第十届“泰迪杯”
数据挖掘大赛C题

报
告

作品名称：基于疫情下的旅游图谱分析

目录

摘要

本次数据挖掘采用KNN与TF-IDF结合模型进行文本分类，运用HMM模型与维特比算法相结合进行旅游产品提取，运用Apriori算法处理关联规则，解析当下疫情旅游图谱。

1 引言

1.1 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，OTA）和游客的用户生成内容（User Generated Content，UGC）数据成为了解旅游市场现状的重要信息来源。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。

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。

本文针对OTA和UGC数据，采用自然语言处理等数据挖掘方法，科学构建旅游产品的热度评价模型，并挖掘出旅游产品间隐含的关联模式，对新冠疫情时期茂名市周边游的发展进行了全面的可视化分析，热度、关联度对于旅游业的资源配置、协同发展都具有重大意义，对于旅游市场未来的开拓方向有参考价值与借鉴意义。

1.2 文本挖掘的目标

1.微信公众号文章分类

在大量的微信公众号推送中，提取出与文旅相关性较强的文章，利用自然语言处理中的文本分类（Text classification)技术力图实现文章内容的自动分类。

2.周边游产品热度分析

景区、酒店、特色餐饮等周边游产品基于 OTA 和 UGC 数据可以进行综合分析，也可以对自身的特定方向进行针对性分析，热度分析则显示出当地的旅游热点，对旅游企业产品的资源优化配置提供引导。

3.周边游产品关联分析

枯燥单一的旅游产品无法满足旅客多变的需求，因此，挖掘出旅游产品之间潜在的关联性，对于引导旅客的消费走向，亦或是完善当地旅游产业的布局，都具有重大意义。

2 相关算法介绍

2.1 TF-IDF 算法

TF-IDF 由词频和逆文档频率组成，可用于计算文本相似度量。主要思想是：文档关键词即会随着它在文档向下走而增加，但在语料库中出现的频率低。则可以很好的区分文档关键词和其他词的关系，适用于分类。

$$\boldsymbol{TF} = \frac{词条在文本出现的次数}{文本所有词条的数量}$$

$$\boldsymbol{IDF} = \boldsymbol{log}\left(\frac{语料库文档总数}{某词条出现文档数 + \mathbf{1}}\right)$$

2.2KNN 最近邻算法

KNN 算法是一种监督学习算法。 KNN 是计算训练集（样本）数据到测试集（待分类）数据的距离，取和测试数据最近 K 个样本数据，从 K 个样本中的数据得出其中占比类别最多的样本数据，即得测试集数据就属于该类别。

K 近邻分类器具有良好的文本分类效果，对仿真实验结果的统计分析表明:作为文本分类器，K 近邻仅次于支持向量机，明显优于线性最小二乘拟合、朴素贝叶斯和神经网络。

KNN 算法步骤:

1. 建立训练集和测试集，设定参数 K

2.计算需要分类的点到其余点的距离（确定距离的方法：欧几里得距离、明可夫斯基距离、曼哈顿距离、切比雪夫距离），本次数据挖掘采用的是欧式距离。

3.距离升序排序，选距离样本点最近的 K 个点

4.加权平均，得到答案

2.3 HMM 模型

HMM(隐马尔科夫模型)是自然语言处理中的一个基本模型，用途比较广泛，如汉语分词、词性标注及语音识别等，在 NLP 中占有很重要的地位。

2.3.1 隐马尔可夫链的介绍

HMM 模型是概率图模型的一种，属于生成模型。在 NER 中定义的实体标签，是一个不可观测的隐状态，而 HMM 模型描述的是由这些隐状态序列（实体标准）生成可观测结果（可读文本）的过程。

1.可观测序列：由所有汉字组成的集合，使用 V_{OBS}表示。V 表示单个汉字，n 为汉字序列的最大值

$$\boldsymbol{V_{OBS}} = \{\boldsymbol{v_1}, \boldsymbol{v_2} \ldots \boldsymbol{v_n}\}$$

2.隐藏状态的集合：由实体命名识别数据的标签组成,使用 Q_{hidden}表示。

$$\boldsymbol{Q_{hidden}} = \{\boldsymbol{q_1}, \boldsymbol{q_2} \cdots \boldsymbol{q_m}\}$$

3.观测到一串自然语言序列文本 O（共有 t 个字），对应的实体标记为 I。

$$\boldsymbol{I} = \{\boldsymbol{i_1}, \boldsymbol{i_2} \ldots \boldsymbol{i_t}\}$$

$$\boldsymbol{O} = \{\boldsymbol{O_1}, \boldsymbol{O_2} \ldots \boldsymbol{O_t}\}$$

4.HMM 状态转移概率:表示在时刻 t 处于状态 q_i的条件下在时刻 t+1 转移到状态 q_j的概率。

$$\boldsymbol{A} = [\boldsymbol{a_{ij}}]_{\boldsymbol{n*m}}$$

$$a_{ij} = P\left(i_{t+1} = q_j \middle| i_t = q_i\right)，\ i = 1,2\ldots n; j = 1,2\ldots n$$

5.HMM 观测概率矩阵：表示时刻 t 处于状态q_j的条件下生存观测v_k的概率。

$$B = [b_j(k)]_{N*M}$$

$$b_{ij} = P(o_t = v_k | i_{t=} q_j), k = 1,2\ldots m; j = 1,2\ldots n$$

6.π是初始状态概率向量：就是由空状态转换为有状态的一个概率。

$$\Pi_i = P(i_1 = q_i). i = 1,2\ldots n$$

给定观测序列$O = \{O_1, O_2 \ldots O_t\}$和模型$\lambda = (\pi, A, B)$,求出最大概率的隐藏序列$X = \{X_1, X_2 \ldots X_t\}$。HMM 模型由初始的概率分布、HMM 状态转移概率分布以及 HMM 观测概率分布三者共同确定。

2.3.2 维特比算法

运用隐马尔可夫模型需要用 Viterbi 解码，。Viterbi 算法是用动态规划的思想去计算 HMM 中的最优路径问题。即给定隐马尔可夫模型的观测序列（文本序列），找到最可能的隐藏状态序列（文本标签），也称为最优路径。假设隐藏状态序列长度为 N，观测序列长度为 T，那么每个观测元素的状态都有 N 种可能，也就是 N^T ，时间复杂度太高。

1.在 t 时刻，隐状态为 i 的所有单个路径的概率最大值：

$$\delta_t(i) = \max_{i_1, i_2, \ldots i_{t-1}} P(i_t = i, i_{t-1}, \ldots i_1, o_t, \ldots o_1 | \lambda)$$

2.递推：

$$\delta_{t+1}(i) = \max_{i_1, i_2, \ldots i_t} P(i_{t+1} = i, i_t, \ldots i_1, o_t, \ldots o_1 | \lambda$$

3.在 t 时刻，隐状态为 i 的所有单个路径的概率最大值的 t-1 个节点：

$$\psi_t(i) = \underset{1 \le j \le N}{argmax}[\delta_{t-1}(j) a_{ji}], i = 1,2\ldots N$$

4.初始化概率矩阵：

$$\delta_1(i) = \pi_i b_i(o_1)\ \ \psi_1(i) = 0 \text{ i=1, 2⋯N}$$

5.递推：　　$\delta_t(i) = \max_{1 \le j \le N}[\delta_{t-1}(j) a_{ji}] b_i(o_1)$

$$\psi_1(i) = \underset{1 \le j \le N}{argmax}\left[\delta_{t-1}(j) a_{ji}\right], t = 2,3 \ldots T;\ i = 1,2 \ldots N$$

6.结束：

$$P^* = \max_{1 \le i \le N} \delta_t(i)$$

$$i_t^* = \arg\max_{1 \le i \le N} [\delta_t(i)]$$

3 数据处理

3.1 数据准备

本文共有五个数据集。

游记攻略集共有50条数据，涉及8个指标，包括“游记ID”、“城市”、“游记标题”、“发布时间”、“出行天数”、“人物”、“人均费用”、“正文”。

酒店评论集共有50条数据、涉及6个指标，包括“评论ID”、“城市”、“酒店名称”、“评论内容”、“入住日期”、“入住房型”。

景区评论集共有50条数据，涉及4个指标，包括“评论ID”、“城市”、“景区名称”、“评论内容”。

餐饮评论集共有50条数据，涉及6个指标，包括“评论ID”、“城市”、“餐饮名称”、“评论等级”、“评论内容”、“标题”。

微信公众号文章共有80条数据，涉及3个指标，包括“文章ID”、“文章标题”、“公众号文章内容”。

3.2 文本预处理

3.2.1 数据清洗

数据清洗(Data cleaning)对数据进行重新处理、检查和校验的过程，目的是去重、纠错，保持数据一致性。

本次数据挖掘的数据清洗的步骤，主要包括文本去重、压缩去词、短句删除、去除无效评论等。

本次数据挖掘在数据清洗过程中：

1、去除完全相同的数据，可被认定为“刷评”。处理的原则是：利用Python程序判断语料库中是否存在重复的中文文本，若存在，则保留一条完全重复文本。

2、去除酒店评论中，评论时间早于入住时间的评论认定为无效评论。处理原则是：利用Python程序在的datetime库判断评论时间是否小于入住时间，若存在，则去除该文本。

3.2.2 中文分词

1.jieba 分词

在文本数据挖掘过程中，一个成熟的中文分词算法往往能够帮助计算机深刻理解负责的中文语言。目前，主流的中文分词工具有：Jieba 分词、NLPIR 分词系统、HanLp 分词、Snow NLP、北京大学 PK Use、哈工大 LTP、p useg、Baidu Lac、等。

本文使用的是 jieba 分词，jieba 库是一款优秀的 Python 第三方中文分词库，jieba 支持三种分词模式：精确模式、全模式和搜索引擎模式。精确模式能够将语句最精确的切分，不存在冗余数据，适合做文本分析，故本文采用的是精确模式。

2.正则化

中文文本常出现数字、字母和标点符号，在使用 jieba 分词时，jieba 会自动的将其划分，增加了 jieba 分词的运行速度，影响后续的文本检索速度，因此，采用正则的手段将其去除。本次正则化采取 re 库的 re.sub 方法。

3. 去停用词

当我们利用 jieba 进行中文分词时，主要是句子中出现的词语都会被划分，而有些词语是没有实际意思的，大量无用的数词、量词、连词、助词等，例如“啊”，“的”、“和”等，对语义作用很小。对于后续的关键词提取就会加大工作量，并且可能提取的关键词是无效的。因此，去除停用词可以有效提高中文文本的检索效率以及中文文本检索的效果，使分词后的结果更加精确，同时去除长度为 1 的词。所以在分词处理以后，我们便会引入停用词去优化分词的结果。本次数据挖掘去停用词采用哈工大的停用词表。

4 文本分类

4.1 自定义训练集

采用爬虫技术在网上抓取文本数据作为训练集。与文化旅游相关的文本数据作为正样本，与文化旅游不相关的数据作为负样本。

4.2 TF-IDF 与 KNN 文本分类

KNN 算法，是一种基本的分类算法。其主要原理是：对于一个待分类数据，将其与一组已经分类标注好的样本集合进行比较，得到距离最近的 K 个样本，K 个样本最多归属的类别，就是待分类数据的类别。

4.2.1 空间距离

KNN 算法的关键是要比较测试集和训练集样本数据之间的距离，则我们采用提取数据的特征值，根据特征值组成一个 n 维实数向量空间/特征空间。然后计算向量之间的空间距离。我们采用了欧氏距离进行空间之间的距离计算。

对于数据 X_i 和 X_j，其特征空间为 n 维师叔向量空间 R^n，则其欧氏距离计算公式为：

$$d(xi,yi)=\sqrt{\sum_{k=1}^{n}(xik-xjk)^2}$$

4.2.2 文本特征

由于计算欧氏距离需要知道文本数据的特征向量，即提取文本关键词，TF-IDF 算法是提取文本关键词的算法之一。该算法由 TF 和 IDF 两部分构成。算法原理于本文 2.1。即某些词在与文化旅游相关的正样本中出现的很频繁，而在负样本出现的不频繁，这些词就是文化旅游相关的关键词。例如，“文化”、“康养”、“乡村旅游”等高频出现在正样本中，而在整个样本集中，并没有频繁出现，则定义这些词为关键词。

本次数据挖掘中，使用自定义训练集提取关键词，利用训练集关键词的词频构造特征向量，再利用空间距离计算公式计算与其他文档的距离，结合 KNN 算法实现文档的自动分类。

5 提取旅游产品

5.1 中文命名实体识别

中文命名实体识别（Chinese Named Enity Recognition），即识别出中文文本中的某些有意义的实体，例如：人名、地名等，将识别的词在文本序列中标注，得到一个具体意义的文本序列。本次数据挖掘中文命名实体识别使用的是隐马尔可夫链（HMM 模型）与维特

比算法相结合的模型。
HMM 模型是概率图模型的一种，属于生成模型。算法介绍于上文 2.4。

隐藏状态集合标签如下。

序号	标	意义
1	B-LOC	地点前缀
2	I-LOC	地点后缀
3	B-ORG	机构组织前缀
4	I-ORG	机构组织后缀
5	B-PRE	人名前缀
6	I-PRE	人名后缀
7	O	无意义标签

基于隐藏状态集合，建立自定义词典，作为该模型的训练集。（仅标注景区地点），使用维特比算法解码。
再使用 python 提取出 B-LOC 与 I-LOC 标签的文字，组成文字序列，最后提取旅游产品。

5.2 提取旅游产品

对酒店评论、酒店评论、餐饮评论旅游产品提取，并合并游记攻略提取出的旅游产品作为总的数据集。由于微信公众号新闻表中数据包含的旅游产品过于少，因此不从中提取旅游产品。去除产品名称相同的旅游产品，仅保留第一条旅游产品名称，并给予相应的 ID

编号。

6 热度评价

6.1 热度指标

旅游产品热度能够反映出旅客在某段时间普遍关注的旅游产品，因此，在定义热度计算公式时，需要综合考虑时间段、产品出现频次、以及产品相关评论的情感倾向。本文选取以下指标进行热度评价，针对每一个旅游产品，分别计算其热度指标：

(1)旅游产品在每一年评论中的出现频次 a，出现频次是热度的重要表现；

(2)旅游产品在每一年评论中的好评数量 b，好评数量是人们对旅游产品的喜好一个重要指标，体现出一个旅游产品的热度；

(3)旅游产品相关评论的情感得分 c，情感值越接近一，对景点的喜好程度越高，也反映出了产品有更高的热度

则旅游产品的热度公式如下式：

$$L = 3a + 2b + ac$$

6.1.1 情感得分计算

中文情感分析是对带有主观意义的中文文本进行分析转化情感程度（情感值）的过程。

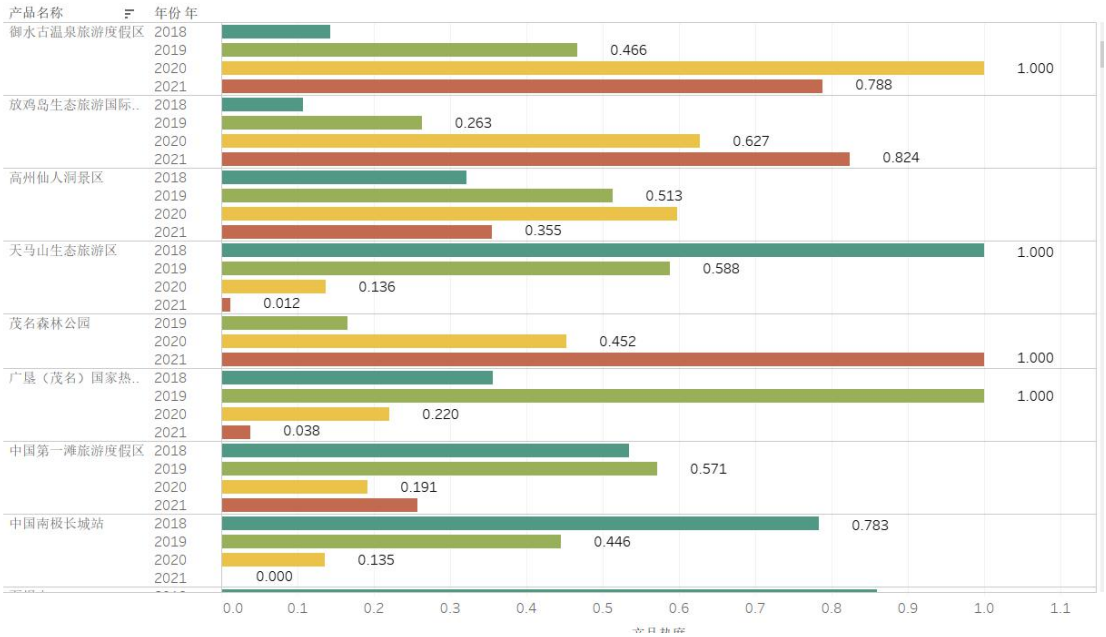
当前的中文情感分析技术主要有三类，一类是根据情感词典配对计算情感数值的方法；一类是使用机器学习的方法，如：Bayes、SVM 等；一类是使用深度学习方法，如：LSTM、CNN 等。由于汉语与英语在语法构造上不同，现今大部分的自然语言处理库基本都针对于英文，国际上通用的 NLP 库在中文文本处理上表现并不优秀，本次使用 SnowNLP 可以方便的处理中文文本内容。SnwoNLP 库中的情感分析方法是基于情感词典配对计算情感数值。

SnowNLP 使用通用的情感分析模型实现，可以进行对本次数据挖掘中评论内容的情感值计算。SnowNLP 计算的情感值得分取值范围在 0-1 之间，越接近 1 表示正面情绪；越接近 0 表示负面情绪。

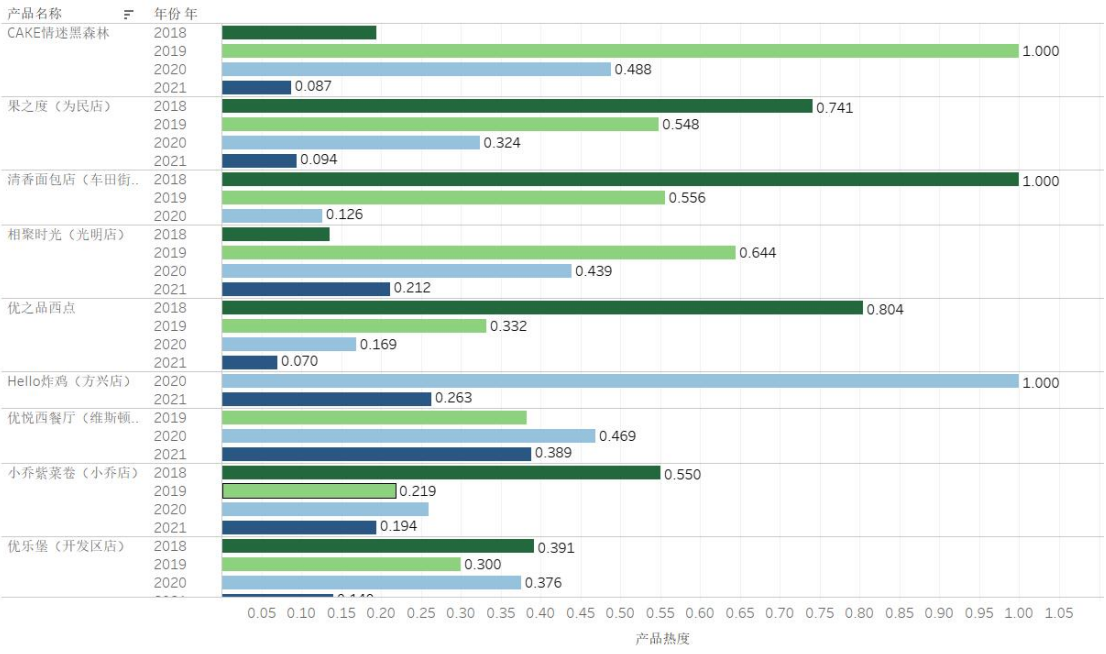
6.1.2 热度计算

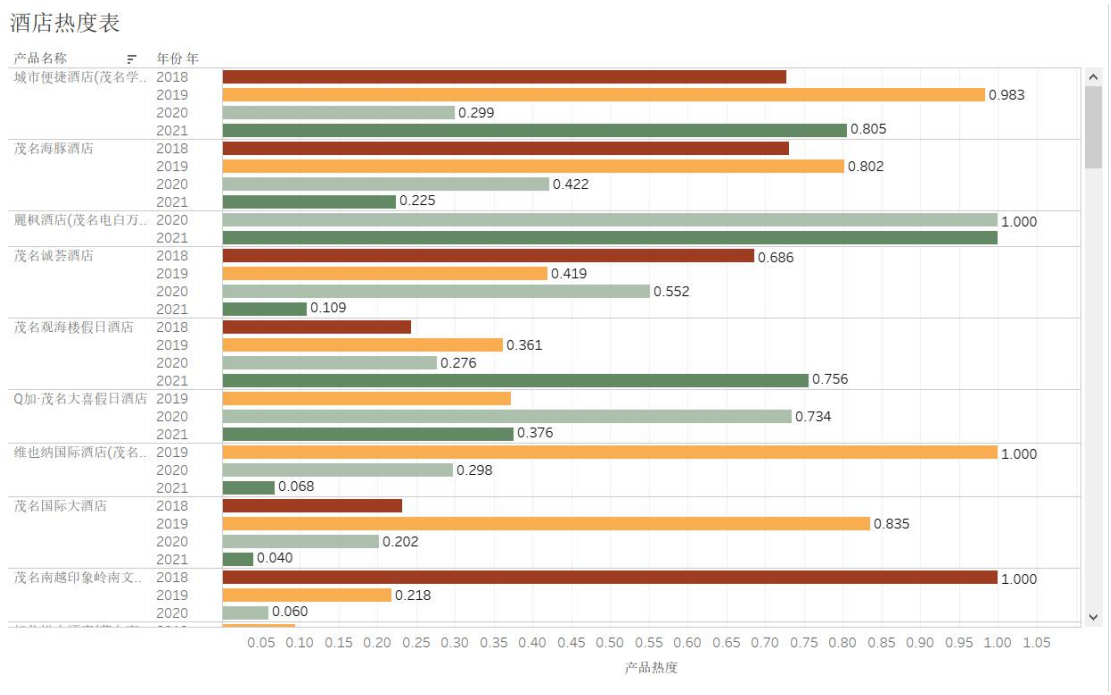
基于以上提出的热度评价指标体系，计算出各旅游产品的年度热度值，如下表：

景区热度表


产品名称
年份 年
御水古温泉旅游度假区
2018
2019
0.466
2020
1.000
2021
0.788
放鸡岛生态旅游国际..
2018
2019
0.263
2020
0.627
2021
0.824
高州仙人洞景区
2018
2019
0.513
2020
2021
0.355
天马山生态旅游区
2018
1.000
2019
0.588
2020
0.136
2021
0.012
茂名森林公园
2019
2020
0.452
2021
1.000
广垦（茂名）国家热..
2018
2019
1.000
2020
0.220
2021
0.038
中国第一滩旅游度假区
2018
2019
0.571
2020
0.191
2021
中国南极长城站
2018
0.783
2019
0.446
2020
0.135
2021
0.000
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1

餐饮热度表


产品名称
年份 年
CAKE情迷黑森林
2018
2019
1.000
2020
0.488
2021
0.087
果之度（为民店）
2018
0.741
2019
0.548
2020
0.324
2021
0.094
清香面包店（车田街..
2018
1.000
2019
0.556
2020
0.126
相聚时光（光明店）
2018
2019
0.644
2020
0.439
2021
0.212
优之品西点
2018
0.804
2019
0.332
2020
0.169
2021
0.070
Hello炸鸡（方兴店）
2020
1.000
2021
0.263
优悦西餐厅（维斯顿..
2019
2020
0.469
2021
0.389
小乔紫菜卷（小乔店）
2018
0.550
2019
0.219
2020
2021
0.194
优乐堡（开发区店）
2018
0.391
2019
0.300
2020
0.376
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
产品热度



7 关联度分析

7.1 数据预处理思维--词袋模型（Bag of Words Model）

BoW(Bag-of-words) 词袋模型是 n-gram 语法模型的特例 1 元模型，该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的。BoW 使用一组无序的单词(words)来表达一段文字或一个文档：

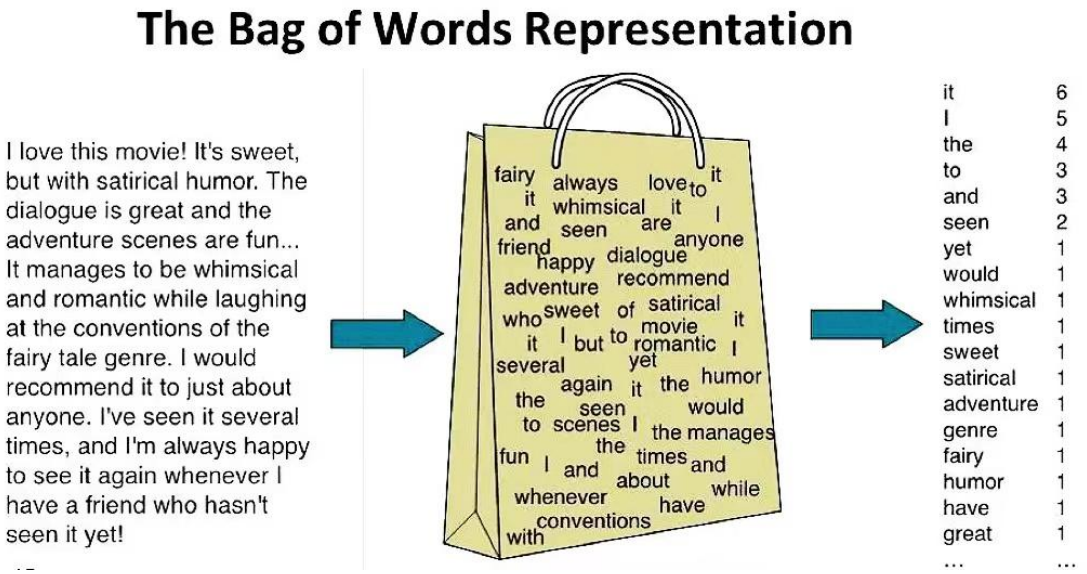


表 1 词袋模型图解

下面给出三个文本如下：I love my boyfriend. Mary likes her boyfirend. Helen likes palying her scoer.

根据三个文本出现的单词，构建词典：{"I": 1, "love": 2,"my": 3, "boyfriend": 4, "Mary": 5,"likes": 6, "her": 7, "Helen": 8,"playing": 9, "scoer": 10}

上面的词典中包含10个单词，每个单词具有唯一的索引，针对每个文本我们使用一个10维的向量来表示。如下：

[1, 1, 1, 1, 0, 0, 0, 0, 0, 0]
[0, 0, 0, 1, 1, 1, 1, 0, 0, 0]
[0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

向量的维度是依据词典中不重复的单个词的数量确定的；向量中词序与原文本词序没有任何关系,是不与词典中的顺序对应的；单个词向量下的数字表示单词在文本中出现的频率（词频）。

7.1.2 词袋模型实现方式

1.Pythoncollections.defaultdict

语法格式：

```
collections.defaultdict([default_factory[,...]])
class defaultdict(Dict[_KT, _VT], Generic[_ KT, _VT]):
default_factory: Callable[[],_VT]
```

该函数返回一个类似字典的对象。defaultdict是Python内建字典类（dict）的一个子类,它重写了方法_missing_(key),增加了一个可写的实例变量default_factory,实例变量default_factory被missing0方法使用，如果该变最存在，则用以初始化构造器;如果没有，则为None。其它的功能和dict一样。

第一个参数为default_factory属性提供初始值,默认为None;其余参数包括关键字参数（keyword arguments）的用法，和dict构造器用法一样。

2.one-hot encoding

One-Hot编码，又称为一位有效编码，主要是采用N位状态寄存器来对N个状态进行编码，每个状态都由他独立的寄存器位，并且在任意时候只有一位有效。

One-Hot编码是分类变量作为二进制向量的表示。这首先要求将分类值映射到整数值。然后，每个整数值被表示为二进制向量，除了整数的索引之外，它都是零值，它被标记为1。

即one hot编码是将类别变量转换为机器学习算法易于利用的一种形式的过程。

对于定类类型的数据，建议使用one-hot encoding。定类类型就是纯分类，不排序，没有逻辑关系。但注意，一般会舍去一个变量，比如男的对立面肯定是女，那么女就是重复信息，所以保留其中一个变量即可。

本项目的产品数据即为定类类型的数据，相互之间没有逻辑关系。所以我们这里对问题2得出的产品数据进行one-hot编码,得出one-hot数组和产品字典编号字典，如图（这里截取一小部分作为示范）：

defaultdict(<class 'int'>, {'鼎龙湾德萨斯水世界': 0, '南国热带花园': 1, '龙眼树': 2, '潘茂名纪念公园': 3, '东湾水': 4, '维纳斯皇家温泉（广东阳西店）': 5, '御水古温泉旅游度假区': 6, '优悦西餐厅（维斯顿店）': 7, '茂名安途四季酒店': 8, '龙胆石': 9, '浪漫海岸旅游度假区': 10, '信宜天鹅湖宾馆': 11, '仙湖山庄': 12, '悦创小鹅鹅桂林米粉（双山一路店）': 13, '椒王火锅（高州店）': 14, '长岗坡渡': 15, '茂名柏丽酒店': 16, '电白郡旧址': 17, '三角圩': 18, '海滨公园': 19, '新安镇': 20, '巽寮湾': 21, '沉香精油': 22, '窦州古城': 23, '盖路步行街大': 24, '尊霸披萨': 25, '茂名茂南高山九树公寓': 26, '仙人阁': 27, '红树林湿地': 28, '茂名诚荟酒店': 29, '陵岛海': 30, '茂名龙栖湾旅馆': 31, '天龙顶国家山地公园': 32, '尊宝比萨（康泰又一城店）': 33, '唐华商务酒店(茂名高铁火车站店)': 34, '仙道观': 35, '高州顺得商务酒店': 36, '博贺湾大': 37, '大路街': 38, '露天矿生态公园': 39, '茂名化州丽登酒店': 40, '镇隆镇': 41, '保利银滩浴场': 42, '茂港区': 43, '古郡水城': 44, '元晟坊蛋糕（南香公园店）': 45, '南三岛': 46, '博贺湾': 47, '稻田公园': 48, '盛香烧鹅（东方市场店）': 49, 'Hello炸鸡（方兴店）': 50, '潭江半岛酒店': 51, '潮州风吹岭石刻群': 52, '顺德火焰醉鹅坊（站北五路店）': 53, '沉香酒': 54, '大排档': 55, '灵王庙': 56, '鳌头古镇': 57, '俗文化馆': 58, '精途酒店(茂名高铁火车站店)': 59, '海洋王': 60, '康小屋': 61, '塔斯汀中国汉堡（民主店）': 62, '浪漫海岸度假区': 63, '沉香林': 64, '御水谷温泉': 65, '庐山村': 66, '大王岭': 67, '海滨公园游乐场': 68, '好莱登商务宾馆(信宜绍秀体育馆店)': 69, '白水瀑布': 70, '摩天小镇': 71, '南华广场': 72, '新湖公园': 73, '文化遗': 74, '麗枫酒店(茂名高铁站店)': 75, '中国第一滩旅游度假区': 76, '御水古温泉景区': 77, '兰欧酒店(信宜绍秀体育馆店)': 78, '东海岛': 79, '温德姆酒店': 80, 'Q加·茂名大喜假日酒店': 81, '袂花镇': 82, '江泽民同': 83, '大通街': 84, '爱可咖啡馆（茂南店）': 85, '露天矿街道': 86, '马贵镇': 87, '长坡镇旧城村': 88, '金鹿园': 89, '覃福庙': 90, '茂名熹龙国际大酒店': 91, '茂名零六六八公寓': 92, '马尾岛': 93, '清香面包店（车田街店）': 94, '云南过桥米线（汇景店）': 95, '徐闻港': 96, '信宜丽晶酒店': 97, '南广场': 98, '东港区': 99, '大唐荔乡': 100, '溪小

表 2 one-hot 数组（一小部分）

```
Out[4]: (array([[0, 0, 0, ..., 0, 0, 0],
                [0, 0, 0, ..., 0, 0, 0],
                [0, 0, 0, ..., 0, 0, 0],
                ...,
                [0, 0, 0, ..., 0, 0, 0],
                [0, 0, 0, ..., 0, 0, 0],
                [0, 0, 0, ..., 0, 0, 0]]),
        defaultdict(int,
                    {'鼎龙湾德萨斯水世界': 0,
                     '南国热带花园': 1,
                     '龙眼树': 2,
                     '潘茂名纪念公园': 3,
                     '东湾水': 4,
                     '维纳斯皇家温泉（广东阳西店）': 5,
                     '御水古温泉旅游度假区': 6,
                     '优悦西餐厅（维斯顿店）': 7,
                     '茂名安途四季酒店': 8,
```

表 3 产品字典编号字典（一小部分）

7.2 关联规则（Apriori 算法）

关联规则是指事物间的相互联系，反映了一个事物与其他事物之间的相互依存性和关联性。如果两个或者多个事物之间存在一定的关联关系，那么其中一个事物就能够通过其他事物预测得到。

一般使用以下三个指标来衡量关联性：

1. 支持度：support，也即旅游产品的流行程度

$$\text{支持度} = \frac{(\text{包含旅游产品A的记录数量})}{(\text{总的记录数量})\ \text{得到了旅游产品A的支持度}} = \frac{(\text{同时包含旅游产品A、B的记录数量})}{(\text{总的记录数量})\ \text{得到了旅游产品(A,B)的支持度}}$$

2. 置信度：confidence，也即出现旅游产品 A 就会出现旅游产品 B 的可能性

置信度（A –> B）= (同时包含旅游产品A和B的记录数) / (包含A的记录数)

3. 提升度：lift,也即当出现一个旅游产品时另一个旅游产品的出现率会增加多少

提升度(A –> B) = 置信度（A –> B）/（支持度A）

7.3 Apriori 算法

Apriori 算法可使用解决于关联挖掘。

生成频繁项集：这阶段需查找所有满足最小支持度的项集（即为频繁项集）。

Apriori 基于以下两条核心原理生成频繁项集：1.如果某个项集是频繁的，则其所有子集是频繁的。2.若子集不是频繁的，则所有包含它的项集都是不频繁的。

生成规则:在上一步产生的频繁项集的基础上生成满足最小置信度的规则，产生的规则称为强规则。

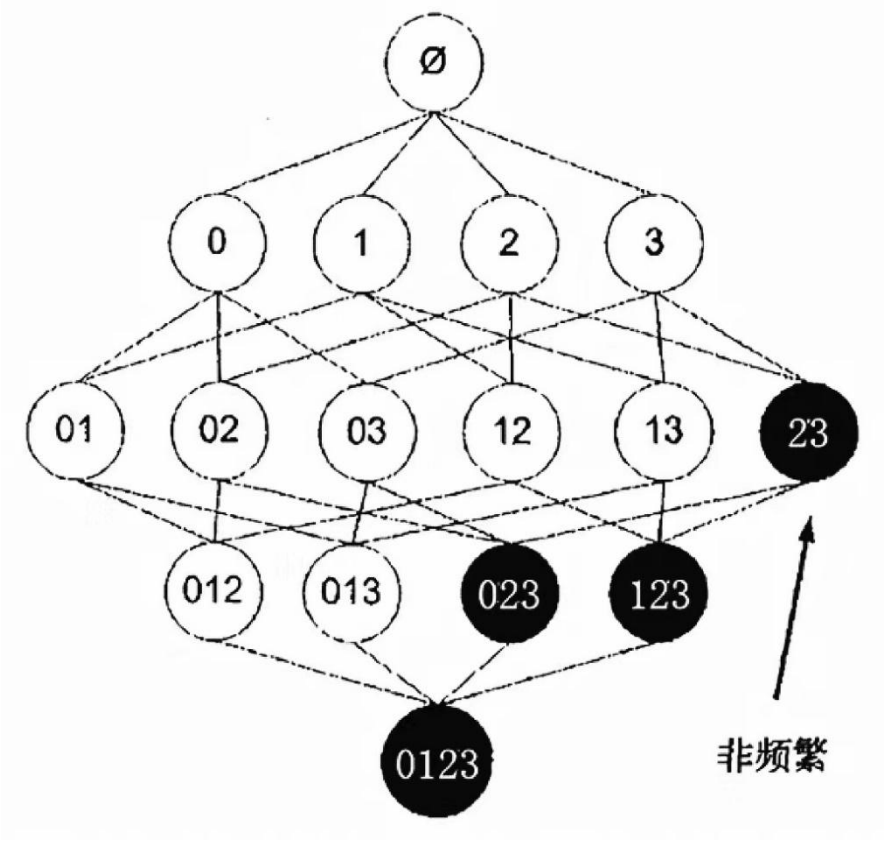


表 4 Apriori 思想

关联规则公式：

$$Connection = （Support * 5 + Confidence * 1/5 + Lift/100）/100$$

其中 Connection 为关联度，Support 为支持度，Cinfidence 为置信度，Lift 为提升度

7.4 关联规则下的旅游图谱

7.4.1 可视化工具--network

networkx 是 Python 的一个包，用于构建和操作复杂的图结构，提供分析图的算法。图是由顶点、边和可选的属性构成的数据结构，顶点表示数据，边是由两个顶点唯一确定的，表示两个顶点之间的关系。顶点和边也可以拥有更多的属性，以存储更多的信息。
对于 networkx 创建的无向图，允许一条边的两个顶点是相同的，即允许出现自循环，但是不允许两个顶点之间存在多条边，即出现平行边。边和顶点都可以有自定义的屈性，属性称作边和顶点的数据，每一个属性都是一个 Key:Value 对。

7.4.2 关联度分析

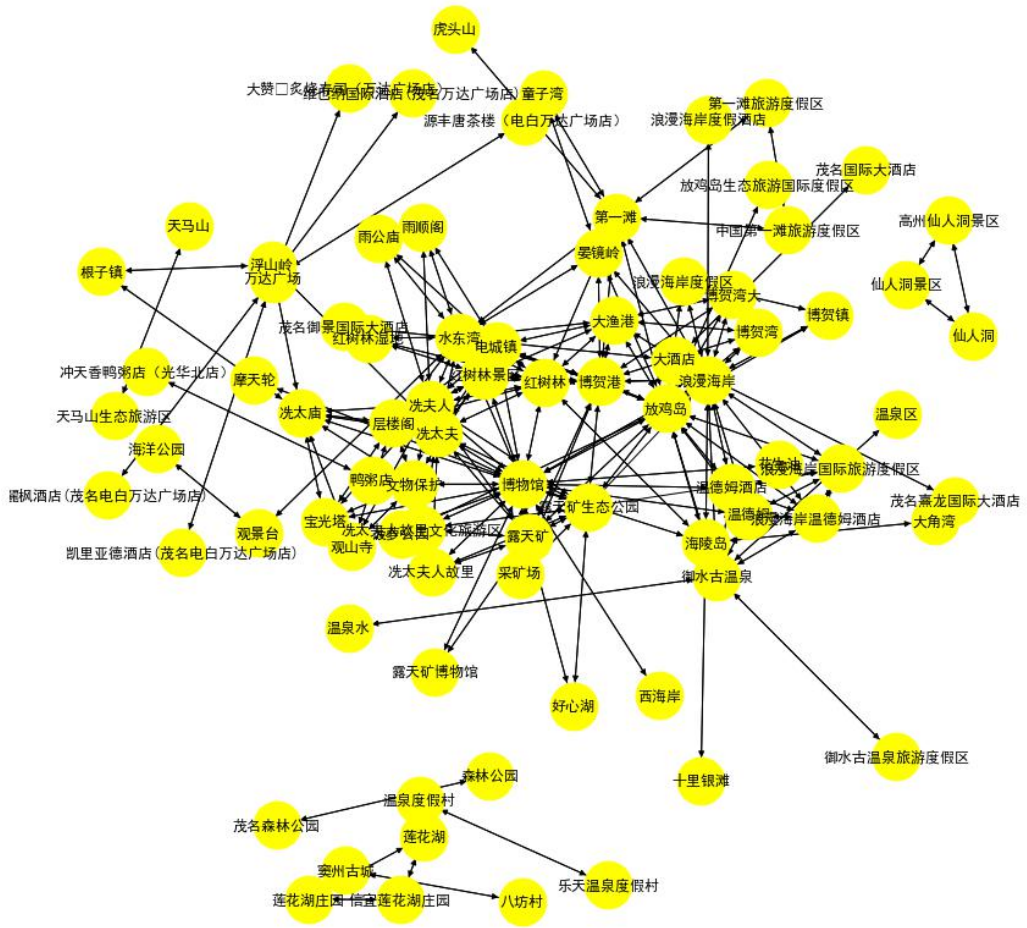


表 5 关联度>0.1 时的旅游图谱

于表 5 关联图可以看出，游客探访古城古镇乡村具有一定的关联度，说明游客对文化

旅游有一定的偏好。推动古城古镇乡村一体发展势在必行，建设文化旅游产业，是旅游行业转型的一道制胜之招。

于表 5 关联图可知，庄园与生态公园、乡村等之间具有一定的关联度，说明游客愈发崇尚养生，崇尚康养旅游。

于表 5 关联图可知，度假村与酒店、餐饮等之间具有一定的关联度，说明游客对便捷旅游也饱含期待。因此，打造休闲旅游一体化，方可促进旅游消费，带动旅游产业经济，吸引更多游客。

2.设置关联度>0.15 时下的旅游图谱，对其进行分析：

于表 6 关联图可知，庄园与生态公园、森林、山等之间具有一定的关联度，温泉、生态旅游区之间具有一定的关联度，说明游客对康养旅游具有一定的向往。在当今时代，受疫情影响，人们越来越注重养生，因此，推动康养旅游也不失为是一种旅游产业转型的好策略，能更好更快的促进当地旅游产业的发展。

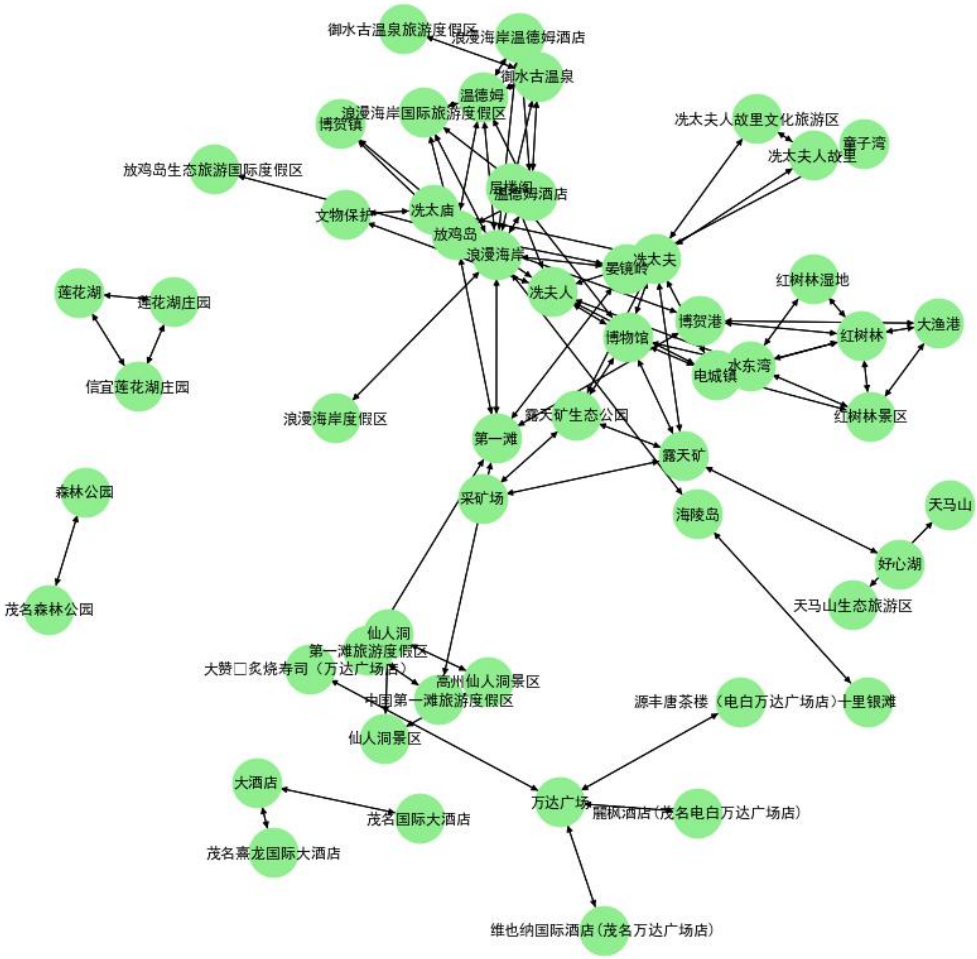
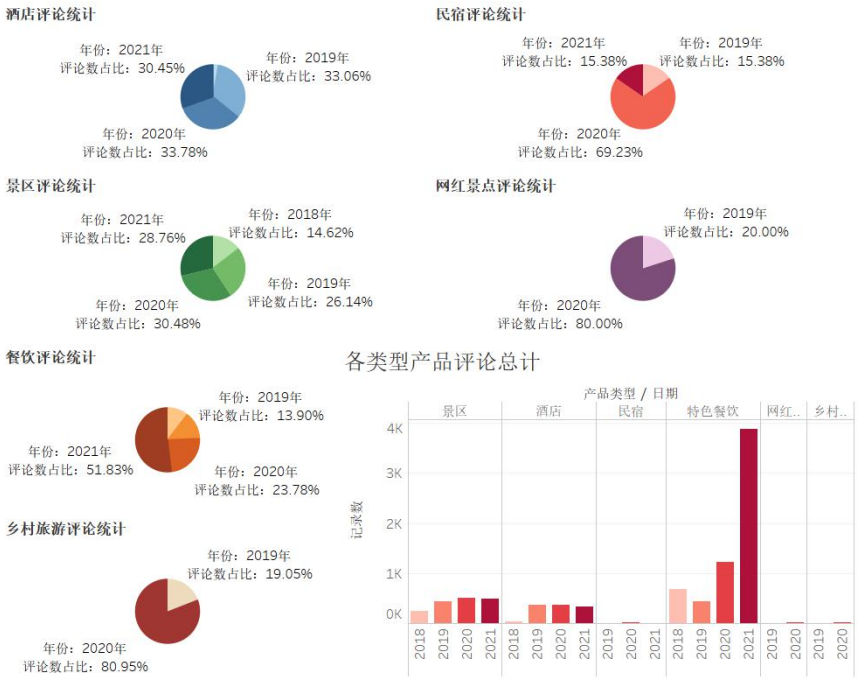


表 6 关联度>0.15 时的旅游图谱

8 分析报告与建议

8.1 新冠疫情前后茂名市旅游产品的变化



新冠疫情后茂名市的旅游产品类型仍然以特产餐饮为主导，而酒店景区的消费者逐步减少，进一步扩大了特色餐饮的消费市场，并且在疫情后也出现了特色餐饮类的旅游产品热度持续高涨的局面，原先热度较高的旅游景区也因疫情原因变得冷清。

8.2 旅游行业发展的政策建议

1.政府部门应当充分发挥出引导作用，支持、鼓励旅游行业的发展并帮助公众树立信心。

通过分析可知，受疫情影响，茂名地区旅游业的发展增速降缓，说明茂名地区旅游业受到重创。其中，促进旅游消费必然会成为疫后恢复经济的首要手段之一，旅游需求消费是人们满足美好生活的刚性需求，而在疫情结束后必然会出现大量报复式的旅游需求反弹。因此，政府部门应大力扶持旅游产业发展，给予旅游产业企业及从业人员培训，提升服务质量，扶持文化和旅游企业渡过难关、推动企业恢复发展、促进文化旅游消费。

2.大力推进文化旅游，提升旅游地区文化涵养，建设具有特色的旅游文化。

在新时代，人民日益追逐美好的生活需要，对文化旅游愈发向往。推动乡村旅游文化建设势在必行，推进乡村与生态、康养、旅游的融合，推进古城古镇与人文的融合，挖

掘当地人文资源，鼓励旅游企业丰富旅游产品创意，改进旅游产业衍生产品设计， 打造乡村与康养旅游相结合、古城古镇与人文风趣相结合的旅游产品。打造让老百姓称赞的好口碑旅游产品，建设具有特色的旅游文化。

3.制定旅游优惠政策，促进旅游业回暖。

疫情后市场必将反弹，旅游业也会迎来复苏阶段，相应的政府当发行旅游优惠政策。发放旅游消费券，吸引客源，才能加速旅游业的反弹。制定一定的优惠策略，促进人流回潮，将热度拉高，对特殊工作人员实现减免政策，带动大众消费，才能加速开拓旅游市场。

4.推进旅游“安全”建设。

疫情后，越来越多的消费者开始重视旅途中的安全性，而确保游客的安全是旅游产业的重中之重，因此，旅游产业应该加强游客的疫情防控和安全防护工作，控制客流量，严格把关卫生条件，巩固安全措施。