基于数据挖掘技术的旅游目的地游客印象分析

摘要

随着网络信息技术的发展，各种社交软件层出不穷，旅游者也越来越习惯于在旅游前上网搜索目的地的评价信息，根据评价对旅游计划作出修改。在旅游目的地的美誉度逐渐变得透明化的今天，各地文旅主管部门和相关企业除了切实做好旅游景点的管理之外，了解游客的评价及旅游体验，切实提高游客的满意度和目的地的美誉度至关重要。对评论文本进行分析是了解游客真实需求的一个常用方法。本文将基于数据挖掘技术对附件中给出的评论文本进行分析，力求从中获得有助于目的地旅游体验提升的信息。

针对问题 1：小组对原始数据进行预处理后，运用 Python 中的 jieba 库对文本进行分词处理，分别统计出各个词语的历史词频和当前词频；接着引入牛顿冷却定律，将热门词的出现频率类比为物体的冷却过程，从而计算出每个词语的热度，选取出每个目的地的热门词 TOP20。

针对问题 2：为方便后续分析，小组在预处理的基础上提取出含有名词的评论。接着建立 LDA 主题模型，对评论中常常相互关联的词语进行归类，将每条评论归为服务、位置、设施、卫生、性价比中的一个方面。评论归类后，可得出每个景区、酒店的每个方面下评论的数目。再对评论进行情感分析，运用李克特 5 分量表将每条评论的情感倾向从“非常不满意”到“非常满意”进行判断，分别赋分为 1-5 分。从而可以得出每个景区、酒店在五个方面中每一方面的具体得分。对于综合得分，小组运用模糊综合评价法获得五个方面的权重，得出每个景区、酒店的综合评价得分。最后用均方误差（MSE）法将计算出的综合得分与附件 2 中的专家评分进行比对，评价模型的准确性。

针对问题 3：小组首先对影响评论有效性的情况进行归类，给出“无效评论”的定义。接着对不同类型的无效评论进行识别与统计。针对简单重复类评论，小组运用余弦相似度对评论的相似性进行量化，识别出相似度过高的评论；针对内容无关或无有效内容的情况，小组运用问题 2 中建立的 LDA 主题模型，将在每个主题中对主题的贡献度较低的评论视为无效评论；针对文本过短的评论，统计文本长度低于 5 个字符的评论。最终统计出无效评论的数目，对评论有效性进行分析。

针对问题 4： 小组引入 TF-IDF 算法，在对评论文本进行分词处理后，统计出高频词，获得各目的地的高频词集；将每一目的地的所有评论合成为一个文档，即为 TF-IDF 算法中的一份“文件”，所有“文件”统称为一个“语料库”；对高频词集内每一词语进行分析，若该词语在它所在的“文件”中出现次数较多，即在该目的地的评论中被多次提及，而在其余“文件”中出现次数很少，那么我们就可以称这个词语能够代表该目的地的一个特征。再根据问题 2 中得出的综合评价得分将目的地进行分级，每个级别选择 3 个景点和酒店进行具体分析。

关键词：文本分析；牛顿冷却定律；LDA 主题模型；模糊综合评价；TF-IDF 算法

Abstract

With the technological development, various social softwares emerge in an endless stream, and tourists are becoming more and more accustomed to searching the Internet for comment information of destinations before traveling. Today, when the reputation of tourist destinations is gradually becoming transparent, cultural and tourism authorities and related enterprises should not only do a good job in the management of tourist attractions, but also understand the feelings of tourists, so as to effectively improve the satisfaction of tourists. This article will analyze the comments given in the attachment based on data mining technology, and strive to obtain information that will help improve the travel experience.

As for question 1: After the team preprocessed the original data, the jieba library in Python was used to segment the text, and the historical and current frequency of each word were counted; **Newton's law of cooling** was then introduced to calculate the heat of each word, and the TOP20 popular words of each destination are selected.

As for question 2: The team extracted comments containing nouns on the basis of preprocessing. Then establish the **LDA topic model**, classify the words that are often related to each other in the comments, and classify each comment as one of the aspects of service, location, facilities, hygiene, and cost-effectiveness. After the comments are categorized, the number of comments in each area of each scenic spot and hotel can be obtained. Then analyze the sentiment of the comments, and use the Likert 5-point scale to judge the sentiment tendency of each comment from "very dissatisfied" to "very satisfied", and assign scores of 1-5 points respectively. For the comprehensive score, the group uses **the fuzzy comprehensive evaluation method** to obtain five weights, and obtains the comprehensive evaluation score of each scenic spot and hotel. Finally, the mean square error (MSE) method is used to compare the calculated comprehensive score with the expert score in Appendix 2 to evaluate the accuracy of the model.

As for question 3: The team first classifies the conditions that affect the validity of the comments and gives the definition of "invalid comment". Then identify and count different types of invalid comments. For simple repetitive comments, the group uses cosine similarity to quantify the similarity of comments and identify comments that are too similar; for situations where the content is irrelevant or has no valid content, the group uses the **LDA topic model** to In each topic, comments with a low degree of contribution to the topic are regarded as invalid comments; for comments with too short text, comments with a text length of less than 5 characters are counted.

As for question 4: The team introduced the **TF-IDF algorithm**. After word segmentation of the comments, the high-frequency words were counted to obtain the high-frequency word set of each destination; all comments of each destination were combined into one document, namely as a "document" in the TF-IDF algorithm, all "documents" are collectively referred to as a "corpus"; each word in the high-frequency word set is analyzed, if the word has been mentioned many times in the comments of the destination, but rarely appears in the remaining "documents", then we can say that this word can represent a feature of the destination.

Keywords: text analysis; Newton’s law of cooling; LDA topic model; the fuzzy comprehensive evaluation method; TF-IDF algorithm

目录

一、 挖掘目标

近年来，随着网络的发展，旅游目的地美誉度逐渐变得透明化，消费者可以通过各种软件得知游客对该目的地的满意度及相关评论。因此，对于各地文旅主管部门和里边有相关企业而言，掌握游客满意度的影响因素，从而提高游客满意度和目的地美誉度，对该地旅游业的长远发展至关重要。

问题 1 要求小组进行景区及酒店的印象分析。根据附件 1 中的景区、酒店的评论文本，以景区（酒店）分类计算出目的地的 TOP20 热门词以及其相关热度。

问题 2 要求小组对景区及酒店进行综合评价。根据附件 1 中的评论文本，建立数学模型，从服务、位置、设施、卫生、性价比五个方面对相应目的地进行评分。为检验模型的准确性，再利用附件 2 中的专家打分结果运用均方误差（MSE）进行评价。

问题 3 要求小组进行文本的有效性分析。针对文本中出现的内容不相关、简单复制修改、内容无效等影响文本有效性的因素，建立数学模型，评价附件 1 中的文本的有效性。

问题 4 要求小组进行目的地的特色分析。小组需要建立数学模型，从附件 1 评论文本中发掘每个目的地的特色和亮点，并选择部分样本，结合建模结果进行具体分析。

二、 问题分析

问题 1 的分析：首先进行数据的预处理，将评论中内容重复的条数删去；继而选择精确分词法对文本进行分词处理；对于热度的计算，考虑到时间较为久远的评论对现今的旅游体验参考价值较低，小组使用牛顿冷却法，区分出当前评论以及历史评论，综合考虑评论日期及词语出现次数确定热度。

问题 2 的分析：题目要求从服务、位置、设施、卫生、性价比五个方面进行评分，因此引入 LDA 主题模型，建立包含上述五个维度的综合评价指标体系，并确定每个主题所包含的关键词。接着运用李克特 5 分量表法对每一条评论概率值最大的主题进行情感分析，从而对该主题进行评分，分值为 1-5 分，其中 1 分表示“非常不满意”，5 分表示“非常满意”，从而得出各个景区、酒店五个方面的得分。最后是计算景区、酒店的综合得分，运用模糊综合评价法建立相应的评语集、因素集，从而得到评判对象的综合评价集，结合上述各个方面的得分计算出各主题的对应权重，从而得出综合评价总分。

问题 3 的分析：题目要求对附件 1 中的评论进行文本有效性分析。首先小组需要综合题目所列的内容不相关、简单复制修改、内容无效等情况给出具体的“无效评论”的定义。其次，再针对不同的情况建立数学模型，以判定无效评论，统计其数目。针对“简单复制修改”这一情况，小组选择利用余弦相似度法查找出相似度较高的评论文本条数；针对“内容不相关”以及“内容无效”的情况，小组引入 LDA 主题模型识别出对应的无效评论；对于“过短评论”，小组统计评论文本长度后作出分析，给出过短评论的长度限制并统计过短评论数。由此得出无效评论的总条数以及无效评论在评论条数中的占比，并据此分析评论文本的有效性。

问题 4 的分析：小组引入 TF-IDF 算法提取目的地的特征。首先，小组统计高频词获得每个目的地的高频词集；然后，同时题目要求分别选取高、中、低三个层次的 3 家景点和酒店进行具体分析，因此需要从问题 2 中的综合评价结果中对景区、酒店进行分层，确定其属于高、中、低层次的哪一个层次。

三、 数据分析过程

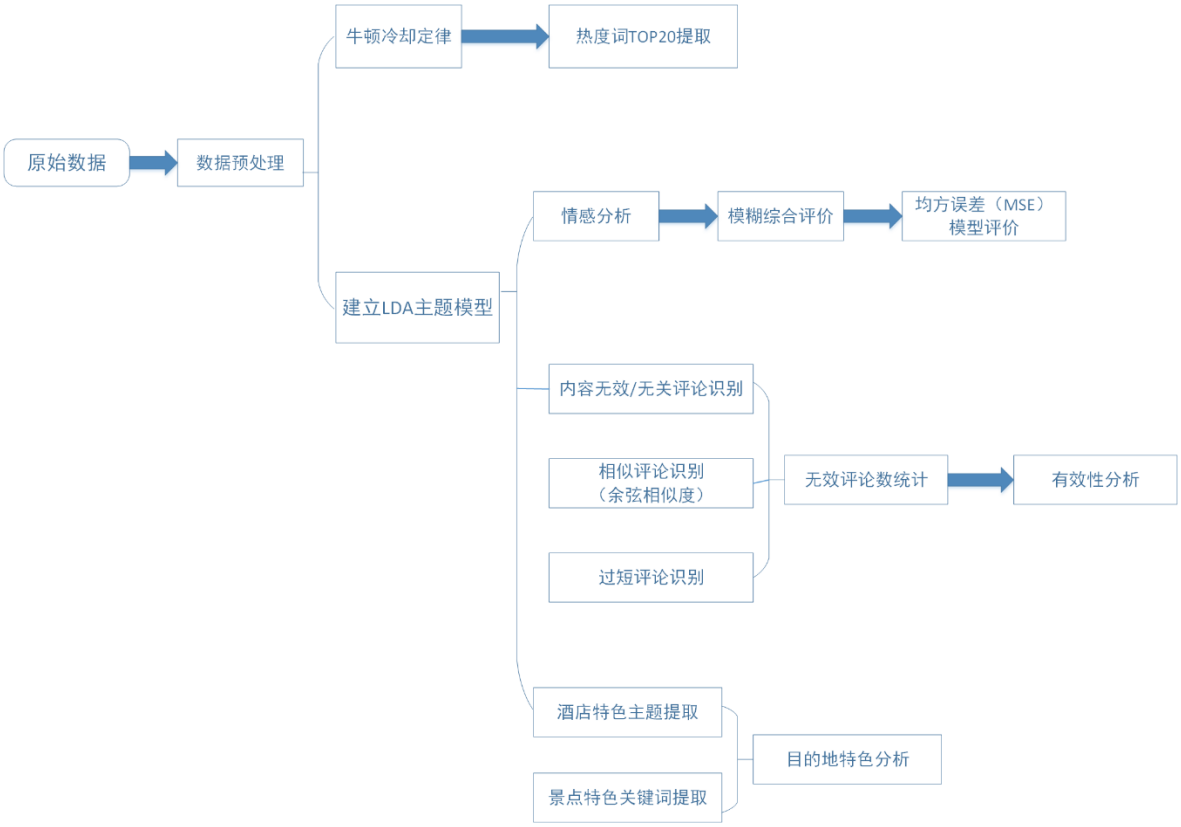

图 1 整体分析流程图

3.1. 问题 1 的分析过程

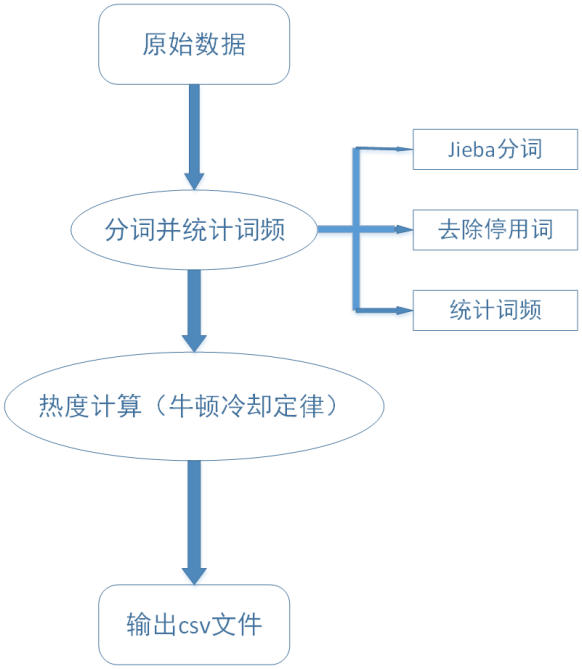

图 2 问题 1 的数据分析流程图

3.1.1. 数据预处理

由于同一个人可能去同一景区旅游多次，或者有游客参考他人的评论进行简单复制，

此类情况中重复多次的评论显然不具有太高的分析价值。考虑到评论中可能出现的文本的简单重复这一现象降低文本分析的有效性的问题，小组首先对附件 1 中的评论进行文本去重，删除内容完全一致的评论。

3.1.2. 文本评论热度词提取

要准确分析评论文本的具体内容、感情色彩，必须对文本中的句子进行分割成一个个词语，进而对词语的出现频率进一步分析。然而，文本中原有的“边界”只有段落和标点符号，词语和词语之间的“边界”是非常模糊的，这无疑对文本的整体文义理解造成一定的困难。

	景区名称	评论日期	评论内容
0	A01	2020/6/16	是亲子游的绝佳场所，门票就是有点贵，不过可以接受，爷爷奶奶不放心小朋友也跟上来了，当天我们十...
1	A01	2020/1/23	**景区差不多，票价偏贵了。大马戏比较精彩，八点的场次，6点40才能检票进入，我们6点多看看...
2	A01	2020/3/22	很有**特色的亲子酒店，房间里的装修很可爱，小孩子特别喜欢，洗漱用品也很有特色，对应的房间还...
3	A01	2020/12/25	有园区的工作人员在那，他会主动给你园区里的地图和表演的时间安排，很周到，上接驳车大概也是34...
4	A01	2020/11/28	周五逃课跟朋友在广州集合！终于如愿以偿的到达欢乐世界。学生票198 需要出示相关证件（校卡或...
...	...	...	...
58663	A50	2015/2/25	还好吧。我们刚刚到瀑布楼遇到一点小意外，打电话到景区办公室要求帮助，景区值班领导马上行动，在...
58664	A50	2015/2/25	山高路远，走的很辛苦。景色宜人爬山很累。
58665	A50	2015/2/22	环境很好，空气非常棒，很适合全家旅游，特别是避暑
58666	A50	2015/2/16	都很方便，价格实惠吧，可以预早就订好。
58667	A50	2015/2/22	旅行社不负责任 到了景点没有与门票售票协调好 等了很久

58668 rows × 3 columns

图 3 部分文本示例

另外，热度的计算不能够简单理解为“词语在评论中出现的次数”。不可否认，评论的日期对于评论的数据分析具有一定意义。由于各旅游目的地在不断进行翻新、提升，以往评论中提及的数据、缺陷可能与现在有所区别，因此，评论日期距离现今日期越远，该评论对现在的旅游者的参考价值越低。在热度计算中，需要综合考虑词语词频以及随时间变化该词语出现频率的变化趋势，计算出具有现实意义的热度词。

以下是提取热度词的具体步骤：

（1） 步骤一：分词

导入 jieba 库，运用精确分词模式对评论文本进行划分。jieba 分词工具是目前最常用的中文分词工具，其中精确分词模式试图将语句进行最精准的切分，且不存在冗余数据，因此最常用于文本分析以及评论数据分析。

以下是部分文本的分词结果：

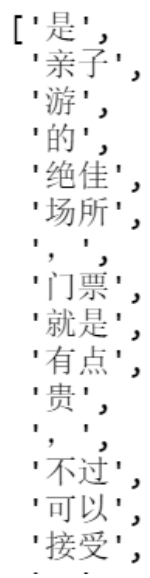

图 4 部分文本分词结果示例

（2） 步骤二：去除停用词

经过分词这一步骤，原文本已经变成多个词语的集合。然而，文本中仍然含有很多无意义的词语，如中文的“的、地、得”等，出现频率非常高，但是无实际意义，对文本的理解没有帮助，因此也需要删除。

小组采用与停用词表相匹配的方式，若该词与停用词表中的词语相匹配，则将该词从现有的词语集合中删除。

通过查找相关资料，小组最终确定的停用词种类包括：标点符号、特殊符号、序号、数字、代词、语气词、关系词等。因为题目是对景区的酒店的分析，评论中包含很多“酒店”,“景区”等对分析帮助不大但出现频率较高的词语，所以我们还在停用词表中添加了这些词：酒店、大堂、前台、地方、景区、景点、门票、票、买票、取票、订票、冯家铭、冯、小孩、孩子、小孩子、小朋友、一家人、老人、时间、小时、游、游玩、旅游、玩、景。以下是部分停用词：

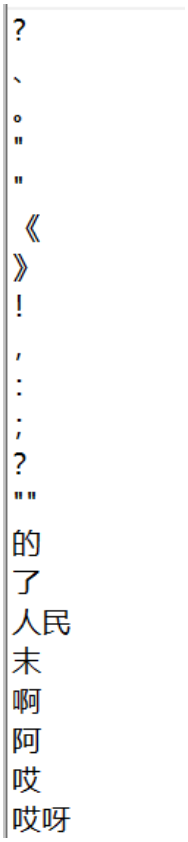

图 5 停用词示例

以下是去除停用词后的分词结果：

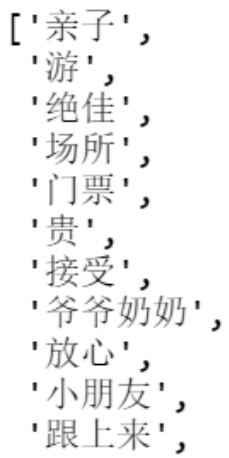

图 6 部分文本去除停用词示例

从分词结果中可见，文本的整体长度被缩短了，相比于原来带有标点符号、“的、地、得”等词语的文本而言可读性降低，但留下来的词语都带有实际意义，是评论文本分析所需要的。

（3） 步骤三：统计每个词语的当前词频和历史词频

考虑到评论时间对评论价值及热度的影响，小组需要将评论分为“当前评论”及“历史评论”两类。

① 对于景区：通过观察数据，发现大部分景区都有 2015-2021 年的数据，但有部分景区存在近期年份数据的缺失。例如 A48，这个景区没有 2020 年及 2021

年的数据。如果直接将 2020 年、2021 年的数据作为当前数据，之前的数据作为历史数据，容易给后续的统计造成误差。因此，小组决定将每个景区数据的最近两年的评论视为“当前评论”，这些评论中各词语出现的频率称为“当前词频”，这些评论对现今的旅游者参考价值较高；最近两年以前的评论视为“历史评论”，这些评论中各词语出现的频率称为“历史词频”，这些评论可能有部分内容过时了。

② 对于酒店：观察数据发现酒店的评论数据均为 2020 年数据，用“年”作为区分标准跨度太大。因此小组决定将每个酒店的 2020 年下半年评论视为“当前评论”，这些评论中词语出现的频率称为“当前词频”，旅游者能从中获得更有价值的信息；2020 年上半年的评论视为“历史评论”，这些评论中词语出现的频率为“历史词频”，旅游者从中获取的信息有可能过时。

以下是部分词频统计结果：

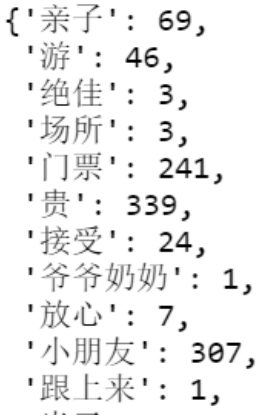

图 7 部分历史词频统计结果

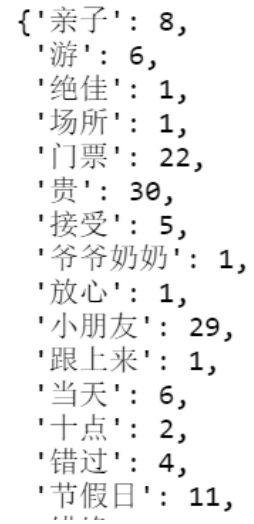

图 8 部分当前词频统计结果

（4） 步骤四：热度计算

牛顿冷却定律揭示了物体的温度随时间变化的规律，表明物体的温度随时间的变化率与物体当前所处自然环境的温度之间的温度差是成比例的。牛顿冷却定律的数学公式为：

$$\frac{\mathrm{dQ(t)}}{\mathrm{dt}} = Q'(t) = -\beta(Q(t) - E) \qquad (1)$$

其中，Q(t)表示温度随时间t变化的函数，即物体在时间t时的当前温度，$Q'(t)$是温度函数的导数，即物体温度变化的速率，E是环境温度。常数β($\beta > 0$)表示物体和环境的温度差与物体自身温度变化速率之间的比例关系，负号表示该物体处于降温过程中。

我们可以把评论热门词的变化过程理解为物体的冷却过程。因此，为求出词语的热度，我们只需要求出上述公式中β的值，即每个词语的负冷却系数即可。

对(1)式两边积分，得

$$\int \frac{Q'(t)}{\mathrm{Q(t)} - \mathrm{E}} dt = \int (-\beta) dt \tag{2}$$

进而求解：

$$\ln(Q(t) - E) = -\beta t + c \tag{3}$$

$$Q(t) = E + Ce^{-\beta t} \tag{4}$$

其中C为常数项，代入点(t_0, Q_0)，得到C的表达式为：

$$C = （Q_0 - E）e^{\beta t_0} \tag{5}$$

将(5)式代入(4)中得到$Q(t)$的表达式为：

$$Q(t) = E + （Q_0 - E）e^{-\beta（t - t_0）} \tag{6}$$

由于物体温度最终会等于环境温度，为方便起见，可将环境温度定义为0，由此得到$Q(t)$的表达式：

$$Q(t) = Q_0 e^{-\beta（t - t_0）} \tag{7}$$

并由此得到负冷却系数β的表达式：

$$\beta = -\ln\left(\frac{Q(t)}{\mathrm{Q_0}}\right)/(t - t_0) \tag{8}$$

将该公式应用于热度词的发现过程，则$Q(t)$表示当前词频，$\mathrm{Q_0}$表示历史词频，$t - t_0$表示当前时间与历史时间之间的时间差。冷却系数越低，则表明该词热度越高。由于在本次分析中时间差均为6年，对热门词的热度排名不产生任何影响，因此冷却系数的定义简化为：

$$\beta = -\ln\left(\frac{Q(t)}{\mathrm{Q_0}}\right) \tag{9}$$

部分词语及其冷却系数如下：

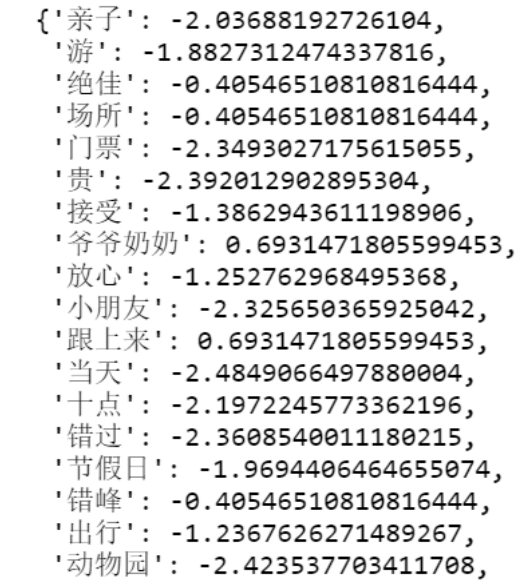

图 9 部分词语及其冷却系数示例

（5） 步骤五：选取出热度最高的 20 个热度词

为使题目所需的热门词的热度这一数值更符合我们平常所理解的“词语越热门，热度值越高”这一现象，小组对冷却系数作如下处理：由于热度最小的词冷却系数最大，因此选取出来的热门词冷却系数一定低于这一词的冷却系数。对此，我们将每个词的热度定义为：热度最小的词的冷却系数与该词的冷却系数之差的 1000 倍。可如下表示：

$$\alpha_i = [\max(\beta) - \beta_i] \times 1000 \qquad (10)$$

计算出每个景区及酒店的词频及相应热度，选取出最热门的 TOP20 词语，每个景区、酒店生成一个 csv 文件：

评论热词	热度
十环	4564.348
区	4488.636
真心	4382.027
垂直	4320.816
儿童	4262.68
长颈鹿	4234.107
型	4219.508
手机	4174.387
挺好玩	4174.387
滑板	4143.135
超	4060.443
大象	4051.785
摆锤	4025.352
好看	3988.984
窗口	3976.562
累	3951.244
U	3912.023
公园	3857.215
本来	3850.148
可惜	3843.03

图 10 景区 A01 热门词 TOP20 及相应热度

评论热词	热度
做	3314.186
大型商场	3091.042
利用	3091.042
东西	3091.042
坐	3091.042
便捷	3091.042
火车	2957.511
公司	2908.721
吃饭	2908.721
陶陶	2908.721
居	2908.721
合适	2908.721
床	2803.36
中餐厅	2803.36
挺不错	2803.36
只能	2803.36
标准	2803.36
浴缸	2803.36
问	2803.36
距离	2803.36

图 11 酒店 H01 热门词 TOP20 及相应热度

3.2. 问题 2 的分析过程

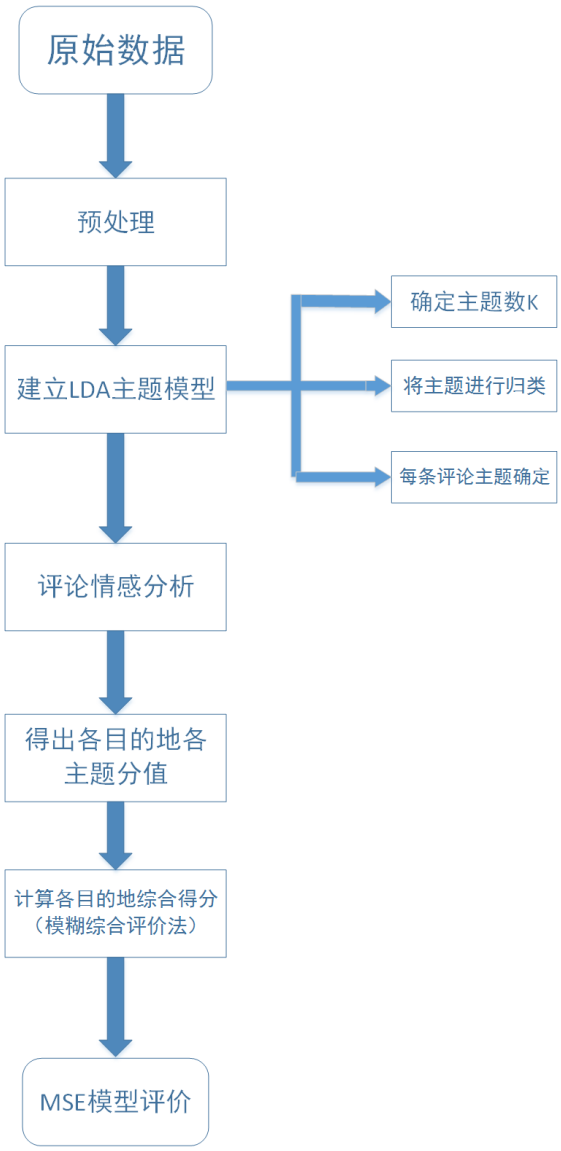

图 12 问题 2 分析流程图

3.2.1. 数据进一步处理

本题要求从评论文本中提取出评论所提及的主题以及相应的评分。通过观察评论文本，我们发现有部分评论仅仅有“还不错”“很好”之类的词语，能够通过情感分析得出情感倾向得分，但是并不能够判断出该评论所评价的是哪一方面。

为减少此类评论对最终模型结果的影响，小组在前文分词结果的基础上，将含有名词的评论标注出来，以供后续分析使用。

3.2.2. LDA 主题模型的建立

LDA 主题模型在文本分析中有广泛应用，常用于进行文本分类。对于每一篇文档，LDA 主题模型可以集中每篇文档的主题，并以概率分布的形式给出。通过分析文档的主题分布，可以根据主题对文档进行主题聚类或文档分类。

建立 LDA 主题模型需要先建立词袋模型。所谓词袋模型，是经文本处理后统计各个词语是否出现，以及出现的次数，而不考虑词语出现的顺序。

对于每一段文本而言，它包含的每一个潜在主题都有各自的关键词分布，这也就是说文本的每一个词语都以一定的概率归属于某个主题。因此，文本中每个词语出现的概率可以表示为：

$$p(词语|文本) = \sum_{主题} p(词语|主题) \times p(主题|文本)$$

即某个词语在某一文本中出现的概率为这一词语在某一主题中出现的概率与该主题在该文本中出现的概率之积对每一个主题求和。

过程图如下：

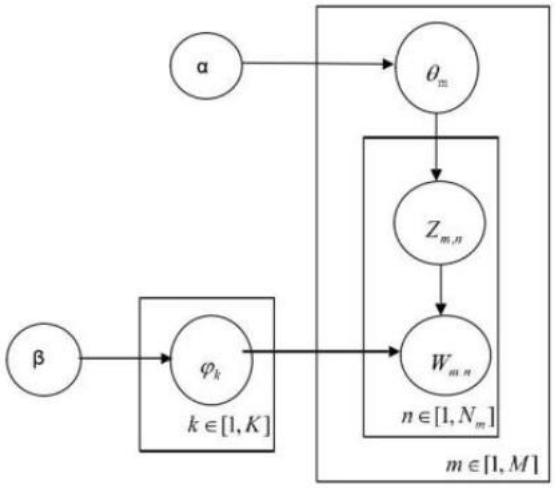

图 13 LDA 主题模型过程图

其中，M为评论总数，K为潜在的主题数，N_m表示在第m篇文档中词语的总数。β和α都是狄利克雷先验参数，但两者面向的对象不同。前者面向的是每个主题下词语的多项式分布，而后者面向的是每篇文档（即每条评论）下的主题的多项式分布。$Z_{m,n}$表示第m篇文档中第n个词语所属的主题，$W_{m,n}$表示在第m篇文档中的第n个词语。隐含变量θ_m表示在第m篇文档下所生成的主题分布，向量维度是n维。另一个隐含变量φ_k表示在第k个主题下所生成的词语的分布，它的向量维度是所有文档所包含的词语的总数。

首先，我们需要确定一个最佳的 K 值（即主题数目），这是建立 LDA 模型的关键，对模型结果将产生非常大的影响，而合适的主题数也能防止模型过拟合。

在 LDA 主题模型的建立中，一致性得分是常见的评判该主题数是否合理的指标之一。一致性得分越高，说明此时模型聚类效果越好，即此时设置的主题数较为合理。一致性得分的具体计算方法如下：

$$coherence(D) = \sum_{(v_i, v_j) \in V} score(v_i, v_j, \epsilon) \qquad (11)$$

其中，V是描述某个主题的词语集合，ϵ是一个平滑因子，以确保返回的值是一个实数。

$$score(v_i, v_j, \epsilon) = log \frac{D(v_i, v_j) + \epsilon}{D(v_j)} \qquad (12)$$

公式（12）是 UMass 度量标准的分数计算方法，其中$D(v_i, v_j)$指包含词语v_i和v_j的评论数目，$D(v_j)$表示包含词语v_j的评论数目。本文将用上述两个公式计算每个主题数的一致性得分（Topic Coherence）。

我们通过设置不同的主题数计算一致性得分，并将数据进行可视化处理，具体主题数与一致性得分的关系如下图所示：

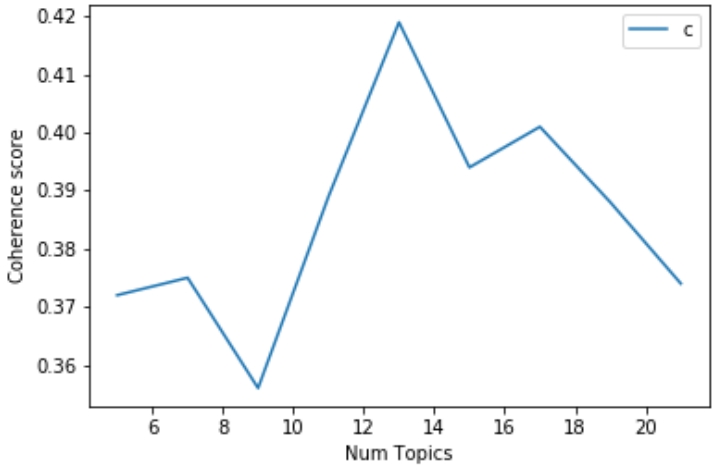

图 14 景区主题数与一致性得分关系图

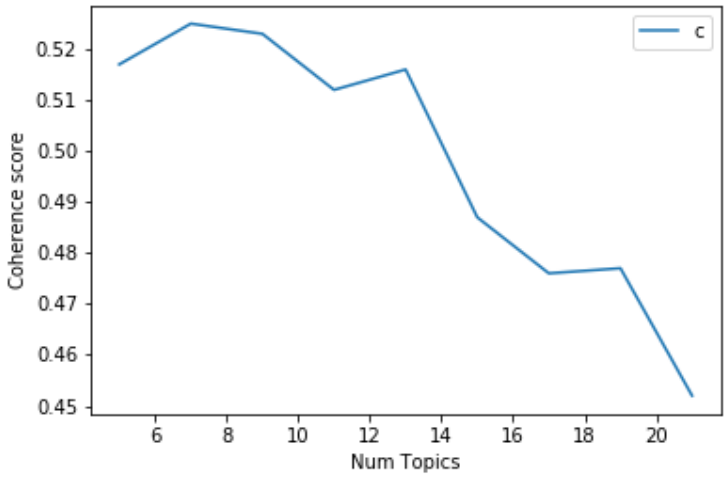

图 15 酒店主题数与一致性得分关系图

分别计算出各主题数下一致性得分的数值，得出当景区评论的主题数设置为 13，酒店评论的主题数设置为 7 时，一致性得分最高，即此时模型最有效。

确定主题数后，利用 Python 中的 gensim 库进行编程，构建 LDA 主题模型并生成主题分布。每个主题所包含的关键词及其出现的概率如下所示：

```
[(0,
  '0.053*"值得" + 0.048*"适合" + 0.030*"环境" + 0.025*"干净" + 0.021*"休闲" + 0.020*"公园" '
  '+ 0.020*"空气清新" + 0.018*"瀑布" + 0.018*"水" + 0.017*"很漂亮" + 0.015*"好去处" + '
  '0.015*"度假" + 0.014*"特色" + 0.012*"很大" + 0.012*"环境优美"'),
 (1,
  '0.126*"风景" + 0.033*"表演" + 0.028*"值得" + 0.024*"一般般" + 0.024*"位于" + '
  '0.022*"晚上" + 0.021*"好看" + 0.020*"建筑" + 0.019*"美丽" + 0.018*"大自然" + 0.017*"旧" '
  '+ 0.015*"值得一看" + 0.013*"节目" + 0.012*"风格" + 0.011*"特色"'),
 (2,
  '0.127*"好玩" + 0.055*"开心" + 0.049*"太" + 0.030*"天气" + 0.024*"项目" + 0.023*"特别" '
  '+ 0.022*"舒服" + 0.022*"夏天" + 0.021*"热" + 0.020*"风景" + 0.018*"乐园" + '
  '0.018*"水上" + 0.017*"刺激" + 0.013*"避暑" + 0.012*"下次"'),
 (3,
  '0.027*"走" + 0.027*"高" + 0.015*"性价比" + 0.014*"免费" + 0.014*"太" + 0.013*"坐" + '
  '0.012*"建议" + 0.012*"观音" + 0.011*"开车" + 0.011*"感觉" + 0.010*"路" + 0.010*"好好" '
  '+ 0.009*"收费" + 0.009*"游客" + 0.008*"栈道"'),
 (4,
  '0.120*"便宜" + 0.049*"价格" + 0.046*"优惠" + 0.045*"网上" + 0.031*"现场" + 0.027*"不用" '
  '+ 0.026*"实惠" + 0.023*"排队" + 0.022*"购票" + 0.016*"订" + 0.015*"购买" + '
  '0.014*"划算" + 0.014*"行" + 0.011*"提前" + 0.010*"窗口"'),
 (5,
  '0.122*"温泉" + 0.080*"服务" + 0.026*"服务态度" + 0.024*"满意" + 0.017*"泡温泉" + '
  '0.017*"热情" + 0.015*"家人" + 0.014*"工作人员" + 0.014*"方便快捷" + 0.014*"态度" + '
  '0.014*"值得" + 0.014*"赞" + 0.012*"推荐" + 0.012*"度假村" + 0.012*"池子"'),
 (6,
  '0.015*"上山" + 0.015*"广东" + 0.015*"导游" + 0.012*"山顶" + 0.012*"时" + 0.011*"走" + '
  '0.010*"到达" + 0.010*"两个" + 0.010*"下午" + 0.009*"坐" + 0.008*"车" + 0.008*"想" + '
  '0.008*"门口" + 0.008*"水质" + 0.007*"下山"'),
 (7,
  '0.053*"环境" + 0.048*"贵" + 0.036*"很好" + 0.031*"没什么" + 0.027*"吃" + 0.022*"东西" '
  '+ 0.019*"太" + 0.018*"差" + 0.018*"感觉" + 0.016*"风景" + 0.014*"价格" + 0.013*"还好" '
  '+ 0.013*"索道" + 0.013*"一家" + 0.013*"海鲜"'),
 (8,
  '0.166*"空气" + 0.049*"设施" + 0.032*"可惜" + 0.021*"完善" + 0.019*"还行" + 0.018*"游乐" '
  '+ 0.014*"有意思" + 0.013*"老板" + 0.012*"负离子" + 0.011*"商业" + 0.011*"房子" + '
  '0.010*"不错" + 0.010*"雨" + 0.008*"楼梯" + 0.007*"遗憾"'),
 (9,
  '0.058*"山" + 0.031*"爬山" + 0.023*"好多" + 0.015*"希望" + 0.014*"太" + 0.014*"美的" + '
  '0.013*"钱" + 0.012*"水" + 0.012*"骑" + 0.010*"活动" + 0.009*"游泳池" + 0.009*"很美" + '
  '0.008*"坑" + 0.008*"长" + 0.008*"感觉"'),
 (10,
  '0.075*"喜欢" + 0.074*"下次" + 0.041*"漂亮" + 0.028*"朋友" + 0.027*"值得" + 0.027*"去过" '
  '+ 0.022*"机会" + 0.021*"字" + 0.021*"开心" + 0.020*"风景优美" + 0.018*"游戏" + '
  '0.017*"环境" + 0.016*"第一次" + 0.016*"感觉" + 0.015*"机动"'),
 (11,
  '0.141*"景色" + 0.080*"美" + 0.026*"感觉" + 0.023*"山上" + 0.022*"值得" + 0.020*"中" + '
  '0.018*"优美" + 0.013*"拍" + 0.012*"交通" + 0.011*"少" + 0.011*"太" + 0.011*"山顶" + '
  '0.010*"特别" + 0.009*"国庆" + 0.008*"村"'),
 (12,
  '0.048*"动物" + 0.032*"爬" + 0.029*"岛" + 0.024*"沙滩" + 0.020*"天然" + 0.019*"氧吧" + '
  '0.016*"表演" + 0.015*"广州" + 0.015*"动物园" + 0.013*"种类" + 0.012*"特别" + '
  '0.012*"景观" + 0.012*"游泳" + 0.011*"鱼" + 0.011*"珠海"')]
```

图 16 景区 LDA 主题模型运行结果

```
[(0,
  '0.038*"服务" + 0.026*"房间" + 0.021*"升级" + 0.020*"入住" + 0.020*"免费" + 0.016*"早餐" '
  '+ 0.015*"热情" + 0.014*"推荐" + 0.014*"满意" + 0.013*"不错"'),
 (1,
  '0.076*"服务" + 0.057*"干净" + 0.051*"卫生" + 0.032*"房间" + 0.020*"特别" + 0.016*"热情" '
  '+ 0.013*"温泉" + 0.012*"整洁" + 0.011*"喜欢" + 0.010*"入住"'),
 (2,
  '0.031*"房间" + 0.019*"位置" + 0.018*"旧" + 0.018*"地理位置" + 0.017*"住" + 0.015*"设施" '
  '+ 0.014*"机场" + 0.013*"早餐" + 0.011*"出行" + 0.011*"出差"'),
 (3,
  '0.030*"住" + 0.022*"舒服" + 0.020*"房间" + 0.017*"感觉" + 0.016*"入住" + 0.014*"性价比" '
  '+ 0.014*"下次" + 0.012*"太" + 0.010*"床" + 0.009*"高"'),
 (4,
  '0.014*"房" + 0.010*"旅游" + 0.009*"服务" + 0.008*"时" + 0.007*"房间" + 0.007*"好评" + '
  '0.007*"游玩" + 0.007*"套房" + 0.007*"选择" + 0.007*"订"'),
 (5,
  '0.124*"不错" + 0.066*"环境" + 0.060*"服务" + 0.054*"早餐" + 0.035*"位置" + 0.032*"交通" '
  '+ 0.028*"设施" + 0.027*"房间" + 0.022*"性价比" + 0.018*"高"'),
 (6,
  '0.024*"房间" + 0.020*"差" + 0.017*"服务态度" + 0.017*"入住" + 0.015*"疫情" + '
  '0.013*"空调" + 0.012*"太" + 0.011*"期间" + 0.010*"点" + 0.009*"体验"')]
```

图 17 酒店 LDA 主题模型运行结果

之后，我们根据各主题出现概率较高的几个关键词，将该主题归属于服务、位置、设施、卫生、性价比五大方面的其中一个。

归类结果如下：

表 1 景区主题归类结果

归类	主题序号
卫生	0
服务	1，5，6
位置	10，11
性价比	2，8，9，12
设施	3，4，7

表 2 酒店主题归类结果

归类	主题序号
卫生	1
服务	0
位置	2
性价比	3，5
设施	4，6

3.2.3.评论情感分析

要对评论文本进行情感分析，首先要对文本所含的词语根据其情感的积极性、消极性进行分类。我们所采用的判断评论情感方法的基本思路是：当一个词普遍被认为跟消极、中性或是积极的情感有关联时，将这个词赋予分值-1、0 或+1，再将整条评论的词语权重加和，去判定消费者对一个文本（一条评论）的感情。

具体步骤如下：

（1） 步骤一：构建词语感情倾向表

构建词语感情倾向表采用词典匹配法，将评论处理后获得的词语集与知网发布的“情感分析用词语集（beta 版）”进行匹配，合并具有相同词语的部分。然后分别根据词语的感情色彩对其赋分，正面词语赋分为 1 分，负面词语赋分为-1 分，从而构建出词语感情倾向表。

（2） 步骤二：修正情感倾向

由于中文表达中时常存在“否定词+肯定评价”或“否定词+否定评价”等情况，单纯按照一个词语的感情倾向进行情感判断可能会出现与原意相反的结果。因此需要将否定词纳入判断范围，如果根据情感倾向判断为正面后，含有奇数个否定词，则情感倾向与原来相反，含有偶数个否定词，则情感倾向与原来相同。于是，我们引入否定词表，对每个正向情感词前面的两个词语进行搜索，若否定词个数为奇数，则将该情感词修正为负向情感词，否则不做修正。

（3） 步骤三：计算每条评论的情感得分并划分情感等级

将整条评论词语的情感得分加和得到每一条评论的情感得分。然后根据得分区间，将情感划分为 5 个等级，等级 1 到 5 分别为非常不满意、不满意、一般、满意、非常满意。最后得到不同景区和酒店的情感分析结果。

3.2.4.五个方面的游客满意度统计

由上述 LDA 主题模型，我们可以计算每一条评论语句中五个主题出现的概率值。为方便统计，我们选取每条评论概率值最高的一个方面，将该评论归属到该主题下。确定评论主题归属后，可获得各方面的游客满意度统计表。

3.2.5.模糊综合评价法得出综合得分

接着运用模糊综合评价法对各目的地的综合得分进行定量分析。基本过程如下：

（1）建立因素集$U=(\mathrm{u}_1,\mathrm{u}_2,\ldots,\mathrm{u}_n)$，其中$\mathrm{u}_1,\mathrm{u}_2,\ldots,\mathrm{u}_n$分别表示各评价因素，即主题；以及评语集$V=(v_1,\mathrm{v}_2,\ldots,\mathrm{v}_m)$，其中$v_1,\mathrm{v}_2,\ldots,\mathrm{v}_m$分别表示不同的评语等级。

（2）设定因素集U中每个评价因素的权重值，用模糊子集$A=(a_1,a_2,\ldots,a_n)$来表示，其中a_i表示每个评价因素所对应的权重，满足条件$\sum a_i=1$。

（3）模糊子集$R_i=(r_1,r_2,\ldots,r_m)$表示评语集V中单个因素的模糊综合评价矩阵，当有多个因素时，其评价矩阵R如下所示：

$$R=\begin{bmatrix}R_1\\R_2\\\cdots\\R_n\end{bmatrix}=\begin{bmatrix}r_{11}&\cdots&r_{1m}\\\vdots&\ddots&\vdots\\r_{n1}&\cdots&r_{nm}\end{bmatrix} \tag{13}$$

最终得到各目的地的模糊综合评价集B，$\mathrm{B}=\mathrm{A}\times R$。

结合题目中的数据，评价因素集为$U=(\mathrm{u}_1,\mathrm{u}_2,\ldots,\mathrm{u}_5)$，其中$\mathrm{u}_1,\mathrm{u}_2,\ldots,\mathrm{u}_5$分别代表景区及酒店的服务、位置、设施、卫生、性价比；评语集$V=(v_1,\mathrm{v}_2,\ldots,\mathrm{v}_5)$，其中$v_1,\mathrm{v}_2,\ldots,\mathrm{v}_5$分别代表 1，2，3，4，5 分，即游客的满意度为非常不满意，不满意，一般满意，满意，非常满意。

对于各因素的权重分配，考虑到游客评论的出发点和侧重点不同，评价次数可以很好地反映游客对该维度的重视程度。因此，我们将用户在某一评价维度的点评次数占总点评次数的比例作为该维度在满意度评价中的权重A。

以酒店 H50 为例，计算得其评判矩阵 R：

```
array([[0.        , 0.08333333, 0.25      , 0.41666667, 0.25      ],
       [0.16666667, 0.01851852, 0.5       , 0.22222222, 0.09259259],
       [0.22222222, 0.08888889, 0.4       , 0.17777778, 0.11111111],
       [0.        , 0.07142857, 0.28571429, 0.42857143, 0.21428571],
       [0.09565217, 0.0173913 , 0.53043478, 0.24347826, 0.11304348]])
```

图 19 H50 的评价矩阵 R

各评价维度的权重A为

$$\mathrm{A}=(0.05,\quad 0.225,\quad 0.1875,\quad 0.05833333,\quad 0.47916663) \tag{14}$$

即为服务、位置、设施、卫生、性价比五个方面的权重值。

其他景区及酒店均如此操作，可得各目的地的综合得分以及各方面得分，具体信息详见附件。

3.2.6.均方误差（MSE）模型评价

均方误差（Mean Squared Error，MSE）是指参数估计值与参数真值之差平方的期望值，可以衡量数据的变化程度。MSE 的值越小，说明该预测模型与实验数据之间具有更好的精确度。

均方误差的表达式为：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y-\mathrm{f(x)})^2 \tag{15}$$

分别将得出的每个景区及酒店的服务、位置、设施、卫生、性价比得分以及综合评价得分与附件 2 中的得分对比，得出该模型计算得出的景区综合评价均方误差为 1.68，服务、位置、设施、卫生、性价比的均方误差分别为 0.77，2.82，1.73，3.44，1.68；酒店的综合评价均方误差为 2.66，服务、位置、设施、卫生、性价比方面得分的均方误差分别为 3.31，2.71，1.83，3.44，0.86。由此可见，可见对于景区，该模型的拟合效果较好，但仍有不小的误差，其中卫生方面的误差最大，可能与主题归类不当、评论数目不足有关；对于酒店，该模型的拟合效果一般，在服务、卫生方面的估计误差较大，或许与评论数目不足、具体情感倾向较难量化有关。相比而言，该模型对于景区的预测效果更好一些。

3.3. 问题 3 的分析过程

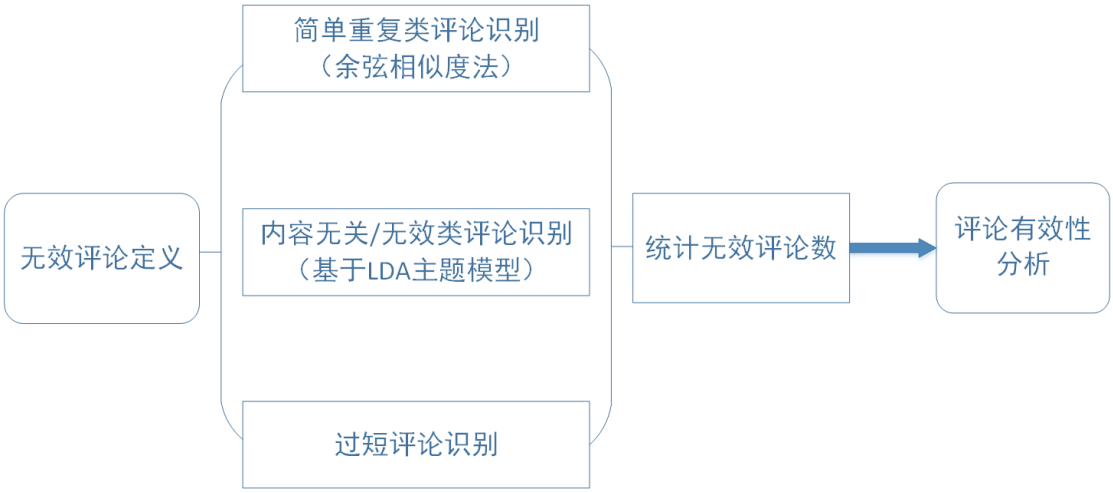

图 19 问题 3 分析流程图

3.3.1. 无效评论的定义

无效评论是指对旅游者参考价值较低、旅游者无法从中获取有效信息的评论。具体而言，无效评论有如下三类：

（1） 简单重复类评论。这类评论是指它是其他评论的简单复制或修改部分词语，影响了评论整体的情感倾向，但没有给旅游者提供除原有评论外的其他有效信息的评论。

（2） 内容无关、内容无效类评论。这类评论是指评论内容与景区、酒店完全无关，或从该评论中完全无法获取有价值有意义的信息的评论。“太棒了”、“还不错”这类评论就是典型的无效评论，旅游者无法从中获得更多关于旅游目的地的信息。

（3） 过短评论。部分评论可能出于某种目的，消费者仅仅留下几个字的评论，例如“不错”“很好”“还可以”等。这种评论虽然能够给到一定的情感倾向信息，并且改变相应词语的词频，但不能给消费者带来太多有效信息，故纳入无效评论。

在本题的分析过程中，我们将满足上述条件之一或更多的评论都视为无效评论。

3.3.2. 简单重复类评论的识别

简单重复类评论文本的识别需要识别评论之间的相似性。相似性需要衡量文本之间的相似水平，高于某个阈值的即视为相似评论。

向量分布空间的余弦相似度模型是识别文本相似性的一个常用模型。具体建模步骤

如下：

一个向量分布空间里 2 个向量夹角的余弦值可以衡量两个向量之间的相似度。余弦值越接近于 0，则夹角趋于 90°，表明两个向量越不相似；余弦值越接近于 1，则夹角越趋于 0°，则两个向量越相似。

对一条评论进行分词、对分词进行编码、计算词频后，得到词频向量$A=(a_1,a_2,\dots,a_n)$以及$B=(b_1,b_2,\dots,b_n)$，计算方法如下：

$$\cos\theta=\frac{\sum_{t=1}^{n}(a_t\times b_t)}{\sqrt{\sum_{t=1}^{n}{a_t}^2}\times\sqrt{\sum_{t=1}^{n}{b_t}^2}}=\frac{A\times B}{|A|\times|B|} \qquad (16)$$

余弦相似度模型的优点在于它适合应用于每一方向的向量映射，而且是从方向上区分每一对照物之间的差异。然而，余弦相似度模型对于每一向量上的单一个体的标准数值不敏锐，这意味着无论该个体是否出现，都可以有效抵消关键字，而不必过于关注事件发生的具体次数或重要性。再者，余弦相似度模型表示相似度的数值取值范围为[0,1]，与相似度成正比，清晰直观。

通过多次试验和人工评定，小组确定以下标准：对于景区的评论，认为相似度高于 0.98 的评论为相似评论；对于酒店，认为相似度高于 0.95 的评论为相似评论。

gensim 模块用于自然语言处理，使用其中 Word2Vec 库将每一条评论文本转换为词向量再计算两个词向量之间的余弦相似度，相似度大于阈值时为几乎相同的评论。

识别出的部分相似评论如下：

---------40---------
带着老人和宝宝一起去的 很不错 一天的时间就足够了 因为是动物园 为动物健康考虑 规定不可以带食品入园 只能在园内购买食品 是可以理解的 但食品的价格实在让人不敢恭维 太贵了 不让带食物入园有点变相让人消费了 除了这点 其他的都很好

带着老人和宝宝一起去的 很不错 一天的时间就足够了 因为是动物园 为动物健康考虑 规定不可以带食品入园 只能在园内购买食品 是可以理解的 但食品的价格实在让人不敢恭维 太贵了 不让带食物入园有点变相让人消费了 除了这点 其他的都很好

图 20 简单重复类评论示例 1

---------222---------
价格实惠取票方便挺好的

挺好的取票方便价格实惠

价格实惠取票方便挺好的

取票很方便，价格实惠

价格实惠取票方便挺好的

取票很方便，价格也实惠

图 21 简单重复类评论示例 2

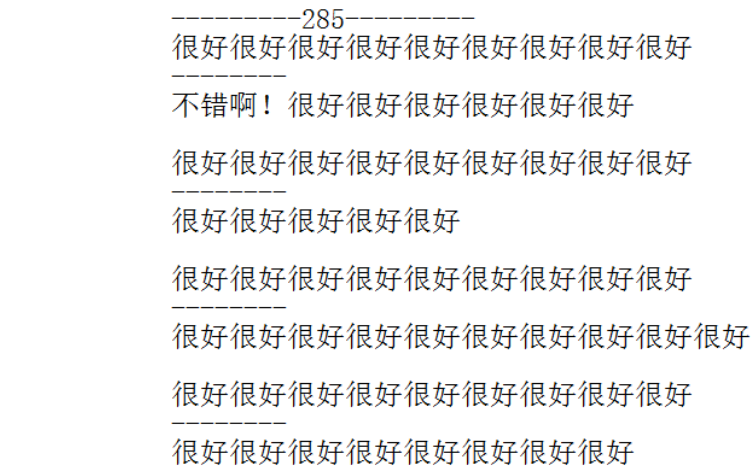

图 22 简单重复类评论示例 3

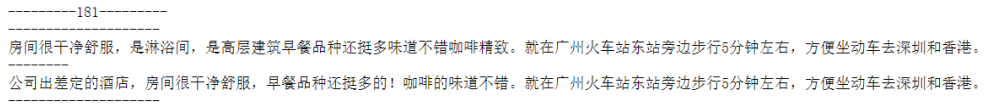

图 23 简单重复类评论示例 4

程序运行结果表明，余弦相似度识别法能够有效识别出内容完全一致（有标点符号或空格不一样，如示例 1）、词语相同但语序不同（如示例 2）、同个词语的重复次数不同（如示例 3）、部分内容完全相同（如示例 4）的简单重复类评论，证明该识别方法有效可行。

对于前 2281 条景区评论，发现其中存在相似评论 59 条，存在相似评论的比例约为 2.587%。从简单重复类评论比例来看，景区评论的有效性较高。

对于前 500 条酒店评论，发现其中存在相似评论 164 条，存在相似评论的比例约为 32.8%，远高于景区评论。由此看来，酒店评论的有效性一般。

3.3.3. 内容无关、内容无效类评论识别

由问题 2 所建立的 LDA 主题模型，评价主题共分为服务、位置、设施、卫生、性价比五个方面，其中每个主题分别对应不同的关键词。根据关键词的出现频率，我们对每一条评论都可以得出其关于每个主题的概率值，以及每条评论对主题的贡献度。

对于内容无关、内容无效类的评论，我们容易发现这些评论对主题的贡献度都非常低，因此旅游者无法从中得到有价值的信息。因此，只要设定一个阈值，使得当每条评论贡献度最高的数值低于该阈值时，该评论则被判定为无效评论。

经观察，主题贡献度低于 0.23 的评论，基本可视作与主题关联不大，可视作无效评论。

经统计，景区本类无效评论共 2265 条，酒店 619 条。

3.3.4. 过短评论的识别

过短评论也会影响评论的有效性，由于评论文本过短，通常不能给到旅游者有价值的信息。因此还需要对评论文本的长度进行分析，去掉过短的评论。

分别统计景区和酒店的评论文本特征，结果如下：

景区评论文本长度的均值为 42.5669，即评论字数平均值约为 43 个字。方差为 3027.5881，说明评论的长度差别很大。偏度为 6.3289，说明数据呈右偏趋势，即评论字数小于平均值的偏多。

文本长度的均值为 35.6619，即评论字数平均值约为 35 个字，比景区评论少 10 个字左右。方差为 3602.9565，说明评论的长度差别很大。偏度为 10.8033，说明数

据呈右偏趋势，即评论字数小于平均值的偏多，且右偏程度高于景区评论。

由于评论字数的右偏程度较大，证明过短评论占据了一定的数目，对评论的有效性有较大的影响。因此小组规定文本长度小于 5 个字符的评论为过短评论，通过统计各文本长度，筛选出景区过短评论 21 条，酒店过短评论 54 条。

3.3.5. 评论文本的有效性分析

经过小组筛选及统计，景区评论共 59106 条，其中，内容无关/无效类评论 2265 条，过短评论 21 条，简单重复类占比约 2.587%，评论有效率 93.55%。

酒店评论共 25225 条，其中，内容无关/无效类评论 619 条，过短评论 54 条，简单重复类评论占比约 32.8%，评论有效率 64.53%。

容易发现酒店评论的有效率远低于景点评论。经观察评论及查阅相关文献，小组认为评论有效率出现差异的原因在于酒店的游玩活动较少，大部分旅游者考察的方面是服务以及舒适度，因此差异性较小，重复评论出现的频率较高。

3.4. 问题 4 的分析过程

图 24 问题 4 的分析流程图

3.4.1. 景点与酒店的分级

问题 2 根据评论文本得出了各个景区及酒店的各维度得分以及综合评价得分。根据综合评价得分，我们将景区及酒店分为“高、中、低”三个等级，其中综合评价得分在前 33.3%的为高级，后 33.3%的为低级，剩余的为中级。

各个景区及酒店的分级结果如下：

表 3 基于模糊综合评价的景点分级结果

等级	景点编号
高	A01, A02, A07, A09, A11, A14, A15, A16, A17, A18, A19, A22, A26, A29, A30, A32, A33
中	A12, A13, A21, A23, A24, A25, A27, A35, A36, A37, A39, A40, A42, A45, A46, A48
低	A03, A04, A05, A06, A08, A10, A20, A28, A31, A34, A38, A41, A43, A44, A47, A49, A50

表 4 基于模糊综合评价的酒店分级结果

等级	酒店编号
高	H01, H04, H07, H10, H11, H14, H17, H19, H21, H25, H26, H28, H32, H33, H36, H40, H42
中	H02, H03, H05, H06, H08, H09, H12, H22, H23, H27, H35, H37, H43, H44, H47, H49
低	H13, H15, H16, H18, H20, H24, H29, H30, H31, H34, H38, H39, H41, H45, H46, H48, H50

3.4.2. 基于 TF-IDF 算法的目的地特征提取

TF-IDF （term frequency–inverse document frequency）算法用以统计一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度，是关键词提取的最具代表性的方法之一，是一种用于信息检索、文本挖掘的常用加权技术。TF-IDF 算法的理论依据是：字词的重要性随着它在文件中出现的次数成正比增加，随着它在语料库中出现的次数成反比减少。

TF-IDF 算法的主要思想是：如果一个词语在某一个文档中出现的频率高，并且在其他文档中出现的频率低，那么这个词语具有很好地类别区分能力，可以用于分类。

对于目的地的特征提取，TF-IDF 算法也可以得到应用。在对评论文本进行分词处理后，统计出高频词，获得各目的地的高频词集；将每一目的地的所有评论合成为一个文档，即为 TF-IDF 算法中的一份“文件”，所有“文件”统称为一个“语料库”；对高频词集内每一词语进行分析，若该词语在它所在的“文件”中出现次数较多，即在该目的地的评论中被多次提及，而在其余“文件”中出现次数很少，那么我们就可以称这个词语能够代表该目的地的一个特征。

具体计算方法如下：

TF（Term Frequency）指的是词频，即某一个词语在某一文档中出现的频率。公式如下：

$$TF_{i,j} = \frac{\mathrm{n}_{i,j}}{\sum_k \mathrm{n}_{k,j}} \qquad (17)$$

其中，$\mathrm{n}_{i,j}$是该词语在文件d_j中出现的次数，$\sum_k \mathrm{n}_{k,j}$表示所有词语在文件d_j中出现次数的总和。

IDF（Inverse Document Frequency）指的是逆向文件频率，通过计算总文件数目除以包含该词语的文件的数目，再取对数。公式如下：

$$IDF_i = \log \frac{|D|}{|\ \{j: t_i \in d_j\}\ |} \qquad (18)$$

其中，$|D|$是语料库中文件的总数；$|\ \{j: t_i \in d_j\}\ |$表示包含词语t_i的文件的数目。

TF-IDF 的值即为 TF 和 IDF 的值的乘积：

$$TF - IDF_{i,j} = \mathrm{TF}_{i,j} \times \mathrm{IDF}_i = \frac{\mathrm{n}_{i,j}}{\sum_k \mathrm{n}_{k,j}} \times \log \frac{|D|}{|\ \{j: t_i \in d_j\}\ |} \qquad (19)$$

3.4.2. 景点特色的发掘

根据之前的分级结果，本文选取了高、中、低层次的各三家景点，共 9 家景点进行特色发掘。分析结果如下：

A01（高）：

```
1[[('马戏', 0.0068471866987599795),
  ('动物', 0.005328241036481537),
  ('过山车', 0.0042192231924923895),
  ('动物园', 0.004145827461524677),
  ('垂直', 0.0030398277231958024),
  ('火车', 0.0027580396369846034),
  ('十环', 0.0022117805377703044),
  ('北门', 0.002201602802846129),
  ('南门', 0.00210458901822699),
  ('熊猫', 0.0017764638876081474)],
```

图 25 A01 关键词 TF-IDF 值 TOP10 表

A01 是一个高分景点。从关键词看，该景点应该是一个综合性的游乐场，有动物园以及马戏表演，还有过山车、垂直过山车等游乐设施。“北门”、“南门”应该是交通方面较方便的出入口。

A07（高）：

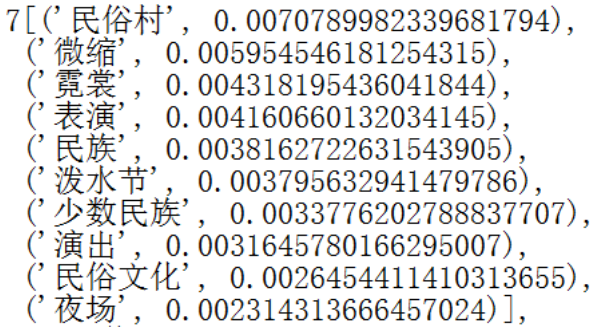

图 26 A07 关键词 TF-IDF 值 TOP10 表

景点 7 是一个充满文化气氛的民族文化村。“微缩”表明该景点浓缩了许多少数民族的特色村落，另外，民俗文化的表演也是吸引游客的因素之一。该景点未来需要继续保持其独特性，用少数民族的特色文化吸引各地游客。

A11（高）：

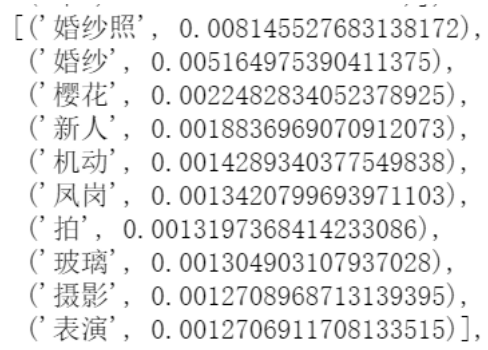

图 17 A11 关键词 TF-IDF 值 TOP10 表

从上面的词语中可以推断出，A11 应该是一个供结婚新人去拍摄婚纱照的地方，可能位于凤岗。“玻璃”也成为关键词之一，可能是有玻璃栈道一类的特色景点。“樱花”也是游客前往该目的地的原因之一。

A21（中）：

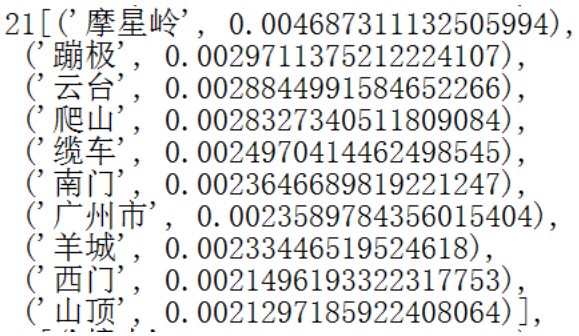

图 27 A21 关键词 TF-IDF 值 TOP10 表

可以推断出该景点位于羊城广州，是一座山。同时，这座山还有“摩星岭”、“蹦极”“缆车”等特色景点以及设施，是游客的旅游好去处。评分为中级的原因可能有景区管理、交通出行、娱乐设施等原因，需要景点根据现实情况作出调整和改进。

A39（中）：

```
[('动物', 0.010795855213964046),
 ('茂名', 0.006992255062999563),
 ('公园', 0.0037514164285005107),
 ('茂名市', 0.003407546313671114),
 ('动物园', 0.003303836466082355),
 ('勇敢者', 0.0021850797071873633),
 ('老虎', 0.0020411282015659646),
 ('勇者', 0.0016708446318789293),
 ('植物', 0.0015591194994548825),
 ('猛兽', 0.0014674402453424306)],
```

图 28 A39 关键词 TF-IDF 值 TOP10 表

从景点 A39 的评论中提取出的 TF-IDF 值最高的 10 个词语中可以推断出该景点应该是一个位于茂名市的动物园，且该动物园以猛兽、老虎较为著名，因此吸引了许多游客前去游玩。

A48（中）：

```
48[('五台山', 0.03609622983645124),
 ('五爷', 0.009788808091241014),
 ('佛教', 0.0069726016807036524),
 ('庙', 0.005227160026456591),
 ('凑', 0.004845377384469278),
 ('五台', 0.0036708030342153804),
 ('菩萨顶', 0.0024472020228102536),
 ('台顶', 0.0024472020228102536),
 ('黛螺顶', 0.0024472020228102536),
 ('塔院寺', 0.0024472020228102536)],
```

图 29 A48 关键词 TF-IDF 值 TOP10 表

从中可以推断出该景点位于五台山，是一个信奉佛教的游客喜爱的旅游地。其中“菩萨顶”“黛螺顶”“塔院寺”是这个景点的特色。该景点要增强竞争力，需要发挥其佛教特色。

A38（低）：

```
[('温泉', 0.015213247463310946),
 ('阳西', 0.010950907685437354),
 ('池', 0.0056990923116471024),
 ('咸水', 0.005333436043799302),
 ('泡', 0.0031548546725189315),
 ('泡温泉', 0.0028373844960826134),
 ('多万平方米', 0.002265705038366349),
 ('毛巾', 0.002118633489790671),
 ('温泉水', 0.0019348784862497778),
 ('阳江', 0.0018908698001080498)],
```

图 30 A38 关键词 TF-IDF 值 TOP10 表

可以推断出，该景点是一个位于阳江的温泉景点，该景点特色为温泉，而且可能温泉占地面积较大。该景点得分较低的原因可能是卫生方面的不足或是设施不够充足，以致游客满意度不高。

A04（低）：

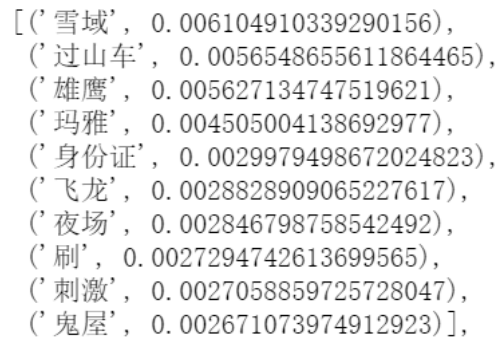

图 31 A04 关键词 TF-IDF 值 TOP10 表

从关键词中可以看出该景点应该是一个游乐场，有过山车、鬼屋等设施，分为“雪域雄鹰”、“飞龙”等主题，进场需要出示身份证。许多游客对于该游乐场的评价为比较刺激。

A43（低）：

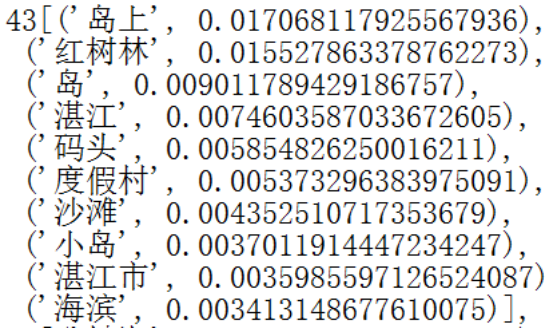

图 32 A43 关键词 TF-IDF 值 TOP10 表

可以推断出该景点位于湛江市海滨，是一个度假村。游客在这里可以住在小岛上，享受沙滩、大海带来的快乐。评分较低可能的原因是卫生不到位、交通不便、设施不齐全等等，需要景点根据自身实际情况进行提升。

3.4.3.酒店特色的发掘

和景点相似，本文选取高、中、低层次的各三家酒店，共 9 家酒店进行特色发掘。结果如下：

H04（高）：

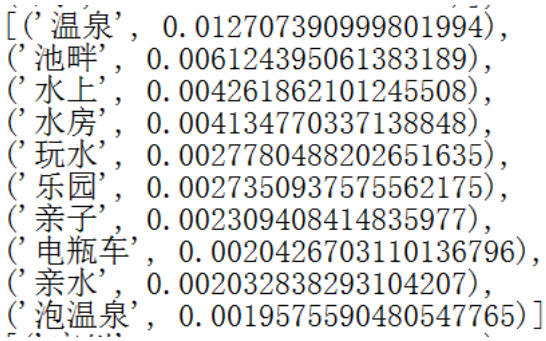

图 33 H04 关键词 TF-IDF 值 TOP10 表

可以推断出这家酒店的特色是温泉，同时配备有水上乐园、电瓶车等设施，方便需要带小朋友的游客前来游玩。这些设施或许就是这家酒店评分较高的因素，该酒店未来也可以从维护和提升设施体验入手加强竞争力。

H07（高）：

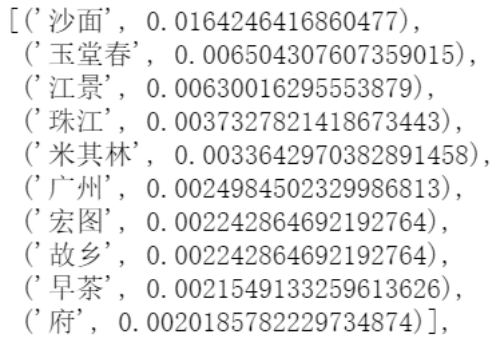

图 34 H07 关键词 TF-IDF 值 TOP10 表

从上面的关键词中可以推断出该酒店位于广州沙面，毗邻珠江。其中的餐饮是它的特色，“玉堂春”应该是餐厅名称，“米其林”表明该餐厅出品获得肯定。游客大多来这里欣赏江景，品尝广州早茶。“故乡”有可能指这里的氛围让人想起故乡的味道。

H10（高）：

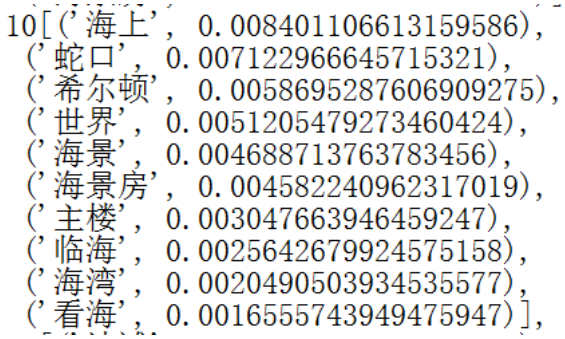

图 35 H10 关键词 TF-IDF 值 TOP10 表

可以推断出该酒店是位于深圳蛇口的，临海。酒店特色是海景房，品牌是希尔顿。该酒店得分较高的原因可能是位置较好，品牌服务较好，设施齐全。

H06（中）：

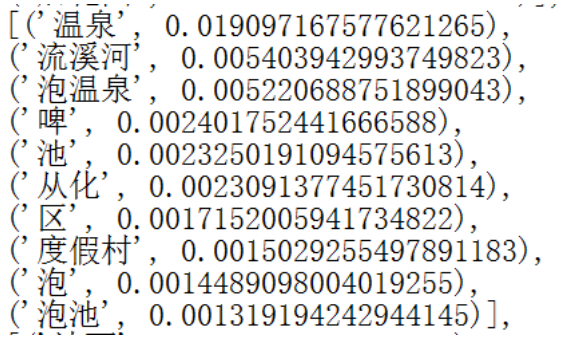

图 36 H06 关键词 TF-IDF 值 TOP10 表

可以推断出该酒店位于从化市的一个度假村，在流溪河流域。游客来到该酒店，可以享受温泉，这也是很多游客提到的特点。该酒店得分为中级的原因可能是位置交通不便，或是设施、服务方面有所不足。

H27（中）：

```
27[('小镇', 0.017231890695338654),
 ('欧洲', 0.008961363302482738),
 ('九龙湖', 0.0037438739628382956),
 ('生态', 0.002259048052797475),
 ('水景', 0.0014975495851353182),
 ('蹦床', 0.001247957987612765),
 ('纸杯', 0.0012322080287986228),
 ('欧式', 0.0011692081935420525),
 ('公主', 0.001076993168458666),
 ('拍照', 0.0010013103794277685)],
```

图 37 H27 关键词 TF-IDF 值 TOP10 表

该酒店位于九龙湖旁的一个欧洲风情小镇。游客在这里可以感受欧洲风情，拍摄照片，还有蹦床等设施。这里的生态环境、水景也是游客常提到的特点。该酒店得分为中，可能是因为服务、设施、卫生等方面有所不足。

H43（中）：

```
43[('小蛮', 0.014775855730884817),
 ('腰', 0.014401379624184358),
 ('塔', 0.013701598608099287),
 ('珠江', 0.008531218207154715),
 ('广州', 0.004684594186872527),
 ('江景', 0.004655292824289746),
 ('花城', 0.0029200891933473866),
 ('江边', 0.0027313076135199728),
 ('临江', 0.0022469990024063975),
 ('华', 0.002085777995248133)],
```

图 38 H43 关键词 TF-IDF 值 TOP10 表

可以推断出该酒店是位于广州市临江路的，可以在酒店眺望广州塔，欣赏珠江夜景。得分为中的原因可能有性价比、服务、卫生等方面的原因。

H13（低）：

```
[('港珠澳', 0.009165369131852456),
 ('大桥', 0.0077082688307636),
 ('韦', 0.004944475452709878),
 ('海景房', 0.00442452110984445),
 ('施', 0.004341490641403795),
 ('明月', 0.0037385058300977124),
 ('梁震', 0.0036179088678364958),
 ('海景', 0.003473019612686067),
 ('长隆', 0.0029074398224956106),
 ('糖水', 0.002129471890971039)],
```

图 39 H13 关键词 TF-IDF 值 TOP10 表

从上面关键词可以推断出，该酒店位于港珠澳大桥附近，应该临海，有海景房。“韦”、“施”“梁震”应该是酒店工作人员的姓名中的字，被多次提及，表明这家酒店服务被顾客肯定，也与其得分较高，属于高等级酒店对应。“长隆”一词也出现在关键词中，表明该酒店附近可能有长隆乐园。

H38（低）：

```
[('沙滩', 0.014457831325301205),
 ('海滩', 0.005732004291671412),
 ('海景房', 0.0034292046759910114),
 ('巽', 0.0034115863540883904),
 ('寮', 0.0034115863540883904),
 ('沙子', 0.00321285140562249),
 ('私家', 0.0018860625717950564),
 ('海边', 0.0018796037012985768),
 ('湾', 0.001664130278227184),
 ('私人', 0.0015981526452692272)],
```

图 40 H38 关键词 TF-IDF 值 TOP10 表

可以推断出该酒店位于巽寮湾的海滩边，配备有私家海景房。来这个景点旅游的游客主要是希望享受沙滩、大海的美丽景色。该酒店评分较低的原因可能是沙滩卫生方面有所欠缺，或是其他（如服务）方面不如竞争对手。

H48（低）：

```
48[('东门', 0.0069058433738209335),
 ('光华', 0.005173671487436936),
 ('国贸', 0.005173671487436936),
 ('老街', 0.003946196213611962),
 ('金', 0.002466372633507476),
 ('出岔', 0.0023979816574961448),
 ('立地条件', 0.0023979816574961448),
 ('日系集', 0.0023979816574961448),
 ('天虹', 0.0023979816574961448),
 ('黄金', 0.0023223218341737258)],
```

图 41 H48 关键词 TF-IDF 值 TOP10 表

可以推断出该酒店位于深圳市东门老街。评论中提到的“光华”“天虹”“黄金”可能是附近的商场名称。该酒店评分较低，可能是地理位置、服务等等因素造成的。

四、 挖掘目标

本文基于数据挖掘技术对旅游目的地的评论文本进行了分析，并得到了一些有价值的信息。通过数据挖掘，小组得到了各个目的地的印象词，提取出热度较高的词语；通过评论文本得出了每个目的地的综合评价得分以及服务、卫生、设施、位置、性价比五个方面的得分，并验证我们得出的评分与专家评分具有一定的一致性；通过挖掘无效评论，得到了景点及酒店的有效评论比例，并进行了有效性分析；对于目的地未来的发展，小组挖掘出每个目的地的特征和亮点，据此给出发展的建议。

总体而言，我们挖掘出的信息具有一定的参考意义，能在一定程度上反映景区及酒店目前的旅游满意度；然而，模型的拟合结果还有待提升，我们得出的综合评价得分与专家打分还存在一定的误差。误差存在的主要原因是情感倾向性的评判还不够准确，这与小组操作过程中的不全面性、模型本身的准确性有关，是我们在后续的数据挖掘中可以进一步探讨的地方。

五、 参考文献

[1]钱成. 基于网络评论的旅游目的地游客满意度分析[D].中南财经政法大学,2019.
[2]唐子豪. 基于改进 LDA 的在线商城垃圾评论识别研究[D].西安理工大学,2020.
[3]王宇超. 基于 CNN 的微博垃圾评论识别方法[D].中北大学,2020.
[4]池毛毛,潘美钰,王伟军.共享住宿与酒店用户评论文本的跨平台比较研究：基于 LDA 的主题社会网络和情感分析[J].图书情报工作,2021,65(02):107-116.
[5]胡汝佳. 基于游客感知的凤凰古城旅游形象的独特属性研究[D].湘潭大学,2020.
[6]曾子明,杨倩雯.基于 LDA 和 AdaBoost 多特征组合的微博情感分析[J].数据分析与知识发现,2018,2(08):51-59.