C 题：“智慧政务”中的文本挖掘应用

摘要

本文旨在设计“智慧政务”中的自然语言处理技术的应用，根据附件提供的群众留言记录及相关部门的答复意见等数据，实现了对群众留言内容的一级标签分类，并完成了对群众留言热点问题的挖掘和对相关部门答复意见的质量评价。对提升政府的管理水平和施政效率具有重大的意义。

针对问题一，利用深度学习**ERNIE模型**算法，构建了关于留言内容的一级标签分类模型。将预处理后得到的训练样本进行模型的训练，然后将训练好的模型对测试集数据进行测试，得到关于一级标签的分类结果，计算得到混淆矩阵，以及模型的Acc值（准确率）为98.37%，通过F-Score评价模型对七类的F1结果进行总评价，得到**F-Score为0.9820**。

针对问题二，定义了三个指标：簇内留言条数、簇内留言总点赞数和簇内总反对数，建立关于群众留言问题的热度评价模型：**热度指数=**w_1***簇内留言条数+**w_2***簇内留言总点赞数+**w_3***簇内总反对数**。首先通过**TF-IDF算法**对预处理后的数据进行特征提取，其次利用**DBSCAN聚类算法**进行聚类，得到留言聚类结果，最后统计每一簇内的留言数据总量、点赞数和反对数，得到了热度排名，并给去排名前5热点问题，其中**小区旁修建搅拌站扰民**排名第一。

针对问题三，从答复意见的相关性、完整性和可解释性三个角度，建立了基于**多特征融合**的答复意见质量的评价模型：**质量评价指标=**w_1***相关性+**w_2***完整性+**w_3***可解释性**。首先通过构建每条答复意见的**字向量**文本，其次通过**编辑距离**求文本相似度，得到相关性指标，然后定义了完整性和可解释性两个特征，最后基于多特征融合的评价模型即可得到答复意见质量的综合评价指数。

关键词：ERNIE；TF-IDF；DBSCAN聚类；编辑距离；多特征融合

目　录

一、问题的重述

1.1 问题的背景

随着人们生活水平的提高，对生活质量也有了相对高的要求，而逐渐流行起来的旅游行业恰巧可以改善人们的生活质量，缓解人们的压力改善人们的心情。而在疫情当下的情况之下，人们对旅游的渴望逐渐增加，但为了疫情的更好的防疫和当地疫情的严重情况，很多人选择的则是在当地选择旅游。既保证了疫情的防控也能满足人们渴望旅游的心情。旅游行业的客流量急剧增加，更推动了当地经济在疫情的影响下的快速复苏。

在疫情影响下的当地旅游环境下，人们旅游的方式已经发生了很大的改变。然而如何对旅游者选择所旅游的地方的因素进行判定，基于对微信公众号文章内容的吸引，旅游地周边产品的喜爱程度在互联网上的数据进行分析。基于一系列的数据分析，本文依靠相关的机器学习方法，在大量的文本数据中挖掘合理的信息，并为微信公众号文章内容与旅游是否相关和旅游周边产品的喜爱度进行合理的分析。

1.2 解决的问题

1. 根据附件中所提供微信公众号推送文章的内容的数据，判断其内容是否与旅游相关，并把判断后的结果以“相关”和“不相关”两类的形式进行保存，结果的形式保存为文件“result1.csv”。

2. 从数据中分别对景区评论、酒店评论、游记旅游、餐饮评论数据进行提取，并鉴别其内容是否有用将提取出的旅游产品和所依托的语料以表 2 的形式保存为文件“result2-1.csv”。建立相关的旅游产品的模型，对提取出的旅游产品按年度进行热度分析，并排名。将结果以表 3 的形式保存为文件“result2-2.csv”。

3. 对问题二中提取出的旅游产品，进行对相应的关联度分析，并且找出以景区、酒店、餐饮等为核心的强关联模式，结果以表 4 的形式保存为文件“result3.csv”。

4. 从旅游图谱和附件数据中的旅游产品的数量的变化，写一封信件向当地

旅游主管部门提出旅游行业能够更好的发展的政策建议。

二、问题的分析

2.1 问题一的分析

附件中给出的数据，对微信公众号新闻内容单独提取出来，问题一所需要解决的问题是根据推送文章根据的内容判别是否与文旅相关，对文本进行分类，可以把该题归为分类的一类题型。

在对数据进行数据预处理，利用结巴库对数据中词组进行分词，并用 TF-IDF 计算词组的权重，在用主成分分析法对数据进行降维，再利用 K-Means 聚类对数据进行分类，对每个分类的簇分别画出词云图，再分析是否与文旅相关，再对其相关的簇对应的数据进行分类，并把结果保存为 result1.csv。

2.2 问题二的分析

第一小问 对附件中酒店名称数据，景区名称数据，餐饮名称数据，游记攻略数据中的信息进行提取。酒店名称数据，景区名称数据，餐饮名称数据可以直接提取出酒店名称、酒店评论 ID 等相应的列，并把列名称转换成产品名称、语料 ID 等，对相应的产品 ID 进行设定。

游记攻略数据没有对应产品数据的列，则需要对这该数据进行单独提取，用 excel 对旅游攻略数据中旅游产品名称提取出来。

第二小问 该问题基于第一小问，需要对其建立相应的评分模型，按年度对产品数据进行时间性的排名，再以相应的评价热度的评价指标进行产品热力度的计算，并对相应的数据进行排序。

2.3 问题三的分析

基于问题二中提取出的旅游产品数据，对提取出来的旅游产品进行关联分析，找出对关联分析的指标，基于 apriori 算法的关联规则：支持度、置信度、提升度[4]，分别对景区、酒店、餐饮等的数据算出其相应的关联度值，并对运行得出

的数据结果保存为文件“result3.csv”。

2.4 问题四的分析

根据历史数据，从旅游图谱的数据关联度分析，可以对疫情前后旅游产品的变化进行分析。而分析疫情前后的数据，则是分别对附件 2018 到 2019 的数据为疫情前的数据，附件 2020 到 2021 的数据为疫情后的数据。从附件中的酒店评论、景区评论、游记攻略、餐饮评论等数据中的产品数量的变化进行叙述。

三、模型的假设

1. 假设附件中的数据可靠性强
2. 假设 K-Mean 聚类时，误差不影响最后数据的分类
3. 不考虑外界因素对文本数据的影响
4. 忽略数据的样本量少，带来的个体独立数据带来的误差

四、符号说明

符号	符号说明
$x_{normalization}$	数据分归一化后的数值
i	旅游产品
$D(i)$	旅游产品的词频
$num1(i)$	i 的数量
$R(i)$	旅游产品热度值
X 、 Y	某两种旅游产品数据
$num(I)$	总词组集的个数
$Support(X \to Y)$	支持度
$Confidence(X \to Y)$	置信度

$Lift(X \to Y)$	提升度

五、模型与建立与求解

附件中给出的数据中是基于茂名城市旅游的数据，该题给出的数据中分别包括 2018 年到 2021 年的数据，对其数据进行相应的拼接，其中酒店评论的数据有 6059 条，景区评论 1203 条，旅游攻略 294 条，餐饮评论 6984 条，微信公众号新闻 6286 条。对所给出的数据进行处理，并解决所提出的问题。

5.1 问题一的模型的建立与求解

5.1.1 数据预处理

数据的预处理，是对数据进行合理的删减处理，使需要分析的数据合理化。

（1）异常值的处理

问题一 对两个附件数据中的微信公众号新闻内容的单独提取，在合并，使用 Jupyter notebook 对数据进行处理，用 python 中的 Data.info()函数对数据进行数据基本信息统计，检查是否含有缺失值和异常值。

```
---  ------  --------------  -----
0   文章ID    6286 non-null   int64
1   公众号标题   6285 non-null   object
2   发布时间    6286 non-null   object
3   正文      6286 non-null   object
```

图 1 数据的基本统计信息

经过 python 代码的运行，发现微信公众号新闻内容含有缺失值和重复值，此时考虑到的则是删除或者忽略，然而问题一所需解决的相关性划分则是根据其文章 ID 来划分的，而每个文章 ID 是唯一确定的，则可以忽略其中的异常值。

（2）文本数据的合并

微信公众号新闻内容中公众号的标题，对划分是否跟文旅相关也有十分大的决定性，由于微信公众号新闻内容的数据中其中正文并未包含与文旅相关的词组，但微信公众号新闻内容中公众号的标题中却含有相关的词组，所以需要对正文跟

公众号的标题进行合并，组成一个新的内容，以方便之后对数据的处理。

（3）文本词组特殊符号的处理

抽取数据的前 5 行数据（如图 2） 图中的数据表示为微信公众号新闻内容所对应的正文和标题部分的数据，发现正文和标题合并的内容中含有较多的英文字母、数字、特定的日期等，为提高提取数据的准确性，需要对其进行删除。

文章ID	公众号标题	发布时间	正文	标题or正文
1001	2018，对自己好一点	2018-01-02 17:28	2017的旅程已经结束\n2018的未来拉开了帷幕\n新的一年里，请对自己好一点\n一辈子很...	2018，对自己好一点\n2017的旅程已经结束\n2018的未来拉开了帷幕\n新的一年里，...
1002	春节机票预订有窍门	2018-01-02 17:28	距离春节还有一个多月的时间，在线旅游网站的春节机票销售火爆，部分航线甚至一票难求。在这里分享...	春节机票预订有窍门\n距离春节还有一个多月的时间，在线旅游网站的春节机票销售火爆，部分航线甚...
1003	冬日旅游知多D	2018-01-03 17:32	960万平方公里的祖国大地，四季都有独特美景\n冬天的旅行也别有一番风味\n但是冬季的严寒气...	冬日旅游知多D\n960万平方公里的祖国大地，四季都有独特美景\n冬天的旅行也别有一番风味\...
1004	2018冬季暖心之旅	2018-01-03 17:32	长按二维码，关注我们\n中心联系人：林小姐13709649096\n刘小姐135000781...	2018冬季暖心之旅\n长按二维码，关注我们\n中心联系人：林小姐13709649096\n...
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		关于粤K27618号大客车排气管“喷火”事件的情况说明\n ...

图 2 提取前 5 行的数据

5.1.2 问题一模型的建立与求解

经过预处理过后的数据，利用 python 中的结巴库对其数据进行分词，观察到数据中有大量的词组不跟文旅相关，且含的数量较多，需要对这些词组进行删除，因此获取当前的中文停用词，对其中的字符串中的词组进行删减。由于结巴库不能准确的对一些特定性的词组进行分词，所以需要对数据中的特定词进行处理，导入相关的特定词组使数据更加准确。

TF-IDF 算法

通过 TF-IDF 算法对文本数据中词组的权重进行计算，TF-IDF 是一种可以用来对信息进行检索和数据挖掘的一种加权技术,可以用来评估一个词组对一个语料库中某一文件的重要程度。词组的重要性随着它出现的次数成正比增加，但却与语料库中出现的频率成反比下降。

TF 可以定义为词组的词频，IDF 为词组的逆文本频率指数，该权重为 IDF 词组逆文档频率，其等于总词组数量除以包含词语的数量，再将得到的商取对数得到。如果包含词条的文档越少，IDF 越大，则说明词条具有很好的类别区分能力[1]。

TF 公式为：

$$TF=\frac{\text{词汇在文本中出现的次数}}{\text{文本词汇的总个数}}\ IDF=\log\left(\frac{\text{语料库中文本的总个数}}{\text{包含该词汇文本个数}+1}\right)$$

即：

$$TF-IDF=TF*IDF$$

得到一个词组的 TF-IDF 值，该词组对文章的重要性越高，其 TF-IDF 值就越大，所以运算出的数据结果出现在最前面的几个词就是文本数据内容的关键词。利用 TF-IDF 算法计算得出的部分数据结果（如图 3）：

```
(0,  83041)     0.0821678253108874
(0,  11621)     0.032537338512450144
(0,  52975)     0.07816529694760235
(0,  95198)     0.05984061360247046
(0,  18851)     0.033202750444437525
(0,  25579)     0.07149191463828795
(0,  95744)     0.029546160451663597
(0,  16662)     0.2204826788449964
(0,  66285)     0.033229692478311215
(0,  55531)     0.07903652789273553
(0,  24600)     0.05880320269465346
(0,  36595)     0.043263993221607816
(0,  91805)     0.11968122720494093
(0,  34500)     0.1455694152408417 2
(0,  16879)     0.05663184556455246
(0,  54246)     0.05663184556455246
(0,  75930)     0.11876770216548517
(0,  4717)      0.025999802910347106
(0,  664)       0.05344280490154544 6
(0,  85253)     0.05176494009184178 7
(0,  7641)      0.0821678253108874
```

图 3 部分数据的 TF-IDF 算法计算的值

运行出的结果的行和列分别是 6286 条和 97527 条，由于数据集数量庞大，数据间的特性存在一定的关联和相关性，且有些数据的内容含有重合，导致数据内容的重复和处理数据的效率低，于是我们可以用更高效的方法处理数据，提取原数据的一部分数据特征来反应当下数据的特征，所以需要用主成分分析（PCA），对数据进行降维处理，把数据的特征数量降维到 50，并计算出每一个降维后的数据，由此可以提高计算的效率和准确性[2]。

基于 K-Means 的数据聚类分析

运行代码后得到 PCA 降维后的数据特征的数据，对其数据进行 K-Means 聚类分析，算出对应的类别。并挖掘和分析该数据，得到对应 K-Means 分类的基本流程图（如图 4）：

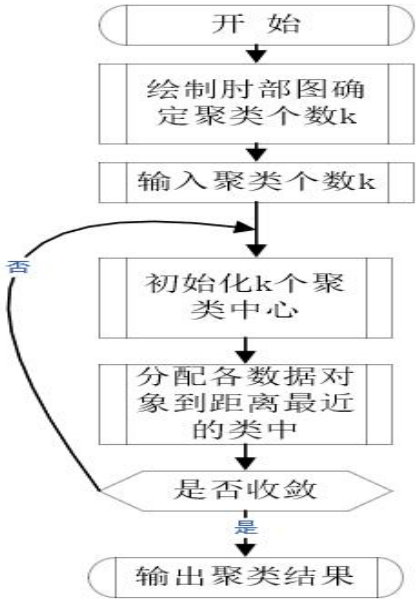

图 4 K-Means 聚类的流程图

（1）确定聚类 k 的个数

常用的方法有肘部法则和轮廓系数法等，肘部法是根据损失值下降平稳的拐点来确定的，轮廓系数法是根据轮廓系数的最大值进行判断。而该题我们选用肘部法则来计算 k 的数量。

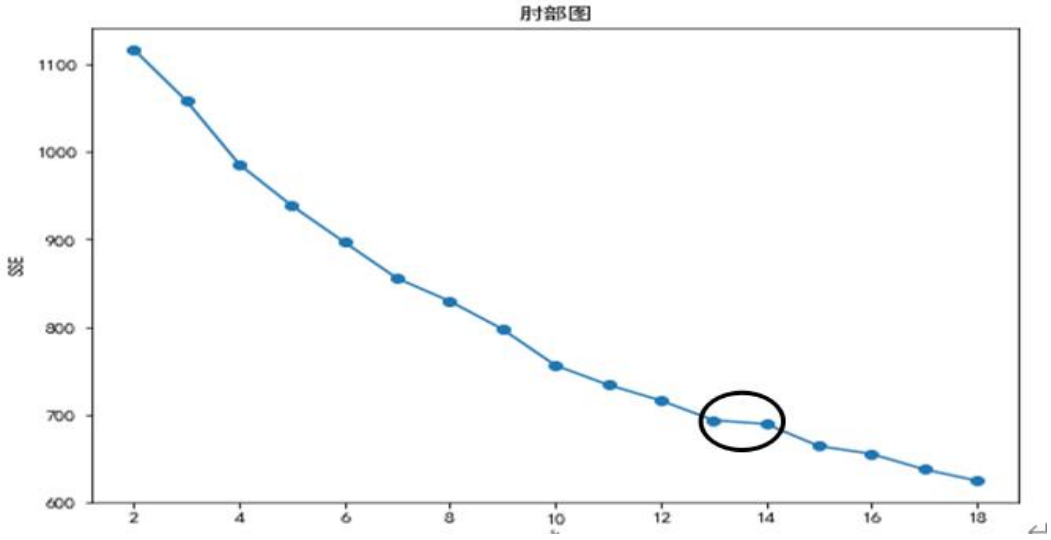

图 5 肘部图

基于图 5 的肘部图，由图中的曲线的平滑程度，在 k 等于 13 的时候，曲线趋于平缓，则可以确定聚类 k 的个数为 13。

（2）K-Means 聚类

确定 k 为 13 之后，可以对特征数量数据进行聚类，分别算出每个簇含有的数据数量。第 0 簇为 118 条，第 1 簇为 282 条，第 2 簇为 796 条，第 3 簇为 290 0 条，第 4 簇为 349 条，第 5 簇为 239 条，第 6 簇为 205 条，第 7 簇为 457 条，第 8 簇为 249 条，第 9 簇为 133 条，第 10 簇为 124 条，第 11 簇为 272 条，第 1 2 簇为 99 条。

生成簇的词云图

分别对 K-Means 聚类的每一个簇的数据生成词云图，如图 6 为第 0 簇到第 5 簇的词云图，图 7 为第 6 簇到第 7 簇的词云图。并展示相应的词云图（图 6、图 7）：

图 6 簇 0 到簇 5 的词云图

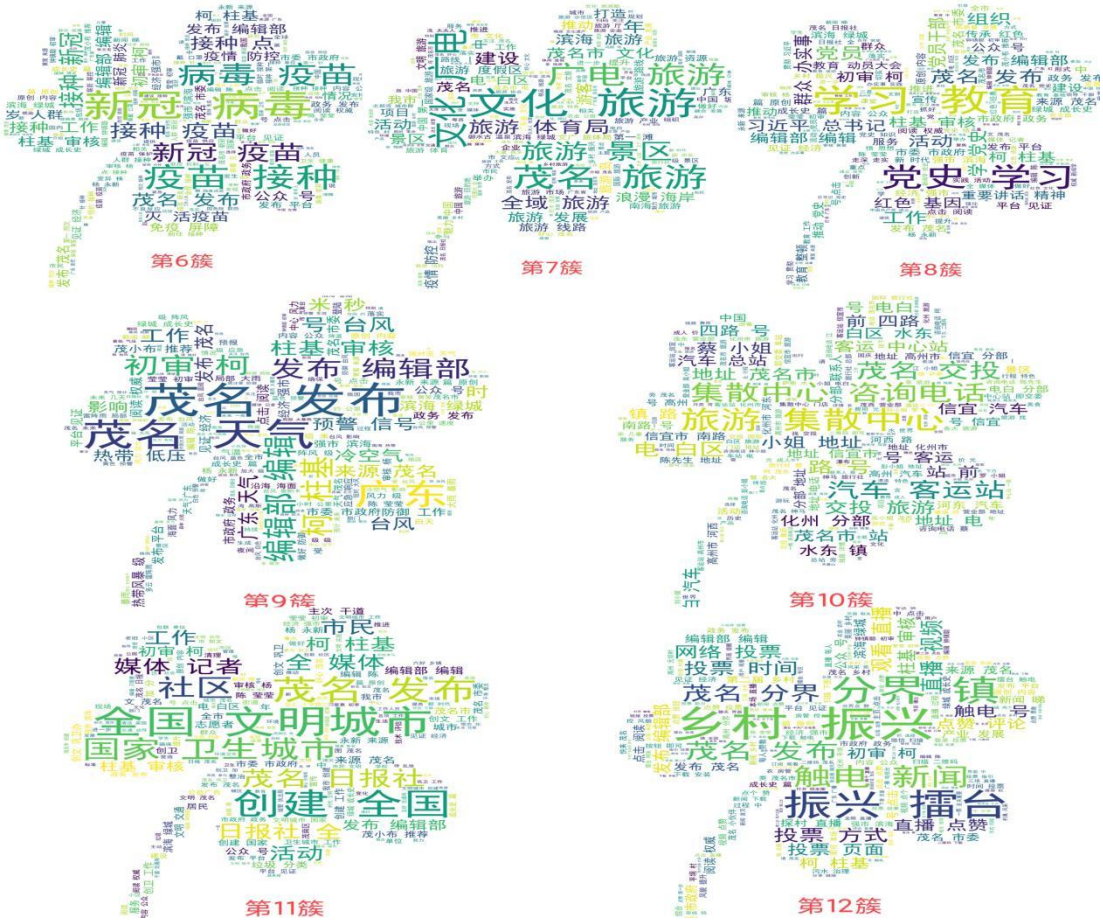

图 7 簇 6 到簇 12 的词云图

根据簇的词云图可以分析得到，第 7 簇和第 10 簇是与文旅有相关性的，除此之外的簇与文旅不相关。从 K-Means 聚类后得到的簇的数量可知，在第 7 簇和第 10 簇的数据分别为 457 条和 124 条，求和为 581 条数据与文旅相关。因此可以根据第 7 簇和第 10 簇的数据对整个微信公众号新闻内容进行分类。

“相关性”与“不相关性”数据的占比为：

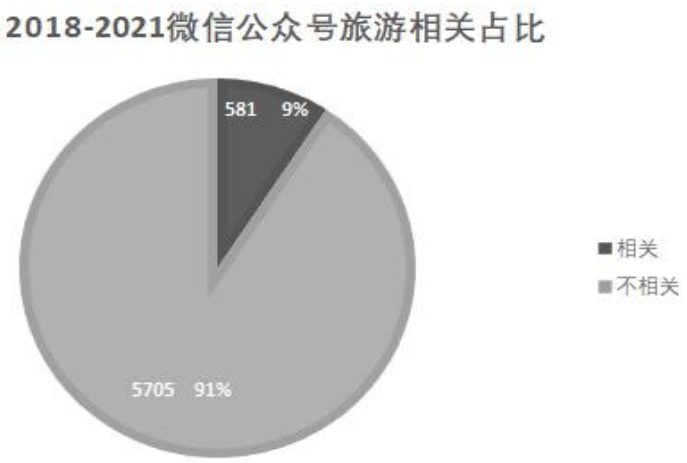

图 8 “相关性”与“不相关性”数据所占比例

把“相关”的数据与“不相关”的数据，提取出来，并保存为 result1.csv。

数据结果的检验

为检验一下“相关”与“不相关”分类的准确性，分别把分好的两类数据，绘画出“相关”与“不相关”的词云图。

图 9“相关”与“不相关”的词云图

由“相关”与“不相关”的词云图中的关键字可以看出数据的分类是准确的。

5.2 问题二的模型的建立与求解

5.2.1 数据预处理

对于问题二中附件的数据，进行一定的处理和检查

1. 是否含有缺失值
2. 是否含有异常值
3. 是否含有重复值
4. 是否需要对文本数据的内容进行，停用词的删减处理

经过对附件数据的运算，附件中酒店评论的数据为 1093 条，景区评论的数据为 1203 条，餐饮评论的数据为 6984 条，正常游记攻略和微信公众号新闻的数据分别含有 294 条和 6286 条数据，但游记攻略中的出发时间、出行天数、人物、人均费用和微信公众号中微信公众号标题的含有缺失值。

```
 0   酒店评论ID  1093 non-null   int64
 1   城市       1093 non-null   object
 2   酒店名称    1093 non-null   object
 3   评论日期    1093 non-null   object
 4   评论内容    1093 non-null   object
 5   入住日期    1093 non-null   object
 6   入住房型    1093 non-null   object
dtypes: int64(1), object(6)
```

```
 0   景区评论ID  1203 non-null   int64
 1   城市       1203 non-null   object
 2   景区名称    1203 non-null   object
 3   评论日期    1203 non-null   object
 4   评论内容    1203 non-null   object
dtypes: int64(1), object(4)
```

```
 0   游记ID     294 non-null    int64
 1   城市       294 non-null    object
 2   游记标题    294 non-null    object
 3   发布时间    294 non-null    object
 4   出发时间    262 non-null    object
 5   出行天数    257 non-null    object
 6   人物       262 non-null    object
 7   人均费用    224 non-null    object
 8   正文       294 non-null    object
```

```
 0   餐饮评论ID  6984 non-null   int64
 1   城市       6984 non-null   object
 2   餐饮名称    6984 non-null   object
 3   评论日期    6984 non-null   object
 4   评论内容    6984 non-null   object
 5   标题       6984 non-null   object
dtypes: int64(1), object(5)
```

```
 0   文章ID     6286 non-null   int64
 1   公众号标题   6285 non-null   object
 2   发布时间    6286 non-null   object
 3   正文       6286 non-null   object
dtypes: int64(1), object(3)
```

图 10 附件数据的相关信息

问题二中所需解决的第一小问是以将提取出的旅游产品名称，旅游 ID 和所依托的语料以表 2 的形式保存为文件“result2-1.csv”，第二小问则是对提取出的旅游产品数据按年度进行热度分析，并排名。两个小问需要解决的问题均与缺失值无关，所以在该题中可以忽略。

5.2.2 问题二模型的建立与求解

第一小问 模型的建立与求解

（1）提取酒店评论、景区评论、餐饮评论的旅游产品

对附件中的数据分别进行提取，把数据中的酒店名称，景区名称，餐饮名称分别以列的形式提取出来，并重新把相应的数据合并，令其新列名称为产品名称。

再用 python 程序对数据的列名称进行处理，把酒店评论 ID、景区评论 ID、餐饮评论 ID 以 1001 的形式转换为酒店评论-1001、景区评论-1001、餐饮评论-1001 的形式。对数据中每以条产品数据的长度进行设置其对应的产品 ID 长度，并设置相应产品 ID。把各个酒店名称、景区名称、餐饮名称统一改成产品名称。

取出酒店评论前 10 行数据进行展示：

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(茂名电白店)
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(信宜绍秀体育馆店)
酒店评论-1008	ID8	茂名海豚酒店
酒店评论-1009	ID9	茂名荔晶大酒店
酒店评论-1010	ID10	茂名海豚酒店

图 11 部分酒店评论数据

对以上的酒店评论数据进行处理的方式分别对景区名称数据，餐饮名称数据进行一样的处理，即可以得到类似的数据表格结果，并对三个数据进行合并处理。

（2）提取游记攻略数据中的产品

游记攻略没有相应的产品名称列，需要构建相应的模型，对数据正文和标题内容中的产品类型的特定词进行提取。首先对数据中的发布时间以年的形式提取出来，再对游记攻略中的旅游产品名称用 excel 进行相应的提取，再对提取出的旅游产品名称进行保存。再以这些提取出来的名称，对游记攻略中每一条文本信息的产品名称分别进行筛选提取，生成相应的产品名称列。并与前面的酒店评论、景区评论、餐饮评论等数据进行合并，再进行保存。

游记攻略提取出的数据（如图 12）：

游记攻略-1290	广东周边游	御水古温泉	宝藏度假地\n 茂名 御水古温泉度假酒店\n广东 茂名 宝藏小众度...	御水古温泉	2021
游记攻略-1291	99%的人不知道御水古温泉还可以这样玩！\n 骨灰级御水古温泉3日2夜攻略\n 御水古温...	御水古温泉	2021
游记攻略-1292	高州行—D1佛山至高州：潘洲公园、观山、高州博物馆、宝光塔、缅茄树、鉴江、濂洲公园、高凉鼓...	博物馆	2021
游记攻略-1293	我们去茂名走马观花（广东⑩）\n走马：路上\n爷爷是一个很重男轻女的传统 广东 人，但在看到...	博物馆	2021
游记攻略-1294	高州行—D4深镇镇&耀新村\n昨晚夜宿仙人洞附近民宿—仙人阁山庄。\n山庄为年轻人回乡创业项...	仙人洞	2021

图 12 部分旅游攻略数据

第二小问模型的建立与求解

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时间

排序，再以情感分析的得分和产品词频两方面指标对产品热度进行计算。

（1）情感分析

情感分析主要可以用来对游记攻略中文本内容数据的分析，通过对文本内容进行相应态度观点的提取、对文本内容的主题分析、含有情感态度的文本词组挖掘出来的一种计算热度的方式。

情感得分可以划分为三类指标中评、好评、差评分别是 0、1、-1，表示对文本内容的好坏程度进行一个好的划分。更加快速的对游记攻略文本内容得分进行处理。其提取出前 5 行的数据（如图 13）进行展示：

语料ID	内容	产品名称	年份	产品ID	情感得分
酒店评论-1001	干净卫生服务好	茂名君悦商务酒店	2018	ID1	1
酒店评论-1002	环境可以，干净！	维也纳国际酒店(茂名电白店)	2018	ID2	0
酒店评论-1003	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...	茂名永利之家	2018	ID3	1
酒店评论-1004	很好.....舒服态度不错	茂名诚荟酒店	2018	ID4	1
酒店评论-1005	#卫生# #设计风格# #酒店餐饮#	茂名华景商务酒店	2018	ID5	1

图 13 部分游记攻略数据的情感得分

基于情感分析得出的情感得分来分析，本题是计算旅游产品的热度，由题中表格中情感得分的形式来表示在情感分析下的一个热度值。

（2）统计旅游产品的词频

在统计出现的某一旅游产品的数量进行提取和分析时，是根据现有的旅游产品数据的数量进行提取，但不同相应的 ID 会含有相同的旅游产品，所以需要对相应的旅游产品进行去重处理，在对去重后的数据进行旅游产品的数量计算。

定义 $D(i)$ 为该旅游产品的词频，i 为某一旅游产品的名称计算词频为：

$$D(i) = num1(i)$$

其中 $num1(i)$ 代表某一旅游产品中出现的数量。

为了避免对旅游产品中的情感得分和旅游频次的差距较大，导致数据的误差大。所以需要对两者数据进行求和，所得出的结果代表其旅游产品的热度。

定义对应的旅游产品热度为 $R(i)$,即为：

$$R(i) = D(i) + x_{normalization}$$

对产品热度的计算，是基于情感分析的得分与旅游产品频次的数量两部分之和来确定其旅游产品的热度，根据问题二中需要保存的形式，此时需要对得出的数据结果值进行 0-1 标准化，把产品旅游热度的数据进行标准化。

（3）0-1 标准化

计算出产品热度后需要对该数据值进行 0-1 归一化处理，使旅游产品的数据更加标准化，更加准确。

该 0-1 归一化的公式为：

$$x_{normalization} = \frac{x - Min}{Max - Min}$$

其中 $x_{normalization}$ 代表为旅游产品热度值的归一化数据，x 为样本的数据，Min 和 Max 分别代表该文本内容旅游产品热度值的最大值和最小值。

运算出的结果为旅游产品的热度值（如图 14），对其结果进行保存，并展示相应的数据结果：

产品ID	产品类型	产品名称	产品热度	年份
ID1	酒店	茂名君悦商务酒店	0.013986	2018
ID2	酒店	维也纳国际酒店(茂名电白店)	0.006993	2018
ID3	酒店	茂名永利之家	0.020979	2018
ID4	酒店	茂名诚荟酒店	0.020979	2018
ID5	酒店	茂名华景商务酒店	0.013986	2018
...	...	...	...	...
ID7861	特色餐饮	三两粉（茂名东汇城店）	0.351682	2021
ID7888	特色餐饮	塔斯汀中国汉堡（民主店）	0.082569	2021
ID8062	特色餐饮	泰哥茶档	0.211009	2021
ID8574	特色餐饮	周黑鸭（东汇城店）	0.039755	2021
ID8718	特色餐饮	小和寿司（光明路店）	0.058104	2021

图 14 部分旅游产品的热度数据

5.3 问题三的模型的建立与求解

5.3.1 建模准备

该问题需要解决的是相关旅游产品的关联度

针对问题二得到的旅游产品数据，对提取出来的旅游产品进行关联分析，并计算其相关的关联值。

由于计算关联度是在某一个 ID 下的多种旅游产品之间的关联度分析，所以该题就是在附件中游记攻略部分数据进行相应的分析。在计算结果之前需要对附件中游记攻略数据中的旅游产品进行提取。用 excel 对每条旅游攻略数据的正文数据部分出现的旅游产品提取出来再进行相应的计算。

one-hot 编码

后面需要对数据的支持度、置信度、提升度进行相应的计算，需要对旅游产品的特殊变量进行数据数字化，即将数据的形式转化为相应的特殊 one-hot 编码，one-hot 编码不存在变量顺序关系的一种类别变量，就可以利用机器学习，直接进行数据间的运算。对提高变量间数据的计算结果的准确性。

已知基于 Apriori 算法的关联规则：支持度、置信度、提升度[4]，可以从这三个方面计算出相应的关联度值。

Apriori 算法:是一种对数据进行挖掘，基于一定规则数据频率词集的算法。本文则是利用 Apriori 算法对基于游记功率的数据频集的内容进行计算[4]。

5.3.2 问题三模型的建立与求解

支持度

支持度表示项集{X,Y}在总词组集里出现的概率，其相应的公式为：

$$Support(X \to Y) = \frac{P(X,Y)}{P(I)} = \frac{P(X \cup Y)}{P(I)} = \frac{num(X \cup Y)}{num(I)}$$

其中 I 表示，文本数据内容分词后的总词组集，num( )表示所求总词组集里特定词组出现的次数。

$num(I)$ 表示总词组集的个数，$num(X \cup Y)$ 表示含有 $\{X,Y\}$ 的词组集的个数。

置信度

在先决条件 X 发生的情况下，再由关联规则“ $X \rightarrow Y$ ”推出 Y 的概率。

其公式为：

$$Confidence(X \rightarrow Y) = P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{P(X \cup Y)}{P(X)}$$

提升度

提升度则是表示在含有 X 的条件下，同时含有 Y 的概率，与 Y 总体发生的概率之比。

其公式为：

$$Lift(X \rightarrow Y) = \frac{P(Y \mid X)}{P(Y)}$$

根据所运行出的结果，分别以支持度、置信度、提升度这三个方面计算出各个产品的关联度值。

部分支持度、置信度、提升度结果数据的展示:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(一滩海鲜档)	(中国第一滩)	0.020408	0.224490	0.020408	1.000000	4.454545	0.015827	inf
1	(中国第一滩)	(一滩海鲜档)	0.224490	0.020408	0.020408	0.090909	4.454545	0.015827	1.077551
2	(一滩海鲜档)	(放鸡岛)	0.020408	0.224490	0.020408	1.000000	4.454545	0.015827	inf
3	(放鸡岛)	(一滩海鲜档)	0.224490	0.020408	0.020408	0.090909	4.454545	0.015827	1.077551
4	(一滩海鲜档)	(水东湾)	0.020408	0.061224	0.020408	1.000000	16.333333	0.019159	inf

图 15 部分支持度、置信度、提升度结果

	support	itemsets
0	0.020408	(0668)
1	0.020408	(一滩海鲜档)
2	0.020408	(东平镇)
3	0.020408	(严家祠)
4	0.224490	(中国第一滩)
...	...	...
466	0.020408	(顺德, 阳江)
467	0.020408	(露天矿博物馆, 雨顺阁)
468	0.020408	(露天矿生态公园, 高州)
469	0.020408	(露天矿生态公园, 高州博物馆)
470	0.020408	(高州, 高州博物馆)

图 16 相应的旅游变量特征的支持度

旅游图谱

用 python 的机器学习对两个附件的分别画出其中的旅游图谱，再对总的数据绘画出旅游图谱分析（图 17、图 18、图 19）。

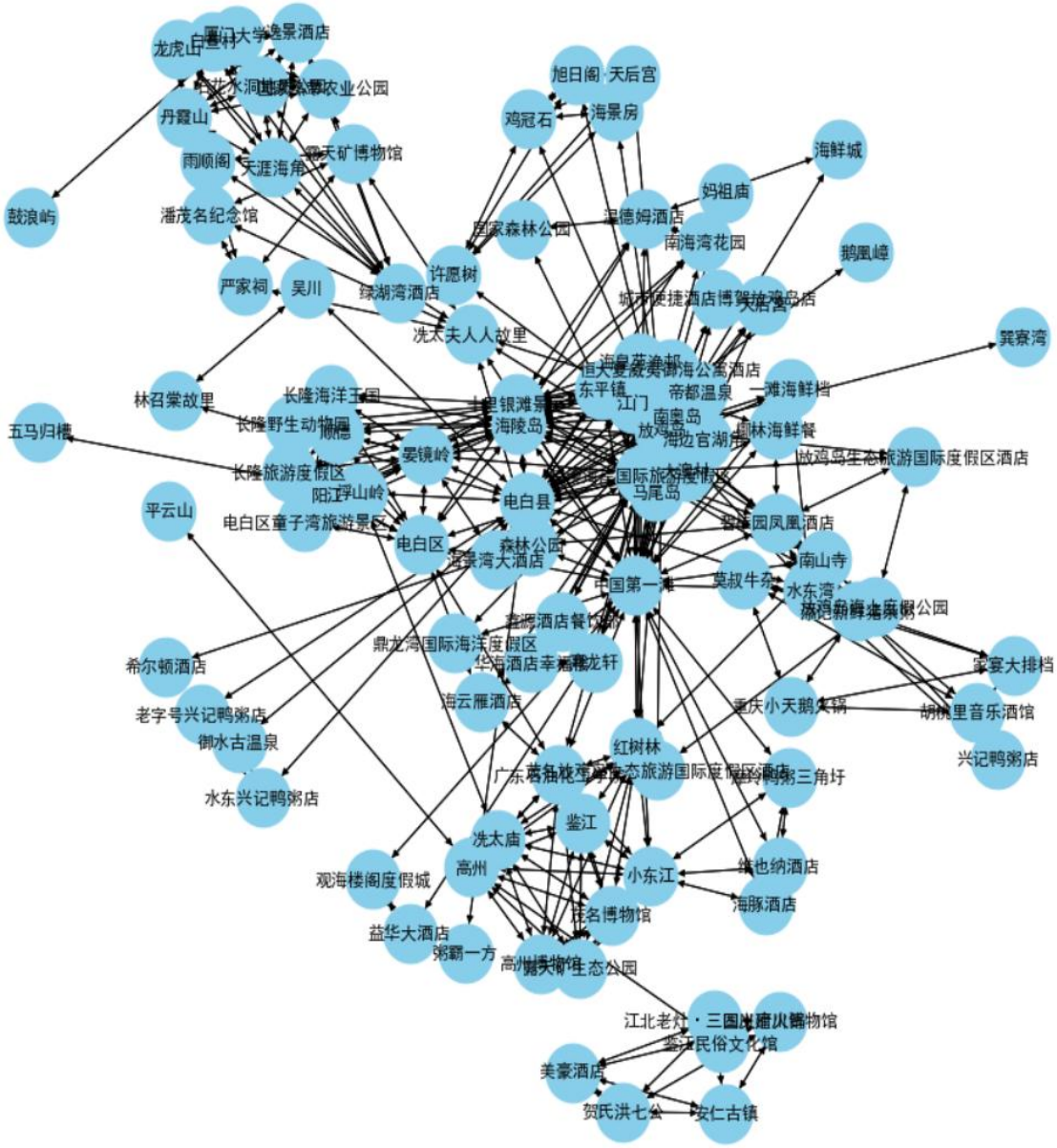

图 17 2018-2019 的旅游图谱

从 2018 到 2019 的旅游图谱的数据，发现图旅游景区、酒店、餐饮的关联度很大、旅游的人流量也均匀分布。

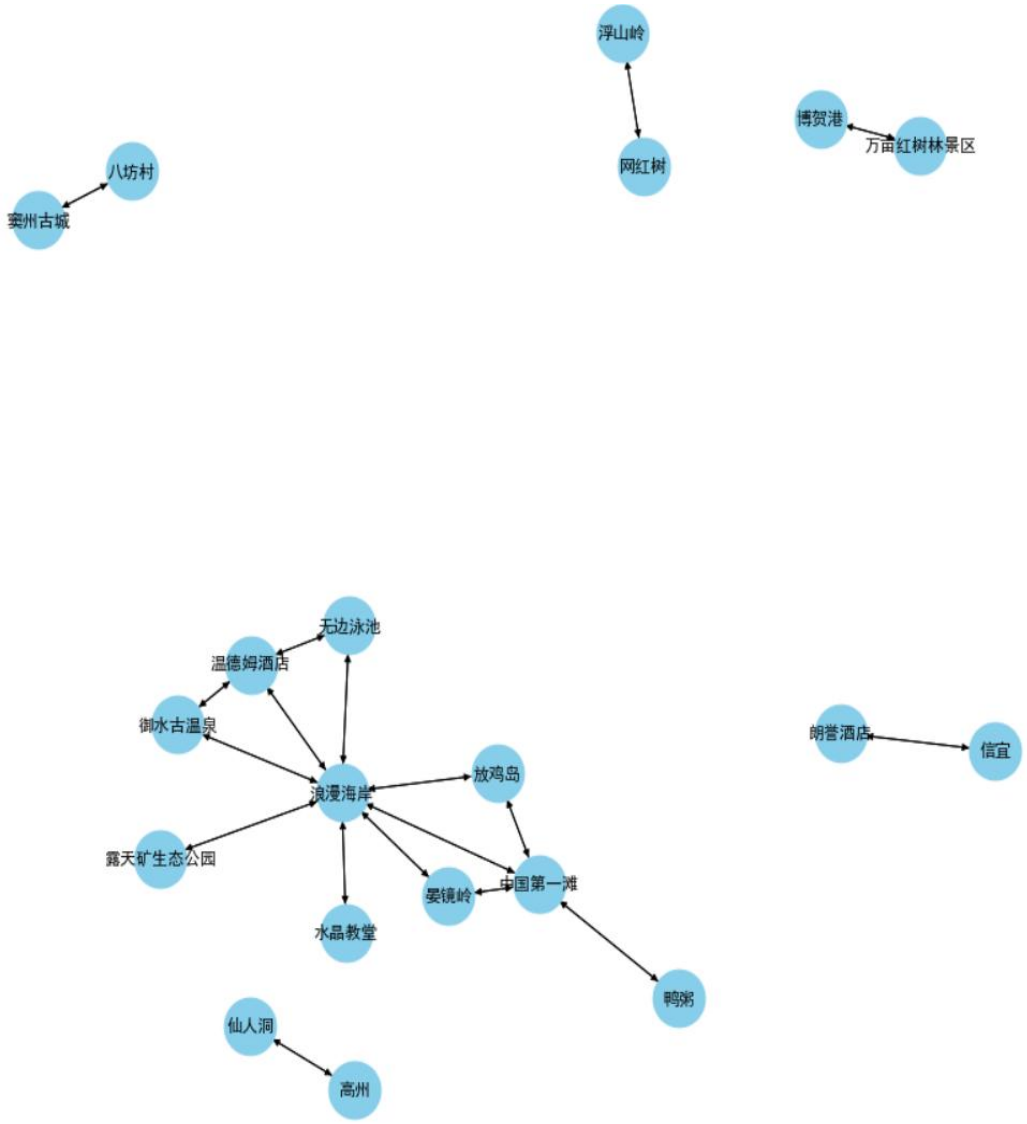

图 18 2020-2021 的旅游图谱

从 2020 到 2021 的旅游图谱的数据，发现图旅游景区、酒店、餐饮的关联度相对较小、旅游的人流量也不是很均匀的分布。中国第一滩、温德姆酒店、浪漫海岸、御水古温泉、露天矿生态公园、宴镜岭、放鸡岛等地方的关联度很强。说明旅游的游客对这几个地方旅游的频率高，相对于该景点周围的餐饮等方面的东西需要提高。

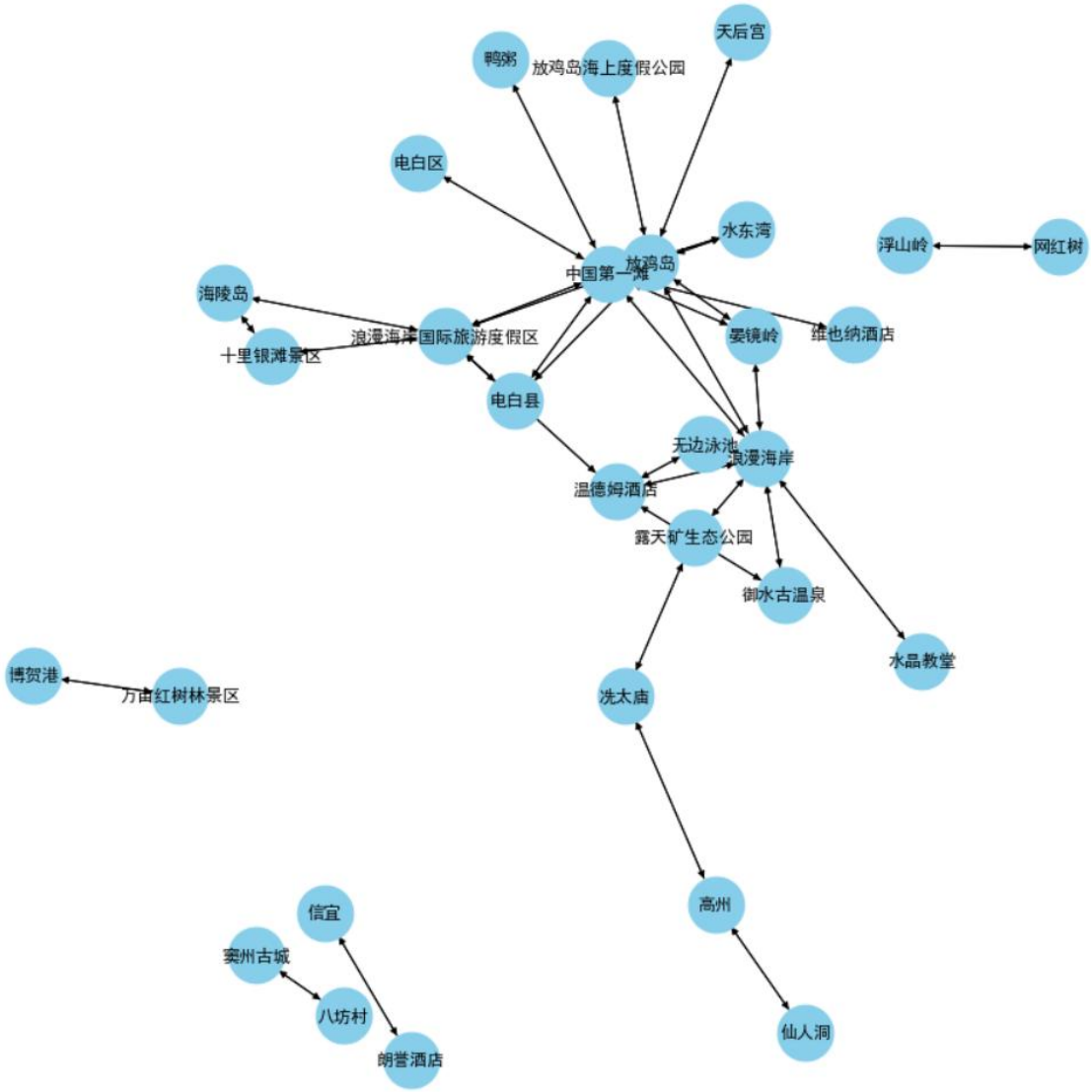

图 19 2018-2021 的旅游图谱

从 2018 到 2021 的旅游图谱的数据，发现图旅游景区、酒店、餐饮的关联度相对较大、旅游的人流量也很均匀的分布，则看该图还是不能对疫情前后的数据进行分析。

根据中展示的数据可分析一滩海鲜档与放鸡岛的关联度是很高的，一滩海鲜档与中国第一海滩的关联度也是较高的，根据部分的关联度值的数据，再对附件中文本内容的分析，体现出该数据也是具有一定的准确性的。

5.4 问题四根据旅游图谱的建议信

5.4.1 数据的分析

根据疫情前后旅游产品的酒店，景区和餐厅的数量分析

分别以附件 2018 到 2019 的数据，作为疫情前茂市的旅游产品等的数据，以附件 2020 到 2021 的数据作为疫情后的数据。以对两个数据的分析，对其做出该对应的旅游图谱或者比较两类数据中酒店评论、景区评论、游记攻略、餐饮评论等数据中的旅游产品的数量变化进行分析。用 excel 对酒店评论、景区评论、游记攻略、餐饮评论等数据中的旅游产品的数量进行计算。2018~2019 年，酒店评论、景区评论、餐饮评论数据中的旅游产品分别有 293 家酒店、539 个景区、1177 家店。2020~2021 年，酒店评论、景区评论、餐饮评论数据中的旅游产品分别有 699 家酒店、664 个景区、5807 家店。

根据疫情前后的旅游图谱分析

从 2018-2019 的旅游图谱中可以看出，相关的酒店，景区和餐厅的关联度高。而 2020 到 2021 的旅游图谱的数据，发现图旅游景区、酒店、餐饮的关联度相对较小，对酒店，景区和餐厅的选择相同性大，主要集中在中国第一滩、温德姆酒店、浪漫海岸、御水古温泉、露天矿生态公园、宴镜岭、放鸡岛等地方的频率高，而在窦州古城、八坊村、万亩红树林景区等地区的频率较低。且疫情后游客选择景区的个数减少。

敬爱的茂名市旅游主管：

您好

很冒昧的给你写这封信，希望您可以花几分钟时间阅读一下。基于疫情前两年后两年的茂名市一些旅游数据的分析，想向您提供一些我自己的建议。

基于疫情前后中的酒店数量的变化，2018-2019 年的 294 家的酒店到 2020-2021 年的 699 家酒店，疫情后酒店的数量增加了一倍多，随着人们选择在本地旅游时选择的增加，酒店的数量也在增加，由于酒店使用的增加迅速，需要增加对酒店防疫安全措施。增加相应的消毒工具，增设旅游者进出酒店的防疫关口，对酒店隔离场所的增建，避免因为疫情的感染而缺少集中隔离的场所，造成人们的恐慌和疫情的扩散。

疫情前后的景区数量的变化，2018-2019 年的 539 个的景区到 2020-2021 年的 664 个景区，数量增加了接近 100 左右的景区，旅游景区的增加，人流量的增加，需要对旅游景区老旧的设施进行重新维护，避免安全事故的发生。对应急药品、对人员处理应急事件的培训进行加强、对防疫意识的加强，需要对人员口罩的佩戴加强提醒。

疫情前后的餐厅数量的变化，2018-2019 年的 1177 家的餐厅到 2020-2021 年的 5870 家餐厅，餐厅的数量增加了三倍的数量，餐厅数量的增加，也反应了人们对餐厅的需求量大，人口的聚集度大，不利于疫情防控。需要对不满足疫情防控要求的餐厅进行彻查，减少疫情期间病毒的传播。对餐厅座椅的间隔距离和餐区的消毒进行严格的规定。

根据 2018-2019 的旅游图谱和 2020-2021 的旅游图谱的关联度分析，2018-2019 的旅游图谱，展示出了旅游各个方面的关联度，发现不管是餐饮、酒店还是景区的关联度还是很大，人流量分布很均匀。再观察 2020-2021 的旅游图谱，可以发现餐饮、酒店、景区的关联度变小，人们旅游的景区也变成相对应的几个景区。由于疫情的影响，茂名市当地的旅游人数也随之增加，相对聚集的地方也单一，如果突发疫情的化，会产生大量的传播。所以需要在相对集中的旅游景点增设可以满足，游客的酒店和餐馆，使游客分散自己的选择，从而达到有效的避免大量的聚集人流。

疫情后相应的景点在中国第一滩、放鸡岛、宴镜岭、御水古温泉、温德姆酒

店、露天矿生态公园、浪漫海岸等游客出行频率较高的旅游景点，进行相应设施的维修，应急隔离的站点也需要增设。坚持对防疫的常态的一个时效性的检查，加强对疫情风险的一个提前的判断，有效的控制旅游景区的人流量，防止人流量的大量集中，加强防疫的宣传和对应防疫操作的引导，加强对游客，游玩之前的健康码和体温的严格检查。培养该景点的员工，对疫情防控的措施进行实际的培训。

XXX

2022 年 4 月 29 号

六、参考文献

[1] 何铠,管有庆,龚锐.一种基于权重预处理的中文文本分类算法[J].计算机技术与发展,2022,32(03):40-45+53

[2]薛薇.Python 机器学习：数据建模与分析[M].2021

[3] https://blog.csdn.net/u011250186/article/details/107407500

[4] https://blog.csdn.net/qq_40851534/article/details/106448884