用户评论的文本挖掘、变量生成与指标优化

——基于携程平台景区游客评论的统计建模

0764 / 与发际线作队

2021 年 5 月 5 日

摘要

本文使用携程平台的景区游客评论进行文本挖掘，生成了多个变量来反映评论的“内蕴价值”，包括文本相似度、情感得分、词性分布、评论长度等，并系统地分析了不同变量的作用方式和效果。通过建立分类模型，本文分析了它们与用户评分的相关程度与相对重要性，进而设计了基于评论指标优化的数据产品，来对景区内部精选评论的选取排序及榜单的准入标准提供有效的改进建议。

第 1 节 背景介绍与研究问题

互联网时代涌现出一批致力于提供便捷旅游服务的电子商务平台，这类平台通过用户上传游记和评论为其他用户提供景点信息和游玩攻略，并通过一定的算法生成景点旅游排名并推出各类精选推荐榜单，这些榜单往往为游客制定出游计划提供参考。为了丰富平台内容，激励用户分享游记，平台往往会推出一些奖励计划激励用户分享游记，由此可能导致部分用户为获取奖励发布缺少实质信息的评论和评分，或是习惯性给出五星好评但在文字中表达着对旅途的不满；也可能存在景区或周边商家为提高自身热度吸引游客，通过大量制造好评数据增加曝光度，从而吸引更多游客实现收益。这两方面因素都会对景点内部精选评论的选择和榜单景点排序的可靠性和有效性造成影响。如果平台充斥着不具参考价值评论和评分，游客根据推荐榜单排行选择景点，到达后如果发现与所看评论有所出入，容易产生对榜单和平台的不信任感，进而转向其他平台。虽然各大平台已通过各种机制筛选出部分优质评论，但难以确定部分无效游记和评分是否会对景点评分和榜单排名造成影响。本文希望对景点的评论进行文本挖掘和变量生成，筛选出与评分紧密相关且能够真实反映游客体验的指标，进而优化精选评论的筛选和榜单排行的方式。

第 2 节 数据的说明与描述

2.1 数据来源与预处理

本文使用的数据来源于携程平台“华北地区踏青景点榜单”，选择其中评论数大于 400 的 25 个景点[1]，利用网络爬虫收集游客的评论。由于携程平台的限制，每个景点最多只能浏览 3000 条评论，因此多于 3000 条评论的景区评论爬取并不完全。本文爬取的数据内容包括游客评分（5 分制）、评论正文、评论时间等关键信息。

在正式的统计建模与数据分析之前，我们需要对文本进行预处理，包括数据清洗和变量生成，这些变量的描述和简单统计性质可见 2.2 节的表 1。步骤如下：（1）在对游客评论数据的清洗中，只保留中文、选用“百度停用词表”来停用无意义的词语，再使用 jieba 分词工具进行中文分词并进行词性标注，生成评论的有效单词数量（即长度）length 变量、名词比例 noun 变量、动词比例 verb 变量、形容词比例 adjective 变量和副词比例 adverb 变量；（2）使用 TextRank 算法[2]计算出所有评论排名前 100 的主题词，再使用单个景点全部评论的词典建立稀疏矩阵并计算每个词的 TF-IDF 值，进而根据向量的余弦值来计算单条评论与主题词之间的相似度，生成文本相似度变量 similarity；（3）根据 SnowNLP[3]工具计算每条评论反映正面情绪的概率，生成变量 sentiment；（4）为了得到更细致的情感分

[1] 包括北京市颐和园、曲阜市尼山圣境、北京市玉渊潭公园、天津市盘山、石家庄市动物园、北京园博园、菏泽市曹州牡丹园等。

[2] Mihalcea, R., & Tarau, P. (2004). Textrank: bringing order into text. *Emnlp*, 404-411.

[3] https://pypi.org/project/snownlp/0.11.1/.

析，使用大连理工情感词典[4]，划分每条评论的情感词，并依据原词典的分类计算“乐”类、“怒”类等 7 类情感词的比例，并将“惊”、“乐”、“好”归为正面情感词，其余为负面情感词，生成 9 个变量依次见表 1。同时，考虑文本的否定词、程度副词及情感词，综合计算评论的情感得分，生成 emotion 变量。

2.2 变量与描述性统计

本文将 2.1 节生成变量的描述与基本统计性质列在表 1 中，其中评论基本信息、评论词性分布、评论情感词分布这 3 类变量的统计性质仅展示盘山景区的结果，因为在第 3 节进行统计建模时主要使用它们。

表 1 变量描述与基本统计量

变量类型	变量符号	变量描述	均值	标准差
评论基本信息	score	评论对景点的评分	4.591	0.713
	length	有效单词数量	18.17	20.81
	similarity	文本相似度	0.054	0.040
	sentiment	评论反映正面情绪的概率	0.793	0.319
	date / year	评论日期 / 年份		
评论词性分布	noun	名词比例	0.286	0.150
	verb	动词比例	0.208	0.145
	adjective	形容词比例	0.126	0.146
	adverb	副词比例	0.099	0.097
评论情感词分布	emotion	情感得分（0 表示中性）	3.526	4.953
	positive	情感词中正面词的比例	0.635	0.449
	surprise	情感词中“惊”类词的比例	0.001	0.024
	good	情感词中“好”类词的比例	0.518	0.444
	happy	情感词中“乐”类词的比例	0.119	0.252
	negative	情感词中负面词的比例	0.099	0.242
	anger	情感词中“怒”类词的比例	0.001	0.024
	disgust	情感词中“恶”类词的比例	0.067	0.196
	fear	情感词中“惧”类词的比例	0.010	0.076
	sadness	情感词中“哀”类词的比例	0.022	0.116
景点基本信息	level	景点质量等级（5A、4A、其他）	5A: 7	4A: 13 NA: 5
	rank	景点榜单排名（1-25）		
	comment	景点游客评论总数量	3520	6739

[4] 徐琳宏，林鸿飞，潘宇等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2):180-185.

第 3 节 数据分析与解读

考虑到不同景点的游客评论存在异质性，组间误差可能会使模型存在内生性问题。因此，3.1 节-3.3 节我们仅选取天津市盘山景区的评论进行文本挖掘，分析方法可迁移到其余景点；3.4 节将基于全榜单进行数据分析。

3.1 关键词分析

在具体的数据分析和统计建模之前，本文先对盘山景区 3000 条游客评论进行关键词分析，方法包括词云图和语义网络分析，以对游客的评论和关注内容从感性上有直观的了解。

首先，本文分别选取正面情感评论（即$emotion > 0$）和负面情感评论（即$emotion < 0$）分别 1730、237 条绘制词云图，如图 1 所示。这里我们过滤掉了一些经常出现但没有明显情感的词：景点、景区和盘山。观察可以发现，正面情感评论中出现较多的情感词包括“不错”、“有趣”、“漂亮”、“优美”等，而负面情感评论则更多反映游玩的具体内容，如“索道”、“缆车”、“门票”、“排队”、“停车场”等，可能是旅游途中较不满意的环节。

图 1 词云图（左为正面情感评论，右为负面情感评论）

此外，我们还可以对关键词进行共现分析。本文使用 TextRank 算法选取所有评论中前 50 个关键词，计算它们两两之间出现在同一条评论中的次数。图 2 绘制了它们之间共现关系的无向知识图谱，知识图谱的布局方式采用 ForceAtlas2 算法[5]，其中共有 50 个结点，904 条边，边的粗细反映共现次数，结点大小反映关键词的总权重。复杂网络的一些统计性质如下：结点平均度数为 36.12，网络直径为 2，平均路径长度为 1.296，平均聚类系数为 0.789，图密度为 0.737 等，这些统计性质反映出这个语义网络比较密集。我们进一步使用 Fast Unfolding 算法[6]进行社区发现，设置模块化解析度为 1，最终迭代得到 3 个社区，用不同颜色表示，意味着社区内部结点的联系相比社区外部的结点更加紧密。观察发现，紫色社区关键词包括“到达”、“爬山”、“山顶”、“风景”等，推断此类评论主要表达成功登顶盘山后的感受；绿色社区关键词包括“何必”、“早知”、“有趣”、“值得”等，推断此类评论主要抒发较为强烈的情感；橙色社区关键词包括“缆车”、“老人”、“孩子”、“体力”等，推断此类评论主要给出游玩的建议，如随行人员中有老人或孩子可考虑乘坐缆车。

---

[5] Mathieu, J., Tommaso, V., Sebastien, H., Mathieu, B., & Muldoon, M. R. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *Plos One, 9*(6), e98679.

[6] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory & Experiment.*

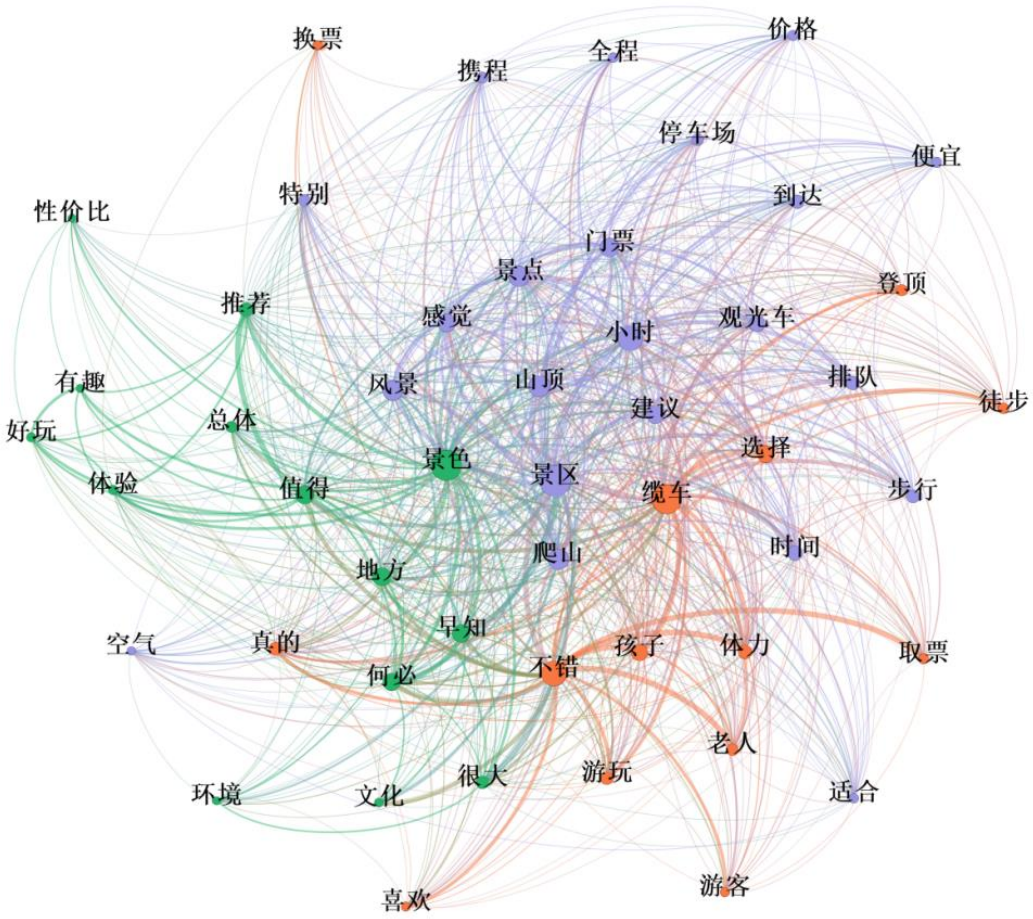


图 2 关键词共现关系知识图谱

词云图和语义网络分析直观地展现了游客的关注点，作为本文后续开展统计分析的铺垫。

3.2 探索性数据分析

此小节我们对利用盘山景区 3000 条评论转化来的文本变量进行探索性数据分析。图 3 和图 4 对除了评论情感得分 emotion 外其余变量进行了可视化分析，由于 2016 年至 2020 年包括了 80%以上的评论，我们仅对这部分评论进行趋势分析。

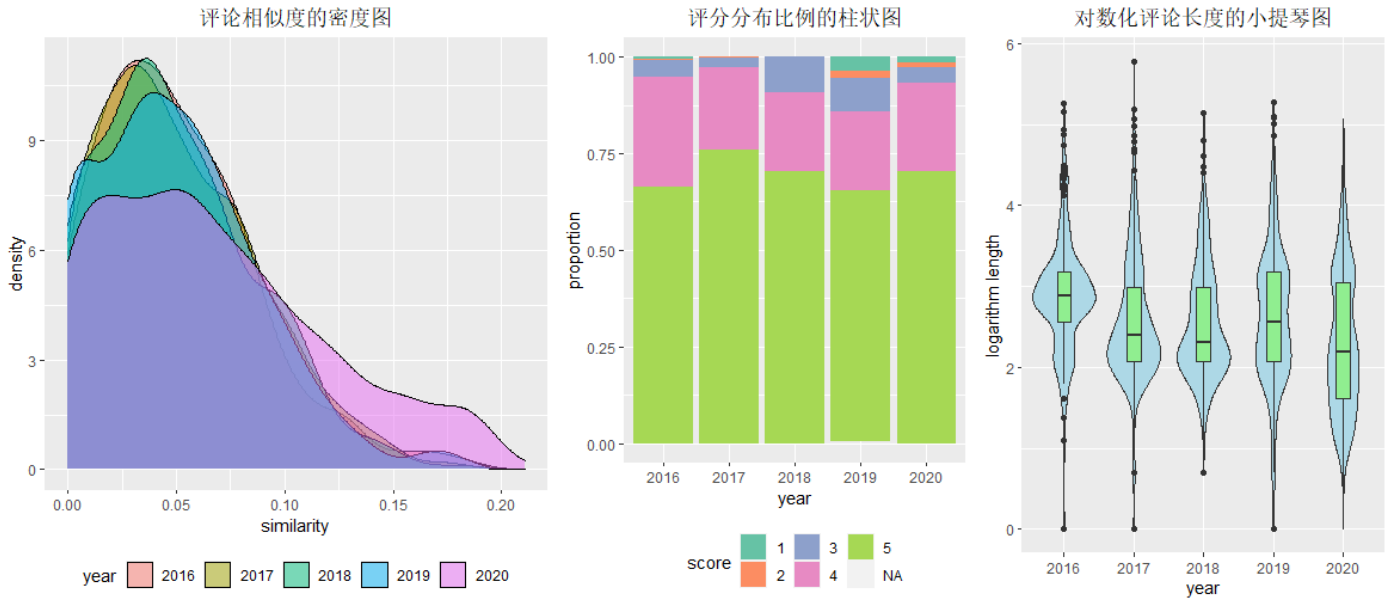


图 3 关键变量的可视化分析

观察图 3 可以发现，文本相似度的分布在 2020 年显著区别于前 4 年，表现出在更大的方向有更多的分布；评分的分布在 5 年内变化不大，但 2019 年、2020 年的“差评”要明显多于此前；评论长度呈现出明显的重尾特点，我们先取对数化处理，再观察小提琴图发现 2016 年的评论长度整体较长且集中，2020 年评论长度的分布更加广泛，表现出“瘦高”的小提琴特征。而图 4 的多个变量柱状图显示名词和动词是评论中的主要词性，且正面情感词显著多于负面情感词，其中“好”类情感词占比最大，且评论反映正面情绪的概率（sentiment）有先下降后上升的特点，这与本文后面对 emotion 的序列分析相吻合。

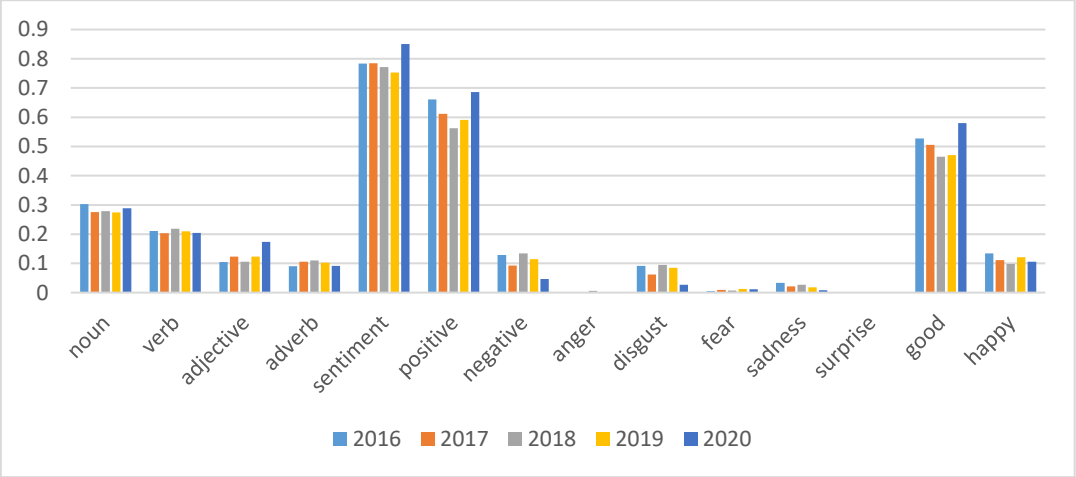


图 4 其他变量随年份变化图

下面重点对情感得分进行时序分析。考虑到 16 年及之前的评论较少，本文取 2017 年 1 月至今、以月为单位的平均情感得分建立长度为 52 的时间序列。图 5 左面板为序列的自相关函数，可以看出序列最多 7 阶的自相关系数显著。在此基础上，为了得到更加平滑的趋势，我们以 7 个区间进行简单滑动平均，结果如图 5 右面板所示，可以发现在 2018 年上半年和 2019 年初的评论情感最为消极。本文选择了其中部分评论进行展示，如“感觉不太好，没什么意思，就是爬山”（2018-03-17）、“饭店服务员态度太差了，也没有早饭！”（2018-04-19）、“去的季节不对，夏末季节去有树有花才好看”（2019-02-10）等。同时可以发现，从 2019 年 9 月开始，评论的情感得分有显著的上升趋势，游客对景点的满意程度有明显的提高。此外本文还对序列进行了季节分解，发现 6 月情感得分的季节性最高，12 月的季节性最低，这反映出淡季（冬天）与旺季（夏天）在盘山景区的旅游体验有明显的差距。

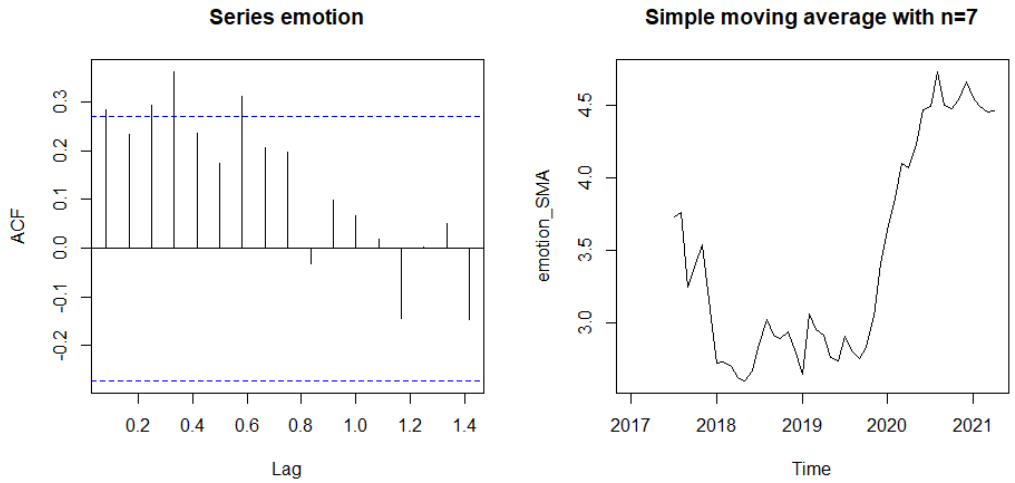


图 5 序列相关与滑动平均

3.3 统计建模与变量重要性

我们的数据产品希望能挖掘出从文本出发的指标来反映对景点的评分，这些变量相比评分更加真实，同时能表现出游客对景点的评价。因此，本文考虑一个二分类模型，将对景点的评分≥4 分的评论作为正类（记为 1，共 2782 条），对景点的评分≤3 分的评论作为负类（记为 0，共 218 条），分别表示“好评”与“中评及差评”。首先，我们建立 Logistic 回归模型来判断哪些变量对两类评论的识别显著相关，而考虑到可能有多重共线性的影响，本文使用双向逐步选择法来对变量进行筛选，分别记两个模型为“原始模型”和“双向选择模型”。这里为了反映出不同变量的作用大小，我们先对所有变量进行归一化到 0 至 1 的范围。回归的系数估计值和显著性见图 6 所示。

	*** noun	verb	adjective	*** adverb	*** similarity	*** sentiment	*** emotion	*** length	positive	negative	anger	. disgust	fear	** sadness	surprise	good	happy
原始模型	-3.16	-1.14	-0.1	-3.06	2.04	1.08	1.44	-4.1	-0.07	2.86	-4.63	-3.27	-2.62	-3.71	0.07	0.07	0.07
双向选择模型	-3.2	-1.12	0	-3.1	2.03	1.12	1.53	-3.95	0	0	0	-0.5	0	-0.95	0	0	0

图 6 Logistic 回归系数估计值和显著性

我们发现，对识别“好评”有显著正向作用的有文本相似度、正面情绪概率和情感得分，而对识别“中评及差评”有显著正向作用的有名词比例、副词比例、评论长度和“哀”类情感词比例。具体而言，文本相似度越高，越有可能是好评，这一定程度上验证了存在景区刷好评的可能；长度越长、副词越多、名词越多，越可能是差评，表明游客在表达不满时，会更全面地描述心情与经历，且聚焦于内容（名词）和增加情感程度（副词）。

使用 Logistic 回归得到的结果并不完美。其一，线性模型比较简单，而在实际中变量之间的交叉效应更有意义，应考虑变量组合；其二，我们需要确定变量的相对重要性，以在数据产品中给每个变量不同的作用权重。

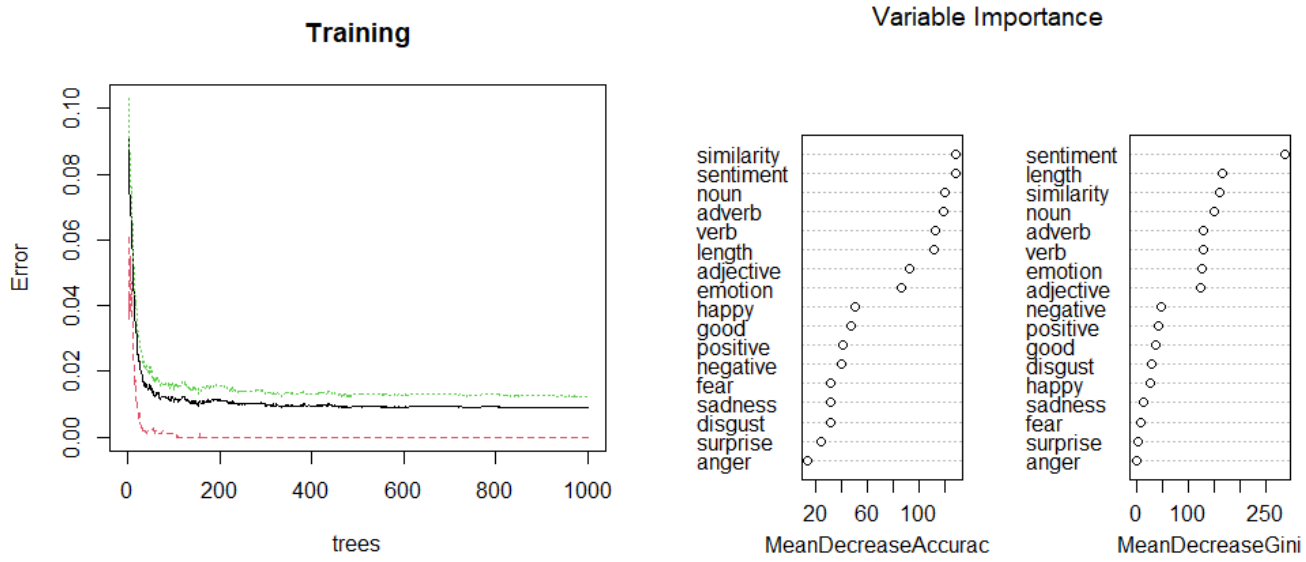


图 7 随机森林模型的训练误差收敛曲线与变量相对重要性

综合以上分析，本文选择随机森林算法[7]来进一步建模，主要利用 bagging 的方法计算变量的相对重要性。在实际建模中，我们取 80%的数据进行训练，20%进行测试，考虑到负类样本占比非常小，我们对训练集中的负样本进行数据增强（重复为原来的 5 倍）。我们设置子树的个数为 1000，训练的误差收敛曲线如图 7 的左面板，同时在测试集中得到 2%左右的误差，表明模型是可靠的，分类效果较好。此时我们观察图 7 最右面板的变量相对重要性排序，以 Gini 指数作为标准，对分类效果最重要的几个变量为正面情绪概率、长度、文本相似度、名词比例、副词比例、情感得分等。

3.4 基于全榜单的变量相关分析

上文选择单个景点的评论建模是为了避免基于景点的选择性偏差，但我们的数据产品也定位于为榜单准入和排序进行服务，因此需要对全榜单景点在每个变量上的平均值进行相关分析。由于这里的分析粒度提升到景点，我们纳入景点级别（level）、评论数量（comment，反映景点热度）等变量。

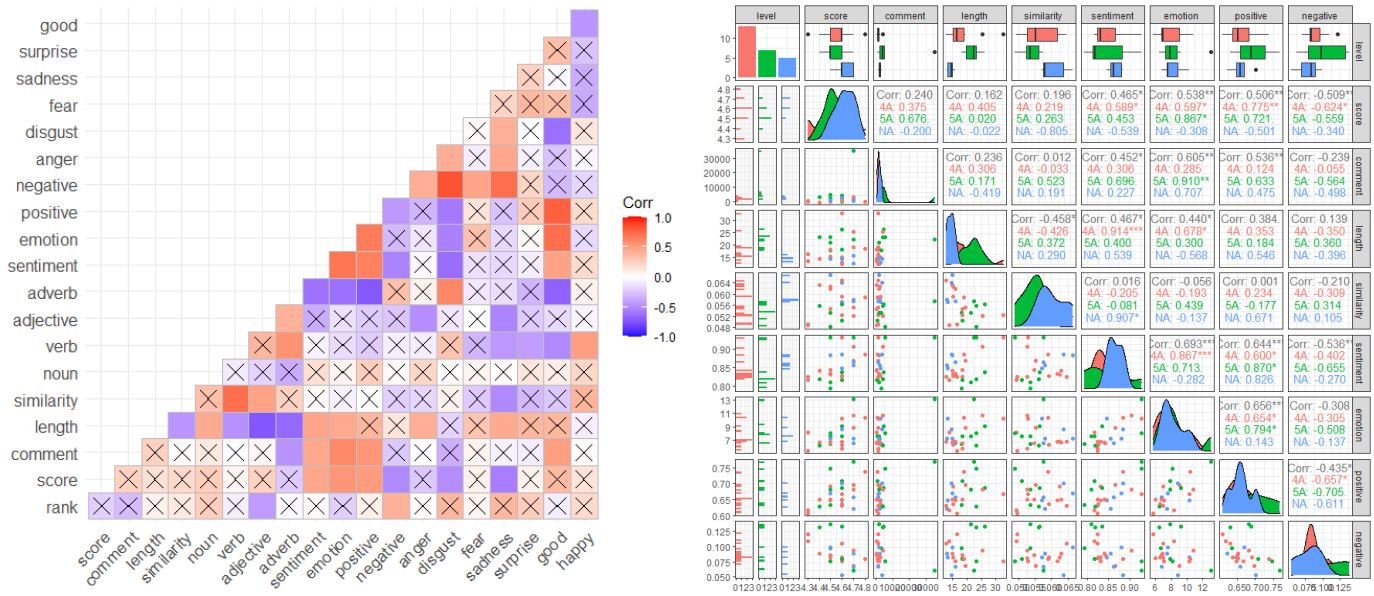


图 8 景点变量均值的相关系数热力图与相关关系矩阵图

首先，我们对一些数值型变量计算相关系数、进行相关性检验并绘制热力图，如图 8 左面板，其中×表示相关性检验不显著。观察发现，文本相似度与动词比例、形容词比例显著正相关，但与长度、“哀”类词比例显著负相关，表明高相似度的评论中多用动词和形容词；评论长度与“怒”类词比例、“哀”类词比例显著正相关，但与副词比例、形容词比例等显著负相关，表明较长的评论更有可能抒发对旅游体验的不满；评论数量与情感得分显著正相关，与副词比例显著负相关，这说明景点的热度确实与旅游的良好体验相匹配；其余变量的分析暂无赘述。

最后，本文将景点的级别纳入考虑（5A、4A 或其他）。我们画出部分变量的相关关系矩阵如图 8 右面板所示，其中绿色代表 5A 级景区，红色代表 4A 级景区，蓝色代表其他景区。观察矩阵对角线的密度图、下三角的散点图和上三角的相关性检验可以发现，非 5A、4A 级景区的文本相似度整体最高；5A 级景区的评论长度整体最长；仅 5A 级景区的情感得分与热度显著正相关；仅 4A 级景区的正面情绪概率与评论长度显著正相关等。当然，受限于样本量较小，这些结果不一定可靠，仅供参考。如需投入商业应用，可考虑进一步增大景点的数量来分析。

[7] Breiman, L. (2001), Random Forests. *Machine Learning 45*(1), 5-32.

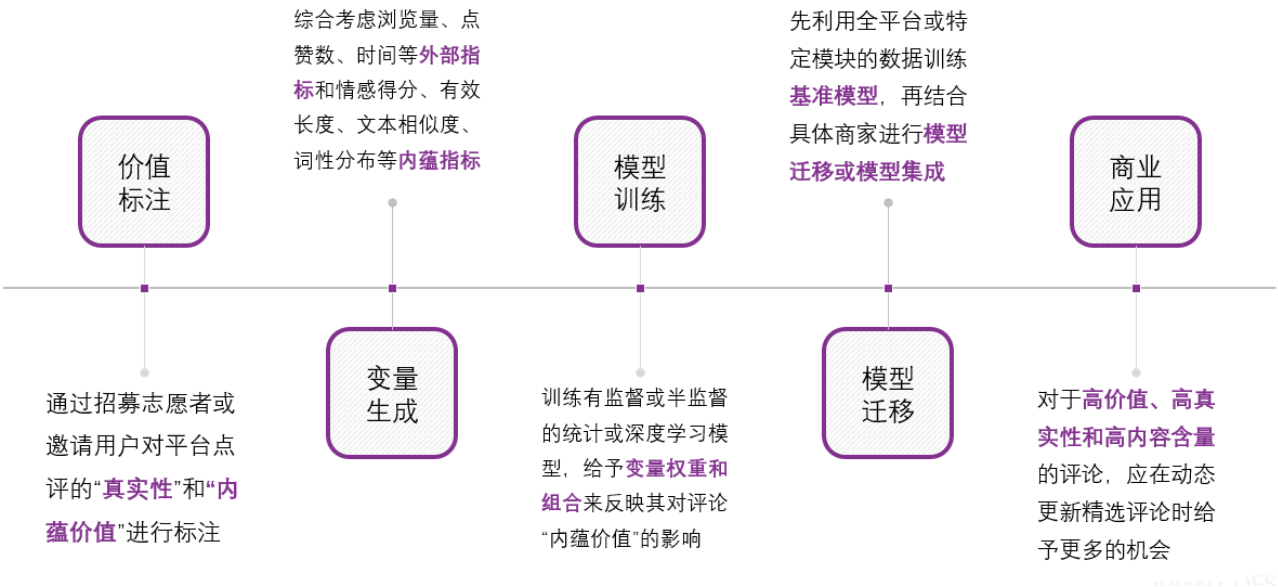
第 4 节 商业应用与总结

4.1 现实意义与实际应用

现有电商平台在提供信息方面极大地便利了用户生活，但同时应注意到，在信息量冗杂特别是存在大量无效信息的情况下，用户可能会被平台所提供的信息误导。以餐饮行业为例，用户通过平台进行消费后，商家往往通过发放现金红包（多见于外卖行业）或小礼物的方式引导顾客给出好评，即使实际上用户并没有达到相应的满意程度，仍会为了得到现金红包给出比实际评价更高的评分。如果新用户因大量不准确的评论产生了不愉快的体验，可能降低对平台的信任感，造成平台用户流失。点评内容与实际价值的一致性日益重要，一方面，许多电商平台将海量的用户点评作为核心竞争力之一；另一方面，用户在进行消费前习惯于参考其他用户的反馈已成为常态。

4.2 数据产品设计

本文建立的统计模型具有很好的迁移性，适用于包含用户点评的绝大部分场景。我们希望实现这样的数据产品：在景区或商家内部筛选优质评论时，除了考虑浏览量、点赞数、评论时间、评分等外部指标，综合考虑情感得分、有效长度、文本相似度、词性分布等内蕴指标，通过人工标注来训练有监督或半监督的统计模型，给予变量不同的权重来反映对评论“内蕴价值”的影响。在数据量较大的情况下，可以使用深度学习模型以得到更加复杂而准确的权重分配和变量组合。考虑到不同商家的异质性，可先利用全平台或特定模块的数据训练基准模型，再结合具体商家进行模型迁移或模型集成。对于高价值、高真实性和高内容含量的评论，应在动态更新精选评论时给予更多的机会。同时，在进行榜单准入和排序时，也可以借助相应的评论指标，使榜单的推荐更加合理。



我们设计的数据产品可以给用户带来更加良好而真实的体验，在服务用户和服务商家之间找到最佳的平衡点，实现用户和商家整体的福利最大化，以智慧科技助力智慧生活。

4.3 模型改进方向

本文设计的数据产品有一定的局限性。首先，本文在数据分析与统计建模部分集中于一个特定的景点，对不同景点的联系讨论不充分，未来可研究不同景点或商家的异质性程度和关键变量。其次，情感得分生成方式并不一定普适，未来可以根据特定行业训练词语对应的情感得分。最后，内蕴指标还应提供更多的选择，充分挖掘文本信息。