摘　要

随着智能手机的普及和互联网的快速发展，在日常经济生活中，网络在线评论越加能够影响人们的决策，通过网络舆论导向引导消费者购买甚至成为一种趋势。网络在线评论不仅能够对消费者购买产生影响，通过对其进行分析也能够得到大量隐藏的有价值的信息。本文将文本分析技术和机器学习模型应用在新能源汽车在线评论分析领域，通过网络评论挖掘消费者对新能源汽车的情感取向，进一步深入挖掘消费者痛点和新能源汽车的优劣之处，为新能源车企和消费者购买决策提供一定的参考。

本文使用绕过汽车之家和易车网反爬虫方式的 Python 网络爬虫技术，获得 68523 条可用数据，并且对数据进行了预先处理。为了得到更好的分词效果，本文获取了百度和搜狗的汽车行业专有词库并将其与数据清洗后的汽车品牌数据合并作为 jieba 分词专有名词库。同时，还合并了包括哈工大停用词库在内的多个停用词库，以实现更好的文本分词，并统计分词词频以绘制词云图。然后使用了 Word2Vec 模型进行分布式词向量训练，并以 TF-IDF 算法所得结果作为权重进行加权计算单个评论文本的空间向量，最终构造特征矩阵作为机器学习分类算法的训练集和测试集数据。然后以生成的特征矩阵进行了逻辑回归和随机森林模型的构建，并以测试集数据计算了模型的准确率、精准率、召回率、F1-score 和 AUC 值，以及绘制了模型的 ROC 曲线作为模型分类效果的评价标准。在构建 LDA 模型前，进行了预训练绘制模型困惑度曲线和一致性曲线，以确定主题数量。然后分别以汽车之家的舒适性、内饰、电耗和操控维度的数据进行了 LDA 模型训练，并提取主题关键词及其权重，进行了结果分析。

关键词： 新能源汽车，在线评论，Word2Vec，机器学习，LDA

Research on Emotion Analysis of Online Reviews of New Energy Vehicles Based on Machine Learning

Abstract

With the popularity of smart phones and the rapid development of the Internet, in daily economic life, online comments can more and more influence people's decision-making, and it even becomes a trend to guide consumers to buy through the guidance of online public opinion. Online comments can not only influence consumers' purchase, but also get a lot of hidden valuable information by analyzing them. In this paper, the text analysis technology and machine learning model are applied in the field of online review and analysis of new energy vehicles. Through online reviews, the emotional orientation of consumers towards new energy vehicles is mined, and the pain points of consumers and the advantages and disadvantages of new energy vehicles are further explored, which provides some reference for new energy vehicle enterprises and consumers' purchasing decisions.

In this paper, Python web crawler technology, which bypasses the anti-crawler mode of car home and Yiche. com, is used to obtain 68,523 available data, and the data is preprocessed. In order to get a better segmentation effect, this paper obtains the auto industry-specific thesaurus of Baidu and sogou, and merges it with the car brand data after data cleaning as jieba word segmentation proper noun library. At the same time, a number of disabled word banks, including Harbin Institute of Technology's disabled word bank, are merged to achieve better text segmentation, and the word frequency of word segmentation is counted to draw the word cloud map. Then, Word2Vec model is used to train distributed word vectors, and the results of TF-IDF algorithm are used as weights to calculate the space vector of a single comment text. Finally, the feature matrix is constructed as the training set and test set data of machine learning classification algorithm. Then, the logistic regression and random forest model are constructed by using the generated feature matrix, and the accuracy, precision, recall, F1-score and AUC values of the model are calculated by using the test set data, and the ROC curve of the model is drawn as the evaluation standard of the model classification effect. Before constructing LDA model, pre-training was carried out to draw model confusion curve and consistency curve to determine the number of topics. Then, the LDA model is trained with the data of comfort, interior decoration, power consumption and control dimensions of car home, and the subject keywords and their weights are extracted, and the results are analyzed.

Key Words： New energy vehicle， Online user reviews， Word2Vec，Machine learning，LDA

目　录

1 绪论

1.1 研究背景和意义

1.1.1 研究背景

随着我国经济的快速增长和人民消费能力及消费意愿的增强，汽车逐步走进千家万户，成为万千家庭不可缺少的消费物品。汽车行业作为我国最大的经济支柱行业之一，在推动我国经济发展以及提升人民生活幸福感方面发挥了极大作用。但是汽车产业的快速发展也伴随着自然资源消耗和环境污染的加剧，这和人与自然和谐发展的优良理念背道而驰。汽车产业发展与生态绿色发展观念相耦合使得新能源汽车得以产生和发展，奠定了汽车行业未来的发展新方向。

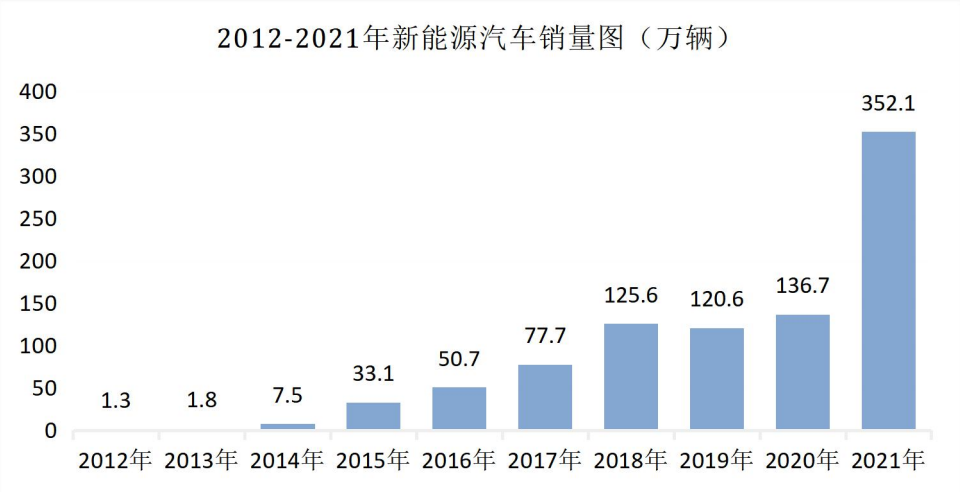


图 1-1　2012-2021 年新能源汽车销量图

发展迅猛的国产新能源汽车在时间轴上晚于国外同行业，但拥有令人期待的未来。如图 1-1 所示，年度新能源汽车销量从 2013 年的 1.3 万辆跃升为 2021 年的 352.1 万辆。尤其是疫情期间，在全球经济低迷的环境下，我国新能源汽车产业仍保持高速发展，2021 年新能源汽车销量环比为 257.6%，成为疫情下我国经济复苏的重要动能。从图 1-1 2012-2021 年新能源汽车销量图可以看出，2012 年到 2014 年我国新能源汽车发展缓慢，销量增长率低。但从 2015 年起我国新能源汽车开始飞速发展，销量迅速增加，国产新能源汽车品牌如比亚迪等迅速兴起，成为我国新能源汽车产业的重要推动力量。根据乘联会数据，2021 年我国新能源汽车渗透率达到 14.8%，意味着在我国每每有 7 辆汽车被购买，新能源汽车必含于其内。我国发展势头极为厉害的新能源汽

车之所以有此成绩，与人民生态环境保护意识的提高、国家环境保护政策和对新能源汽车的补贴及优惠政策息息相关。新能源汽车相比于传统汽车具有污染小更环保、政策补贴更省钱、噪声小更舒适以及更加节能等优势，契合了当前生态保护趋势。

截至 2021 年 6 月，我国的网民数量已经越过 10 亿大关，网络普及率更是高达 71.6%，此项数据的来源为第 48 次《中国互联网络发展状况统计报告》[1]。网络在人们的生活、消费等方面产生了重大影响。消费者可以通过网络发表自己对商品的评价，也可以通过网络评价揭开商品包装的面纱获取其他消费者对商品的真实感受以利于自己的购买决策。

1.1.2 研究目的及意义

根据《2019 新能源汽车消费市场研究报告》，在线评论对消费者是否购买该车的影响是所有能影响消费者购买与否的因素中最为重要的一点，当然要首先排除消费者本身的自我肯定。新能源汽车作为新兴产业，许多消费者对其知之甚少，更加希望能够通过网络获取其他消费者对于其真实客观的评价，以形成对新能源汽车的正确认知。新能源车企也同样希望能够从消费者的真实评价中获取消费者预期以对自身产品和研发方向做出及时调整，适应市场发展。所以从大量的消费者评价信息数据中获取真实有效的信息具有很强的现实意义。

文本挖掘是自然语言处理（NLP）中的热点方向，近些年来一直被国内外学者不断深入研究，形成了很多具有理论价值和实际应用价值的成果。文本挖掘与机器学习的结合更加促进了其快速发展与应用。

本文将首先利用网络爬虫获取各大汽车网站中新能源汽车板块的最新消费者评价信息数据，再通过数据预处理形成有价值可分析的在线评论语料库，通过 jieba 分词，引入停用词及建立专用词典等方式进行词库建设。通过 word2vec 方法将词句映射到数字空间，再导入机器学习模型进行分析，最终得到有价值可应用的模型结果和分析结论，为消费者从大量在线评论中获取有价值的信息提供契机。新能源汽车虽然相比于传统汽车有其独有的优势，但是其发展历程短，存在许多需要优化的方面，如续航能力及充电便利度等，通过挖掘消费者在线评论中蕴藏的信息，可帮助车企分析用户需求，为其自身调整和战略制定提供一定的参考。

1.2 文献综述

1.2.1 新能源汽车研究现状

使用氢能、电能等新型能源作为汽车燃料的及根本动能的车被人们唤为新能源汽车。与占据市场主导地位的传统的燃烧汽油的和燃烧柴油的车辆相与对比，新能源汽车具有环境友好等优势，在近几年内发展迅速，俨然成为了汽车制造行业的未来趋势。国内外学者的主要研究重点是新能源汽车的技术升级和未来发展。鉴于本文不涉及任何自然科学技术方面的研究，故参考文献多为对新能源汽车发展的研究。

在关于新能源汽车的消费市场的研究中，乔靖玚等[2]运用统计学方法对北京市新能源汽车市场进行了调查研究，并对消费者偏好进行了分析。通过数据分析指出 2033 年新能源汽车将在汽车市场占据主要地位，对新能源车企提出了改进建议。丁红萍等[3]依托 4C 理论指出新能源汽车推广营销的关键在于提高消费者对新能源汽车的价值认知。张厚明[4]分析了疫情后我国新能源汽车市场的复苏态势和发展难点。李千千[5]研究了营销策略对新能源汽车在互联网急速发展的形势下市场竞争的影响，重点研究了大数据和新媒体背景下的汽车销售契机。张译等[6]从消费者角度出发分析了新能源汽车购买影响因素，分析结果表明电池技术进步、基础设施建设、国家政策及新能源汽车自由特点对消费者购买意愿有显著影响。叶曼曼[7]通过实证分析指出对消费者购买新能源汽车决策有影响的因素主要有产品的实质刺激、文化因素和个人因素等。

许多学者从不同角度以不同方法对新能源汽车政策进行了研究解读。李晓敏等[8]使用时间序列协整模型和误差修正模型在控制新能源汽车价格，电池价格以及充电桩数量等变量的条件下对财政补贴、购置税减免、不限行不限购及政府和公共机构采购四个政策进行了量化评估，结果显示四种政策对新能源汽车市场的发展均有正向推动作用，其中财政补贴的效果最大，且在 2012-2016 年逐渐效果增强。Sierzchula 等[9]以 2012 年 30 个欧洲国家的政策和新能源汽车市场份额数据为样本，分析了购买补贴与新能源汽车市场份额之间的量化关系，结果表明购买补贴对新能源市场的快速发展有显著的正向影响。Han 等[10]分析了我国两个阶段的新能源汽车补贴计划对其市场渗透率的影响，并认为我国的新能源汽车补贴计划在短期内对新能源汽车保持成本优势非常重要。谢青等[11]从政策工具和创新价值链两个维度对与我国新能源汽车相关的 37 项政策文本进行了全面分析，结果显示环境政策工具在政策工具中使用最多，表明新能源汽车的环境友好性是其受到广泛关注的重要原因。

作为汽车行业未来的发展重心，新能源汽车发展趋势被广泛研究。杨荣华[12]研究了产业融合背景下新能源汽车的发展，指出了包括电池技术，控制技术、软件生态及汽车以太网等的发展对新能源汽车的的整合推进作用。赵雨[13]研究了我国新能源汽车在发展过程中的技术、市场、产业和政策方面的障碍并指出了核心技术、政策和基础设施对新能源产业发展的支柱作用。辜文杰等[14]基于大数据对我国新能源汽车的发展问题和发展前景进行了研究分析，并讨论了大数据在新能源汽车发展中的应用。高春晓等[15]对全国 337 个城市进行了新能源汽车市场分析，在建立城市新能源汽车发展趋势的模型时分析并采用了机器无监督学习中的二阶聚类的方法，将聚类后的城市根据其新能源汽车发展潜力分为四级，不同级别间新能源汽车平均销量和渗透率差异明显。结果显示新能源汽车发展潜力呈现区域特征，区域内部核心城市带动能力强。

1.2.2 文本挖掘研究现状

文本挖掘是指从大量文本内容中获取有价值、可应用、可理解的知识的过程，是自然语言处理（NLP）的重点研究方向。文本挖掘由三部分组成，第一部分是文本挖掘基础方向包括机器学习算法等。第二部分是文本挖掘技术，主要有文本处理、分类和聚类以及主题模型分析等。第三部分是应用层，主要内容有文本信息访问和文本知识发现等。很长时间以来，国内外学者在文本挖掘领域进行了大量研究，提出了许多实用的理论和技术。

Zhang[16]利用 one-hot 进行了词语的向量化表示，缩短了文本挖掘模型参数训练时间。但是在词汇量增加的情况下，one-hot 算法会导致词向量维度爆炸，且数据稀疏性很高。Mikolov 等[17]提出了三层神经网络的 word2vec 模型，在文本的 one-hot 向量化表示基础上能训练出质量更高的文本词向量。Word2Vec 模型能够解决 one-hot 编码时向量维度极度稀疏的问题，并且能够在数字空间中体现出中文词汇间的相关性，为进一步文本分析的技术应用提供可能。阎亚亚[18]利用 5 种算法比较了文本挖掘的词袋模型和 TF-IDF 特征选择在文本分类方面的差异，发现在短文本分类种使用 TF-IDF 的效果要优于词袋模型。毕云杉等[19]提出了基于（ERNIE）知识增强语义表示的文本分类方法，ERNIE 方法第一步先通过 ERNIE 模型得到具有强语义关联的词向量，再利用卷积神经网络获得层次更加深的文本特征的表达，最后通过机器学习方法进行文本分类，该模型相比于传统的 BERT 模型能更好地对中文文本进行有效分类。梁顺攀等[20]设计了并行神经网络模型 TC-ABlstm，它结合卷积神经网络（CNN）和 BilSTM 模型，改善了传统卷积神经网络记忆能力欠缺的

问题，能有效捕获文本的全局信息。在文本挖掘领域，主题提取也是热门研究方向，其在搜索引擎、推荐系统及广告预测等方面被广泛应用。在文本主题提取方面，Blei 等[21]提出了 LDA 模型，用先验分布为狄利克雷分布的隐变量作为文档的主题分布模拟文档的生成。此后关于主题提取的研究大多是建立在 Blei 的 LDA 模型基础之上。

文本挖掘在新能源汽车方面也得到了一定应用。张胜[22]搭建了双向长短时记忆神经网络和条件随机场用于提取汽车评论关键词，并基于多层感知机和 Apriori 频繁项集设计了新能源汽车外观感性工学定量设计方法。余帆[23]运用情感分析、词频统计和语义网络图等方法挖掘消费者对新能源汽车的价值认知，分析结果表明，空间、动力和舒适性是影响新能源汽车用户的主要因素。张瑾[24]使用文本挖掘和主题提取方法对新能源汽车评论文本进行了知识获取和主题识别，以此为依据分析了新能源汽车的营销策略。在分析政策文本中的高频词汇和主题强度等方面时，赵公民等[25]使用了 LDA 主题模型进行主题挖掘。

1.3 研究内容和方法

1.3.1 研究内容

本研究主要通过机器学习、深度学习及文本挖掘相关方法，对新能源汽车近年来在线评论进行分析，获取影响消费者对新能源汽车价值认知的因素，为新能源汽车的发展提供建议。研究过程主要分为四个步骤，首先搜集资料，阅读文献，了解我国新能源汽车的发展进程和发展概况，对汽车之家及易车网等大型网络平台进行数据收集，简要分析多家大型汽车交易平台的在线评论文本，了解消费者评价大体倾向、评论文本的规范性及在线评论涉及的维度以确定后续分析的方法。然后阅读相关文献，了解机器学习及文本挖掘方面的相关算法，评估适用于本研究分析的算法并进行深入学习。了解在新能源汽车评论分析方面文本分析的应用概况和应用方法，确定研究用户在线评论的方法，具体有：数据收集与预处理方法、评文本分词技术、文本向量化技术、主题提取技术、评论情感正负倾向分析方法、词语相似度计算方法等。然后，编写网络爬虫，获取汽车之家等大型汽车交易平台的在线评论数据并进行数据预处理，获得更高质量的数据。然后进行分词，使用 word2vec 和 one-hot 等方法进行文本向量化。最后，首先编写机器学习和文本挖掘算法，代入部分数据进行算法调试。然后将数据处理后的向量化文本输入算法模型进行处理，获取算法结果并进行分析。

概括来说，本研究就是利用 Python 爬虫获取汽车之家及易车网等平台的在线评论数据，通过分词及词向量的转化构建机器学习及情感分析数据集。利用机器学习方法构建并训练文本分类模型，并通过主题提取获取新能源汽车在线评论的主题词并进行分析。

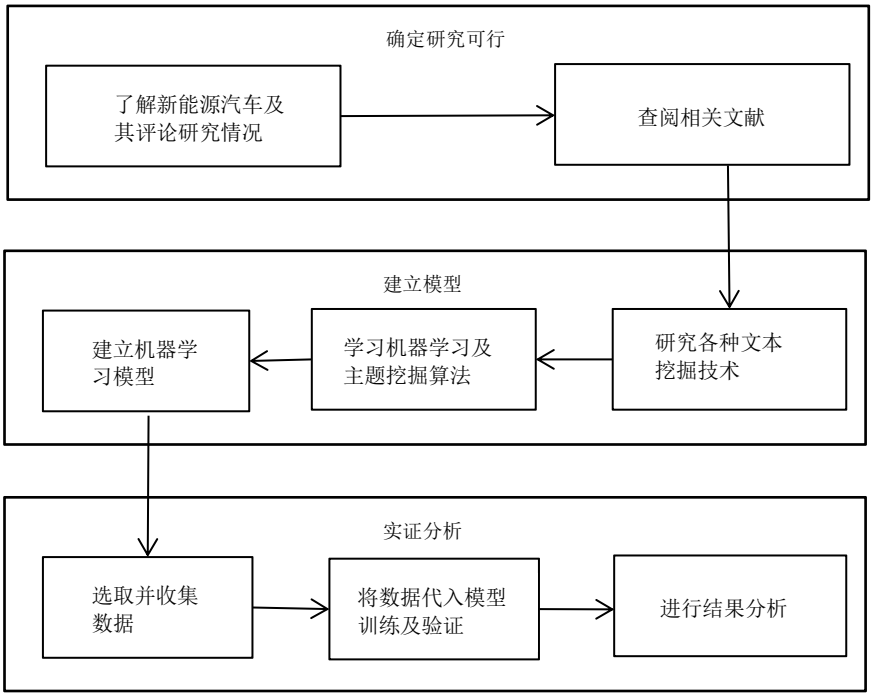


图 1-2 研究框架

1.3.2 研究方法

本论文主要采用理论学习法和实证分析法进行研究。

理论学习法：从官方平台上搜集有关互联网发展状况、新能源汽车发展报告等信息，获取课题背景信息；从专业文献查询平台（如知网、万方、维普等）搜集与研究内容相关的国内外研究文献，包括在新能源汽车消费市场、政策分析、未来发展等以及在线用户评论研究、在线评论情感分析研究、文本挖掘算法等，通过仔细研读文献掌握本研究需要的各种技术和算法的基本原理和运作步骤。

实证分析法：使用 Python 爬虫获取汽车之家和易车网两个大型汽车在线交易网络平台的用户在线口碑评论数据并对其进行一定的先手处理，包括去除无效词、分词、添加停用词等，将分词后的数据进行词云图展示；使用 word2vec 等方法进行词语的向量化转换，并使用机器学习方法进行文本分类，然后使用潜在狄利克雷模型（LDA）提取评论主题词。

2 相关理论

2.1 爬虫介绍

爬虫是一段程序，能够按照一定的规则，成批量地从网络上采集用户所需要的数据。爬虫获取数据的过程与用户直接访问万维网相似，但不同的是用户直接访问只能获取少量的数据，只对数据需求量少的任务有效，而爬虫能够批量地自动获取互联网上所需数据并进行整理，适用于有大量数据需求的任务。本研究将使用易车网等大型平台海量的用户在线评论数据，因此将利用爬虫进行数据批量获取，并将获取的数据保存为规范的 csv 格式数据以便进行后续分析。

爬虫主要由调度器、URL 管理器、解析器、下载器和数据库五个部分组成[26]。图 2-1 展示了 Python 爬虫的基本架构与调度过程。

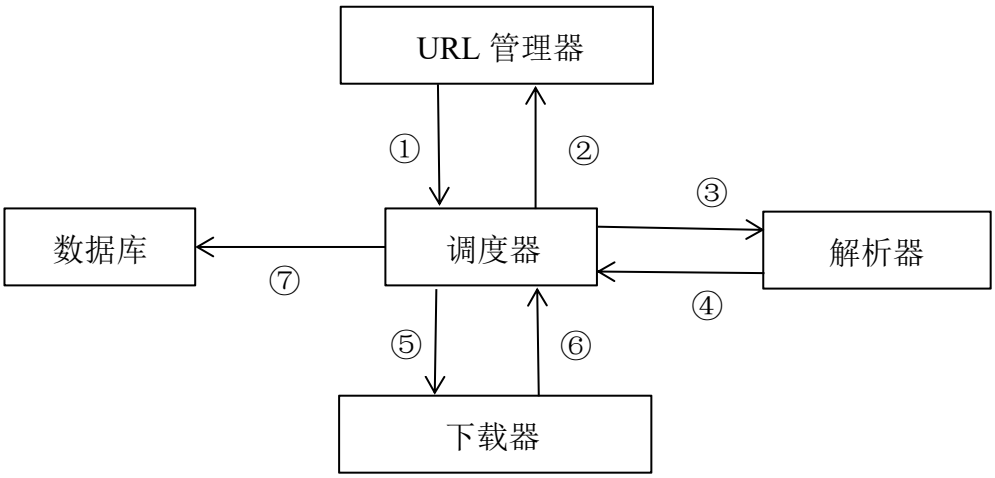


图 2-1 爬虫基本架构

调度器主要负责协调其他爬虫组件的工作。URL 管理器用于管理爬取数据过程中的 URL 地址，包括未爬取的 URL、已爬取的 URL 和新增 URL,通过 URL 管理器还可以防止出现网页的循环抓取。下载器用于根据 URL 管理器传入的 URL 地址获取目标网页并将其下载为文本文件，Python 中的下载器主要有 urlib2 和 requests，均可用于网页获取。解析器用于分析下载器获取的网页文件，根据需要获取有价值的网页信息。解析器一般采用 DOM 树方式解析网页文件，再使用 CSS、正则表达式及 XPath 等方式匹配具体内容，Python 中采用 DOM 树解析方式的库包主要有 Htmlparser、Beatifulsoup、html.Parser 和 Lxml 等。数据库主要用于存储网页文件、URL 地址及从网页文件中解析出的具体有价值的数据等，数据库既可以是 mysql 和 Oracle 等专业数据库，也可

以是 csv 文件、txt 文件及 json 文件等形式。

爬虫工作过程为首先通过 URL 管理器判断是否存在待爬取的 URL 地址，若存在则返回 URL 并传入到下载器，利用网页下载器将传入 URL 对应网页的源代码拷贝到本地，再使用解析器分析网页源代码，利用 CSS、正则表达式或 XPath 等匹配所需数据，最后将所有有价值的信息保存到数据库，进行下一轮数据爬取，直到 URL 管理器中不存在待爬取的 URL。图 2-2 展示了 Python 爬虫的一般流程。

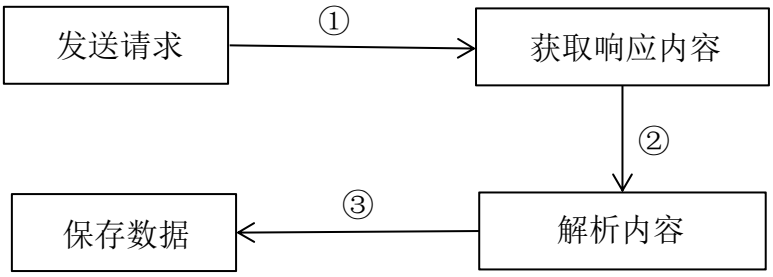


图 2-2 爬虫一般流程

2.2 jieba 分词介绍

20 世纪 80 年代梁南元教授提出了一种基于查字典方式的中文分词，是中文分词最早的方式。中文分词就是在保证文本语义不变的前提下，将中文文本划分为单独的词语。比如将句子“如果光已经忘了将前方照亮，你会握着我的手吗”切分为“如果/光/已经/忘了/将/前方/照亮/你/会/握着/我的手/吗”。

现阶段 NLP 领域进行文本分词常用的有基于字符串匹配、基于统计和基于理解的分词方法。本文使用的分词方法是 Python 的 gensim 模块的 jieba 分词，它是一种基于最大概率的分词方法。由于拥有自带的存储了两万多条数据的词典的加持，对于需要分词的文本，jieba 分词首先依据字典进行拆分，得到每个字的成词方式，再进行后续分词操作。jieba 分词的主要流程如下：

（1）根据 jieba 分词自带词典和传入的个性化自定义词典构建分词文本的前缀词典；

（2）根据前缀词典构建关于分词文本的有向无环图（DAG）；

（3）利用动态规划搜索概率最大 DAG 的的文本划分路径，根据此路径进行文本分词；

（4）对于不属于自带词典和自定义词汇中的词，使用隐马尔可夫模型（HMM），利用维比特算法找到概率最大的隐状态序列。

图 2-3 展示了 jieba 分词的框架结构。

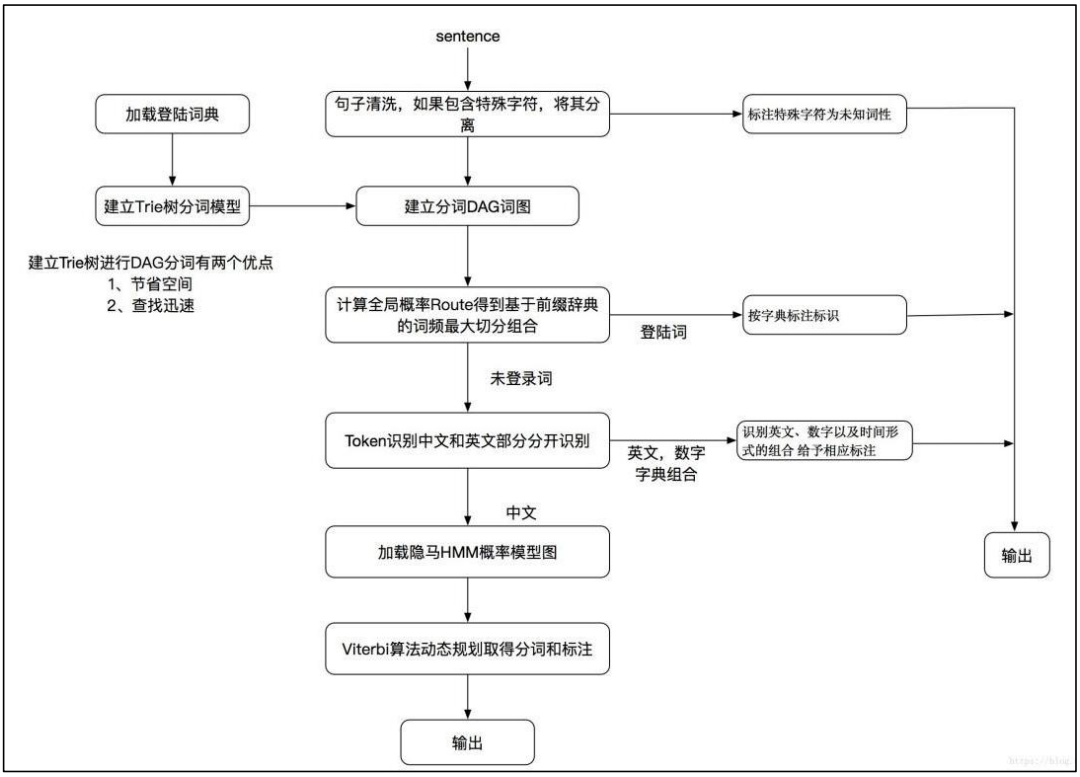


图 2-3 jieba 分词框架图

精确模式、全模式、搜索引擎模式和 paddle 模式是 jieba 分词中包含的四种分词模式。精确模式 jieba 分词时尽可能将文本划分地更加精确，是文本挖掘和文本分析中最常用的 jieba 分词模式。全模式就是快速地扫描待分词文本中的所有词语。搜索引擎模式是对精确模式分词结果中的长词进行再次拆分，适用于搜索引擎工作时对待搜索语句的分词。paddle 模式通过其训练出的双向 GRU 模型进行文本划分，该模式的基石是 PaddlePaddle 框架。除此之外，paddle 模式也支持词性标注[27]。

2.3 Word2Vec 模型

2013 年 Google 研究团队的 Mikolov 等提出了结构如图 2-4 所展示的具有单个隐层的能够高效训练分布式文本词向量的神经网络模型 Word2Vec[28]。Word2Vec 模型能够将分词后的中文文本转化为空间中的数值向量，相比于维度众多且稀疏的 one-hot 算法，Word2Vec 能有效降低文本向量的维度并且能够保留词语间的语义相关性，计算词语间的余弦相似度，Word2vec 模型的这种特性使得其能够应用于本文分析领域，推动了机器学习算法在文本分析中的应用，图 2-4 展示的是 Word2Vec 模型的。Word2Vec 中有两种训练方式，分别是跳字模型（Skip-gram）和词袋模型（CBOW）。给定单词，通过 Skip-gram

模型便可以一定的概率预测其前后文的词语，此模型适用于文本数据量大的模型训练，其训练过程如图 2-5 所示。给定某个前后文，使用 CBOW 模型便可预测当前的单词，CBOW 模型的训练过程如图 2-6，该模型适合在文本数据少的时候使用。

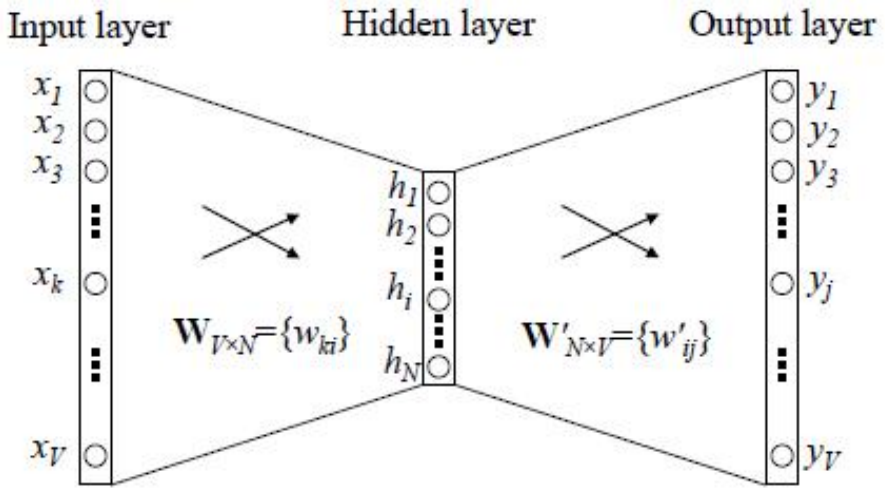


图 2-4 Word2Vec 模型结构

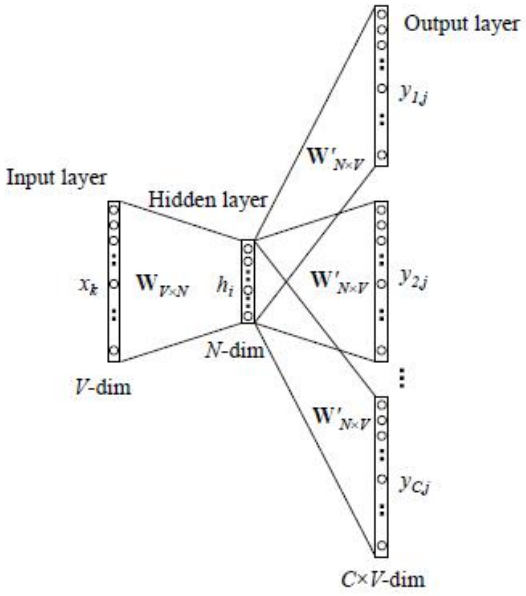


图 2-5 Skip-gram 模型训练

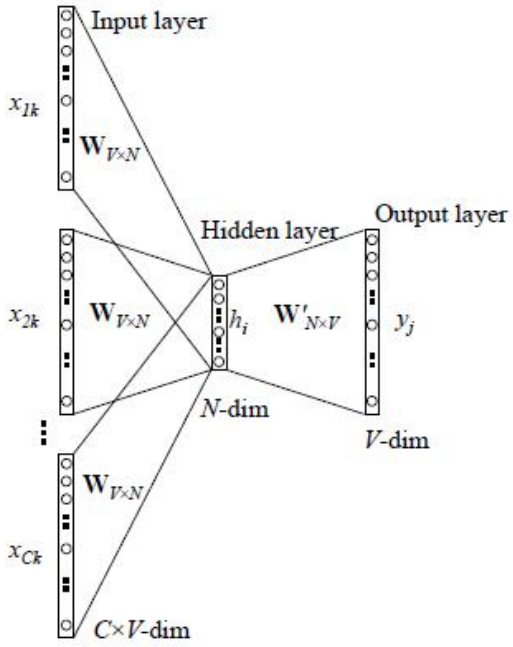


图 2-6 CBOW 模型训练

Word2Vec 训练后会生成语料库，后续生成特征矩阵时会通过该语料库输出分词后的数据对应的空间向量。并且训练 Word2Vec 模型时所用的数据量越大，停用词库越丰富，Word2Vec 模型生成的分布式词向量越准确，词间相关性更强[29]。Word2Vec 模型的实现步骤如下：

(1) 进行文本分词并去除停用词，将文本中的句子拆分为相互独立的单词，并还原词性；

(2) 遍历步骤(1)中的单词列表，统计词频并以此构造词典；

(3) 根据步骤(2)中的词频构造哈夫曼树形结构；

(4) 根据步骤(3)得到的结构计算节点的二进制码；

(5) 输入叶子节点的词向量，以中间节点词向量为参数，进行初始化；

(6) 选择一个模型训练词向量。

其训练复杂度[30]为：

$$Q = C \times \left(D + D \times \log_2 V \right) \tag{2-1}$$

其中，Word2vec 模型输入层的窗口的长度用 C 表示，词向量维度用 D 表示，训练语料中的词典的数目用 V 代表。

设有词典 $vocab = \{t_i | i \in 1 \cdots N\}$ 和文档 $d_i = \langle w_1, w_2, \cdots w_j \rangle$，N 为词向量维度，首先用 Word2Vec 模型训练词典 vocab，然后生成 d_i 中单词 w_i 的词向量，根据式 2-2 得到 d_i 的向量表示 $R(d_i)$：

$$R(d_i)=\sum_t word2vec(t)\ where\ t\ \in d_i \tag{2-2}$$

其中 $word2vec(t)$ 是单词 t 的词向量。

如果文本中单词的重要程度不都是相同的，可以由式 2-3 得到 d_i 加权向量表示 $weight\_R(d_i)$：

$$weight\_R(d_i)=\sum_t word2vec(t)\times w_t where\ t\in d_i \tag{2-3}$$

其中 w_t 表示词汇 t 所占权重。

2.4 逻辑回归

由多元线性模型推广而得的逻辑回归模型属于机器学习中的无监督学习算法类型，其实质是一个含单隐层的神经网络模型，可用来解决二分类问题和多分类问题。逻辑回归是一种广义线性回归模型，其实质是对多元线性回归模型 $y=w^Tx+w_0$ 做一个函数变换，该变换函数一般为 sigmoid 函数 $h(x)=\frac{1}{1+e^{-x}}$。函数 $h(x)$ 的有两条重要性质，一是预测结果可以作为分类概率，二是任意阶可导。函数 $h(x)$ 的图像如图 2-7 所示。

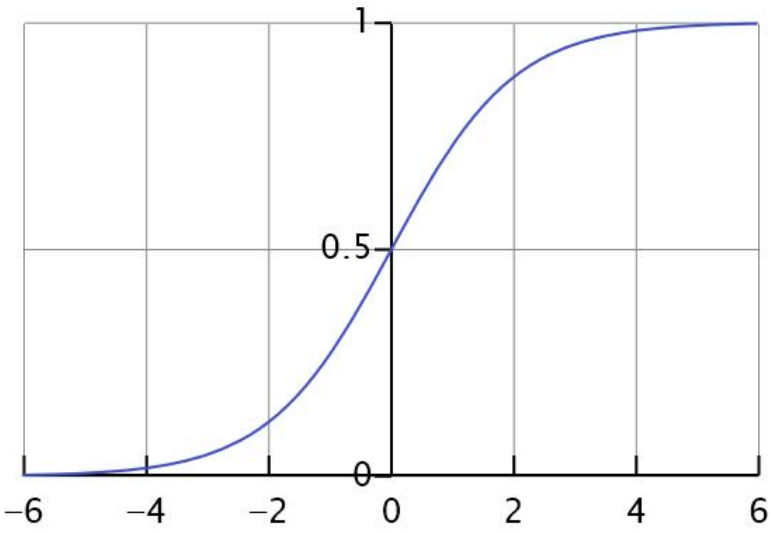


图 2-7 Sigmoid 函数

逻辑回归背后的含义是一组观察数据 $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ 以概率 $h(x)$ 独立同分布地产生：

$$p(y\mid x)=\begin{cases}h(x) & y=1\\ 1-h(x) & y=0\end{cases} \tag{2-4}$$

求解逻辑回归模型一般使用极大似然法，即已知一组观察数据

$\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$，通过求解参数（$w,w_0$），使得模型 $y=\frac{1}{1+e^{-(w^Tx+w_0)}}$ 观察到该组数据的可能性最大。其求解过程如下：

（1）随机产生一组参数(w,w_0)

（2）求解模型产生观察数据的概率：

$$\begin{aligned}L(w,w_0)&=p(y_1\mid x_1;w,w_0)\times p(y_2\mid x_2;w,w_0)\times...\times p(y_n\mid x_n;w,w_0)\\&=\prod_{i-1}^{n}p(y_i\mid x_i;w,w_0)\end{aligned}\tag{2-5}$$

（3）对公式（2-5）取对数得：

$$l(w,w_0)=\sum_{i=1}^{n}p(y_i\mid x_i;w,w_0)\tag{2-6}$$

（4）由公式（2-4）得：

$$p(y_i\mid x_i)=h[(2y_i-1)(w^Tx+w_0)]\tag{2-7}$$

（5）由公式（2-6）和公式（2-7）得：

$$l(w,w_0)=-\sum_{i=1}^{n}\ln(1+e^{-(2y_i-1)(w^Tx+w_0)})\tag{2-8}$$

逻辑回归求解即求一组参数(w,w_0),是的公式（2-8）取最大值。

（6）通过梯度下降法求得参数(w,w_0)使公式（2-9）取最小值：

$$E(w,w_0)=\sum_{i=1}^{n}\ln(1+e^{-(2y_i-1)(w^Tx+w_0)})\tag{2-9}$$

逻辑回归模型求解得损失函数[31]为：

$$J(\omega)=-\frac{1}{n}\sum_{i=1}^{n}[y_i\times w^T\times x_i-\ln(1+e^{w^Tx_i})]\tag{2-10}$$

2.5 随机森林

随机森林是集成学习算法的一种，它通过构造多个决策树并以投票方式得出模型分类预测结果。由于抗干扰能力强并且有能够平衡误差的优点，故随机森林在对缺失数据较多或不平衡的数据集分类时有良好的分类效果[32]。随机森林由多个决策性较弱的决策树构成，是一种常见分类器，每一个决策树的结构都类似一棵树，含有叶节点和非叶节点。常见构造决策树的划分方法依据有信息增益、信息增益比和 Gini 系数。决策树构成如图 2-8 所示。

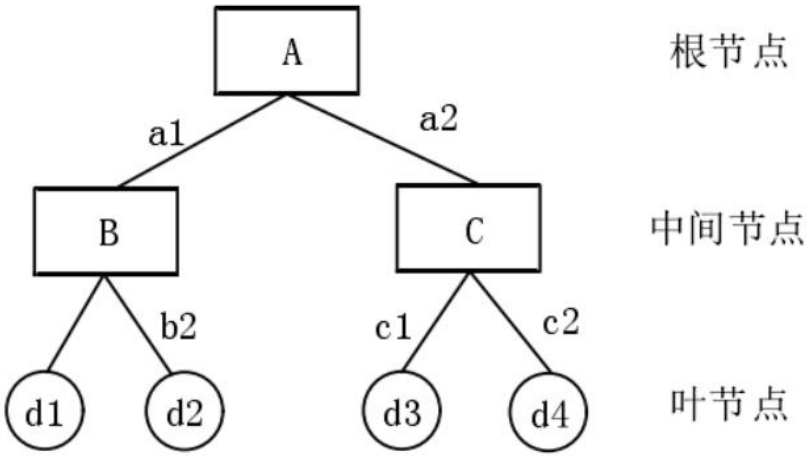


图 2-8 决策树构成图

图 2-8 展示了随机森林的训练和分类过程，随机森林的构造步骤如下：

（1）随机森林的构造过程包括两种随机，一种是使用自举重采样（bootstrapping）方法生成多组待使用数据，每次有放回地抽取 1 个样本并保留，进行 N 次后获得一个样本集，重复此过程直到样本集的数量达到目标要求，此方法即 bootstrapping 采样；

（2）第二种随机体现在弱分类决策树的构造过程中，具体表现为从输入数据所有维度中随机选取 m 个特征作为当前节点的候选特征，然后再从候选特征集中选择分类效果最好的特征作为决策特征；

（3）得到所需数目的决策树后，在对于输入的样本分类时，随机森林（RF）首先将该样本输入到每一个决策树获取其决策结果并进行结果统计，随机森林的分类结果即为统计频率最高的类别。

随机森林的性能可以用边缘函数和泛化误差评价[32]。假设随机森林中有 k 棵 CART 树 $f_1(x), f_2(x), \ldots f_k(x)$。训练样本集合 $T(X,Y)$，X 为样本特征，Y 为对应的类别。其边缘函数为：

$$mg(X,Y) = av_k\{I[f_k(X)=Y]\} - \max_{j \neq Y} av_k\{I[f_k(X)=j]\} \quad (2\text{-}11)$$

公式（2-11）中 I 为特征函数，j 为分类错误的样本所属的类别，av_k 是表示求均值。泛化误差为：

$$PE^* = P_{X,Y}[mg(X,Y)<0] \quad (2\text{-}12)$$

其中 PE^* 表示概率值。随机森林边缘函数为：

$$mr(X,Y) = P[f_k(X)=Y] - \max_{\substack{j \neq Y \\ j=1}} P[f_k(X)=j] \quad (2\text{-}13)$$

其中 $P[f_k(X)=Y]$ 为随机森林分类正确的概率，$\max\limits_{\substack{j\neq Y\\ j=1}} P[f_k(X)=j]$ 为将样本判断错误的概率的最大值。

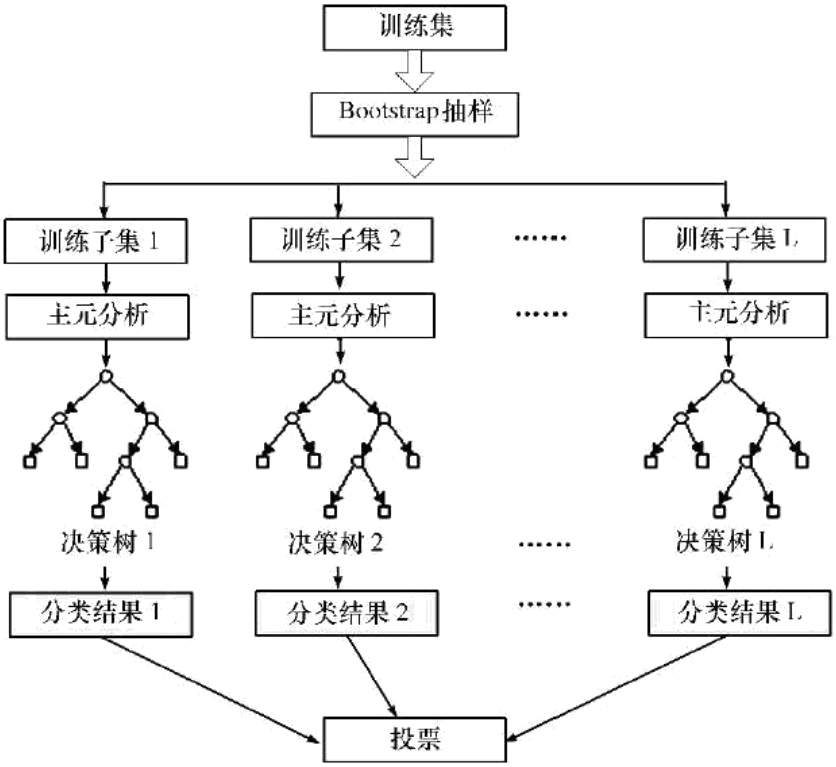


图 2-9 随机森林训练分类过程

2.6 TF-IDF 算法

词频-逆向文件频率（TF-IDF）是文本挖掘领域常用的加权算法。TF-IDF 算法可以计算词汇在语料库文档中的关键程度。如果某个单词频繁出现于某一文档中但极少在语料库其他文档中有所使用，表明该词语具有很好的区分文档的能力，可以作为语料库的部分关键词看待，此即为 TF-IDF 的内在思想。在 TF-IDF 算法的计算逻辑中，如果一个单词在某文档中多次出现，表明其对语料库更为重要，但若该词在语料库的很多文章中均存在，那么其对于语料库的重要程度将会降低。如某个语料库共 1000 篇文档，每个文档含有的单词数量均超过 1000，语料库中包含有单词“加速度”。在该语料库的某篇文章中，“加速度”出现次数越多，该词对语料库有更为重要的影响。但是当“加速度”只出现在语料库中的 40 篇文档中时比该词出现在 200 篇文档中时对于语料库来说，其更为关键。

计算某个词 w_i 的 TF-IDF 权重的过程[33]如下：

（1）计算词 w_i 的词频 TF：

$$TF_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2-14}$$

其中，$n_{i,j}$ 表示词 w_i 在文档 d_j 中的出现次数，TF_i 即为词 w_i 的频率。

（2）计算词 w_i 的逆向文件频率 IDF：

$$IDF_i = \log(\frac{|D|}{1+|j:w_i \in d_j|}) \tag{2-15}$$

其中，$|D|$ 表示语料库共有多少篇文档，$|j:w_i \in d_j|$ 表示有多少文档中含有单词 w_i。

（3）计算词 w_i 的 TD-IDF：

$$TF-IDF_{w_i} = TF_i \times IDF_i \tag{2-16}$$

TF-IDF 算法与 Word2Vec 算法结合生成用于机器学习的分布式词向量的过程如图 2-10 所示。

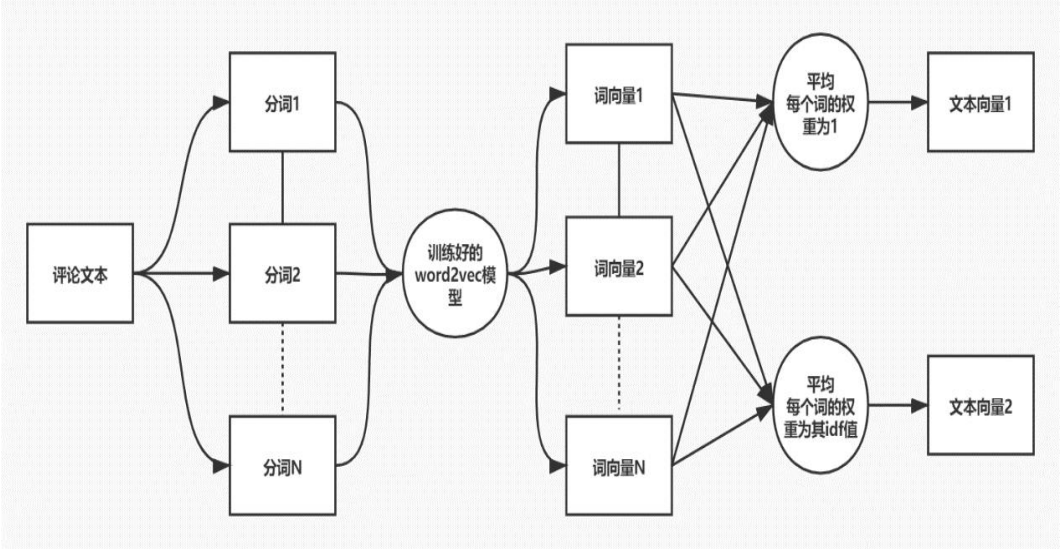


图 2-10 文本词向量的生成过程

2.7 潜在狄利克雷（LDA）模型

Blei 等人于 2003 年提出了可以用来挖掘文本语料库中潜在主题的 LDA 模型[34]，该模型也可通过概率模型生成语料库中的主题文档。LDA 模型的基础是词袋模型，在 LDA 主题模型中，语料库中的每一篇文档中的词汇都是由该文档的主题以一定的概率分布从词袋中产生。通过 LDA 模型生成文档时首先以一定的概率生成文档的主题，再通过文档主题以一定的概率产生该主题下的词汇，进而生成整篇文档。图 2-11 展示的是 LDA 主题模型示意图。

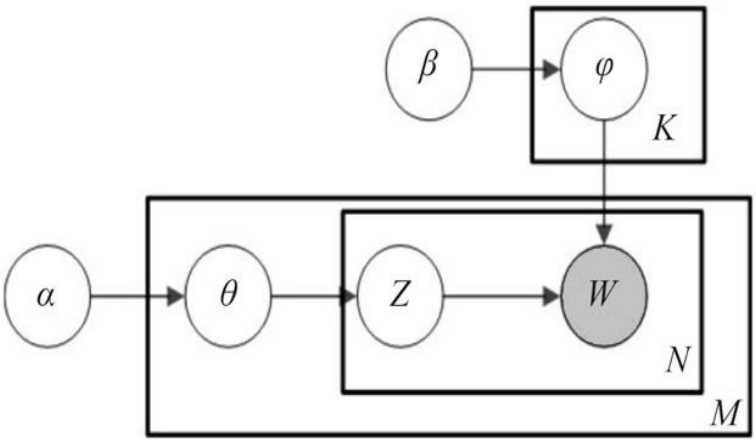


图 2-11 LDA 主题模型

其中，W 表示观察到的词语[35]，M 代表语料库中包含的文档数量。W 所属的潜在主题为 Z，N 表示语料库中总词汇量，K 表示某篇文档一共有多少个潜在文本主题，θ为服从超参数为α的狄利克雷分布的维度为 $M*K$ 的矩阵，用以表示文档的主题分布。φ为超参数β的主题的词汇狄利克雷分布，为 $K*N$ 矩阵。

LDA 主题模型的训练过程为：

（1）对语料库中的每个词随机赋予主题 K；

（2）对语料库中的每个词进行吉布斯采样，重新赋予所属主题；

（3）重复步骤（2）直至吉布斯采样收敛；

（4）计算“主题-词汇”的共现矩阵。

LDA 主题模型的生成过程为：

（1）生成某篇文档 d 的主题分布 θ_d，该分布由超参数为α的狄利克雷分布产生；

（2）从（1）中生成的主题分布中生成文档的中的第 j 个词的分布 $Z_{d,j}$；

（3）从超参数为β的狄利克雷分布中生成主题 $Z_{d,j}$ 的词分布 $\phi_{Z_{d,j}}$；

（4）从词分布 $\phi_{Z_{d,j}}$ 中生成文档 d 的第 j 个词 $w_{d,j}$。

3 数据获取和预处理

3.1 数据获取

本研究所使用的数据为汽车之家和易车网等大型网络平台的新能源汽车在线评论。因为数据量巨大，维度众多，人工收集数据耗时长，故笔者使用Python 爬虫技术在许可范围内进行数据收集整理。新能源汽车诞生发展多年，受到越来越多消费者的青睐，汽车之家和易车网在新能源汽车在线评论板块也积累了大量数据，但二者提供的数据形式有一定差异。易车网在新能源汽车的空间、外观、内饰、动力、操控、舒适性以及续行方面均有评分数据，分值范围为 1-10 分，所提供的文本评论为非规范化数据，消费者对汽车的使用体验直接做出完整的评价。汽车之家提供的口碑数据比易车网的数据规范的多，包括在空间、动力和操控等八个维度的分值范围为 1 至 5 的用户单独评分数据，同时汽车之家也提供了在这八个维度上的规范的用户文本评论数据。汽车之家提供的不同维度的评论数据在后续新能源汽车不同维度上的主题挖掘分析中更有应用价值。

笔者使用了 Python 爬虫的 requests 下载器对网页数据进行抓取。在获取汽车之家和易车网的在线评论数据时，笔者均遇到了以字体混淆替换为主要方式的反爬虫技术，此反爬方式主要是使用自定义的字体文件替换网页中的部分字体数据使其在非正常浏览器访问情况下不可读，也无法直接获取可分析的数据。以汽车之家口碑数据为例，通过检查其口碑网页源代码，发现其评论数据中存在大量不可见文本，进一步分析得知其嵌入了自定义的 myfont 字体，且每次访问时其字体编码动态变化。笔者的解决方案是每次访问其口碑数据时都通过网页源代码中的自定义字体 url 动态解析自定义字体文件，替换原网页中的不可见文本，进一步使用 XPath 获取所需数据。通过编写 Python 爬虫代码，笔者共爬取了汽车之家 2014 年至 2022 年的新能源汽车规范化的口碑数据共 95480 条，获取了易车网的非规范化数据共 4790 条以进行下一步数据清洗和分析工作。部分爬虫代码见附录。

3.2 数据预处理

通过爬虫批量获取的数据存在大量缺失值、重复值等，无法直接用于后续数据分析和模型构建，需对数据进行预处理使其成为准确便于分析的数据形式。本文在将数据进行清洗之后，使用 Python 的 jieba 包对评论数据进行了

分词处理，并加入了汽车行业专有名词及停用词，使分词数据更加准确，可用性更强。在对评论数据进行分词处理后，本文使用 word2vec 模型将分词后的评论数据进行分布式向量化表示及构建特征矩阵，为后续模型构建提供数据支持。

3.2.1 数据清洗

数据清洗能够去除脏数据，提高数据质量，便于后续数据分析处理和模型构建。本文数据处理的工具为 excel 和 Python 语言。首先利用 excel 对数据进行了初步处理，删除了数据中包含的对后续分析无用的字段，对所有数值型属性中的明显错误数据，如评分数据中的负值和超过评分上限的值，对其进行了删除操作，并根据汽车之家提供的最满意和最不满意两个维度的数据新增两列数据用以表示该条数据的用户情感倾向。本文将用户情感倾向分为负向情感和正向情感两类，分别用 0 和 1 表示。

本文使用了 excel 和 Python 程序相结合的方法对爬取的数据进行了处理。字数过少的用户评论可能存在代刷等情况，或者在后续分词时加入停用词后获得的词汇过少或无词汇，导致该条数据在使用 word2vec 进行词向量转化时无实际价值并使得数据处理量过大，故首先使用 Python 的 DataFrame 数据结构将每个维度字数少于 20 的用户评论进行删除。然后对于各个维度的评分数据利用 Python 的 Imputer 包使用均值进行空值填充，再删除包含空白文本评论数据的数据行。因为后续使用 word2vec 模型进行分布式词向量的转化时采用的是连续词袋模型（CBOW），为了防止将单词与数字或字母的词向量相似度过高，故在进行数据预处理时对所有文本评论数据中的字母和数字进行了剔除。经过数据清洗后，汽车之家的可用评论数据由 95480 条下降到了 64581 条，易车网的可用评论数据由 4790 条下降到了 3942 条。虽然数据量有所减少，但是其数据是不存在空值、缺失值及重复值等的更高质量的数据。

3.2.2 数据分词

本文采用 Python 的 jieba 分词进行中文文本的切分并加入多种专有词库和停用词库以获得更好的分词结果。在进行文本分词前，首先下载百度和搜狗的汽车行业专有名词词库，但是由于百度词库的文件类型为.bdict，搜狗词库的文件类型为.scel，无法直接导入 jieba 分词中作为专有名词使用。故首先编写代码将两种不同类型的文件解码合并为 txt 文件，并将预处理后的爬虫数据中的汽车品牌名去重后加入 txt 文件作为专有名词的一部分导入 jieba 分词。

专有名词的引入对文本切分效果有明显提升，如在未加入专有名词前对句子“广汽埃安的车型非常丰富”的分词结果为“广/汽/埃/安/的/车型/非常/丰富”，加入专有名词后的分词结果为“广汽埃安/的/车型/非常/丰富”，文本切分更加精确。停用词的加入对于减少数据冗余，提高文本特征的质量有很大作用。本文下载并合并了四种停用词库，分别是百度停用词库 baidu_stopwords.txt、哈工大停用词库 hit_stopwords.txt、四川大学机器智能实验室停用词库 scu_stopwords.txt 和中文停用词表 cn_stopwords.txt。通过停用词减少了分词中的无用词汇。如在未加入停用词前对句子“哈哈哈，我最喜欢的就是它的外观了”的分词结果为“哈哈哈/，/我/最/喜欢/的/就是/它/的/外观/了”，加入停用词后的分词结果为“最/喜欢/外观”，相比于未加入停用词前的分词结果更加简洁，对于后续分布式向量的转化时保留词语间语义相关性有重要意义。

在准备好专有词汇和停用词库后，下一步是对预处理后的爬虫数据进行文本划分。本文对汽车之家提供的规范化的文本数据的八个维度及最满意和最不满意维度数据分别进行了分词并进行词频统计，将词频统计结果保存为 csv 文件以备后续使用。表 3-1 展示的是分词后各个维度词频统计排序后前 5 个单词及其频数。此外，依据词频统计结果绘制了最不满意和最满意两个维度的词云图，如图 3-1 和图 3-2 所示。

表 3-1 各维度词频统计排序前 5 单词及其词频

维度	前 5 个单词及其词频
操控	（方向盘,44798）（感觉,23890）（刹车,22287）（转向,15384）（驾驶,14235）
电耗	（充电,28720）（续航,27626）（公里,21840）（跑,21099）（能耗,18859）
内饰	（设计,24651）（感觉,15946）（不错,13818）（喜欢,13668）（做工,10802）
外观	（设计,35108）（喜欢,21132）（感觉,15928）（好看,15267）（车身,15206）
动力	（加速,31672）（起步,25388）（模式,20167）（高速,17830）（提速,17416）
空间	（后排,43708）（后备箱,31301）（乘坐,21314）（宽敞,17644）（座椅,17267）
舒适性	（座椅,55373）（舒服,23644）（不错,23549）（感觉,22879）（空调,18075）
最满意	（外观,27389）（车子,22867）（感觉,20190）（动力,19855）（喜欢,14727）
最不满意	（空间,9855）（续航,9110）（内饰,8759）（充电,7457）（座椅,6420）

图 3-1 最满意维度词云图

图 3-2 最不满意维度词云图

3.2.3 分布式词向量转化

本文采用 Word2Vec 方法将文本转化为空间数值向量，为后续机器学习算法提供数据基础。对于 Word2Vec 模型训练而言，训练数据集越大，模型训练效果越好。因此，本文将易车网 3942 条整体评价数据和汽车之家 64581 条数据均代入模型进行训练，而不是仅使用后续生成特征矩阵的数据。其中汽车之家每条数据包含空间、内饰、动力等八个维度以及最满意和最不满意维度

的数据，故用以训练 Word2Vec 模型的数据共 649752 条。在数据预处理阶段，首先对获取的评论数据进行清洗，然后剔除文本评论中的数字和字母。之后借用 Python 的 jieba 分词在加入专有名词和停用词的条件下对易车网评论数据和汽车之家所有文本评论维度数据进行分词并将数据分别保存为不同的 csv 文件以备后续使用。在本步骤中，首先将预处理阶段生成的所有 csv 类型维度分词数据合并，并以 list of list 格式将所有数据保存为 json 文件以便模型调试时高效重复使用。

训练数据生成后，本文调用了 Python 的 gensim 模块的 Word2Vec 模型进行分布式词向量训练。由于本文使用的数据集体量不大，故训练时选择了词袋模型（CBOW）。在本文进行模型训练时的部分代码为 model=Word2Vec(sentences=list_all,sg=0,size=300,window=5,alpha=0.001,min_count=5,hs=0,negative=10,Iter=30,cbow_mean=1)，调试的参数如表 3-2 所示：

表 3-2 Word2Vec 模型训练参数

参数	取值	备注
sentences	list of list 对象	输入数据 list of list 对象
sg	0	sg 取 0 表示使用 Word2Vec 的 CBOW 模型
Size	300	分布式词向量维度
window	5	所取上下文单词个数
alpha	0.001	学习率
min_count	5	设置最少词频，词频过少的单词不导入模型训练
hs	0	hs 取 0 表示使用 negative sampling 负采样加速方法
negative	10	负采样个数
Iter	30	模型迭代次数
cbow_mean	1	cbow_mean 取 1 表示上下文词向量采用均值计算

模型训练完成后将其保存为.model 文件以备后续调用。完成训练后使用 model.wv.most_similiar 方法可以查看与某一单词含义相近的单词，表 3-3 展示了与空间、动力、内饰、满意和加速最相近的 5 个单词及其相似度。Word2Vec 模型训练出的词向量在数字空间中的距离和实际文本中的相似度具有强相关性，此性质在表 3-3 中也有所体现。

表 3-3 单词相似度

单词	最相近的 5 个词及其相似度
空间	('宽敞',0.788),('宽松',0.707),('宽裕',0.695),('前后排',0.688)，('后备箱',0.6500527262687683)
动力	('提速',0.732),('起步',0.714),('强劲',0.712),('加速',0.695),('充沛',0.682)
内饰	('高档',0.725),('用料',0.694),('材质',0.684),('配色',0.675),('中控',0.652)
满意	('挺不错',0.668),('优秀',0.606),('还行',0.589),('不错',0.557),('喜欢',0.531)
加速	('提速',0.915),('加速度',0.806),('输出',0.741),('超车',0.707),('动力',0.695)

训练好的 Word2Vec 模型可以可以查看某个词对应的空间向量，如单词

“加速”对应的向量为“（-2.4035823 , -0.27448818 ,2.3305948 , -0.53101707 , -2.6116195 .....0.6104133 , -0.8369216 ,0.7508109 , 2.1985836 , 2.087963）”，由于词向量维度为 300，故展示部分数据。通过模型的方法 model.wv.similarity 可以计算将词汇转化为分布式词向量后两个单词之间的相似度，例如单词“动力”和“加速”的词向量相似度为 0.6955，“动力”和“空间”的词向量相似度为-0.0409。通过导入所有清洗后的全文本数据进行模型训练，得到的 Word2Vec 模型保留了一定的单词间语义相关性，训练效果显著。

4 在线评论分析和模型构建

4.1 新能源汽车在线评论分析

4.1.1 新能源汽车维度评论分析

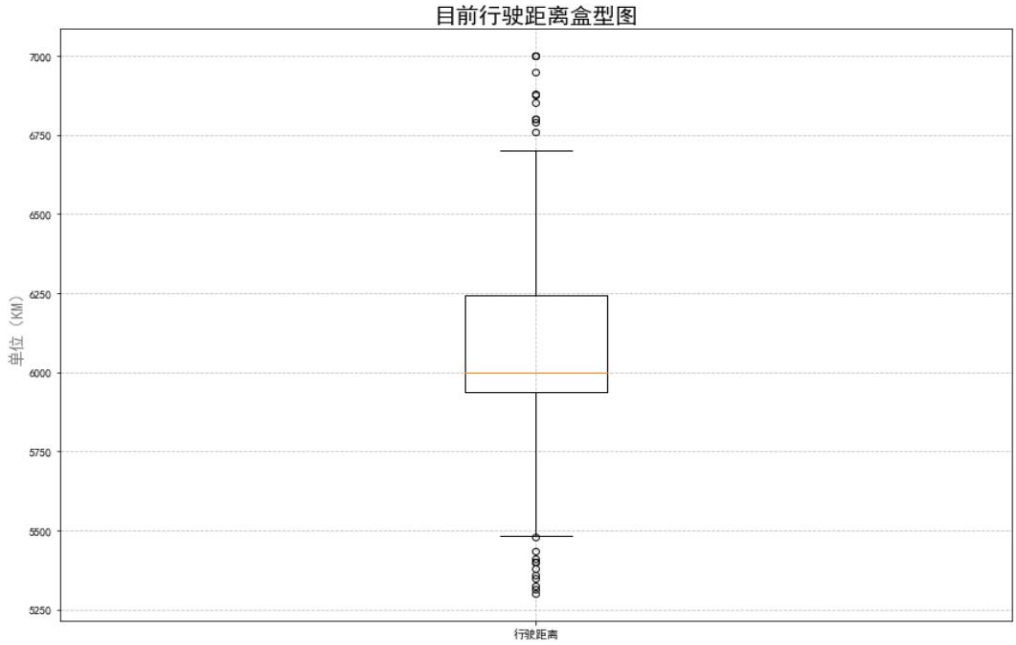


图 4-1 用户评论时行驶距离箱线图

本文采用的是通过爬虫获取的汽车之家和易车网 2014 年至 2022 年的新能源汽车用户在线评论数据，经过数据清洗后用于本部分评论整体分析以及后续建模分析。图 4-1 展示的是用户评论时已经购买的汽车的行驶里程的箱线图。本图展示了分析所用数据中的用户评论时的行驶里程情况，用户最大行驶里程为 7000 公里，最短里程为 5300 公里，所有用户行驶里程的平均值为 6079 公里，且行驶里程数据集中在 5450 到 6700 公里之间。由此可见，用户在汽车之家和易车网评论模块分享新能源汽车的用车体验和主观感受时对其所评论的新能源汽车的品牌和车型已经有一定的了解，其在线评论数据对新能源汽车的文本分析有一定的真实性和可靠性。

图 4-2 用户评分雷达图

图 4-2 展示的是汽车之家 64581 条数据中空间、内饰等八个维度用户评分均值的雷达图。由雷达图可知，用户最满意新能源汽车的外观，对其的评分均值接近 10 分。用户最不满意的是新能源汽车的内饰以及电耗情况。由此可见，新能源汽车虽然在外观动力等方面表现抢眼，但是内饰是其一大诟病。电能作为新能源汽车的动力来源，与其综合续航里程直接相关，也是许多消费者在考虑购买燃油车和新能源车时可能倾向于燃油车的重要因素。新能源汽车诞生发展多年，其续航能力虽然得到了很大提升，蔚来 ET7 的纯电续航里程甚至达到了 1000 公里，但是大多数新能源汽车的续航里程还是较短，且在冬季气温低的情况下续航里程降低较多，导致用户的里程焦虑更甚。新能源车企应该在汽车内饰升级和续航提升上投入更多资源，着力改善新能源汽车的不足。

4.1.2 新能源汽车销量时间趋势分析

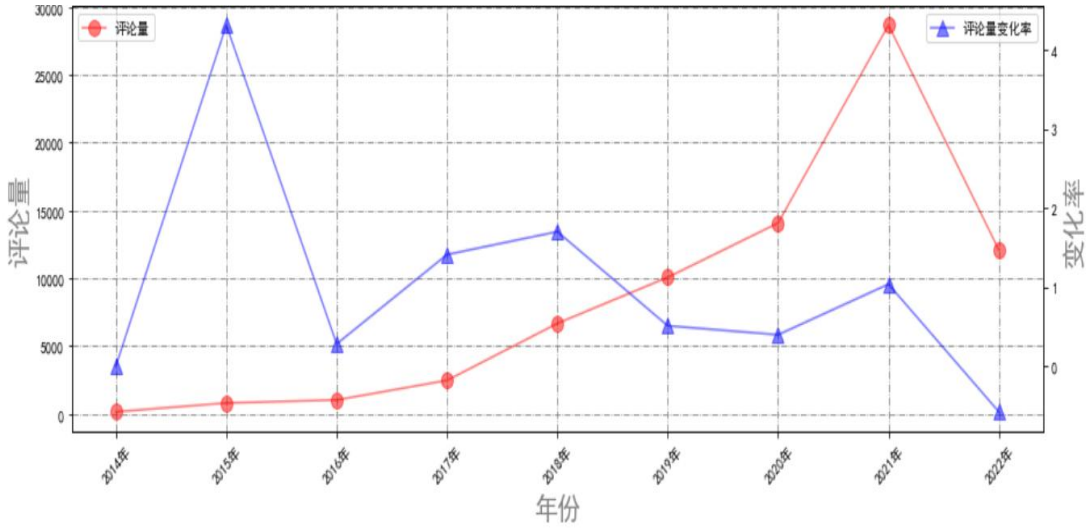


图 4-3 2014-2022 新能源汽车评论量及其变化率

2014 年之前我国新能源汽车发展缓慢，消费者对新能源汽车缺乏了解，用户评论数据少且质量不高，故本文所使用的用户评论数据的时间间隔为 2014 年到 2022 年。图 4-3 展示的是 2014 年到 2022 年新能源汽车评论数量及其变化趋势，进而反映 2014 年到 2022 年我国新能源汽车的销量情况及其变化率。由于 2022 年的评论数据只有第一季度，故其评论数量增长率为负值。从图中可以看出 2014 年至 2015 年评论数量增长率为 432%，与第一节研究背景中图 1-1 展示的新能源汽车销量增长率一致。从 2014 年到 2022 年，我国新能源汽车市场发展迅速，增长率处于较高水平，评论数据量由 2014 年的 150 条到 2021 年的 28668 条，由此反映出这几年间我国新能源汽车市场发展迅速，这与我国的新能源汽车政策息息相关，也与人民的环保意识增强有关。新能源车企要抓住发展机遇占据市场，提高消费者满意度，就要从消费者痛点出发，进行汽车升级优化。

4.1.3 不同省份新能源汽车销量情况分析

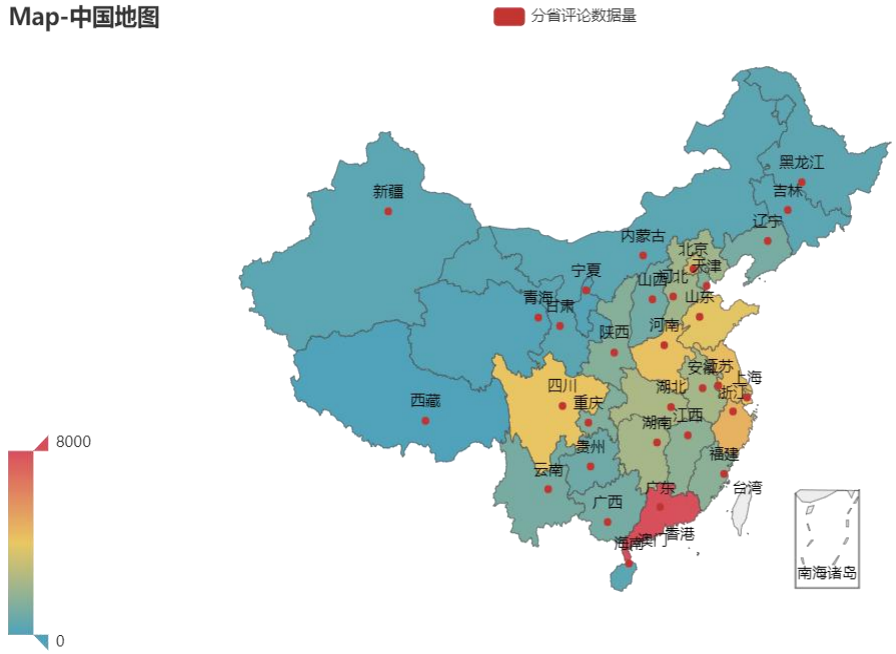


图 4-4 不同省份新能源汽车销量分布图

近几年来，我国新能源汽车市场在近几年里发展迅速，但是地区之间相对不平衡。图 4-4 展示的是我国不同省份新能源汽车销量分布图，从图中可以看出东南沿海地区新能源汽车发展迅速，而广阔的西部地区发展缓慢。相比于西部地区，我国东南地区经济发展水平高，人口稠密，消费潜力大。且对于新能源汽车而言，东南部地区人口众多，人口出行量大，充电网点更加密集，消费者对于新能源汽车有更深的认知，其发展有相比于西部地区有天然优势。

由评论数据分析可知，我国新能源汽车评论数排名前五的省份分别是广东、浙江、江苏、河南和上海，均为东部省份。其中广东省的评论数最多，达到了 13.3%，由此间接反映出广东省新能源汽车销量最大。评论数目最少的省份是西藏，仅占总数的 0.017%，由此也可以看出东西部之间新能源汽车发展的差距。

4.1.4 新能源汽车价位分析

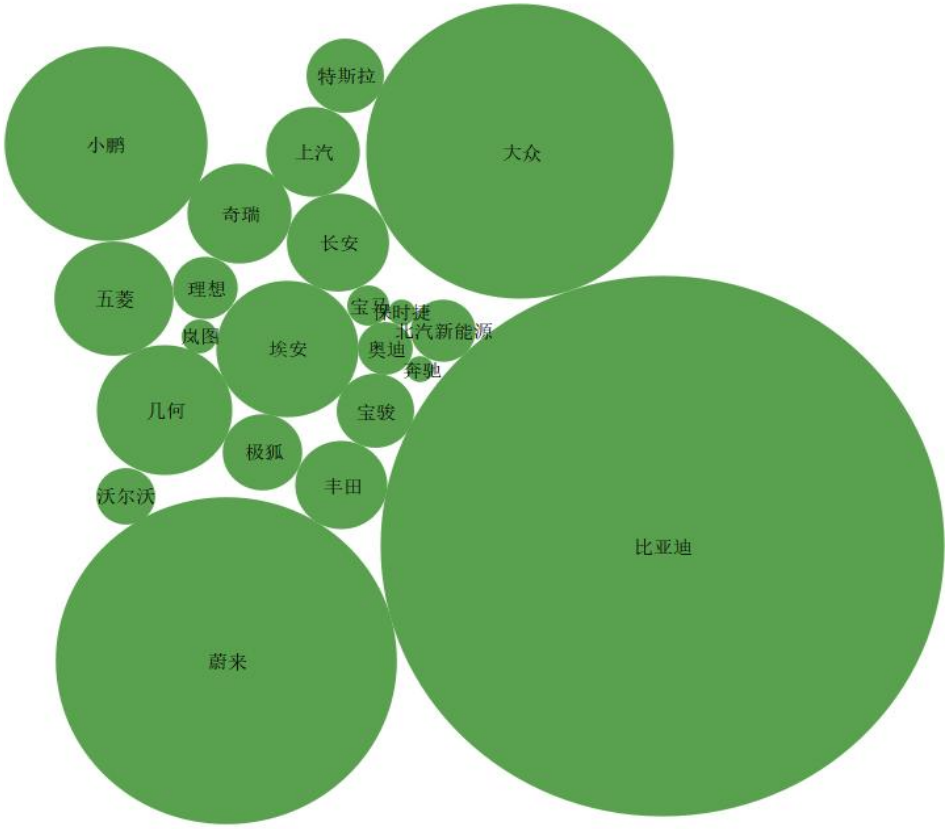


图 4-5 新能源品牌销量气泡图

图 4-5 展示的是新能源汽车品牌销售量气泡图，从图中可以看出销量排在我国新能源汽车市场中前四的品牌分别是比亚迪、蔚来、大众和小鹏，其中比亚迪、蔚来和小鹏均为国产新能源汽车品牌。在本文所用数据中，比亚迪的销量最大，占据总销量的 45.5%。

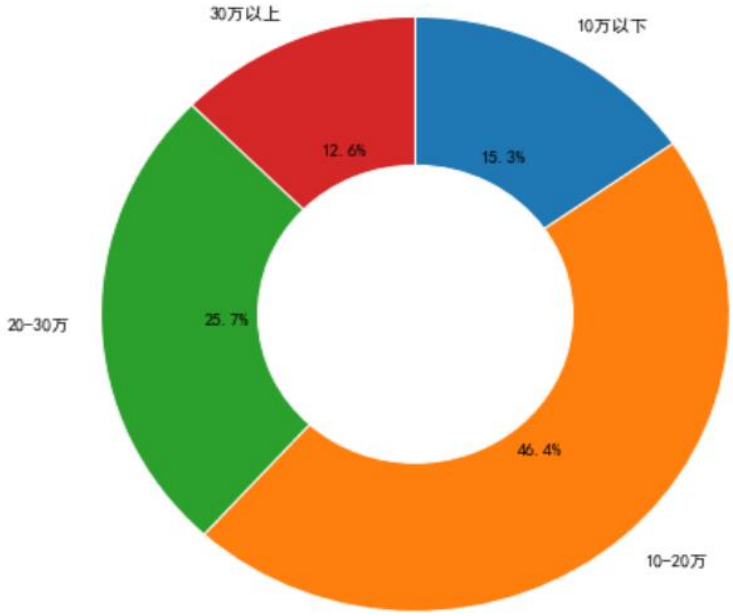


图 4-6 新能源汽车价位比例图

图 4-6 展示的是不同价位的新能源汽车评论数占比，本文将汽车价格分为四个等级，分别是 10 万以下、10-20 万、20-30 万和 30 万以上。从图中可以看出，销量最大的价位是 10-20 万，这个价位的新能源汽车性价比高，既能满足用户日常出行需求，又在多数用户能接受的价格范围之内，符合大众购车现状。30 万以上的新能源汽车销量最低，车型也较少，属于豪华型的新能源汽车，其中保时捷 Taycan 的价格甚至超过百万。10 万元以下的新能源汽车中较受欢迎的有奇瑞小蚂蚁和五菱宏光 MINIEV 等。

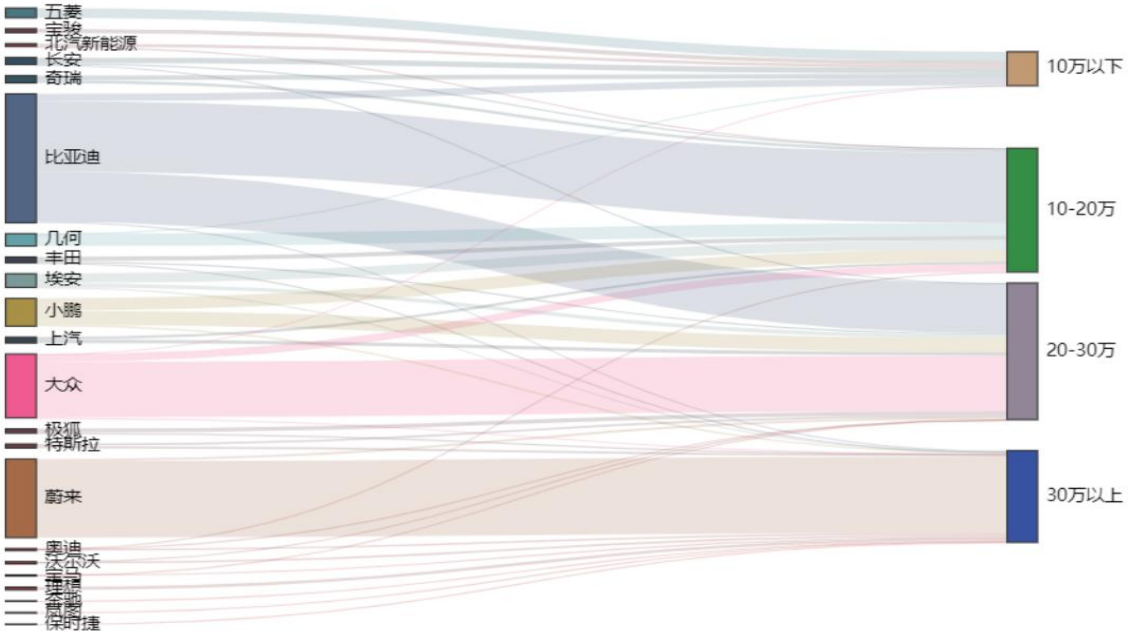


图 4-7 新能源汽车品牌价位桑基图

图 4-7 展示的是新能源汽车品牌和价位的桑基图，与上一节相同，本节仍将汽车价位分为四个等级。可以看出，比亚迪的销量最多，其次是蔚来、大众和小鹏。从图上知，比亚迪的汽车价位大多在 10-20 万和 20-30 万两者之间。蔚来汽车的价位基本处于 30 万以上。大众的价位在 20 至 30 万之间。10 万以下价位的新能源汽车品牌主要有五菱宏光 MINIEV、北汽新能源和奇瑞小蚂蚁等。30 万以上价位的主要有蔚来、保时捷和沃尔沃等豪华车型。

可见，桑基图左右两边展示的信息与图 4-5 和 4-6 的信息相一致，且不同品牌的新能源汽车价位分层明显，消费者进行购车选择时可以先根据预算确定所购买的品牌范围，再进行横向比较。

在本部分中，本文使用了所获取数据的多个属性，从多个角度对新能源汽车评论进行整体分析得出了相应的结论并在下一步研究中对评论数据进行更深层次的信息挖掘。

4.2 模型构建

4.2.1 特征矩阵生成

本文首先清洗所抓取的汽车之家和易车网的在线口碑数据用于后续分析及建模。然后结合 Word2Vec 和 TF-IDF 算法计算单篇口碑文档的空间向量并生成特征矩阵用以训练和评估机器学习算法。

本文将汽车之家最满意和最不满意维度数据用于机器学习分类算法。但为了增加数据量以提高 Word2Vec 模型的训练效果，用于训练 Word2Vec 模型的数据量和机器学习算法训练和测试的数据集不一致。通过合并汽车之家空间、内饰等八个维度的数据、最满意和最不满意维度的数据以及易车网的整体评论数据，使得用于训练 Word2Vec 模型的数据量成倍增加，且训练模型时词向量设置为 1*300 的维度。

建立特征矩阵时，首先合并汽车之家最满意和最不满意维度的数据，进而使用 Python 的 jieba 分词进行文本划分并引入专有词库和停用词库以获取更高质量的分词结果。然后从训练好的 Word2Vec 模型中提取所有评论分词的空间向量。进而采用简单平均或加权平均的方法由分词词空间向量计算每篇评论口碑的整体向量。本文采用 TF-IDF 算法作为词向量的权重进行加权平均。由于进行评论文本的分词结果的词向量转化时未进行去重，故所得结果与引入词频（TF）的结果一致，故由词向量计算文本向量时使用 IDF 作为权重的结果即为使用 TF-IDF 作为权重的结果。本文计算了所用数据的分词结果的每个词的逆向文档频率（IDF）作为权重进行词向量加权得到评论文本词向量，

并将所有文本的词向量结果合并最终得到维度为 129162*300 的特征矩阵，其中 129162 是评论文本数目，300 为单个词向量的维数。

4.2.2 模型建立及分析

本文首先获取了数据清洗后汽车之家最满意维度和最不满意维度的数据及其情感标识数据各 64581 条。合并数据后加入专有词汇和停用词进行分词处理，进而使用以全数据进行训练的 Word2Vec 模型进行分布式词向量的转换，并计算分词结果中的每个词的 TF-IDF 值，以其作为权重对每一篇评论文本分词结果的词向量进行加权平均获得每篇评论的文本向量，然后将文本向量转化为特征矩阵并去除空值，得到维度为 129162*300 的机器学习算法所需数据。进行机器学习模型训练和测试前，先使用 sklearn.model_selection 中的 ShuffleSplit 方法打乱特征矩阵中的数据条并将其划分为 70%的训练集和 30%的测试集。

对于文本分类问题，所用的衡量指标主要有准确率、精确率、召回率和 F1-score 等，此类指标均可由混淆矩阵经过一定的计算得到。混淆矩阵为 2*2 的矩阵，由 TP、FP、TN 和 FN 构成。表 4-1 为混淆矩阵构成。

表 4-1 混淆矩阵

真实类别	预测类别	
	正例	负例
正例	TP	FN
负例	FP	TN

基于混淆矩阵可计算文本分类效果评价指标。其中准确率的计算如(4-1)，精确率的计算如（4-2），召回率的计算如（4-3），F1-score 的计算如（4-4）。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4-1}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{4-2}$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \tag{4-3}$$

$$F1 - score = \frac{2P * R}{P + R} \tag{4-4}$$

除了上述衡量指标，本文还采用了 AUC 值和 ROC 曲线来评价使用的文本分类模型的效果。ROC 曲线的横坐标为 FPR，纵坐标为 TPR，从大到小设置分类模型的概率阈值，获取其 FPR 和 TPR，绘制的曲线即为 ROC 曲线。ROC 曲线越往左上方靠近，即曲线右下方的面积越大，表示模型的分类效果

越好。

本文将划分出的占比 30%的测试集输入到训练好的逻辑回归模型，解析出了 92.172%的分类准确率。在对该逻辑回归模型进行交叉验证时发现其准确率均为 92%左右。通过调用 sklearn.metrics 的 classification_report 得到逻辑回归模型的精确率、召回率和 F1-score 如表 4-2 所示。在正向情感样本上训练的逻辑回归的精确率为 0.93，召回率为 0.91，F1-score 为 0.92；在负向情感上训练的逻辑回归模型的精确率为 0.92，召回率为 0.93，F1-score 为 0.92。

表 4-2 逻辑回归模型评价指标

	precision	recall	f1-score
正向情感	0.93	0.91	0.92
负向情感	0.92	0.93	0.92

逻辑回归模型的混淆矩阵如图 4-8 所示。可知，训练的逻辑回归分类模型正向情感样本中，预测正确的样本共 16870 条，预测错误的样本为 1611 条；负向情感样本中预测正确的样本为 17572 条，预测错误的样本为 1314 条。

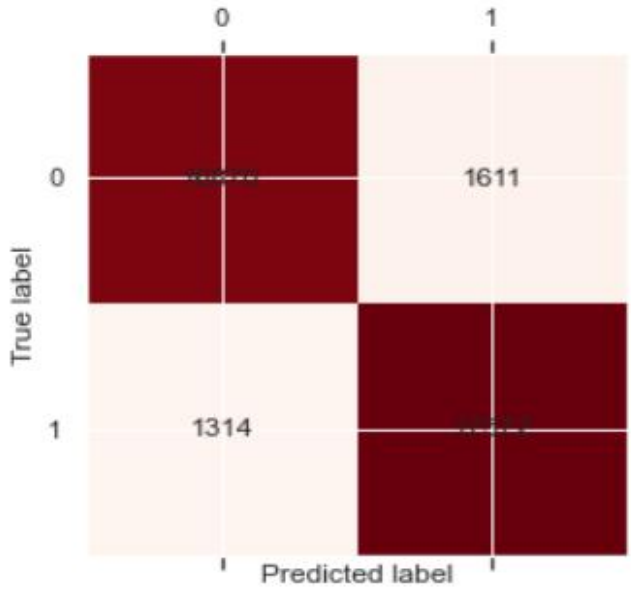
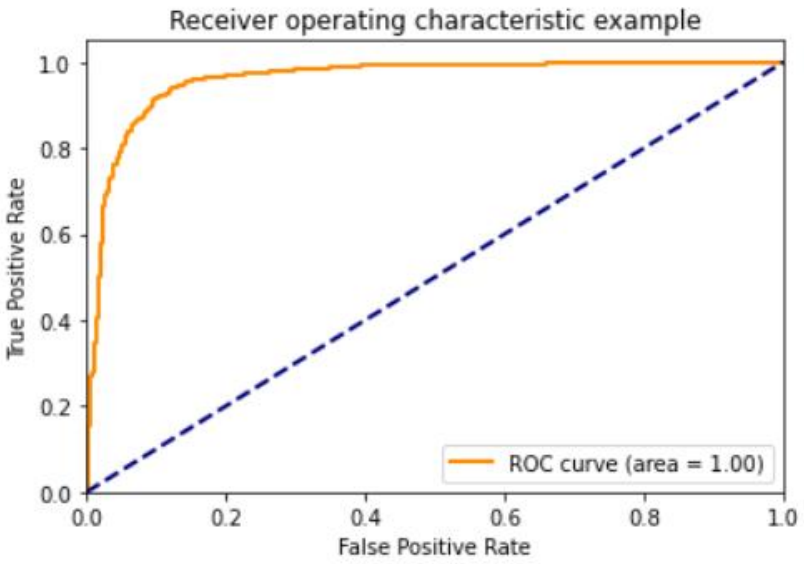


图 4-8 逻辑回归模型混淆矩阵

逻辑回归 ROC 曲线如图 4-9 所示，通过计算得知其 AUC 值为 0.921，可知其分类效果良好。



4-9 逻辑回归 ROC 曲线

在使用处理好的数据进行随机森林训练时，通过预训练得到了部分参数的优化选择。本文首先通过改变模型最大样本数的取值进行随机森林预训练并计算和展示模型训练效果如图 4-10 所示。由图知随机森林的最大样本数设置为 35 较合适。然后通过变更随机森林中每棵决策树的最大深度得到了图 4-11 的结果，由图分析可知决策树的最大深度设置为 16 较合适。

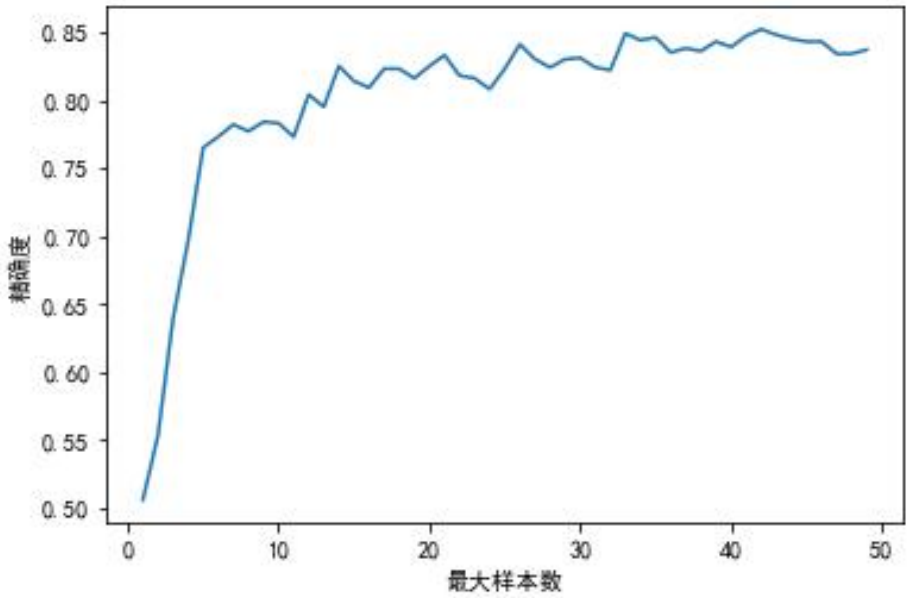


图 4-10 最大样本数不同时的模型效果

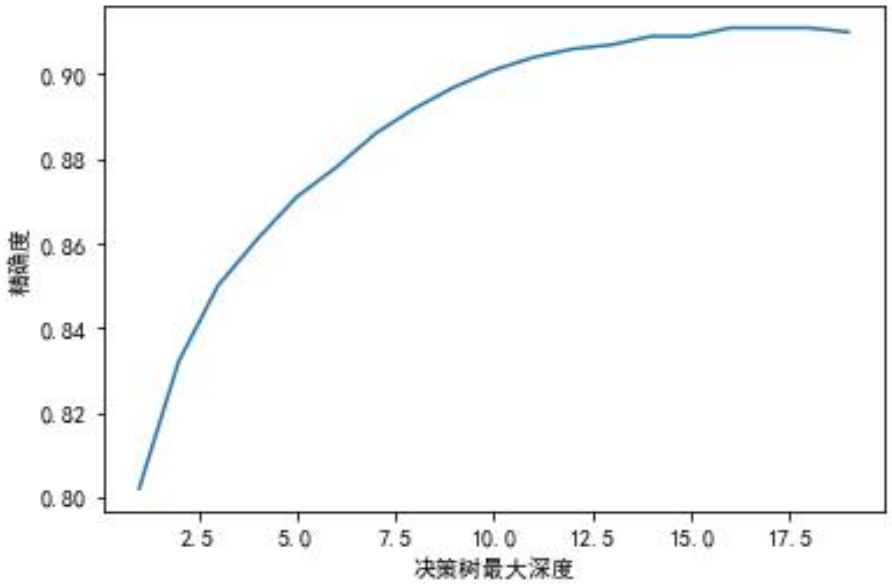


图 4-11 最大深度不同时的模型效果

由以上分析，本文将决策树最大样本数置为 35，最大深度置为 16 调用机器学习库 sklearn. en semble 的 Random Forest Classifier 训练随机森林模型。通过调用 sklearn. metrics 的 classification report 计算所训练的随机森林模型的精确率、召回率和 F1-score，如表 4-3 所示。由表可知训练的随机森林在正向情感和负向情感样本上的精确率为 0.87，召回率为 0.87 和 0.88，F1-score 为 0.87。

表 4-3 随机森林模型评价指标

	precision	recall	f1-score
正向情感	0.87	0.87	0.87
负向情感	0.87	0.88	0.87

随机森林模型的混淆矩阵如图 4-12 所示。可知，训练的随机森林分类模型正向情感样本中，预测正确的样本共 16066 条，预测错误的样本为 2415 条；负向情感样本中预测正确的样本为 16547 条，预测错误的样本为 2339 条。通过计算得知随机森林模型 AUC 值为 0.905，可知其分类效果良好。

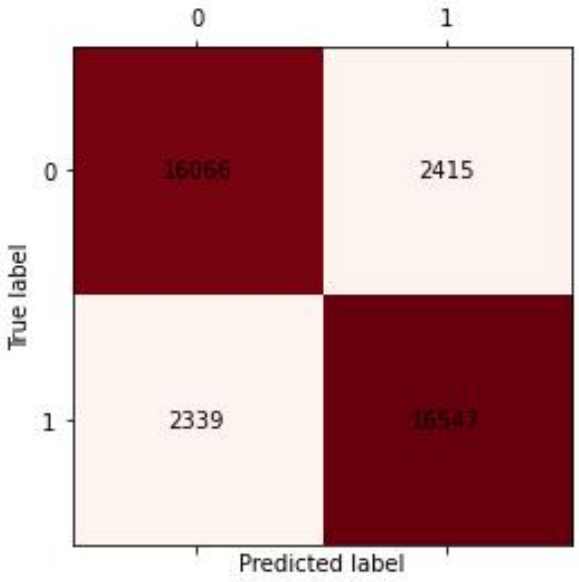


图 4-12 随机森林模型混淆矩阵

对比逻辑回归模型和随机森林模型的分类效果可知，在使用不同的衡量指标时，逻辑回归模型分类效果均比随机森林更优，可知逻辑回归模型比随机森林模型更适合本文对新能源汽车在线评论的情感分类。

4.2.3 LDA 主题分析

本文调用了 gensim. models 的 LdaModel 方法进行 LDA 主题模型训练，在进行模型训练前需确定模型的主题个数、α值和β值。通常情况下，将α值和β值设定为等值，本文进行主题模型训练时使用自动设定的方法确定α值和β值。在进行 LDA 模型训练前先通过绘制模型预训练的困惑度和一致性曲线，从曲线中分析判断合适的主题个数。在主题挖掘时训练出的模型在识别文档中的潜在主题时具有概率上的不确定性，通常用困惑度衡量。困惑度越低，即不确定性越小，该模型的效果就越好[36]。在 LDA 中，困惑度的计算公式如（4-5）和（4-6）所示。

$$perplexity = e^{-\frac{\sum \log[p(w)]}{N}} \tag{4-5}$$

$$p(w) = \sum zp(z \mid d) * p(w \mid z) \tag{4-6}$$

主题模型的可解释性用一致性衡量，一致性越高，模型的训练效果越好。综合主题模型困惑度和一致性，选择主题个数时应该尽可能使得困惑度更低，一致性更高。

由第 4 章的图 4-2 用户评分雷达图可知，用户评分较低的维度为舒适性、内饰、电耗和操控。故本文对舒适性、内饰、电耗和操控四个维度进行主题

提取分析，发掘消费者最关注的点和最不满意的点。从而找到新能源汽车的改进建议方向。

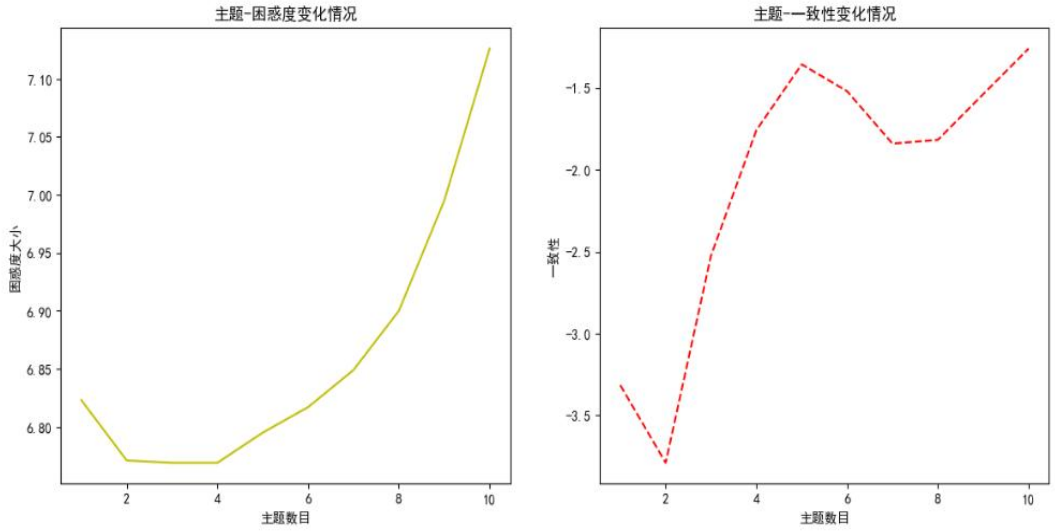


图 4-13 舒适性维度困惑度和一致性曲线

对于口碑评论舒适性维度，不同的主题数目对应的模型困惑度和一致性结果如图 4-13 所示。由此分析后设置模型训练时主题个数为 4。

表 4-4 舒适性维度前五个关键词及其权重

序号	关键词及其权重
主题 1	(0.078,噪音),(0.038,隔音),(0.022,发动机),(0.021,胎噪),(0.019,行驶)
主题 2	(0.088,舒适),(0.062,座椅),(0.024,感觉),(0.013,坐),(0.012,效果)
主题 3	(0.017,驾驶),(0.017,味道),(0.017,异味),(0.016,新车),(0.013,通风)
主题 4	(0.081,空调),(0.029,加热),(0.028,空气),(0.025,过滤),(0.019,按摩)

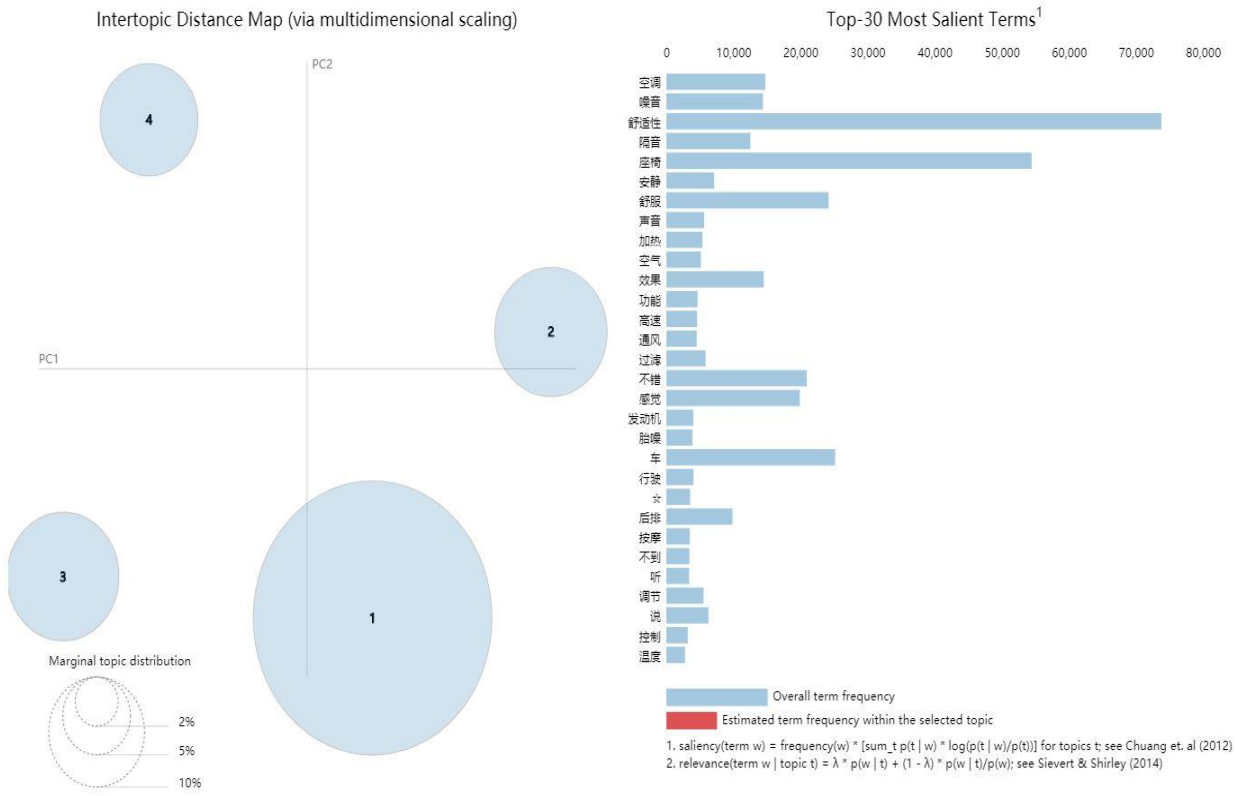


图 4-14 舒适性维度模型可视化

训练好 LDA 模型后输出模型各个主题对应的前 5 个关键词及其权重如表 4-4 所示。对模型数据进行可视化结果如图 4-14 所示，图中展示了不同主题间的距离地图及舒适性维度词频，可知对舒适性维度以主题数量为 4 进行主题提取时效果较好。

由主题 1 的关键词及其权重可知车辆行驶过程中的噪声是用户重要的不满意因素。由主题 2 的关键词可知座椅舒适度对消费者关于新能源汽车的舒适度体验影响较大。由主题 3 和主题 4 的关键词可知，新能源汽车的空气过滤和新车散发的异味影响了消费者的驾乘体验。综上，用户在新能源汽车舒适性维度评分均值较低，用户体验感较差。通过主题分析可知新能源汽车的噪声、座椅、异味和空气过滤是影响消费者舒适度的重要因素，车企应该在驾乘人员舒适性方面做出改善以提高消费者满意度。

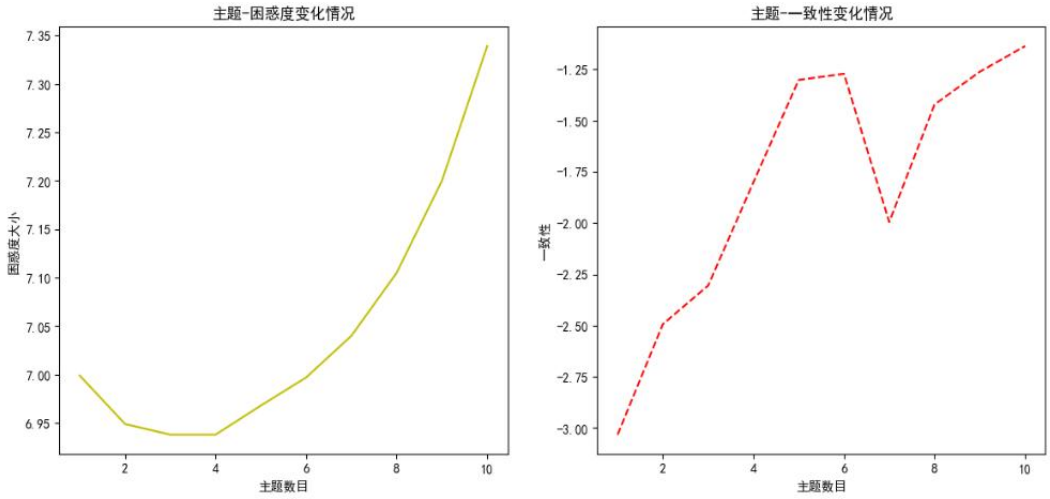


图 4-15 内饰维度困惑度和一致性曲线

对于内饰维度，其主题数目和模型困惑度及一致性的关系如图 4-15 所示。由此分析后设置模型训练时主题个数为 5。

表 4-5 内饰维度前五个关键词及其权重

序号	关键词及其权重
主题 1	(0.177,内饰),(0.029,设计),(0.019,做工),(0.017,用料)(0.014,风格)
主题 2	(0.030,座椅),(0.028,中控),(0.016,真皮),(0.016,设计),(0.012,方向盘)
主题 3	(0.093,氛围灯),(0.067,颜色),(0.014,真皮),(0.014,搭配),(0.012,光线)
主题 4	(0.050,功能),(0.029,控制),(0.016,操作),(0.013,空调),(0.012,大屏)
主题 5	(0.020,行车记录仪),(0.017,异味),(0.016,异响),(0.013,内置),(0.012,音响)

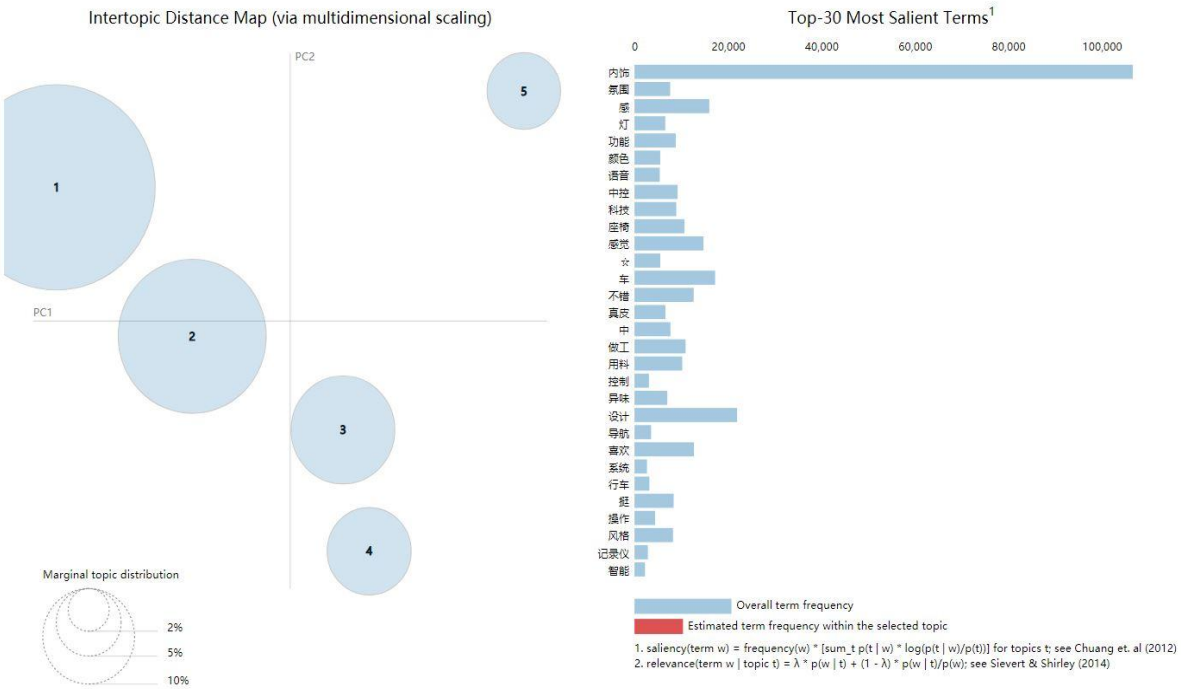


图 4-16 内饰维度模型可视化

将内饰维度的数据输入到主题模型后，训练的模型中每个主题对应的前 5 个关键词及其权重如表 4-5 所示。对模型数据进行可视化结果如图 4-16 所示，可知对内饰维度以主题数量为 5 进行主题提取时效果较好。

结合主题 1 和主题 2 的关键词分析可知在新能源汽车内饰方面，设计和用料做工是用户主要的关注点。设计和用料对于用户在驾乘汽车时的视觉和触感有较大影响，但目前许多新能源汽车车型在内饰设计方面较为简略甚至差，使得其较大程度地影响了消费者的主观感受。由主题 3、主题 4 和主题 5 的关键词可知新能源汽车内部光线搭配，功能设计和视觉质感是用户较为关注的方向。汽车内饰主要影响消费者驾乘汽车时的视觉体验，很大程度上影响了消费者在购车时的的决策，车企应该重视新能源汽车的内饰设计，满足用户多样化、个性化需求，从而提高市场竞争力。

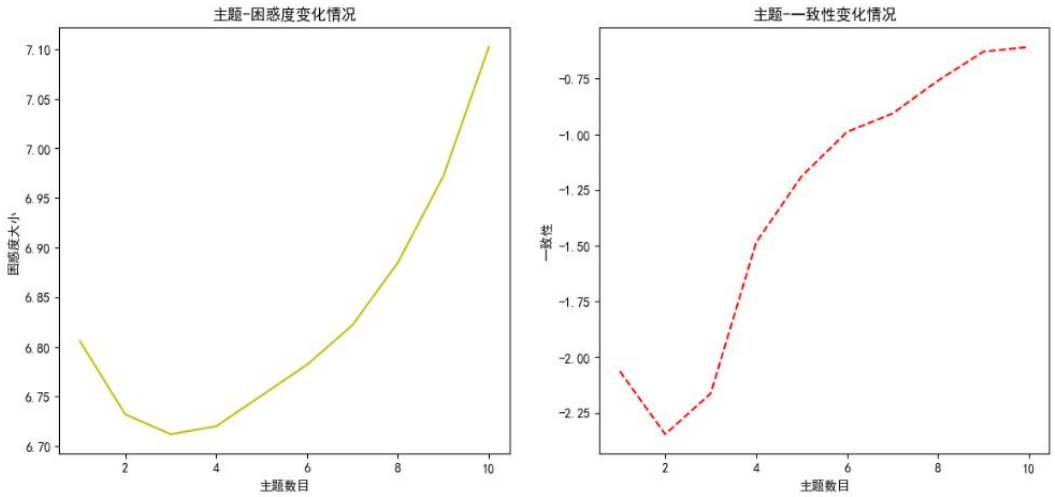


图 4-17 电耗维度困惑度和一致性曲线

在电耗维度上预训练主题模型时，通过计算困惑度和一致性并绘制折线图如 4-17 所示。由此分析后设置模型训练时主题个数为 4。

表 4-6 电耗维度前五个关键词及其权重

序号	关键词及其权重
主题 1	(0.083,电耗),(0.025,市区),(0.019,高速),(0.019,百公里),(0.017,省电)
主题 2	0.054,续航),(0.025,电池),(0.024,空调),(0.021,里程),(0.018,冬天)
主题 3	(0.104,充电),(0.019,充满),(0.014,电站),(0.014,快充),(0.014,免费)
主题 4	(0.145,模式),(0.057,换电),(0.024,发动机),(0.013,动力),(0.013,节能)

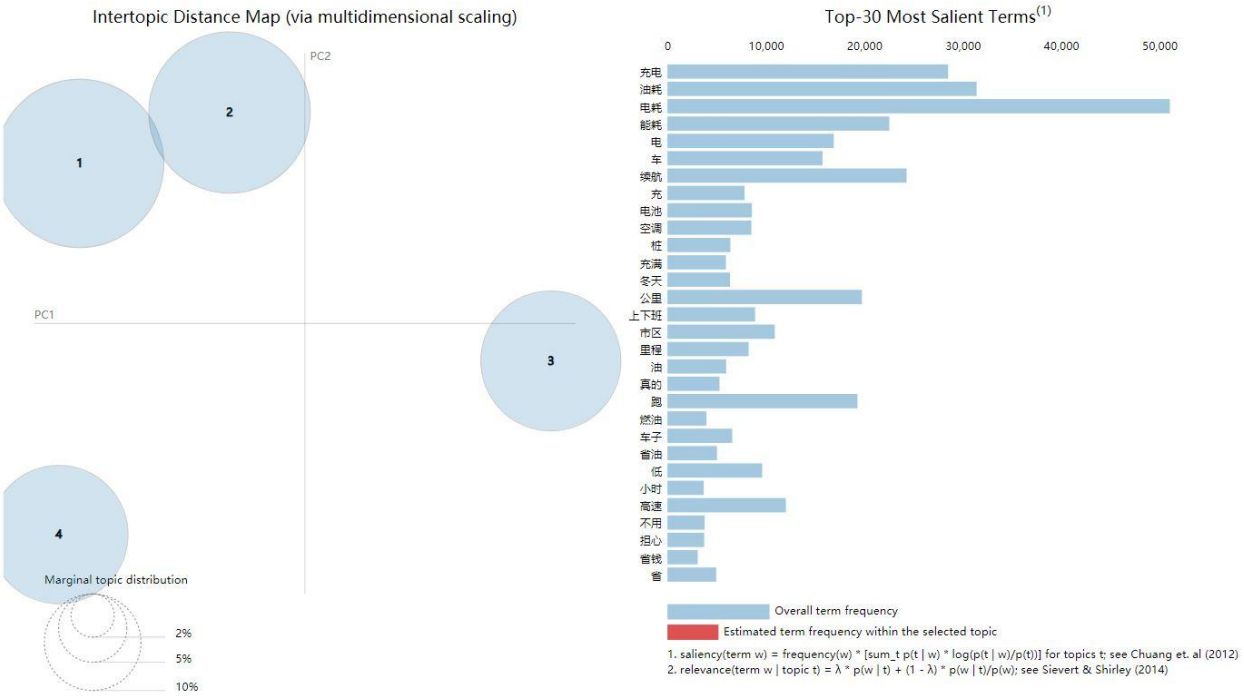


图 4-18 电耗维度模型可视化

通过困惑度和一致性曲线确定最终主题数目后训练出的电耗维度模型的各个主题下的关键词及权重如表 4-6 所示。图 4-18 是最终模型的可视化结果。由电耗维度 4 个主题的关键词及其权重可知，在电耗方面，用户更加关注新能源汽车的续航能力、高速电耗、充电便捷度、充电时间和汽车动力。随着新能源汽车的发展，其续航能力逐步提高，但是里程焦虑仍旧是用户在新能源汽车和传统油车间徘徊的主要原因，尤其是在冬季环境温度低的条件下，新能源汽车的续航能力会有一定的损失。在汽车动力方面，新能源汽车表现良好，某些品牌的汽车百公里加速甚至能够超越跑车，凸显了新能源汽车动力方面的优势。充电便捷度和充电时间对消费者的购买决策也有很大影响。目前我国新能源汽车发展较好的区域是沿海城市，其经济发展水平高，用户对新能源汽车的接触率高，且沿海城市人口密度较大，充电桩布局范围广，使得新能源汽车进一步占据了市场。但我国广大西部地区面积辽阔，但充电网点稀疏，是新能源汽车在西部地区发展缓慢的重要影响因素。目前车企在充电时间上的研发投入也较多，快充技术得到了进一步发展，提高了其市场竞争力。但是在电耗方面，车企还需进一步投入资源，懂车帝联合中国电动汽车百人会、巨量算数发布的《中国新能源汽车市场洞察报告 2021》指出续航能力是消费者购买新能源汽车时最为看重的因素。

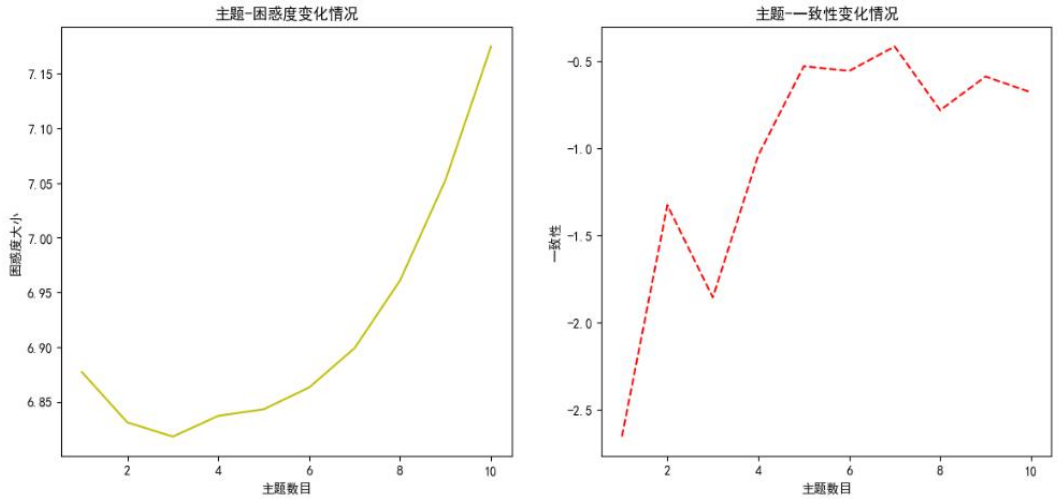


图 4-19 操控维度困惑度和一致性曲线

对于操控维度，其主题数目和模型困惑度及一致性的关系如图 4-19 所示。由此分析后设置模型训练时主题个数为 7。

表 4-7 操控维度前五个关键词及其权重

序号	关键词及其权重
主题 1	(0.138,模式),(0.051,运动),(0.037,驾驶),(0.023,调节),(0.015,切换)
主题 2	(0.035,电车),(0.011,停),(0.009,声音),(0.008,双叉臂),(0.008,升级)
主题 3	(0.056,刹车),(0.034,底盘),(0.014,悬架),(0.012,稳),(0.011,悬挂)
主题 4	(0.053,换挡),(0.041,可调),(0.021,旋钮),(0.018,磨合),(0.017,泥泞)
主题 5	(0.057,辅助),(0.051,功能),(0.026,倒车影像),(0.021,主动),(0.018,配置)
主题 6	(0.147,操控),(0.077,方向盘),(0.026,转向),(0.018,驾驶),(0.018,精准)
主题 7	(0.028,习惯),(0.028,操作简单),(0.015,开车),(0.012,电动车),(0.012,设计)

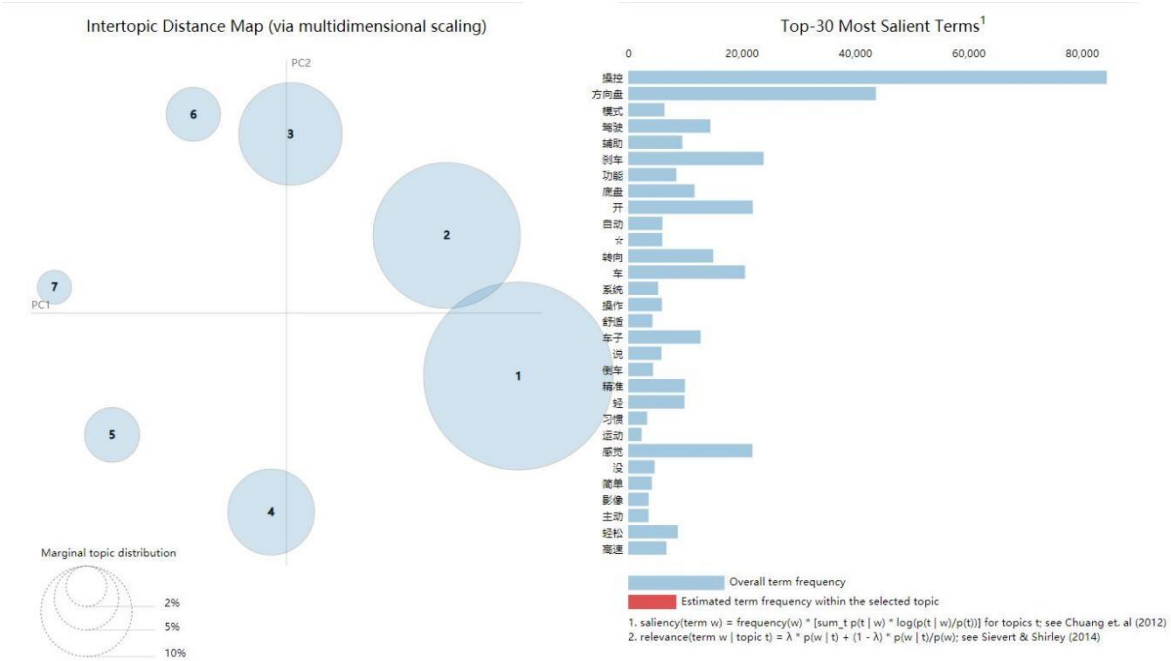


图 4-20 操控维度模型可视化

表 4-7 列出了操控维度主题挖掘后的关键词和对应的权重，图 4-20 便是该模型的效果图示。

通过对操控维度的评论文本进行主题提取可知用户在新能源汽车操控维度更为关注的因素有运动性、模式切换、稳定性及驾驶辅助等。新能源汽车作为新兴汽车产业，其消费人群偏向年轻化，此类消费者对于汽车的运动性关注较多，更偏向于选择动力强劲，有一定运动能力的汽车，具有运动、代步等模式切换的汽车更受青睐。驾驶辅助功能对消费者选择也有很大的影响，新能源汽车相比于传统油车，在辅助驾驶方面具备相当的优势，其软硬件生态都有良好表现，使得新能源汽车具有更强的市场竞争力。

通过建立主题模型对新能源汽车评论数据中用户评分较低的舒适性、内饰、电耗和操控维度进行潜在主题挖掘并进行结果分析，本文得到了新能源汽车在其短板维度中用户的关注重点和其影响消费者体验的因素，为新能源汽车优化升级提供一定的参考建议。

5 结论与展望

5.1 结论

随着互联网的快速发展和普及，网络评论数量呈井喷式增长。在线评论越来越能够影响消费者的购买决策，通过对大量在线评论进行分析，可以挖掘消费者痛点，帮助车企了解市场发展和自己产品的不足，为后续资源投入和产品更新方向提供一定的参考。对消费者而言，通过对在线评论的文本挖掘，可以帮助其快速了解产品画像，做出购买决策。本文使用 Python 爬虫获取了汽车之家和易车网两个大型汽车网络平台的消费者关于新能源汽车的在线评论数据，通过数据清洗、文本分词和词汇的分布式向量转化获得得到可用于机器学习的训练和测试数据集。然后以特征矩阵为基础训练和测试机器学习分类算法，并进一步建立主题模型对在线口碑数据进行潜在信息获取，分析消费者痛点和新能源汽车的劣势。通过数据分析，本文得出了如下结论：

1) 不同地区间新能源汽车市场发展差异明显，东部沿海地区发展迅速，市场占有率逐步上升，西部地区新能源汽车发展较为缓慢。

本文的研究基于汽车之家和易车网的在线口碑数据，据绘制的不同省份新能源汽车销量地图分析，新能源汽车在我国的发展存在地区间极不平衡的问题，我国新能源汽车的主要市场在沿海一带以及经济发达的城市，广阔的西部地区新能源汽车销量寥落。由于许多一线城市存在车辆限行及摇号难等问题，且这些城市人口出行需求量大，车辆充电较为便捷，因此相比于一些经济发展较为落后的地区，新能源汽车的市场的发展更好。

2) 国产新能源汽车市场占有率高，发展迅速，且价位为 10-20 万的车型居多。

由第五节的图 5-5 新能源汽车品牌销量气泡图可知在我国的新能源汽车市场中国产新能源汽车品牌占有率较高，说明我国新能源汽车产业发展态势良好，未来可期。由图 5-6 新能源汽车价位比例图可知，10-20 万的新能源汽车销量占比最大，达到了 46.4%，这个价位的新能源汽车性价比高，既能满足用户日常出行需求，又在多数用户能接受的价格范围之内。

3) 在使用机器学习算法进行新能源汽车口碑数据情感分类时，逻辑回归和随机森林模型均有较高的准确率，但从各个评价角度看来逻辑回归的表现均优于随机森林。

本文使用 jieba 对清洗后的数据进行分词并，进而以计算评论数据分词的

TF-IDF，然后使用全数据进行 Word2Vec 模型训练，将最满意维度和最不满意维度的数据进行分布式词向量转化，并以 TF-IDF 为权重对评论数据分词进行加权，得到评论文本向量，然后将文本向量转化为维度为 129162*300 的特征矩阵。以特征矩阵为数据基础构建机器学习的逻辑回归和随机森林模型，并进行通过混淆矩阵、精确率、召回率、F1-score、AUC 值等进行模型分类性能评估。经模型评估发现，所构建的逻辑回归模型的准确率为 92.172%，随机森林模型的准确率为 87.4%。通过从不同角度评价两个模型的分类效果，逻辑回归的分类效果均比随机森林表现更好，可知逻辑回归模型更适合于本文的新能源汽车在线评论文本分类。

4) 消费者对新能源汽车的外观、动力和性价比最为满意，对新能源汽车的舒适性、内饰、电耗和操控性最不满意。

通过绘制汽车之家各维度的评分数据均值的雷达图发现，在新能源汽车的外观、动力和性价比角度，消费者评分均值较高。可以看出，新能源汽车相比于传统油车在外观上有一定的革新，受到消费者青睐。并且新能源汽车的动力比同等价位的油车更为强劲。从雷达图可以看出，新能源汽车在内饰、电耗及舒适性方面的评分均值较低，消费者在这些方面有更高的需求。通过在电耗、内饰等维度的主题提取可以发现，在内饰方面，消费者较为关注且不满意的有用料做工、设计风格和功能配置等。新能源汽车的续航里程一直以来是影响其市场拓展的重要因素，通过主题分析发现满电续航、充电时间、充电桩数量及冬季续航能力下降程度是新能源汽车最受关注也最能影响消费者决策的因素。新能源汽车应当持续投入资源进行技术研究，专利研发，以自主可控的技术进一步缓解消费者的里程焦虑。

5.2 展望

1) 由于本文爬取的易车网评论数据为用户对汽车的整体评价，并未分维度进行汽车评论，且清洗后的数据只有 3942 条，相比于汽车之家的 64581 条数据过少，故本文对易车网数据的使用仅限于使用其分词后的数据构建 Word2Vec 模型，并未用于其他分析。但在实际文本挖掘中，非结构化数据占多数。在后续研究分析中，应该更加注重非结构化数据的收集和使用，使得本文研究有更强的泛化性。

2) 本文仅使用 68523 条数据条评论来构建分布式词向量 Word2Vec 模型，导致部分生僻词语的词频较低，未来可通过网络爬虫获取更多汽车之家和易车网更多最新的评论数据以及如太平洋汽车等多种网站的新能源汽车口

碑数据，对 Word2Vec 模型进行训练并进行深层次的情感分析和主题挖掘。

3) 由于本文共有 129162 条口碑数据用于训练和测试机器学习模型，对每条数据进行标识情感取向将耗费大量时间，故本文将最满意维度的所有数据假定其情感取向均为 1 ，最不满意维度数据的情感取向均为 0。本文后续应该考虑构造更为精细的模型以适应真实情况下中性评论占比较多的数据来检测更细粒度的情感。并且后续应该对模型参数做多次改变试及采用交叉验证方式进一步提高模型精确度。

参考文献

[1] 中共中央网络安全和信息化委员会办公室，中华人民共和国国家互联网信息办公室．第 48 次[R]．北京：中国互联网络信息中心，2021．

[2] 乔靖扬, 段贺颖, 吴争强, 等. 北京市新能源汽车市场现状调研[J]. 现代商业, 2021(30):19-23.

[3] 丁红萍, 吴志昱. 基于顾客价值认同的新能源汽车营销策略探讨[J]. 现代营销(学苑版), 2021(09):57-59.

[4] 张厚明．我国新能源汽车市场复苏态势及推进策略[J]. 经济纵横, 2021(10):70-76.

[5] 李千千. “互联网+”背景下新能源汽车的营销策略探究[J]. 中国商论, 2021(19):58-60.

[6] 张译, 秦佳良. 消费者购买新能源汽车意愿影响因素分析[J]. 经济研究导刊, 2021(20):41-45.

[7] 叶曼曼．新能源汽车购买影响因素实证分析[J]. 中国管理信息化, 2021,24(17):174-176.

[8] 李晓敏, 刘毅然, 靖博伦. 产业支持政策对中国新能源汽车推广的影响研究[J]. 管理评论:1-11.

[9] Sierzchula W., Bakker S., Maat K., et al. The influence of financial incentives and other socio-economic factors on electric vehicle adoption [J]. Energy Policy, 2014,68(5):183-194.

[10] Han H., Ou X. M. Du J., Feit M. D., et al. Will subsidies drive electric vehicle adoption? measuring consumer preferences in the U.S. and China[J]. Transportation Research Part A: Policy and Practice, 2015, 73(3): 96-112.

[11] 谢青, 田志龙. 创新政策如何推动我国新能源汽车产业的发展——基于政策工具与创新价值链的政策文本分析[J]. 科学学与科学技术管理, 2015,36(06):3-14.

[12] 杨荣华. 产业融合背景下的新能源汽车技术发展趋势研究[J]. 时代汽车, 2022(01):119-120.

[13] 赵雨. 关于新能源汽车发展存在的障碍及其解决措施研究[J]. 时代汽车, 2021(17):133-134.

[14] 辜文杰, 付宽. 基于大数据分析的新能源汽车行业发展研究[J]. 内燃机与配件, 2022(05):163-165.

[15] 高春晓, 李慧. 新能源汽车城市发展潜力区域特征显著[J]. 汽车纵横, 2021(10):45-47.

[16] R Johnson, Z Tong. Effective use of word order for text categorization with convolutional neuworks [C]. Eprint Arxiv,2014:103-112.

[17] Mikolov T,Kai Chen, et al. Efficient estimation of word representations in vector space[J].Computer Science,arXiv:1301.3781,2013.

[18] 阎亚亚. 词袋模型和 TF-IDF 在文本分类中的比较研究[J]. 电脑知识与技术, 2021,17(28):138-140.

[19] 毕云杉, 钱亚冠, 张超华, 等. 基于 ERNIE 模型的中文文本分类研究[J]. 浙江科技学院学报, 2021,33(06):461-468.

[20] 梁顺攀, 豆明明, 于洪涛, 等. 基于混合神经网络的文本分类方法[J]. 计算机工程与设计, 2022,43(02):573-579.

[21] David M. Blei, Andrew Y. Ng, et al. Latent dirichlet allocation, journal of machine learning reasearch 3,p993-1022,2003.

[22] 张胜. 基于用户评论挖掘的新能源汽车外观感性工学设计方法研究[D]. 广东工业大学, 2021.

[23] 余帆. 基于文本挖掘的新能源轿车用户情感分析[J]. 物流工程与管理, 2022, 44(01): 137-140.

[24] 张瑾. 基于评论文本挖掘的新能源汽车营销发展分析[D]. 河北经贸大学, 2020.

[25] 赵公民, 吕京芹, 武勇杰. 基于LDA 模型的新能源汽车政策文本量化分析[J]. 科技和产业, 2021, 21(01): 49-55.

[26] 池毓森. 基于 Python 的网页爬虫技术研究[J]. 信息与电脑(理论版), 2021,33(21):41-44.

[27] 刘洋, 余甜, 丁艺. 一种新的基于最大概率路径的中文分词[J]. 计算机与数字工程, 2022, 50(03): 591-596.

[28] 王琛. 基于 word2vec 情感分析系统的研究[D]. 长春大学, 2020.

[29] 接磊. 线上商品用户评论的情感分析系统研究与实现[D]. 辽宁大学, 2020.

[30] 张谦, 高章敏, 刘嘉勇. 基于 Word2vec 的微博短文本分类研究[J]. 信息网络安全, 2017(01):57-62.

[31] 王龙田. 基于逻辑回归与支持向量机对信用评分模型的研究[D]. 大连理工大学, 2021.

[32] 马辉. 基于随机森林的光伏电站结构故障诊断与分类研究[D]. 西安理

工大学, 2021.

[33] 严石. 基于改进 TF-IDF 和 fastText 算法的文本分类研究[D]. 安徽理工大学, 2020.

[34] 张枫叶. 基于 BERT 和 LDA 的阅读软件评论情感分析研究[D]. 曲阜师范大学, 2021.

[35] 童佳妮. 基于情感分类和主题挖掘的评论文本研究[D]. 浙江工商大学, 2021.

[36] 应昊东. 基于文本挖掘的新能源汽车评论情感分析研究及应用[D]. 东华大学, 2021.

附录 A 部分原始数据和代码

1、实验使用的原始数据（部分）

表 A.1 汽车之家在线评论

品牌名称	购买城市	购买时间	裸车购买价	目前行驶	空间(评分)	动力(评分)
小鹏	浙江	2021/12/1	19.5	4600	5	5
小鹏	江苏	2021/12/1	19.2	5000	5	5
小鹏	福建	2021/9/1	19.29	8300	5	5
小鹏	福建	2021/11/1	19.8	7200	5	3
小鹏	北京	2021/12/1	15.29	3083	5	5
小鹏	广东	2021/10/1	19.58	12000	5	4
小鹏	上海	2021/12/1	19.79	1400	5	5
小鹏	广东	2021/12/1	22.9	3000	4	4
小鹏	福建	2022/1/1	19.29	4200	5	5
小鹏	广东	2021/12/1	19.79	3500	4	4
小鹏	福建	2021/12/1	17.29	9006	5	5
小鹏	福建	2021/11/1	19.29	4000	5	5
小鹏	浙江	2021/12/1	19.29	2134	5	4
小鹏	重庆	2021/12/1	17.83	2816	5	5
小鹏	四川	2021/12/1	23.72	3258	5	5
小鹏	陕西	2022/2/1	16.47	550	5	5
小鹏	河南	2022/2/1	16.47	1159	5	5
小鹏	河南	2022/2/1	19.33	852	5	5
小鹏	广东	2022/1/1	19.79	4800	5	4
小鹏	江苏	2022/2/1	18.79	500	5	3
小鹏	广东	2021/11/1	20.53	5689	5	5
小鹏	湖北	2022/2/1	19.5	2300	3	3
小鹏	广东	2021/9/1	17.29	4828	4	4
小鹏	湖北	2021/12/1	19.79	1500	4	4

表 A.2 汽车之家在线评论

品牌名称	满意内容	不满意内容

小鹏	最满意的地方就是全场景智能语音了，识别率很高，而且区分了音区，主驾-副驾-后排。这样唤醒人的指令就不会因为其它地方的说话而干扰。通过持续的 ota,可以通过语音下达的指令越来越多，比如最近更新的 3.1 版本，可以通过语音降地锁了（非常好用，原先需要用手机 APP 降锁。另外一个的地方就是辅助驾驶了，作为一个买车前只有 100 公里左右驾驶经验的纯菜鸟，辅助驾驶可以让我在路上学到什么情况下车在车道正中。ngp 在高速和高架上表现很好，节约了大量的精力来控制方向和踩油门，当出现特殊情况，司机能有更丰富的精力和精神来处理。希望今年将会 OTA 的城市 NGP 能带来新的惊喜对了，肯定有很多人对辅助驾驶有疑虑，特别是老司机。我可以这么说，路上大部分事故都是不守规矩导致的，比如超速，变道不打灯，闯红灯，人行道不减速，不按规定使用远光灯，不保持安全距离，随意加塞等。这些不守规矩的行为，很多都来自有老司机，他们有自己的一套开车理论。辅助驾驶参与到交通行为中，我认为并不会制造一个新的毒瘤司机，甚至是一个更安全的交通参与者。当然，辅助驾驶不是自动驾驶，握好方向盘，持续关注路况也是必要的。现实 4600 公里使用下来，正确使用辅助驾驶确实能减少司机的疲劳和增加驾驶的安全性	座椅不舒服，第一次开去金华，200 公里的路，感觉腰难受(可能身体太虚了也有关系)。买了个腰靠，腰是舒服了，脖子开始难受了，只能再买一个颈枕。这两个东西放上去之后，整体上就舒服很多了，估计是因为睡眠空间需要座椅躺平的设计，导致不是很符合人体工学（也有可能是我太虚了）后排无法放倒，这也导致 P5 的运载能力不强，比较长的就不用考虑用 P5 运输了。如果安装了冰箱，那运载能力还要再打一个折扣，春节回杭州，不光后备箱满满，后排座位上也塞满了，如果座椅能放倒，那应该能载更多的东西。
小鹏	已经是胖虎 5000 公里的深度体验的车主啦，来说说最近的感受不是最满意但是最值得高兴的地方，就是随着天气回暖，能耗明显降低了，续航里程直线上升，夜间掉里程的情况还有，但是已经到一天 3km 以内了，期待夏天的到来，看看能否有更多的突破使用了 3 个月，最爱不释口的，就是你好小 p 啦，因为小 p 真的能解决好多事情下班后喊声小 p 回家，会为你打开空调，调整最舒适的座椅，播放收藏夹里的歌，打开家的导航你好小 p，打开遮阳帘——太阳太晒啦——我好热，3 段流畅的对话，可以让小 p 连续 3 个指令，不用重新唤醒午间——我想休息会，冥想小 p 会把座椅放下，关闭窗户，打开香氛，关闭灯光，播放蛙叫虫鸣，戴上眼罩，轻松入睡停车？别人是疑难杂症交给人工，而我停车的疑难杂症，全交给小 p 的自动泊车啦，车位窄？车距短？侧方位？喊一句小 p 停车，松开刹车，收拾包包等待小 p 入库就可以啦！其实有很多人质疑，尤其没体验过智能语音的还会觉得，语音控制是个鸡肋，按键可以完成，语音会影响驾驶安全其实小 p 很聪明哦，影响驾驶操作安全的指令，需要在车辆停止的时候，才可以执行，或者必需手动执行。而行驶中就让自己的双手握住方向盘，换换音乐，听个笑话，看看股票，空调开窗，座椅调节，就全交给小 p 吧？	如果一定要选最不满意的，还是想说一下续航的问题。随时续航是电车通病，购车时选择的是 550 版本，和名字一样，是 550km 的续航。因为比较喜欢 18 寸米其林轮胎，续航影响，充满电只有 535 续航了。但是！！！这是商家计算的无阻力，无障碍，空载得来的结果吧，切换为 wltp，535 在充 90%的情况下立刻变 396 了，然后在 396 续航的情况下，还要根据温度，车的载重，路况等等再进行打折，春节满载回家的时候，200 公里就没电了，也是最差劲的一次。希望后续电池能够得到更好的优化，里程能够有进一步的提升

表 A.3 易车网在线评论

车型	空间评分	外观评分	内饰评分	动力评分	操控评分	舒适性评分	油耗/续航评分
比亚迪汉	8	10	8	10	8	8	8
比亚迪汉	8	10	10	10	8	10	8
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	10	8	10	8
比亚迪汉	10	10	10	10	10	8	10
比亚迪汉	10	10	10	10	10	10	10

比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	8	8	8	8	8	8	8
比亚迪汉	8	10	10	10	8	8	8
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	10	10	10	2
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	8	8	8	8	6	6	6
比亚迪汉	10	8	10	10	10	10	8
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	8	8	8	8	8	8	8
比亚迪汉	10	10	10	10	10	10	10
比亚迪汉	10	10	10	8	8	10	6

表 A.4 易车网在线评论

车型	点评内容
比亚迪汉	各位朋友大家好，我买这款车是 2020 款尊贵版汉 ev 当初买车这款车也是出于省油环保再加上北京可以上绿牌不限行，家里有一台燃油车也是北京的指标，由于北京限牌家里一辆车不够用只能再买一辆倒着用，买电车我看了第一款就是小鹏 p7 当时感觉小鹏 p7 操作起来还挺满意的，颜值也很棒，可惜价钱也比较贵超出预算一大块，就放弃了，再去小鹏的路上听另外一个买车的人说比亚迪的电车也挺不错，我就想也去比亚迪看看，到了比亚迪店里服务还挺不错很热情，门口第一台车就是比亚迪汉，我老婆被它造型吸引了说这不比小鹏还漂亮吗？接下来试驾感觉跟小鹏区别不是特别大，但是价钱差距还是挺大的，汉各方面都还不错，就定了，当时候这辆车全部算下来花了 25 万多，给我优惠了两万裸车价 24 万多点，还有保险不到一万反正加起来花了 25 万多点感觉还能接受，现在都跑了一年多了说说我的综合感受吧，先说说续航我是居住在北京，北京夏季续航 550，冬季续航 6-7 折很正常～(-5℃左右） 冬季续航北京的冬天相当寒冷，平均气温只有﹣5℃，极度严寒时，气温会下降到﹣10℃以下，这种情况下，我得汉 ev 续航，5-7 折之间徘徊，也就是 300-400 公里之间非常满意了，一到两周充一次电，非常的不错。冬季来说，虽然在开空调暖风的情况下，平均冬季能耗可以达到 17kwh/100km+，即便这样，实际续航也能突破 300 公里甚至接近 400 公里，一周一冲也不是不能接受，毕竟汉 ev 的快充表现还是相当不错的～在北京这种大型城市日常使用，汉 ev 绝对是一款居家好手～夏季 7 天 1 充（剩余电量 30%左右）冬季 5 天 1 充（剩余电量＜30%） 偶尔家用充电桩，平常国网快充 ，国网快充桩 60kw，30%-100%，需要 1 小时。家用慢充桩，3.5kw 版 0-100%需要 28 小时。 目前全年平均能耗 15kwh/100km，每月充电成本 300 元以内。 首先空调是不能省的，二十几万买了汉，空调省下来的还不如去买秦～不过来说，缓加速，少踩刹车，确实可以省电～值得注意的是，这里说的缓加速并不是速度慢，而是不要地板油加速，要缓慢提升速度～ 慢充就不说了，这里重点说一下在共用快充桩的充电体验。汉 ev 已经把公共充电桩玩儿的比其他品牌专属充电桩还明白了。无论是 120kw 版本还是 60kw 版本的国网快充桩，汉 ev 与其他品牌同时，同电量开始充电，汉 ev 永远是最先充满的。空间表现本人 180cm，体重 87kg，可以说是个敦实的小墩墩～汉 ev 的宽大车身对于我来说相当友好～横向空间直接拉满～其实前排空间表现更重要的是座椅调节，汉 ev 的主驾驶座椅可以很低～这就对司机来说非常舒服了我前排调整好驾驶位置后，我在后排腿部空间还有两拳，这样的空间表现简直堪称优越，后备汉 ev 的运载能力也已经在为我丈母娘搬家的任务中体现出来了～两床被子，一个轮椅，锅碗瓢盆，洗漱用品，高压锅电饭锅，通通塞下，空间上汉 ev 真的在同价位是非常有实力的，无论是轿厢空间还是后备箱空间都是家用首选～更重要的在乘坐的横向空间，给了墩墩们非常友好的乘坐体验～真的要点赞！ 驾驶起来完全可以满足日常需求，即便是激烈驾驶也不会觉得很拉夸，具体动力情况可以参照 2.0T 家用车型，具体动力感觉和凯美瑞、雅阁、迈腾差不太多～家用完全足够。保养第一次半年或 1.2w 公里，保养费用根据买车时优惠政策，2020 款赠送 3 年或 6 次保养，其后基础保养费用在 230 元 。维修仅对冬季洗车车门进水结冰无法开启车门问题和原厂自带的跑偏问题做过四轮定位（免费）我选择的是北京地区比较有实力的经销商，所以售后服务我感觉还是非常不错。热情周到，技术水平好这些基本素质，大经销商都是具备的。而且在解决问题说靠谱的经销商也绝不拖泥带水互相扯皮。就我的维修保养经历而言，北京地区前三的经销商还是非常不错的，选对经销商，服务才能到位。

2、实验各部分代码（部分）

（1）爬虫模块

```
import re
import csv
import requests
import threading
from lxml import etree
from retrying import retry
from collections import Counter
from fontTools.ttLib import TTFont
from fake_useragent import UserAgent

class Spider(object):
    def __init__(self):
        self.headers = {
            'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.121 Safari/537.36'}

    @retry(stop_max_attempt_number=5)
    def _parse_url(self, url):
        """url 请求"""
            response = requests.get(url, headers=self.headers, allow_redirects=False, timeout=5)
            return response
    def get_bast_cmap(self):
        """字体处理"""
        ttfont = TTFont('yc-ft.woff') # 读取字体文件
        # 读取映射表 映射网页中的加密的字符串
        bast_cmap = ttfont['cmap'].getBestCmap()
        new_bast_cmap = {} # hex 转为十六进制
        for key, value in bast_cmap.items():
            # 将 unicode 转换为中文
            new_bast_cmap[hex(key)] = str(value).replace('uni', r'\u').encode('utf-8').decode('unicode_escape')
        return new_bast_cmap
    @retry(stop_max_attempt_number=3)
```

```
    def get_koubei_url(self, url):
        """
        获取车系口碑详情页 url
        :param url:车系 url
        :return:口碑详情 url
        """
        response = self._parse_url(url=url)
        result = etree.HTML(response.text)  # 转换数据类型为 HTML,方便使用 xpath
        for i in range(1, int(page[0])+1):
            print(f'第：{i}页数据获取中...')
            response = self._parse_url(url=f'{url}-{i}.html')
            result = etree.HTML(response.text)  # 转换数据类型为 HTML,方便使用 xpath
            # 详情页 url
            url_list = result.xpath("//div[@class='cm-content-moudle']/a/@href")
            url_list = ['https://dianping.yiche.com'+i for i in url_list]
            for koubei_url in url_list:
                yield koubei_url

    def get_data(self, html, url):
        """提取口碑详情"""
        result = etree.HTML(html)
        username = self.is_null(result.xpath('//div[@class="dj-user"]/text()'))  # 用户名
        car_name = self.is_null(result.xpath('//div[@class="c-info-title"]/text()'))  # 车型
        date_time = self.is_null(result.xpath('//div[@class="tc-date"]/span/text()'))  # 提车时间
        car_price = self.is_null(result.xpath('//div[@class="fd-bot"]/div/text()')).replace('\n', '').replace(' ', '')  # 裸车价
        fd_txt = result.xpath('//div[@class="fd-bot"]/div/text()')  # 购车地
```

```
            fd_txt = fd_txt[1] if len(fd_txt) > 2 else '-'
            data = [[username, car_name, date_time, car_price, fd_txt, score, date_time1, scores[0], scores[1], scores[2], scores[3],
                     scores[4], scores[5], scores[6], tcid, url]]
            return data

        def save_csv(self, filename, data):
            """保存数据"""
            filename = 'D:\\大四\\毕业论文\\毕业论文 41807292\\数据\\易车\\易车口碑数据.csv'
            with open(filename, "a+", encoding='gbk', errors='ignore', newline="") as f:
                k = csv.writer(f, delimiter=',')
                with open(filename, "r", encoding='gbk', errors='ignore', newline="") as f1:
                    reader = csv.reader(f1)
                    if not [row for row in reader]:
                        k.writerow(['用户名', '车型', '提车时间', '裸车价', '购车地', '综合评分', '口碑发布时间', '空间评分', '外观评分',
                                    '内饰评分', '动力评分', '操控评分', '舒适性评分', '油耗/续航评分', '点评内容', '地址'])
                        k.writerows(data)
                    else:
                        k.writerows(data)

        def main(self, url):
            # 字体加密映射表
            new_bast_cmap = self.get_bast_cmap()
            # 获取车系所有口碑详情 url
            for koubei_url in self.get_koubei_url(url=url):
                html = self.fount_replace(self.get_html(url=koubei_url), new_bast_cmap)
                data = self.get_data(html=html, url=koubei_url)
                self.save_csv(filename='', data=data)
```

（2）分词模块

```
# 导入相关的软件包
import pandas as pd #导入 pandas
import jieba
import jieba.posseg as pseg

# 获取汽车数据
data = pd.read_csv('D:\\大四\\毕业论文\\毕业论文 41807292\\数据\\汽车之家\\汽车之家新能源\\汇总数据\\汽车之家新能源汇总.csv')
df = data.copy()
# 获取专有词汇数据
data_proprietary_vocabulary = pd.read_csv('D:\\大四\\毕业论文\\毕业论文 41807292\\数据\\汽车之家\\分词-专有词汇-停用词\\专有词汇\\专有词汇汇总.csv')

# 加入分词专有词汇
list_word = []
for i in data_proprietary_vocabulary.iloc[:,0]:
    list_word.append(i)
map(jieba.add_word,list_word)

# 加入停用词
data_stop = pd.read_csv('D:\\大四\\毕业论文\\毕业论文 41807292\\数据\\汽车之家\\分词-专有词汇-停用词\\停用词\\停用词.csv')
list_stop = [i for i in data_stop.iloc[:,0]]
print(list_stop)

# 获取汽车评论最满意内容
list_satisfied = df.iloc[:,13].astype(str)
# 对最满意维度的第一条评论进行分词
sd = jieba.cut(list_satisfied[0])
for i in sd:
    if i not in list_stop:
        print(i)
# 对最满意内容进行分词
```

```
list_words = []
set_words = set()
dict_words = dict()
list_generator = map(jieba.cut,list_satisfied)
for i in list_generator:
    for word in i:
        if word not in list_stop:
            list_words.append(word)
set_words = set(list_words)
m = 0
for key in set_words:
    m+=1
    print(m)
    dict_words[key] = list_words.count(key)
key = dict_words.keys()
value = dict_words.values()
data_count_frame = pd.DataFrame({'分词':key,'计数':value})
# 保存数据
data_count_frame.to_excel('最满意维度分词统计数据.xlsx')
```

（3）Word2Vec 词向量转化模块

```
# 导入库包
import pandas as pd
from gensim.models import Word2Vec
import jieba
import jieba.posseg as pseg
import json

# 语料准备
# 读取分词结果 list of list 对象
list_all = []
m=0
with open('wordCut.json','r') as f:
    while 1:
        m+=1
        print(m)
```

```
            d=f.readline()
            if not d:
                break
            list_all.append(json.loads(d))
            #         print(type(json.loads(d)))

    # 将分词结果输入 word2vec 模型进行训练
    model = Word2Vec(list_all,sg=0,
size=300,window=5,alpha=0.001,min_count=5,hs=0,negative=10,iter=30,cbow_
mean=1)
    # 保存模型
    model.save('word2vec.model')
```

（4）模型构建模块

```
    # 导入库包
    import pandas as pd
    from gensim.models import Word2Vec
    import jieba
    import jieba.posseg as pseg
    import json
    import time
    import numpy as np
    # 生成特征矩阵
    # 读取分词结果 list of list 对象
    list_all = []
    # 加载模型
    model = Word2Vec.load('word2vec.model')
    # 定义函数获取特征矩阵
    def getVector_v1(cutWords, word2vec_model):
        count = 0
        article_vector = np.zeros( word2vec_model.layer1_size )
        for cutWord in cutWords:
            if cutWord in word2vec_model:
                article_vector += word2vec_model[cutWord]
                count += 1
        return article_vector / count
```

```
    # 获取特征矩阵
    startTime = time.time()
    vector_list = []       #特征矩阵
    i = 0
    for cutWords in list_all:
        i += 1
        if i % 1000 == 0:
    #             pass
                print('前%d 篇评论形成词向量花费%.2f 秒' % (i, time.time() - startTime))
        vector = getVector_v1(cutWords, model)
        vector_list.append(vector)
    #         print(vector)
    X = np.array(vector_list)
    print('Total Time You Need To Get X:%.2f 秒' % (time.time() - startTime) )

    # 保存数据
    X.dump('sat_dissat_vector.txt')
    from sklearn.preprocessing import LabelEncoder
    # 标签编码
    labelEncoder = LabelEncoder()
    y = labelEncoder.fit_transform(data['emotion'])

    # 模型构建
    # 逻辑回归
    from sklearn.linear_model import LogisticRegression
    from sklearn.model_selection import train_test_split
    # 划分数据集
    train_X, test_X, train_y, test_y = train_test_split(X, y, test_size=0.3, random_state=0)
    # 进行逻辑回归模型训练
    logistic_model = LogisticRegression()
    logistic_model.fit(train_X, train_y)
    logistic_model.score(test_X, test_y)
```

```
from sklearn.metrics import classification_report
# 精确率、召回率、F1-score
report = classification_report(test_y, logistic_model.predict(test_X), labels=[0, 1], target_names=['正向情感', '负向情感'])
from sklearn.metrics import roc_auc_score # 计算 AUC 值
# AUC
roc_auc_score(test_y, logistic_model.predict_proba(test_X)[:, 1])

# 随机森林
from sklearn.ensemble import RandomForestClassifier
#实例化模型
rf = RandomForestClassifier(
    criterion='entropy',    #使用信息增益作为标准建立随机森林
    n_estimators=1000,      #建立 1000 棵随机树
    min_samples_split=10, # 定义至少多少个样本的情况下才继续分叉
    min_weight_fraction_leaf=0.02 # 定义叶子节点最少需要包含多少个样本(使用百分比), 防止过拟合)
# 训练模型
rf.fit(train_X, train_y)
# 预测
y_pred = rf.predict(test_X)
print('分类预测值：',y_pred)
```

（5）主题提取模块

```
# 导入库包
import pandas as pd
from gensim.models import Word2Vec
import jieba
import jieba.posseg as pseg
import json
import time
import numpy as np
import warnings
warnings.filterwarnings('ignore')
```

```
# 在 gensim 包中使用 LDA
from gensim import corpora, models
from gensim.test.utils import common_texts
from gensim.corpora.dictionary import Dictionary
import gensim.corpora
from gensim.test.utils import common_corpus, common_dictionary
from gensim.models.coherencemodel import CoherenceModel

# 读取分词结果 list of list 对象
list_all = []
# 根据文本列表创建一个语料库，每个词与一个整型索引值对应
word_dict = corpora.Dictionary(list_all)
# 词频统计，转化成空间向量格式
corpus_list = [word_dict.doc2bow(text) for text in list_all]
model_list = [] # 模型列表
perplexity = [] # 困惑度
coherence_values = [] # 一致性
num_topic = 10 #主题个数

for num_topics in range(1,num_topic+1):
    lda_model = models.LdaModel(corpus = corpus_list, id2word = word_dict, random_state = 1, num_topics = num_topics, passes = 2, alpha='auto')
    model_list.append(lda_model) # 不同主题个数下的 lda 模型

    # 模型对应的困惑度（越低越好）
    perplexity_values = lda_model.log_perplexity(corpus_list)
    print('第 %d 个主题的 Perplexity 为：' % (num_topics ), -round(perplexity_values, 3))
    perplexity.append(-round(perplexity_values, 3))
    # 模型对应的一致性（越高越好）
    coherencemodel = CoherenceModel(model = lda_model, corpus = common_corpus, coherence = 'u_mass')
    coherence_values.append(-round(coherencemodel.get_coherence(),3))
    print('第 %d 个主题的 Coherence 为：' % (num_topics), -round(coherencemodel.get_coherence(),3))
```

```
# 训练最终模型
import pyLDAvis.gensim_models
lda = models.LdaModel(corpus = corpus_list, id2word = word_dict, random_state = 1, num_topics = num_topic_fin, passes = 1, alpha='auto')
# 打印各主题下对应单词
for topic in lda.print_topics(num_words=10):
    print(topic)
```

在学取得成果

一、 在学期间所获的奖励

东凌奖学金	经济管理学院	2018.11
人民奖学金三等奖	北京科技大学	2019.11
人民特等奖学金	北京科技大学	2020.11
优秀三好学生	北京科技大学	2020.11
人民奖学金三等奖	北京科技大学	2021.11

二、 在学期间发表的论文

三、 在学期间取得的科技成果

致　谢

在本文研究的过程中，本人得到了多个组织和个人的重要建议和指导，故于此向所有对本论文工作有贡献及帮助的人士和单位表达感谢。

首先，感谢北京科技大学在学习资源和疑难解答方面提供的有力帮助。在本文研究的各个阶段，包括题目选定、任务书和选题报告的撰写、研究模型的学习以及论文写作等，均需要查阅大量资料。学校与知网、万方等文献查阅平台合作，为我们的研究和论文的撰写提供里所需的各种资料。此外，学校图书馆丰富的书籍资料也对本人的论文研究有很大的帮助，让我能够及时获取信息解决所遇到的问题。

其次，我要感谢我的论文导师和本科生导师刘学娟副教授。刘老师及时准确地告知我们学校的具体安排和各项要求，督促并指导我们按计划进行工作。在论文写作开始前，刘老师便与我们进行了直接的沟通交流，并且提供了许多对我们帮助很大的资料和文件。在确定研究方向和题目时，刘老师提出了很多宝贵的意见和建议，帮助我修正研究框架，使研究更具意义。当我向刘老师请教我的实验或思路中出现的问题时，她都及时耐心地帮我解答，使我能够进入到下一个研究环节。还要感谢刘老师帮助检查任务书、选题报告、论文、提出改进建议，使研究成果更加完善。

再次，我要感谢所有引用在论文中期刊及其作者和所在期刊、以及研究过程中参考的所有知识网站为研究设想提供理论基础，为本研究的实验提供技术支持。

最后，我要感谢我的女朋友在我论文撰写期间的陪伴，她给了我很大的支持和帮助，让我有信心解决各种问题，有条不紊地进行我的研究工作，这篇文章的完成有她很大一部分的贡献。

再次向以上组织和个人表示感谢，因为有她们的支持和协助，论文工作才得以顺利进行，进而得到较为完善的成果。