“智慧政务”中的文本挖掘与综合分析

摘要

政府在为公民提供服务的过程中往往会沉淀海量的数据，这些数据与社会经济、公民生活密切相关，随着各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门带来了极大挑战。为更好地发挥数据资源在社会治理模式转变过程中的战略作用，本文利用自然语言处理和文本挖掘方法，实现互联网公开来源中的群众问政留言分类，留言热点问题挖掘以及对群众留言答复意见的评价。

针对问题一：首先将附件 2 中的非结构化数据进行去重去空，然后统计各类别留言数量分布情况，发现存在分布不平衡情况，对其采用“回译”的方法增强数据，针对留言中存在冗余句子的情况，利用 TextRank 算法抽取关键句，最后进行中文分词、停用词处理等工作。然后基于 TF-IDF 权重提取留言特征关键词，形成词汇-文本矩阵，构建机器学习和深度学习模型对留言进行一级标签分类，最后借助 F1 值检验模型效果。

针对问题二：将附件 3 中的留言数据进行预处理后，结合命名实体识别和修正规则，识别出留言主题中的地点/人群，由于存在语言表述差异，对实体进行相似度计算。然后运用 K-Means 聚类方法对留言主题和留言详情进行聚类，同时利用手肘法找出最佳聚类簇 K，形成 K 个主题类，最后形成相同地点、主题的事件簇。最后利用层次分析法构建留言热度评价指标体系，从事件爆发度和群众作用度两个方面构建 4 个评价指标，计算每个事件的得分，给出热度排名前五的事件。

针对问题三：从内容维度、认知维度和情感维度三个角度出发，构建留言回复评价指标体系，设计了完整性 F11、相关性 F12、可证实性 F21、及时性 F22、充实性 F31、礼貌性 F32 六个评价指标，分别计算留言回复在每个指标下的得分，最后利用归一法将各个指标得分范围限定在[0-100]，再计算平均得分，进行降序排序后得到每条留言回复质量评分。

关键词：文本挖掘；机器学习；神经网络；K-Means；命名实体识别；层次分析法

Abstract

In the process of providing service for citizens, the government tends to collect vast amounts of data, which is closely related to the social economy and citizens' lives. With the quantity of various public opinions related text data rising, departments which mainly rely on artificial to do text classification and hot spot identification face great challenges. In order to better play the strategic role of data resources in the process of social governance mode transformation, this paper uses the methods of natural language processing and text mining to realize the automatic classification, the clustering and the evaluation of people's political comments data from open sources on the Internet.

For question 1: at first, we delete those repeated and empty messages from unstructured data in annex 2, then make out the message number distribution. We find the distribution is uneven, so we adopt the method of "translation" to enhance data amount. Considering the redundant sentences in the message, we use TextRank algorithm to extract keywords. And then, we carry on the Chinese word segmentation, stop words processing, etc. Based on the weight of TF-IDF, we extract the characteristic keywords of the message, form word bag and construct lexical text matrix. Finally, we use Support Vector Machine, Convolutional Neural Networks and Recurrent Neural Network classification models to classify these messages, and F-Score is used to test the recall and accuracy of the classification.

For question 2: after the data in annex 3 are numerically preprocessed, we use the named entity recognition and correction rules to identify the location/population. Due to the different expressions in language, the similarity degree between the entities is calculated. Then, we cluster the message subject and message details via K-Means clustering method. Meanwhile, the best cluster K is found by the elbow method to form K topic categories. Therefore, the event clusters of the same location and topic are formed. Finally, the Analytic Hierarchy Process is used to construct the evaluation index system of message heat, and four evaluation indexes are designed from the aspects of event outbreak degree and mass action degree. According to these indexes, we calculate the score of each event, and obtain the top five hot events.

For question 3: we build a message reply evaluation index system. From perspectives of content dimension, cognitive dimension and emotional dimension, including integrity F11, relevance F12, verifiability F21, timeliness F22, substantiality F31, politeness F32. Then, we calculate the score of the message reply under each index separately, and use the normalization method to limit the score range of the index to [0-100], then calculate the average score. Finally, we obtain the message quality score after descending order.

Keywords: text mining; machine learning; neural networks; K-Means; named entity recognition; AHP

目录

1 问题概述

1.1 挖掘目标

《国家信息化发展评价报告（2016）》数据显示，中国在信息产业规模、信息化应用效益等方面获得显著进步，信息化发展指数排名近 5 年得到快速提升，位列全球第 25 名，首次超过了 G20 国家的平均水平。在技术、商业、政策等多重因素的推动下，也开始涌现出一些新的特点，尤其是在作为国家信息化重要组成部分的政务信息化领域表现最为突出。

本次挖掘目标是利用 16350 条政府平台的公众留言信息数据，其中包含结构化和非结构化文本数据，在对留言数据进行基本的预处理、中文分词、去除停用词后，对其进行深度挖掘与综合分析：①根据留言主题和留言详情，基于题目所给的分类标签，分别采用传统机器学习模型中的支持向量机和深度学习模型中的 CNN、RNN，构建关于留言内容的一级标签分类模型；②构建留言热度评价指标，采用命名实体识别方法和我们制定的修正规则识别出特定地点/人群，使用 K-Means 算法对平台留言进行聚类，得到相应类别中的热点地点、热点人群以及热点事件；③构建留言回复评价体系，对平台留言的回复情况进行三个层次的评价，使用层析分析法计算每条留言得分，便于评价及监管政务服务。

1.2 挖掘意义

政府与公民、企业进行互动时，在提供服务的过程中会沉淀海量的数据，这些数据与社会经济、公民生活密切相关，具有数量庞大、涉及面广、动态精准、可用性强等特点。随着各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门带来了极大挑战。不完善的数据分析工具，导致了其所掌握的大量信息资源没有被充分利用，造成了数据资源的浪费，不利于更好地发挥数据资源在社会治理模式转变过程中的战略作用。

在大数据环境下，政府部门希望能高效利用数据，从中挖掘到有效信息，有效提升政务服务质量和效率。因此，发展“智慧政务”已成为提升政府科学管理水平和社会治理能力的重要手段，而建立基于文本挖掘技术的智慧政务系统已成为社会治理创新发展的新趋势，为政府管理提供了可靠的监控信息和辅助决策，为广大群众的生活提供了高效的民政服务，实现政府决策科学化、社会治理精准化、公共服务高效化。

2 分析方法和过程

2.1 问题一

2.1.1 处理流程

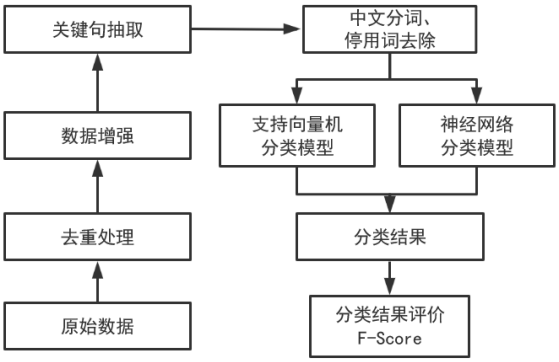


图 1 问题一处理流程

首先对问题一进行数据清洗和预处理，将数据处理成为需要的格式，由于经典的支持向量机分类模型只给出了二分类的算法，在数据挖掘的实际应用中，一般多分类问题是通过多个二类支持向量机的组合来解决，在数据量较多的情况下，矩阵的存储和计算将耗费大量的机器内存和运算时间。基于此，我们针对未来数据量多的情况，同时构建神经网络分类模型，它在大规模数据量处理的性能表现得更好。所以，针对不同的数据量使用不同模型进行数据分类，能够提高分类效果与用户体验。最后，我们使用 F1 值来评价各分类模型的性能。

2.1.2 数据预处理

（1）数据去重

题目所给的数据有一些重复数据，可能是同一用户多次提交问题导致，这些重复数据的特点是留言用户、留言主题、留言详情、一级分类是相同的，但是留言时间不同，同一问题的出现会增加政府工作人员的工作量，工作人员只需要处理一次这个问题就可以了，所以针对这些重复数据，去除时间较久的留言，保留时间较近的留言。由于重复数据比较少，先在附件 2 中依据留言用户的留言编号对数据进行排序，取出发表多条留言的用户以及其留言时间、留言详情、留言位置（在 excel 表中的行数），以留言用户为“key”，留言时间、留言详情、留言位置合成 list 作为“value”，对比其留言内容的一致性，标记相同留言中时间较久的留言位置，删除该条留言，最后将去重后的数据写入新的 excel 表格中。去重代码见附件 quchong.py，去重后数据存入 quchong_data.xls 中。

（2）数据增强

经过统计我们发现，每个类别中的数据量差别较大，其中城乡建设数据量最多，有 1962 条，交通运输最少，只有 585 条，为了保持数据量的一致性，我们采用回译的方法对不足 1500 条的数据进行数据增强。“回译”是指将留言数据先翻译成外文再翻译成中文的操作，它利用不同语言间的差别形成数据的差异化，我们采用百度翻译 api 对数据进行了多次翻译，使其达到数据量翻倍的效果，最终每个类别的数据量大致在 1500~2000。进行数据增强处理的代码见附件 enlarge.py，增强后的数据文件为 zeng_data2_quan.xls。

（3）关键句抽取

由于留言详情中会出现冗余的句子，会影响分类方法的准确度，所以使用 TextRank 算法对留言详情进行关键句抽取。TextRank 算法由 PageRank 算法改进而来，其主要思想[1]是：首先确定文本的单词或句子为文本单元，然后把文章分成若干个已定义的文本单元，以这些文本单元作为节点，节点间是否具有相似性关系确定边的存在，相似度的数值作为权重，进而构成文本网络图，将文本量化，并以矩阵的形式表示，然后对生成的矩阵进行迭代收敛计算，依据节点的计算值大小对节点进行排序，取排名靠前的单词或句子，组成集合，形成关键词集或摘要句子集。

基于 TextRank 的关键句抽取通过选取文本中重要度较高的句子形成文摘，其主要步骤如下：

① 预处理：将输入的文本或文本集的内容按照逗号、句号分割得到句子集 $T=[S_1,S_2,\ldots,S_m]$，构建图 G =（V,E），其中 V 为句子集，对句子进行分词和去除停止词处理，得到关键词集 $S_i=[t_{i,1},t_{i,2},\ldots..,t_{i,m}]$，其中 $t_{i,j}\in S_j$ 是保留后的候选关键词。

② 句子相似度计算：构建图 G 中的边集 E，基于句子间的内容覆盖率，给定两个句子 S_i、S_j，采用如下公式进行计算：

$$\text{Similarity}(S_i,S_j)=\frac{|\{tk \vee tk\in Si \wedge tk\in Sj\}|}{log(|Si|)+log(|Sj|)} \qquad (2\text{-}1)$$

若两个句子之间的相似度大于给定的阈值，就认为这两个句子语义相关并将它们连接起来，即边的权值 $W_{ij}=\text{Similarity}(S_i,S_j)$。

③ 句子权重计算：根据公式，迭代传播权重计算各句子的得分。

④ 抽取文摘句：将③得到的句子得分进行倒序排序，抽取重要度最高的 T 个句子作为候选文摘句。

⑤ 形成文摘：根据字数或句子数要求，从候选文摘句中抽取句子组成文摘。关键句抽取的代码为附件中的 textrank_sentence.py，抽取结果写入 key_sentence.xlsx 中。

（4）中文分词及停用词去除

在对用户留言信息进行挖掘分析之前，需要把非结构化的文本信息转化为结构化、便于计算机识别的信息。中文词的特点是词与词之间没有明显界限，所以在提取关键词之前需要进行分词并去除停用词。

①中文分词---jieba 分词

jieba 分词是基于统计词典的一种分词，先构造一个前缀词典；然后利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置，构造一个有向无环图；通过动态规划算法，计算得到最大概率路径，也就得到了最终的切分形式[2]。

结巴分词支持三种分词模式：全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；精确模式，试图将句子最精确地切开，适合文本分析；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

本文采取精确模式，可以识别一般实体名称，如“北京大学”，符合本实验要求。

②去除停用词

对留言内容进行分词处理后，存在大量标点符号以及出现次数多但是不体现语句关键意义的虚词，去除这些词语有助于提高关键词提取的精度。目前常用的中文停用词表有：哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词等，我们对这些词表进行了整理与归纳，得到所用停用词表 stopwords.txt，经过匹配去除分词后的停用词，处理结果如下：

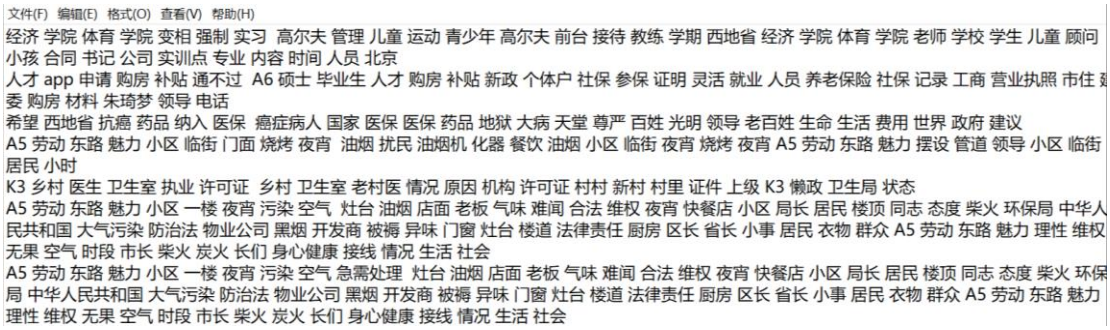

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
经济 学院 体育 学院 变相 强制 实习 高尔夫 管理 儿童 运动 青少年 高尔夫 前台 接待 教练 学期 西地省 经济 学院 体育 学院 老师 学校 学生 儿童 顾问 小孩 合同 书记 公司 实训点 专业 内容 时间 人员 北京
人才 app 申请 购房 补贴 通不过 A6 硕士 毕业生 人才 购房 补贴 新政 个体户 社保 参保 证明 灵活 就业 人员 养老保险 社保 记录 工商 营业执照 市住 建委 购房 材料 朱琦梦 领导 电话
希望 西地省 抗癌 药品 纳入 医保 癌症病人 国家 医保 医保 药品 地狱 大病 天堂 尊严 百姓 光明 领导 老百姓 生命 生活 费用 世界 政府 建议
A5 劳动 东路 魅力 小区 临街 门面 烧烤 夜宵 油烟 扰民 油烟机 化器 餐饮 油烟 小区 临街 夜宵 烧烤 夜宵 A5 劳动 东路 魅力 摆设 管道 领导 小区 临街 居民 小时
K3 乡村 医生 卫生室 执业 许可证 乡村 卫生室 老村医 情况 原因 机构 许可证 村村 新村 村里 证件 上级 K3 懒政 卫生局 状态
A5 劳动 东路 魅力 小区 一楼 夜宵 污染 空气 灶台 油烟 店面 老板 气味 难闻 合法 维权 夜宵 快餐店 小区 局长 居民 楼顶 同志 态度 柴火 环保局 中华人民共和国 大气污染 防治法 物业公司 黑烟 开发商 被褥 异味 门窗 灶台 楼道 法律责任 厨房 区长 省长 小事 居民 衣物 群众 A5 劳动 东路 魅力 理性 维权 无果 空气 时段 市长 柴火 炭火 长们 身心健康 接线 情况 生活 社会
A5 劳动 东路 魅力 小区 一楼 夜宵 污染 空气 急需处理 灶台 油烟 店面 老板 气味 难闻 合法 维权 夜宵 快餐店 小区 局长 居民 楼顶 同志 态度 柴火 环保局 中华人民共和国 大气污染 防治法 物业公司 黑烟 开发商 被褥 异味 门窗 灶台 楼道 法律责任 厨房 区长 省长 小事 居民 衣物 群众 A5 劳动 东路 魅力 理性 维权 无果 空气 时段 市长 柴火 炭火 长们 身心健康 接线 情况 生活 社会

图 2 数据预处理后的数据

2.1.3 基于传统机器学习的文本分类模型

（1）文本向量化

由于留言信息为文本类型，需要通过特征提取，对文本进行向量化，才能够使用机器学习方法进行后续的分析。在完成数据预处理后，我们使用 Term Frequency-inverse Document Frequency (TF-IDF)方法提取留言信息的特征。TF-IDF[3]方法为最常用的文本特征向量化的方法，用于评估词对一个文件集或一个语料库中的一份文件的重要程度，词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降，其计算公式如下：

$$W_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \qquad (2\text{-}2)$$

其中，$W_{i,j}$ 为文档 j 中第 i 项的权重，N 为集合中的文档数量，$tf_{i,j}$ 为文档 j 中第 i 项的词频，df_i 为集合中第 i 项的词频。

除此之外，为了消除数据量纲的影响并提高模型收敛速度，需要在提取特征后进行归一化处理。由于概率模型不关心变量的值，而是关心变量的分布和变量之间的条件概率，所以不需要进行归一化处理，而像支持向量机、线性回归、K 最近邻等算法，我们需要使用欧几里得 L2 范数进行了归一化[4]：

$$w_{norm} = \frac{w}{\|w\|} = \frac{w}{\sqrt{w_1^2+w_2^2+\cdots w_n^2}} \quad (2\text{-}3)$$

（2）卡方检验

在给出的留言信息中，经去重处理后训练数据条数为 9206，文本向量化处理后，特征词维度达 25 万之多，其中包含了很多对分类器训练贡献度不高的词，大大影响分类器性能和数据处理效率，为了选取优质特征，删除无关和冗余特征，实现降低数据集特征维度和防止过拟合，我们需要进行特征选择。本文选取的方法为卡方检验（CHI）[5]，卡方检验利用统计学中的假设检验思想，以卡方统计量来衡量特征词和类别之间的相关程度，词语 t 和类别 c 之间的卡方统计模型如下[6]：

$$\chi^2(t, c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (2\text{-}4)$$

在本文中，A 表示包含特征词 t 且属于类别 c 的留言数，B 表示包含特征项 t 但不属于类别 c 的留言数，C 表示属于类别 c 但不包含特征项 t 的留言数，D 表示既不属于类别 C 也不包含特征项 t 的留言数，N 表示语料中的留言总数。$\chi^2(t, c)$的值越大，特征词 t 与类别 c 的相关性就越强，此时特征词 t 所包含的与类别 c 相关的鉴别信息就越多。表 1 展示了经过以上步骤后得到的每个留言类别中最相关的词语 Top3。

表 1 各类别中最相关的词语

一级标签	最相关的单词	最相关的双词
交通运输	．的士 ．快递 ．出租车	．的士 司机 ．出租车 司机 ．出租车 管理
劳动和社会保障	．退休 ．职工 ．社保	．退休 工资 ．劳动 关系 ．退休 人员
卫生计生	．独生子女 ．医生 ．医院	．独生子女 父母 ．再婚 家庭 ．社会 抚养费
商贸旅游	．垄断 ．电梯 ．传销	．电梯 故障 ．小区 电梯 ．传销 组织
城乡建设	．住房 ．小区 ．公积金	．棚户区 改造 ．公积金 贷款 ．住房 公积金
教育文体	．学生 ．教师 ．学校	．教师 资格证 ．代课 教师 ．培训 机构
环境保护	．环保局 ．排放 ．污染	．环评 报告 ．噪音 扰民 ．噪音 污染

（3）模型选择

考虑到 TF-IDF 特征具有高维稀疏性，本文选择 LinearSVC 作为基准模型，再选取逻辑回归、随机森林和多项式贝叶斯分类算法进行效果对比，下面主要说明 SVM 的原理与实现过程。

原始的 SVM(Support Vector Machine, SVM)是一种二分类模型，目的是寻找一个超平面来对样本进行分割，分割的原则是间隔最大化，最终转化为一个凸二次规划问题来求解。这种分隔面模式可以有效克服样本分布的冗余和过拟合等因素的影响，具有良好的泛化能力。对于非线性的样本是通过核函数将其映射到高维空间，在高维空间将非线性问题转化为线性可分的问题，在这个空间构造一个最优超平面（如图 3），将两类样本无错误地分开, 且要使两个类别(标记为 y∈{–1,1})的分类空隙大，从而达到分类的目的[7]。常见核函数的表达式：

多项式核函数： $k(x_i, x_j) = (x_i^T x_j)^d$, d≥1

高斯核函数：$k(x_i, x_j) = \exp(-\frac{||xi-xj||^2}{2\alpha^2})$， α>0

sigmoid 函数：$k(x_i, x_j) = \tanh(\eta<x_i, x_j>+\theta)$

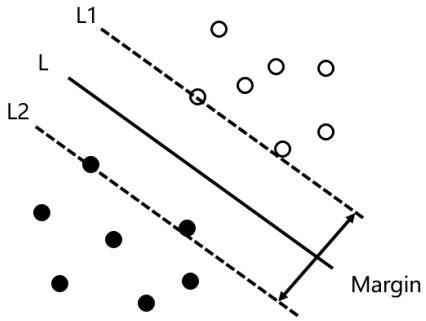


图 3 支持向量机的最优超平面

很明显，我们的任务是构建一个多分类模型，而过往已经有很多研究将 SVM 推广到多分类问题应用中，并有了很好的结果，比如：一对多多分类、一对一多分类、有向无环图多分类、二叉树多分类[8]。本文使用的 LinearSVC 是基于“one-vs-the-rest”实现多分类，用若干个二分类分类器的组合解决一个多分类问题，每次对一个目标类别和剩余其他类别的集合进行二分类，通过循环，完成多分类。这一方法虽然时间复杂度高，运算时间长，但分类效果较好。

本文使用机器学习工具包 scikit-learn[9] 来完成分类器的构建，LinearSVC 是使用 liblinear 实现的算法，相比于 SVC 在选择惩罚参数和损失函数时有更大的自由，对数据量较大的模型有更好的表现，通过 fit 方法将训练数据及其类别标签传入建立模型，predict 方法在已经建好的模型里，传入测试数据并输出预测的类别。我们将导入的模型命名为 model，那么使用函数 model.fit(train_x,train_y) 训练模型，使用训练好的模型预测数据的函数为 clf.predict(text_x)，其中 train_x、train_y 为训练数据集，text_x 用于评价训练出来的模型好坏。另外我们使用“网格搜索”[10]进行模型参数优化，先设置好参数列表，通过网格搜索找出最佳参数，如果最佳参数落在网格边缘则继续调整参数列表，直至最佳参数落在网格中间，使用最佳参数能使该模型能对数据有更好的拟合，具体代码见 svm.py。

2.1.4 基于深度学习的文本分类模型

基于 PyTorch 实现多个模型对中文文本进行分类，分别为使用[2,3,4]kernel size 的卷积神经网络（Convolutional Neural Networks, CNN）模型和循环模块为 LSTM 的双向循环神经网络（Recurrent Neural Network, RNN）模型，最终进行分类结果的比较，具体代码见

text_classify 目录。

（1）文本向量化

为了将语料输入神经网络进行训练，我们首先要将自然语言符号表示成计算机能够理解的数字形式。一个自然的想法是把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。这就是独热编码形式(One-Hot)。独热编码虽然方便易懂，但也有显而易见的不足：首先，One-Hot 编码的维数由词典长度而定，过于稀疏，存在降维难问题，给计算造成了很大不便；其次，One-Hot 编码下任意两个词之间都是孤立的，丢失了语言中的词义关系。

Word Embedding 解决了这两个问题。Word Embedding 矩阵给每个单词分配一个固定长度的向量表示，这个长度可以自行设定，比如 300，实际上会远远小于字典长度（比如 10000）。此外，通过大量的预训练获得的 Word Embedding 可以将一些句法信息等也编码进来，体现出词与词之间的关联性。

（2）CNN 模型

相比 RNN 网络，卷积神经网络最大的优势是可以并行计算效率大幅提高。该训练模型由嵌入层、卷积层、池化层、全连接层所构成，不同层相关定义以及功能表达所下。

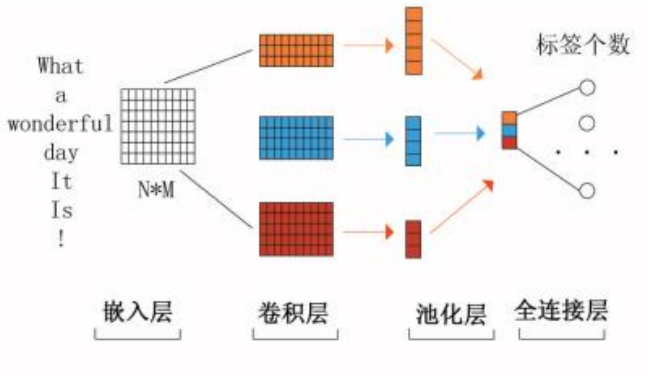


图 4　CNN 训练模型结构图[11]

①嵌入层

训练集中的每个单词可以由词向量进行表示，由所有单词向量组成嵌入矩阵$M \in R^{n*d}$其中 n 表示单词的个数，d 表示单词向量的维度。训练经过数据预处理后表示成索引数字，通过嵌入层映射将索引转化为词向量表示。假设单个样本经过数据预处理后由 s 个单词组成，那么对于每个单词，通过与嵌入矩阵 M 相乘，得到该单词对应的词向量表示。该样本向量矩阵便可由 s 行 d 列的矩阵$A \in R^{s*d}$表示。

②卷积层

卷积层用于提取样本向量矩阵特征。通过不同卷积核得到不同的特征集合，本文定义卷积核的宽度为词向量维度 d，卷积核高度 h 为可变参数。假设存在一个卷积核，由宽度为 d，高度为 h 的向量矩阵 w 所表示，对于一个样本，经过嵌入层之后可以得到向量矩阵$A \in R^{s*d}$，那么卷积操作可以用如下公式表示：

$$o_i = w \cdot A[i{:}\, i+h-1], i = 1,2\ldots, s-h+1 \qquad (2\text{-}5)$$

其中$A[i{:}\, i+h-1]$表示 A 的第 i 行到第 i+h-1 行，在通过激活函数 F 激活后,得到所需的特征。激活函数公式为：$c_i = f(o_i + b)$，其中 b 为偏置项。对于一个卷积核，可以得到特征$c \in R^{s-h+1}$，总共 s-h+1 个特征。通过使用更多高度不同的卷积核，可以得到更丰富的特征表示。

③池化层

池化层是降低卷积层的输出结果以及提取更深层次的特征表达，不同高度的卷积核得到

的特征集合大小不同，本文对每个特征集合使用池化函数，采用 1-maxpooling，提取出特征集合中的最大值。对于每一个卷积核输出特征为该特征集合的最大值，公式如下：

$$\hat{c} = \max\{c\} \tag{2-6}$$

对所有卷积核使用 1-maxpooling，将所有输出特征值级联，得到该文本最终的特征向量表示。

④全连接层

经过池化层级联后的特征向量与标签集$|L|$个神经元进行全连接，作为模型的输出层。同时该模型采用 sigmoid 函数作为模型的输出函数。公式为：

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{2-7}$$

通过设定不同阈值来判定该文本是否属于该标签。

⑤正则化

为了预防训练模型过程中出现过拟合的现象，本文采用 Srivastava 等人[12]提出的 dropout 方法，通过将一定比例隐藏层中的神经元权重设置不工作。从而降低模型的计算，同时也一定程度上避免了模型在训练过程中出现的过拟合现象。

本文使用 3 个 kernelsize 为[n，embeddingsize]（n 取 2、3、4）的卷积核进行卷积操作，卷积后接 max_pool，最后将 3 个卷积输出 concate 后经全连接层得到最后输出。

（3）RNN 模型

RNN 神经网络借助 hiddenstate 对信息的存储、传递特别适合处理文本序列数据，不过存在长城依赖及梯度爆炸问题。LSTM 通过引入门控机制对 RNN 进行改进，可有效缓解相关问题。双向 RNN 模型是 Schuster 在 1997 年提出的，目的是解决单向 RNN 无法处理后文信息的问题，单向的 RNN 只能在一个方向上处理数据，则双向循环神经网络的双向 LSTM 模型的基本思想是提出每一个训练序列向前和向后分别是两个循环神经网络（RNN），而且这两个都连接着一个输出层。下图展示的是一个沿着时间展开的双向循环神经网络:

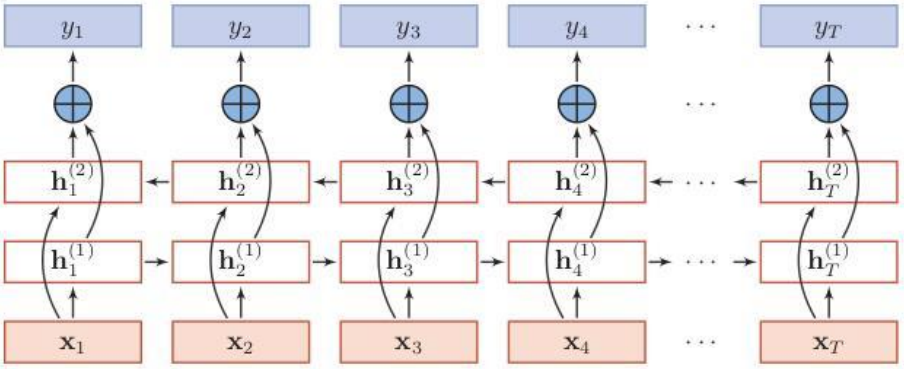


图 5 双向 RNN 结构图[13]

其中自前向后循环神经网络层的更新公式为：

$$\overrightarrow{h_t} = H(W_{x\overrightarrow{h_t}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h_{t-1}} + b_{\overrightarrow{h}}) \tag{2-8}$$

自后向前循环神经网络层的更新公式为：

$$\overleftarrow{h_t} = H(W_{x\overleftarrow{h_t}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h_{t+1}} + b_{\overleftarrow{h}}) \tag{2-9}$$

两层循环神经网络层叠加后输入隐藏层：

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h_t} + W_{\overleftarrow{h}y}\overleftarrow{h_t} + b_y) \tag{2-10}$$

双向 LSTM 神经网络（Bi-direction long short-term memory neural network）模型是结合

双向 RNN 和 LSTM 两个模型的优点形成的新模型，简单来说就是用 LSTM 单元替换掉经典双向 RNN 模型中的循环单元。

2.2 问题二

2.2.1 处理流程

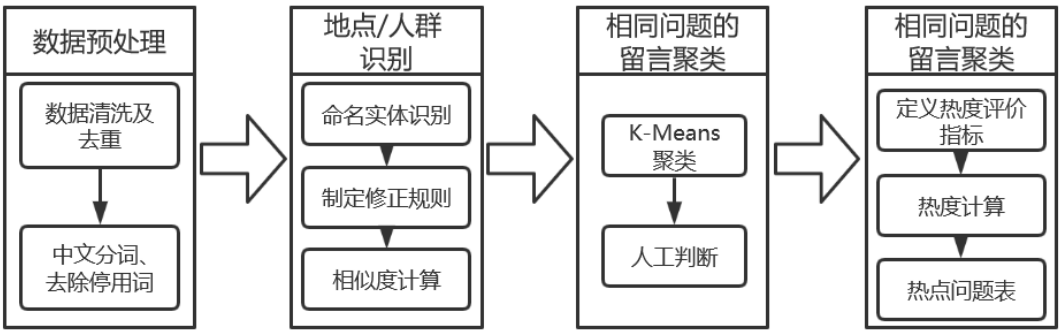


图 6 问题 2 处理流程

我们首先对数据进行清洗和预处理，通过命名实体识别的方法对每条留言主题中的地点和人群进行抽取，为了使结果更加准确我们加入了一系列修正规则，同时由于语言表述的差异，相同地点会有不同的表述形式，这些差异会对结果造成较大影响，所以进行了相似度计算，将真正意义上相同地点/人群的留言编号，然后利用 K-Means 对留言主题和留言详情进行聚类，结合上一步的结果对具有相同类别和相同地点/人群的留言进行人工判断其是否为同一事件。最后，结合相应结果和热度评价指标，计算每一事件的热度，构建热点问题表。

2.2.2 特定地点/人群抽取

由于留言详情中可能出现多个地点或人群对结果造成干扰，所以我们采用命名实体识别(Named Entity Recognition, NER)来抽取每条“留言主题”中的特定地点/人群。现有的大量研究证明，统计机器学习方法中的条件随机场(Conditional Random Field, CRF)方法一般是 NER 任务的常用选择，其优点在于其为一个位置进行标注的过程中可以利用到此前已经标注的相邻位置的信息，从而能够通过解码得到最佳序列。但是由于附件 3 中的数据包括许多虚构地名，如果直接用传统的 CRF 方法不会得到太好的效果，所以我们加入了一些人工规则帮助识别，实现代码见 ner.py。具体流程描述如下：

步骤一：首先选择北京大学收集并且已标注过的 1998 年 01 月的《人民日报》语料库作为标注语料，该部分语料主要用于提取常用地名和常用组织名，帮助构建相对应的知识库。将训练语料进行转换后，利用 CRF 模型对转换后的语料进行训练，由此得到最终的 CRF 模型参数。

步骤二：利用 jieba 分词对附件 3 留言主题进行分词和词性标注，转换数据模式，并利用上一步得到的 CRF 模型进行地名/人群命名实体的识别。

步骤三：通过挖掘未被识别到的和识别不完整的样本数据的内部特点和上下文特征，设计了大量的修正规则，在上一步识别的基础上，进行二次识别，对识别结果进行了修正，由此得到最终的地名/人群。

整体框架如图 8 所示：

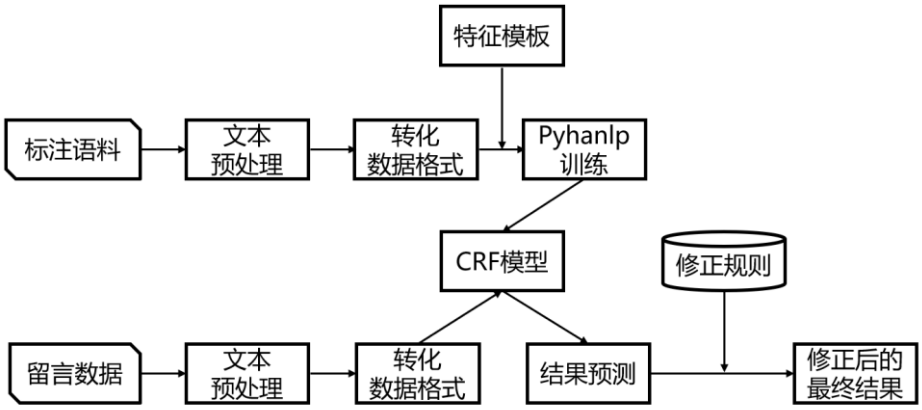


图 8 特定地点/人群抽取框架

由此，从留言主题中抽取出了唯一具有代表性的地点/人群实体，由于存在一些表述上的差异，不能直接将在语义上相同地点/人群实体的留言归为同一类，所以本文采用计算实体之间的相似度进行归类，具体计算方法在 2.3.3 中有所介绍，若相似度高于 0.9，则将对应的地名/人群视为相同，并将结果存入 location_extract.xlsx 对应的列中。

2.2.3 K-Means 聚类模型

由于 jieba 分词一般会把有特殊意义的词语分开，比如“魅力之城”会被拆分成“魅力”和“之城”两个词语，失去了它原本的意义，所以我们使用关键短语识别将留言内容和留言主题中的关键短语抽取出来，同时因为留言主题一般都简要概括了留言内容，所以提高了留言主题的权重。

在计算关键词的 TF-IDF 后，使用 K-Means 进行聚类，K-Means 的基本原理[14]如下:

K-Means 聚类需要事先指定聚类的类别中心 K，首先随机选取 K 个样本分别作为初始划分的簇类中心，然后根据相似性度量函数采用迭代的方法，计算未划分的样本数据到每个聚类中心点的距离，并将该样本数据划分到与之最近的那个聚类中心所在的簇类中，对分配完的每一个簇类，通过计算该簇类内所有数据平均值不断移动聚类中心，重新划分聚类，直到类内误差平方和最小且没有变化时为止。在每一次迭代过程中，模型都要判断每个样本数据是否正确划分到簇类中，若不正确，重新调整。当全部数据调整完后，再修改簇类中心，进行下一次迭代计算。如果某一次迭代过程中每个数据样本都分配到正确的族类中，则不再调整聚类中心。聚类中心稳定不再变化，标志目标函数收敛，算法结束，最后评价聚类结果。

聚类算法流程图如图 7 所示：

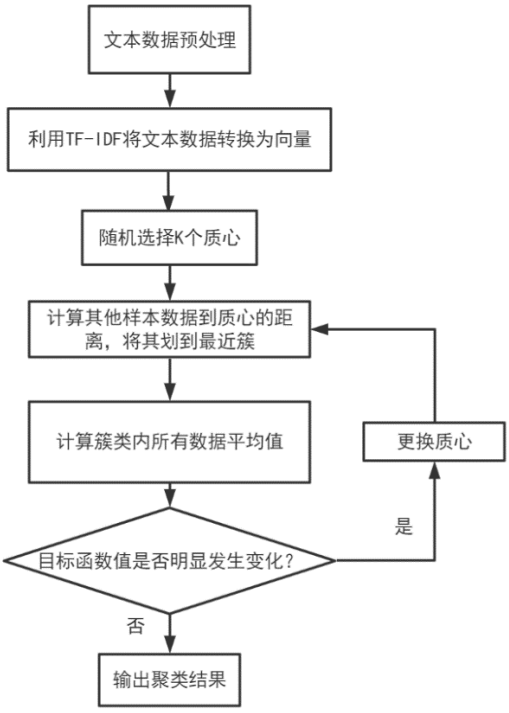


图 7　K-Means 聚类算法流程

在 K-Means 模型算法中，常见使用欧氏距离（也称欧几里得度量（Euclidean Metric））计算各点之间的距离，欧氏距离指在 m 维空间中两个点之间的真实距离，或者向量的自然长度（即该点到原点的距离），欧氏距离的计算公式如下：

$$d_{12} = \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \qquad (2\text{-}11)$$

考虑欧几里得距离的数据，使用误差平方和（Sum of the Squared Error,SSE）作为聚类的目标函数，两次运行 K 均值产生的两个不同的簇集，SSE 最小的那个说明聚类结果更加好。误差平方和的计算公式如下：

$$SSE = \sum_{i=1}^{K} \sum_{x \in Ci} dist(Ci, x)^2 \qquad (2\text{-}12)$$

其中 K 表示 K 个聚类中心，Ci 表示第 i 个中心，dist 表示的是欧几里得距离。SSE 的结果表示每个聚类簇中各个点到簇中心距离的和，经过不断迭代，找到距离之和最小的那个簇中心，就是最终的聚类中心点，实现代码见 kmeans_cluster.py。

2.2.4 热度指标

（1）热度评价指标设定

在查阅文献的基础上，结合网络问政平台上的留言信息，从留言事件本身和群众态度两个方面来构建了最初的热度评价指标体系，经专家评定保留下最重要的指标。最终构建的留言问题热度评价指标体系共分为 3 层：第一层为留言问题总热度 F；第二层包括留言事件的爆发度 F1、群众对留言的作用度 F2；第二层指标下，进一步包括若干指标 F11 到 F22。

表 2 留言问题热度评价指标体系

一级指标	二级指标	三级指标
热度指数 F	事件爆发度 F1	留言数量 F11
		事件持续时间 F12
	群众作用度 F2	群众参与度 F21
		群众情绪强度 F22

（2）末级指标值计算

末级指标量化对于指标体系建立的科学性、计算精确性、评价可靠性来说是至关重要的。鉴于上述四个指标涉及的都是客观数据，其量化的具体方式如下：

留言数量=某个问题的留言数量

事件持续时间=最后一条留言时间-第一条留言时间

群众参与度=点赞数+反对数

群众情绪强度=点赞数-反对数

为了更加便于计算，需要将上述末级指标做同趋化处理，这里采用指标无量纲化的方法对末级指标进行标准化。公式如下[15]：

$$A = \frac{a_{ij} - a_{imin}}{a_{imax} - a_{imin}} \tag{2-13}$$

上述公式中，a_{imax}、a_{imin}是各级指标评价值的最大值和最小值。计算后，得到[0-1]区间内的数值，为使数据更加直观、标准，将所有数值扩大 100 倍，得到[0-100]区间内的最终值，并进行比较分析，得出归一化的结果。

（3）利用层次分析法方法计算指标权重

层次分析法（Analytic Hierarchy Process，AHP）是美国运筹学家萨第（T.L.Saaty）提出的一种结合了定性和定量方法的系统分析方法[16]。本文根据这一方法确定留言问题热度评价体系下的各指标权重，具体过程如下。

①构造判断矩阵[17]。根据各个层次等指标之间的重要性程度之比，参照 Saaty 给出的 9 个重要性等级（见表 3），确定模型两两要素之间的比值，从而构造判断矩阵，见表 4。

表 3　1-9 标度的含义

标度	含义
1	表示两个元素相比，同样重要
3	表示两个元素相比，前者比后者稍重要
5	表示两个元素相比，前者比后者明显重要
7	表示两个元素相比，前者比后者强烈重要
9	表示两个元素相比，前者比后者极端重要
2,4,6,8	表示上述相邻判断的中间值
倒数	若元素 i 与元素 j 的重要性之比为 aij，那么元素 j 和元素 i 重要性之比为 aji=i/aij

表 4 判断矩阵

目标或准则	A1	A2	A3
A1	A11=A1/A1	A12=A1/A2	A13=A1/A3
A2	A21=A2/A1	A22=A2/A2	A23=A2/A3
A3	A31=A3/A1	A23=A2/A3	A33=A3/A3

②确定指标权重。根据表 4 中所列的判断矩阵，利用和积法计算出目标层和准则层的各个判断矩阵的最大特征值及其特征向量，具体计算方法如下：

a. 根据判断矩阵进行每列向量度求和，以及归一化处理

$$N_{ij} = \frac{A_{ij}}{\sum_{j=1}^{n} A_{ij}} \text{ (i, j=1,2,3…n)} \tag{2-14}$$

b. 对矩阵的每一行的向量进行求和

$$T_{ij} = \sum_{i=1}^{n} T_{ij} \text{ (i, j = 1,2,3 ... n)} \tag{2-15}$$

c. 对矩阵T_i进行归一化处理得到指标的层次单权重系数

$$W_{ij} = \frac{T_i}{\sum_{I=1}^{n} T_i} \text{ (i, j=1,2,3…n)} \tag{2-16}$$

③一致性检验。由于判断矩阵的指标的选取是根据调查研究和专家经验判断得到的，带有一定主观性；所以，为了避免这种主观性产生的错误，需要进行判断矩阵的一致性检验，并引入平均随机一致性指标 RI。具体检验方法为：用随机一致性比率 CR=CI/RI，当 CR＜0.1 时，判断矩阵具有满意的一致性，即权重计算正确。其中“CI”为判断矩阵的一致性指标。具体的计算过程如下：

a. 计算最大特征根λmax

$$\lambda max = \sum_{i=1}^{n} \frac{(AW)_i}{nW_j} \tag{2-17}$$

其中，$(AW)_i$表示向量 AW 的第 i 个元素。

b. 一次性指标 CI 的计算

$$CI = \lambda max - n/n - 1 \tag{2-18}$$

c. 一致性判断

根据 Saaty 提供的 RI 的对应值（见表 5），来查找对应的一次性指标。然后，根 CR=CI/RI 计算出一致性比率，进行一致性判断。当 CR＜0.1 时，则对应的判断矩阵通过一致性检验，反之就需要对判断矩阵进行适当的调整。

表 5 RI 值(Saaty)

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45

（4）留言问题热度计算

本文根据各级指标所确定的权重和客观数据值，计算出所有末级指标值，然后对各级指标进行加权求和。

$$S = \sum_{i=1}^{n} F_i W_i \tag{2-19}$$

公式(2-19)中的F_i为第 i 个指标的得分；W_i为第 i 个指标的权重；n 为指标的个数。S 为该留言问题的热度得分。

2.3 问题三

这一问题涉及到的重要理论是政府回应，它是指政府对公众所表达的诉求进行反应。新公共行政理论的代表人物弗雷德里克森[18]认为，“公共行政的核心价值，或者公共行政的精神，不仅包括了在一般意义上对公共的承诺，也包括了在具体意义上对具体的公民和公民团体的回应”。而政府网络回应则是互联网时代背景下政府回应公众诉求的一种典型表现形式，以网上“智慧政务”平台为代表，它是人大代表履职的练兵场，也是监督“一府一委两院”工作的重要窗口，只有知晓了老百姓最真实的生活状态以及社会中存在的大小问题，才能够更好地完善相关政策。之所以需要对政府工作人员对市民的留言回复进行评价和监督，一方面是因为留言所涉及的事件关系民生，另一方面是可以发现政府部门工作的短板，同时对这些诉求进行督办来提高群众幸福感。基于此，我们设定了一系列指标来评价留言答复意见，本文下面将从剔除表中无效信息，然后构建评价指标的量化方法，最后进行指标权重的确定进行展开介绍。

2.3.1 无效留言的处理

在市民留言和回复文本的数据库中，有些问题描述不清或过于简短，导致工作人员无法给出合理的解答，这类问题通常称为留言答复数据中的坏数据。在数据预处理中，我们需要将坏数据剔出评价系统，才能使基于市民问题和回复文本对政府人员们的工作评判具有公平性。在许多情况下，无效问题的文本长度较短，且与一般问题的文本长度有明显差距。以下对每个答复质量的评价都是基于预处理后的留言回复信息数据库。另外，本文引入权重系数，

对每一个答复的计分进行加权求和，得到留言 K 回复质量的评价得分，记为 F（K）。

2.3.2 构建评价指标体系

指标之间既要避免重复交叉，又要有内在逻辑关联，同时答复意见的质量由很多种因素决定，必须从不同角度对其进行描述，我们查阅了有关“问答平台答复意见质量的影响因素”的相关文献，Zhu Z，Bernhard D 等[19]采用专家分析、用户调研以及比较法，提炼并总结了判定问答平台答复意见质量的十三个维度，并建立了问答平台多维质量评估模型。Kim 和 Oh[20]根据提问者的评论总结了判断最佳答案的评价标准，并提炼出一个分析框架，其中包括内容价值、认知价值、社会情感价值、信息源价值、外在价值等，并分析了各个标准的影响因素。由此，结合现有文献研究和本问题的实际情况，本文构建了三项二级指标和六项三级指标帮助对留言回复的好坏进行量化评估，如下表所示：

表 6 留言热度评估指标体系

一级指标	二级指标	三级指标	说明	来源
答复意见质量评估指标(F)	内容维度(F1)	完整性(F11)	答案的结构是否具有完整。	蒋楠,王鹏程.社会化问答服务中用户需求与信息内容的相关性评价研究——以“百度知道”为例[J].信息资源管理学报,2012,2(03):35-45.
		相关性(F12)	答案是否符合所回复问题的主题。	马芳．信患检索中的相关性研究[J]科技情报开发与经济，2009(19)：89．90
	认知维度(F2)	及时性(F21)	答案是否回复得比较及时。	杨子武.SNS 网站质量评价基本指标集分析[J].科技信息,2011(14):390.
		可证实性(F22)	答案中是否包含相关法律条文及参考内容，供验证或进一步拓展。	杨开平, 李明奇, 覃思义. 基于网络回复的律师评价方法[J]. Computer Science, 2018, 45(9):237-242.
	情感维度(F3)	充实性(F31)	答案体现出的回答者的努力认真程度和回答者的诚恳度。	孙晓宁，赵宇翔，朱庆华. 基于SQA 系统的社会化搜索答案质量评价指标构建［J］. 中国图书馆学报，2015，41(4) : 65－82.
		礼貌性(F32)	使用礼貌用语，以及对他人感情和观点的尊重程度。	Zhu Z , Bernhard D , Gurevych I . A Multi-Dimensional Model for Assessing the Quality of Answers in Social Q&A Sites[C] // International Con-ference on Informa-tion Quality. DBLP, 2009.

2.3.3 评价指标量化方法

（1）完整性

一条完整的回复应该具有四个部分：称呼、对情况的了解、回复内容和回复日期。本文通过人工归纳规则来判断回复是否具有上述四个部分，以此衡量留言回复的完整性，具体实施过程如下：

① 称呼：对回复进行分词，若前 10 个分词中含有中文冒号“：”，则该回复具有称呼。

② 对情况的了解：人为归纳词表，若该回复的分词中含有该词表中的词，则该回复具有这一部分。

③ 回复内容：同②。

④ 回复日期：若留言回复的结尾含有如“2020 年 1 月 24 日”或“2020.01.24”这两类内容，则该留言回复具有回复日期。

根据数据具体情况，通过 Integrity 目录下的代码文件得到每条回复意见中包含完整性结构的情况，得到完整性 F11 的指标得分。

（2）相关性

工作人员给出的留言答复必须针对市民留言提出的问题，也就是在答复意见和留言主题间存在一定的相关性。为此，我们建立文本相似度评价函数 TEXTSIM，并将 TEXTSIM（A_i，A_j）作为答复意见 A_i 与留言主题 A_j 的一个相似度指标。TEXTSIM（A_i，A_j）一般都是基于词语间相似度 WORDSIM（W_i，W_j）的，本文使用改进的词袋模型[21]（bag of words, BOM）来计算评价函数 TEXTSIM，该模型泛化性强，效率高，比较轻量级，结果范围在 0~1，该值越大，则表示 A_i 和 A_j 越相似。

对于每条答复意见，我们先进行分句，每个分句需要去除所有的标点符号、分词、去停用词，由此计算得到的第 n 个分句 A_{in} 与留言主题 A_j 之间的相似度，由于大部分分句和留言主题的相似度会比较低，所以我们只考虑每条答复意见中的分句与留言主题中相似度最高的句子的相似度，即:

$$F12 = \max\{TEXTSIM（A_i，A_j）\} \tag{2-20}$$

（3）及时性

对于每一条留言回复，设留言时间为 T，回复时间为 t。我们用留言时间与回复时间的差值来衡量留言回复的及时性，即:

$$F21=T-t \tag{2-21}$$

（4）可证实性

经过对数据的观察，我们把包含参考内容的回复意见分成两大类，如下表所示

表 7 部分数据特征示例

类别	特征	实例
显示地引用法律条文	带有“《》”	2018 年 6 月 7 日，A 市 A3 区人民政府办公室下发了《关于 A 市 A3 区既有多层住宅增设电梯实施方案》的通知。该方案明确了增设电梯条件及流程。详情可咨询政务服务中心住建局受理申请老旧小区加装电梯窗口，咨询电话：0000-00000000。
隐示地引用相关政策、文件	带有“根据、依据、规定”之类的词语	网友：您好！2014 年 10 月份开始，退休工资改称为“退休人员基本养老保险”。文件规定：2014 年到 2024 年 10 年间退休人员基本养老保险按 2014 年 9 月工资计算，满 35 年工龄按在职工资 90%计算，每少一年减 1%。

根据数据具体情况，我们采用正则表达式匹配相关特征值，得到每条回复意见中引用的次数，具体实现代码见 verifibility.py。

（5）充实性

回复内容的详细程度与回复的文本长度有直接的关系[22]。简短内容的回复信息量一般不够，评分应该较低；同时，较长文本的回复评分不应该过高。因此，考虑使用对数函数来量化回复的文本长度与评分的关系，建立“回复内容是否充实” F31 的评价项。

$$F31 = log_m L_i - \frac{1}{N_t}(log_m T_i) \tag{2-22}$$

其中，N_t指超过一定字数的留言量，L_i为针对第 i 个问题回复的文本长度，T_i为第 i 个问题回复超过规定字数的部分，m 为常数。经过讨论，确定 1000 字左右的回复效果最佳，为了将指标值限制在 10 以内，确定 m=2，实现代码见 count.py。

（6）礼貌性

我们根据答复意见中使用的文明词语，构建了常用礼貌词表，见 polite_words.txt，由于这些词语通常分布于开头和结尾，并不与文本长度成正相关，所以我们直接将每条答复意见中礼貌词的频数作为 F32 的值。

最后，我们采用了 2.2.4 中的归一化方法对各项指标进行了处理，代码见 normalization.py。

2.3.4 利用 AHP 计算指标权重

层次分析法是一种简便、灵活而又实用的多准则决策方法。根据已经建立的指标体系，构造两两比较判断矩阵$B = (b_{ij})_{n\times n}$，得到各指标的相对权重，b_{ij}表示指标 i 与指标 j 比较的两两相对重要程度。层次分析法可以充分考虑指标之间的相对重要性，提高权重设计的科学性。对计算指标权重（特征向量）W进行一致性检验。最后计算组合权重，得出一级指标相对于上级总体指标的权重分配值。具体实施过程在 2.2.4 小节中已详细介绍。

3 实验结果分析

3.1 实验环境

表 8 实验环境

实验环境	详情
软件环境	Window10(64 位)操作系统
硬件环境	Intel(R) Core(TM) i7-8550U
	CPU @ 1.80GHz 2.00 GHz
编程语言	Python 3.6
机器学习框架	Scikit-Learn 0.22
深度学习框架	Pytorch1.1.0

3.2 实验评估指标

根据题目要求和模型评估需要，我们总共用到了查准率（Precision）、查全率（Recall）、两者的调和平均数 F1 值对模型进行评估，其中 F1 值被用来作为问题一中留言分类模型的评估指标。以上指标的详细定义如下:

$$\text{查准率 P} = \frac{\text{分类正确的留言数}}{\text{分类正确的留言数+分类错误的留言数}} \tag{3-1}$$

$$\text{查全率 R} = \frac{\text{分类正确的留言数}}{\text{分类正确的留言数+未被识别的留言数}} \tag{3-2}$$

$$\text{F1} = \frac{1}{n}\sum_{i=1}^{n}\frac{2*P_i*R_i}{P_i+R_i} \tag{3-3}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

3.3 实验一

3.3.1 原始数据分析

（1）类别分析

题中给出了 9211 条留言数据，其中含有 620 条重复数据，对数据去重后每一类别的数据分布情况如下：

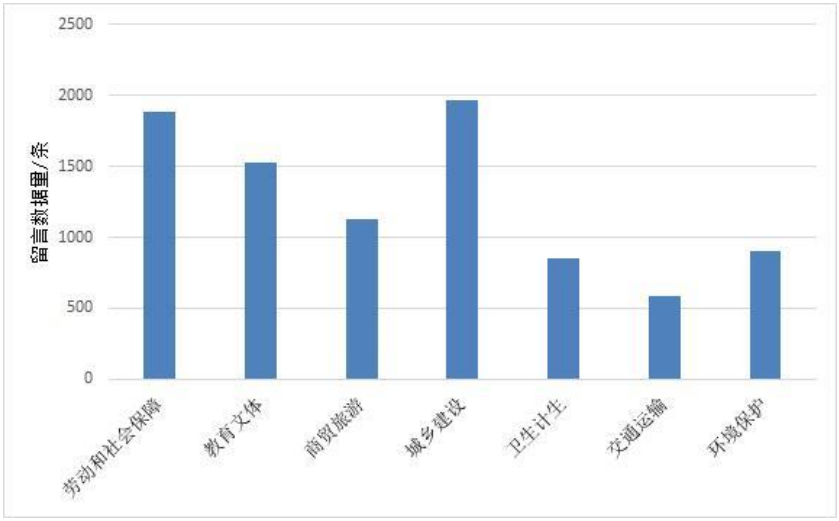


图 9 留言类别数据分布

可以看到“商贸旅游”，“卫生计生”，“交通运输”，“环境保护”这四类数据数量与其他类别相差较多，所以针对这几个类别进行数据增强，增强后的数据如下表：

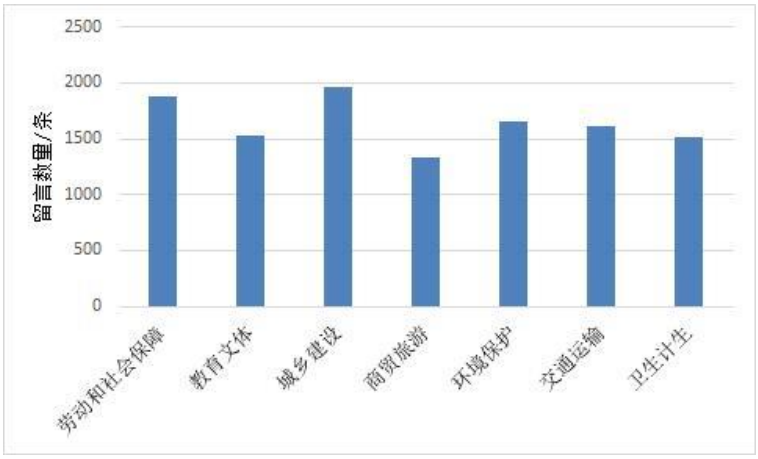


图 10 增强后的留言类别数据分布

（2）各类别关键词分析

对数据进行预处理后，计算每个词的 TF-IDF 值，结合类别标签，制成每个类别的关键词（TOP30）云图，如图 11 所示。

（1）“交通运输”关键词

（2）“劳动和社会保障”关键词

（3）“卫生计生”关键词

（4）“商贸旅游”关键词

（5）“城乡建设”关键词

（6）“教育文体”关键词

（7）“环境保护”关键词

图 11 各类别关键词的词云图

3.3.2 基于传统机器学习模型的分类结果

（1）不同模型下的分类效果

前文已有提到我们选取了线性支持向量机、逻辑回归、随机森林和多项式贝叶斯算法构建一级标签分类模型，将数据集按照 8：2 分为训练集和测试集，同时采用五折交叉验证对比每次结果的 F1 值对不同模型进行评估，并选用箱型图直观展示结果，实验结果如图 12。

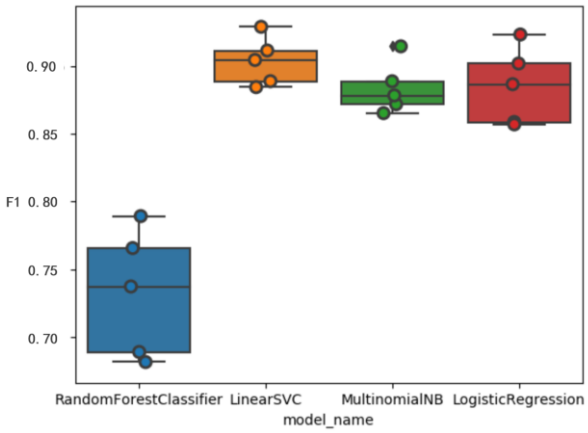


图 12 四种模型分类结果箱型图

从图中可以看出，与我们的预想一致，线性支持向量机模型的平均 F1 值是最高的，达到了 90.2%，且分布密集无离散点，而随机森林分类器的准确率是最低的，分布也较离散，因为随机森林属于集成分类器，一般来说集成分类器不适合处理像文本数据这样的高维数据，因为文本数据有太多的特征值，使得集成分类器难以应付，另外两个分类器的平均准确率都在 85%以上。所以接下来我们重点研究了在线性 SVM 上的分类效果。在实验过程中，经反复次调整参数，并记录每次实验结果，最终确定了最优参数组合，并应用于（2）中。

（2） 不同特征数下的分类效果

在对数据、模型和超参数进行了优化后，最后我们研究了在使用卡方检验降维时，选择不同的特征数对分类器的分类效果的影响，结果如图 13。

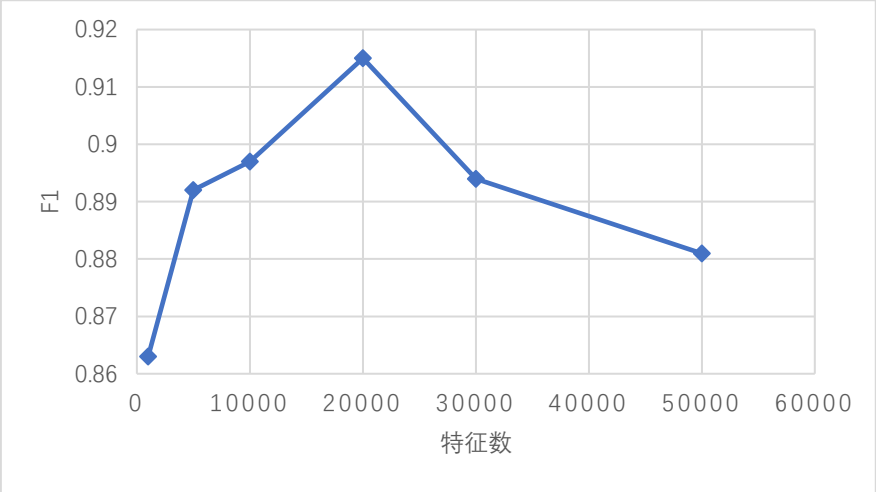


图 13 不同特征数下分类效果

从上图可以得出以下结论：

①分类器的分类效果普遍随着特征数的增加而先变优然后再变差。

②最好的分类效果出现在特征数为 20000，F1 值为 91.5%。

（3） 混淆矩阵分析

利用已训练好线性支持向量机模型对测试数据集的 x_test 进行分类预测得出分类结果 y_pred。然后用 x_test 及 y_test 测试数据集对该模型进行评估，我们利用 y_test 和 y_pred 画出混淆矩阵，X 轴为预测值的留言分类，Y 轴为真实值的留言分类，如图 14 所示：

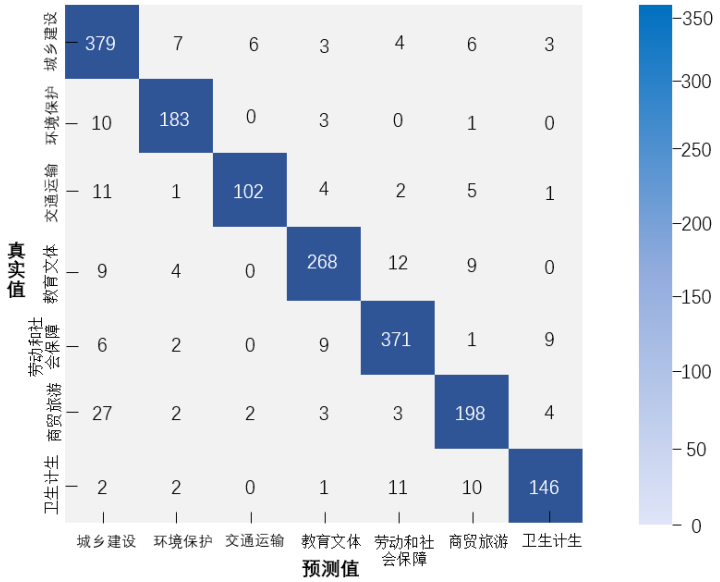


图 14 线性支持向量机模型混淆矩阵

从混淆矩阵的结果可以看出，城乡建设类的留言均有被分为其他六类的留言，而环境保护、交通运输、教育文体和商贸旅游类更容易被判定为城乡建设类的留言，卫生计生类中被判定为劳动和社会保障类的留言居多，为了找出判别错误的原因，以劳动和社会保障预测为卫生计生为例，我们列出了部分判断错误的留言如下：

劳动和社会保障 预测为 卫生计生 ： 9 例.

	一级标签	留言
154	劳动和社会保障	E7县麻塘山乡尖山村尿毒症患者请求低保，巨额的医疗费用像大山一样压的我喘不过气来，后来在E7...
5532	劳动和社会保障	听听楚雅二医院做事的非正式工作人员的心声，如今中南大学楚雅二医院又在争评全国优质三甲医院，大...
6780	劳动和社会保障	A市什么时候能实行独生子女护理假，很多省份都已经试水独生子女护理假了，宁夏、广西、福建他们都...
6652	劳动和社会保障	F市医保卡何时能落实异地联网，去年去A市看病，说是能联网了，结果还是要跑到地区去签各种文件盖...
6259	劳动和社会保障	J11市白内障治疗新农合能报销多少，2月15号有J市眼科专家下乡去村里义疹，由于我父亲视力模...
6108	劳动和社会保障	请问西地省现在是否有城市户口和农村户口之分，我是E市非转农我老婆是A市A2区农村户口我现在是...
6295	劳动和社会保障	J市东华医院挂床住院，虚增医疗费用套取国家医保基金，2017.2.25在J市东华医院医生的免...
5506	劳动和社会保障	感谢毛市长让我看到了希望，还望继续关注其中的一个错假冤案，2010年元月9日上午9时30分许...
6120	劳动和社会保障	独生子女补贴80元是否已经发放到退休工资里，本人于2005年退休，今年已经63岁了，当年响应...

图 15 部分判断错误的留言

从图中了解到两个类别存在交叉的部分，比如留言中提到“医疗费用”、“医院”、“看病”、“医保基金”等词，导致被判定为卫生计生，两个类别的交叉界限还不够明确，需要更大量的数据来训练模型以达到更好的效果。

3.3.3 基于深度学习模型的分类结果

分别在 CNN 模型和 RNN 模型中训练 10 个 epoch，每个 epoch 后在验证数据集上校验，保存 F1 最好的那个模型，最后测试集上的结果如下：

表 9 各分类模型 F1 指标值

	CNN	**RNN**
F1	0.9160	0.9099

可以发现 CNN 模型和 RNN 模型的 F1 值都接近于 0.91，其中 CNN 分类模型的 F1 值略高于 RNN。

进一步对 CNN 和 RNN 每个 epoch 时的 F1 指标值进行分析，可以发现在前 5 个 epoch 时 CNN 模型的分类效果都显著由于 RNN。随着训练数据量的增加，CNN 模型和 RNN 模型的分类效果越来越接近，但 CNN 模型的分类效果始终更优。

表 10 各分类模型每个 epoch 时的 F1 指标值

	CNN_F1	**RNN_F1**
epoch: 1	0.7842	0.6504
epoch: 2	0.8423	0.8013
epoch: 3	0.8958	0.8505
epoch: 4	0.8945	0.8749
epoch: 5	0.9067	0.8711
epoch: 6	0.9065	0.9081
epoch: 7	0.9072	0.8928
epoch: 8	0.9115	0.9011
epoch: 9	0.9147	0.9002
epoch: 10	0.9160	0.9099

CNN 和 RNN 分类模型的梯度优化算法都采用的 Adam，从图 14 可以发现，各分类模

型训练时的 loss 值逐渐递减，CNN 模型的 loss 值相对较高。

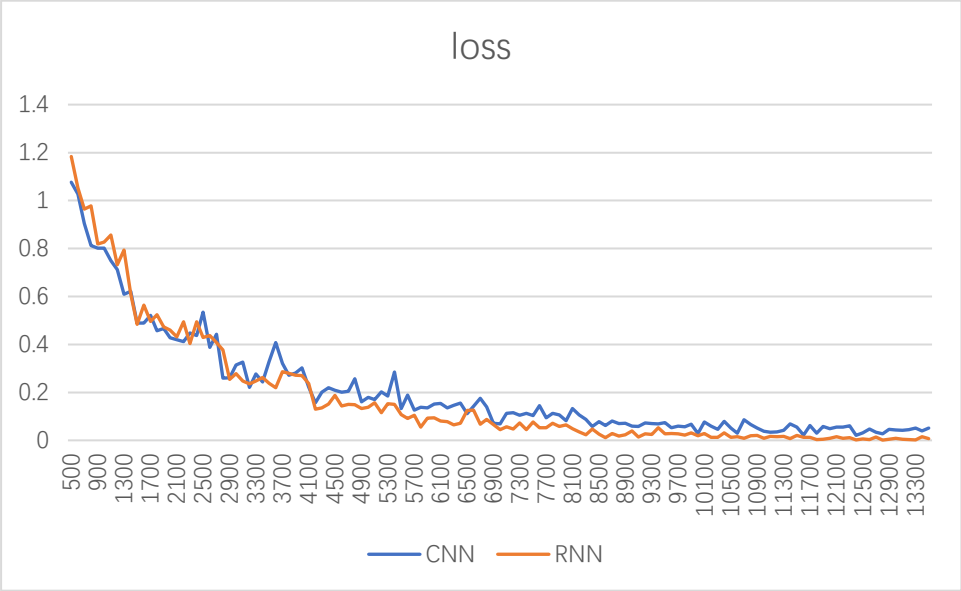


图 16 各分类模型训练时的 loss 值

3.3.4 创新点

本文分别构建了基于传统机器学习和基于深度学习的一级标签分类模型，尽管从结果看来，SVM 和 CNN 两种方法的 F1 值只有轻微差异，前者甚至不需要特别高端的运行硬件，在计算上更加便捷，但我们仍然保留了基于深度学习的分类模型。因为在未来，随着网络问政平台的普及和民众参与度的提高，将来需要处理的一定是海量数据，这时深度学习的优势就会显现出来，随着数据量的增加，传统机器学习模型性能没有很大提升，然而后者会相比前者有更好的分类效果。

3.4 实验二

3.4.1 特定地点/人群识别结果

以准确率和召回率为标准，我们训练出了最优 CRF 模型并结合人工规则对预测结果进行修正，最终得到了下图所示的数据。

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	地点/人群
210425	A00097039	A1区才子佳	2019/4/25	我是A市A1	0	0	A1区才子佳郡
262348	A00097039	子佳郡小区3	3/6/26 23:1	以往恶性循环	0	0	A1区才子佳郡
281101	A00011221	A1区朝阳街	2019/3/20	您好！我们	0	0	A1区朝阳街道解放东路
289808	A00011221	A1区朝阳街	2019/3/20	尊敬的A市	0	1	A1区朝阳街道解放东路
191062	A00027934	请公开A1区	2019/5/13	尊敬的冯意	0	0	A1区东成大厦
203369	A00027934	请求公开A	2019/3/26	尊敬的夏文	0	0	A1区东成大厦
233529	A00027934	A1区东成大	2019/7/15	行政复议请	0	0	A1区东成大厦
265260	A0001501	A1区恒大江	2019/11/3	您好！向您	0	0	A1区恒大
234943	A00028238	A1区恒大君	2019/4/17	A市A1区马	0	0	A1区恒大

图 19 预测结果修正

但由于存在错别字、语序不一致等问题，不能直接进行地点的归类，所以我们使用“地名相似度.py”计算了两两地点/人群实体之间的相似度，当相似度在 0.9 以上则给与它们同一个类别号，部分结果如下：

286897	A0006857	迟迟不给员	2017/7/19 10:0	欠费月份高	0	0 A市巴辛国际早教中心 217
221684	A0002431	8早教中心投	2017/7/17 19:0	动局黑名单	0	0 A市巴辛国际早教中心 217
200265	A0002936	7际早教中心	2017/7/11 15:0	四节早教	0	0 A市辛巴国际早教中心 217

图 20 部分预测结果修正

从图中可以看出，市民在提到“A 市巴辛国际早教中心”时，有误写为“A 市辛巴国际早教中心”的情况，但通过我们的处理它们都被标为了同一地名类别号。

3.4.2 聚类中心结果分析

(1)聚类簇选取

K-Means 聚类需要事先指定聚类簇 K，为了确定最好的聚类簇，本实验同样使用 SSE (sum of the squared errors，误差平方和)的值来确定最优解，这种方法也叫手肘法[23]。手肘法中 SSE 具体的计算方法和上面提到的相同，SSE 是所有样本的聚类误差，代表了聚类效果的好坏。

手肘法的核心思想是：随着聚类数 K 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和 SSE 自然会逐渐变小。并且，当 K 小于真实聚类数时，由于 K 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大，而当 K 到达真实聚类数时，再增加 K 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 K 值的继续增大而趋于平缓，也就是说 SSE 和 K 的关系图是一个手肘的形状，而这个肘部对应的 K 值就是数据的真实聚类数。K-Means 聚类的手肘图如下图所示：

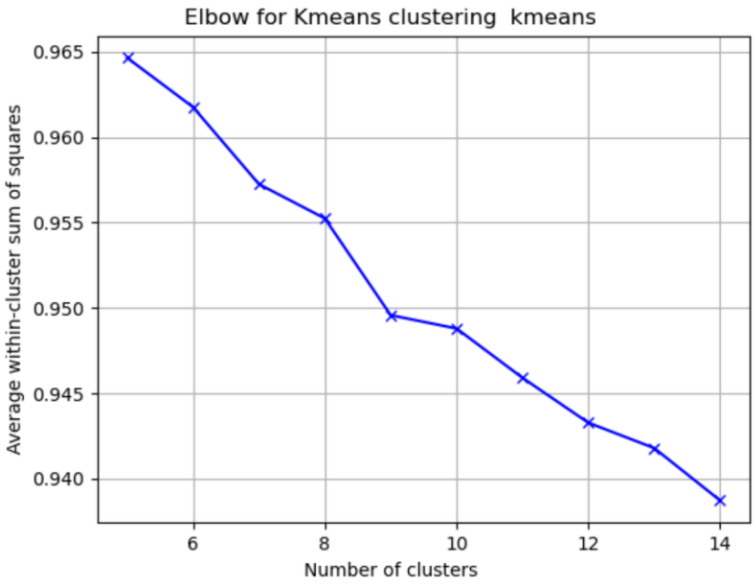


图 17 K-Means 聚类的手肘图

可以看到，在 K=9 时，折线发生了明显的弯折，在聚类数小于 9 时误差平方和较高，而聚类数从 9 变化到 10 时误差平方和变化幅度明显放缓，这说明 K=9 时误差平方和 SSE 下降幅度降低，聚类簇 K 最佳选择为 9。

(2)聚类结果分析

从聚类结果表 11 来看，留言数据可大致分为 9 个主题：①与交通安全相关②与公司、政府事务相关③与地方建设规划相关④与购房买房、开发商相关⑤与学生、学校相关⑥与物业、街道相关⑦与楼盘、合同相关⑧与地铁、高铁建设相关⑨与噪音、油烟扰民相关。

表 11 留言数据聚类主题

类别	类别关键词
聚类簇 1	路口,车辆,大道,公交车,道路,交通,路段,红绿灯,行人,车道,路灯,时间,交警,人行道,马路,线路,安全隐患,方向,周边,车流量
聚类簇 2	公司,领导,医院,部门,有限公司,人员,时间,单位,情况,工资,电话,派出所,项目,合同,平台,微信,员工,老百姓,企业,信息,
聚类簇 3	街道,村民,政府,领导,国家,周边,公园,建设,政策,背景,中心,老百姓,部门,规划,群众,房屋,居民,土地,力度,百姓
聚类簇 4	购房,车位,景园,滨河,商品房,资格,职工,合同,房子,项目,开发商,领导,人才,政策,政府,房产,工作,社保,有限公司,户口
聚类簇 5	学校,幼儿园,学生,中学,孩子,小学,家长,教育局,老师,学院,学费,领导,小孩,时间,费用,教师,周边,情况,部门,宿舍,
聚类簇 6	物业,业主,电梯,物业公司,街道,业委会,物业费,领导,居民,开发商,部门,房屋,车库,商铺,生活,住户,服务,政府,垃圾,国际
聚类簇 7	业主,开发商,房屋,楼盘,领导,质量,全体,房子,政府,交房,合同,电梯,房产证,车位,维权,部门,街道,户型,住宅
聚类簇 8	地铁,号线,出入口,线路,规划,时间,高铁,领导,居民,交通,周边,手机,中心,广场,建议,施工,部门,建设,公交车,扰民
聚类簇 9	居民,噪音,扰民,部门,生活,环境,油烟,新城,城管,声音,街道,周边,领导,工地,麻将馆,住户,夜宵,老人,渣土,通宵,门面,污染

从聚类簇包含的留言数量来看，第三类关于地方建设规划的留言比较多，超出了其他类别数量的两倍，是总留言量的大约 1/3，关于噪声、油烟投诉扰民类的较少。

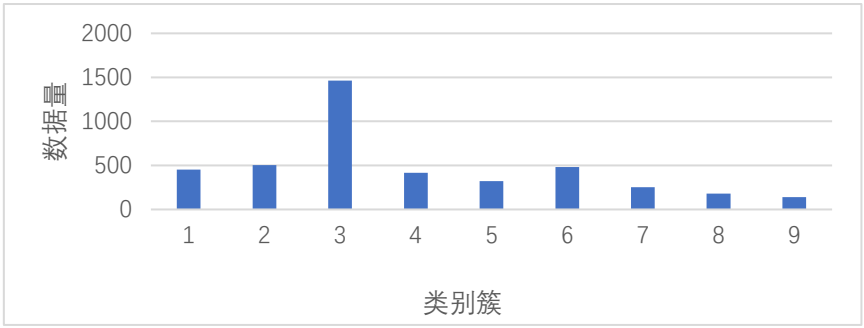


图 18 各聚类簇的留言数量

接下来我们对多次出现的地点/人群所对应的事件进行筛选，如果他们属于同一聚类类别，则人工判断是否为同一事件，若不属于同一聚类类别，说明他们属于同一事件的可能性极小，可不需人工判断，最后我们得到了每条留言的问题 ID。

3.4.3 热度计算结果

采用层次分析法对中表 2 的各个指标进行计算，得到其权重如下所示：

表 12 留言主题热度指标权重

权重 w	事件爆发度 F1 0.667	群众作用度 0.333	总权重
留言数量 F11	0.750		0.500
事件持续时间 F12	0.250		0.167
群众参与度 F21		0.333	0.111
群众情绪强度 F22		0.667	0.222
CR	通过一致性检验		

对各留言主题的热度指数进行计算，列出排名前 5 的热点问题，见表 13，相应热点问题对应的留言信息，保存到了“热点问题留言明细表.xls”。

表 13 热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	242	97.2	2020/1/7 至 2019/1/18	A 市地铁	A 市地铁线规划
2	226	66.3	2019/9/8 至 2019/1/18	A 市公交	A 市公交线路规划
3	247	48.3	2019/9/1 至 2019/4/8	A 市伊景园滨河苑	违规捆绑销售车位
4	33	47.4	2020/1/26 至 2019/11/2	A 市丽发新城	附近修建搅拌厂带来噪音和尘土污染
5	209	36.7	2019/9/27 至 2018/11/15	A 市人才	A 市人才购房补助发放问题

3.4.4 创新点

在识别特定地点/人群方面，本文采用 CRF 与规则相结合的实体识别方法，先训练 CRF 模型进行初次识别, 再利用人工制定的规则进行二次识别, 这样做的好处是：①很好地解决了无法识别留言主题中的未登录词问题；②避免完整地名被拆分成多个名词；③传统命名实体识别中不能直接输出地名和群体（如学生、教师、老人等）的组合，所以需要辅以规则帮助识别。

在对相同问题的留言归类方面，本文首先利用 K-Means 聚类将所有留言分为 9 个类别，如果不在一个类别里的留言则不考虑它们反映相同问题的可能性，只对同一类别且具有相同地点/人群的留言进行判断，这样大大地降低了人工筛选的负担。

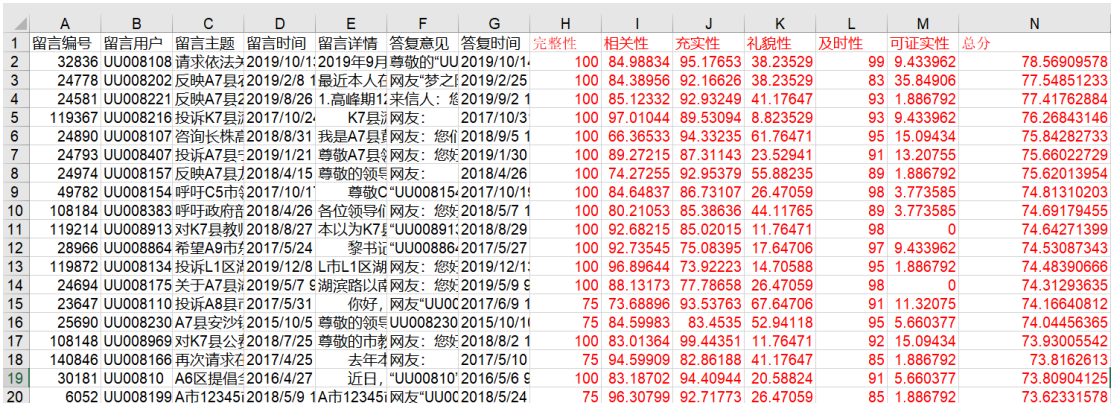
3.5 实验三

采用层次分析法对中表 6 的各个指标进行计算，得到其权重如下所示：

表 14 留言答复评价指标权重

权重 w	内容维度 F1 0.49	认知维度 F2 0.20	情感维度 F3 0.31	总权重
完整性 F11	0.33			0.163
相关性 F12	0.67			0.327
及时性 F21		0.67		0.133
可证实性 F22		0.33		0.067
充实性 F31			0.50	0.155
礼貌性 F32			0.50	0.155
CR		通过一致性检验		

依据前面所给的指标计算方法，得到每个指标计算值，为了消除量纲的影响，对各结果进行归一化处理，使平均值在 0~100，依据得分降序排列，得到每条答复意见的评分，结果存至 question3.xls 中，部分结果见图 21 所示。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	完整性	相关性	充实性	礼貌性	及时性	可证实性	总分
2	32836	UU008108	请求依法为	2019/10/1	2019年9月	尊敬的“UU	2019/10/1	100	84.98834	95.17653	38.23529	99	9.433962	78.56909578
3	24778	UU008202	反映A7县(	2019/2/8 1	最近本人在	网友“梦之[	2019/2/25	100	84.38956	92.16626	38.23529	83	35.84906	77.54851233
4	24581	UU008221	反映A7县2	2019/8/26	1.高峰期1	来信人：尊	2019/9/2 1	100	85.12332	92.93249	41.17647	93	1.886792	77.41762884
5	119367	UU008216	投诉K7县3	2017/10/2	K7县3	网友：	2017/10/3	100	97.01044	89.53094	8.823529	93	9.433962	76.26843146
6	24890	UU008107	咨询长株潭	2018/8/31	我是A7县[	网友：您们	2018/9/5 1	100	66.36533	94.33235	61.76471	95	15.09434	75.84282733
7	24793	UU008407	投诉A7县[	2019/1/21	尊敬A7县[	网友：您好	2019/1/30	100	89.27215	87.31143	23.52941	91	13.20755	75.66022729
8	24974	UU008157	反映A7县[	2018/4/15	尊敬的领导	网友：	2018/4/26	100	74.27255	92.95379	55.88235	89	1.886792	75.62013954
9	49782	UU008154	呼吁C5市[	2017/10/1	尊敬C	“UU00815	2017/10/1	100	84.64837	86.73107	26.47059	98	3.773585	74.81310203
10	108184	UU008383	呼吁政府督	2018/4/26	各位领导们	网友：您好	2018/5/7 1	100	80.21053	85.38636	44.11765	89	3.773585	74.69179455
11	119214	UU008913	对K7县教[	2018/8/27	本以为K7[	“UU00891	2018/8/29	100	92.68215	85.02015	11.76471	98	0	74.64271399
12	28966	UU008864	希望A9市[	2017/5/24	黎书记	“UU00886	2017/5/27	100	92.73545	75.08395	17.64706	97	9.433962	74.53087343
13	119872	UU008134	投诉L1区[	2019/12/8	L市L1区湖	网友：您好	2019/12/1	100	96.89644	73.92223	14.70588	95	1.886792	74.48390666
14	24694	UU008175	关于A7县[	2019/5/7 9	湖滨路以[	网友：您好	2019/5/9 9	100	88.13173	77.78658	26.47059	98	0	74.31293635
15	23647	UU008110	投诉A8县[	2017/5/31	你好，	网友“UU0(	2017/6/9 1	75	73.68896	93.53763	67.64706	91	11.32075	74.16640812
16	25690	UU008230	A7县安沙[	2015/10/5	尊敬的领导	UU008230	2015/10/1(	75	84.59983	83.4535	52.94118	95	5.660377	74.04456365
17	108148	UU008969	对K7县公[	2018/7/25	尊敬的市委	网友：您好	2018/8/2 1	100	83.01364	99.44351	11.76471	92	15.09434	73.93005542
18	140846	UU008166	再次请求在	2017/4/25	去年才	网友：	2017/5/10	75	94.59909	82.86188	41.17647	85	1.886792	73.8162613
19	30181	UU00810	A6区提倡[	2016/4/27	近日，	“UU00810	2016/5/6 9	100	83.18702	94.40944	20.58824	91	5.660377	73.80904125
20	6052	UU008199	A市12345[	2018/5/9 1	A市12345[	网友“UU0(	2018/5/24	75	96.30799	92.71773	26.47059	85	1.886792	73.62331578

图 21 答复意见评分情况

4 总结与未来展望

利用大数据、人工智能和文本挖掘技术将民众反映的“大事小情”汇总、整理和分类，设定热度评价指标，挖掘对影响市民生活、社会发展的普遍性和苗头性问题进行综合处理，对群众反映的行业热点问题进行专题研判，最后通过留言回复评价模型对政府人员的工作进行评估，确保真正做到“民有所呼，我有所应”，这一系列过程不仅是我们本次研究要实现的任务，也构成了当今网络问政平台必不可少的要素和功能。构建一个好的网络问政平台是民众行使知情权、参与权、表达权和监督权的重要渠道，架起了政府与民众沟通的桥梁，是帮助地方政府如何更好地履行职责的一个重要课题。而其中最大的一个难题是网民留言的多样性，包括表述方式的多样性、文化水平的多样性、民众情绪的多样性等，这都对处理网民留言造成了巨大的挑战，需要在技术上不断探索与研究。

通过对目前网络问政平台的留言进行研究，我们得到了以下启示：政府的网络问政从来不是一个纯粹的互联网技术问题，而是新技术背景下基层政府与基层社会的的关系问题，是一个基层治理的综合性问题。虽然互联网技术构筑了社会公众与基层政府沟通互动的新的渠道和桥梁，但是，政府将如何利用这些新的渠道和桥梁来回应社会关切、解决社会问题从而维护公众的合法合理利益，仍带有较强的不确定性。互联网技术并不能“自动”带来基层政社关系的革新，关键仍然在于政府对技术的运用方式与处理态度。所以在未来，除继续加强对现有网络问政相关人员进行全面系统的教育和培训外，各地应积极探索如何突破现有体制和政策性障碍，加快推进网络问政职业性、专业化人力资源队伍建设，优化政府网络回应机制。

参考文献

[1] 李娜娜. 基于 TextRank 的文本自动摘要研究[D].山东师范大学,2019.
[2] 结巴分词详细讲解.[EB/OL]. https://www.cnblogs.com/palace/p/9599443.html, 2018-09-06/2020-04-21.
[3] Hong T P, Lin C W, Yang K T, et al. Using TF-IDF to hide sensitive itemsets[J]. Applied Intelligence, 2013, 38(4): 502-510.
[4] 刘建伟, 付捷, 汪韶雷,等. 坐标下降 l_2 范数 LS-SVM 分类算法[J]. 模式识别与人工智能, 2013(05):60-66.
[5] Abdelwadood Mesleh． Chi square feature extraction based svms arabic language text categorization system[J]．Journal of Computer Science，2007，3 (6)：430-435．
[6] 高宝林,周治国,杨文维, 等.基于类别和改进的 CHI 相结合的特征选择方法[J].计算机应用研究,2018,35(6):1660-1662.
[7] 白光祖, 何远标, 马建霞, et al. 利用小样本量机器学习实现学术文摘结构的自动识别[J]. 数据分析与知识发现, 2014, 30(7).
[8] 陆彦婷，陆建峰，杨静宇. 层次分类方法综述[J]. 模式识别与人工智能，2013，26（12）：1130-1138.
[9] Pedregosa F, Gramfort A, Michel V, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2013, 12(10):2825-2830.
[10] 王兴玲,李占斌. 基于网格搜索的支持向量机核函数参数的确定[J]. 中国海洋大学学报(自然科学版), 2005,35(5):859-862.
[11] 万磊,张立霞,时宏伟.基于 CNN 的多标签文本分类与研究[J].现代计算机,2020(08):56-59+95.
[12] Srivastava N.，Hinton G.，Krizhevsky A.，et al. Dropout:A Simple Way to Prevent Neural Networks from Overfitting[J]. The Journal of Machine Learning Research，2014，15（1）:1929-1958.
[13] 罗锋. 基于深度学习的文本情绪多标签分类方法研究[D].山西大学,2019.
[14] 冯超. K-means 聚类算法的研究[D].大连理工大学,2007.
[15] 曹学艳,张仙,刘樑,方宽,段飞飞,李仕明.基于应对等级的突发事件网络舆情热度分析[J].中国管理科学,2014,22(03):82-89.
[16] 哈罗德·拉斯韦尔．社会传播的结构与功能［M］．何道宽，译．北京：中国传媒大学出版社,2012:45.
[17] 刘言君,黄婷,林建君.基于层次分析法的体育场馆类 APP 传播效果的评价研究[J].浙江体育科学,2018,40(06):46-52.
[18] 张秋立. 公共行政的精神[J]. 长沙大学学报, 2006(06):88-90.
[19] Zhu Z , Bernhard D , Gurevych I . A Multi-Dimensional Model for Assessing the Quality of Answers in Social Q&A Sites[C]// International Conference on Information Quality. DBLP, 2009.
[20] Soojung Kim, Sanghee Oh. Users' Relevance Criteria For Evaluating Answers In A Social Q&a Site[J]. Journal of the American Society for Information ence and Technology, 2009, 60(4):716-727.
[21] https://ai.baidu.com/ai-doc/NLP/ek6z52frp
[22] 杨开平, 李明奇, 覃思义. 基于网络回复的律师评价方法[J]. Computer Science, 2018, 45(9):237-242.
[23] 吴广建,章剑林,袁丁.基于 K-means 的手肘法自动获取 K 值方法研究[J].软件,2019,40(05):167-170.