面补上 0，以此来把构建好的词向量转化成相同长度的新序列。

接着，划分训练集与测试集。本文参考胡甜甜等人的研究，随机将 80%的样

本划分为训练集，剩下 80%的样本划分为测试集。

最后，分别构建 GRU 与 Bi-LSTM 模型。

（3）评价指标

本文采用的评价指标为准确率（Precision），召回率（Recall）和 F1 值，其计算公式如下：

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{3.1}$$

$$R = \frac{T_P}{T_P + F_n} \times 100\% \tag{3.2}$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{3.3}$$

其中T_P为真正例，即模型识别为正例的实体中实际也为正例的数量。F_P为假正例，即模型识别为正例的实体中实际为反例的数量。F_N为假反例，即识别为反例实际为正例的数量。T_N是真反例，即模型识别为反例的实体中实际也为反例的数量。预测类别见表 3.3。

表 3.3 预测类别

真实情况	预测情况	
	正例	反例
正例	T_P（真正例）	F_N（假反例）
反例	F_P（假正例）	T_N（真反例）

3.3.3 实验结果及分析

本文分类算法均使用相同的语料和相同的配置环境完成的，模型结果对比如表 3.4 所示。

表 **3.4** 分类模型对比

序号	模型名称	准确率 P/%	召回率 R/%	F1 值/%
1	LDA+SVM	0.85	0.83	0.84
2	LDA+XGBoost	0.86	0.86	0.86
3	GRU	0.89	0.91	0.90
4	BiLSTM	0.91	0.91	0.91

由表可知，两种传统的机器学习方法——LDA+SVM 与 LDA+XGBoost 模型

的效果相差不大，两种深度学习方法——GRU 与 BiLSTM 模型的效果也类似。但是深度学习方法的结果要明显优于传统的机器学习方法，在本文使用的两种深度学习方法中，BiLSTM 模型的效果相对更好。因此，本文选取 BiLSTM 模型作为分类模型。

3.4 小结

在确定将 BiLSTM 模型作为本题的分类模型之后，我们将剩余未标注的公众号进行上述预处理过程后传入训练好的 BiLSTM 模型，最终得到与文旅相关的公众号 2217 篇，与文旅不相关的公众号 4069 篇。分类结果如图 3.2 所示。将分类结果写入“result1.csv”。

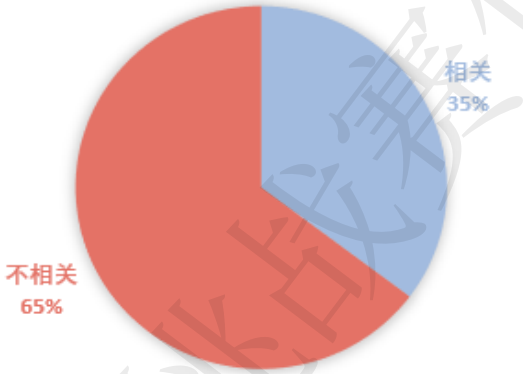

图 3.2 分类结果

4 周边游产品热度分析

4.1 问题分析

本问题要求从附件提供的 OTA、UGC 数据中提取包括景区、酒店、网红景点、民宿、特色餐饮、乡村旅游、文创等旅游产品的实例和其他有用信息，并建立旅游产品的多维度热度评价模型，对提取出的旅游产品按年度进行热度分析。

从文本数据中提取出旅游产品实例等是一个典型的命名实体识别(Named Entity Recognition, NER)问题。解决 NER 问题的方法主要用三种：基于规则的方法、基于特征模板的方法、基于神经网络的方法。

基于规则的方法常用的是利用手工编写的规则，将文本与规则进行匹配来识别出命名实体，在构建规则的过程中往往需要大量的语言学知识，不同语言的识别规则不尽相同，而且需要谨慎处理规则之间的冲突问题；此外，构建规则的过

程费时费力、可移植性不好。

基于特征模板的方法利用大规模语料来学习出标注模型，从而对句子的各个位置进行标注。基于特征模板的方法需要大规模带标签的语料库，往往难以获得，使用成本较高。

基于神经网络的方法即机器学习（Machine Learning，ML）和深度学习（Deep Learning，DL）。神经网络方法使得模型的训练成为一个端到端的整体过程，而非传统的 Pipeline，不依赖特征工程，是一种数据驱动的方法，得到很多研究者青睐。

针对本问题，本文使用在旅游领域中应用广泛、效果较好的解决命名实体识别任务的深度模型——BERT+Bi-LSTM+CRF 模型来解决本问题。

4.2 实验配置

本实验软硬件配置如表 4.1 所示。

表 4.1 实验软件硬件设置

类别	配置
硬件	**GPU**：NVIDIA Tesla P100-16GB **CPU:**Intel Xeon Platinum 8260M **操作系统**：Ubuntu 16.04 Linux64 位 **内存**：60G **显存**：16G
软件	Python:3.6 Tensorflow:1.12.0

4.3 旅游产品识别

4.3.1 数据预处理

此问题的数据预处理过程与第三章的数据预处理过程类似。首先，将游记与第三章中提取出的与文旅相关的公众号的标题加入到正文中，共同作为语料。其次，对加入的公众号进行清洗，删除上一章中定义的噪声文本。

由于 BERT+Bi-LSTM+CRF 模型属于有监督学习，需要在语料中先人工标注

出一些想要提取的类别。因此，本文参考旅游产品的六要素——食住行游购娱来进行标记，具体分为美食店铺（restaurant）、美食（food）、酒店（hotel）、交通方式（transportion）、旅游景区（tourism）、购物（purchase）和娱乐活动（festival）七类。

由于该标注问题较为复杂，我们首先邀请行业专家对团队成员进行培训并请他标注一些语料作为我们标注的黄金标准。随后，参照这些标准，我们随机抽取了 300 篇游记中的 50 篇进行人工标注。本文将 BIO 作为标注规则，其中 B 表示当前位置的字在其所组成的实体中位列首位。而 I 则表示当前位置的字在其所组成的实体中位列非首位的其他位置。O 表示当前位置中不属于任何实体的字或者标点。在此任务中，标签一共分为 15 类，分别是：无所属实体“O”，美食店铺实体开头“B-restaurant”，美食店铺实体其他部分“I-restaurant”；美食实体开头“B-food”，美食实体其他部分“I-food”；酒店实体开头“B-hotel”，酒店实体其他部分“I- hotel”；交通方式实体开头“B-transportion”，交通方式实体其他部分“I-transportion”；旅游景区实体开头“B-tourism”，旅游景区实体其他部分“I-tourism”；购物实体开头“B-purchase”，购物实体其他部分“I-purchase”；娱乐活动实体开头“B-festival”，娱乐活动实体其他部分“I-festival”。标注示例如表 4.2 所示。

表 4.2 标注示例

中文字符序列	市	博	物	馆	迁	至	…
标签序列	B-tourism	I-tourism	I-tourism	I-tourism	O	O	…

4.3.2 实验步骤

首先，本文随机将 70%的样本划分为训练集，20%的样本划分为测试集，剩下 10%的样本划分为验证集。其次，将划分好的数据传入 BERT+Bi-LSTM+CRF 模型，根据结果对参数进行微调，最终使用的参数 BERT 如表 4.3 所示。

表 4.3 BERT 参数设置

版本	**网络层数**	**隐藏层**	**头**	**参数**
BERT-Base	12	768	12	110M

模型其他参数设置见表 4.4。

表 4.4 模型训练参数

参数名称	数值
max_seq_length	128
train_epochs	50
train_batch_size	16
learn_rate	1e-5
dropout_rate	0.5
clip	0.5
bilstm_size	128
warmup_proportion	0.1

4.3.3 实验结果与分析

在实际处理本文的数据时，我们发现交通方式（transportion）、购物（purchase）和娱乐活动（festival）在数据中的数量占比较少，模型无法识别出这三类产品实体。此外，我们还发现对于旅游景区（tourism）一类，可以细分为旅游景区（tourism）和更细粒度的旅游景点（spot）。因此，结合本研究的实际情况，我们把要识别的产品类别调整为美食店铺（restaurant）、美食（food）、酒店（hotel）、旅游景区（tourism）和旅游景点（spot）五类。

模型评价指标选取准确率（Precision），召回率（Recall）和 F1 值，具体模型结果如表 4.5 所示。

表 4.5 模型结果

类型	准确率 P/%	召回率 R/%	F1 值/%
food	77.19%	81.48%	79.28%
hotel	94.19%	96.69%	95.42%
restaurant	87.50%	77.78%	82.35%
spot	75.76%	80.65%	78.12%
tourism	90.79%	94.95%	92.83%
综合	89.16%	92.37%	90.74%

可见，各类型以及整体的 F1 值都在 78%以上，最高的 F1 值可以达到 95%，

模型结果较好。因此本研究将该训练好的模型用于旅游产品识别，最终识别出68366个产品实体，识别结果样例如表4.6所示。

表4.6 识别结果样例

语料ID	提取实体	类型
公众号文章3009	3	spot
公众号文章4574	##hp##v疫苗	spot
景区评论2058	)、竹海	spot
餐饮评论1253	大爱	tou
公众号文章1479	、	spot
餐饮评论1617	黑森林	tou
公众号文章4498	、婚纱摄影基	spot
公众号文章4395	放鸡岛	tou
公众号文章4395	放鸡岛旅游度假区	tou

从表中可见，提取出的许多产品实体是错误实体，需要对这些提取出的错误实体进行删除。本文按产品出现的次数来判断该产品是否属于错误产品实体，将次数的阈值分别设置为1-10之后，发现将阈值设置为5的效果最好，即将出现次数少于5的产品实体当作错误产品实体进行删除。并且还有一些实体名称不一样，但是指的是一个意思，需要进行实体对齐工作。本文使用余弦相似度算法与人工判断相结合的方法来进行实体对齐工作。实体对齐样例如表4.7所示。在进行对齐的过程中，我们发现在旅游景区（tourism）中出现了诸如“文化节”等娱乐类型的实体。因此，本文新增乡村游和文化艺术（festival）一类，并手动将这些实体增加至乡村游和文化艺术（festival）类。

表 4.7 实体对齐样例表

实体	相似实体
维也纳酒店(茂名沿江大厦店)	维也纳 3 好酒店(茂名沿江大厦店)
温德姆酒店	茂名温德姆至尊酒店 浪漫海岸温德姆酒店 茂名浪漫海岸温德姆酒店
放鸡岛	放鸡岛旅游度假区
古郡水城景区	古郡水城 高州古郡水城

数据处理之后最终删除了 27679 个产品实体，剩余 40286 个产品实体。各数据类型数量分布如图 4.1 所示。

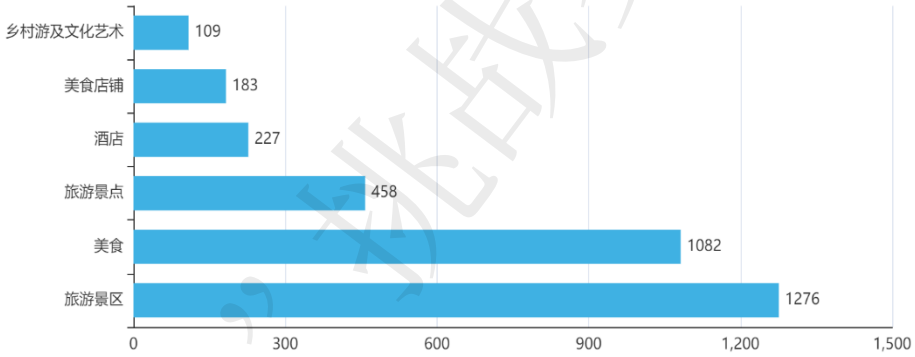

图 4.1 各数据类型数量分布图

在为每个产品分配 ID 后，将结果存于“result2-1.csv”，数据样例如表 4.8 所示。

表 4.8 产品提取数据样例

语料 ID	产品 ID	产品名称
公众号文章-6766	ID0001	大坡镇
公众号文章-6934	ID0001	大坡镇
公众号文章-2798	ID0002	高州粤龙山
公众号文章-3228	ID0002	高州粤龙山

4.4 产品热度分析

4.4.1 定义热度指标

“热度”是公众的观点、话题或某一关键词所受的关注程度。从信息论角度理解，热度是文本所包含的信息量，通过搜索引擎指数来衡量。网络舆情中的热度通过媒体资讯中所能检索到的关键词条目数量来反映，它体现了某话题的受关注程度。综上，热度可以体现为某个属性提升与集中的程度，是“量”的表现。

旅游学术界在使用“热度”一词时，基本沿用了舆情分析中“热度”基本内涵，特别在区域旅游的热度研究中，表现的很明显。旅游景区相关研究成果中的“热度”，主要是沿用信息论视角定义，代表性的有“游记文本集合中所体现的游客对景区的偏好程度”，“景区受游客的关注情况”，通过关键词及文本数量来测量；也有引申为“突出程度”的，代表性的有“旅游资源使用度”，通过旅游客流量来衡量这一指标，由旅游景区的接待人数来表征“热度”。

具体到本题现有的评论、游记及公众号数据。由于与文旅相关的公众号往往是政府单位、旅行社以及较为有名的头部博主来经营，其受众更广，更新更快，且人们浏览公众号的频率比游记、评论的频率更高。因此本文将从公众号中提取出的产品赋予更高的权重。

本文综合考虑产品频次与产品出现的文章来源类别，将产品热度公式定义为：

$$H = \frac{0.2 * \alpha + 0.8 * (0.3 * \beta + 0.3 * \gamma + 0.4 * \delta)}{\varepsilon} \tag{4.1}$$

产品热度公式符号定义如表 4.9 所示。

表 4.9 热度公式符号定义

符号	定义
H	产品热度
α	产品该年出现的总次数
β	包含该产品该年的评论数
γ	包含该产品该年的游记攻略数量
δ	包含该产品该年的公众号文章数量
ε	该年所有产品的数量

4.4.2 热度计算

在定义了产品热度计算公式之后，由于热度数值相差较大，本文将其进行了归一化，归一化公式如下：

$$H = \frac{H - \min(H)}{\max(H) - \min(H)} \tag{4.2}$$

最后，将结果存于“result2-2.csv”，2018 年热度数据样例如表 4.10 所示。

表 4.10 2018 年旅游产品热度数据样例表

产品名称	出现频次	热度
荔枝	109	1
蛋糕	110	0.85
茂名浪漫海岸国际旅游度假区	84	0.77
御水古温泉	79	0.72
清香面包店（车田街店）	87	0.67
放鸡岛	75	0.67
海鲜	78	0.67
果之度（为民店）	76	0.59

2019 年热度数据样例如表 4.11 所示。

表 4.11 2019 年旅游产品热度数据样例表

产品名称	出现频次	产品热度
茂名浪漫海岸国际旅游度假区	153	1
荔枝	108	0.77
第一滩	98	0.64
放鸡岛	98	0.63
御水古温泉	78	0.53
中国城	68	0.49
CAKE 情迷黑森林	78	0.47
海鲜	74	0.47

“泰迪杯”挑战赛优秀作品

2020 年热度数据样例如表 4.12 所示。

表 4.12 2020 年旅游产品热度数据样例表

产品名称	出现频次	产品热度
荔枝	310	1
北部湾城	216	0.7
冼夫人好心文化	189	0.61
第一滩	210	0.61
茂名浪漫海岸国际旅游度假区	190	0.57
Hello 炸鸡（方兴店）	143	0.39
放鸡岛	116	0.34
御水古温泉	111	0.33

2021 年热度数据样例如表 4.13 所示。

表 4.13 2021 年旅游产品热度数据样例表

产品名称	出现频次	产品热度
荔枝	2576	1
分界镇	204	0.08
相聚时光（化州店）	199	0.07
茂名浪漫海岸国际旅游度假区	192	0.07
草原郎烤羊	200	0.07
友情有意音乐餐厅	200	0.07
罗非鱼	181	0.07
椒王火锅（电白店）	200	0.07

5 旅游产品关联分析与旅游产品图谱构建

5.1 问题分析

根据语料库的结构可以发现，酒店、餐饮、景区评论是语料的主要组成部分，三者之间的关联是所有旅游产品关联间的核心关系，除此之外，本文在第四章中使用命名实体识别提取到的其他旅游产品也可以作为补充进行关联。通过识别同

一条评论（同一篇游记攻略、同一篇公众号文章）中共同存在的实体，来判断实体之间存在着关联关系。在酒店、餐饮、景区评论中，存在着大量的方面级的评价，本文认为这些方面级的评价可以刻画到本地旅游图谱中，作为旅游产品实体的一个属性关系。

综上，本文从两个方面刻画知识图谱：

（1）以 Apriori 算法提取出上文所提取出实体之间所隐含的关联规则。具体来说，依据本文在第四章周边游产品热度分析中通过命名实体识别所提取出的旅游产品实体，通过 Apriori 算法计算在每条评论、每篇游记及每篇文旅相关的公众号文章中共同出现的旅游产品实体来挖掘它们潜在的关联。

（2）利用依存句法分析，找到对酒店、餐饮、景区三种实体评价的方面级内容，作为实体的属性关系加入到知识图谱中。通过对酒店、餐饮、景区的评论进行依存句法分析，并筛选出具有实际含义的 S_V_O（主谓宾三元组），找到有实际含义的主语-宾语组合加入到实体的关系中，作为对实体评论的可视化分析。

5.2 关联分析

给定两个实体r_i和$r_j(i \neq j)$，他们之间的置信度为：

$$conf(r_i \rightarrow r_j) = \frac{|r_i \cap r_j|}{|r_i|} \tag{5.1}$$

支持度为：

$$supp(r_i \rightarrow r_j) = \frac{|r_i \cap r_j|}{|r_i| + |r_j|} \tag{5.2}$$

其中，$|r_i|$代表包含实体的r_i的节点的数量，$|r_i \cap r_j|$代表同时包含实体r_i和r_j的节点的数量。置信度的含义是指在所有出现了实体r_i的节点中，同时出现一条实体r_j的概率。由于实体之间的置信度是不对称的，即通过置信度计算得到的初步关联关系是有向的，因此需要将两个方向的不同置信度值进一步处理得到一个统一的关联度值。参照F_1值的定义方法，定义两条实体之间的关联度为：

$$R(r_i, r_j) = \frac{2 \times conf(r_i \rightarrow r_j) \times conf(r_j \rightarrow r_i)}{conf(r_i \rightarrow r_j) + conf(r_j \rightarrow r_i)} \tag{5.3}$$

由此我们将置信度公式中两个实体有向关系转化为统一的双向关系，并以

$R(r_i, r_j)$作为本文的关联度值进行判断。

本文以公式 5.3 作为依据，提取每一条评论、每一篇游记攻略以及每一篇文旅相关公众号文章出现的产品作为一条记录，将其输入关联分析模型。以游记 ID 为 1280 的一篇游记攻略为例，如表 5.1 所示，游记 1280 中提到了 6 个旅游产品，则将这 6 个产品作为一条记录输入，以此类推，得到前文所提取出实体的所有数据，进行关联度计算。

表 5.1 游记攻略-1280

产品 ID	产品名称
ID1322	灯楼角
ID0164	海口埠
ID0105	开平碉楼
ID0787	茂名市露天矿生态公园
ID0297	牙象大地艺术公园
ID0439	晏镜岭

最终得到了 1919 条关联关系，涉及到 673 个不同实体之间的相互关系。考虑到置信度过低的关系并没有参考意义，我们将最低置信度要求设置为 0.3，剔除了置信度小于 0.3 的实体关系，最后得到的关联度分布如图 5.1 所示。

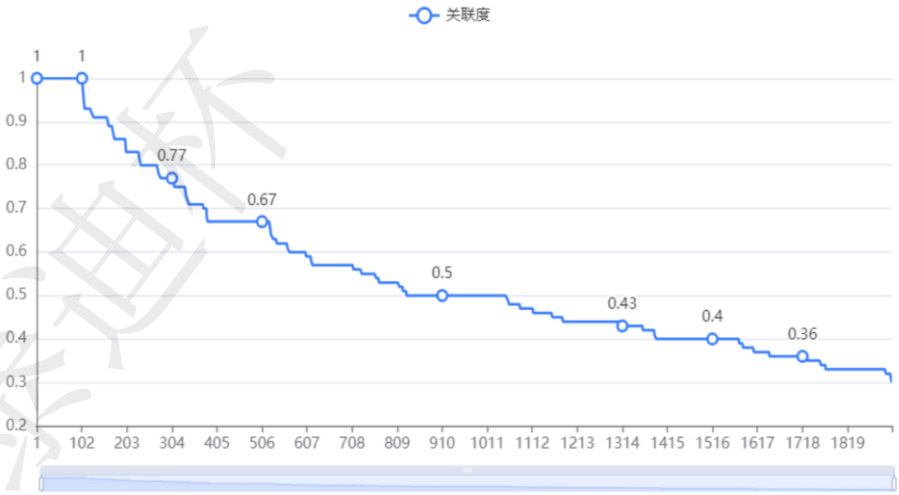

图 5.1 关联度分布折线图

将有关系的实体对应的两个产品 ID、关联度以及实体之间的关联类型导入

“result3.csv”文件中，就得到了赛题第三题所需要的输出结果，表 5.2 展示了一小部分结果数据的样式。

表 5.1 数据结果样例

产品 ID	产品 ID	关联度	关联类型
ID0010	ID1032	0.91	旅游景点-旅游景点
ID0026	ID0159	0.91	美食-美食
ID0649	ID0987	0.91	旅游景区-旅游景区
ID0817	ID0895	0.77	乡村游及文化艺术-旅游景点
ID0101	ID0262	0.71	旅游景区-酒店

5.3 旅游产品评价属性提取

本文选择百度开源的依存句法分析系统 DDparser 对酒店、餐饮、景区评论文本进行语料分析。如表 5.3 所示，依存句法分析可以从文本中提取出七种精炼的句法结构或句子中的词汇之间的依存关系。

表 5.3 依存句法结构

标签	数据格式	实例	说明
S_V_O	（主语，谓语，宾语）	（我，吃，鱼）	主谓宾结构
V_CMP	（核心词，补语）	（吃，多久）	动补结构
ADV_V	（修饰词，核心词）	（不能，吃）	动词及其修饰词
ATT_N	（修饰词，核心词）	（美国，首都）	名词及其修饰词
F	（核心词，方位词）	（7 点，后）	方位词限定核心词
DOB	（主谓语，谓语，间接宾语，直接宾语）	（我，给，他，吃）	双宾语结构
Phrase	（核心词）		一个词作为整体

本文需要的关系是每条评论对应实体各个方面的评价，例如景区评论中可能对对应实体的环境进行了评价，表 5.4 展示了对一个评论进行依存句法分析的结果示例，加粗的部分出现了我们需要的方面级评价对象“环境”以及评价“不错”，是我们可以利用的部分。

表 5.4 依存句法分析结果示例

句法分析结果示例	[[**['环境/n', '不错/a', None], 'SVO']**, [['里面/f', '很大/a', None], 'SVO'], [['水质/n', '干净/a', None], 'SVO'], [['很/zg', '干净/a'], 'ADV_V'], [[None, '值得/v', '去/v'], 'SVO'], [['一/m', '去/v'], 'ADV_V']]

阅读观察了大量依存句法分析的结果后，我们认为表 5.3 所展示的七种句法结构中，只有 S_V_O 结构中的大部分内容是有实际意义的，可以找到有关方面级别的评价。我们筛选出有意义的 S_V_O 结构的分析结果，并与其所对应评论的旅游产品进行匹配，最后找到酒店、餐饮、景区三类实体评论中的方面级评价。表 5.5 展示了三个评价提取结果的样例。

表 5.5 旅游产品属性提取样例

评论 ID	产品名称	产品 ID	属性		
酒店评论 1003	茂名永利之家	ID0511	环境-不错	卫生-好	性价比-很高
餐饮评论 1949	元晟坊蛋糕（南香公园店）	ID1230	环境-好	价格-美丽	环境-适合
景区评论 2035	广垦（茂名）国家热带农业公园	ID1301	环境-好	品种-多	交通-便捷

5.4 旅游产品图谱构建

本文使用 NOSQL 图形数据库 Neo4j 制作知识图谱。

（1）首先，将第四章命名实体识别所得到的旅游产品实体作为节点“Entity”，并设置产品 ID、产品名称（name）、产品类型三个属性，初步得到旅游产品图谱中所有的节点。

（2）设置关联分析算法所得到的 1919 条关系名为“RELATE_TO”，并设置关联度、关联类型两个属性。得到初步的旅游图谱如图 5.2 所示：

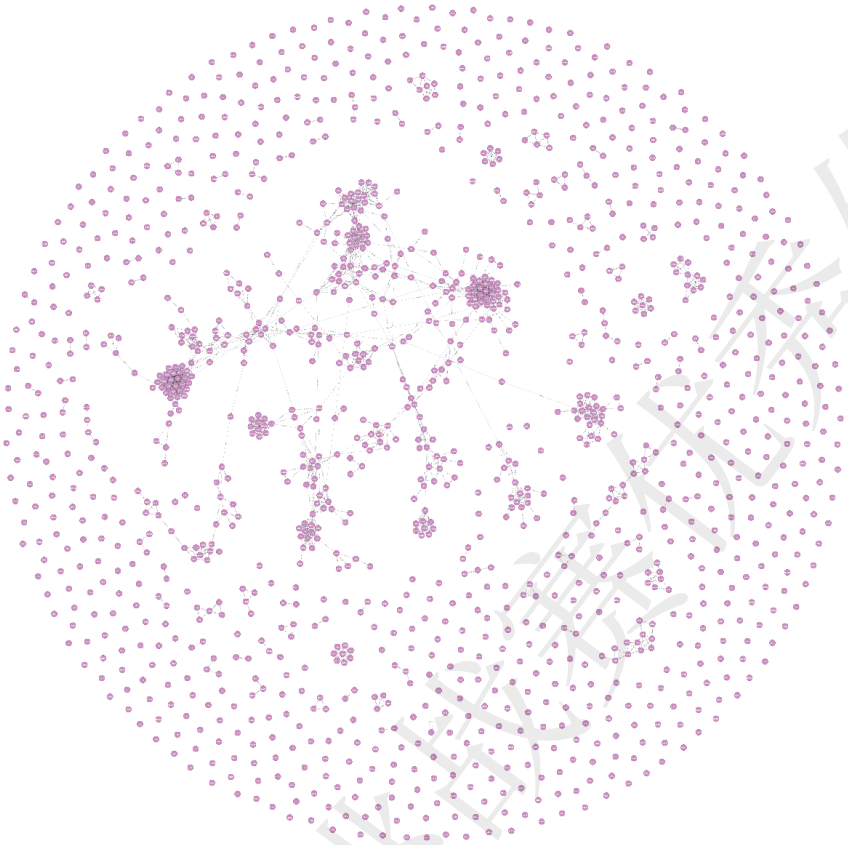

图 5.2 旅游产品图谱总览

旅游产品图谱局部如图 5.3 所示：

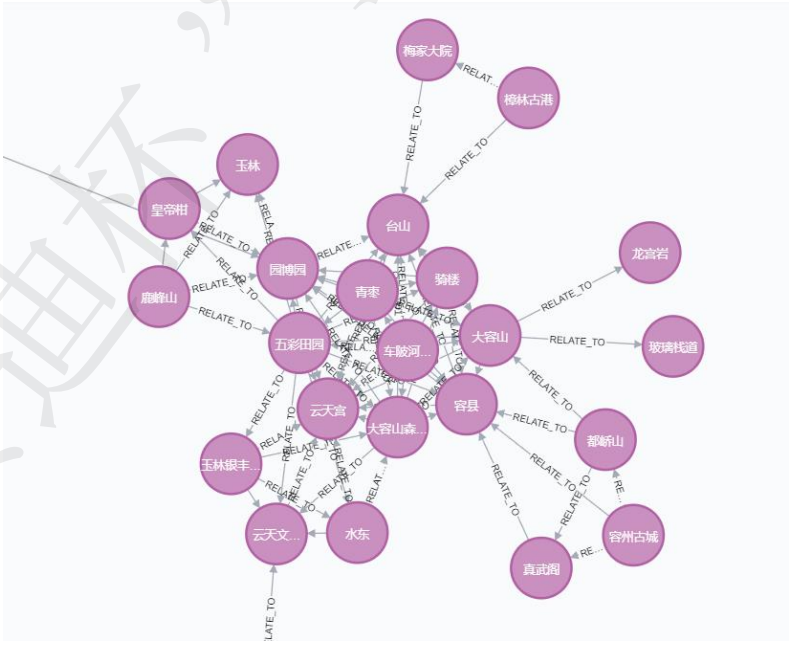

图 5.3 旅游产品图谱局部图

（3）将依存句法分析中所筛选出的酒店、餐饮、景区实体对应评论中的方面级评价导入到旅游产品图谱中，将实体的评价方面作为关系名，评价内容作为节点名在知识图谱中表现出来，如图 5.4 是汉宫传承隆江猪脚饭（上排桐店）的方面评价可视化结果：

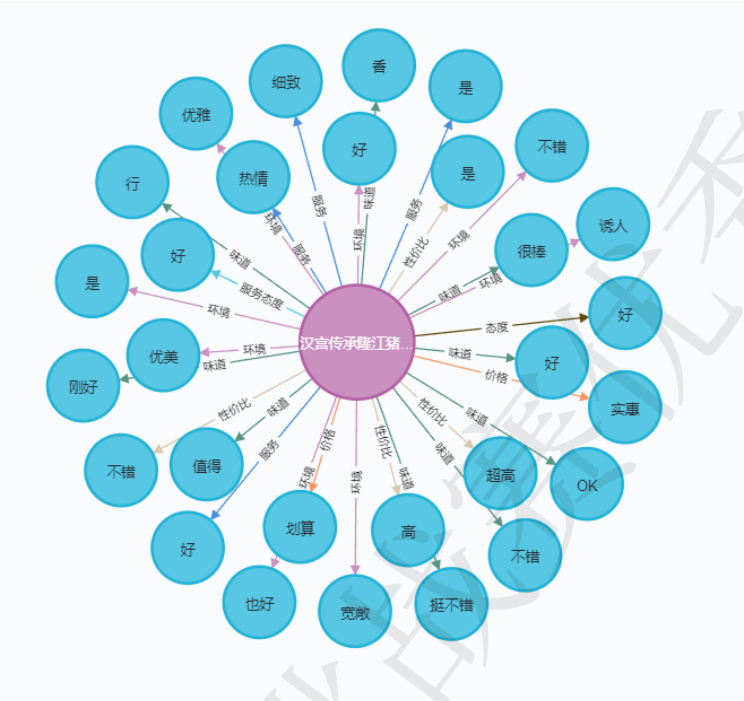

图 5.4 汉宫传承隆江猪脚饭（上排桐店）方面评价可视化

6 旅游产品需求变化分析及政策建议

6.1 需求变化分析

为了探究新冠疫情前后茂名旅游产品的变化。本文首先通过网络爬虫，在百度地图上爬取前文从游记等内容中提取出的产品实体的经纬度坐标，将爬取到的坐标与产品热度相结合，使用 ArcGIS 软件绘制旅游产品热度分布地图，以期直观地探究新冠疫情前后茂名旅游产品的变化。2018 至 2021 年的茂名旅游产品热度分布地图如 6.1 所示。

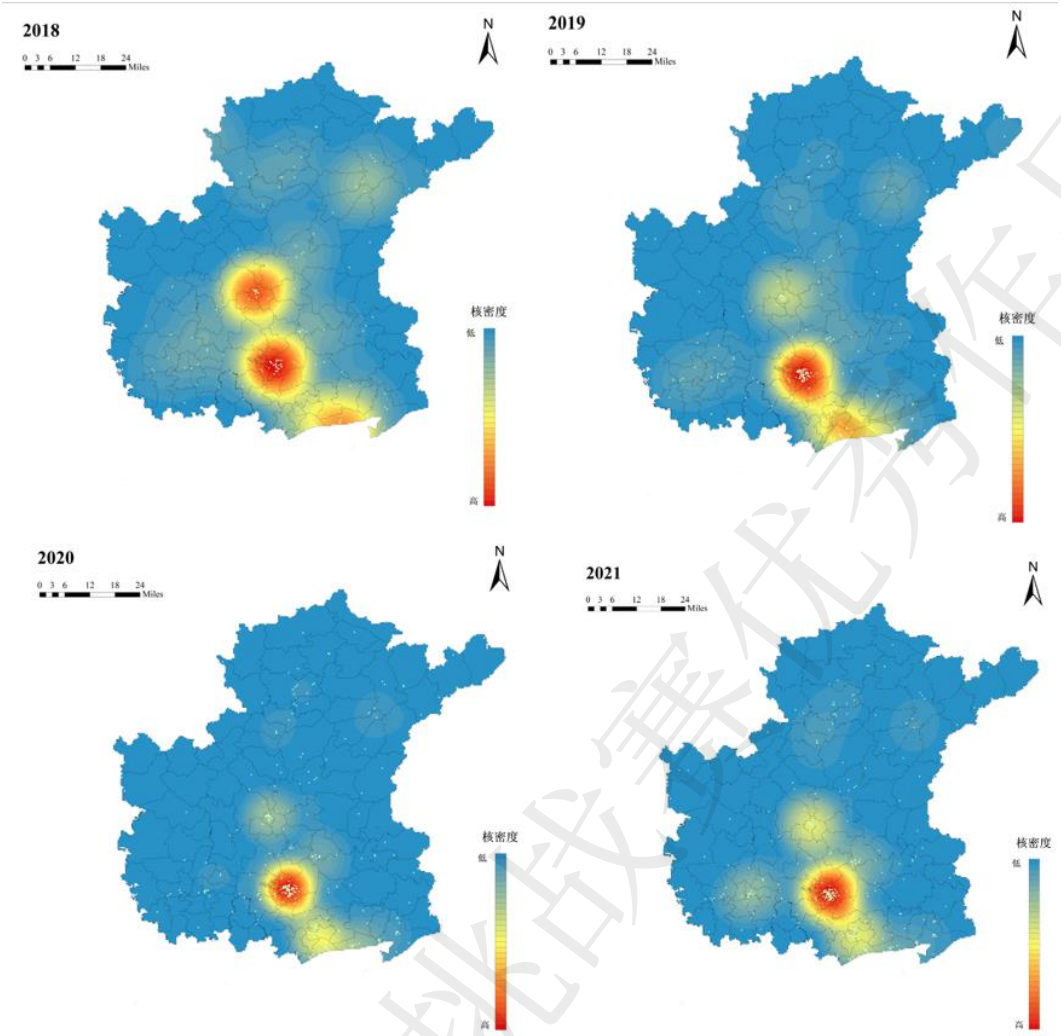

图 6.1 2018 至 2021 年的茂名旅游产品热度分布地图

从图中可以直观的看出，首先，2020 年较 2019 年，茂名旅游热度区域明显减少，且直到 2021 年也没能恢复至 2019 年的水平，说明新冠疫情对茂名旅游产业的冲击巨大且至今未能完全恢复。其次，2021 年的热度图较 2018 年有明显回升，且与 2019 年相似度较高，说明茂名市在疫情之后出台的各项对旅游行业的扶持政策是有效的。第三，从 2018 年和 2019 年的热度图可以看出茂名旅游市场呈现出集聚发展趋势，从多核心旅游市场向双核心甚至单核心市场转变。最后，从图中可以看出 2021 年较 2018 年，地图上出现的旅游产品更分散，说明在后疫情时代，相较于热门景点，游客们更倾向于分散的周边游。

除了旅游产品热度分布地图之外，本文根据前文提取出的产品实体及其热度，分别绘制了 2018 年至 2021 年的产品实体词云图，如图 6.2-6.5 所示。

图 6.2 2018 年旅游产品热度词云图

图 6.3 2019 年旅游产品热度词云图

图 6.4 2020 年旅游产品热度词云图

图 6.5 2021 年旅游产品热度词云图

从图中可以发现，荔枝、第一滩等产品在绝大多数年份都占据了相当重要的地位，而这两者正是茂名市的特色产品与独有景区，因此，茂名市应坚持发展这些特色产业。

最后，本文就提取出的旅游产品来源进行了分析，分析结果如表 6.1 所示。

表 6.1 旅游产品来源

来源 / 年份	餐饮评论	酒店评论	景区评论	游记攻略	公众号
2018	1108	31	408	108	2561
2019	709	418	492	925	2172
2020	2213	433	507	1434	4949
2021	6708	378	543	913	13276

从表中可以看出，在 19 年之后，从公众号、餐饮评论中提取出的旅游产品数量急剧上升，说明了在后疫情时代，对于旅游产品的宣传更多的依赖于公众号等互联网进行开展，发力互联网的宣传阵地对于疫情期间的旅游产品宣传非常重要。

6.2 政策建议

尊敬的茂名市文化广电旅游体育局的领导们，您们好！至今，新冠疫情仍然在各地偶有发生，严重影响了人民群众的正常工作与生活，给各地的经济带来了不小的困境。我们团队基于 2018 年至 2021 年的茂名市酒店、餐饮、景区评论，游记攻略以及微信公众号文章数据，对其中出现的茂名旅游产品进行了提取、热度计算、结果分析并构建知识图谱，希望从中发掘出茂名市旅游业潜在的需求变化与发展脉络，给出能够支持茂名市旅游经济发展的建议。现将团队工作结果以及相关建议汇报如下。

首先，基于消费者的在线评论数据，分析得到了茂名市旅游产品间的关联关系，为茂名市的旅游产品组合设计提供了方向。具体而言，我们基于茂名旅游产品的关联关系构建了旅游产品图谱，并从中挖掘出了一些具有较强关联关系的旅游产品集合，包括具体的景区-酒店、景区-餐厅关系等。由于这些产品及其关系是我们从 UGC（用户生成内容）和 OTA（在线旅游）数据中挖掘的，因此这些强关联的产品组合是被消费者和市场所认可的。对此，我们建议贵单位应主动引导旅行社等旅游产品供应商在设计产品路线时，将关联性强的产品进行组合打包以满足消费者的需求，挖掘潜在的旅游路线，积极给予政策扶持，促进茂名市旅

游行业快速恢复。

其次，我们将产品地理坐标与热度相结合，使用 ArcGIS 软件绘制 2018 年至 2021 年茂名市旅游产品热度分布地图，通过对比这四年间的产品热度分布及演化，为茂名市的景点规划及布局提供了思路。第一，我们发现新冠疫情对茂名市旅游产业的冲击巨大且至今未能完全恢复，因此，**有关部门应该在满足疫情防控需要的基础上，采取更加积极主动的政策，刺激旅游业更快的恢复**。第二，我们发现 2021 年的热度图较 2018 年有明显回升，且与 2019 年相似度较高，这说明茂名市在疫情之后出台的各项对旅游行业的扶持政策是有效的，这些**积极有效的政策应该继续坚决地贯彻下去，以达到更好的效果**。第三，从地图中可以看出茂名市存在三个旅游热点中心：茂名市区、电白区和高州市。**我们建议贵单位总结这些热点中心的旅游发展经验，充分发挥热点中心以点带面的辐射作用，带动茂名市旅游行业的整体协调发展。**第四，我们从 2018 年和 2019 年的热度图中发现茂名市旅游市场呈现出集聚发展趋势，从多核心旅游市场向双核心甚至单核心市场转变。我们认为过度的聚集效应在疫情时代的旅游行业是不可取的，**贵单位应有意引导茂名市旅游成多中心的发散态势，分散疫情风险的同时带动更多地方经济的快速复苏**。这一点从 2021 年和 2018 年的图中可以得到验证，2021 年地图上出现的旅游产品更分散，说明在后疫情时代，相较于热门景点，游客们更倾向于分散的周边游。因此，我们建议贵单位不仅要争取各方资源，为民众建设更多的周边公园，满足人们对于分散的周边游的需求；而且需要与旅行社等旅游服务提供机构积极沟通，指导他们更多拓展周边游市场。

再次，我们根据前文提取出的产品实体及其热度，分别绘制了 2018 年至 2021 年的产品实体词云图，为茂名市旅游线路中特色产品的设计提供了参考。在词云图中，我们发现茂名市的特色产品与独有景区等始终占据了非常重要的位置，说明大家对于这些含有茂名特色的产品非常喜爱，接受程度很高。因此，**我们建议贵单位积极引导，大力扶持这些特色产业**，争取把这些特色产业发展为茂名市的优势产业，打造茂名市的特色名片，成为大家心目中的首选地点。

然后，通过对提取出的旅游产品来源及内容进行分析，发现在 19 年疫情爆发之后，从公众号、餐饮评论中提取出的旅游产品数量急剧上升，这为旅行商选

择推广旅游产品的渠道提供了方向。研究结果表明在疫情之后，互联网已成为旅游消息发布的重要阵地，很多与旅游相关的消息都通过互联网转发给大众。因此，我们建议**贵单位更加积极主动地利用公众号等互联网平台来发布一手的相关消息**，对茂名市的旅游产品进行积极宣传，提升茂名旅游的知名度。而且从餐饮评论的涨幅要远远高于酒店评论和景区评论的涨幅可以发现，相比于进行外出旅游，人们对于饮食的需求增长的更多更快，餐饮行业是最先从疫情的打击中恢复的行业。因此，我们建议**贵单位可以更多的介绍茂名市的特色美食，指导餐饮机构如何防范疫情，在疫情防控常态化的情况下更好的满足消费者反弹的需求**。

最后，我们团队在从百度地图整理产品经纬度坐标时，发现很多旅游产品的景点位于茂名市周边城市，如湛江、阳江、玉林、台州、云浮、梧州、来宾，佛山、桂林、清远、广州、河源、崇左等。因此，我们建议贵单位在满足疫情防控要求的前提下，指导旅行社考虑和茂名市的周边热点产品相结合。

7 总结与展望

7.1 论文主要工作总结

本论文的主要研究工作是从酒店、餐饮、景区评论以及游记攻略、微信公众号文章提取出旅游产品的实体、热度以及关联度分析并进行知识图谱的刻画，最终通过分析数据形成对茂名市政府旅游业发展的政策建议。研究工作主要分为四个部分，第一部分首先是人工对 2018-2021 年间的 6286 篇微信公众号文章进行标注，判断文章是否与文旅相关；之后通过两个机器学习算法 SVM 与 XGBoost 以及两个深度学习算法 Bi-LSTM 与 GRU 对预处理后的公众号文章语料进行测试，综合考虑准确率、召回率和F_1值，选择最优的 Bi-LSTM 算法对剩余的微信公众号文章进行预测，最后识别出 2217 篇与文旅相关的微信公众号文章。

第二部分为了获取评论、游记攻略以及公众号文章中的相关旅游产品实例，本文采取命名实体识别的方法，构建 BERT+Bi-LSTM+CRF 模型从上述三类语料中提取旅游产品的实体，并在实体对齐的基础上，通过构建以频次为基础的实体热度计算公式计算旅游产品实体每年的热度值。

第三部分是设计并制作茂名旅游产品图谱。主要工作是利用 Apriori 算法计

算第二部分提取到实体之间的关联度，在此基础上制作知识图谱；此外，本文使用依存句法分析获取了酒店、餐饮、景区三类实体评论中的评价方面-评价二元组，将评价方面作为关系名，评价作为节点名绘入知识图谱当中，一定程度上完善知识图谱的广度。

第四部分是前三部分所得到的结果的分析和针对性的政策建议。为了保证政策建议的科学性、合理性与针对性，我们结合第二部分茂名旅游产品实体热度的结果，使用网络爬虫获取了美食店铺、酒店、旅游景点、旅游景区四类实体的经纬度坐标，结合热度值在 ArcGIS 软件中以年度为单位绘制产品热度分布地图，试图找到旅游地点热度在四年间的变化规律。

7.2 今后研究工作展望

本论文主要以 UGC、OTA 语料为研究对象，主要研究疫情前后茂名旅游产品的一系列特征。虽然各部分的实验效果都比较理想，但由于时间限制以及团队人员限制，本文也存在一些可以深入研究和改进的地方，下面将阐述一些研究的难点及下一步的研究工作。

引入情感分析完善方面级评价的提取工作。可以利用情感极性量化旅游产品实体的评价，准确的捕捉旅游产品的质量信息。

把文本挖掘方法应用到不同的领域中。随着 Web3.0 时代的接近，各种平台的崛起，会产生越来越多的用户生成内容。这就需要自然语言处理工作者们与时俱进，提出更好的解决文本挖掘分析的方法。

提取更复杂、细颗粒度的实体关系。用深度学习从中文文章中提取出头实体-关系-尾实体三元组是一项困难耗时的工作，需要大量的人工标注语料、高性能的硬件支撑和时间。基于深度学习的关系抽取能够刻画出更加具有分析意义的知识图谱，从而挖掘隐含的旅游产品关联信息，促进地方旅游业的发展。

参考文献

[1] 唐培, 何建民, 冯学钢. 文化冲突对中国入境旅游需求的影响 [J]. 地理科学, 1-9.

[2] M N P, LUCILA O-M, W C W. Natural language processing: an introduction [J]. Journal of the American Medical Informatics Association : JAMIA, 2011, 18(5):

[3] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展 [J]. 软件学报, 2006, 09): 1848-59.

[4] 贾澎涛, 孙炜. 基于深度学习的文本分类综述 [J]. 计算机与现代化, 2021, 07): 29-37.

[5] JELODAR H, WANG Y L, YUAN C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey [J]. Multimed Tools Appl, 2019, 78(11): 15169-211.

[6] 吴江, 侯绍新, 靳萌萌, 等. 基于 LDA 模型特征选择的在线医疗社区文本分类及用户聚类研究 [J]. 情报学报, 2017, 36(11): 1183-91.

[7] 宁宁. 基于 LDA 和句向量的文本分类研究 [D]; 天津理工大学, 2021.

[8] 付静, 龚永罡, 廉小亲, 等. 基于 BERT-LDA 的新闻短文本分类方法 [J]. 信息技术与信息化, 2021, 02): 127-9.

[9] 胡吉明, 付文麟, 钱玮, 等. 融合主题模型和注意力机制的政策文本分类模型 [J]. 情报理论与实践, 2021, 44(07): 159-65.

[10] 康宸, 郑山红, 李万龙. 融合 LDA 主题模型和二维卷积的短文本分类 [J]. 计算机应用与软件, 2020, 37(11): 127-31+53.

[11] 郑腾, 吴雨川. LDA 特征扩展的多类 SVM 短文本分类方法研究 [J]. 武汉纺织大学学报, 2019, 32(02): 72-6.

[12] 段友祥, 张晓天. 基于主动学习的 SVM 评论内容分类算法的研究 [J]. 计算机与数字工程, 2022, 50(03): 608-12.

[13] 卢增祥, 李衍达. 交互支持向量机学习算法及其应用 [J]. 清华大学学报(自然科学版), 1999, 07): 93-7.

[14] 丁勇, 秦晓明, 何寒晖. 支持向量机的参数优化及其文本分类中的应用 [J]. 计算机仿真, 2010, 27(11): 187-90.

[15] 李程雄, 丁月华, 文贵华. SVM-KNN 组合改进算法在专利文本分类中的应用 [J]. 计算机工程与应用, 2006, 20): 193-5+212.

[16] 巩知乐, 张德贤, 胡明明. 一种改进的支持向量机的文本分类算法 [J]. 计算机仿真, 2009, 26(07): 164-7.

[17] 陈振宇, 刘金波, 李晨, 等. 基于 LSTM 与 XGBoost 组合模型的超短期电力负荷预测 [J]. 电网技术, 2020, 44(02): 614-20.

[18] 叶丹, 李智, 王勇军. 基于 SPLDA 降维和 XGBoost 分类器的行为识别方法研究 [J]. 微电子学与计算机, 2019, 36(06): 35-9.

[19] 龚维印, 王力. 基于卷积神经网络和 XGBoost 的文本分类 [J]. 通信技术, 2018, 51(10): 2337-42.

[20] 李文慧, 张英俊, 潘理虎. 改进 biLSTM 网络的短文本分类方法 [J]. 计算机工程与设计, 2020, 41(03): 880-6.

[21] 范昊, 何灏. 融合上下文特征和 BERT 词嵌入的新闻标题分类研究 [J]. 情报科学, 1-8.

[22] 李超凡, 马凯. 基于注意力机制结合 CNN-BiLSTM 模型的电子病历文本分类 [J]. 科学技术与工程, 2022, 22(06): 2363-70.

[23] 王伟, 孙玉霞, 齐庆杰, 等. 基于 BiGRU-attention 神经网络的文本情感分类模型 [J]. 计算机应用研究, 2019, 36(12): 3558-64.

[24] 冼广铭, 王鲁栋, 曾碧卿, 等. 基于 LDA 和 BiGRU 的文本分类 [J]. 计算机技术与发展, 2022, 32(04): 15-20.

[25] 刘浏, 王东波. 命名实体识别研究综述 [J]. 情报学报, 2018, 37(03): 329-40.

[26] 杨飘, 董文永. 基于 BERT 嵌入的中文命名实体识别方法 [J]. 计算机工程, 2020, 46(04): 40-5+52.

[27] 王子牛, 姜猛, 高建瓴, 等. 基于 BERT 的中文命名实体识别方法 [J]. 计算机科学, 2019, 46(S2): 138-42.

[28] 赵平, 孙连英, 万莹, 等. 基于 BERT+BiLSTM+CRF 的中文景点命名实体识别 [J]. 计算机系统应用, 2020, 29(06): 169-74.

[29] 陈赟, 古丽拉·阿东别克, 马雅静. 旅游领域实体和关系联合抽取方法研究 [J]. 计算机工程与应用, 1-13.

[30] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53(03): 582-600.

[31] JI S X, PAN S R, CAMBRIA E, et al. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications [J]. IEEE Trans Neural Netw Learn Syst, 2022, 33(2): 494-514.

[32] PAULHEIM H. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods [J]. Semant Web, 2017, 8(3): 489-+.

[33] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45(04): 589-606.

[34] 崔妍, 包志强. 关联规则挖掘综述 [J]. 计算机应用研究, 2016, 33(02): 330-4.

[35] 赵洪英, 蔡乐才, 李先杰. 关联规则挖掘的 Apriori 算法综述 [J]. 四川理工学院学报(自然科学版), 2011, 24(01): 66-70.

[36] 张黎, 徐蔚然. 中文分词研究 [J]. 软件, 2012, 33(12): 103-8.

“泰迪杯”挑战赛优秀作品