题　　目　基于 LSI 主题模型和 BERT 模型的疫情背景下的周边游需求图谱分析

队员姓名

1. 吕昆

2. 袁钰喜

3. 廖礼航

摘　　要：

如今，我们身处大数据时代，以文本形式产生的数据是了解对应市场的重要信息来源。本文使用给定的的在线旅游和游客的用户生成内容数据，围绕公众号文章分类，旅游产品热度，本地旅游图谱构建做了数据挖掘与数据分析。

对于问题 1，首先判断其本质上是二分类问题，本文通过 LSI 算法提取微信公众号的主题词，并通过主题词与旅游词库的重合度来判断与旅游相关与否。对于问题 2，本文首先对热度指标进行了定义，然后基于 BERT 模型进行了情感分析，并同其他模型进行了比较，并结合游记中花费，人数等指标结合热度定义，进行了热度排序。对于问题 3，本文首先对关联度进行了定义，基于产品名称在其他产品评论的出现频次等指标，算出了关联度，之后基于 paddlepaddle 进行了知识图谱的构建。

关键字： 文本分类；　BERT；　LSI；　情感分析

Today, we are in the era of big data, and the data generated in the form of text is an important source of information for understanding the corresponding market. This paper uses the given user generated content data of online tourism and tourists, and does data mining and data analysis around the classification of official account articles, the popularity of tourism products, and the construction of local tourism maps.

For question 1, first of all, it is judged that it is a binary classification problem in nature. This paper extracts the subject words of wechat official account through LSI algorithm, and judges whether it is related to tourism through the coincidence degree between the subject words and the tourism thesaurus. For question 2, this paper first defines the heat index, then analyzes the emotion based on the Bert model, compares it with other models, and ranks the heat in combination with the indexes such as cost and number of people in the travel notes and the definition of heat. For question 3, this paper first defines the relevance degree, calculates the relevance degree based on the frequency of product names in other product reviews, and then constructs the knowledge map based on paddlepaddle.

Keywords:Text classification;　BERT;　LSI;　Emotional analysis

目录

插图

表格

1. 引言

1.1 挖掘背景

茂名市作为广东的地级市，它拥有全地形特征的中国优秀旅游城市，共有 23 个 A 级景区，其中 4A 级景区 9 个，3A 级景区 14 个，旅游资源丰富。

近年来互联网和信息行业的飞速发展，网络数据具有数据量巨大，种类繁多，价值密度底等显著特征。如何在大量数据中，通过算法找到相关问题隐藏信息成为数据挖掘的关键。并且以文本形式生成内容数据成为主要的数据来源。大数据时代的来临使得数据的规模和复杂性都出现爆炸式的增长，促使不同应用领域的数据分析人员利用数据挖掘技术对数据进行分析。文本挖掘，是数据挖掘领域重要的组成部分。简单地说，文本挖掘就是通过 NLP、机器学习等方法从大量的文本资料中发掘出有价值的信息。无论是微信聊天记录，还是新闻文章，文本挖掘的应用领域非常广泛。

1.2 挖掘意义

随着 2019 年底新冠席卷全球，对旅游产业也造成了巨大的冲击，为了分析疫情背景下茂名的周边游需求情况，对疫情前后的茂名市酒店，景点，餐饮数据进行了数据挖掘与数据分析，不仅对工作人员快速准确的处理评论、减少工作强度等有帮助，而且还找到疫情对茂名旅游产品的影响，以及构建旅游图谱，即以旅游产品为中心，扩散出与其相关的其他产品，比如酒店，机票，餐厅，签证，景点等。

基于此，我们通过对原始数据的特殊字符处理，文本去重，文本分词，去除停用词得到实验数据，然后对微信公众号文章进行分类，找到与旅游相关的文章，之后提取旅游产品，进行情感分析，体现产品热度，找出强相关产品，进行本地旅游图谱构建。目标是希望通过数据挖掘，对部门提出旅游行业发展的建议，进一步提升游客对茂名旅游产品的总体印象评价。以及构建旅游图谱，推动景区智能化信息管理、游客高效查询与决策以及旅游企业精准营销等方面的发展。

1.3 实验环境配置

实验硬件环境配置为：GPU 计算型 GN7|GN7.5XLARGE80(配置一颗 NVIDIA T4),80 内存。操作系统为 Windows Server 2019 数据数据中心版 64 位中文版。开发环境位 Python 3.9, 采用的深度学习框架为 paddlepaddle 百度飞桨框架。

2. 算法介绍

2.1 TF-IDF 算法

在信息检索中，TF-IDF（术语频率 – 逆文档频率的缩写）是一种数字统计，旨在反映单词对集合或语料库中的文档的重要程度，是当今最受欢迎的术语加权方案之一。它由两部分组成，第一部分，TF 是词频 (Term Frequency)，由词频除以文章总词数；第二部分，IDF 是逆向文件频率 (Inverse Document Frequency)，由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。具体公式如下：

$$TF-IDF\left(w_i\right)=\frac{n_{ij}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{j: w\_i \in d\_j\}|}$$

其中，d_j 表示文本，$n_{i,j}$ 表示文本 d_j 中该词出现的次数，$|D|$ 表示语料库中的文本总数，$|\{j; w_i \in d_j\}|$ 表示 包含该词的文本数 。

2.2 Word2vec 算法

Word2vec 是由 Google 的 Mikolov 在 2013 年提出的基于神经网络的深度学习算法。该算法将每个单词表示为实数值的向量，即所谓的词向量。Word2vec 算法的基本构思是基于 bengio 三层神经网络语言模型的改进。Word2Vec 是构建了一个 $f(x) \rightarrow y$ 的映射, 其中 x 是句子中的词语, y 是上下文中的另一个词语, f 是语言模型。这个 f 其实简单来说就是个神经网络, 该神经网络是个很简单的结构, 只有一个线性的隐藏层。我们的目标不是输出多么优秀的结果, 而是需要训练中的权重矩阵。Word2vec 算法由两个模型组成：Cbow 模型和 Skip—gram 模型。

2.2.1 Cbow 和 skip—gram 模型

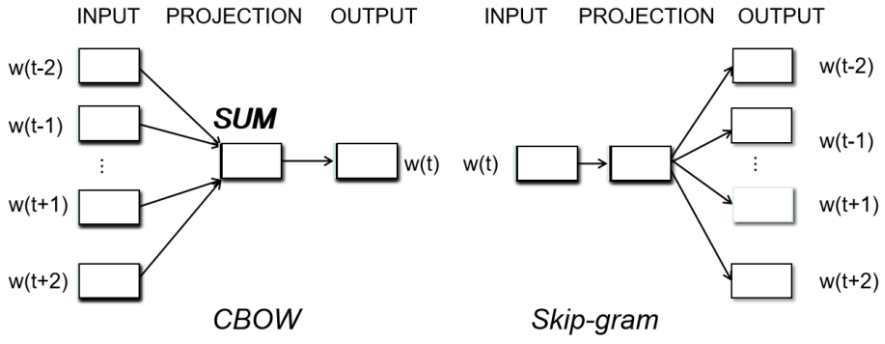

图 1　Cbow 和 skip-gram 模型

从图 1 可以看出，Cbow 和 Skip-gram 模型均包含输入层、投影层和输出层。其中，CBOW 模型通过上下文来预测当前词，Skip-gram 模型则通过当前词来预测其上下文。简单的说 cbow 就是已知周围词预测中心词，而 skip-gram 是已知中心词预测周围词。

2.3 LSI 算法

LSI 算法是一种简单实用的主题模型。LSI 是基于奇异值分解的方法来得到文本的主题的。首先，介绍一下 SVD 分解.

奇异值和特征值有相似的重要意义，都是为了提取出矩阵的主要特征。假设 A 是一个 $m□n$ 阶矩阵，如此则存在一个分解 m 阶正交矩阵 U、非负对角阵 Σ 和 n 阶正交矩阵 V 使得:

$$\mathrm{A} = \mathrm{U} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T$$

Σ 对角线上的元素 $\Sigma_{i,j}$ 即为 A 的奇异值。而且一般来说，我们会将 Σ 上的值按从大到小的顺序排列。他解决了分解中只能针对方阵而没法对更一般矩阵进行分解的问题，因此，我们可以表示为：

$$A = U\Sigma V^T$$

其中 A 是一个 m*n 的矩阵，U 是一个 $m * m$ 的矩阵，Σ 是一个 $m * n$ 的矩阵，V 是一个 $n * n$ 的矩阵。当我们把矩阵 Σ 里的奇异值按从大到小的顺序呢排列以后，很容易就会发现，奇异值 减小的速度特别快。大部分奇异值都很小。因此，我们就可以用前面 r 个奇异值来对这个矩阵做近似，即：

$$A = U\Sigma V^T$$

不同的是，U 是一个 $m * r$ 的矩阵，Σ 是一个 $r * r$ 的矩阵，V 是一个 $n * r$ 的矩阵, 且 $r << m, r << n$。在本文中，A 即是 TF-IDF 值的矩阵。Σ 是非负对角阵且对角线元素是 A 的奇异值按大小排序后保留前 K 个奇异值的矩阵 (K 为假定的主题数)。$V(i,j)$ 可以理解为对于文本和主题的相关度。LSI 就是通过奇异值分解的方法来得到文本的主题，最后返回最相近的主题。

2.4 BERT 模型

BERT 算法它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的 masked language model(MLM) 以致能生成深度的双向语言表征。BERT 是基于 Transformer 模型建立的一个双向编码器。Transformer 模型起源于机器翻译领域，抛弃了循环神经网络中循环式网络结构方法，利用注意力机制构建每个词的特征，通过分析词之间的相互影响，得到每个词的特征权重。

BERT 模型的主要结构图如下：

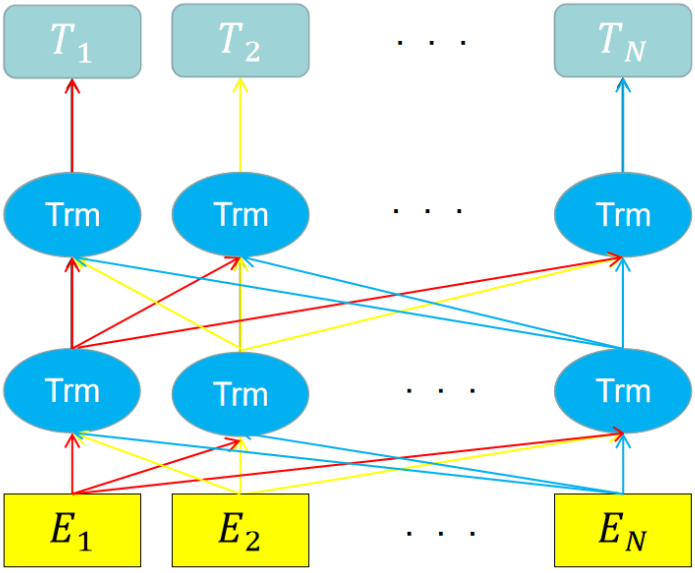

图 2　BERT 模型结构图

其中 E_i 表示中文字符，trm 表示 12 层双向的 Transformer 编码器，最后得到文本字符语境化的向量表示。

2.4.1 BERT 的输入

BERT 模型的输入表示通过对相应的字向量、文本向量以及位置向量进行求和来构造，即输入的是字向量、段向量以及位置向量的总和。其中文本向量的取值在模型训练过程中自动学习，用于刻画文本的全局语义信息，并与单字/词的语义信息相融和合。由于出现在文本不同位置的字/词所携带的语义信息存在差异，因此，BERT 模型对不同位置的字/词分别附加一个不同的向量以作区分。具体流程见图 3:

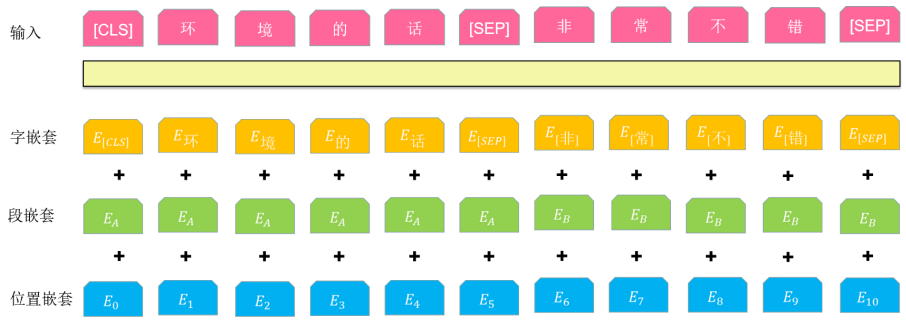

图 3　BERT 模型的输入

其中为了完成具体的分类任务，除了单词之外，输入的每一个序列开头都插入特定的分类 $[CLS]$, 该分类对应的最后一个 Transformer 层输出被用来起到聚集整个序列表征信息的作用。单词外还把分割 $[SEP]$ 插入到每个句子后，以分开不同的句子。

2.4.2 BERT 的预训练

在 BERT 模型之前的语言模型的预训练中，如 Word2vec 和 ELMO 等模型多数是采用从左往右和从右往左预测当前词、上一个词或者下一个词。但是 BERT 模型利用了双向语言模型，其本质就是利用 Transformer 构造一个多层双向的编码网络。为了有效的训练双向语言模型，BERT 采用两项预训练任务目标，即随机屏被词预测和下一句预测。

随机遮盖词预测:

BERT 模型中随机遮盖输入序列中一些词，将其替换为统一标记 [MASK], 设定随机遮盖每个句子中的 15% 的词，对遮盖的词进行预测。其目标是训练 BERT 模型，使得 BERT 模型生成的每个词的表征都参照上下文信息。

随机遮盖词预测训练方法，虽然能够使得 BERT 生成具有上下文信息的词表征，但是这种训练方法造成预训练和微调不一致。BERT 模型针对该缺点进行改进，将 80% 的遮盖词用 [MASK] 进行替换，10% 的遮盖词随机替换为其他词，剩余的 10% 的遮盖词的不改变。通过这种设定，使得 Transformer 编码器无法区分需要被预测的词和需要被替换的词，而强制每个词的表达都需要参照上下文信息。

下一句判断:

BERT 模型中随机输入语料库中句子对，句子对中含有句子 A 和句子 B, 将 50% 句子对中的句子 B 随机替换掉。Bert 模型预测与句子 A 构成句子对的句子 B。该训练方法的目的是使得 BERT 模型具有能够获取句子间的信息的能力。

2.4.3 BERT 的微调

针对特定的任务，Bert 模型可以连接一个额外的输出层，学习最小数量的模型参数。

针对旅游评论情感分析，大概分为了四个步骤：1、在大量通用语料上训练一个 LM(Pretrain)。2、在旅游领域上继续训练 LM(Domain transfer)3、在任务相关的小数据上继续训练 LM(Task transfer)4、在任务相关数据上做具体任务 (fine-tune) 对于参数 Learning rate 需要尽可能的小，以避免灾难性遗忘；Number of epochs: 选择 3 或者 4；weighted decay 使用 warmup 搭配线性衰减。

3. 数据处理

3.1 数据描述

总体来看，给定数据分为 2018-2019 以及 2020-2021 以疫情前后作为时间划分的数据表，表中包含酒店评论，景区评论，餐饮评论，游记攻略以及公众号文章，涉及评论 ID，名称，日期，内容等指标，都是未经加工的数据。通过对数据的简单观察可以发现，他们中含有无意义评论，内容重复评论数据，且评论中常含有大量标点符号，无意义词汇以及部分评论中含有特殊符号。进行去重处理的原始数据有公众号文章，其中 2018-2019

公众号文章 620 条，去重删除短句后 481 条；2020-2021 公众号文章 5665 条，去重筛选后 4490 条。

3.2 数据预处理

文本预处理的过程，主要包括特殊字符处理，文本去重，文本分词，去停用词, 标注。具体流程图如下图。

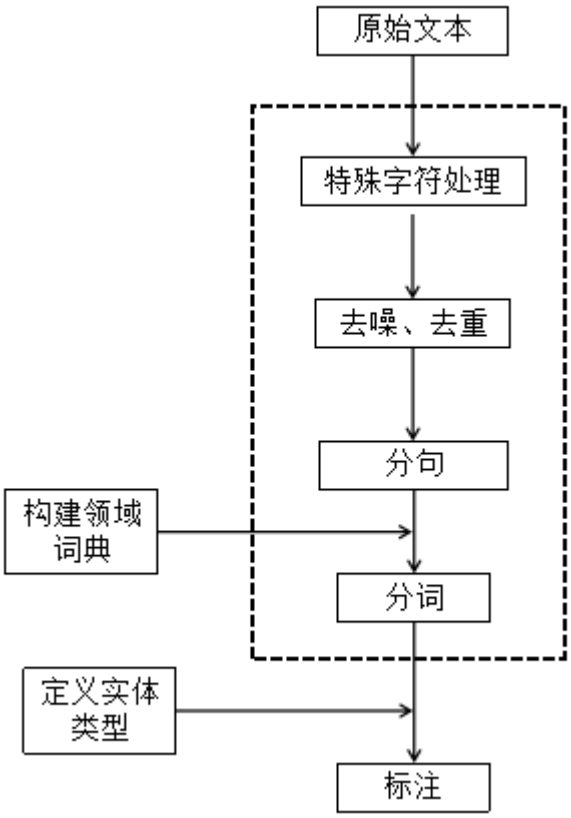

图 4　数据处理流程图

3.2.1 特殊字符处理

在原始数据中，有部分评论含有特殊字符，因此先对数据进行了特殊字符处理，通过设置正则匹配的表达式。正则表达式，也称为 Regex，是一个字符序列，用于匹配某些文本内的字符串模式。匹配模式后，你可以在该模式上应用不同的功能。例如，你可以替换字符串上的值，并且根据正则表达式模式，可以在文本中添加或删除值，可以在文本内部搜索值等。设置正则匹配的表达式后，把文本中匹配到的字符替换成空格，最后去除空白最后得到了只保留中英文，数字和符号的数据。

3.2.2 文本去重

因为在原始评论数据中，包含有评论内容完全一致论 (具体内容见表二)，或者内容相似的评。由于这类评论往往是通过复制粘贴产生的，往往意义不大，因此我们对本文对一致的评论假定评论时间最早的有效，删除其他评论。

表 1　重复及相似评论示例

产品 ID	餐饮名称	评论内容
ID267	围炉海鲜烤肉自助 (金港湾店)	好吃不贵
ID267	围炉海鲜烤肉自助 (金港湾店)	好吃不贵
ID251	椒王火锅 (电白店)	给蜡笔小新好评
ID251	椒王火锅 (电白店)	给蜡笔小新好评

在实际操作上，我们使用 python 的去重函数判断并去除重复评论，并且保留评论时间最早的评论。

3.2.3 文本分词

在中文中, 词是最小的语言单位。中文分词，其意思是将中文的一句话按照一定的规范分割成单独的词，这也是处理中文文本与英文文本的不同之处，因为英文文本单词中间带有空格，可以起到分词的效果，中文只有标点符号对语句进行断句。因此需要利用中文分词的方法对中文语句进行识别。

目前的分词工具有 Jieba 分词、NLPIR 分词系统、pyltp 等等。本文采用了 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，是最常用的分词工具之一。jieba 的主要功能是做中文分词，可以进行简单分词、并行分词、命令行分词，当然它的功能不限于此，目前还支持关键词提取、词性标注、词位置查询等。因此，本文采用了 jieba 进行分词。

3.2.4 除去停用词

在自然语言处理过程中, 有一类对于分类结果没有参考价值的词, 通常将其称之为停用词。就中文而言, 存在着介词、连词、语气助词以及各种各样的标点符号, 例如“恩”、“呵”、“哦”、“!”等。对于这一类词, 在文本预处理的过程中需要直接将其过滤。目前比较常见且公认的停用词表有: “哈工大停用词表”、“百度停用词表”等。

为了更加有效的去除停用词，我们首先合并了哈工大停用词表，百度停用词表，四川大学机器智能实验室停用词库，中文停用词表。之后为了进一步优化结果，对词频过高且没有参考价值的词加入了停用词并再进行去除停用词不断更新停用词词库。具体加入的部分停用词如下表 (针对微信公众号文章)：

表 2　加入的部分停用词示例

停用词	停用词	停用词
陈小姐	林小姐	点击
时	茂名	小姐

3.2.5 标注

按照上述规则对原文本进行标记, 将每一个字与标签单独放到一行进行表示, 并且在字和标签之间用空格隔开, 为了训练要求, 每个句子后面添加空白行, 部分标注数据如表 4。

表 3　部分标注数据示例

文本	标注	文本	标注
贾	B-SA	登	I-SA
峪	I-SA	国	I-SA
家	I-SA	森	I-SA
林	I-SA	公	I-SA
园	I-SA	位	O
于	O	阿	B-LOC
勒	I-LOC	泰	E-LOC
地	I-LOC	区	O

4. 微信公众号文章分类

4.1 问题分析

问题需要将微信公众号文章分类为与旅游相关与旅游不相关两类，其本质是一个二分类问题。问题的关键在于对数据主题的抽取以及对主题是否和旅游相关性高。对与文本数据的主题抽查，可以使用 LSI 算法，LDA 主题模型等提取文章潜在主题。通过每个支持该主题的词的概率值表示, 概率值越大, 说明该词与主题的关联程度越大 [8] 。主题强度的分布反映每个主题在文本库中的相对分量, 主题强度可以衡量主题在文本库中的受关注程度, 计算公式如下:

$$P_b = \frac{\sum_i^N \theta_b}{N}$$

其中, P_k 为第 k 个主题的强度, N 代表文档数量, θ_{ki} 为第 k 个主题在第 i 篇文档中的概率。

其次是对抽取的主题进行旅游相关主题进行比较，判断是否相关

4.2 基于 LSI 算法的主题抽取

公众号评论内容 s 冗长，即使通过预处理也难以判断中心思想，因此我们需要通过算法对每条公众号文章进行主题抽取 LSI 是基于奇异值分解的方法来得到文本的主题的。前面介绍了 LSI 算法的实现过程，在实际抽取过程中，我们以分词后的结果构建词典并建立 LSI 模型，然后对每条微信公众号文章抽取出来的主题我们只保留了与文章最相关的 5 个主题，以免若相关主题干扰旅游相关性判断。微信公众号文章主题抽取结果实例如下表：

表 4　公众号文章主题抽取结果实例

分词后部分文章	主题词
··· 公园落地动工广东滨海旅游公路段全省开建 ···	旅游，滨海，美丽，童子，综合
··· 秀丽湖泊山丘森林莲花湖独具魅力自然景观 ···	旅游，莲花，庄园，信宜，汽车

4.3 旅游相关性分类

与文旅相关性较强的主题有旅游、活动、节庆、特产、交通、酒店、景区、景点、文创、文化、乡村旅游、民宿、假日、假期、游客、采摘、赏花、春游、踏青、康养、公园、滨海游、度假、农家乐、剧本杀、旅行、徒步、工业旅游、线路、自驾游、团队游、攻略、游记、包车、玻璃栈道、游艇、高尔夫、温泉等等。但与旅游主题相关的主题远不止与此，问题的关键是找到旅游相关主题词库。

为此我们找到了三个旅游词库并且采用准确率最高的词库来作为判断微信号文章主题与旅游是否相关的主题词库。词库 (部分) 见下表：基于以上旅游主题词库，对提取出

表 5　旅游词库 (部分)

主题词	评价	主题词	评价
风景	好	天气	热
景色	好	门票	贵
排队	久	瀑布	壮观

的公众号文章主题词进行匹配判断，如果有三个词在旅游主题词库中，我们便定义为相关，反之不相关。

5. 产品热度分析

5.1 问题分析

通过旅游大数据进行热度分析能更好地了解旅游需求、游客行为、游客满意度等旅游问题。旅游产品提取的关键在于预处理的准确性，需要提取包括景区、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，将提取出的旅游产品和所依托的语料，所以在本节不过多阐述。将重点放在对周边游产品热度分析，其关键在于热度的定义，需要根据已有数据，尽可能多的考虑热度影响因素，确保热度的有效性。

5.2 定义热度指标

热度，反映了特定人群在某段时间普遍关注的问题和事物。在所给数据中，需要综合考虑游客正负向情感，评论条数，时间段等一系列因素。通过查阅相关文献后，以及对数据观察分析，对热度相关指标及其旅游产品热度做出了如下定义：

(1) 该目的地最早评论日期与最晚评论日期的间隔天数 m, 热度高的产品评论往往在较短时间内集中产生；

(2) 该产品评论的正向情感次数 a 与负向情感次数 b, 评论的条数越多，评论中情感词数越多, 也反映出更多的关注度。

(3) 对游记中涉及到的产品的出行人数 s，出行天数 d，及其花非 p 都是衡量热度的重要指标, 对于热度高的产品，人们往往具有结队出行，游玩以及消费欲望。

则目的地的热度 (L) 公式如下式:

$$
\begin{aligned}
L &= \mathbf{W}\sqrt{\frac{\left(10+\frac{a+b}{5}\right)}{4\left[1+\log_2(\mathrm{m}+5)\right]}} \\
&= \left(\frac{s}{4\mathbf{E(s)}}+\frac{d}{4\mathbf{E(d)}}+\frac{p}{2\mathbf{E(p)}}\right)\sqrt{\frac{\left(10+\frac{a+b}{5}\right)}{4\left[1+\log_2(\mathrm{m}+5)\right]}}
\end{aligned}
$$

其中，**W** 表示权重, 由出行人数 s, 花费 p, 出行天数 d 构成,**E()** 表示取均值.

通过观察数据，我们可以发现游记中有缺失数据，因此对游记中没有的涉及的产品取权重 $\mathbf{W}=1$, 对出行人数 s，出行天数 d，及其花非 p 有部分缺失的数据，对缺失部分取均值填充。

5.3 基于 BERT 模型的旅游产品情感分析

5.3.1 模型评估指标

模型评估指标使用准确率 (P), 召回率 (R) 和 $F1$ 值, 其中 TP 为正确识别的实体个数, FP 是识别出的不相关的实体个数, FN 是数据集中存在且未被识别出来的实体个数。

$$
\begin{aligned}
P &= \frac{TP}{TP+FP}\times 100\% \\
R &= \frac{TP}{TP+FN}\times 100\%
\end{aligned}
$$

通常精确率和召回率的数值越高, 代表实验的效果好, 然而有时精确率越高, 召回率却越低, 所以需要综合考量他们的加权调和平均值, 也就是 $F1$ 值, $F1$ 值定义如下:

$$F1 = \frac{2PR}{P+R} \times 100\%$$

5.3.2 BERT 情感分析

为了计算影响热度中的正负向情感次数，我们对旅游产品评论做了情感分析。该任务属于自然语言处理中的文本分类任务，首先我们的测试集包含 9700 多条带标签的酒店评论，利用这些评论进行模型训练和验证，计算了模型的准确率、召回率等指标并且与传统的分类算法结果做了模型评价指标的比较，具体结果如下表或图：

表 6　不同算法在训练集上的表现

方法	准确率	召回率	精确率	F1
支持向量机	0.505	0.505	0.498	0.501
BP 神经网络	0.587	0.587	0.547	0.502
朴素贝叶斯	0.589	0.589	0.55	0.491
adaboost	0.811	0.811	0.812	0.808
随机森林	0.944	0.944	0.947	0.944
catboost	0.977	0.977	0.977	0.977
梯度提升树	1	1	1	1

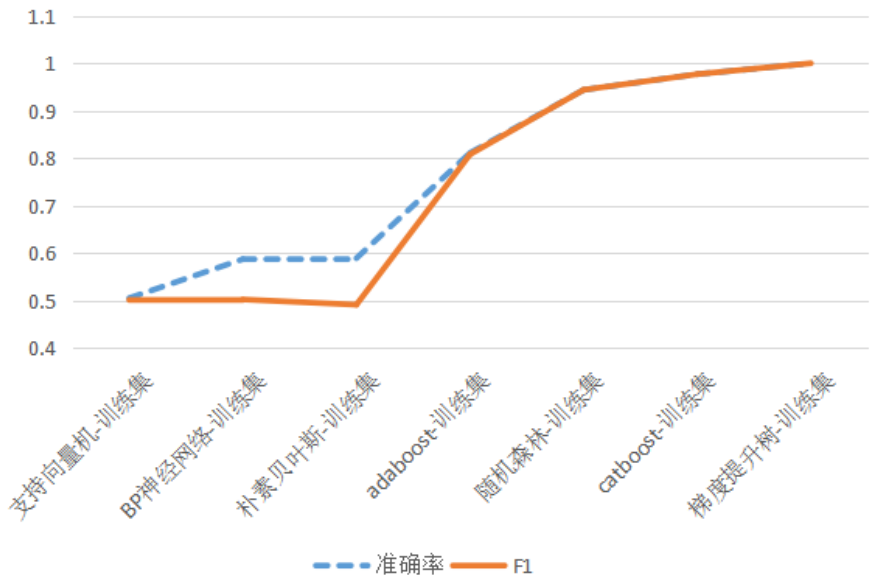

图 5　不同算法在训练集上的表现

表 7　不同算法在测试集上的表现

方法	准确率	召回率	精确率	F1
支持向量机	0.617	0.617	0.556	0.547
BP 神经网络	0.463	0.463	0.507	0.472
朴素贝叶斯	0.55	0.55	0.515	0.525
adaboost	0.6	0.6	0.557	0.556
随机森林	0.523	0.523	511	0.516
catboost	0.487	0.487	0.528	0.495
梯度提升树	0.557	0.557	0.541	0.547

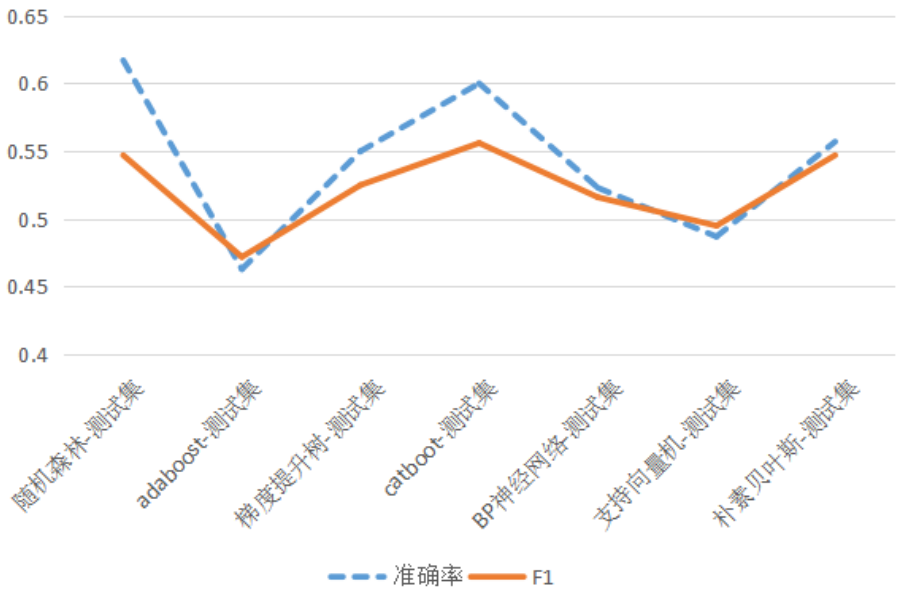

图 6　不同算法在测试集上的表现

可以发现，虽然部分算法在训练集上表现不错。但是总体来看，他们普遍都在测试集表现上差强人意。因此，本文采用基于 BERT 模型的情感分析，并且在测试集上表现优异，具体数值如下表：

表 8　BERT 在测试集上的表现

	准确率	损失函数值
BERT 模型	0.90	0.0988

为了分析疫情前后的旅游情感倾向，我们分别统计了 2018-2019 和 2020-2021 的酒店、餐饮、以及景区的正负向情感频次，具体数据如下图：

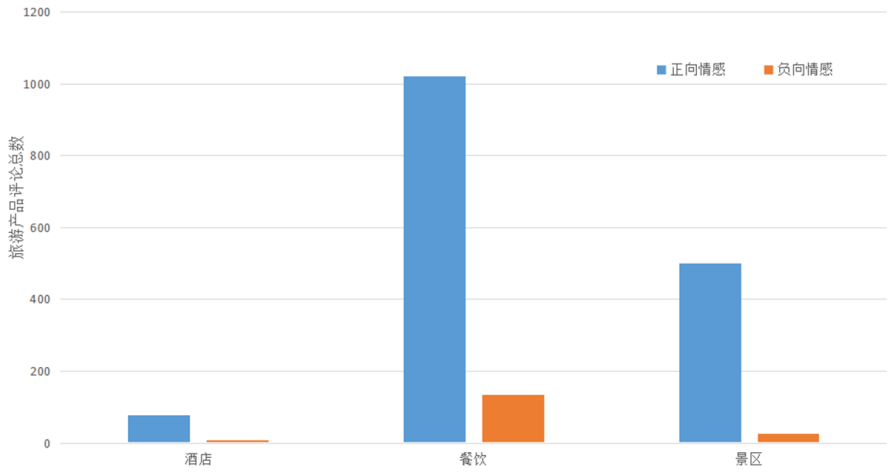

图 7　2018-2019 旅游产品情感分布

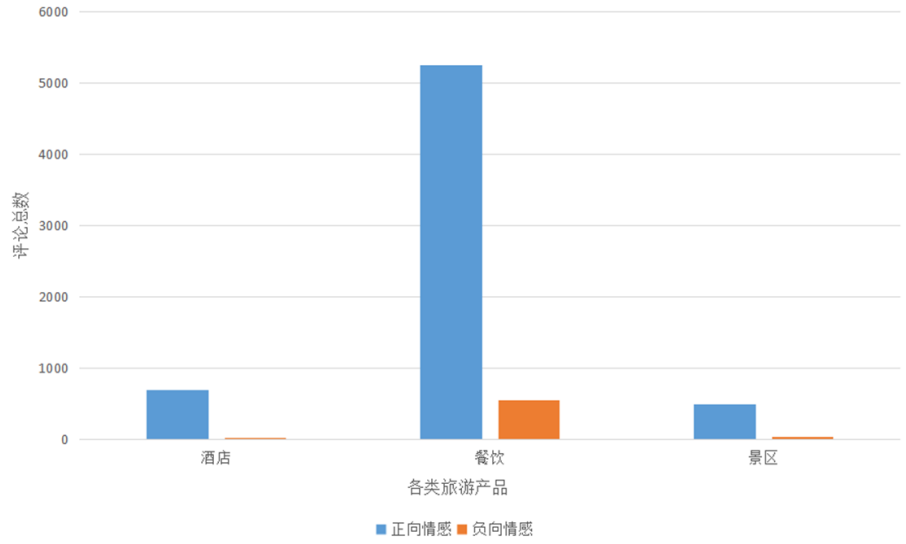

图 8　2020-2021 旅游产品情感分布

根据情感分布图，我们可以发现负向情感在分布中占比较少，并且在时间段上，2020-2021 年时间段的负向情感占比较 2018-2019 略微下降。可以间接说明当地对于疫情控制较好，并且对旅游产业的重视度较高。

5.4 热度计算与分析

基于上一节 BERT 模型的情感分析结果，我们得到了每个产品的正向负向情感频次 a, b，以及通过对日期格式的评论时间最大值减去最小值得到评论间隔天数 m，使用产品名称在游记中进行匹配，得到旅游产品的人数 s, 花费 p，出行天数 d。至此，与热度相关的指标值全部得到，通过计算，得到热度，部分产品热度如下图：

疫情前后部分旅游产品热度对比如下图：

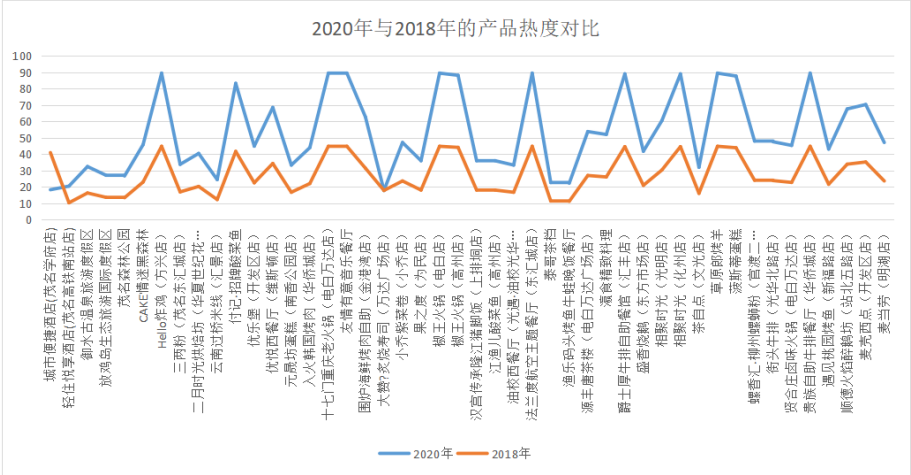

图 9　疫情前后部分旅游产品热度

通过对疫情前后的产品热度对比，我们不难发现：2020-2021 旅游产品热度较 2018-2019 产品热度整体上有提升，这个结果也与之前的情感分析结果相似。虽然新冠疫情不同于以往的旅游危机事件，从时间尺度上和空间尺度上给旅游业带来的冲击是前所未有的。但是，一方面，“二战”之后 7 0 多年逐渐形成的世界旅游经济的运行逻辑被打破；另一方面，旅游业在重启过程中也将发生结构性变革。全球旅游业将进入“二战”以来最为关键的产业重塑期。值此之际，对旅游领域中有关新冠疫情等文献进行深入分析，对于我国乃至全球旅游的研究与旅游业发展都具有十分重要的意义。数据中挖掘出信息，随着年份的增长旅游产品的热度有着上升的趋势，产品热度高的随着年份的增长增幅也大。随着全球新冠疫苗接种的推广，旅游业随之进入震荡式复苏，但未来发展依然存在较大的不确定性。

6. 本地旅游图谱构建与分析

6.1 问题分析

为了构造知识图谱，需要找出以以景区、酒店、餐饮等为核心的强关联模式，而构造知识图谱的基础任务是从文本信息中抽取三元组，即找出文本中的实体嵌套以及关系，即实体和关系抽取是基础，实体对齐是知识融合的保障。通过构建旅游领域的知识图谱并结合智能问答、个性化推荐等上层应用，可以促进旅游行业智能信息服务技术的快速发展，从而带来更高的经济效益。

6.2 旅游产品关联分析

6.2.1 旅游产品关联度定义

通过观察我们发现，不同旅游产品名称会出现在不同旅游产品评论中，这是产品关联的重要标志，因此，本文定义产品 i 与产品 j 的关联度如下 l：

$$l_{i,j} = \frac{n_{i,j}}{m_i}$$

其中，$n_{i,j}$ 表示产品 i 的名称在产品 j 的评论中出现的频次，m_i 表示产品 i 的评论总频数。得到如下总体结果：

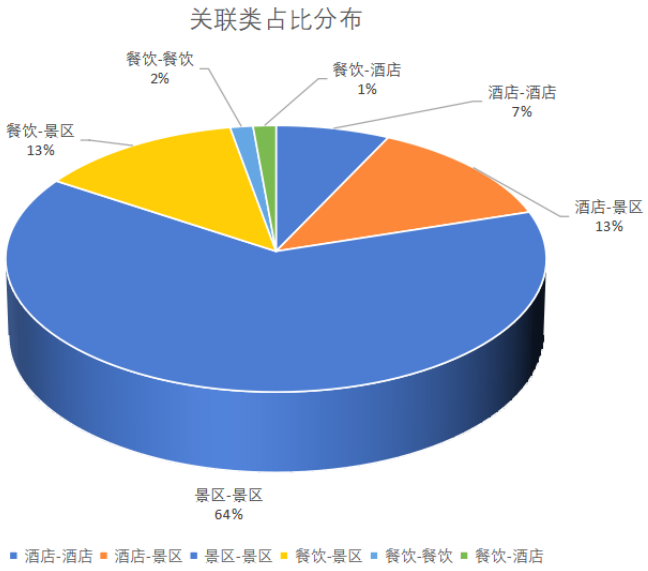

图 10　关联占比分布

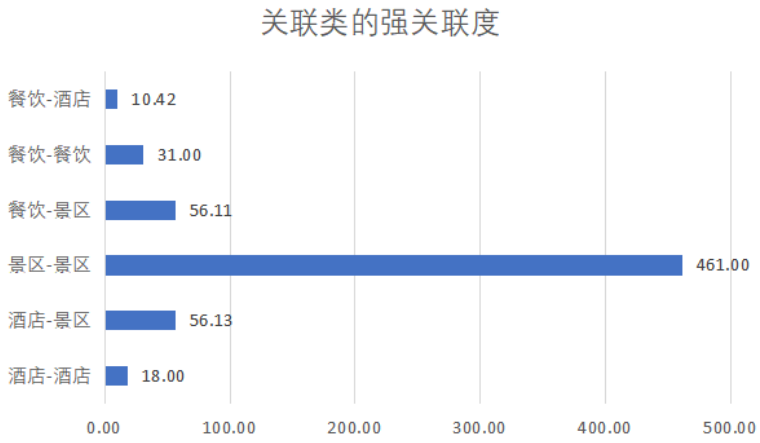

图 11　关联强度情况

通过关联强度与关联占比的分析，我们不难发现，景区 -景区的关联类型联系最紧密，

其次是景区 -餐饮。可以间接说明游客对茂名市旅游景点的喜爱，游客的旅游往往是按景点分布为行进路线的，因此景区 -景区的关联类型联系最为紧密。

6.3 旅游知识图谱构建

6.3.1 基于 BERT 的三元组抽取

知识图谱中的数据通常以“实体 -关系 -实体”或“实体 -属性 -属性值”的关系三元组存储, 形成一个图状知识库，因此从非结构化文本信息中抽取关系三元组是构建知识图谱的关键任务。而旅游领域的文本信息中存在着大量的实体嵌套和关系重叠现象，例如表 7 文本中包含的以 (subject，predicate，object) 形式的关系三元组有 [(杭州西湖风景区，所在城市，杭州),(杭州西湖风景区，著名景点，苏堤春晓), (苏堤春晓，所属景区，杭州西湖风景区), (苏堤春晓，所在城市，杭州)]

表 9　旅游领域关系重叠样例

text	杭州西湖风景区内的苏提春晓因苏轼而知名
EPO	(杭州西湖风景区，著名景点，苏堤春晓)
	(苏堤春晓，所属景区，杭州西湖风景区)
SEO	(杭州西湖风景区，所在城市，杭州)
	(杭州西湖风景区，著名景点，苏堤春晓)

在实体抽取过程中，其中作为景点名称的实体“杭州西湖风景区”中嵌套作为城市名称的“杭州”，Luo 等人提出基于注意力机制的 $Att-BiLSTM-CRF$ 模型进行化学领域命名实体识别，使用 $B/I/O$ 标签加实体类型来区分实体的开头、中间和结尾或者判断是否为实体。这种命名实体识别的方法无法将实体“杭州西湖风景区”中的“杭”同时标注为景点名称的开头和城市名称的开头，因而无法解决实体嵌套问题。

在关系抽取的过程中，表 7 中文本包含 EPO 和 SEO 两种关系重叠形式。其中，EPO 是指句子中至少有两个关系三元组，并且至少有两个关系三元组以相同或者相反的顺序共享一对实体；SEO 是指句子中有多个关系三元组，并且至少有两个关系三元组共享一个相同的实体。由于管道抽取模型存在误差积累和实体冗余会造成模型性能大幅下降，不能够有效处理关系重叠问题。Zheng 等人提出的联合抽取模型将问题转化为序列标注问题，生成标注序列后将关系标签合并为实体三元组时采用就近组合的方法，虽然能够从句子中抽取出多种关系，但并不能处理实体出现重叠的关系抽取问题。

其次，对三元组中的实体和关系名称进行补充是营商环境文本知识抽取的重要步骤，作为政策名称实体如果不使用规则，则很容易被拆分抽取部分作为实体而导致三元组错

误。

根据语义角色标注和依存句法分析结果和以上三元组抽取规则，总结实体关系抽取算法的具体步骤如下：Step1：输入文本语句；Step2：构建依存句法树；Step3: 根据{”SBV”,”VOB”,”FOB”,”POB”,”CMP”,”COO”} 句法关系提取三元组，如果提取出三元组，跳转至步骤四；如果没有提取出三元组，跳转至步骤五；Step4：对提取的三元组进行扩展，将定中关系包含于其中；Step5：对定中关系中进行分类，分别为名词组合和动词组合，对其中的动词组合根据主谓宾关系提取三元组。

6.3.2 基于 paddlepaddle 的知识图谱构建

知识表示与推理是受人类解决问题的启发，将知识符号化，为智能系统获取解决复杂任务的能力而进行的过程和方法，知识图谱技术是知识表示与推理的重要研究内容。知识图谱 (Knowledge Graph KG) 是以图的形式表现客观世界中的实体 (概念、人、事物) 及其之间关系的知识库，通常表示为三元组 (h, r, t) 的集合，其中 h、t 表示实体，r 表示实体之间的关系，每个三元组 (h, r, t) 代表一条事实，即“头实体”和“尾实体”之间存在有向关系。图 1 给出了知识图谱的一个典型示例。

在知识图谱相关研究中，知识表示是知识应用与获取的基础，是贯穿知识图谱的构建与应用全过程的关键，也是知识图谱相关研究的热点内容。基于知识图谱的知识表示学习 (Knowledge Representation Learning, KRL)，也称知识图谱嵌入 (Knowledge Graph Embedding, KGE)，是对知识图谱中的实体和关系完成分布式表示的过程，通过将实体和关系映射到低维向量空间来间接捕获它们的语义。

基于 paddlepaddle 的知识图谱如下图：

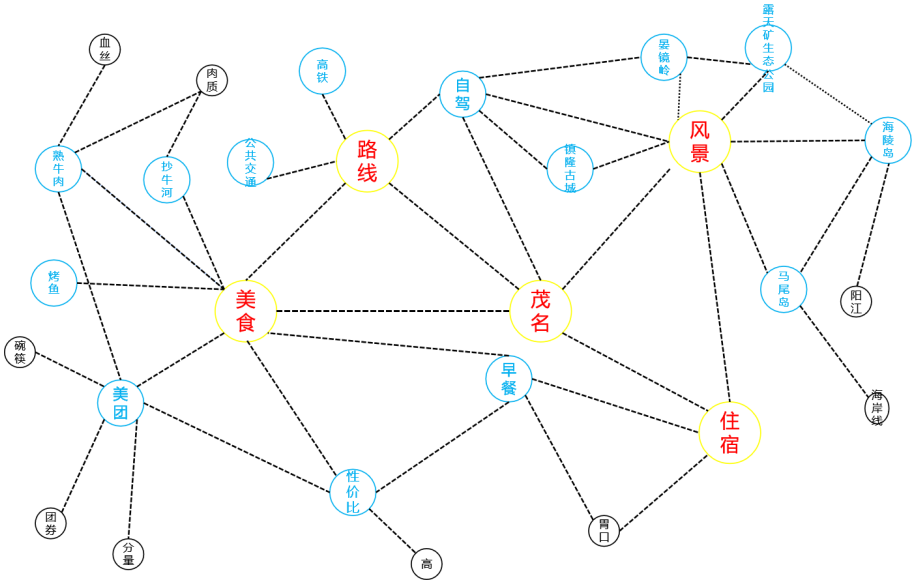

图 12　不同算法在测试集上的表现

7. 总结与建议信

7.1 总结

大数据时代的来临迫切需要我们不仅从海量的数字数据中挖掘有效的信息，还得分析以文本形式出现的数据来源，而大数据的规模和复杂度也对我们挖掘信息提出了进一步的要求。当然本文注重研究的是文本数据的挖掘，主要应用了这几种算法：TF-IDF 算法，Word2vec 算法，BERT 模型，LSI 算法。IF-IDF 算法是一种数字统计，主要反映词对集合或语料库中的文档的重要程度，有利于提取关键词或主题句为文本删繁去重，减少冗余做准备。Word2vec 是基于神经网络深度学习算法，该算法将每个单词表示为实数值的向量，就是人们常说的词向量，而 Word2vec 作为 Cbow 和 skip-gram 模型的输入层有助于深度挖掘文本中关联词或预测其上下文。

LSI 算法是一种简单实用的主题模型，LSI 是基于奇异值分解的方法得到文本主题的，通过这种算法用于文本的相似度有着举足轻重的作用。BERT 算法不同于传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的掩蔽语言模型以致生成深度的双向语言表征。该算法打破了 CNN 和 RNN 只能捕获单一的近邻关系和序关系的能力边界，利用注意力机制来预测和判断文本中的不同位置上、不同尺度下的关系。一种好的算法想要发挥其最大的作用免不了对数据进行预处理，而肉眼直观可见茂名市地区文旅文本评论中出现了大量特殊符号，无意义且内容重复的数据。因此需要对数据进行处理。首先通过正则表达式找到特殊字符串将匹配到的字符替换成空格，然后只保留中英文，数字和符号的数据，其次是文本中包含重复的评论，很明显就是游客偷懒复制加粘贴其他人的评论想省事省力，但重复评论过多对于关键字或主题词的权重造成了影响，也对输出结果造成了偏差应当剔除。最后就是文本的分词和标注，本文采用了市面上比较成熟且好用的分词工具 Jieba 分词，jieba 分词这是数据分析的第一步也是关键一步，分词的好坏直接影响整个文本挖掘的走向，当然也不能马虎大意分词完了就拿去套用算法或模型这是不对的，还需要人工的校对和修改，不断地改正词库或调参直到分词满意为止，分完词之后就是简单的训练数据要求的标注。到了这里为止数据的预处理和算法模型也已经准备到位，下面就是实际操作应用。

针对问题一：判断微信公众号与旅游相关本质上是二分类问题，首先对部分数据进行人工的标记划分为相关或不相关，然后通过 LSI 算法识别微信公众号的主题词，并算出微信公众号的主题词占微信公众号的比例，就可以通过逻辑回归进行有监督的分类问题划分。

针对问题二：通过旅游大数据进行热度分析能更好地了解旅游需求，从每个产品中可能包括但不限于游客的正负向情感次数，评论条数，时间，位置关系等一系列因素，在考虑多方面的影响下定义了热度 L 计算公式。可以很好地反映旅游产品的热度大小，还可

以对比不同旅游产品的热度关系。

针对问题三：找出产品间的关联模式，当然我们考虑产品的关联时涉及的因素比较多，比如产品的物理位置关系，游客的行程路线顺序关系，以及游客本身对产品的偏好程度也是需要处理的对象，相对而言微信公众号或游记中的同一评论对不同产品的评论也是潜在的关联只是不易被发现。这里我们在处理产品的关联度时把产品自身出现的评数当成权重，出现的越多表示关联权重越大。而关联强度中明显景区 -景区的关联与其他关联关系中不在同一级别下，这也很好说明了，在当下人们的生活压力大周末或假期外出旅游偏向于体验大自然的无限风光以减少压力。

7.2 建议信

尊敬的区旅游主管部门:

您好！

我们是参加数据挖掘竞赛的成员，通过对茂名市的在线旅游和游客的用户生成内容数据进行数据挖掘和数据分析，我们得到茂名市旅游产品的总体信息，希望可以和您分享一下，以进一步提升游客对茂名旅游产品的印象与评价。

通过茂名旅游产品热度分析，我们发现茂名旅游产品 2020-2021 年热度较 2018-2019 整体上提升。可以说茂名市相关部门对于疫情防控做得相当到位，并且在防控的同时，也推进了当地旅游事业的发展，这一点做得很好，同样也可以在情感分析中发现，大量评论都是带正向情感的，可以说游客对茂名旅游产品评价也是比较高的，但是我们分析了其中部分负向评论，酒店中有数条反应隔音不好的问题，又或者卫生条件差等评论。对此我们建议涉及到的酒店加装隔音棉，保证用户休息，加强员工管理，加强清洁打扫。景点评论中反应部分景区餐厅有点贵，景区水质不佳，不好玩等负向评论，建议加强景点餐饮管理，尽量做到物有所值，有条件更做到物超所值，保证游客花钱的同时有良好的体验，并且加强旅游卫生管理，提高垃圾桶的数量，树立警示牌，包围好我们共同的家园。在餐饮评论中，我们发现有评论显示味道越来越差，难吃，服务态度不好，卫生不好等一系列问题，在旅游产品类中反响最差，建议相关部门加大对餐饮行业的检查力度，并且加强卫生管理，员工素质培养等等，做好反思。

通过对旅游产品关联度分析，我们发现部分产品有关联，这些关联往往在地区上有所体现，建议有关联的旅游产品进行相关联动活动，比如购买联票优惠等策略，加强游客游玩欲望。

我们的分析和建议到这就结束了，相关数据会通过邮件发送给您，希望通过我们共同的不懈努力，茂名市旅游产品会越做越强，越做越大。

参赛成员

2022/4/28

参考文献

[1] 徐春, 李胜楠. 融合 BERT-WWM 和指针网络的旅游知识图谱构建研究 [J/OL]. 计算机工程与应用:1-10

[2] 刘美茹. 基于 LSI 和 SVM 的文本分类研究 [J]. 计算机工程,2007(15):217-219.

[3] 殷纤慧, 古丽拉·阿东别克. 基于多特征注意力卷积神经网络的旅游领域实体关系抽取 [J]. 东北师大学报: 自然科学版,2022,54(1):79-83.

[4] 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/OL]. 计算机工程与应用:1-13

[5] 徐有为, 张宏军, 程恺, 廖湘琳, 张紫萱, 李雷. 知识图谱嵌入研究综述 [J/OL]. 计算机工程与应用:1-25

[6] 任瀚, 张怡. 新冠疫情冲击下旅游相关研究的进展与展望 [J]. 资源开发与市场,2022,38(2):231-238