基于集成数据挖掘方法的“智慧政务”文本挖掘

摘要

随着网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升。针对传统人工处理网络问政平台的群众留言分类、挖掘某一时段群众留言热点问题、留言答复意见质量评价的工作量大、低效率和不准确等问题，依据机器学习和深度学习等理论，利用数据清洗、文本向量表示、改进卷积神经网络的文本多分类器、F-score 评价等方法，高效和自动地解决了留言内容和标签的分类问题；采用数据预处理方法、文本标签向量化方法、改进 DB-SCAN 聚类方法、LDA 问题主题抽取方法、EMB-CRF 命名实体识别方法、层次分析法，建立了留言向量标签相似度计算的模型，提出了热度值计算公式，实现了热点问题的生成和热度值的排序，准确和及时解决了热点问题的挖掘问题。利用文献资料调研分析和曲线的趋势分析及层次分析法，确定留言答复质量评价指标体系及其权重；利用 python 爬虫技术和神经网络监督学习方法和模型，为题目附件 4 每条留言的各个指标打分；最后利用综合评价方法，获得留言答复质量的评价结果，有效解决了留言答复意见质量评价的问题。文中提出的文本挖掘思想、方法和技术及形成的评价方案可为改进政务服务的质量提供理论和技术支撑。具体如下：

针对问题一，利用正则表达式去重、结巴分词、去除停用词等数据清洗方法对题目附件 2 的留言详情进行数据预处理，得到 9172 条有效留言数据，并用 Skip-gram 模型进行特征向量转化，作为卷积神经网络多文本分类器的输入。利用卷积神经网络和机器学习中的朴素贝叶斯方法进行了留言的多文本分类，并用查准率、查全率、 F1 值、F1 均值、准确率 5 个人指标比较了它们的分类效果，得到卷积神经网络分类器在文本的多分类中优于朴素贝叶斯方法。特别利用 Skip-gram 模型结合随机生成的词向量，类比彩色图片中的三通道，使用改进的双通道文本表征方式，形成了基于改进卷积神经网络的双通道多文本分类模型，发现 F1 均值提高了 2.7%分，表明改进模型的分类效果更好。

针对问题二，采用问题 1 预处理数据的方法，对题目附件 3 中留言数据进行预处理。利用 Skip-gram 模型与 MinHash 算法分别对留言主题与留言详情进行向量标签化，并构造留言向量标签相似度计算模型；在此基础上，利用改进的 DB-SCAN 的 Label-Vec 密度聚类方法实现对留言详情的聚类，本次聚类总共得到 197 个留言簇。对每个留言簇利用 LDA 的问题主题抽取方法生成留言问题主题，利用 EMB-CRF 模型对问题主题进行地点和人群识别；利用层次分析构建影响留言热度的指标体系并求得指标的权重；再提出各指标的打分规则，依据题目附件 3 数据为指标打分；接着利用构造的热度计算公式得到相应留言簇的热度值，最终求得排名前 5 的“热点问题表.xls”与“热点问题留言明细表.xls”，详见附录附

件 1 与附件 2。

针对问题三，利用文献资料调研分析、层次分析法和曲线的趋势拟合分析等方法，分析研究和发现影响答复意见质量的机理，并根据指标确定的原则和方法确定了留言答复意见的质量评价指标体系。设置了 5 个一级评价指标和 10 个二级评价指标，并用层次分析法确定了每个指标的权重。利用 python 爬虫在“四川问政”平台抓取精选栏留言回复、答复评分等数据，通过数据整合、数据降维与转化，得到各个二级指标的主观得分；再选用神经网络监督学习方法为二级指标打分。最后利用综合评价方法，得到留言答复质量的综合评价方案，即将留言评价的每个指标的权重与得分的乘积相加得到该留言答复的质量分数，求取所有留言答复质量平均分；并依据提前设定的评语集，得到本次题目附件 4 中政务留言综合答复质量评价结果为“质量较高”。

关键词：智慧政务 改进 CNN　Label-Vec 聚类　LDA 主题抽取 层次分析法

目录

1 挖掘目标

1.1 挖掘背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，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的工作带来了极大挑战。随着大数据、云计算、人工智能等现代信息技术的发展，面对大量的数据和信息，人们越来越倾向于利用计算机对数据和信息进行处理，不但可以提高相关操作的效率，还可以在一定程度上提高相关操作的准确度。

信息挖掘和检索、自然语言处理是目前数据管理的关键技术，而文本分类则是这些技术进行操作的重要基础，其成为了人工智能（AI）子领域自然语言处理(NLP)的一个重要分支，因此建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率都具有极大的推动作用。

1.2 挖掘目标

目标一、 群众留言分类。在处理网络问政平台的群众留言时，当前的处理方法是工作人员首先按照一定的划分体系(问题题目附件 1 提供的内容分类三级标签体系)对留言进行分类；然后将群众留言分派至相应的职能部门处理。请你们针对目前大部分电子政务系统还是依靠人工根据经验处理中存在的工作量大、效率低、且差错率高等问题，依题目附件 2 给出的数据，建立关于留言内容的一级标签分类模型，并考虑用 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2\times P_i \times R_i}{P_i + R_i} \qquad (1\text{-}1)$$

其中 P_i 为 i 类的查准率， R_i 为 i 类的查全率。

目标二、热点问题挖掘。热点问题是指某一时段内群众集中反映的某一问题，它的及时发现有助于相关部门进行有针对性地处理，提升服务效率。请根据题目附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表. xls”。

目标三、答复意见评价。根据题目附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

2 问题分析

2.1 问题一分析

根据题目附件 2 的数据可知群众留言一级分类标签总共为 7 个，因此我们建立的标签分类模型是要解决一个文本多分类问题。因此第一问要做的工作就是，首先对题目附件 2 中的留言详情数据清洗，包括去除字母、数字、汉字以外的其他字符，jieba 进行分词，去除停用词等过程；随后进行特征向量表示，构造文本分类器，将数据分为测试数据和训练数据，分别进行模型的训练和测试；最后再利用 F-Score、查准率、查全率对分类器的留言分类效果进行检验，最终评价构造的文本分类模型的好坏。

2.2 问题二分析

根据题意，本题要求对热点问题进行挖掘，以便及时发现热点问题。为此，首先在问题一的基础之上，我们应对题目附件 3 中的留言详情和留言主题进行数据清洗等一系列操作；然后考虑留言详情和留言主题的具有长短文本的差异性，分别选择相应的算法将文本表示转化为特征向量，并计算每条留言的相似度；其次再选取相应的聚类算法对留言详情进行聚类，在得到留言聚类簇之后应进行问题主题的抽取，得到相应的类别留言的问题主题，在问题主题中包括我们需要的地点人群和问题描述；接着选取命名实体模型对问题主题的内容进行地点人群识别，地点人群识别后的问题主题内容作为问题描述；最后构建热度评价指标体系，提出热度值计算公式求得相应留言簇的热度值，并按照热度值进行排序截取热度值前 5 的热点问题，形成热点问题表.xls 与热点问题留言明细表.xls。

2.3 问题三分析

根据题意，要求我们针对题目附件 4 的相关留言的答复意见进行质量评价并实现。首先我们应进行广泛文献调研，找到影响答复意见质量的指标，如答复的相关性、完整性、可解释性、响应性等因素；然后选用基于层次分析法综合确定影响答复质量因素级，查阅文献专家经验构造评判矩阵并确定权重；然后在“四川问政”平台上用 python 爬虫抓取各个二级指标的主观得分情况，选用监督学习方法进行模型训练；其次利用该模型为每个指标进打分，将每个指标的权重与得分的乘积相加，可得到该留言答复的质量分数；最后求得所有留言答复质量分平均值，依据提前设定的评语集，得到政务留言答复质量评价结果。

3 模型假设与符号说明

3.1 模型假设

数据输入是有限个，即卷积神经网络（CNN）节点有限；

样本量足够大，能够满足 CNN 训练；

产生的序列足够短，能够正常处理池化；

留言详情对应的问题主题符合狄利克雷分布的先验分布；

任意问题主题对应的单词也符合迪利克雷分布的先验概率。

3.2 符号说明

关键符号	符号说明
C_{h*v}	卷积核窗口 h 行词向量维度 v 列的卷积核
$f(x)$	神经元激活函数
P_i	第 i 类的查准率
R_i	第 i 类的查全率
$\bar{F}_1$	所有类别的 F_1 的均值
$Jaccard(A,B)$	留言文本集合 A、B 的 Jaccard 相似系数
$Simde_{i,j}$	留言详情相似度
$SimTh$	留言主题相似度
$Sim(\text{Messages}_i, \text{Messages}_j)$	留言 i 与留言 j 的相似度
V	LDA 主题抽取的词典集合
T_C	留言详情问题主题集合

4 问题一模型建立与求解

4.1 建模和求解思路

（1）分类模型构建。首先读取题目附件 2 中全部的留言详情和一级标签数据，将得到的数据的 90%作为模型训练，10%作为模型测试；然后对读取的数据利用正则表达式对特殊字符进行清洗；接着进行结巴分词；其次我们利用网上现有的停用词表进行合并整理并去重，提取一个比较全面的停用词表用于去除数据中存在的大量虚词、名词、标点符号等；最后对得到了一种非结构化文本数据转成词向量，构造一个 M. embedding Size 大小的随机矩阵，构建之后再构造卷积神经网络文本多分类模型，并将随机矩阵作为卷积神经网络模型的输入，本次设计的群众留言分类架构图如图 4-1。

（2）分类模型评价。利用查准率(Precision)、查全率(Recall)、F1 值、$\bar{F}_1$均值、准确率(ACC)五种指标评价我们构建的卷积神经网络文本多分类模型与朴素贝叶斯分类模型的分类效果。

（3）对模型进行调优。利用 Skip-gram 模型结合随机生成的词向量，类比彩色图片中的三通道，使用改进的双通道文本表征方式，形成了基于改进卷积神经网络的双通道多文本分类模型，从而 ACC 和 F1 值都得到了极大的提高。

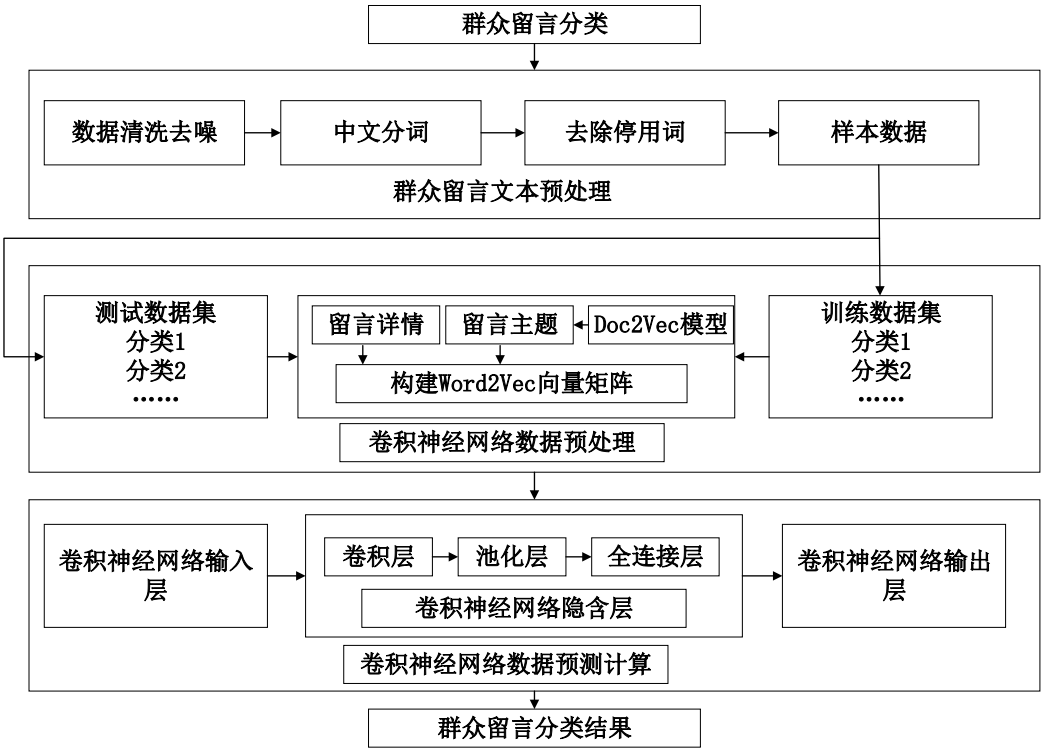


图 4-1 群众留言自动分类架构图

4.2 基于卷积神经网络的文本多分类模型的建立

包括数据清洗、分词和词性、去停用词、文本特征选择、文本表示、文本相似度、构建模型和模型验证调优。

4.2.1 数据预处理

（1）分词

分词是本次文本挖掘预处理过程中必不可少的步骤，常见的中文分词程序主要有中科院的张华平开发的 NLPIR 中文分词软件、结巴分词、清华大学实验室研制的 THULAC 中文词法分析包，通过文献调研发现三种分词方法在长文本分词中，如果综合考虑准确率、速度和使用的难易度的特性效果后，结巴分词效果更佳[1] ，因此本次数据挖掘选用结巴分词方法。

结巴分词支持三种不同的分词模式，本次综合考虑三种分词模式的分词效果选用了精确模式对数据集进行处理。

（2）去除停用词

在题目附件 2 中的留言数据中，存在大量的虚词、助词和没有特定含义的动词、名词、标点符号等，这些词语对文本分析起不到任何帮助，因此我们首先去掉这些“停用词”。去除停用词操作合并在分词过程中，即在分词时，增加查询停用词这一操作，如果查询到，则不作为分词结果。

本次数据挖掘过程，首先根据网上现有资源，对“哈工大停用词词库”、“百度停用词表”、“中文停用词表”等多种停用词表进行合并整理并去重，之后提取了一个比较全面的停用词表，一共 1908 个停用词，包括标点符号、中英文字符、使用频率较高的单汉字和无意义的词汇。从语料库中将这些对文本分类没有辨识度的停用词去除掉，可以降低特征的维度、提高关键词密度。

4.2.2 Skip-gram 词向量模型

经过前面文本预处理之后得到了一种非结构化文本数据，这类数据计算机无法直接处理，因此我们还需要将它转化为词向量。本次数据挖掘选用了 word2vec 工具包的 Skip-gram 模型进行词向量的转化和训练，Skip-gram 模型[2]如图 4-2。

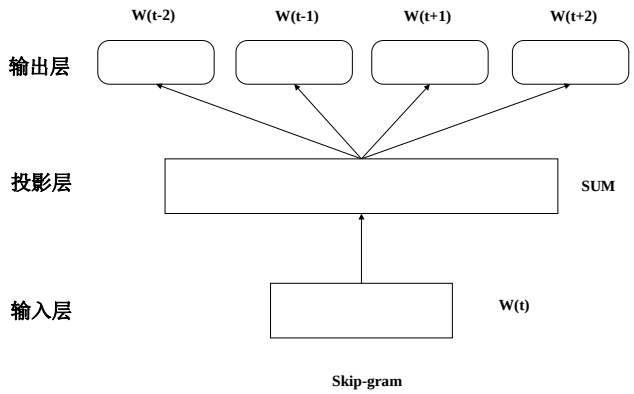


图 4-2 Skip-gram 模型

Skip-gram 模型是根据当前词语来预测上下文的词。它主要是将词分布式地映射到低维空间，并且，该低维空间中词向量的位置关系很好地反映了它们在语义上的联系，这样能很好地反映文本的特征。把做好后的词向量做成二维矩阵，作为改进卷积神经网络的输入数据。

假设留言文本中，经过数据预处理后，长度最长的留言包含，n 个词，该留言中的第 i 个词所对应的词向量是 $v_i \in R^{\mathrm{d}}$，那么卷积神经网络的输入就是由 n 个 d 维向量组成的 $n \times d$ 的二维矩阵。

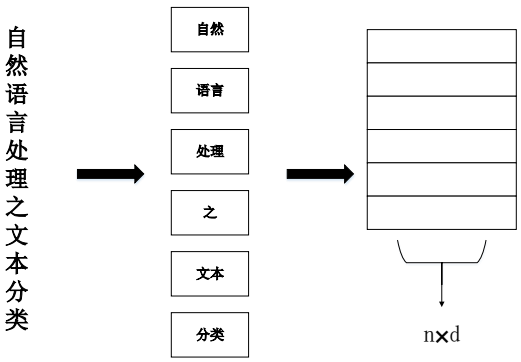


图 4-3 输入的二维矩阵

图 4-3 是文本数据经过预处理后，一个句子的矩阵表示，它是由句子中的所有词汇的词向量纵向拼接在一起，可以表示为：

$$v = v_1 \otimes v_2 \otimes \cdots \otimes v_n \qquad (4\text{-}1)$$

其中，$\otimes$ 是纵向拼接操作符；v 是一个样本留言的矩阵表示。

4.2.3 卷积神经网络（CNN）的文本多分类模型

CNN 模型包括输入层、隐含层以及输出层，利用梯度下降法最小化损失函数对权重参数逐层反向调节，通过迭代训练来提高模型分类效果。

本文设计的 CNN 模型如图 4-4 所示。

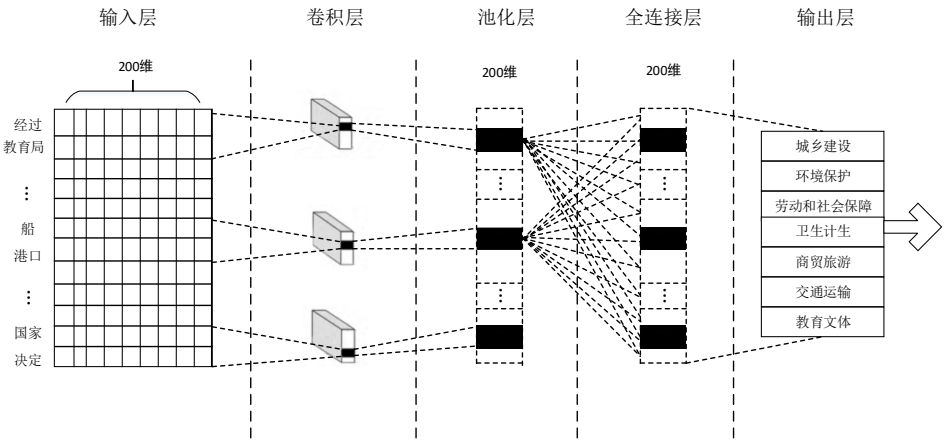


图 4-4 CNN 模型设计

（1）输入层设计

在利用卷积神经网络进行训练过程中由于使用梯度下降方法来进行学习，卷积神经网络的输入特征需要在输入层进行标准化处理。处理过程中将文本中经过分词处理以后的词对应的词向量依次排列形成特征矩阵作为输入数据传入卷积神经网络进行训练。每个词向量存储在利用 Skip-gram 网络结构提前训练好的词向量模型中，假设文本中有 n 个词，每个词向量维度为 v，那么这个特征矩阵就是 $n\times v$ 的二维矩阵。

（2）卷积层设计

通过内部包含的卷积核进行特征提取，特征提取的计算方法为[3]：

$$S_i = f(C_{h*v}T_{i:i+h-1} + b) \qquad (4\text{-}2)$$

其中 C_{h*v} 为卷积核，行数 h 为卷积核窗口大小，列数 v 为词向量维度，T 为文本特征矩阵，每个卷积核会依次与 h 行 v 列的特征矩阵做卷积操作，b 为偏置量。$f(x)$ 为神经元激活函数，在训练过程中为了防止神经元特征信息丢失以及克服梯度消失问题，设计中采用 Leaky ReLU 方法作为激活函数[4]：

$$f(x) = \max(0, x) + \gamma(0, x) \qquad (4\text{-}3)$$

式（4-3）为固定较小常数，通过卷积核特征提取后得到特征图：

$$S = [S_1, S_2, \cdots, S_{m-h+1}] \qquad (4\text{-}4)$$

在卷积层的设计过程中，考虑到一个卷积核提取特征存在不充分性的问题，在卷积层中包含了 C_{3*300}、C_{4*300} 以 C_{5*300} 3 种不同大小的卷积核，每个卷积核的操作模式设置为相同，每种特征图各提取出 100 张。最终在卷积层的输出端得到共 300 张特征图。

（3） 池化层设计

在卷积层进行特征提取后，由于特征图的维度还是很高，因此需要将特征图传递至池化层，通过池化函数进行特征选择和信息过滤。通过池化函数将特征图中单个点的结果替换为其相邻区域的特征图统计量，池化过程与卷积层扫描特征

图的过程相同[5]。在实验中采用最大池化函数(Max Pooling)对卷积核获取的特征保留最大值，同时放弃其它特征值。

（4）全连接层设计

对提取的特征进行非线性组合得到输出，全连接层本身不具有特征提取能力，主要用来整合池化层中具有类别区分性的特征信息，在实验中采用 Leaky ReLU 函数作为全连接层神经元的激励函数。

（5）输出层设计

使用多类交叉熵函数(Multi-class Cross Entropy) 作为损失函数以及归一化指数函数(Soft max)作为激活函数输出特征分类标签，完成文本分类任务。

4.2.4 分类算法性能评价指标

本次挖掘采用查准率(Precision)、查全率(Recall)、 F1 值、$\bar{F}_1$均值、准确率(ACC)五种指标来衡量文本分类器的效果，表 4-1 所示。

表 4-1 分类结果的混淆矩阵

预测 / 真实	预测为正例	预测为反例
真实为正例	TP（真正例）	FN（假反例）
真实为反例	FP（假正例）	TN（真反例）

（1）查准率 P 是指分类器预测为正且预测正确的样本占所有预测为正的样本的比例，计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{4-5}$$

（2）查全率 R 是指分类器预测为正且预测正确的样本占所有真实为正的样本的比例，计算公式如：

$$R = \frac{TP}{TP + FN} \tag{4-6}$$

（3）准确率 ACC 是综合评价分类结果的一个指标，准确率越高，则分类器越好，计算公式如下：

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4-7}$$

（4） F1 值是综合了 P 和 R 的一个指标，一般计算公式[7]，如下：

$$F_\beta = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \tag{4-8}$$

此时 F_β 是基于 P 和 R 的加权调和平均，$\beta > 0$ 度量了 R 对 P 的相对重要性，通常取 $\beta = 1$，此时公式退化为标准的 F_1 值，公式具体为：

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{4-9}$$

（5）留言文本是一个多分类结果，则 F_1 值的均值为 $\bar{F}_1$，公式为：

$$\bar{F}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{4-10}$$

其中 $0 \le F_1 \le 1$，当 P=1，R= 1 时，F1 值达到最大为 1，P 和 R 是一对矛盾的度量，当 P 高时，R 往往会偏低；当 R 高时，P 往往偏低，因此，在使用 F1 值越接近 1，则 $\bar{F}_1$ 值越接近 1，说明分类模型的性能越好。

4.3 基于卷积神经网络的文本多分类模型的求解与评价

（1）数据清洗

在题目附件 2 留言详情文本中，总共清洗出 9210 条留言作为样本数据，再按照每类留言平均分成 10 等份，每次实验抽取各类留言的 8 份组成训练集，剩余 2 份组成测试集，每次实验训练集数据为 7638 个，测试集数据为 1842 个。

（2）文本表示

将文本映射为词向量，造一个 M. embedding Size 大小的随机矩阵，M 是字典 dic 的大小 embedding Size 词向量的位数，我们设定为 128，并将随机向量矩阵作为卷积神经网络模型的输入。

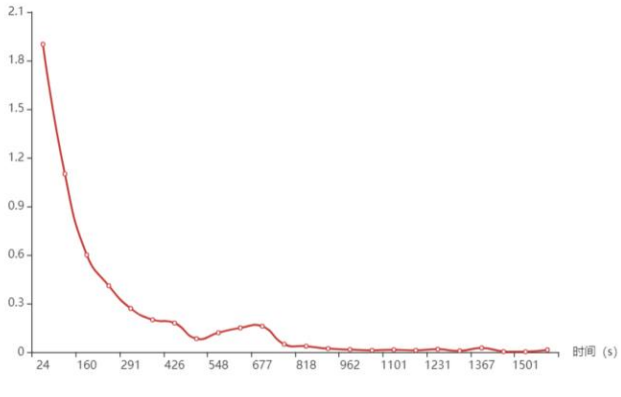
（3）模型训练与测试

在 CNN 模型完成之后，再将样本数据输入其中进行测试，本次数据挖掘采用十折交叉验证法进行 CNN 模型能力评估，将数据集划分为训练集和测试集，训练集用于模型训练，测试集用于评估模型性能，CNN 超参设置见表 4-2。

表 4-2 CNN 超参数设置表

超参数	描述	值
词向量	输入词向量维度	300
微批次	计算梯度和参数更新时一次使用的样例数量	32
正则化	避免在训练时发生过拟合，采用 L2 正则化	0.0001
学习率	学习率过低会导致模型参数更新变化太小，学习率过大会跳过局部极小	0.00001
激活函数	防止神经元特征信息丢失以及克服梯度消失	LeakyReLU
权重初始化	使用小随机数来初始化各网络层的权重，以防止产生不活跃的神经元	ReLU
dropOut	在训练过程中按照一定概率防止模型过拟合	0.5

最终得到卷积神经网络的分类器在留言分类文本数据集上的训练过程，并求得各评价指标的值，取 10 次实验结果的均值进行评估。模型训练过程中的 Loss 曲线如图 4-5 所示，ACC 曲线如图 4-6 所示。



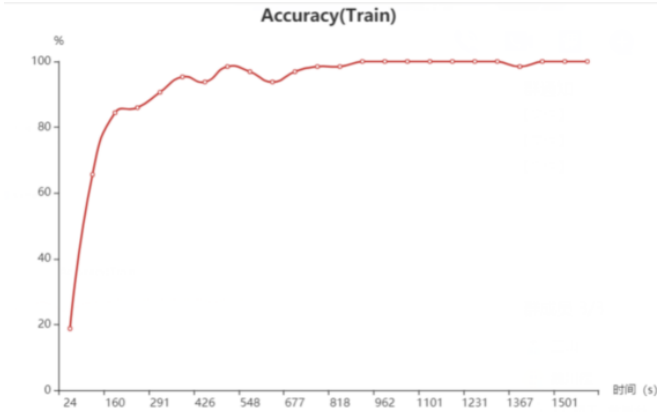


图 4-5 CNN 训练过程中的 Loss 曲线　　**图 4-6 CNN 训练过程中 ACC 曲线**

在图 4-5 中横坐标为训练时时长，纵坐标为 Loss 绝对值，可以发现训练过程中 Loss 曲线有明显下降的趋势，且后期开始慢慢收敛。

由图 4-6 中横坐标为训练时时长，纵坐标为 ACC 绝对值，ACC 指标呈快速上升趋势。可以发现随着训练进行，损失函数 Loss 明显降低，而 ACC 则明显上升，两者呈明显反比，符合预期。

为了对比 CNN 模型的分类效果，本次也选取了传统机器学习朴素贝叶斯文本分类方法[6] 作对比，最终分类效果如表 4-3 和表 4-4 所示。

表 4-3 CNN 与朴素贝叶斯分类结果对比

指标	留言分类
测试集	1842
字典大小	600
类别数	7
朴素贝叶斯（F1 均值）	0.85
CNN（F1 均值）	0.9
朴素贝叶斯（ACC）	0.86
CNN（ACC）	0.91

表 4-4 CNN 与朴素贝叶斯分类测试对比

分类测试结果 类别	CNN 分类模型			朴素贝叶斯分类模型			对应留言文本数目
	查准率	查全率	F-Score	查准率	查全率	F-Score	
城乡建设	0.87	0.9	0.88	0.82	0.83	0.83	799
环境保护	0.96	0.91	0.93	0.86	0.87	0.86	408
交通运输	0.88	0.91	0.84	0.8	0.67	0.73	292
教育文体	0.94	0.96	0.95	0.89	0.89	0.89	680
劳动和社会保障	0.93	0.94	0.94	0.88	0.91	0.9	688
商贸旅游	0.86	0.83	0.84	0.85	0.83	0.84	456
卫生计生	0.92	0.93	0.92	0.86	0.87	0.86	360

由表 4-3 和 4-4 可知，最终构建的卷积神经网络多文本分类测试结果，$\bar{F}_1$均值和 ACC 可以达到 90%，且在文本预处理和特征词项，都一致的条件下，基于

卷积神经网络模型的分类算法的分类精度比朴素贝叶斯算法高，可见本次挖掘所构建的基于卷积神经网络模型的文本分类器，不仅可行，而且有着更好的分类效果。

4.4 基于卷积神经网络的文本多分类模型的改进

通过前面的 CNN 分类器测试结果发现，卷积神经网络的词向量表征方式，过于简单。词向量的元素取自随机生成的，在[-1，1]区间上服从均匀分布，不能刻画词和词之间的相似性。

因此我们对模型进行改进调优，利用 Skip-gram 模型结合随机生成的词向量，类比彩色图片中的三通道，使用改进的双通道文本表征方式，形成了基于改进卷积神经网络的双通道文本多分类模型，改进双通道文本表征方式如图 4-7。

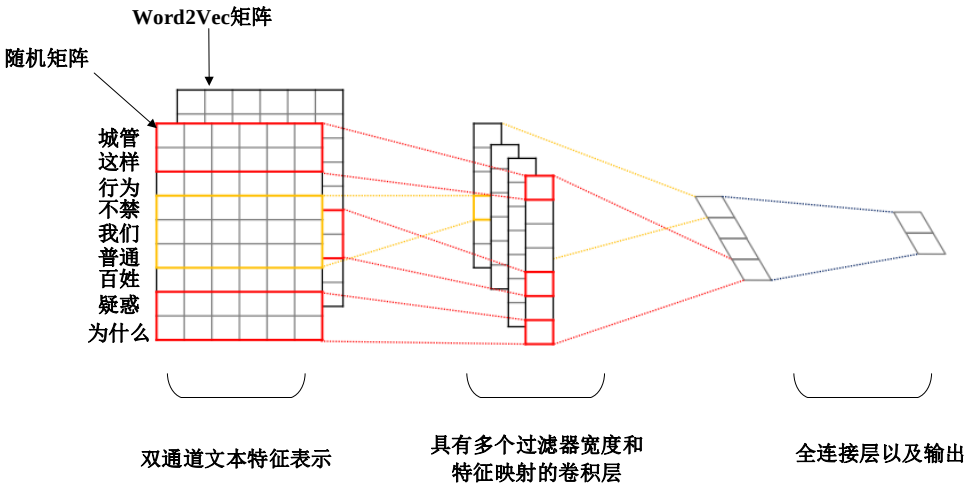


图 4-7 改进 CNN 双通道文本表征方式

最终改进前与改进后分类效果对比如表 4-5、4-6 和图 4-7。

表 4-5 改进 CNN 测试对比结果

测试集	1842
字典大小	600
类别数	7
未改进 CNN(ACC)	0.91
未改进 CNN(F1 均值)	0.9
改进 CNN(ACC)	0.934
改进 CNN(F1 均值)	0.927

表 4-6 CNN 改进前后各个留言类分类结果

分类测试结果 类别	CNN 分类模型（未改进）			CNN 分类模型（改进后）			对应留言文本数目
	查准率	查全率	F-Score	查准率	查全率	F-Score	
城乡建设	0.87	0.9	0.88	0.88	0.93	0.91	799
环境保护	0.96	0.91	0.93	0.95	0.92	0.95	408
交通运输	0.88	0.91	0.84	0.93	0.91	0.87	292
教育文体	0.94	0.96	0.95	0.97	0.98	0.96	680
劳动和社会保障	0.93	0.94	0.94	0.96	0.96	0.96	688
商贸旅游	0.86	0.83	0.84	0.87	0.87	0.88	456
卫生计生	0.92	0.93	0.92	0.96	0.96	0.95	360

由表 4-5 与表 4-6 可以发现，改进的双通道文本表征方式，比未改进前有了更好的效果：ACC 提高了 2.4%，$\bar{F}_1$均值提高了 2.7%。

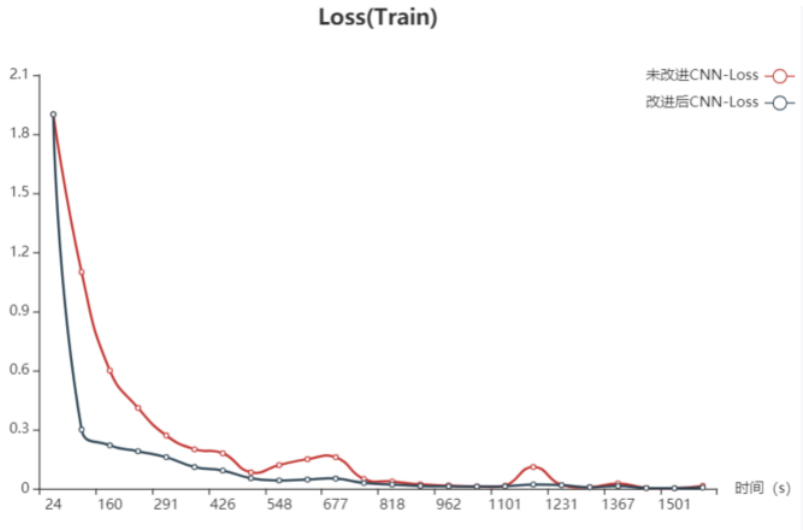


图 4-8 改进前与改进后 Loss 曲线对比

由图 4-8 可以发现，改进后的 CNN 对数损失(Loss) 曲线为蓝色，收敛速度明显快于未改进的 CNN 橙色曲线。

最终得出结论，深度学习模型在特征提取方面有天然的优势，Skip-gram 模型生成的词向量可以弥补了随机向量词之间缺乏联系的不足，双通道相比较单通道，输入特征更丰富，构造的分类器效果更佳。

5 问题二模型建立与求解

5.1 建模和求解思路

根据问题 2 中的热点数据挖掘问题的分析，首先对题目附件 3 留言数据进行预处理、分词、去除停用词之后；然后构建了留言文本向量化和留言相似度计算模型，分别实现对留言详情和留言主题的标签向量化和相似度计算；之后再构建 Label-Vec 密度聚类模型实现对留言的聚类，聚类过程如图 5-1，针对聚类结果再构建基于 LDA 的问题主题抽取模型，得到聚类后的留言簇的相应热点留言问题主题，留言问题主题中包含地点人群和问题描述；接着构建 EMB-CRF 模型对抽取的热点留言问题主题进行地点和人群进行命名实体识别提取，剩余的问题主题内容作为问题描述，再建立影响留言热度的指标体系并使用层次分析法求得各个指标的权重；最后通过构造的热度计算公式得到相应留言簇的热度值。

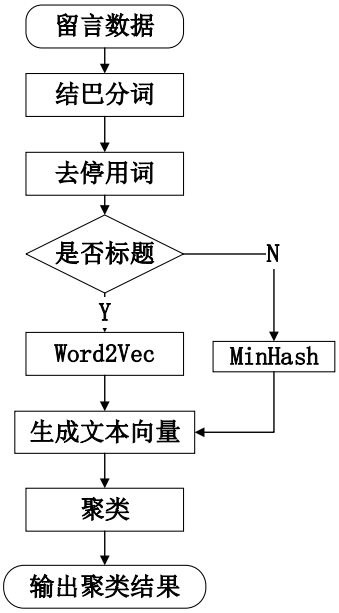


图 5-1 留言聚类过程

5.2 模型的建立

5.2.1 文本标签向量模型

（1）Jaccard 系数

Jaccard 系数是衡量两个文本之间相似度的系数。我们将留言文本进行分词，得到单词定集合 A 和集合 B， A、B 中共有的元素个数占 A、B 总元素个数的比重即为 A 与 B 的 Jaccard 相似系数。即

$$Jaccard(A,B)=|A\cap B|/|A\cup B| \qquad (5\text{-}1)$$

该系数的值域为 0 至 1， 系数越接近 1，两个文本之间的相似度越高。

（2）MinHash 算法

本次选用 MinHash 对留言详情进行相似度计算，理由是因为此方法的复杂度较低和标签简便可以降低后面聚类的比较次数。MinHash 的原理为，假设存在集合 A、B， 将集合 A、B 中的元素经过哈希函数哈希之后，如果其中具有最小哈希值的元素既在 A∪B 中也在 A∩B 中，那么 hmin(A) = hmin(B) ，集合 A 和 B 的相似度就可以表示为集合 A、B 的最小哈希值相等的概率[8]，即：

$$Jaccard(A,B)=Pr[\ hmin(A) = hmin(B)\ ] \qquad (5\text{-}2)$$

MinHash 算法计算留言详情标签流程为：

1）提取到的留言详情特征值向量和文档组成矩阵，通过多个哈希函数将特征值向量矩阵哈希成签名矩阵，签名矩阵中的数据都是降维后的结果。

2）将矩阵中的行号进行随机变换并排列，取出某一文档对应的列中排在最前面的 1， “1”对应的单词可以代表该文档，经过多次哈希随机变换就可以选取出多个特征值来代表这个文档。从而对文档实现了降维，高维度的文档被压缩为 n 个整数表示。

3）计算两两文档之间的 Jaccard 系数，得到留言详情之间的相似度。

（3）留言相似度计算

通过前面步骤 MinHash 算法，我们得到了留言详情文本向量。因为留言主题单词量不多，采用 MinHash 算法降维后会丢失很多信息，不适合短文本的提取，所以针对留言主题本次采用基于将采用基于 word2vec 工具包的 Skip-gram 的向量标签化方法。

设留言详情 MinHash 后标签的相似度记为 $Simde_{i,j}$，在计算留言主题相似度时，Jaccard 相似度可以直接衡量标签的重合程度，因此本次将 Jaccard 相似度作为标签相似度的一部分，记为 $SimTh^{J}$，对于不相同的单词，可以通过 embedding 的向量的余弦相似度来计算出单词间的相似度，其计算过程为：

Step1 输入两个不同的向量标签 M 和 N 以及相似度阈值 T。

Step2 通过异或操作求得 M 和 N 中各自独有的单词向量对应的向量集，记为 $M^{'}$ 及 $N^{'}$。

Step3 对于 $M^{'}$ 及 $N^{'}$ 中的任意一对单词组 (m_i,n_j)， 计算其余弦相似度 S。

Step4 如果 Step3 中的 $S>T$，则使总相似度 $SIM+S$。

Step5 重复 Step3 及 Step4，直至算法遍历所有单词对，输出最终相似度 SIM。

上述算法的 SIM 在留言主题标签的相似度计算中记为 $SimTh^{C}$，留言主题标签的相似度计算公式表示为：

$$SimTh=\frac{1+\gamma}{\frac{1}{SimTh^{J}}+\frac{\gamma}{SimTh^{C}}} \qquad (5\text{-}3)$$

其中 γ 为 theme 关于 Jaccard 相似度及余弦相似度的超参。将 theme 的 Jaccard 相似度与 details 标签的相似度按照系数累加，得到向量标签模型的相似度计算方法，其公式可以表示为：

$$Sim(\text{Messages}_i, \text{Messages}_j) = \alpha Simde_{i,j} + (1-\alpha) SimTh_{i,j} \qquad (5\text{-}4)$$

其中 α 是 title 相似度和 details 相似度的权重系数。

5.2.2 基于改进的 DB-SCAN 的 Label-Vec 密度聚类模型

由于每日产生的群众留言数是一个不确定值，而且往往有大量留言对应簇的样本数很低，这些类别的留言容易影响聚类精度，K-Means 等聚类方法难以剔除噪声数据的干扰，并且也难以选择合适的中心点数量。

基于上述原因，本次挖掘设计了基于改进 DB-SCAN 的 Label-Vec 密度聚类方法，并针对留言文本特点设计了合理的密度计算公式，从而实现了高效高精度的群众留言文本聚类。

本次设计的 Lab-Vec 密度聚类模型，解决了 DBSCAN 的复杂度较高、大样本集中聚类收敛时间过长等问题，通过采用桶思想，设定桶标签来实现 ε-邻域估计，聚类过程如图 5-2。

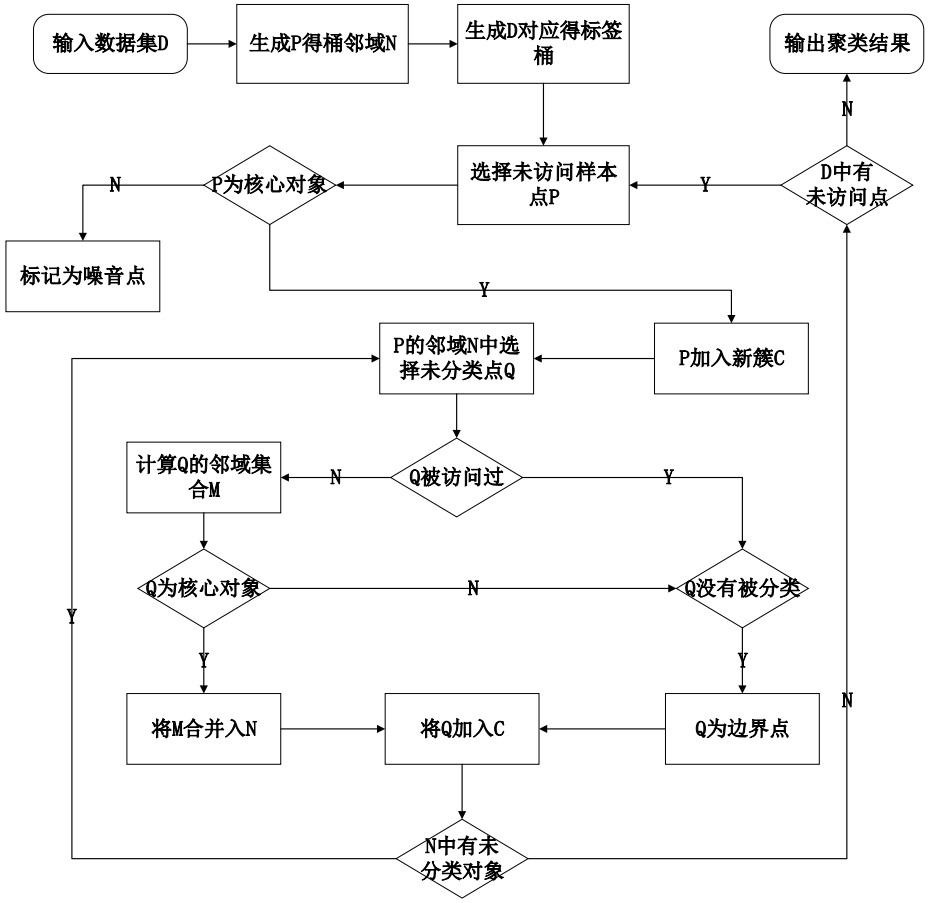


图 5-2 Lab-Vec 密度聚类流程图

输入：留言样本集 D，标签分割区间大小 A，标签哈希函数 H 及对应的标签

桶个数 m，邻域参数$(\delta, MinPts)$，样本距离计算公式见公式（5-4）。

输出：簇划分 C

Step1 初始化参数

初始化标签桶集合$\{M_{\mathrm{i}}=\varnothing \mid i=1,2\cdots m\}$、核心对象集合$\Omega=\varnothing$、未访问样本集合 T=D、聚类簇数 K=0、簇划分$C=\varnothing$。

Step2 更新标签桶

1) 对于样本集 D 中的样本$x_1, x_2, \cdots, x_m$，按照标签分割区间大小λ，将样本x_i的标签向量分割成$\lfloor |text|/\lambda \rfloor$个不同的标签子集。

2)对x_i的每个标签子集，通过标签哈希函数 H 生成每个区间对应的签名，组成签名集Key_{i}。

3)对于任意一个标签桶M_{j}，如果样本x_i的签名集Key_{i}中包含签名值 j，则将样本 x 加入标签桶M_{j}，则：

$$M_{\mathrm{j}}=M_{\mathrm{j}}\cup\{x_i \mid Key_{\mathrm{i}}\cap\{Key_{M_{\mathrm{j}}}\}\neq\varnothing,\ i=1,2,\cdots,m\} \qquad (5\text{-}5)$$

Step3 更新核心对象集合

1) 对于样本集 D 中的样本$x_1, x_2, \cdots, x_m$，计算样本x_i的所有桶邻接对象集：

$$\Gamma_i=\sum_{j=1}^{m}(M_{\mathrm{j}} \mid M_{\mathrm{j}}\cup\{x_i\}\neq\varnothing) \qquad (5\text{-}6)$$

2)计算样本x_i的$\delta-$邻域$N_\delta(x_i)$，即，在桶邻接对象集中的样本x_{j}，如果x_{j}的桶邻接对象集Γ_j与x_i的桶邻接对象集Γ_i交集个数$|\Gamma_j\cap\Gamma_i|\geq\delta$，则将样本$x_j$加入$x_i$的$\delta-$邻域$N_\delta(x_i)$。

3)如果$\delta-$邻域$N_\delta(x_i)$的样本个数$|N_\delta(x_i)|\geq MinPts$，则将样本$x_i$加入核心对象集合$\Omega$，即$\Omega=\Omega\cup\{x_i\}$。

Step4 判断是否结束

如果核心对象集合$\Omega=\varnothing$，则算法结束，并输出簇划分 C，否则进入 Step4。

Step5 初始化样本簇

在核心对象集合Ω中随机选择一个核心对象 o，初始化当前簇核心对象集$\Omega_{cur}=\{o\}$，初始化簇类别序号 K=K+1，初始化当前样本簇集合$C_{\mathrm{k}}=\{o\}$，更新未访问样本集合$T=T-\{o\}$。

Step6 判断簇生成结束

如果当前簇核心对象集$\Omega_{cur}=\varnothing$，则簇生成结束，更新簇划分$C=\{C_1, C_2, \cdots, C_K\}$，更新核心对象集合$\Omega=\Omega-C_K$，进入 Step3。

Step7 计算核心对象邻域

在当前簇核心对象集Ω_{cur}中随机选择一个核心对象 o'，根据样本距离计算公式（13）计算核心对象 o'的$\delta-$邻域$N_{\delta}(o')$，令$\Delta=N_{\delta}(o')\cap T$，更新当前样本簇集合$C_K=C_K\cup\Delta$，更新未访问样本集合$T=T-\Delta$，更新当前簇核心对象集$\Omega_{cur}=\Omega_{cur}\cup(\Delta\cap\Omega)-o'$，进入 Step5。

上述步骤聚类过程如图 5-2 所示。

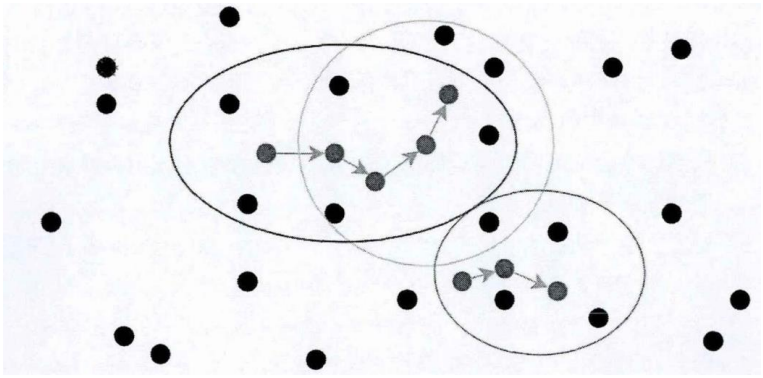


图 5-3 Label-Vec 法聚类过程

如图 5-3 所示，图中的椭圆形标志为不同的标签桶。对于每一个样本点，通过计算该样本点与所在的所有标签桶中的其他样本点的相似度，Lab-Vec 算法可以快速地寻找到核心对象。通过核心对象集， Lab-Vec 确认密度相连区域，从而完成聚类。

5.2.3 基于 LDA 的问题主题抽取模型

对于每一条群众留言，其对应的问题主题是所有问题主题集合的一个混合概率分布，而留言的每一个问题主题也是单词集合上的一个概率分布，则在 LDA 问题主题模型中，某条留言中第i 个单词是ω的概率为[9]：

$$P(\omega i)\sum_{j=1}^{T}P(\omega i\,|\,zi=j)P(zi=j) \qquad (5\text{-}7)$$

上式中，参数 T 为主题的个数、参数 z 为某个特定的主题、参数ωi指文档中的第i个单词为ω、参数 zi 指文档中的第i个单词所属的主题为 z，LDA 的问题主题抽取过程如下：

Step1 确定留言详情的问题主题

对于在某一聚类簇 C 中的所有留言详情，通过分词得到同一簇下文本集合的单词集合，并将单词集合作为 LDA 中的词典 V。对于数据集中的所有文本对象，采用 TF-IDF 算法统计词典中所有单词的权值，并将该聚类簇 C 中的核心文本的标题中权值超过给定阈值ε的单词作为主题，ε的选择需要保证每一条留言详情对应的初始问题主题个数不少于 1 个。

Step2 初始化迪利克雷超参

将 Step1 中得到的每一条留言详情的初始问题主题个数作为θ_d的迪利克雷

分布的初始参数[9]。并随机初始化迪利克雷分布 β_k 的超参数 β。

Step3 求解问题主题集合

求解 LDA 收敛后的 θ_d 和 β_k 并得到聚类簇 C 下所有留言对应的留言详情问题主题集合 T_C。

Step4 确定热点话题

对于聚类簇 C 下所有留言主题预处理后，按照 TF-IDF 权值进行排序，将权值最高的句子作为聚类簇 C 的留言热点问题。

Step5 求解所有聚类簇留言热点问题

对于每一个聚类簇按照上述步骤生成所有簇对应的留言热点问题。

5.2.4 基于 EMB-CRF 的中文地点人群命名实体识别

在提取到了热点问题之后，我们需要对热点问题中的地点和人群进行识别提取，目前中文命名实体识别领域主要有基于规则和基于统计两类方法。基于规则的方法通常需要对命名实体的构成规则进行归纳总结，通常需要维护大量的规则，缺乏适应性。因此本文采用的是基于统计的方法进行命名实体识别。具体使用了字位置嵌入的增强型条件随机场用于中文命名实体识别算法，用于提取地点人群。

（1）字位置嵌入表示

中文字标注命名实体识别中，引入词边界信息对于实体识别效果有提升[10]。因此，本文使用了字位置嵌入进行热点问题地点人群的识别。

字位置嵌入是字嵌入与词嵌入的折中，具体为：字加上字在词中的位置作为字的表示，使用这种表示来训练嵌入向量。首先我们对每个前面提取的每个簇对应的留言热点问题进行分词；然后对词中的每个字根据其位置信息进行表示，如词“天安门”与“蓝天”中，“天”的表示分别为“天 0"和“天 1”。

（2）EMB-CRF 命名实体识别

本次引入嵌入表示的增强型条件随机场模型(以下简称为 EMB-CRF)。该模型在传统的特种工程上，结合深度学习方法得到的嵌入特征，改进了条件随机场模型。于是，本文使用字位置嵌入与 EMB-CRF 模型结合，用于热点问题地点人群的命名实体识别。图 5-4 是一个使用了字位置嵌入的 EMB-CRF 中文命名实体识别模型。

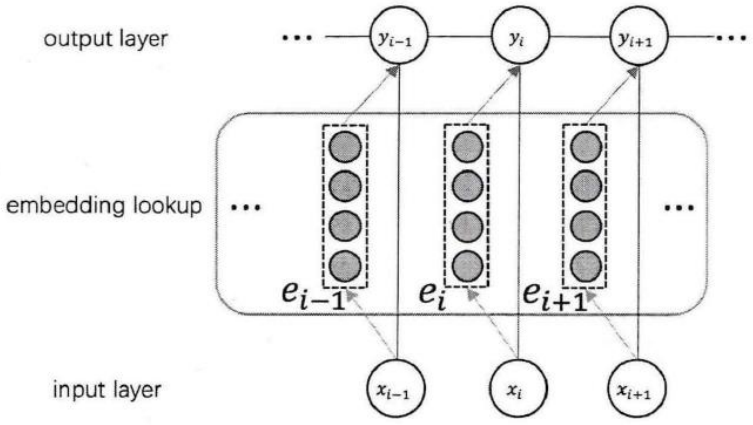


图 5-4 中文命名实体识别 EMB-CRF 结构

图中圆角矩形框中，表示对输入的查找操作(将输入的文字转换为对应的字位置嵌入)，之后下方虚线箭头表示将对应的嵌入 e_i 作为参数送入输出层；输入层 x_i 与输出层 y_i 之间的实心圆球连接表示传统的特征函数。

为与改造的结构相适应，本文将 e_i 作为特征引入到条件随机场的模型中。与传统的特征函数不同，词嵌入特征是低维稠密向量特征，可以使用简单的线性变换的方式将其与传统特征整合。对原始条件随机场模型改进如下：

$$P(y \mid x)=\frac{1}{Z(x)}\exp\left(\sum_{k=1}^{K}\theta_k F_{\mathrm{k}}(y,x)+\sum_{i=1}^{n}\sum_{j\in L}\mathrm{I}(\mathrm{y}_i=j)W_{\mathrm{j}}^{T}\mathrm{e}_{\mathrm{i}}\right) \qquad (5\text{-}8)$$

其中， $Z(x)$ 称为规范化因子，可表示为

$$Z(x)=\sum_{y}\exp\left(\sum_{k=1}^{K}\theta_k F_{\mathrm{k}}(y,x)+\sum_{i=1}^{n}\sum_{j\in L}\mathrm{I}(\mathrm{y}_i=j)W_{\mathrm{j}}^{T}\mathrm{e}_{\mathrm{i}}\right) \qquad (5\text{-}9)$$

$\mathrm{I}(\mathrm{y}_i=j)$ 函数表示当 $\mathrm{y}_i=j$ 时取 1，否则返回 0；e_i 为词 x_i 对应的词嵌入参数；W_j 表示标注类别为 j 时对应的权重系数向量。作如上改进的原因是：基于一元得分是构成特征函数的集合中激活函数权值之和这一原理，将低维稠密向量特征经过线性变换后与传统一元得分求和，从而引入词嵌入特征。这样，本文就建立了一个可以同时使用传统特征和词嵌入特征的条件随机场模型。

5.2.5 热度评价指标模型的建立

在对留言进行聚类和 LDA 问题主题抽取后，通过查阅文献和相关专家经验，我们将影响留言的热度因素划分为三个：留言互动数（点赞数与反对数总和）、留言持续时间、所在分类簇留言条数。

（1）用层次分析法计算每个影响因素在留言热度贡献的权重

1）采用 Saaty 的标度矩阵，构造留言热度影响因素比较矩阵 W 如下：

表 5-1 留言热度影响因素比较矩阵

M	留言互动数	留言持续时间	所在分类簇留言条数
留言互动数	1	1/3	1/5
留言持续时间	3	1	1/3
所在分类簇留言条数	5	3	1

$$M=\begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{5} \\ 3 & 1 & \frac{1}{3} \\ 5 & 3 & 1 \end{bmatrix} \quad (5\text{-}10)$$

2）每个影响因素权重计算

依据层次分析法，通过 matlab 计算对矩阵计算可得最大特征根 $\lambda_{\max}=5.37$，M 的最大特征值对应的特征向量为 $[0.2347,0.1565,0.6088]^{T}$，计算一致性指标：

$$CI=\frac{\lambda_{\max}-\mathrm{n}}{n-1}=0.0938$$

计算随机一致性比率，3 阶矩阵 $RI=1.12$

$$CR=\frac{CI}{RI}=0.0837<0.1$$

可知判断矩阵式（5-10）具有满意的一致性，因此矩阵 W 的特征向量归一化后可以作为权向量为 $[0.0347,0.0965,0.8688]^{T}$。

3）计算留言的热度值

依据 2）求得的留言热度各个影响因素的指标权重，则可得出群众留言热度值计算公式（5-11）为：

$$S=H_{1(\text{interaction})}W_1+H_{2(d\text{uration})}W_2+H_{3(n\text{umber})}W_3 \quad (5\text{-}11)$$

公式（5-11）中，$H_{1(\text{interaction})}$ 代表所在分类簇的互动数指标的得分，其计算公式如下：

$$H_{1(\text{interaction})}=\text{点赞数}+\text{反对数} \quad (5\text{-}12)$$

$H_{3(n\text{umber})}$ 代表所在分类簇的留言条数指标得分，计算如下：

$$H_{3(n\text{umber})}=\text{簇内留言条数} \quad (5\text{-}13)$$

$H_{2(d\text{uration})}$ 代表所在分类簇的持续时间指标的得分，计算如下：

$$H_{2(d\text{uration})}=\frac{\text{簇内留言条数}}{\text{最后一条留言时间}-\text{第一条留言时间}} \quad (5\text{-}14)$$

W_1、W_2、W_3 分别代表留言互动数、留言持续时间、所在分类簇留言条数所占的权重， S 代表该留言话题热度的最终得分。

由于主观性数据的值处于 0—100 之间，然而我们经过挖掘得到的客观性数据往往能达到几万不等，二者之间不在同一层次上因而无法比较。

因此本次还引入无量纲化方法，将获得的数据进行标准化，其目的是将不同量纲的数据统一为相同的量纲，从而能进行下一步热度值的计算。

无量纲化有以下两个公式(5-15)和(5-16)，两公式的试用对象不同。公式 (5-15)适用于正向性的效益型指标，公式(5-16)适用于负向性的成本性指标。

$$A=\frac{a_{ij}-a_{i\min}}{a_{i\max}-a_{i\min}} \tag{5-15}$$

$$A=\frac{a_{i\max}-a_{ij}}{a_{i\max}-a_{i\min}} \tag{5-16}$$

公式(5-15)和(5-16)中的$a_{i\max}$代表数据值中的最大值，$a_{i\min}$代表数据值中的最小值。计算得出的值在 0—1 之间，由于热度评价值的标准为 0—100 分的范围，所以将标准化后得到的数值扩大 100 倍，最终得到的指标分数全为 0—100 区间的值。

5.3 模型的求解与分析

5.3.1 算法思想与求解步骤

（1）算法思想

对于本题的热点数据挖掘，首先对题目附件 3 的留言详情和留言标题进行数据预处理，利用正则表达式等方法除去换行符、制表符、标点符号等除字母、数字、汉字以外的其他字符；之后再利用 jieba 进行分词，利用停用词表去除停用词，随后我们将题目附件 3 的留言主题和留言详情定义为两个标签 theme 标签与 details 标签；再将 theme 标签与 details 标签分别使用 Skip-gram 模型与 MinHash 算法进行向量标签化，分别计算 theme 标签与 details 标签的相似度，并将两者结果累加得最终到该条留言向量标签模型的相似度。

利用改进的 DB-SCAN 的 Label-Vec 密度聚类方法实现对留言详情进行聚类，得到了各个留言聚类结果，随后再利用 LDA 的问题主题抽取方法，最终生成热点留言问题主题，其包含地点人群和热点问题描述。对抽取的留言热点问题主题使用 EMB-CRF 模型进行地点和人群命名实体识别提取，剩余的问题主题内容作为热点问题描述；然后构建影响留言热度的指标体系，用层次分析法求得各个指标的权重，通过构造的热度计算公式得到相应留言簇的热度值；其次按照留言簇内留言热度值的高低进行排序，得到排名和排名前 5 的热点问题表.xls；最后按照题目所给出的格式，得到热点前 5 的热点问题留言明细表.xls。

（2）求解步骤

关于热点问题挖掘求解过程如下图 5-5 所示。

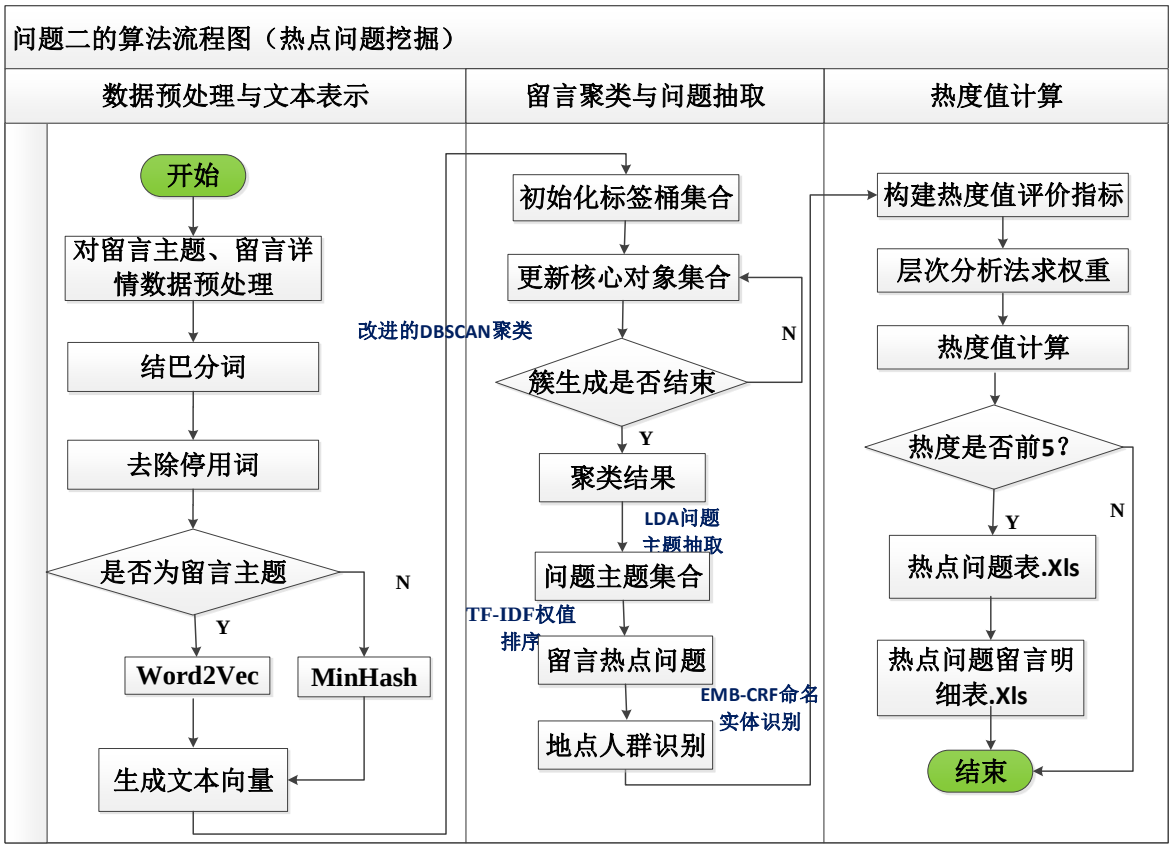


图 5-5 问题二热点问题挖掘求解流程图

5.3.2 求解结果与分析

在求解此热点数据挖掘问题时，利用 python3.6 和 Matlab12.0 编程求解，程序见附录附件 3，其运行环境见表 5-2。

表 5-2 运行环境说明

硬件环境	操作系统	Windos10
	主频	2.79GHZ
	内存	8GB
软件环境	软件	python3.6
	软件	Matlab12.0
	重要算法	Label-Vec 聚类算法
	重要算法	LDA 问题主题抽取

结果一 ： 基于 Label-Vec 密度聚类的留言归类结果

对题目附件 3 的数据去重清洗，得到 4208 条有效留言数据，并将数据代入图 5-5 的算法求解步骤中，利用 python 求解得到 Label-Vec 密度聚类结果，本次聚类总共获得 197 个留言簇， Label-Vec 密度聚类图见图 5-6。

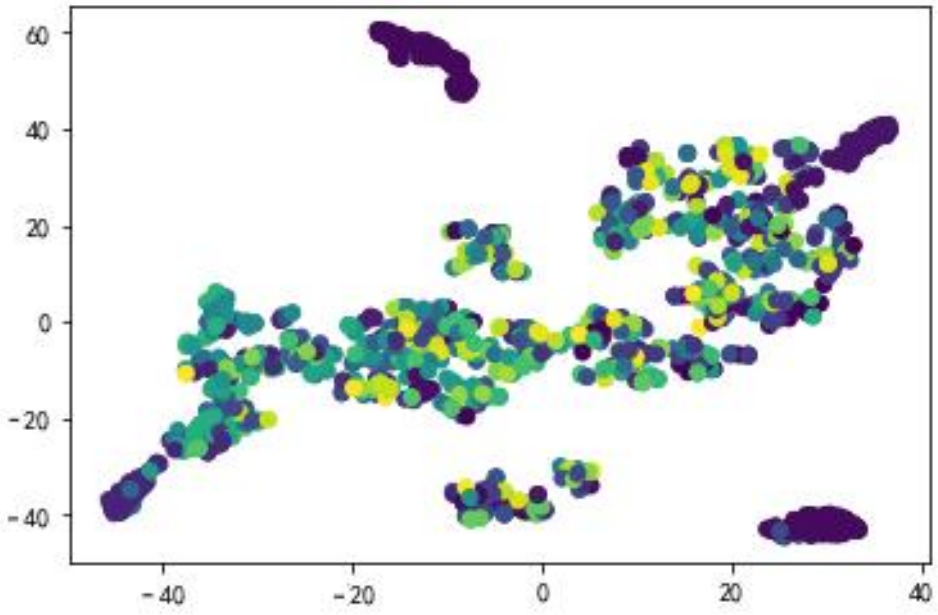


图 5-6 Label-Vec 密度聚类图

根据图 5-6 可以发现本次改进的 Label-Vec 密度聚类模型，在留言详情长文本聚类效果符合预期，且有较明显的优势。

最终将 197 个聚类结果绘制的热点留言绘制分布条形图和分布饼图，如图 5-7、图 5-8。

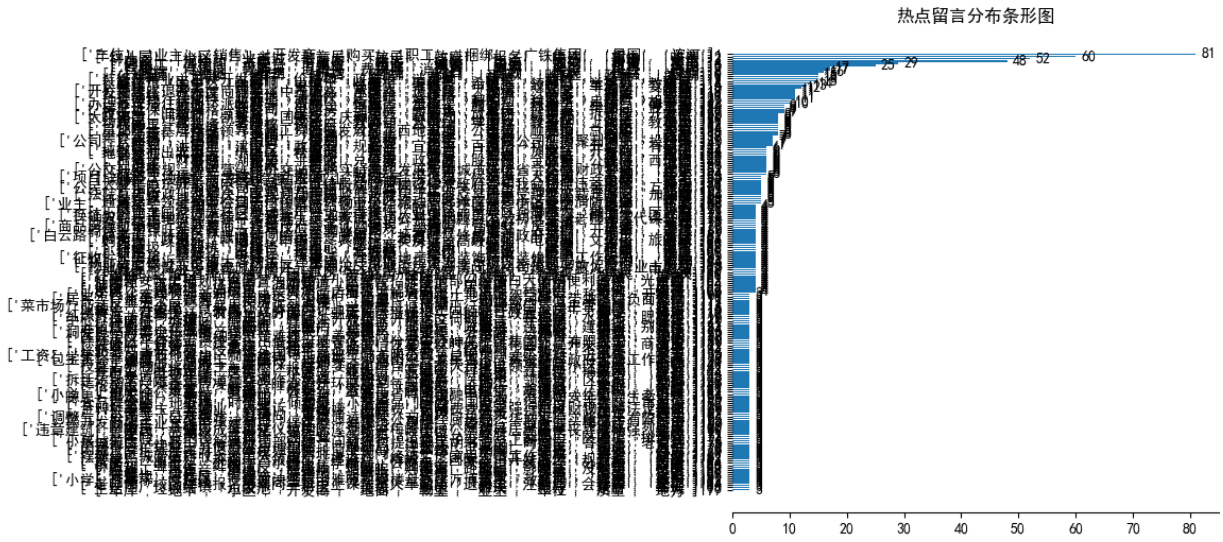


图 5-7 留言分布条形图

通过图 5-7 中我们可以发现，留言条数排名前五的依次为 81 条、60、52、48、29 条，并可以得到依据留言条数排名的热点留言话题关键词。

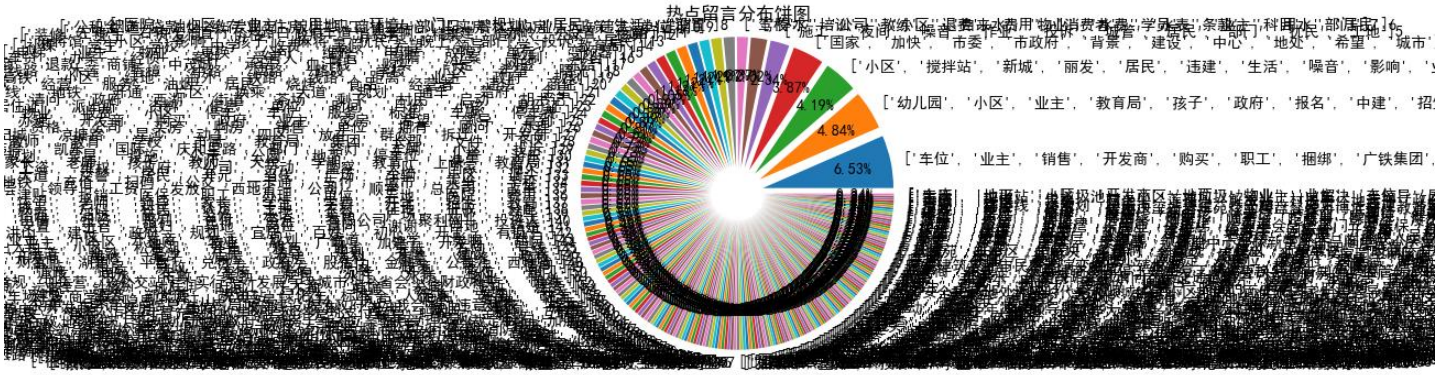


图 5-8 热点留言分布饼图

通过图 5-8 可知，聚类中留言条数最多的话题为“捆绑销售车位”，其留言比例占到整个留言的 6.53%。

结果二：基于 LDA 问题主题抽取与热度值计算结果

根据图 5-5 的算法求解步骤，将聚类结果带入 LDA 问题主题抽取模型中，利用 python 求得相应留言簇的问题主题。问题主题包括地点人群和问题描述；然后进行地点人群识别求解和相应留言簇的热度值求解；其次按照热度值进行排序取前五的热度值形成热点问题，并把热点问题绘制词云图如图 5-7 所示。

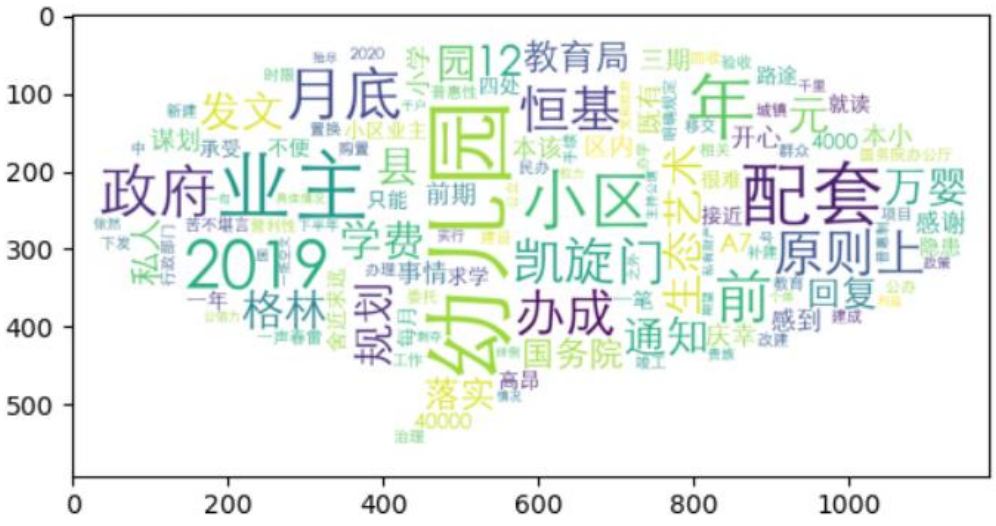


图 5-7 前 5 的热点话题词云图

最后求得排名前 5 的热点问题表.xls 如表 5-3，和热点问题留言明细表.xls 如表 5-4，具体详见附录附件 1 与附件 2。

表 5-3 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	1	94.88	2019/07/07 至 2019/09/06	伊景园滨河苑职工	集团坑客户购房金额并捆绑销售车位
2	2	87.62	2019/03/03 至 2019/12/20	A7 县三一街区/邦盛水岸御园儿童	幼儿园的入学问题
3	3	83.98	2019/01/08 至 2019/07/08	A 市地铁/医疗中心	建议加快建设力度
4	4	82.41	2019/11/02 至 2020/01/15	A2 区丽发新城小区	违建搅拌站灰尘和噪音扰民
5	5	80.93	2019/03/28 至 2020/01/03	A 市 A3 区咸嘉湖路熊家湾巷/A5 区万科金域华府别墅区/A2 区竹塘西路—新姚路	项目凌晨施工噪音严重扰民

表 5-4 热点问题留言明细表

分类编号	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
1	188801	A909180	投诉滨河苑针对广铁职工购房的霸王规定	2019/8/10:00	尊敬的张市长，您好！我叫李建义，来自湖北仙桃...	0	0
1	190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019/8/23 12:22	投诉伊景园.滨河苑开发商捆绑销售车位...	0	0
1	191001	A909171	A 市伊景园滨河苑协商要求购房同时必须购买车位	2019/8/16 9:21	商品房伊景园滨河苑项目是由 A 市政府办牵头为广铁集团铁路职工定向销售的楼盘...	1	12
1	195511	A909237	车位捆绑违规销售	2019/8/16 14:20	对于伊景园滨河苑商品房，A 市广铁集团违规捆绑车位销售至今...	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	188887	A00085665	关于 A7 县恒基凯旋门万婴格林幼儿园办普惠园的咨询	2019/7/17 10:19:08	敬的领导：您好！我是家住恒基凯旋门三期业主，凯旋门...	1	4
2	198132	A00063808	A3 区中海国际社区幼儿园无法满足实龄幼儿就读需求	2019/8/12 15:32:27	A3 区中海国际社区有一个普惠幼儿园和一个公立幼儿园...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

通过图 5-7 与表 5-3、5-4 可知当前的最热的问题为“小区旁高铁噪音扰民”问题相关部门，应针对性对此前 5 个问题进行处理，这样才能有效的提升最大服务效率。

6 问题三模型建立与求解

6.1 建模和求解思路

根据 2 问题分析中对答复意见的评价分析，我们选择层次分析法对留言答复意见的质量进行模糊综合评价，首先我们广泛查阅文献和相关政务、问政平台对留言答复质量的指标刻画并改进[11]，设置了 5 个一级评价指标——相关性、可解释性、完整性、移情性、响应性，10 个二级评价指标（见图 6-1）；然后通过文献调研和专家经验，利用层次分析法确定每个指标的权重；随后在“四川问政”平台上用 python 爬虫抓取各个二级指标的主观得分情况，随后构建神经网络模型进行监督学习，并进行模型检验；对题目附件 4 的答复意见数据进行预处理等操作，选取了 2812 条答复意见数据，依据构造的监督学习模型对答复意见数据进行为每个指标打分，最终依据构造的答复质量计算公式（式 6-1），得到每条留言的答复质量得分，再求均值得到整个数据的综合答复质量得分，并依据答复质量评语集，最终得到政务留言答复质量评价结果。

6.2 模型的建立

6.2.1 答复意见质量评价体系构建

（1）评价体系框架

通过文献调研影响评价质量的因素，可得答复意见质量评价体系一级指标五个、二级指标十个如图 6-1 所示。

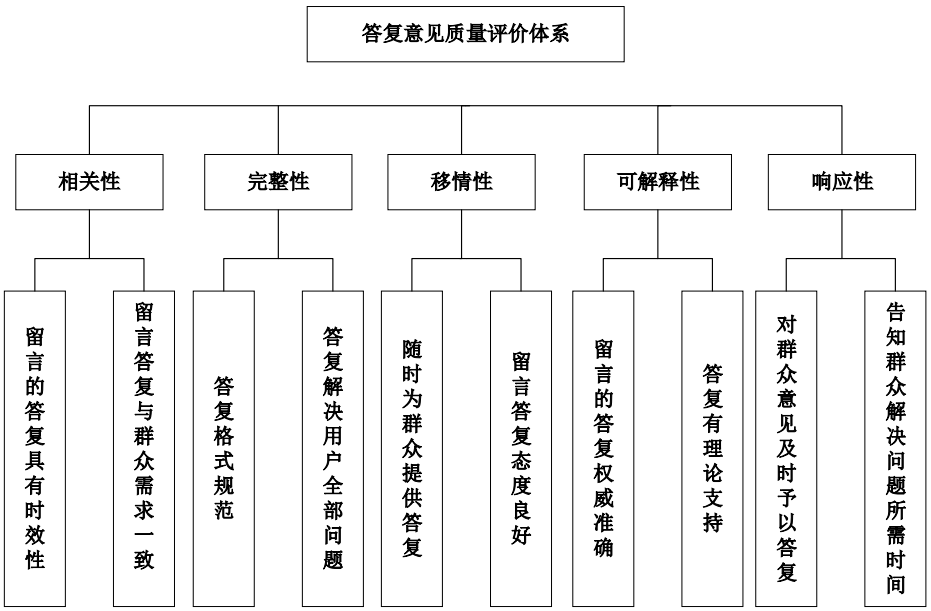


图 6-1 答复意见质量评价系统框架

（2）按照以上的答复意见评价体系框架，评价答复意见质量的指标体系如

下表 6-1 所示。

表 6-1 评价答复意见质量的指标体系

	一级指标	二级指标
答复意见质量评价体系	U1 相关性	U11 答复时效性
		U12 答复与需求一致
	U2 完整性	U21 答复格式规范
		U22 答复解决群众全部问题
	U3 移情性	U31 答复态度良好
		U32 随时为群众答复
	U4 可解释性	U41 答复有理论支持
		U42 答复权威准确
	U5 响应性	U51 对群众意见及时答复
		U52 告知群众解决问题所需时间

以上每个二级指标含义如下：

答复时效性：留言答复时间是否在有效时间内答复。

答复与需求一致：留言答复内容是否与用户问题一致。

答复格式规范：留言答复是否满足一定的规范。

答复解决群众全部问题：留言答复内容是否解决群众全部问题。

答复态度良好：留言答复态度是否良好。

答复有理论支持：留言答复是否有理论支持，是否给出文件、法律条款等支持。

答复权威准确：留言答复是否具有权威性。

对群众意见及时答复：群众发出问题后等待回复的时间是否低于平均回复时间。

告知群众解决问题所需时间：留言答复是否告知群众问题解决所需时间。

6.2.2 评价指标权重的确定

判断矩阵中的打分通过咨询专家获取，专家通过填写咨询表确定指标间的相对重要性，未通过一致性检验的打分表与专家沟通商议后重新调整打分。本研究共征询了 10 位专家的意见，其中大学教师 7 人，博士生 3 人。权重计算以专家一的打分为例进行的说明。

本模型的评价系统通过层次分析法确定权重，按照上述评价指标体系，设计判断矩阵，准则层判断矩阵如下：

表 6-2 准则层判断矩阵

	相关性	完整性	移情性	可解释性	响应性
相关性	1	3	5	1	3
完整性	1/3	1	5/3	1/3	1
移情性	1/5	3/5	1	1/3	3/5
可解释性	1	3	5	1	3
响应性	1/3	1	5/3	1/3	1

判断矩阵的一致性检验：

判断思维的一致性是指专家在判断指标重要性时，各判断之间协调一致，不致出现相互矛盾的结果。检查专家判断思维的一致性时，一致性判断指标 CI 越小表明判断矩阵的一致性越好。衡量不同阶判断矩阵是否具有满意一致性时使用随机一致性比率 CR，当其小于 0.1 时，认为判断矩阵具有满意一致性。

对此计算矩阵，计算的最大特征根 $\lambda_{\max}=5.13$，M 的特征向量为：$[0.3457,0.1152,0.0781,0.3457,0.1152]^T$。计算一致性指标，因为 $CI=(\lambda_{\max}-n)/(n-1)$，则对于 n=5 的表的矩阵数据，我们可以得到其一致性比率为：$CR=\frac{CI}{RI}=0.0293<0.1$，所以一致性检验通过。

表 6-3 相关性判断矩阵

	答复时效性	答复与需求一致
答复时效性	1	1/3
答复与需求一致	3	1

对此计算矩阵，计算的最大特征根 $\lambda_{\max}=2$，M 的特征向量为：$[0.25,0.75]^T$。

表 6-4 完整性判断矩阵

	答复格式规范	答复解决群众全部问题
答复格式规范	1	3/7
答复解决群众全部问题	7/3	1

对此计算矩阵，计算的最大特征根 $\lambda_{\max}=2$，M 的特征向量为：$[0.3,0.7]^T$。

表 6-5 移情性判断矩阵

	答复态度良好	随时为群众答复
答复态度良好	1	1/3
随时为群众答复	3	1

对此计算矩阵，计算的最大特征根 $\lambda_{\max}=2$，M 的特征向量为：$[0.25,0.75]^T$。

表 6-6 响应性判断矩阵

	答复有理论支持	答复权威准确
答复有理论支持	1	2
答复权威准确	1/2	1

对此计算矩阵，计算的最大特征根 $\lambda_{max}=2$，M 的特征向量为：$[0.667,0.333]^T$。

表 6-7 相关性判断矩阵

	对群众意见及时答复	告知群众解决问题所需时间
对群众意见及时答复	1	5/7
告知群众解决问题所需时间	7/5	1

对此计算矩阵，计算的最大特征根 $\lambda_{max}=2$，M 的特征向量为：$[0.4167,0.5833]^T$。

根据各层次的单层次权重的计算，可以获得“答复意见质量评价体系”评价指标体系的权重如下表：

表 6-8 答复意见质量评价体系权重

一级指标	权数	二级指标	权数	总权重
相关性（U1）	0.3457	U11 答复时效性	0.25	0.0860
		U12 答复与需求一致	0.75	0.2592
完整性（U2）	0.1152	U21 答复格式规范	0.3	0.0346
		U22 答复解决群众全部问题	0.7	0.0806
移情性（U3）	0.0781	U31 答复态度良好	0.25	0.0195
		U32 随时为群众答复	0.75	0.0586
可解释性（U4）	0.3457	U41 答复有理论支持	0.667	0.2306
		U42 答复权威准确	0.333	0.1151
响应性（U5）	0.1152	U51 对群众意见及时答复	0.4167	0.0480
		U52 告知群众解决问题所需时间	0.5833	0.0672

6.2.3 基于层次分析法与监督学习的答复意见质量综合评价

（1）答复质量值计算

根据表 6-8 确定的指标权重和各末级指标的得分，从末级指标开始依次对各层级指标进行加权求和。根据计算公式（6-1）最终得出待验证实际留言答复意见的质量值，计算公式（6-1）为：

$$q=\sum_{i=1}^{n} y_i w_i \qquad (6\text{-}1)$$

公式（6-1）中的 y_i 代表第 i 个指标得分，这里指标代指二级指标；w_i 代表第 i 个指标的权重； n 指标的个数；q 代表实际留言答复意见的质量值。

由于主观性数据的值处于 0—100 之间，然而我们经过挖掘得到的客观性数据往往能达到几万不等，二者之间不在同一层次上因而无法比较。类似于在 5.2 节中的热度评价指标模型和计算公式（5-14）与（5-15），引入无量纲化方法，将获得的数据进行标准化。

（2）指标得分值确定

打分规则一：对答复与需求一致（U12）打分

本次使用了计算留言详情与答复意见的相似度 $Sim(\text{Messages}_i, \text{Messages}_j)$，具体计算公式为式（5-14），计算方法同理文本聚类过程计算两条留言详情相似度，并将相似度分数转化为 0—100 之间。

打分规则二：对答复时效性（U11）打分

首先设定留言等待时间的对应得分区间，随后使用答复时间减去留言时间得到等待时间，最终依据设定的等待时间区间获得 U11 指标的得分。

打分规则三：对 U21、U22、U31、U32、U41、U42、U51、U52 打分

本次类比问题一，利用监督学习的方法得到影响答复意见指标的分数，具体实现过程如下图 6-2：

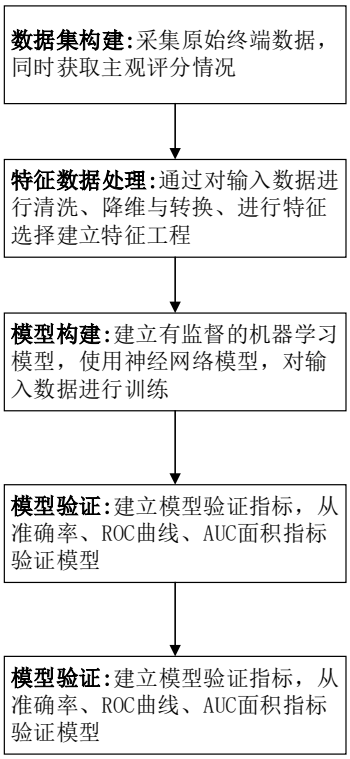


图 6-2 监督学习指标打分整体框架

Step1 数据集构建

本次使用了网络爬虫技术，通过 Python 的 Scrapy 框架对“问政四川”政务平台内的的精选留言问题、官方回复、答复评价、答复评分、服务态度等数据进行抓取，抓取网址： (https://ly.scol.com.cn/welcome/showlist?keystr=wzrd)，通过数据整合、数据降维与转化、数据特征标准化等工作，最终得到主观关于 U21、U22、U31、U32、U41、U42、U51、U52 指标的评分情况。

Step2 机器学习模型构建

本次类别问题一构建神经网络模型，其具有监督学习机制，拥有强大的拟合能力，设计算法如下：

1)根据得到的数据集确定此模型输人层的神经元个数。根据终端参数特征，

全连接神经网络包含 1 个输入层、1 个输出层和 2 个隐含层。

2)初始化神经网络权重并设置学习率。运用服从标准正太分布的随机数来设置初始权重，选择一个作为基准，确定一个最好的算法方案。

3)使用 Sigmoid 函数作为激活函数，该函数能够对数据进行非线性变化，将输人值压缩到 0~1，使之有强大的解析数据的能力。Sigmoid 函数公式为：

$$Sigmoid\left(x\right)=1/\left(1+e^{x}\right)$$

4)反向传播。本步骤需要定义损失函数，计算损失函数关于神经网络中各层的权重的偏导数(梯度)，使用梯度下降的方式优化神经网络参数。

5)重复以上过程，直到模型收敛为止。

随着训练次数的增加，模型的损失值震荡减少。当模型损失值不再变化时，表示训练结束，说明模型已经收敛。

Step3 模型验证

为了验证本方案机器学习的神经网络模型的算法可行性，本次采用准确率、ROC 曲线和 AUC 面积模型相结合的方法来衡量模型优劣的指标。与准确率和 ROC 曲线相比较，AUC 面积更能保证模型的准确度，如图 6-3 所示。图 6-3 中横坐标表示假正例率，纵坐标表示真正例率，完美预测的 AUC 大小为 1，表明该模型 AUC 值接近 1。AUC 越大，正确率越高。无论是准确率还是可信度，均有大幅提升，评估效果较好。

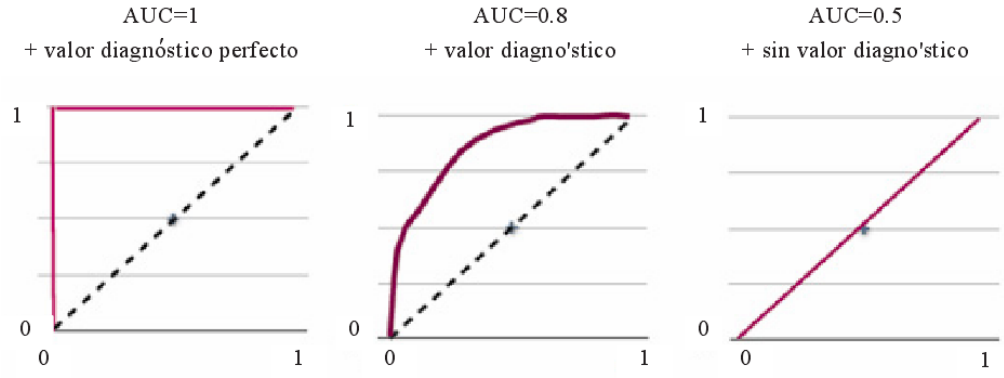


图 6-3 AUC 模型

Step4 模型进行指标打分

在通过模型检验了其正确性之后，便通过构造的监督学习神经网络模型，依据题目附件 4 的留言详情与留言答复，对 U21、U22、U31、U32、U41、U42、U51、U52 指标进行打分，最终或者各个指标，在本次附件 4 所给答复意见的质量指标得分，最终再依据公式（6-1）求得题目附件 4 中 2812 条有效留言的各自答复质量评价得分与整体答复质量平均得分。

（3）建立评语集

根据对主客观数据的无量纲化处理，以及公式（6-1）计算出的答复质量值，建立评语集，以百分制为评价区间，即群众留言答复意见的质量值在[0~100]之间

[12]，v(评语集) $=\{V_1$质量很高、V_2质量较高、V_3质量一般、V_4质量低 $\}$对应的分数集本次取$F=\{\ 100-85,\ 85-70,\ 70-55,\ 55$ 以下 $\}$。

V_1、V_2说明答复意见的质量很高，相关部门工作效率和问政服务公众很满意。V_3说明答复质量值一般，有待改进和提升。V_4答复意见质量较低，需要密切关注，采取措施进行改进服务质量。

6.3 模型的求解与分析

首先依据打分规则一、二在题目附件 4 留言数据中对指标 U11、U12 打分；其次依据规则三，在构造的神经网络监督学习模型通过模型验证后，依据此模型对题目附件 4 中 2812 每条留言中的 U21、U22、U31、U32、U41、U42、U51、U52 规则打分，最终得分结果以及加权分结果如下表 6-9 所示（由于篇幅有限仅展示前后 5 条留言得分数据）。

表 6-9 各条留言最终答复意见质量值

留言编号	U11	U12	U21	U22	U31	U32	U41	U42	U51	U52	加权分
2549	70	80	90	80	90	58	76	88	60	89	77.99
2554	58	90	90	90	90	57	82	77	58	82	79.85
2555	60	90	90	90	90	57	77	62	57	70	76.28
2557	57	90	90	80	90	57	78	70	57	50	75.03
2574	56	90	90	90	90	55	80	77	56	82	79.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
181267	57	30	32	57	67	32	52	53	31	40	43.83
181603	42	37	62	85	72	27	62	80	28	88	55.95
184423	82	90	90	85	90	32	68	70	31	70	73.91
185799	7	90	90	90	90	56	76	76	7	68	70.51
185986	4	90	90	90	90	56	63	83	4	40	66.04
平均值	62	86	92	91	62	48	59	78	53	88	84.06

上述表 6-9 展示了每一条留言编号所对应的二级标签的所获得的评分，根据每个二级标签的得分数乘上对应的权值就可以计算出该条留言的一个总的得分情况，根据这个总的得分情况所在区间就能识别出该条留言的答复质量情况，实现对留言答复的质量评价。

最终求得每条留言的答复意见质量得分后，依据评语集可得到留言的答复质量等级，各个留言答复质量分布如图 6-4 所示。

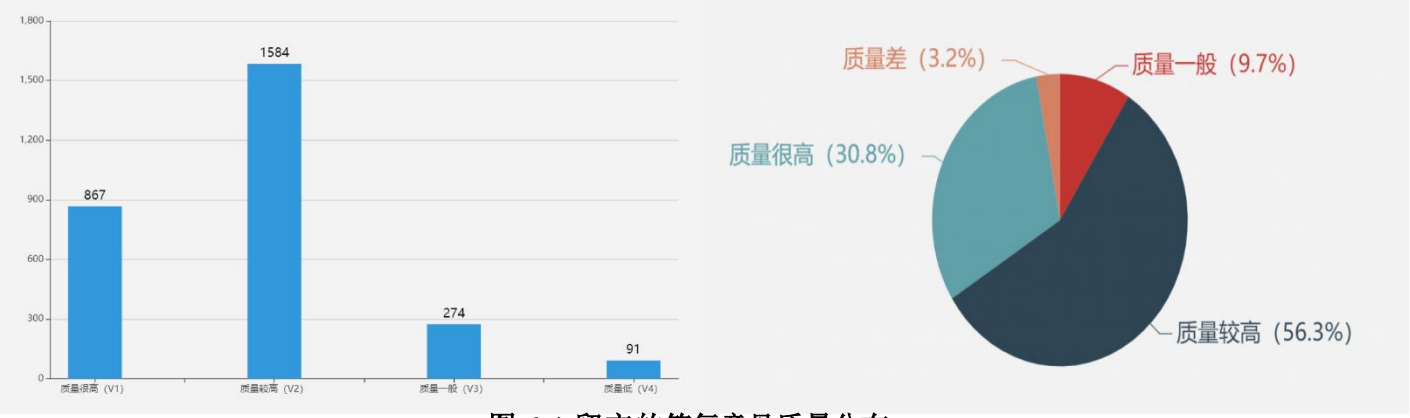


图 6-4 留言的答复意见质量分布

由上述表 6-9 的得分情况可知，求平均值后整体的留言意见答复质量为 84.06 分，对应评语集为答复质量较高；可以发现答复与需求一致 U12、答复格式规范 U21、答复态度良好 U31 三个指标分数相比其他指标得分普遍较高，则表明在留言答复的时候在答复与需求一致、答复格式规范、答复态度良好方面做得很好。

从表 6-9 中，我们也可以看出：虽然对留言答复得比较快速(10-20 天内)，并且答复都比较人性化，但是最主要的关于留言答复的完整性，切题方面并没有做得很好，而这才是衡量一条答复对于留言的响应处理质量高的标准，因此每个政务人员应该提升自己的专业素养，更好服务于民。

通过图 6-4 可以发现，整体留言评语集分布在答复质量较高与质量很高占到了 87.1%，且依据表 6-9 答复质量平均得分 84.06，可以综合判定本相关部门答复意见质量较高。

7 总结

本文通过深度学习、机器学习等技术，使用卷积神经网络分类构造器、DBSCAN 聚类、层次分析等方法构造了多文本分类模型、热点提取模型以及答复意见的评价系统来解决“智慧政务”中的文本挖掘问题，得到以下结论：

（1）对于“智慧政务”中文本多分类问题而言，文本预处理和特征词项都一致的条件下，基于卷积神经网络模型的分类算法的分类精度略比朴素贝叶斯算法高一些，可见本次挖掘所构建的基于卷积神经网络模型的文本分类器对于“智慧政务”的文本分类有着不错的效果。对于词向量生成而言，使用 Skip-gram 模型生成的词向量一定程度上弥补了随机向量词之间缺乏联系的不足，更能体现词向量之间的联系。最后我们改进了模型，使用改进的双通道文本表征方式，双通道相比单通道，输入特征更丰富，而深度学习模型在特征提取方面有天然的优势，两者得到了很好地融合。

（2）对于留言详情的向量转化采用 MinHash 算法，该方法的复杂度较低和标签简便可降低留言详情相对较多的文本聚类的比较次数。针对留言主题采用基于 Skip-gram 的向量标签化方法以保证重要信息不会过度丢失。通过我们设计的 Lab-Vec 密度聚类模型来实现文本的聚类，解决了 DBSCAN 的复杂度较高、大样本集中聚类收敛时间过长等问题。通过 LDA 的问题主题抽取和 EMB-CRF 命名实体识别来抽取话题以及地点人群可以很好地给决策者提供热点话题的情况，让决策者能从大量数据中快速找到热点问题，实现“智慧政务”。

（3）采用文献调研，我们建立了含 5 个一级指标、10 个二级指标的答复意见质量评价体系，能够从相关性、完整性等各个方面系统全面地评价留言回复的质量。评价指标权重通过征询专家意见确定，我们通过建立准则层判断矩阵并且进行一致性判断得出我们的指标权重是十分合理的。最后通过答复质量值计算以及建立评语集可以让决策者以更加直观的形式，观看到留言回复情况以及目前的留言回复存在哪些问题，帮助决策者进行管理。

综上所述，我们的模型能够高效、准确地解决“智慧政务”中的文本挖掘问题，实现真正的智慧政务管理。

8 参考文献

[1]白璐.基于卷积神经网络的文本分类器的设计与实现[D].北京交通大学，2018.

[2]黄鹤，荆晓远，董西伟，吴飞.基于 Skip-gram 的 CNNs 文本邮件分类模型[J].计算机技术与发展，2019，29(06)：143-147.

[3] 孙璇. 基于卷积神经网络的文本分类方法研究[D]. 上海：上海师范大学，2018.

[4]卢玲，杨武，杨有俊，等.结合语义扩展和卷积神经网络的中文短文本分类方法[J]. 计算机应用，2017， 37 (12)： 3498- 3503.

[5] Shen Y，He X，Gao J， et al. A Latent Semantic Model with Con volutional - Pooling Structure for Information Retrieval [J]. Pro-ceecdings of the 23rd ACM International Conference on Conference on Information and Knowledge Management-CIKM 14，2014： 101-110.

[6]马小龙.网络留言分类中贝叶斯复合算法的应用研究[J].佛山科学技术学院学报(自然科学版),2013,31(02):43-47+68.

[7] Zhou Zhihua. Machine Learning [M]. Beijing： Tsinghua University Press，2016， 30-32.周志华、机器学习[M].北京：清华大学出版社，2016， 30-32.

[8]王安瑾.一种基于 MinHash 的改进新闻文本聚类算法[J].计算机技术与发展，2019，29(02)：39-42.

[9]童昱强. 基于数据挖掘的网络新闻热点发现系统设计与实现[D].北京邮电大学,2019.

[10]何声欢. 面向社交媒体的中文词法分析研究[D].南京理工大学,2018.

[11]赵卉. 兰州市“微博问政”效果评价研究[D].兰州大学,2019.

[12]郑丽丝.微博时代政府公信力的挑战与机遇[J].法制与社会,2013(30):175-176.

附录

附件 1：热点问题表.xls（排名前 5 的热点问题）

附件 2：热点问题留言明细表.xls（相应热点问题对应的留言信息）

附件 3：python 源代码（包含 3 个问题的源程序代码）