目 录

[一、绪论 1](#_Toc40815987)

[（一）研究背景及意义 1](#_Toc40815988)

[（二）国内外研究综述 1](#_Toc40815989)

[（三）本文工作 2](#_Toc40815990)

[（四）本文创新点 3](#_Toc40815991)

[二、情感分析基础理论 4](#_Toc40815992)

[（一）词向量 4](#_Toc40815993)

[（二）机器学习 5](#_Toc40815994)

[（三）依存句法分析 6](#_Toc40815995)

[三、数据来源及情感词典的构建 7](#_Toc40815996)

[（一）数据来源 7](#_Toc40815997)

[（二）情感词典的构建 8](#_Toc40815998)

[四、细粒度情感分析 10](#_Toc40815999)

[（一）属性词提取 10](#_Toc40816000)

[（二）属性词和情感元组抽取 12](#_Toc40816001)

[（三）情感计算 16](#_Toc40816002)

[（四）实验及结果分析 17](#_Toc40816003)

[五、用户类型与满意度的相关分析 20](#_Toc40816004)

[（一）情感标签处理 20](#_Toc40816005)

[（二）机器学习流程 20](#_Toc40816006)

[（三）实验及结果分析 21](#_Toc40816007)

[六、总结 24](#_Toc40816008)

[（一）全文回顾 24](#_Toc40816009)

[（二）不足和未来展望 24](#_Toc40816010)

[参考文献 25](#_Toc40816011)

基于酒店评论的细粒度情感分析研究

# 一、绪论

（一）研究背景及意义

随着电子商务的快速发展，越来越多的人进行网上购物，并且大多用户愿意在网上留下评论，分享自己的使用体验和对商品的看法，留下的各式各样的评论文本已成为消费者在选择购买产品或服务时的重要参考依据，同时语料内容涵盖了大量的管理和决策支持信息。一方面，从买方来讲，在线评论的情感极性反映了使用者对商品的满意度，用户更愿意相信客观的使用评价，而不是商家的宣传资料，以此来判断商品的质量和商家的可信度，从而决定是否购买；另一方面，站在卖方角度，能够获取用户的反馈信息，发现自身和竞争对手的差距，这有助于商家及时做出决策，弥补自身短板，提高竞争力。

在酒店领域，携程、美团、去哪儿等应用将酒店搬到了互联网，提供了更加方便快捷的服务，人们习惯于在网络上预订酒店并留下了具有很高参考价值的评论语料。本文研究的则是通过情感分析来挖掘用户的需求，同时为卖家分析出酒店的不足之处和需要改进的地方，从而进行决策优化，以便进一步提高其产品或服务的质量。目前，已经有部分研究通过文本挖掘、情感分析的方法挖掘酒店评论数据，取得了不错的成果，但已有工作粒度较粗，只假设一段评论含有一种感情倾向。而大量在线评论中，往往存在着对多个方面表达多样情感的现象。如何有效建模、分析出用户对产品或服务属性方面细粒度需求的情感偏向具有重要意义、也具有很大的挑战。

（二）国内外研究综述

1．粗粒度情感分析

近多年情感分析一直都是文本挖掘领域的热点研究问题之一。情感分析的目标就是对带有情感色彩的文本进行挖掘和分析，辨别其情感极性。早期情感分析的研究大都是在词语层面的倾向性分析，比如情感词的辨别和情感词极性的分类。之后，基于篇章、段落、句子的情感分析技术也逐渐多了起来，这些研究也叫做粗粒度情感分析研究。

可以把粗粒度情感分析看坐一个典型的分类问题。分类问题有正、负二分类，也有多分类，包括正面、负面、中性等。解决途径可以概括为两种：情感词典的方法、机器学习的方法。分类方法包括监督学习、无监督学习以及半监督学习。其中，机器学习方法大致过程为语料预处理、特征提取、特征选择、分类模型处理；情感词典方法是指利用词语的感情色彩来分析句子的情感极性，首先找到句中的正负情感词，然后结合极性词语的个数和程度副词的程度值，通过计算得到句子的情感极性。

2002年，Turney[18]提出了分析评价词情感倾向的方法，首先判断文中程度副词和修饰词的极性，然后根据加权求和的最终结果来判断句子的情感倾向；Pang[19]等人首次使用有监督的学习方法将电影评论分为正、负两类，分别使用了贝叶斯、最大熵与支持向量机作为分类器，结果显示支持向量机的分类效果更好。

2．细粒度情感分析

然而，粗粒度情感分析只是在句子层面的粗略分析，并未考虑情感词所对应的具体方面，用户无法了解商品在各个方面的具体表现。细粒度的情感分析能够抽取与识别粒度上更细的情感信息。细粒度情感分析包含了两方面的任务：命名实体识别和文本分类。

Kobayashi[17]等人提出了一个情感倾向的表达方法，即用五元组来表示各个情感表达，包括意见发起者、评价对象、评价中心词、评价强度、极性，根据五元组就可以计算出每个评价对象的情感倾向；刘丽[7]针对传统粗粒度情感分析忽略具体评价对象，提出结合条件随机场和语法树剪枝的方法对产品评论进行细粒度情感分析；张玥[14]提出了面向产品评价文本的细粒度情感标注体系；黄胜等[9]提出了一种基于双层CRFs模型的细粒度意见挖掘，构建了<意见陈述，主题词，意见词，情感>四元组；王山雨[16]研究产品评论语料中的产品属性抽取方法，设计了一个功能包括评价对象抽取、极性词识别、评价对象极性判断等的细粒度情感分析系统；蓝天广[6]提出了显式属性和隐式属性分开识别的方法；李长江[5]采用了依存句法分析进行细粒度情感分析，并做了酒店检索系统。

（三）本文工作

本文研究了中文情感分析在酒店评论领域的应用，包含了细粒度和粗粒度两方面的情感分析，主要任务有：

1. 构建了酒店领域的属性词典和情感词典。基于词向量的方法抽取属性词，属性词典包含了价格、环境、服务等用户关心的点。情感词典则在常用的中文情感词典基础上，用词向量的方法抽取酒店评论中的情感词，添加到情感词典中。

2. 使用依存句法分析得到句子的语法结构，并结合属性词典、情感词典抽取出情感词、属性词、修饰词组成三元组，进一步来计算酒店用户在各个属性上的情感值，最后计算用户在各方面的满意度。

3. 使用传统方法对整条评论进行粗粒度情感分析。使用多种机器学习的方法进行有监督学习，并对比各方法的准确率。最后使用准确率高的分类器算出酒店的好评率。

4. 以具体酒店为例，从不同出游类型的角度出发，结合两种情感分析得到的满意度和好评率，进一步挖掘不同用户对各方面的需求差别，指出酒店的不足之处，并提出相关建议。

（四）本文创新点

1．采用word2vec词向量方法抽取属性词，简单快捷的完成特征抽取，构建了较为详细的属性词典，为细粒度情感分析打下基础。

2．采用句法依存分析，并结合情感词典，取得了较好的情感元组抽取效果，通过对<属性词，情感词，修饰词>三元组的分析，准确的计算出了酒店在各个属性上的情感倾向。

3．从粗粒度和细粒度两方面对酒店领域的评论文本进行情感分析，根据计算出的情感指标进一步得到酒店的好评率以及用户对酒店各方面的满意度。

# 二、情感分析基础理论

在酒店领域评论的情感分析中，文章使用有监督的学习算法对评价语料进行粗粒度的情感分类，包括了评价语料的预处理、特征选择、构建分类模型等步骤。在细粒度情感分析中，包括了属性词抽取、依存句法分析的知识。本章节将对文章采用的文本情感分析技术基础理论知识进行简单的介绍。

（一）词向量

词向量技术将自然语言中的词语转化为稠密的向量，相似的词会有相似的向量表示，这样的转化方便挖掘文字中词语和句子之间的特征。生成词向量的方法从一开始基于统计学的方法（共现矩阵、奇异值分解）到基于不同结构的神经网络的语言模型方法。到现在为止已经有很多成熟的词向量模型，这里介绍本文用到的两种词向量模型：word2vec和TF-IDF。

1．word2vec

2013年，Google团队发表了word2vec工具。word2vec[4]的本质是一个神经网络语言模型，它有两个模型：跳字模型（skip-gram）和连续词袋模型（CBOW）。Skip-gram模型是在已知当前词的情况下，预测其上下文；CBOW模型是是在已知当前词的上下文的情况下，预测当前词。此外，word2vec词向量可以较好地表达不同词之间的相似和类比关系，本文借助word2vec模型来寻找语料中与种子词语义相近的词语，再加以人工筛选，即可得到较完善的属性词典。

2．TF-IDF

TF-IDF是一个权重，能够体现一个词的重要程度，一个词在文档中出现的频率越高，其重要程度越高，但在语料库中出现的频率越高，重要程度越低。例如，像“的”、“了”、“是”这种词，虽然在文本中出现的频率高，但不能说它们很重要，相反的它们的重要性很低，这是因为这些词在每个文档都出现了，就变得不重要了。

词频 (TF) 指的是统计的语料中词语出现的次数与所有词语的总数的比值。但比值要进行归一化处理 (一般是词频除以文章总词数), 来避免它偏向长的文件，其计算公式为：

|  |  |  |
| --- | --- | --- |
|  |  | （1）式 |

逆文件频率 (IDF) 的核心思想是：如果含有词条w的文档越少, IDF越大，则表示词条具有很好的文档辨识能力。对特定词语的IDF计算方法为所有文档的条数与有某词的文档条数的比值的对数，在实际操作中，我们会加平滑因子，防止统计数为0这种情况出现：

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | （2）式 |

如果一个词语在语料库中有较高的频率，同时具有低的的逆文档频率，则能得到高权重的TF-IDF。所以，TF-IDF可以筛选出重要的词语：

|  |  |  |
| --- | --- | --- |
|  |  | （3）式 |

在训练文本较多的情况下，TF-IDF特征量化的效果更好，相比于word2vec更有优势。因此本文在机器学习中运用TF-IDF进行特征选择，将得到的权重矩阵输入分类器中进行训练。

（二）机器学习

机器学习中有很多分类算法，可以运用在情感分析中。常用的算法有：朴素贝叶斯分类法、支持向量机、k近邻、神经网络算法等。下面主要介绍本文运用的朴素贝叶斯和支持向量机。

1. 朴素贝叶斯

它的思想是：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。它的基础是贝叶斯理论，公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （4）式 |

表示在事件B发生的条件下事件A发生的概率。

贝叶斯分类器的流程为：

1. 设x={a1,a2,a3,…,am}为一个待分类项，而每个a为x的一个特征属性。x是样本，a是样本中的各个属性；
2. 有类别集合C={y1,y2,…,yn}，C是标签;
3. 计算P(y1|x),P(y2|x),…,P(yn|x)。计算在各个样本属性下，某个类别的概率;
4. 如果P(yk|x)=Max{ P(y1|x),P(y2|x),…,P(yn|x)}，则x∈yk。取概率最大的作为最终的分类结果

那么现在的关键就是如何计算第3步中的各个条件概率。根据贝叶斯理论有如下推导：

|  |  |  |
| --- | --- | --- |
|  |  | （5）式 |
|  |  | （6）式 |

1. 支持向量机

它常用来解决分类和回归问题，其学习目标便是寻找最大间隔超平面，将文本划分开，进而达到分类的目的。以下图为例，平面上有两种类型的点，分别为圆圈和叉。现在想找到分类依据将这两类分开。可以发现两者有明显的界限，之间有无数条直线可以将它们分开，这些直线就是超平面。

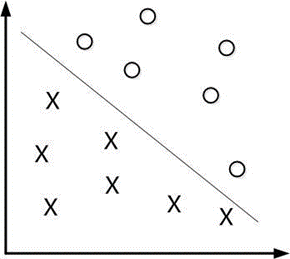


图1 超平面

这样的超平面有无数个，支持向量机的目的就是找到最优的超平面，这个超平面应满足离两边数据的间隔最大。根据此要求，可以得出求解最优超平面的目标函数为：

|  |  |  |
| --- | --- | --- |
|  |  | （7）式 |

而直接求解上面的方程比较困难，而上述问题中，给定了一些约束条件，因此可以使用拉格朗日乘子法，最后优化目标可以写为：

|  |  |  |
| --- | --- | --- |
|  |  | （8）式 |
|  |  | （9）式 |

（三）依存句法分析

在评论中，存在情感对象和情感词距离很远，一个情感词作用于多个情感对象的情况。如果仅通过分词和词性标注来获得属性词和情感词，往往存在着不足。例如评论“服务、交通、卫生都很差”，用常规的抽取方法只能得到<卫生，差>，而忽略了“差”同样修饰“服务”和“交通”。而通过句法依存分析处理，可以看到句子中各词语的依赖关系，判断出“服务”、“交通”、“卫生”属于并列关系，可以完整的抽取出<服务，差>、<交通，差>、<卫生，差>三个元组。

对于中文的句法分析，需要先对文本进行分词和词性标注，然后分析词语间的关系，得到各个词之间的依赖关系，从而获得句法结构。本文采用依存句法分析得到句子结构，在此基础上完成元组的抽取工作。

# 三、数据来源及情感词典的构建

（一）数据来源

本文的研究数据来自携程网上的酒店评论，爬取了大连星海广场周边多家酒店的评论文本，共计57734条数据。爬取了每个用户的昵称、评论、打分、出游类型、评论时间等，存放于JSON文件中。部分评论数据如下表所示：

表1 评论数据表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 用户昵称 | 出游类型 | 评论时间 | 打分 | 评论内容 |
| 哆啦有个梦^\_^ | 独自旅行 | 2019-10-17 | 5.0 | 离星海广场特别近，交通方便，下楼就有便利店 夜景也挺好看的 总之好评 |
| 妙禾小友 | 情侣出游 | 2019-08-22 | 4.3 | 房间与照片一样，比如家汉庭快捷酒店大很多，有厨房有电磁炉但没有锅碗刀叉做不了饭。 |
| 妍小胖 | 家庭亲子 | 2020-01-05 | 5.0 | 酒店公寓的地理位置交通很方便，房间宽敞明亮整洁，酒店公寓的布置非常温暖舒适，价格也非常实惠，是一次很满意的住宿。以后还会再来入住. |
| 135079\*\*\*\* | 商务出差 | 2019-12-02 | 3.8 | 这个酒店式公寓窗户能看到星海广场，步行十分钟就能到喷泉那里，很方便。每天有阿姨过来打扫卫生，基本上就是倒倒垃圾，床品是不换的。因为没有阳台，所以洗衣服没地方晾。除了位置好，别的一般，三居价格适中。 |
| 雷仔儿 | 商务出差 | 2019-11-22 | 3.2 | 住过最差没有之一的酒店，首先床头柜和桌子上的插座没一个好的，手机都没地方充电。洗澡的那个花洒一天到晚有滴水声，感觉拍鬼片一样，洗澡的时候居然有墙皮往下掉！！！！估计连锁就是这样被搞垮的 |
| 哎呀荷 | 情侣出游 | 2018-07-08 | 3.8 | 床头掉皮的特别严重，被子上、枕头上，都是掉下来的屑，感觉很不卫生啊！用成这样应该考虑换一个新的了。灯开关设计的也不是很好，只有厕所灯和走廊灯可以在一进门的位置控制，卧室灯分在床两侧控制，很不方便。 |
| 孤剑小楼听雨轩 | 独自旅行 | 2019-10-17 | 4.5 | 酒店的地段非常好，走几步就是西安路商圈，逛街啊，交通啊都挺方便的。不过呢就是那个床头真的是有点太旧了。 |
| 马大哈的马 | 朋友出游 | 2019-10-06 | 5.0 | 订的是两个卧室的，五个朋友一起，床很大，三个人一个床也够睡，因为就住一晚上，也没有太大要求，干净就好，楼下就有罗森，买东西很方便 |

爬取过程中，由于携程网的反爬虫机制，遇到了较大的困难。电脑端的网址反爬虫机制十分完善，研究对比发现，手机端的网页比电脑端的爬取方便一点，于是爬取手机端的网页链接。携程网采用了Ajax动态请求加载，利用谷歌浏览器中的开发者工具找到含有用户评论的请求文件，找到url，headers，以及 form data这些参数。为防止IP被封，使用了代理IP，最终爬取成功。

（二）情感词典的构建

用户对商品进行评价时，往往会从多个角度出发，评价商品的各个属性，在酒店领域也是一样。例如用户会评论酒店的位置是否便利，环境是否卫生干净，或者设施是否完善等。用户在选择酒店时也会想了解酒店各个方面的表现情况，如果自身对卫生要求较高，就会选择卫生评价高的酒店。在对酒店评论进行细粒度情感分析时，要结合酒店领域特色构建专属的属性词典和情感词典。情感词典不只是指有情感色彩的词语，如动词、名词和形容词等，也需要有对情感词的修饰词如否定副词、程度副词等。特征词典则涵盖了酒店的属性词。

1. 情感词典

在文本挖掘领域中，有研究人员整理好的通用的情感词典，并且标注了情感词的词性。常用的中文情感词典有台湾大学总结的情感词典、董振东提供的知网情感词典等。本文结合了这两个情感词典作为基础情感词典。

同时，考虑到酒店领域有着特有的情感词汇，如：便利、温馨、豪华、昂贵等，利用训练好的word2vec将词转化为词向量，寻找与种子词语义相近的情感词，具体的做法会在下一章进行。去重后，将新的情感词添加到基础情感词典中，构建酒店领域的情感词典。

表2 种子词表

|  |  |
| --- | --- |
| 情感 | 种子词 |
| 积极 | 周到、整洁、实惠、干净、近、优雅、安静、便捷、细致、划算、有礼貌 |
| 消极 | 偏僻、凌乱、不周、陈旧、简陋、糟糕、臭、恶劣 |

2. 程度副词词典

本文使用了知网情感词典中的程度副词词典，词典中包含了220个程度副词，程度词既有加强作用也有减弱作用，通过赋予他们不同的权重，细化的计算情感值。部分的程度副词如下表所示：

表3 程度副词表

|  |  |
| --- | --- |
| 程度词 | 程度权重 |
| 倍加、非常、极、十分、更格外 | 1.5 |
| 略微、稍微、未免、有点、轻微 | 0.5 |

3. 否定词词典

否定词的个数较少，下表为部分否定词：

表4 否定词表

|  |  |
| --- | --- |
| 否定词 | 权重 |
| 不、没、无、非、莫、弗、毋、勿、未、否、不曾、未必、没有、不要、难以、未曾 | -1 |

# 四、细粒度情感分析

现实生活中，用户对商品进行评价时会从多个角度出发，以酒店为例，用户进行评论时，会从价格、环境卫生、设施、服务等方面进行评价。例如这个评论文本：“酒店位置很好，服务人员也非常热情，就是价格有点贵。”很显然，我们不能直接判断用户对这个酒店的整体情感倾向，因为用户对酒店的位置、服务质量、价格三个方面分别给出了评价，并且情感倾向有好有坏。因此，想要准确的把握用户的情感倾向就需要进行细粒度的情感分析。

细粒度情感分析可以分成两个任务：情感对象的识别和情感倾向分析。前者可以看成是属性词提取（命名实体识别），后者则是分类任务。本文使用word2vec进行属性词提取，使用依存句法分析并结合构建好的酒店情感词典进行情感倾向分类。

（一）属性词提取

本文根据酒店用户常见的关注点，将情感对象划分为六类，分别是环境、服务、设施、价格、交通、整体。首先将所有语料进行分词，去停用词后进行词频统计，找出频率较高的属于情感对象的种子词，并归类，结果如下表所示：

表5 种子词表

|  |  |  |
| --- | --- | --- |
| 对应属性类型 | 属性词 | 词频 |
| 环境 | 环境 | 8608 |
| 卫生 | 5766 |
| 服务 | 服务 | 12234 |
| 前台 | 6749 |
| 态度 | 1288 |
| 设施 | 设施 | 4218 |
| 装修 | 1188 |
| 价格 | 性价比 | 2674 |
| 价格 | 1631 |
| 交通 | 位置 | 7705 |
| 交通 | 4735 |
| 出行 | 1821 |
| 整体 | 房间 | 18090 |
| 酒店 | 15558 |
| 总体 | 905 |

之后使用训练好的word2vec词向量模型，找到上表种子词的近义词，以“位置”一词为例，找到与其语义最相近的10个词，如下所示：

表6 语义相似词表

|  |  |
| --- | --- |
| 词语 | 语义相似度 |
| 地理位置 | 0.897 |
| 地点 | 0.823 |
| 周边环境 | 0.788 |
| 地段 | 0.773 |
| 周围环境 | 0.755 |
| 环境 | 0.701 |
| 条件 | 0.617 |
| 风景 | 0.615 |
| 景色 | 0.601 |
| 服务态度 | 0.546 |

可以发现，识别效果不错，‘地点’、‘地段’、‘地理位置’等可以划为‘交通’一类的属性词，但也出现了‘环境’、‘服务态度’等不合适的词汇，因此需要进行人工筛选。之后像这样反复几次人工筛选、再迭代之后，可以较全面的找出所有语料中与‘交通’相关的属性词。其他属性类别也按此方法操作，最终得到如下属性词分类表：

表7 属性词表

|  |  |
| --- | --- |
| 属性类别 | 相关词汇 |
| 环境 | 环境 卫生 装修 隔音 通风 光线 采光 视野 风景 景观 海景 景色  周边环境 周围环境 卫生条件 |
| 服务 | 服务 态度 客服 阿姨 员工 经理 前台 店家 素质 保洁 接待 打扫  收拾 整理 服务员 服务生 服务质量 办事效率 |
| 价格 | 价钱 价格 价位 费用 房费 房价 性价比 |
| 设施 | 设施 配置 家电 设备 物品 家具 用品 用具 装饰 空调 生活用品 基础设施 |
| 交通 | 交通 位置 地理 地点 地段 周边 周围 附近 出行 公交 地铁 公交车  公交站 商业区 市中心 公交车站 地理位置 |
| 整体 | 整体 总体 总之 酒店 公寓 宾馆 总的来说 |

（二）属性词和情感元组抽取

1. 基于依存句法的抽取规则

依存句法分析可以确定句子的句法结构或者句子中词汇之间的依存关系。简单来说，就是判断在句子中的“主谓宾'、“定状补"的成分，可以清晰的展现句子的结构。通过句法依存分析可以准确找到情感词作用的情感对象以及修饰词作用的情感词，避免了直接抽取造成的关系冲突问题。

本文通过哈工大研发的语言计算平台（LTP）[10]实现了句法依存分析。通过平台提供的开源代码库即可实现句法依存分析。此外，平台官网提供了在线生成依存关系图的功能，可以直观地看到句子的结构。

哈工大语言计算平台总结表示，一共有14种句子依存关系[10]，如下表所示：

表8 依存关系表

|  |  |  |
| --- | --- | --- |
| 关系标签 | 关系类型 | 例子 |
|  | 主谓关系 | 他吃过饭了 (他 <– 吃) |
|  | 动宾关系 | 他吃过饭了 (吃 –> 饭) |
|  | 间宾关系 | 我送她一束花 (送 –> 她) |
|  | 前置宾语 | 我每道菜都吃 (菜 <– 吃) |
|  | 兼语 | 她帮我写作业 (帮 –> 我) |
|  | 定中关系 | 蓝月亮 (蓝 <– 月亮) |
|  | 状中结构 | 格外动人 (格外 <– 动人) |
|  | 动补结构 | 吃完饭 (吃 –> 完) |
|  | 并列关系 | 猫和老鼠 (猫 –> 老鼠) |
|  | 介宾关系 | 在贸易区内 (在 –> 内) |
|  | 左附加关系 | 猫和老鼠 (和 <– 老鼠) |
|  | 右附加关系 | 兄弟们 (兄弟 –> 们) |
|  | 独立结构 |  |
|  | 核心关系 |  |

本文借用了上表的七种关系。首先通过SBV、VOB、ATT、COO、FOB这五种关系，结合情感词典和属性词典抽取出<情感词，属性词>二元组；然后通过ADV、CMP两种关系，结合修饰词典抽取出情感词的修饰词语，最终得到<情感词，属性词，修饰词>三元组。下面以酒店领域的语料为例，使用平台提供的句法依存树介绍<情感词，属性词>的抽取关系：

1. SBV关系

SBV是主谓关系，如果属性词作主语以SBV关系依赖于作为谓语的情感词，可以认为该属性词和情感词是对应的。例如评论“价格非常便宜”，属性词“价格”和情感词“便宜”之间具有SBV关系，能够得到<价格，便宜>这一对应元组。

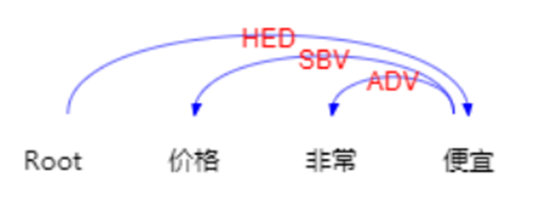


图2 SBV图解

1. VOB关系

VOB是动宾关系，如果属性词作宾语以VOB关系依赖于作动词的情感词时，可以认为该属性词和情感词是对应的。例如评论“没有空调”，属性词“空调”和情感词“没有”之间具有VOB关系，通能够得到<空调，没有>这一对应元组。



图3 VOB图解

1. ATT关系

ATT是定中关系，如果情感词作定语以ATT关系依赖于作中心语的属性词时，可以认为该属性词和情感词是对应的。例如评论“相对便宜的价格”，属性词“价格”和情感词“便宜”间具有ATT关系，能够得到<价格，便宜>这一对应关系。

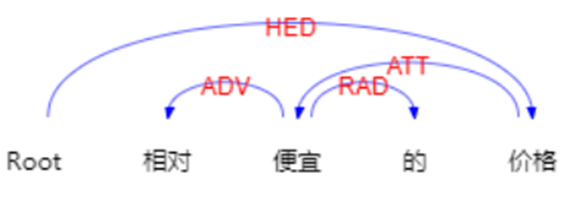


图5 ATT图解

1. COO关系

COO是并列关系，包括属性词的并列和情感词的并列。利用COO可以通过具有并列关系的词语找到对应的情感词或属性词。例如评论“酒店的设施、环境都不错”，“设施”和“环境”具有并列关系，同时“设施”由“不错”进行修饰，所以“环境”和“不错”也具有修饰关系。能够得到<设施，不错>、<环境，不错>这两个对应关系。

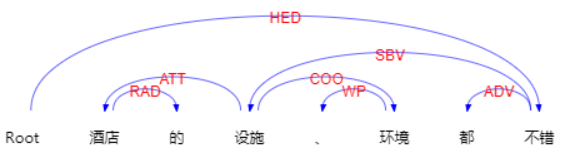


图6 COO图解

1. FOB关系

FOB是前置宾语关系，如果属性词作宾语以FOB关系依赖于作谓语的情感词时，可以认为该属性词和情感词是对应的。例如评论“床单没有整理”，属性词“床单”和情感词“整理”间具有FOB关系，能够得到<床单，整理>这一对应关系。

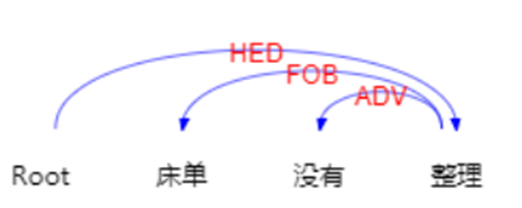


图4 FOB图解

可以看到，通过<床单，整理>来判断句子的情感显然是错误的，因为否定修饰词“没有”对句子的情感有巨大的影响。因此，只抽取出情感词和属性词是不够的，还需要抽取出作用在情感词上的修饰词。下面介绍能够找出修饰词的ADV、CMP两个关系：

（1） ADV关系

ADV是状中关系，利用ADV可以得到与情感词对应的修饰词。还以评论“床单没有整理”为例，可以看到否定词“没有”以ADV关系依赖于情感词“整理”，因此加入否定词后可以抽取出<床单，整理，没有>。

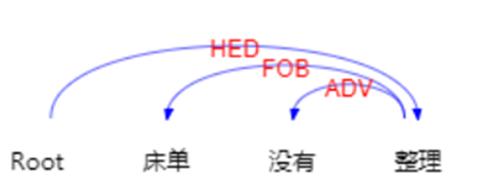


图7 ADV图解

（2） CMP关系

CMP是动补关系，可以利用CMP关系抽取属性词与情感词对，也可以利用CMP关系抽取修饰情感词的修饰词。对于评价语句“房间整理不错”，情感词“整理”是动词，修饰词“不错”通过CMP关系依赖于情感词，同时，属性词“房间”通过SBV关系依赖于情感词，因此可以抽取出<房间，整理，不错>。



图8 CMP图解

2. 潜在特征属性问题

现实中，用户评论有时会省掉针对的评价属性，只用情感词，而我们依然可以从情感词中知道潜在的特征属性。例如“酒店服务很好，就是有点旧'。例句中“旧”是指酒店的设施陈旧，针对的特征属性是酒店的设施。但是如果仅用依存关系进行抽取，这一情感元组就会被忽略掉。

情感词和属性词具有潜在的关联，从特定的情感词可以反向找到潜在的属性词。本文将能够推断潜在属性词的情感词进行了总结，如下表所示：

表9 潜在情感词典

|  |  |
| --- | --- |
| 属性类别 | 相关词汇 |
| 环境 | 潮 暗 脏 卫生 干净 整洁 舒适 舒服 温馨 潮湿 昏暗 明亮  宽敞 漏水 反味 难闻 发霉 干净利落 |
| 服务 | 热情 贴心 耐心 友善 冷漠 体贴 周到 亲切 细心 细致 和蔼 |
| 设施 | 新 旧 齐全 陈旧 老旧 破旧 很全 |
| 价格 | 贵 实惠 便宜 划算 优惠 不贵 经济 涨价 物所超值 |
| 交通 | 方便 便利 快捷 |

3. 属性词与情感元组抽取算法

|  |
| --- |
| 算法1：情感三元组抽取算法  输入：评论语句S，属性词典T，情感词典Q，修饰词典X，潜在属性词典C  输出：情感三元组集合List[d1,d2,…,dn]  使用LTP提供的模块进行分词、词性标注、依存句法分析，得到分词集合W[w1,w2,…,wn]和对应的关系集合R[r1,r2,…,rn]  FOR wi in W:  IF wi在属性词典T中，且对应的ri为SBV、VOB或FOB关系，且依赖的词语在情感词典Q中：  抽取出属性词和情感词  IF 集和W中有以ADV或CMP关系依赖于此情感词的修饰词：  抽取出修饰词，构成字典d{属性词，情感词，修饰词}，并添加到集合List中  ENDIF  ELIF wi在情感词典Q中，且对应的ri为ATT或CMP关系，且依赖的词语在属性词典T中：  抽取出属性词和情感词  IF集和W中有以ADV或CMP关系依赖于此情感词的修饰词:  抽取出修饰词，构成字典d{属性词，情感词，修饰词}，并添加到集合List中  ENDIF  ELIF wi在属性词典T中，且ri为COO关系：  通过COO找到与wi并列的属性词，引用并列词汇抽取出的情感词和修饰词加入到新的字典d  中，并添加到集合List中  ELIF wi在情感词典Q中，且ri为COO关系：  通过COO找到与wi并列的情感词，引用并列词汇抽取出的属性词和修饰词加入到新的字典d  中，并添加到集合List中  ENDIF  ENDFOR  FOR wi in W:  IF wi是未抽取的情感词，且wi在潜在属性词典C中：  抽取出该情感词，并根据其对应的潜在属性，添加属性词  IF集和W中有以ADV或CMP关系依赖于此情感词的修饰词:  抽取出修饰词，构成字典d{属性词，情感词，修饰词}，并添加到集合List中  ENDIF  ENDIF  ENDFOR  RETURN List |

以评论语句“性价比很高的酒店，住着很舒服，服务人员也都非常热情，就是周边交通很不方便”为例，抽取的结果如下：

表10 抽取词对表

|  |
| --- |
| {’属性词’ : ’性价比’ , ’情感词’ : ’高’ , ‘修饰词’ : [‘很’]} |
| {’属性词’ : ’交通’ , ’情感词’ : ’方便’ , ‘修饰词’ : [‘很’，‘不’]} |
| {’属性词’ : ’环境’ , ’情感词’ : ’舒服’ , ‘修饰词’ : [‘很’]} |
| {’属性词’ : ’服务’ , ’情感词’ : ’热情’ , ‘修饰词’ : [‘非常’]} |

（三）情感计算

上文得到的每个元组都含有一个属性词、一个情感词和修饰词。我们不能简单地根据情感词来判断情感倾向，需要考虑修饰词对情感倾向的影响，这就需要数值化修饰词的影响程度。具体做法如下：

|  |
| --- |
| 算法2：情感倾向计算  输入：List[d1,d2,...dn],正向情感词典，负向情感词典，程度副词词典，否定词词典，qlist[]  输出：Dict{属性类别:情感值}  FOR di in List:  IF di中的情感词在正向情感词典中：  情感值q=1  ELIF di中的情感词在负向情感词典中：  情感值q=-1  ENDIF  IF di中的修饰词在程度副词词典中：  根据其程度权重w计算q=q\*w  ENDIF  计算修饰词中否定词的个数，记为count  IF count为偶数：  q=1\*q  ELIF count为奇数：  q=-1\*q  ENDIF  根据di中的属性词的所属类别，进行归类，得到字典{属性类别:q}，添加到集合qlist中  ENDFOR  事实上一条评论中可能对一个属性类别进行了多次情感表达，可能有的属性类别没涉及到。因此，要对qlist进行整合，得到这条评论在六个属性类别的情感倾向:  IF 一个属性类别有多个情感值，则将各个值相加，结果记为total：  IF total>0:  则最终的情感值q=1  ELIF total<0:  则最终情感值q=-1  ENDIF  向字典Dict中添加键值对{属性类别:q}  ELIF 评论未涉及到此属性类别：  最终情感值q=0  向字典Dict中添加键值对{属性类别:q}  ENDIF  RETURN Dict |

以上一节的评论为例，根据其抽取的结果进一步得到细粒度情感倾向：

表11 单条评论分析结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 环境 | 服务 | 设施 | 价格 | 交通 | 整体 |
| 1 | 1 | 0 | 1 | -1 | 0 |

（四）实验及结果分析

以“大连星海假日酒店”的评论为例，该酒店共8939条评论，来挖掘用户对环境、服务、设施、价格、交通、整体六个方面的满意度。通过前面介绍的算法，可以得到该酒店每个用户对这六个方面的情感倾向，部分结果如下表所示：

表12 前十条评论的运算结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 评论内容 | 环境 | 服务 | 设施 | 价格 | 交通 | 整体 |
| 出门就是星海公园 星海广场也很近。很方便 楼下有便利店 公交车也多 。适合溜溜达达悠闲的逛，整洁卫生，可以放心入住 | 1 | 0 | 0 | 0 | 1 | 0 |
| 很棒的酒店。很适合出差和家庭住，房间很整洁，设备很齐全，还有冰箱，乐视电视有会员，可以看电影，前台很热情，水很热 | 1 | 1 | 1 | 0 | 0 | 1 |
| 服务态度环境都不错，就是这个房间的设施有点差强人意了,冲洗功能就完全坏了用不了；大冬天的，暖气好像不给力，空调也不给力，幸好有两床被子。 | 1 | 1 | -1 | 0 | 0 | 0 |
| 干净，服务态度好，空调坏了。离公交站挺近，过街就是。 | 1 | 1 | -1 | 0 | 1 | 0 |
| 酒店位置好，离星海公园步行很近，交通便利，出行方便，附近餐馆很多，吃饭方便，酒店环境好，卫生干净，设施齐全，无线信号强，前台服务态度热情，推荐! | 1 | 1 | 1 | 0 | 1 | 0 |
| 非常普通，我定的是没有窗户的，不通风！只能说勉勉强强比最差的小旅馆，小旅店强那么一点点！地理位置是优势！周围吃饭交通都很方便！一分钱一分货吧，没有什么性价比！ | 0 | 0 | -1 | -1 | 1 | 0 |
| 以前出门都看好如家酒店，但是这次来大连，定的如家酒店，卫生设备老旧 马桶 喷头都坏了 隔音效果差 隔壁说话都听得到 如家要赶紧整治 不然以后大家都不来住了 | -1 | 0 | -1 | 0 | 0 | 0 |
| 整体不错，地理位置不错，里星海公园很近，马路对面就是。去圣亚海洋世界方便。酒店服务前台很热情，就是旅游旺季价格偏贵 | 0 | 1 | 0 | -1 | 1 | 1 |
| 酒店房间很大，周围吃饭很方便，交通也方便，前台也比较热情，美中不足是个别房间的无线信号很差，水龙头声音太吵 | 0 | 1 | -1 | 0 | 1 | 0 |
| 酒店不好 无窗的房间 装修也破旧了 卫生也不太干净，总之住着不是很舒服 除了离星海公园近之外无其他优点了，最好不要住无窗房间 空气不流通 | -1 | 0 | -1 | 0 | 1 | -1 |

再按照如下公式对所有结果进行计算即可得到每个属性特征的满意度：

某属性的满意度=情感值为1的个数/(情感值为1的个数+情感值为-1的个数)

结果如下：

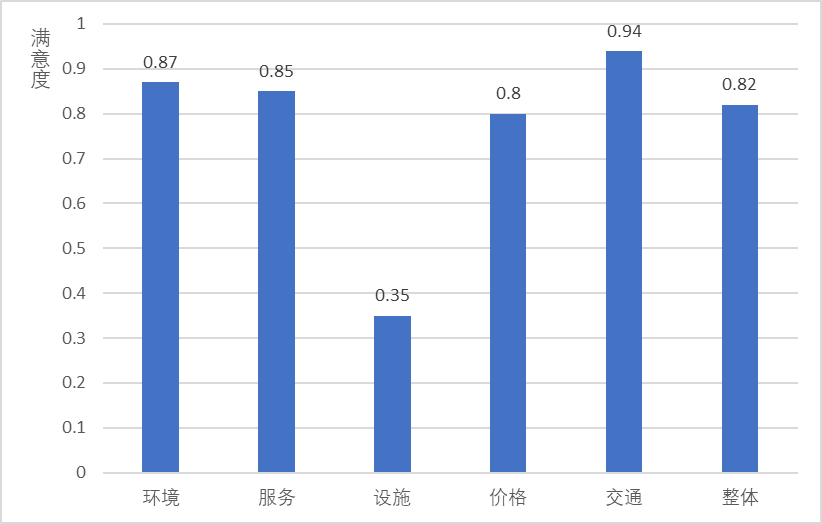


图9 各属性满意度柱状图

可以发现，用户对“交通位置”的满意度比较高，说明酒店的地段不错占据较好的位置，该酒店临近星海广场，具有独特的地理优势；在“环境”、“服务”、“价格”等方面，用户满意度都很高，没有太大的差别，说明该酒店整体上管理的不错；但在“设施”方面，用户的满意度仅为0.35，是非常低的，说明设施这块是酒店的短板，从用户的评论内容来看，设施包括了空调、热水供应、暖气、无线网等。因此，总体来说酒店的软件实力不错，但硬件设施不行，商家应尽快弥补自身短板，升级改善酒店的设施，可以更换新的床上用品、提高无线网速，购置新的空调、热水器，淘汰老旧设备等。

# 五、用户类型与满意度的相关分析

前文从属性层面进行了细粒度的情感分析，但结果较为粗略，为进一步发掘酒店的不足，本章从用户类型的角度出发，探究不同出游类型的用户在各属性方面的满意度的差别，并结合用户的好评率进行分析。好评率的计算需要用到传统的粗粒度情感分析，判断整条评论的情感极性。实验根据携程网上用户的评分来计算情感标签，采用有监督的机器学习方法进行情感分类。

（一）情感标签处理

本文爬取的实验数据是不含情感标签的，但可以根据用户的评分来添加情感标签。携程网的酒店评分满分为5分，本文拟设定一个边界分数，如果评分高于此分数，则评论情感倾向为正，否则为负。接下来的问题就是如何确定这个边界分数。

对爬取的57734条数据进行统计，发现分数不低于4.0的评论占52.3%，高分的比重占了一半。通过观察发现，评分在4以上的评论具有明显的正向情感；评分在3-4之间的评论大都是既有好的一面也有坏的一面，但负面居多；评分在3以下的评论都具有的负向情感。可以初步判断边界分数在3-4之间。

为了使正负语料尽可能的平均，以3.8为分界线，把评分小于3.8的归为负向评论。正向语料共32516条，负向语料共25218条，占比如图10所示。最终正负语料各取25000条，共计50000条评论语料。

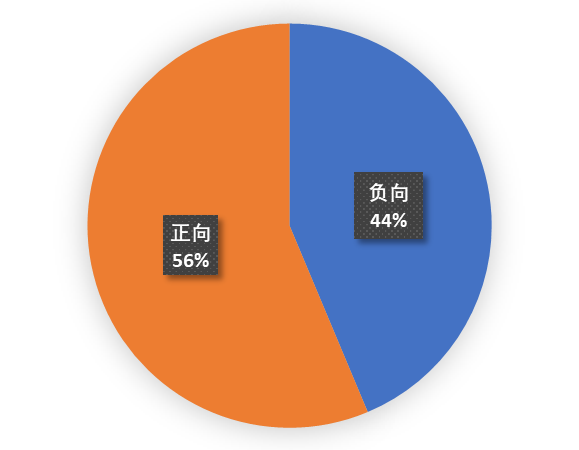


图10 正负文本饼图

（二）机器学习流程

将已标注分类的文本按一定比例划分为训练集和测试集。通过对训练集的学习，基于机器学习的分类算法构建出分类器，利用分类器学到的知识指导测试集的分类。本文使用机器学习进行文本分类的方法大致有以下几个步骤：

1. 文本预处理，包括对语料进行分词，去停用词等。
2. 按一定比例划分测试集和训练集，用TF-IDF算法进行文本特征表示。TF-IDF能够评估词语的重要程度，获得词语的权重。
3. 选取合适的模型进行训练，构建分类器，输入处理好的特征向量训练分类器。

④ 将测试文本输入分类器，输出情感倾向。对比各分类器的准确率、召回率等相关指标，选出效果最优的。

（三）实验及结果分析

实验中以75%作为训练语料，25%作为测试语料，并采用交叉验证的方法得到结果数据。实验采用了朴素贝叶斯、支持向量机、ANN、逻辑回归四种分类器，并根据预测准确率进行了对比，结果如下：

表13 分类器对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分类器 | 朴素贝叶斯 | 支持向量机 | ANN | 逻辑回归 |
| 准确率 | 91.6% | 92.5% | 89.9% | 91.3% |
| 召回率 | 88.7% | 90.3% | 83.2% | 90.1% |
| F值 | 0.89 | 0.91 | 0.85 | 0.90 |

可以发现，各分类器差距不是很大，召回率也都比较高，因此可以用准确率来评估各分类器。其中，支持向量机的准确度最高，达到了92.5%。我们用训练好的支持向量机模型计算一个具体酒店的评论文本，还以“大连星海假日酒店”为例，计算该酒店总体的好评率：好评率=正向评论数/总评论数。该酒店的好评率为81.8%。

该酒店好评率不是很高，为进一步找出酒店的不足，将细粒度与粗粒度的分析结合起来，从出游类型的角度对比用户的满意度和好评率。各方面评论的占比以及用户在各方面的满意度和好评率如下所示：

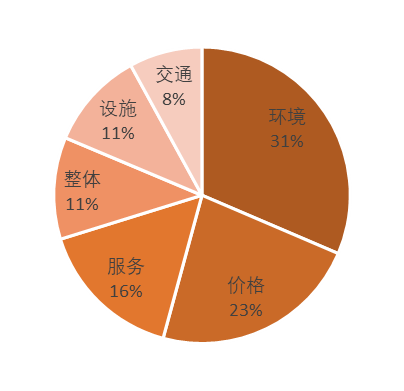
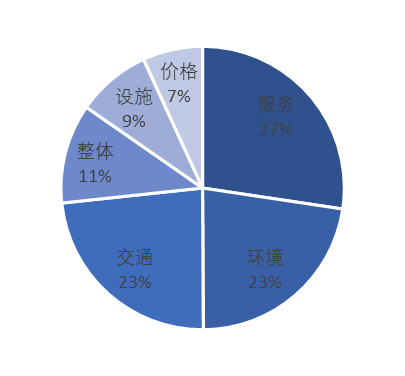


图11 商务出差各属性评论占比 图12 家庭出游各属性评论占比

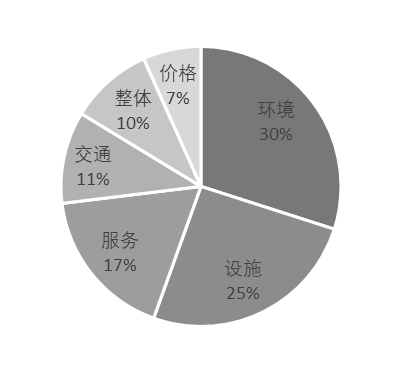
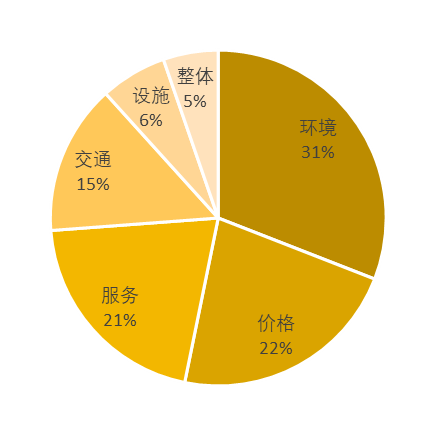
 

图13 情侣入住各属性评论占比 图14 独自旅行各属性评论占比

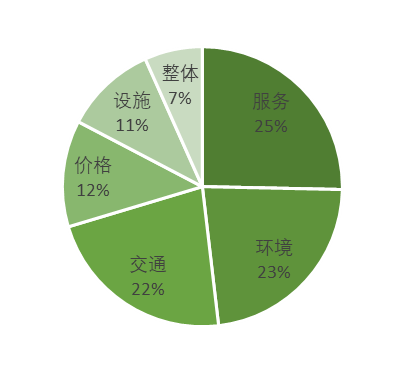


图15 朋友出游各属性评论占比

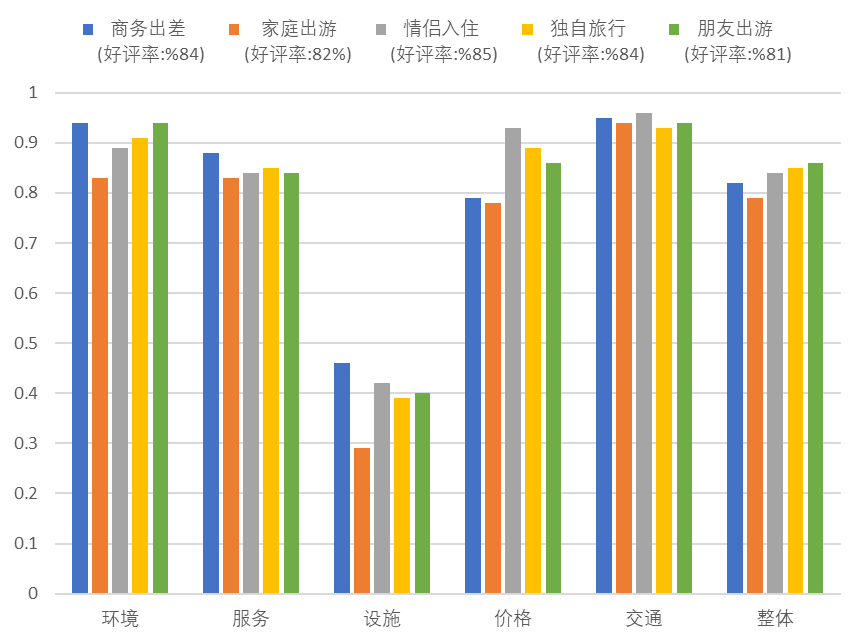


图16 满意度、好评率对比图

通过饼图可以发现，无论哪种类型的用户，对酒店的卫生环境都比较关心，说明环境对用户评价有着很大的影响，是衡量一个酒店好坏的基本要素。除环境外，每种类型用户的关注点也存在着差别，商务出差的用户更关注服务、环境卫生以及交通的便利程度，而对价格的关注度较少；家庭出游的用户则更关心酒店的卫生环境和性价比；情侣更关注酒店的环境和设施，而对价格关注度较少；独自旅行的用户更关注酒店的环境、价格和服务，对设施的关注度较少；朋友出游的用户则更关注服务和交通。

再结合满意度、好评率对比的柱状图，可以发现所有出游类型给出的整体好评率相差不大，都在82%左右，但不同类型的用户在酒店各方面的满意度有所差别。例如在价格方面，情侣对酒店的满意度最高，说明情侣对价格高低不是很在意，他们更关心住的舒不舒服，而家庭出游的用户对价格的满意度最低，说明他们更希望拥有实惠的套餐，追求高性价比；在环境方面，家庭出游用户的满意度最低，考虑到家庭用户中有孩子或老人，说明他们对卫生环境的要求更高；在交通方面，各类型用户的满意度基本一致，均处在较高的水平，说明酒店的位置优势是吸引用户的一个点；同时，可以看到各类型用户对设施的满意度都很低，均在50%以下，说明设施这块是酒店的短板，酒店应急需改进。

基于上述分析，对大连星海假日酒店提出以下建议：

1．可以针对不同类型的用户提供不同的入住套餐。比如：对于商务出差、情侣这种能够接受高价格的用户，可以向其推荐更为高档的房间，提高价格的同时提高房间的舒适度和设施条件，也可以额外提供早餐服务；对于家庭出游用户则推荐经济实惠、装修朴素的房间，确保环境卫生、设施齐全即可，可以为携带孩子的家庭提供儿童专用的洗漱用品。

2．应尽快弥补短板，全面改善酒店的设施。可以从用户反馈较多的问题着手，例如：房间的隔音差、暖气和热水的供应、空调老旧等问题，同时要注意酒店的清洁以及床上用品及时更换。加上自身优越的地理位置，可以很大程度的提高竞争力。

# 六、总结

（一）全文回顾

随着电子商务的迅猛发展，越来越多的人在网上购物，每人都有表达自己的机会。因此互联网充斥着大量的评论数据，而这些评论数据反映着用户的情感倾向。如果掌握了用户评论的情感倾向就可以了解用户的观点，挖掘产品的潜能，为商家提供决策支持，这使得互联网亟须自动化的情感倾向分析工具。本文工作总结如下：

1. 构建了酒店领域的属性词典和情感词典。使用word2vec词向量模型识别出属性词生成属性词典，包含了价格、环境、服务、设施、交通、整体六个用户关心的方面；情感词典则结合了知网情感词典和台湾大学情感词典，在此基础上，用词向量的方法从评论语料中抽取酒店领域的情感词，添加到情感词典中。

2.借助哈工大语言技术平台提供的方法进行依存句法分析，得出句子的结构，并结合属性词典、情感词典以及修饰词词典抽取出<属性词，情感词，修饰词>三元组来计算酒店在各个属性上的情感倾向。

3. 使用传统方法对整条评论进行粗粒度情感分析。使用多种机器学习的方法进行有监督学习，并对比各方法的准确率。

（二）不足和未来展望

本文对评论文本情感倾向分析进行了初步的研究，取得了一定的成果，同时也存在着一些不足之处：属性词典构筑的不够完善；由于细粒度的情感分析缺乏标注好的训练语料，未采用机器学习的方法，使用机器学习的方法会更快速、准确。

在如今网络发达的环境下，评论文本也越来越多，用户的表达方式和用语也出现了新的变化，比如会套用一些网络现象来表达自己的情感。目前我们做的工作还是远远不够的，还有很多值得改进的地方，例如：构建更完善精准的词典、提高分词效果等。这些都有助于提高情感分析的准确度。

参考文献

1. 何立峰，基于在线评论的酒店顾客需求分析研究，青岛大学，2019。
2. 高楠，在线评论对酒店消费者购买决策影响研究，旅游纵览，2019。
3. 敦欣卉， 基于微博的药物评论细粒度情感分析，吉林大学，2018。
4. 唐明、朱磊，基于word2vec的一种文档向量表示，计算机科学，2016。
5. 李长江，基于酒店中文评论情感倾向分析，华南理工大学，2016。
6. 蓝天广，电子商务产品在线评论的细粒度情感强度分析，北京邮电大学，2015。
7. 刘丽、王永恒、韦航，面向产品评论的细粒度情感分析，计算机应用，2015。
8. 雷龙艳，中文微博细粒度情绪识别研究，南华大学，2014。
9. 黄胜，Web评论文本的细粒度意见挖掘技术研究，北京理工大学，2014。
10. 李正华，汉语依存句法分析关键技术研究，哈尔滨工业大学，2013。
11. 刘楠，面向微博短文本的情感分析研究，武汉大学，2013。
12. 杜振雷，面向微博短文本的情感分析研究，北京信息科技大学，2013。
13. 施寒潇，细粒度情感分析研究，苏州大学，2013。
14. 张玥，面向产品评价的细粒度情感分析技术研究，哈尔滨工业大学，2013。
15. 张冬梅，文本情感分类及观点摘要关键问题研究，山东大学，2012。
16. 王山雨，面向产品领域的细粒度情感分析技术，哈尔滨工业大学，2011。
17. e.a.N.Kobayashi、K.Inui， Extracting Aspect-evaluation and Aspect-of Relations in Opinion Minging，Proceedings of EMNLP-CONLL, 2007。
18. P. D. Turney，Thumbs up Or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews，Proceeding ACL Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002。
19. L. L. Pang、S. Vaithyanathan，Thumbs Up? Sentiment Classification Using Machine Learning Techniques，Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002