问题选择	2020 MCM/ICM公司 摘要表	团队控制编号
C		2003717

占领市场：全面探索亚马逊评论和评级

最近，Sunshine公司计划在在线市场上推出三种新产品，并且我们的团队需要对给定的数据提供一些见解，并提供战略建议以改善产品的未来销售和声誉。具体任务分为两个大问题。

对于问题1，我们首先通过删除不必要的信息，对所有文本进行标记以及来应用数据清理技术对所有词进行词法去除和词干处理。然后，我们应用LDA主题模型对评论所关注的内容进行直观描述。接下来，我们通过时间和交叉分析可视化评论数量，评论长度，星级和有用投票之间的关系。结果表明，具有较高帮助等级的评论往往会伴随着较高的等级和较长的评论时间。此外，在早期阶段，每个评论的平均有帮助票数远远超过后期。同时，收视率，评论时长和其他指标也有很大波动。这些都表明在冷启动阶段保持良好品牌形象的重要性。

关于问题2(a)，我们提出了三个Sunshine公司要重点关注的指标：1.加权评级比率，它表示每个评级通过有用投票数加权产生的比率；2.评论的加权情感评分，其中我们应用逻辑回归来根据帮助程度计算每个概念词的评分及其加权总和；3.偏好向量，我们根据LDA的结果为每个产品分类七个属性，建立包含每个属性相关术语的字典，并基于随时间衰减的加权词频统计估计这些属性上人们的偏好比率。结果，Sunshine公司可以分配不同的努力来改善不同的产品功能。

在问题2(b)中，我们认为产品的声誉与平均星级，评论权威性和销量有关，其中我们假设销售量与固定时间窗口内的评论次数成正比。因此，在这一固定时间窗口内的声誉可以看作是这一时期有关比率和评论的特征的共同贡献，并且经过计算，结果表明，对于吹风机和安抚奶嘴，它们的声誉得分在早期会增加，并且倾向于以后会保持稳定，而微波的微波会一直保持增长。

关于2(c)，我们继续使用2（b）中随时间变化的信誉作为评分和评论的综合指标，并应用嵌套的两层LSTM模型来预测其对评论序列的价值，考虑到这个指数考虑到几乎每一个信息量给定的特点（销售额包括在内）。

关于2(d)，我们考虑了评论的连锁反应。在分析了不同等级的每月平均数量随时间变化的趋势之后，我们得出结论，等级为5的评论倾向于引发更多的评论。除此之外，我们出乎意料地观察到1级的评论数量与评论的长度显著相关，并且在Granger因果检验之后，我们发现，在大量低星级评级后，短评论往往会在两个月内滞后。

关于2(e)，我们专门为情感词整理字典，并为每个情感词分配分数。然后，我们计算所有评论的新情感得分，并将评论分为五个等级。在比较星级和评论等级的混淆矩阵后，我们发现映射不对称性是：一些具有强烈情感词的评论具有中等评分，而具有较弱情感词的评论具有极高评分。我们通过可视化来解释这一现象，对于这些评论，更大的机会是积极的词汇和消极的情感词汇出现，或者强烈的情感词汇被描述产品属性的词汇取代。

最后同样重要的一点是，我们总结了解决方案的优缺点，并向Sunshine公司的市场总监介绍了自己的见解，目的是帮助Sunshine公司在在线市场上取得领先。

关键字： 逻辑回归LSTM； LDA主题模型； Granger原因测试；情绪分析。