第八届“泰迪杯”
数据挖掘挑战赛

泰
迪
杯

作品名称：基于文本挖掘的“智慧政务”系统

基于文本挖掘的“智慧政务”系统

摘要：随着互联网、云计算、自然语言处理技术的快速发展，政府及管理部门也正朝着一种以互联网等为技术支撑，通过智能化等方式，促进“智慧政务”系统工作开展。而网络问政平台逐渐成为了群众反映民生民意的重要渠道，面对问政留言的文本数据量剧增，留言划分归类、群众关注的热点问题挖掘给相关部门的工作带来了极大挑战。因此，本文将探寻适合“智慧政务”平台留言分类模型和热点问题挖掘的方法，并对留言回复质量构建评价体系。

在数据预处理阶段，我们对数据进行分词和去停用词处理，通过多种算法的比较，最终选择 TF-IDF 算法获得文本的词向量表示。

针对问题一，我们分别尝试了传统机器学习中的朴素贝叶斯模型和深度学习中的 CNN 模型和 LSTM 模型对比其中分类结果的准确率，发现传统机器学习的方法更适用于本次的分类任务。为了进一步提升准确率，我们再选取了逻辑回归多分类器中的 OVR 和 OVO 形式进行分类。多次试验测算证明，逻辑回归多分类器的 OVO 形式的准确率提升了超过 20%，该模型的 F1-score 值较高，验证了逻辑回归多分类模型的有效性。

针对问题二，通过使用基于贝叶斯信息准则的聚类分析方法，聚类得到留言反映的各类问题。然后，我们根据“留言的权重大小、关键词的竞争力、主题内容”等因素，构造了“智慧政务”专有的热度测算方法来衡量聚类后的问题热度，最终发现热点问题。

针对问题三，通过构建一个综合评价体系评价留言回复质量。我们对留言回复与留言内容通过 word2vec 处理得到词向量，并通过余弦相似度等方法计算两者的相关性、回复的完整性、可解释性和时效性，综合上述几个评价指标，经过归一化处理后，得到留言问政平台回复质量完整的评价体系。

关键词：TF-IDF 算法；逻辑回归多分类器；贝叶斯信息准则；word2vec

"Smart Government Affairs" System Based on Text Mining

Abstract: With the rapid development of the Internet, cloud computing, and natural language processing technologies, the government and management departments are also moving towards a technology-based support, through intellectualization and other means, to promote the work of the "smart government" system. The online questioning platform has gradually become an important channel for the public to reflect people's livelihood and public opinion. In the face of the sharp increase in the amount of textual data on the questionnaire, the classification of the message and the mining of hot issues of concern to the people have brought great challenges to the work of relevant departments . Therefore, this article will explore methods suitable for message classification models and hotspot mining on the "smart government" platform, and build an evaluation system for the quality of message responses.

In the data preprocessing stage, we perform word segmentation and stop word processing on the data. Through comparison of various algorithms, we finally select the TF-IDF algorithm to obtain the word vector representation of the text.

For problem 1, we compared the accuracy of the classification results of the Naive Bayes model in traditional machine learning and the CNN model and LSTM model in deep learning, and found that the traditional machine learning method is more suitable for this classification task. In order to further improve the accuracy, we chose the OVR and OVO forms in the logistic regression multi-classifier for classification. Multiple tests have proved that the accuracy of the OVO form of the logistic regression multi-classifier has been improved by more than 20%. The F1-score value of this model is high, which verifies the effectiveness of the logistic regression multi-classification model.

For the second problem, by using the clustering analysis method based on Bayesian information criterion, the clustering can get all kinds of problems reflected by the message. Then, based on factors such as "weight of message, competitiveness of keywords, subject content" and other factors, we constructed a "smart government" proprietary heat measurement method to measure the heat of the problem after clustering and finally found hot issues.

In response to question three, the quality of message responses was evaluated by constructing a comprehensive evaluation system. We use word2vec to process word replies and message content to obtain word vectors, and calculate the correlation, completeness, interpretability, and timeliness of the two through cosine similarity and other methods. After the treatment, a comprehensive evaluation system for the quality of the reply to the message inquiry platform was obtained.

Key words: TF-IDF algorithm; logistic regression multiple classifier; Bayes information criterion; word2vec

目录

图目录

表目录

第1章 绪论

1.1 研究背景及意义

随着互联网、云计算、自然语言处理技术的快速发展，一个以海量信息和数据挖掘为特征的大数据时代已经到来，高科技技术也在风驰电掣般的发展，大数据已经渗透到我们生活中的方方面面，悄然推动着人类的生产生活、社会管理方式向着“智慧”的方向发展。与此同时，政府及管理部门也正朝着一种以互联网、云计算、数据挖掘、深度学习等为技术支撑，以政府信息高效化处理为核心，通过互联化、物联化、智能化的方式，促进智慧政务系统工作开展，实现政务信息强度整合、深度应用之目标的新模式——智慧政务发展。网络技术的进步实现了许多政务事项网上办理，落实了全流程“数据跑”，为群众带来了“最多跑一次”甚至“零次跑”的便利。智慧政务已逐渐成为新一轮社会治理创新与发展的潮流。

而近年来，随着微信、微博、市长信箱、阳光热线等方式逐渐成为了群众反映民生民意的重要渠道，网络问政平台越来越受群众关注，群众问政留言的文本数据量剧增，然而目前，大部分网络政务系统平台还是依靠人工凭借经验处理，存在工作量大、效率低，且错误率高等问题。因此，这给靠人工进行留言划分归类、群众关注及急需处理的热点问题整理和意见答复的相关部门的工作带来了极大挑战。

目前，数据潜在的隐形价值得到人们的关注，越来越多的人开始重视数据挖掘技术的发展和研究，从而自然语言处理技术的研究也得以快速提高。早在上一世纪，自然语言处理就已经进入到了计算机科学的人工智能分支的研究领域。自然语言处理系统的用途范围较广：如机器翻译、文本分类等。作为处理海量信息重要手段，文本分类技术使信息查找方便快捷，可以准确地为需求者提供所需的信息。该技术已经流行用于垃圾邮件分类、舆情监测、新闻分类等场景[1]。

基于以上背景，本文将利用自然语言处理和文本挖掘技术构建问政留言分类和热点问题测度模型，为提高政府部门对群众问政留言处理的工作效率和热点追踪，并探究留言答复质量的各个因素之间的内部关系，建立一套评价体系，完善答复质量和方案，进而对“智慧政务”更好地运行提出一些针对性的改善建议。

1.2 技术路线

本文首先查阅各类文献，了解文本挖掘和自然语言处理的相关理论。接着我们对文本数据进行预处理，对文本分词和去停用词处理后采用 TF-IDF 算法将文本转化为向量数据，进而进行模型的构建和分析。针对问题 1 的文本分类问题，本文构建了朴素贝叶斯分类器、CNN 分类器等将传统机器学习方法和深度学习的方法进行对比，选择适合本文的分类器；针对问题 2 的热点问题挖掘，利用基于贝叶斯信息准则的 K-means 聚类分析的方法挖掘出相似问题，并计算热度，挖掘出“智慧政务”的热点问题；针对问题 3，我们从相关性、完整性、可解释性等角度构建了网络问政平台回复的评价体系，对留言回复质量进行评判。最后，根据实验结果，为“智慧政务”体系更好地发展提出相应地结论建议。

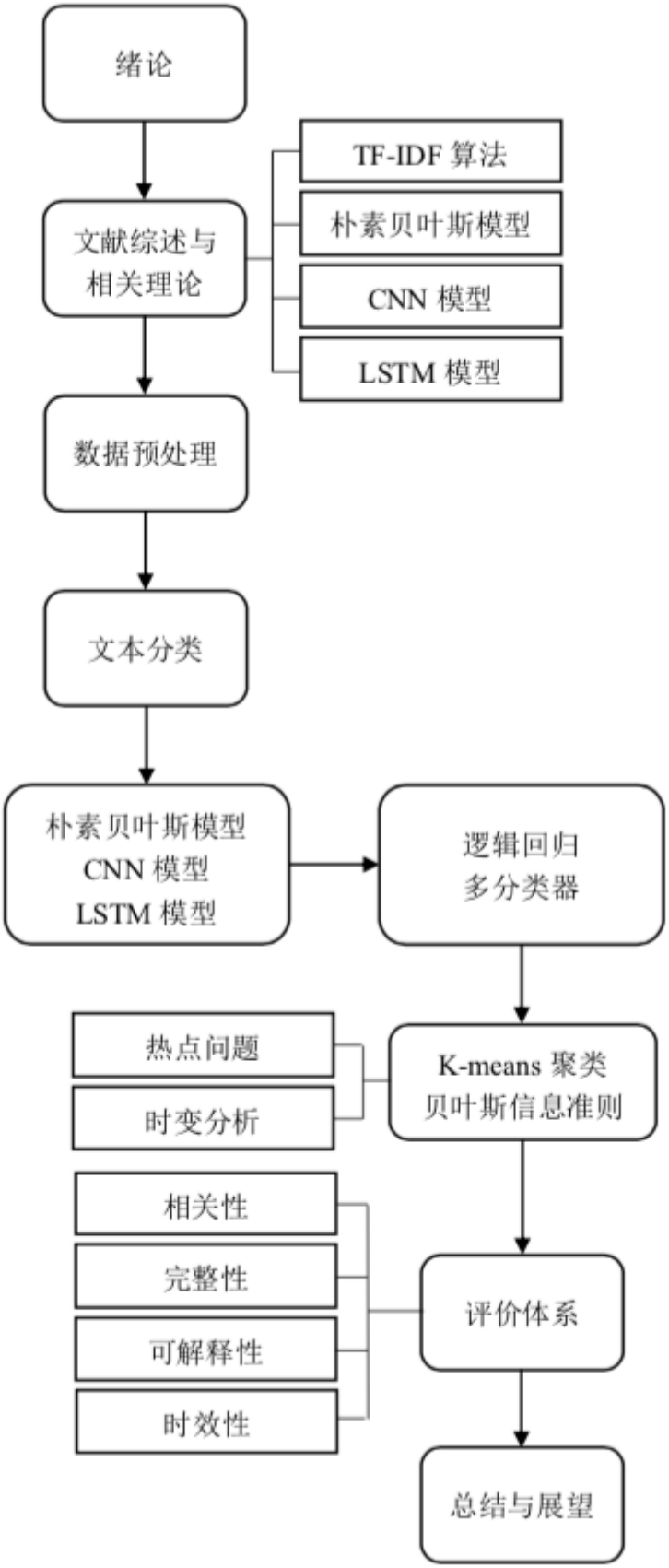
绪论
文献综述与
相关理论
TF-IDF 算法
朴素贝叶斯模型
CNN 模型
LSTM 模型
数据预处理
文本分类
朴素贝叶斯模型
CNN 模型
LSTM 模型
逻辑回归
多分类器
热点问题
时变分析
K-means 聚类
贝叶斯信息准则
相关性
完整性
可解释性
时效性
评价体系
总结与展望

图 1.1　技术路线图

1.3 文献综述及相关理论

随着大数据时代技术的不断创新和深化，云计算、大数据、互联技术正悄然地改变着我们生活的环境：世界的基础结构正在向“智慧”模式不断发展，人类的生产生活及各行各业的管理环境正逐渐联结成一张张“智慧”的数据网。目前，我国的城市管理体系已向智慧城市方向发展，推广之，政府部门的管理形态也在顺应时代的发展，向智慧政务迈进。然而，政府问政平台面临的留言归类处理和热点问题的追踪的工作量大、效率低，智慧政务建设亟需改进。针对此类问题，我们搜集了专家分别对文本分类、热点问题分析、网络回复评价等方面的研究，进行了进一步地探究。

1.3.1 文本分类

李寿山(2010)提出，针对文本分类问题，目前主流的方法是基于机器学习的分类方法。利用统计及其学习分类方法学习标注样本，利用学习好的分类器测试非标注样本，这种方法在性能上比其他基于规则的方法有着明显的优势，因此选择合适的分类方法成为分类研究的一个重点问题[2]。王国薇(2019)研究介绍了，目前较为主流的传统机器学习文本分类方法和深度学习文本分类方法，包括朴素贝叶斯、K-最近邻、决策树，神经网络方法包括：卷积神经网络(CNN)、双向长短时记忆网络等模型方法，利用集成学习方法进行了实验，并指出高效、准确地对所需要的文本进行分类尤为重要[3]。

1.3.2 热点问题探究

李情情(2016)提到了热点和热度值的相关定义。热点：在过去或者当前某一时间段内，被比较多人关注或集中关注的信息点。热度值：对信息点的点击量与关注度的量化，反映了用户对哪个话题感兴趣。可见热度值比较小的热点话题可能会被其他热度值大的热点话题影响，因此，热点话题取决于热点和热度值的两个因素[4]。文中给出了如(1)所示的热度值计算公式：

$$H_T = \log(\sum_{j=1}^{N}(N\_traW_j^2 + N\_comW_j) * w_j) \tag{1}$$

$$w_j = \frac{\delta}{t_j - t_0 + 1} + \sigma \cdot S_j \tag{2}$$

公式(1)中，N_traW_j 代表第 j 篇博文的转发数，第 j 篇博文的评论数是 N_comW_j，w_j 是特定用户的权重，即第 j 篇博文作者代表的权重，该话题一天发表的博文总数为 N。在公式(2)中，t_j 表示第 j 篇微博的发布时间，t_0 表示一个话题的第一篇微博，即起始时间，以分钟为单位，并且满足 $t_j \geq t_0$。

由于近年来对热点问题的挖掘研究越来越得到关注，热度值计算公式也在不断改进，同时，在不同的文献研究中，有不同版本的热度计算方式。

孙胜平(2011)定义了微博热度计算公式，如公式(3)所示：

$$HD_i = \log_{10}(fl_i + 1) + \sqrt{re_i} + cm_i \tag{3}$$

公式(3)中，re_i、cm_i和fl_i分别表示微博D_i的转发数，评论数和其发布者的粉丝数[5]。

柏建普等人(2013)研究发现话题影响力可以反映话题的热度，并定义了话题影响力公式，如公式(4)所示：

$$f_T = \sum_{i=1}^{n} f_{w_i} \tag{4}$$

公式(4)中，f_{w_i}代表每个词语w_i的影响力，又满足$f_{w_i} = \sum_{j=1}^{m} f_{b_j}$，每条微博影响力用$f_{b_j}$来表示，将其定义为微博的用户关注度，以此计算话题热度值[6]。

而本文的热度值计算方法将在第 4 章详细描述，在这不再赘述。

计算热点问题的热度前提是在留言中挖掘出相关的问题，因此在热点问题发现的整个流程中，文本聚类是重要的组成部分。聚类分析是以相似性为基础，在一个聚类中的模式之间比不在同一聚类中的模式之间具有更多的相似性，聚类分析的方法当前应用较为广泛。目前比较成熟的聚类算法，可以分为划分法(Partitioning Methods)、层次法(Hierarchical Methods)、基于密度的方法(density-based methods)、基于网格的方法(grid-based methods)、基于模型的方法(Model-Based Methods)。[7]而文本聚类的主要过程为：先记录分词并计算权重；再构建 N 维空间向量模型；在空间向量模型中通过相似性计算得到最相近的向量文本，然后将相似的文本合并得到各类，最后进行验证。基于以上分析，在聚类时应当选取合适的方法，并利用其优点改进不足，提高聚类质量，完成热点挖掘的重要步骤。

1.3.3 回复评价系统

杨升平(2018)等人建立了判定回复信息质量优劣的准则，由此，建立了律师回复信息质量评价模型。对数据库中各个律师的问答文本进行了量化分析，结果表明，该模型能够很好地评估律师的回复质量。[12]在张钫炜(2018)中运用了评价系统对训练文本进行分析和标注，进而构建情感词库，然后采用分类算法中的支持向量机(SVM)来实现对测试文本情感倾向的分析。[13]

综上所述，学者们已用多种不同类型的模型对“文本分类”、“热点问题挖掘”和综合评价体系进行了研究和阐述。基于此，本文针对留言“文本分类”构建不同类型的分类器比较选择性能较好的类别；再者，针对“热点问题发现”，我们将选择 K-means 聚类分析结合贝叶斯信息准则，完成文本的聚类和热度指标的测算；最后，为问政平台的留言回复质量构建综合评价体系，能够对留言回复的质量进行合理地评价。

第2章 数据预处理

2.1 数据筛选

针对附件 2，由于“留言编号”、“留言用户”、“留言时间”在分类中的作用不大，因此，本文选择了“留言主题”、“留言详情”及“一级指标”作为可用数据。“留言主题”即为留言详情的主要内容，在分类中起到重要导向作用，我们将两者指标合并，可以提高关键词的比重，使得后续分类更加精确。附件中的“一级指标”有 7 类，为了后续的分类处理时的便利，本文直接将其赋予“0-6”的数字作为其分类标签。

针对附件 3，“留言编号”相当于每一条留言特定的索引，在处理数据过程中，数字序号即可代替其功能。而“留言用户”由于出题方给的信息不足，暂且无法根据留言者的身份来判断该用户是否为特定用户，无法赋予不同权重。因此，这里本文剔除“留言编号”和“留言用户”两类指标的数据。

针对附件 4，“留言编号”在此也就相当于每一条留言特定的索引，在处理数据过程中，数字序号即可代替其功能。“留言用户”的信息不足，也只是充当每条留言的识别作用，因此删除“留言编号”和“留言用户”两列指标数据。由于附件 4 需要自己构建评价体系，接下来的预处理过程中，暂先不对文本进行 TF-IDF 的向量化处理。

2.2 分词

由于中文没有明显的分界符，词与短语之间的分界定义模糊，因此在文本中提取信息时需要先进行分词。分词作为数据挖掘的最关键的流程之一，分词的正确性对后续建模分析的精确度有着重要的作用。针对中文里很多词句不同的人有不同的理解方式，同一字词在不同语句的含义不同等问题，本文采用目前 Python 中最好的中文分词组件——jieba 对文本数据进行分词。jieba 分词中，首先通过对照典生成句子的有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取。对于未登陆词（不在词典中的词）使用 HMM 进行新词发现。

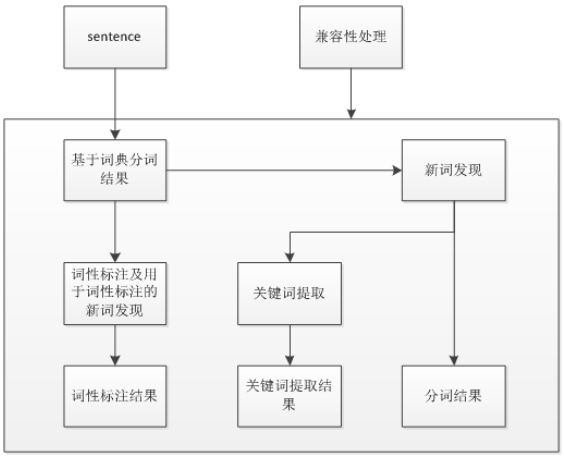

图 2.1　jieba 分词整体工作流程图

2.3　去停用词

停用词是指人类语言中包含的极其普遍的功能词，与其他词相比，这些词没有实际含义，另一类是应用十分广泛的词汇词，但这样的词搜索引擎无法保证能够给出真正相关的搜索结果，难以帮助缩小搜索范围，同时可能还会降低搜索的效率。而在 jieba 分词之后，结果中仍然包含有停用词，本文取用中科大等多个官方公布的停词表，就分词后的停用词进行删除。分词并去除停用词后的部分样本数据，如图 2.2 所示：

分词
经济 学院 体育 学院 变相 强制 实习　书记 您好 西地省 经济 学院 体育 学院 一名 大四 学生 系里 实习
经济 学院 强制 学生 外出 实习　经济 学院 强制 届 电子商务 企业 物流 专业 实习 企业 物流 专业 实习 6 月
请求 地铁 2 线 梅 溪湖 CBD 处 增设 站　领导 游 A3 区山 A3 区 山下 湖 大内 堵 半个 小时 停 好车 联想 上周
人才 app 申请 购房 补贴 通不过　朱琦梦 年 月 落户 并于 年初 A6 区 首次 购房 硕士 毕业生 符合 人才 购房 补
职能部门 相互 推诿 A5 区 植物园 社区 山水 熙园 小区业主 苦不堪言　尊敬 胡书记 您好 A5 区洞 井 街道 植物
市星沙 大道 开元 路 路口 优化　市民 住 星沙 市区 上班 市民 住 市区 经开区 上班 市区 环线 万家 丽路 三一
电建 星 湖湾 强制 业主 收房　星 湖湾 洋房 二期 首批 质量 小区 品质 诉求 书 星 湖湾 洋房 二期 购房者 怀着
A1 区 万国 城 moma 未经 业主 同意 强建 养老院　A1 区 万国 城 moma 三期 业主 年 月 日 晚间 发现 位于 小
A1 区 万国 城 新建 养老院 违规　A1 区 万国 城 业主 新建 养老院 几点 投诉 1 建 民意调查 方式 调查 调查 A1
冰雪 天气 A7 县 校车 停运　接通 明天 校车 停运 学生家长 自行 接送 请问 教育局 校车 出事 承担责任 家长
反对 L 燃气 垄断　5 家里 换 液化气 洗澡时 发现 漏气 打个 电话 人接 晚上 七点 停止 营业 搞 全家 讲卫生 昨天 打个 电话
l 重点 文物 文昌阁 寂寞 荒废 几十年 图　文物 国家 再生 文化 资源 我市 悠久 历史 文明 见证 市级 重点 文
寒冬 来临 请 关注 南方 供暖　寒冬 来临 南方 供暖 小视　写 博客 关注 百姓 呼声 现实 抒发 情感 小资 情调 意义 不
西地省 企业 退休 人员 养老金 调整 高工 倾斜 应分 正高 付高　前 几年 企业 退休 人员 养老金 调整 高
城乡居民 大病 保险 工作 指导 意见 实施 执行　国家 发展 改革 委 卫生部 财政部 人力资源 社会保障
请 H 市人 社局 公布 工勤 技能 岗位 考试分数 最低 合格 线　基层 没关系 事业单位 工勤 职工 参加 全
省 社保局 企业 退休 人员　今日 接到 B5 县 居委会 电话 通知 身份证 退休证 居委会 报 登记 年 元月

图 2.2　部分分词结果图

2.4　TF-IDF

由于目前计算机只能对数字进行理解识别，为了方便我们后面使用机器学习等方法来训练模型，我们需要将自然语言相关的文本形式数据数值化处理。

本题的数据量较大，特征维度较大，为了确定最佳方案，我们分别使用

Word2Vec 算法，TF-IDF 算法与 One-hot 编码算法对语料进行词嵌入处理，并将测得性能进行了对比。经过实验发现使用 One-hot 编码由于任意两个词之间变得孤立，丢失了语言中的词义关系；且附件 2 给的数据量 9200 多条，远远不够深度学习 Word2vec 模型训练的样本数目，数据的后续建模准确率不高。为了保证各个分类器之间的可比性，本文在第 3 章训练分类模型时，均选用 TF-IDF 算法对文本词向量化。

在 TF-IDF 算法中，TF 就是记录某些词语在一篇文章中的频率，IDF0 值就是记录某些词语在所有文章中出现的比例。该算法根据提取的特征词计算特征值，TF-IDF 值就是 TF 和 IDF0 的结合值，由于一个越大越好、一个越小越好，直接相乘不合适，所以将 IDF0 值变个形式——IDF=ln(1/IDF0)，这样得到的 IDF 就是越大越好。TF-IDF=TF*IDF，其值越大，分配得到的权重越高，这个词反映的特征则越强、重要程度也越高。

我们通过 sklearn 的 TfidfTransformer 和 CountVectorizer 模块来计算文本 tf-idf 的值。其中 sklearn 的 tf-idf 计算公式略有不同：

$$idf(t,d) = \log\frac{n_d + 1}{1 + df(t,d)}$$

$$tf - idf(t,d) = tf(t,d) * (idf(t,d) + 1)$$

其中，TfidfTransformer 是直接对 tf-idf 做归一化。而 TfidfTransformer 默认使用 L2 归一化，通过与一个未归一化特征向量 L2 范数求出比值，使得返回向量的长度为 1。

第3章　留言文本分类

在处理网络问政平台的群众留言时，工作人员需要按照一定的划分体系对留言进行分类，以便后续将群众反映的问题分派至相应的职能部门处理。根据当前已有的分类数据，构建合适的文本分类模型对政府部门工作效率起着显著的作用。

本文考虑到 TF-IDF 算法得出的特征矩阵具有高维稀疏性，因此在选用传统机器学习的模型分类器作比较时以线性模型、回归模型为主。

3.1 分类器比较选择

3.1.1 朴素贝叶斯分类器

给定类标号 y，朴素贝叶斯分类器在估计类条件概率时假设属性之间条件独立。条件独立假设可形式化地表述如下：

$$P(X \mid Y=y)=\prod_{i=1}^{d} P(X_i \mid Y=y)$$

其中每个属性集 X={X_1，X_2，…，X_d}包含 d 个属性。

分类测试记录时，朴素贝叶斯分类器对每个类 Y 计算后验概率：

$$P(Y \mid X)=\frac{P(Y)\prod_{i=1}^{d} P(X_i \mid Y=y)}{P(X)}$$

为了得到可靠稳定的分类模型，本文这里采用 5 折交叉验证法。我们将数据集划分为 5 份，轮流选择其中一份作为验证集，每次训练时将另外四份作为训练集放入贝叶斯分类器中进行训练，最后利用验证集来测试训练得到的模型。

经过 5 次对朴素贝叶斯模型的训练，计算并得到拟合测试集时的准确率。将 5 次计算得到的值求平均，准确率的值越高则模型越理想。

计算结果表明，五折分层交叉验证的拟合度结果如表 3.1 所示：

表 3.1　五折分层交叉验证的拟合度结果表

交叉验证次数	Accuracy
第一次	65.418%
第二次	66.992%
第三次	65.201%
第四次	66.992%
第五次	64.135%

五折交叉验证的平均准确率为：65.748%，结果表明训练好的分类模型具有一定的准确性，但是该模型离我们预期的准确率仍有待提高。

3.1.2 CNN 分类器

CNN（卷积神经网络）是将多维数据用卷积核做多层卷积运算、池化、tanh

激活降维，再将多个层融合之后采用 softmax 算法得出输出结果的过程。除了在计算机视觉方向应用很广，CNN 最适合做的是文本分类，由于自然语言处理时任务的输入的是句子或者文档的矩阵表示形式，矩阵的每一行代表一个词向量。

基于 tensorflow，本文的 CNN 训练模型定义如下：

1、利用 tensorflow 自带的 VocabularyProcessor 将句子转化成词向量的 list，再转成矩阵图，这样就将文字问题转化成了图像问题；然而 tensorflow 的卷积转化操作需要批次(batch)、宽度(width)、高度(height)和频道(channel)这样一个四维张量，这样就够成了第一层：embedding 层。

2、接下来做卷积和池化取最大特征，因为卷积时采用了不同大小滤镜，结果必然是不同形状张量，因此需要加一层处理，将结果合并成一个大的特征向量。

3、加入 dropout 层，防止过拟合。

4、通过执行矩阵乘法获得最终类别，计算出 Loss 和 Accuracy，用于训练时纠偏。[10]

3.1.3 LSTM 分类器

长短时记忆网络(Long Short Term Memory Network, LSTM)，是一种改进之后的循环神经网络，可以解决 RNN 无法处理长距离的依赖的问题。原始 RNN 的隐藏层只有一个状态，其对于短期的输入非常敏感。LSTM 再增加了一个状态，让它来保存长期的状态，称为单元状态(cell state)。

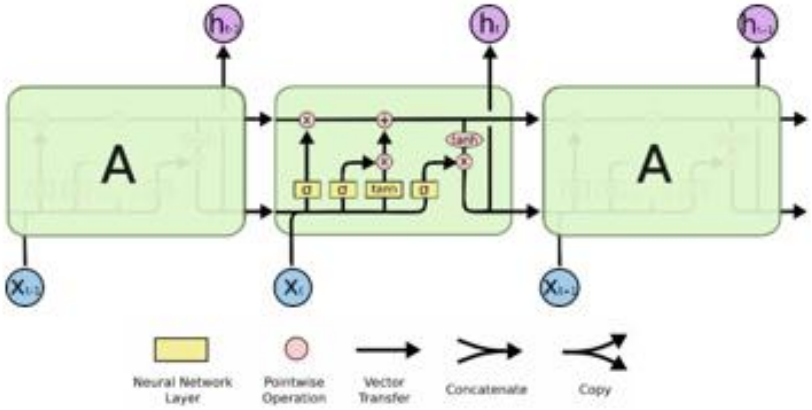

图 3.1 LSTM 结构

本文采用 LSTM 分类文本的原理，如图 3.2 所示。整个模型划分为两个部分：

第一部分：词的特征工程提取处理。

将数据按比例划分训练集和验证集，把每个句子生成相应的 mask 向量，用以标记每个输入文本的实际长度（在后期的模型中根据 mask 向量将 padding 为 0 部分所对应的隐藏层输出砍掉）。

第二部分：改进 RNN 的分类器。

每个词经过 embedding 后，进入 LSTM 层，然后经过一个时间序列得到 n 个隐藏 LSTM 神经单元的向量，这些向量经过 Mean pooling 层之后，得到一个向量 h，最后通过 Softmax 函数得到一个类别分布概率向量，取概率值最大的类别作为最终的输出结果。[9]

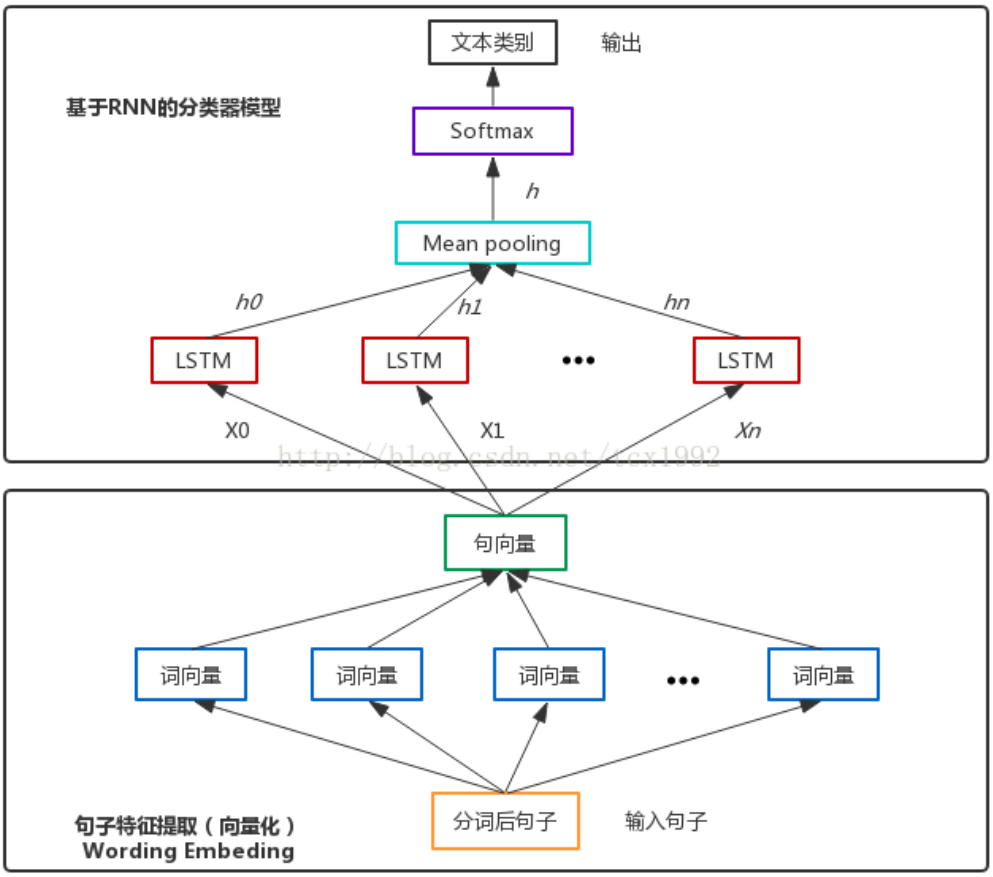

图 3.2　文本分类示意图

3.1.4　实验结果

本文主要采用 Python 环境进行实验。其中用到的 Python 库有：

Tensorflow：Google 开源的第二代用于数字计算(numerical computation)的软件库，近些年已经成为深度学习较为主流的系统框架。本文将使用该库进行有关深度学习模型的建立和训练。

Numpy：Python 语言的一个扩展程序库，支持高维度的数组矩阵运算，此外也针对数组运算提供大量的数学函数库。对于本次实验的高维度数据较为适用。

Sklearn：机器学习中常用的第三方模块，对常用的机器学习方法进行了封装。本文将调用该模块对数据和传统机器学习的方法进行处理及模型的建立。

在训练 CNN 和 LSTM 时设置的相关参数，如表 3.2 所示：

表 3.2　训练时的相关参数

参数	参数值
词嵌入维度	200
Batch_size	32
epochs	10
dropout	0.2
激活函数	Softmax
张量长度	7

表 3.3 CNN 分类器准确率

交叉验证次数	Accuracy
第一次	45.318%
第二次	53.467%
第三次	48.908%
第四次	51.293%
第五次	52.447%

表 3.4 LSTM 分类其准确率

交叉验证次数	Accuracy
第一次	55.418%
第二次	49.492%
第三次	50.081%
第四次	51.342%
第五次	48.151%

对 CNN 和 LSTM 模型也进行了 5 折交叉验证并得到模型的准确率，如表 3.3、表 3.4 所示。

通过上述建立传统方法的模型和深度学习方法进行的实验比较分析，本文观察发现在平台所给的样本量下，传统机器学习模型的分类结果优于深度学习方法。但考虑到 66%的准确率不够高，我们对数据进行进一步地建模分析。

3.2 逻辑回归分类器

经过上述多种分类器的比较，传统机器学习分类表现的性能更佳。因此，本文再尝试使用机器学习算法中性能较为优越的逻辑回归分类器，为寻找更优方案。

普通的逻辑回归算法针对二分类问题，而本文处理的数据为多分类问题。对于此，我们选用改进后的逻辑回归作为分类器的算法。改进后的方式主要为两种：OVR(One VS Rest)和 OVO(One VS One)。

OVR 主要是指将多个分类结果（假设为 n）分成是其中一种分类和其它分类的和，直接根据每个类别，都建立一个二分类器，带有这个类别的样本标记为 1，带有其他类别的样本标记为 0。这样便可以有 n 种分类的模型进行训练，最终选择得分最高的类别（概率值最大）作为分类结果。

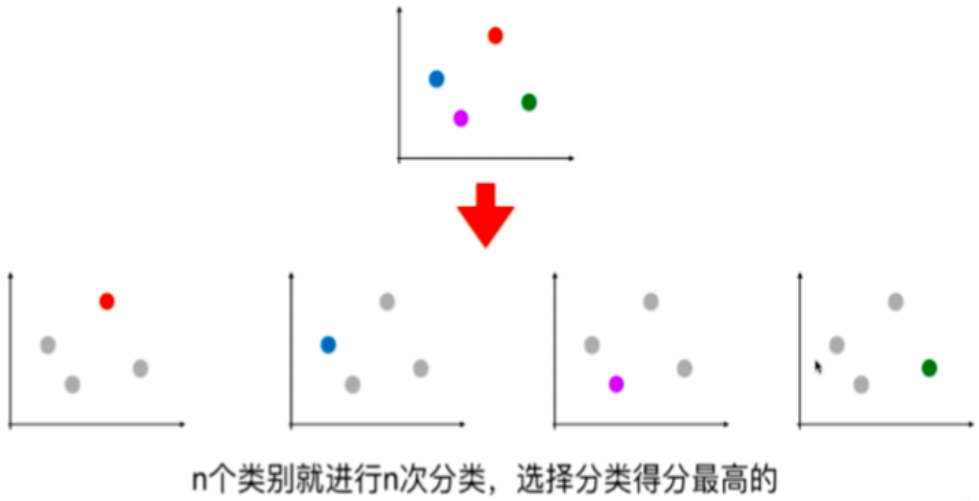

图 3.3　OVR 示意图

OVO 主要是将 n 个数据分类结果任意两个进行组合，然后对其单独进行训练和预测，最终在所有的预测种类中赢数最高的即为分类结果，这样的分类方式最终将训练分为 n(n−1)/2 个模型，虽然计算时间相对较长，但每次训练各个种类之间不混淆也不影响，分类结果较为准确。

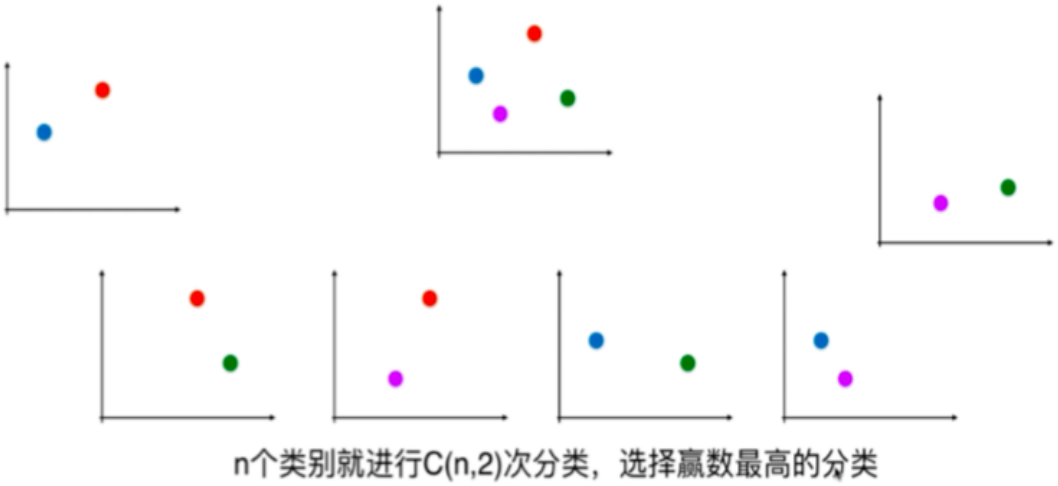

图 3.4　OVO 示意图

在 sklearn 的 LogisticRegression 中提供了 2 种不同的惩罚项(penalty)参数选择，其中 L1 相当于 lasso 回归，L2 相当于 ridge 回归。由于 lasso 回归会使参数矩阵较 ridge 回归更加稀疏，且 L2 的拟合能力要比 L1 的拟合能力强，在文本的高维度数据条件下，本文选择了“L2”参数。在传统分类器中，本文的样本数量较大，且样本量远大于特征量，我们将 dual 参数设置为 False，选择目标函数为原始形式不进行对偶化转换处理，减小误差。

对于 sklearn 中的 OVR 和 OVO 的多分类方式设置的参数形式有些不一，应当将参数 multi_class 设置为 ovr 和 multinomial，而且 solver 这一参数也必须指定为 newton-cg。我们将 multi_class 分别设置为 ovr 和 multinomial，进行训练，并比较其中的结果。

我们首先将数据集也进行了测试集和训练集的划分，采取五折交叉检验的方式对模型进行训练，并得到各次测试的准确率，如表 3.5 所示：

表 3.5　逻辑回归多分类器的准确率

模型 / 准确率	OVR	OVO
第一折	87.039%	87.547%
第二折	86.985%	88.203%
第三折	89.419%	90.349%
第四折	90.005%	91.494%
第五折	88.357%	88.193%
均值	88.361%	89.157%

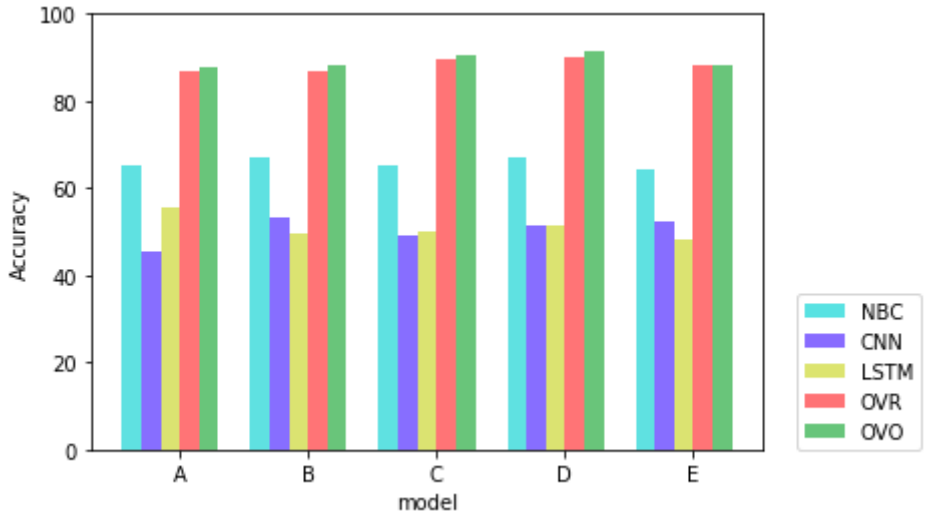

图 3.5　模型准确率对比图

对比上述四类准确率，逻辑回归的准确率最高，因此我们选定逻辑回归多分类器的 OVO 形式作为我们的文本多分类模型。

出题方给出了分类通常使用的评价指标为 F-Score：

$$F-Score=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

其中 P_i 为第 i 类的查准率(Precision)，R_i 为第 i 类的查全率(Recall)。

查准率：

$$\text{Precision}=\frac{TP}{(\text{TP}+\text{FP})}$$

查全率：

$$\text{Recall}=\frac{TP}{(TP+FN)}$$

TP：真阳性（true positive）；FP：伪阳性（false positive）；TN：真阴性（true negative）；FN：伪阴性（false negative）。

基于此，我们再次对选定的逻辑回归多分类器选用评价指标 F1-Score 进一

步对模型进行判断和比较。

表 3.6　逻辑回归多分类器的 F1-score

折数	F1-Score
第一折	0.8747
第二折	0.8720
第三折	0.9034
第四折	0.9049
第五折	0.8819
均值	0.8874

据表 3.6 可知，逻辑回归 OVO 多分类器的 F1-score 评价指标均值为 0.8874，该分类器的性能较好，适合“智慧政务”问政平台的留言分类。

第4章　热点问题挖掘

4.1 K-means 聚类

为了有助于政府相关部门提升“智慧政务”体系下的服务效率，及时发现热点问题并能够有针对性地处理，避免热点问题的扩散给城市管理带来不良的影响，成为了当下“智慧政务”亟待解决的问题。

热点问题发现系统的核心部分是对留言文本内容的聚类分析。经过聚类后的文本的各个类别均反应了一个问题，通过进一步地热度指标的测算和比较分析，并抽取出各条留言中的文本，就可以发现留言文本中的热点问题。

文本挖掘的一个重要领域就是聚类分析。聚类分析作为一种无监督的机器学习方法，由于不需要前期的数据训练拟合过程，且不用预先对文本数据进行手工标注分类，因此在对未知的热点问题发现与探索时，运用聚类分析的算法具有一定的灵活性和较高的自动化处理能力。

在数据预处理阶段，我们已经对文本数据进行了处理，因此这里可以直接进入聚类分析部分。现阶段的文本聚类分析的方法有很多，没有一种方法普遍适用于任何数据集。由于我们利用 TF-IDF 算法对文本向量化处理后，特征维度较大，为了避免计算量过于庞大造成模型的泛化能力低下，本文选取了较为简单、聚类效果较好的 K-means 聚类分析的方法。

K-means 聚类分析算法是现在聚类算法中最常用的一种算法。K-means 算法是一种典型的基于划分的聚类算法，该聚类算法的基本思想是在聚类正式开始前预先设定类簇数目 k，并在数据集中随机地选择 k 个对象作为 k 个初始类簇的中心，对于文本集中剩余的每个对象，根据对象到每一个类簇中心的欧几里得距离，划分到最近的类簇中；全部数据归入类簇中将重新计算每个类簇的中心点，再计算每条文本数据和新的类簇中心的距离，将数据点重新划分至最近的类簇；不断重复直至所有样本都不再重新分配为止。其中欧几里得距离计算公式如下：

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

由于 K-means 聚类分析的 k 值需要预先设定，所以在对 k 值选取时具有一定的随机主观性，会影响模型的泛化性能。时志芳(2013)提到通过使用模型选择法来得到聚类数目，而且有实验证明，BIC 准则估计聚类数目类的误差很小[11]。据此，我们利用贝叶斯信息准则来判断聚类数目作为预定 k 值得标准，减少人为判断的主观影响。

贝叶斯信息准则(BIC)用于模型选择，近似等价于基于贝叶斯因子的模型选择方法。该方法引入了模型参数的相关惩罚项，防止模型训练时，由于参数数量的增加导致模型复杂度的增加因而增大了似然函数，造成过拟合现象，这样就避免了模型精度过高而形成的模型复杂度过高。贝叶斯信息准则的公式为：

$$BIC=k\ln(n)-2\ln(L)$$

其中，k 为模型参数个数，n 为样本数量，L 为似然函数。$k\ln(n)$ 惩罚项为

了避免文本特征向量较为稀疏的情况下出现维数灾难的现象。

在此，向量文本输入的数据集 X={x_1,x_2,…x_n}满足独立同分布，设定后验概率 f(M_i|w)，我们定义 k 值的集合 W(k)为：

$$W(k)=\{f(M_i\mid w)\mid w=(\lambda_1,\lambda_2,...,\lambda_k),w\in\Omega_k\}$$

在 BIC 准则中，ln(L)表征了 $f(M_i\mid w)$ 函数模型在 X 上的极大似然后验概率值。当函数模型 f 满足 $\hat{f}=\arg\max_{f\in M(k)} BIC(f)$ 时，BIC 达到最大值[11]，基于此选定聚类的数目。根据计算，并根据实际情况，在选定的 k 值附近进行调参，我们得到了以下的聚类结果，由于类簇较多，我们抽取了一部份数据结果进行展示。聚类结果如所示。

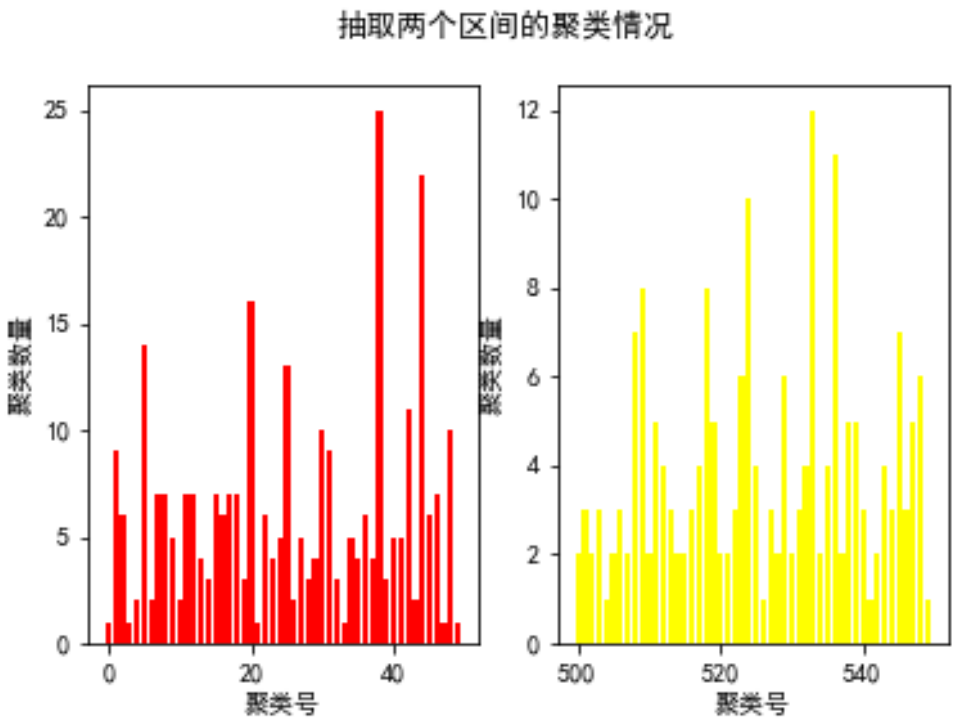

图 4.1　聚类结果图

4.2　热度指标测算

问题的热度就是反应了某一段时间内群众对该类问题的关注程度，热度越高，群众对问题越关心，达到一定程度即成为热点问题。

通过大量的文献查阅，针对“智慧政务”系统中热点问题的挖掘，本文总结了影响问题热度的 3 个主要关键因素：该条留言的权重大小；关键词的竞争力；主题内容。

在“智慧政务”的网络问政平台中，每条留言的权重大小体现在留言的点赞量和反对量。据生活经验和大量的调查研究，我们得知非热点问题的留言在平台中不会引起大量网友和群众的关注，也不会引发群众对该问题的评论。此处的点赞数和反对数理解为群众对该条留言对应的政策现象的支持和反对现象。因此这

两个指标反应问题的热度情况权值应等同。

由于每条留言所含有的关键词数量不同，且每个关键词对话题热度的影响力不同，因此我们将上文聚类好的每一类问题利用TF-IDF算法抽取前5个关键词，计算每个关键词的TF值，即词频。对此我们计算每一条留言中关键词的权重，即每条留言的关键词的词频和，作为关键词的竞争力。

因此，我们为“智慧政务”网络问政平台提出了如下的热度计算公式：

$$\text{热度}=\mathrm{p}\cdot\log\left(\sum_{\mathrm{i}=1}^{n}\left(\text{点赞量}_i^{\,2}+\text{反对量}_i^{\,2}+\sum_{j=1}^{k}tf_j\right)\cdot(t_i-t_0+1)\right)$$

其中，$p=\dfrac{s_k}{\sqrt{\sum s_k^2}}$，$s_k$为上文聚类后第k类问题的留言条数。$t_i$表示第i条留言的发布时间，$t_0$表示一个热点问题的第一条留言，即起始时间，以分钟为单位，并且满足$t_i \geq t_0$，最后加一是为了防止$t_i=t_0$该项等于0，造成指数计算出现无穷的情况。

据此，我们对聚类好的问题进行了热度的测算，测算热度热度值前五的部分数据结果如图4.2所示。由于所需的结果数据量较少，对地点/人群和问题描述的提取，本文不再采用文本挖掘中的方法进行自动提取。

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	195	4.66	2019/8/15	A市A5区汇金路五矿万境附近居民	小区物业管理十分糟糕，交通以及各种住房安全问题亟待解决
2	186	4.483	2019/4/11	A市金毛湾小区居民	A市金毛湾配套入学与开发商描述不符，消费者权益受到损害，且屡次维权失败
3	372	4.465	2019/1/11	A市58车贷特大集资诈骗案受害者	A市58车贷案审查情况投诉
4	890	4.365	2019/9/1	A市A4区绿地海外滩二期居民	小区建设不合理，旁边的赣州高铁噪音扰民
5	213	3.061	2019/6/19	A市富绿物业丽发新城业主	A市富绿物业丽发新城强行断业主家水及周边搅拌站噪音扰民

图4.2　热度结果图

第5章　留言回复的评价体系

随着网络问政平台的关注度提高，政府部门也意识到提高“智慧问政”体系下的行政办事效率的重要性，如何利用好问政平台的留言信息和回复信息对于“智慧政务”的发展变得越来越重要。

然而，政府人员在该平台的回复信息质量存在很大差异，使得群众对政府平台的工作质量评价不一。为了提高政府的工作质量，建立一个留言回复的评价体系既可以为留言回复进行质量的评判，也可以为政府工作人员的工作质量进行量化考核提供参考依据。

在该网络问政平台有些留言的回复是无用信息，比如以下留言主题和答复：

留言主题一：请求处理 K4 县城锦豪雍景园小区商品房的费用问题。

回复：已收悉。

留言主题二：请求解决蒋垄家火车噪音问题

回复：网友：您好！留言已收悉

留言主题三：请求解决 L1 区华都小学小孩上学无公车的问题

回复：L1 区网宣办 2015 年 9 月 18 日

上述留言的答复严重影响了平台群众的体验感和回复信息的质量。例如，在留言一和二中，平台的回复只给出了“已收悉”，但是没有具体列出解决方案和留言问题可答复的部门及相关联系方式。在留言三中政府人员回复了“L1 区网宣办 2015 年 9 月 18 日”，完全答非所问，政府部门的答复只是反映了该条留言已被查阅，政府工作人员应当给出具体的情况和答复信息。

本文根据出题方给的问政平台系统历史数据库中的留言回复信息，提取政府平台的留言回复信息质量优劣的特征，研究建立通过综合不同角度对留言回复信息进行量化描述，建立衡量回复内容相关性、完整性、可解释性和及时性等方面的综合评价体系的数学模型。该模型的研究建立有助于对网络问政平台留言的回复信息进行评价，监督政府的工作质量，使得群众能够享受更加优质的“零次跑”问政服务。

5.1　留言回复质量优劣特征

对于在海量的回复数据下，如何识别出留言的最佳回复，如何评价留言的回复质量，是“智慧政务”的网络问政平台发展亟需解决的重要问题。根据我们的经验和给出的数据的浏览判断，我们对该平台的留言回复质量的优劣特征进行归纳总结，得到以下作为本文评判的参考特征。

(1)良好回复质量的参考特征：1)回复内容和留言主题及详情的具体信息较为匹配，其关键词信息和留言内容的词语语义相似度高；2)回复内容的长度适中，适宜网友网上浏览时的长度；3)留言的回复内容充实，具体内容具体分析；4)留言回复的时效性，回复时间较为及时，可尽早处理留言问题。

(2)较差回复质量的参考特征：1)对不同留言使用同一回复，或回复内容的相似度很高；2)回复内容过于简短、无实质内容或回复过长内容冗杂，例如“您好，已收悉”等；3)回复内容套用固定模板，复制修改采用其他留言回复格式，例如

“……详情咨询***部门”等；4)留言回复时间间隔过长，失去了回复的意义。

在海量数据之下，如何基于这些特征利用自然语言处理和文本挖掘技术对留言回复进行评价也是一个极具研究价值的问题。下面从不同因素对网络问政的留言回复质量进行量化描述，从而建立评价体系模型。

5.2 留言回复质量的量化

5.2.1 相关性

留言回复内容是否和对应的留言详情相关性大小，需要充分考虑两者内容的匹配度。根据匹配度得出相关性可以判断留言回复质量，可以认为在留言回复语句中所用的词语和留言中关键词的语义相似度越高，对留言反映的问题的答复越多，则回复的质量越高。回复过程中，要避免提供过多无效信息而影响评判，如下面的回复，字数虽多但相关性小信息量少，因此没有解决问题。

为此，我们定义了函数 sim(i)来描述留言中留言回复 w_i 与留言内容 p_i 的相似度。本文中 sim(i)的计算方法基于 gensim 库中的 Word2Vec 包。

与传统的词向量相比，word2vec 训练出的词向量避免了词语表示的“维数灾难”和“高稀疏”的问题，实现了连续 bag-of-word 模型和计算词向量的 skip-gram 结构来将文本中的词语表示为词向量。这些词向量可以作为词的特征应用到自然语言处理问题中。由于本文构建的质量评价体系是自己的函数定义，因此这里我们利用 word2vec 训练词向量避免高稀疏性对我们的函数结果造成影响。有文献检验过这种方式适用于在海量的网络回复数据中计算词语的相似度。[12]

利用 word2vec 算法对出题方给出的网络问政平台系统的数据进行训练，得到该个文件的文本数据的词向量，并计算每一条留言和其对应的回复内容的相似度 sim(i)。

$$sim(i)=\frac{w_i \bullet p_i}{\|w_i\|\times\|p_i\|}$$

其中，w_i 和 p_i 是留言回复和留言内容所有词的向量之和表示为：

$w_i=\sum_{j=1}^{n} m_j$，其中$m_j=(x_{1j},x_{2j},x_{3j},\cdots,x_{nj})$，$p_i=\sum_{k=1}^{s} n_k$，其中$n_k=(x_{1k},x_{2k},x_{3k},?\quad x_{nk})$，n 表示向量训练时设置的维度。

5.2.2 完整性

留言回复内容的完整程度和详细性与回复的文本长度有直接的关系。当留言回复内容过于简短，可以认为其回复信息量一般不够，因此评价体系所给的评分应当较低；同时，对于留言回复中较长文本的回复评分也不应该过高。通过对附件 4 数据的浏览，有一部份数据的回复以“您好，已收悉”为模板，该类模板过于简短，在整体评价体系中会对完整性指标影响较大。因此，可以考虑使用对数函数结合上四分位数来量化回复的文本长度与评分的关系，建立“回复内容完整

性”的评价指标 comple(i)。若留言答复的文本长度和上四分位数的标准差越大，完整性指标的得分越低。

$$comple(i)=\log\frac{1}{\left|len(x_i)-0.75\dfrac{\sum_{i=1}^{n}len(x_i)}{n}\right|}$$

其中，x_i 为每一条留言回复的文本内容，$len(x_i)$ 为每一条留言内容文本长度。

5.2.3 可解释性

留言回复的有些较差质量的回复中会出现很多固定词语的模板，如“已收悉”；且政府工作人员在同时回复多条留言时或许会采用相同的模板或者复制自己之前已经回复过的内容，则都表明留言回复的质量不高，对留言内容的可解释性不足。为此，我们建立文本可解释性评价函数 explain，并将 explain(i)作为回复内容 X_i 与总体的一个相似度指标，该值越大，表示 X_i 与总体均值越相似，套用模板等现象越明显，该条留言回复质量越低。下面介绍本文定义的文本可解释性评价函数 explain 的计算方法。

首先，将上文中训练得到的 word2vec 词向量 $(x_{1k},x_{2k},x_{3k},\cdots,x_{nk})$ 相加，得到每条留言回复的向量 $(X_{1k},X_{2k},X_{3k},\cdots,X_{nk})$，对每条留言的向量求和并求均值 $(w_{1k},w_{2k},w_{3k},\cdots,w_{nk})$，作为该类数据集留言回复的模板，将每条留言和数据集的模板向量求相似性。

$$\exp lain(i)=\frac{X_{ik}\bullet w_{ik}}{\|X_{ik}\|\times\|w_{ik}\|}$$

本文相似度计算利用了词语间向量的求和平均相似度，能够减小同义词和多义词在向量中无法识别的情况，解决文本数据中存在的自然语言问题。

5.2.4 时效性

留言回复时间的及时性对回复信息质量评判也尤为重要，回复得越晚，该类问题的解决的效率越低，若问题严重且具有连锁效应，则会对社会造成负面影响。因此，留言的时效性在质量评价时也是一个重要的参考特征。时间越短，指标得分应当越高。本文定义时效性的评价函数为 tm。

$$tm(i)=\frac{1}{t_j-t_i+1}$$

其中，t_j 为留言回复的时间，t_i 为留言发布的时间，以天为单位，并且满足 $t_j\geq t_i$，为了防止留言回复及时在同一天回复的情况因此通过分母+1，避免出现

分母为 0 的情况。

5.2.5 综合评价体系

评价问政网络回复的关键在于如何建立对留言回复质量的量化评价模型，以在海量数据下利用计算机对留言回复进行智能评分，同时可以进行工作人员的工作质量的量化考核。根据上述量化方法，本文定义了适用于此的综合评价体系。

由于相关性、完整性和可解释性均是对文本内容的挖掘分析，其对回复质量的衡量权值应当一样，因此，在通过上述的量化处理后，我们对所得数据再次进行归一化处理，避免三类指标在综合评价模型测算时各指标处于不同的数量级，消除奇异样本数据导致的不良影响。

而时效性在上述所有指标中，发挥的作用略有不同，当时效性已经失去作用，留言问题已被解决时，即使上文的三个指标接近完美，那在现实的评价体系中，该条回复也不再发挥实际作用。基于此，本文定义了留言回复的评价体系：

$$p = 1000 \bullet tm(i) \bullet (sim(i) + comple(i) + \exp lain(i))$$

根据以上评价体系的构建过程，我们测算出了留言回复质量的指标评分，部分数据，如图 5.1 所示：

	答复意见	答复时间差	留言时间	答复时间	留言主题	留言详情	指标评分
0	你好。请向当地民政部门询问。 2017年8月18日	2330	2014-06-14	2017-08-18	咨询低保、残疾人补助的相关问题	\n\t\t\t\t\t\n\t\t\t\t\t 关于低保，残疾人补助，不切实际，落实不到...	0.578369
1	网友“A00087964”： 你好。残疾人的相关政策，请到残联具体咨询。 若是生活困难，符合...	894	2011-04-02	2013-09-12	残疾人可以把福利提高一点吗?	\n\t\t\t\t\t\n\t\t\t\t\t 残疾人真的很可怜，请政府支持，谢谢。\...	0.731671
2	“UU0081103”：您好！您于2018年11月14日在涉涟网络留言发布的帖文已收悉，我局...	2978	2018-11-14	2019-05-28	咨询M5市农村合作医疗保险减免政策	\n\t\t\t\t\t\n\t\t\t\t\t关于农村建档立卡贫困户缴费农合医疗是否免费享...	1.363973
3	UU0082326： 您好！您需要的两个数据现说明如下： 1、2012年https:...	1187	2013-07-24	2014-07-02	咨询l市2012年两个数据	\n\t\t\t\t\t\n\t\t\t\t\t张局长： 你好！打扰了。我因研究洞庭湖生态...	1.430564

图 5.1 评价体系结果图

第6章　未来展望

6.1 文本向量化

在本文分类中，TF-IDF 算法的维度问题，无法解决维度过大的问题，而且在对文本分类时，文本之间的上下文信息，无法彻底考虑其中的关系。虽然留言文本的分类可能不是正确分类，但是本文的逻辑回归多分类器在大多数情况下还是位于返回的正确分类集合中。

6.2 热点聚类

在热点问题聚类时，由于聚类是一种非监督学习，簇心的选取是个非常随机的过程，导致 k 值相同的情况下聚类的结果每次都不一样，但又不能取平均值作为结果，所以聚类的好坏很难被评价出来。在每次聚类后的结果虽然略有不同，但是大致情况一致。

6.3 未来改进

由以上论述可知，我们的分类器可以较为准确地分类。下一步，我们希望继续改进我们的分类模型，在问政平台积累的数据更多的条件下，数据量适用于深度学习训练。利用深度学习中的多头注意力机制捕获与多个分类类别相关的不同单词，多角度多层面获取重要的文本分类中的信息，结合胶囊网络与双向长短期记忆网络(BiLSTM)方法进行特征融合，通过胶囊网络既提取局部的文本特征；又通过 BiLSTM 在同层神经网络之间相互传递，体现相隔较远的文本之间的前后联系，从而提取文本的全局特征。通过特征融合的方式可以利用不同网络模型的各自优点，获得不同层次上的文本信息特征，提高分类效果。并利用上述网络捕获到的文本之间更深层次的关系，将网络间文本输出向量作为聚类的输入，使得聚类效果更加准确。

参考文献

[1] 王振. 基于机器学习的文本分类研究与实现[D]. 南京邮电大学，2018.

[2] 李寿山，黄居仁. 基于 Stacking 组合分类方法的中文情感分类研究[A]. 中文信息学报，2010.

[3] 王国薇. 基于深度学习的文本分类方法研究[D]. 新疆大学，2019.

[4] 李情情. 基于话题热度的微博推荐算法研究[D]. 山东师范大学，2016.

[5] 孙胜平. 中文微博客热点话题检测与跟踪技术研究[D]. 北京交通大学，2011.

[6] 柏建普，田芳. 基于语义分析的微博热点话题发现技术研究[J]. 内蒙古科技大学学报，2013..

[7] https://blog.csdn.net/xiaojimanman/article/details/44977889

[8] https://blog.csdn.net/ray0354315/article/details/77053620

[9] https://blog.csdn.net/wendaoliutou/article/details/95628890

[10]https://blog.csdn.net/ray0354315/article/details/77053620?utm_source=blogxgwz2

[11]时志芳. 移动投诉信息中热点问题的自动发现与分析[D]. 北京邮电大学，2013.

[12]杨开平，李明奇，覃思义. 基于网络回复的律师评价方法[A]. 计算机科学，2018.

[13]张钫炜. 基于评价系统的评论类文本情感倾向性分析[A]. 语言文字应用，2018.