基于 NLP 技术的智慧政务系统构建

摘要

随着互联网的普及，中国公民得以通过网络问政渠道来参与民主政治，但政府从该渠道收到的各类留言文本数据量庞大，这给人工处理文本数据的相关部门带来极大的挑战。从大量的文本数据中迅速的将留言进行分类，提取热点问题并给出有效的答复意见，对提高政府发现问题和解决问题的效率和质量，提升政府的管理水平有着重要的意义。

针对问题一，我们对数据进行文本分类，在数据去重后，利用 R 语言的 gsub 函数对数据进行进一步清洗，选择了基于中科院的 ictclas 中文分词算法做分词处理，接着过滤停用词和提取特征，其次使用 KNN、贝叶斯、SVM、随机森林、决策树、神经网络等算法分别构建文本分类器，而后选取 F-Score 作为评价指标，比较其大小来评估模型的优劣，最后选出分类效果最好的分类器，建立关于留言内容的一级标签分类模型。

针对问题二，我们将数据进行文本聚类后，对每个类进行主题词抽取。而后根据抽取的主题词来分析热度影响因素并对相关指标进行量化及标准化处理，接着基于层次分析法和熵值法提出两步定权法来确定指标权重，最后根据各评价指标及其权重构建综合评价体系。

针对问题三，我们首先对留言详情和答复意见分别进行特征提取和文本向量化，通过欧式距离计算答复意见与留言详情间的文本相似度来度量答复意见的相关性。其次提出对文本向量化后的答复意见进行完整性的判断可以利用答复意见中的文明用语、答复时间、解决方案等，此外，我们认为可以将答复时长作为衡量答复有效性的指标，最后利用熵值法确认指标权重，可以构建出评价模型，对答复质量进行评估。

关键词：文本分类　K-means 文本聚类　熵值法　文本相似度

目录

一、引言

随着互联网的发展，网络成为中国公民行使知情权、参与权、表达权和监督权的新渠道，依托互联网的大平台，网络问政风生水起，网络表达成为中国民主建设的新方式，一方面使得政府的信息更加公开透明，另一方面增进政民之间的沟通，促进政民关系。在网络问政渠道不断增加和便利性不断提高的情况下，政府通过网络问政平台收到的反馈和意见不计其数，涵盖了各类社情民意相关的留言。然而各类留言文本数据庞大，只依靠人工对公民留言进行分类、发现热点问题以及进行有效的答复将耗费大量的时间及人力，相关部门面临极大的挑战，存在着工作量大，效率低，且出错率高等问题。

在当今互联网时代，智慧政务系统融入云计算、大数据、人工智能等先进技术，可以让政府部门采集到更广泛的信息，处理信息的能力也变得更加强大，从而使政府整合资源的提供服务的效率得到大幅度提升。因此，我们根据文本挖掘方法，建立基于自然语言处理技术的相关模型，帮助相关政府部门对公民提出的问题进行快速有效的分析，并从大量的文本数据中迅速的发现热点问题，此外，对政府部门的答复意见建立评估方案，提高政府发现问题和解决问题的效率和质量，提升政府的管理水平。

二、问题分析

2.1 问题一的分析

针对问题一，对附件二的数据进行文本分类，在文本数据通过清洗、分词、去停用词、特征选择降维等处理后，使用不同的算法分别构建文本分类器，而后比较评价指标的大小来评估模型的优劣，选出分类效果最好的分类器，建立关于留言内容的一级标签分类模型。

2.2 问题二的分析

针对问题二，首先对附件三的数据进行文本聚类，通过词性标注、词频统计、文本压缩抽取主题词，将特定时间反映特定地点或者特定人群问题的留言进行归类，其次定义热度评价指标体系，分析热度影响因素，包括时间特征、聚集特征、内容特征、关注特征，接着对相关指标进行量化及标准化处理，最后利用层次分析法和熵值法确认指标权重，构建模型确认评价结果。

2.3 问题三的分析

针对问题三，对附件四的数据进行数据预处理后，分别对留言详情和答复意见进行特征提取和文本向量化，而后计算答复意见与留言详情间的文本相似度，作为衡量答复意见相关性的指标。接着对文本向量化后的答复意见进行结构完整性的判断，其中，结构完整包括文明用语、答复时间、解决方案等，此外，将答复时长作为衡量答复及时性的指标，最后利用熵值法方法确认各个指标的权重，构建出评价模型，对答复质量进行评估。

三、数据预处理

3.1 基本步骤图

图 3.1　数据预处理步骤

以问题一为例：

3.2 数据清洗

由于文本数据常常存在随意性大、数据混杂、不规范性等问题，因此我们需要在文本分析前对数据进行清洗工作。

（1）缺失、重复数据的处理

若数据空缺或无效，将其视为缺失值；若留言内容完全相同，则将其视为重复数据。在对本次 9210 条数据进行审核时，我们发现未存在缺失数据的情况，但存在重复数据 158 条，因此我们利用 R 语言进行数据清洗，最终得到有效数据共 9052 条。

（2）标点符号、特殊符号等无用信息的删除

考虑到中文文本中标点符号、空格等信息对于后续文本分析没有任何帮助，并且可能降低分词工作效率，因此我们利用 R 语言的 gsub 函数对数据进行进一步清洗。

3.3 文本分词

词汇处理是自然语言处理的基础。汉语的书写以汉字作为基本单位，词与词之间没有明显的形态界限，要进行汉语的计算机处理，首先必须将汉语的词与词分割开，即分词。因此，在文本挖掘过程中，首先应对文本进行分词处理，即将单个文本表示为词向量的过程，分词结果的准确性对后续文本挖掘算法有着不可忽视的影响。本文选择了基于中科院的 ictclas 中文分词算法，对文档中的留言详情进行分词处理，并进一步进行词性标注等工作，为后续文本分析做准备。使用 R 语言进行分词处理的部分结果如下：

> traindata[6280,]$留言详情
[1]“我是 M2 县印塘乡的一名普通村民朱灿，2019 年 10 月 10 日，发现我儿子朱诺杭(3 岁)误服了几粒鱼肝油，送去 M2 县人民医院进行治疗。在医生间内部意见没有统一的情况下，我们听从一名医生要求同意进行洗胃操作，但洗胃过程中，儿子抢救无效死亡。医院开始并不承认有任何责任，西地省楚雅司法鉴定中心在 2019 年 11 月 28 日出具了尸检报告，显示我儿子就是因为洗胃过程中，发生了肺气肿、支气管痉挛导致呼吸障碍死亡，就是因为洗胃操作不当(很有可能是将胃管插入了气管中)导致的死亡，死亡原因证实后，我们多次到 M2 县卫生局、M2 县人民医院要求给个说法，但 M2 县人民医院直至今日都未进行正式的协商与回应，作为一个普通的群众，作为一个亡儿的父亲，深深的感受到了医院之强势，与维权的不易!恳请领导督促相关部门予以积极解决，早日让我及家人从痛苦中走出来，感激不尽!”

图 3.2　留言详情

>segmentCN(traindata[6280,]$留言详情)									
[1]	“我”	“是”	“M”	“2”	“县”	“印”	“塘”	“乡”	“的”	“一名”
[11]	“普通”	“村民”	“朱”	“灿”	“2019 年”	“10 月”	“10 日”	“发现”	“我”	“儿子”
[21]	“朱”	“诺”	“杭”	“3 岁”	“误”	“服”	“了”	“几粒”	“鱼肝油”	“送”
[31]	“去”	“M”	“2”	“县”	“人民”	“医院”	“进行”	“治疗”	“在”	“医生”
[41]	“间”	“内部”	“意见”	“没有”	“统一”	“的”	“情况”	“下”	“我们”	“听从”
[51]	“一名”	“医生”	“要求”	“同意”	“进行”	“洗胃”	“操作”	“但”	“洗胃”	“过程”
[61]	“中”	“儿子”	“抢救”	“无效”	“死亡”	“医院”	“开始”	“并”	“不”	“承认”
[71]	“有”	“任何”	“责任”	“西”	“地”	“省”	“楚”	“雅”	“司法”	“鉴定”
[81]	“中心”	“在”	“2019 年”	“11 月”	“28 日”	“出具”	“了”	“尸”	“检”	“报告”
[91]	“显示”	“我”	“儿子”	“就是”	“因为”	“洗胃”	“过程”	“中”	“发生”	“了”
[101]	“肺气肿”	“支气管”	“痉挛”	“导致”	“呼吸”	“障碍”	“死亡”	“就是”	“因为”	“洗胃”
[111]	“操作”	“不当”	“很”	“有”	“可能”	“是”	“将”	“胃”	“管”	“插入”
[121]	“了”	“气管”	“中”	“导致”	“的”	“死亡”	“死亡”	“原因”	“证实”	“后”
[131]	“我们”	“多次”	“到”	“M”	“2”	“县”	“卫生局”	“M”	“2”	“县”
[141]	“人民”	“医院”	“要求”	“给”	“个”	“说法”	“但”	“M”	“2”	“县”
[151]	“人民”	“医院”	“直至”	“今日”	“都”	“未”	“进行”	“正式”	“的”	“协商”
[161]	“与”	“回应”	“作为”	“一个”	“普通”	“的”	“群众”	“作为”	“一个”	“亡”
[171]	“儿”	“的”	“父亲”	“深深”	“的”	“感受”	“到”	“了”	“医院”	“之”
[181]	“强势”	“与”	“维权”	“的”	“不易”	“恳请”	“领导”	“督促”	“相关”	“部门”
[191]	“予以”	“积极”	“解决”	“早日”	“让”	“我”	“及”	“家人”	“从”	“痛苦”
[201]	“中”	“走”	“出来”	“感激不尽”						

图 3.3　文本分词结果

由图 3.3 可以看出，在使用内置分词词典对文本数据进行初步分词时，文本中的词基本能被区分开来，但是仍然存在一些如“印塘乡”、“支气管痉挛”“人民医院”这样的专属名词和地名等词语不能很好的识别出来，为解决这个问题，我们根据留言所涉及的领域，导入搜狗词库中医学类、交通类、行政区地名等相关词典，并通过人工标注等方法对分词结果进行进一步优化，优化结果见下（图 3.4）：

>segmentCN(traindata[6280,]$留言详情)								
[1] “我”	“是”	“M2 县”	“印塘乡”	“的”	“一名”	“普通”	“村民”	“朱”
[10] “灿”	“2019 年”	“10 月”	“10 日”	“发现”	“我”	“儿子”	“朱”	“诺”
[19] “杭”	“3 岁”	“误”	“服”	“了”	“几粒”	“鱼肝油”	“送”	“去”
[28]“M2 县”	“人民医院”	“进行”	“治疗”	“在”	“医生”	“间”	“内部”	“意见”
[37]“没有”	“统一”	“的”	“情况”	“下”	“我们”	“听从”	“一名”	“医生”
[46] “要求”	“同意”	“进行”	“洗胃”	“操作”	“但”	“洗胃”	“过程”	“中”
[55]“儿子”	“抢救”	“无效”	“死亡”	“医院”	“开始”	“并”	“不”	“承认”
[64]“有”	“任何”	“责任”	“西地省”	“楚”	“雅”	“司法”	“鉴定”	“中心”
[73] “在”	“2019 年”	“11 月”	“28 日”	“出具”	“了”	“尸”	“检”	“报告”
[82]“显示”	“我”	“儿子”	“就是”	“因为”	“洗胃”	“过程”	“中”	“发生”
[91] “了”	“肺气肿”	“支气管痉挛”	“导致”	“呼吸”	“障碍”	“死亡”	“就是”	“因为”
[100] “洗胃”	“操作”	“不当”	“很”	“有”	“可能”	“是”	“将”	“胃”
[109]“管”	“插入”	“了”	“气管”	“中”	“导致”	“的”	“死亡”	“死亡”
[118]“原因”	“证实”	“后”	“我们”	“多次”	“到”	“M”	“2”	“县卫生局”
[127]“M2 县”	“人民医院”	“要求”	“给”	“个”	“说法”	“但”	“M2 县”	“人民医院”
[136]“直至”	“今日”	“都”	“未”	“进行”	“正式”	“的”	“协商”	“与”
[145]“回应”	“作为”	“一个”	“普通”	“的”	“群众”	“作为”	“一个”	“亡”
[154]“儿”	“的”	“父亲”	“深深”	“的”	“感受”	“到了”	“医院”	“之”
[163]“强势”	“与”	“维权”	“的”	“不易”	“恳请”	“领导”	“督促”	“相关”
[172] “部门”	“予以”	“积极”	“解决”	“早日”	“让”	“我”	“及”	“家人”
[181]“从”	“痛苦”	“中”	“走”	“出来”	“感激不尽”			

图 3.4　优化结果

经过测试，最终分词的准确率达到 94.6%，可以很好的满足后续文本分析要求。

3.4 停用词过滤

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉的某些字或词，分词处理后的词条中含有大量的停用词，这些停用词不仅携带的文本信息量较少，而且还对其他实词起到一定的抑制作用，降低了分类系统的处理效率和准确度，因此，文本预处理过程有必要将所有的停用词过滤。本文根据“哈工大停用词词库”、“四川大学机器学习智能实验室停用词库”、百度停用词表“等各种停用词表，整理去重出了一个含 1622 个停用词的词表，并据此词表对上一步的分词结果进行停用词过

滤处理。

>Train_CN2[6280]							
[1]“M2 县”	“印塘乡”	“一名”	“村民”	“朱”	“灿”	“2019 年”	“10 月”
[9] “10 日”	“发现”	“儿子”	“朱”	“诺”	“杭”	“3 岁”	“误”
[17] “服”	“几粒”	“鱼肝油”	“送”	“M2 县”	“人民医院”	“治疗”	“医生”
[25] “间”	“内部”	“意见”	“统一”	“情况”	“听从”	“一名”	“医生”
[33] “同意”	“洗胃”	“操作”	“洗胃”	“过程”	“中”	“儿子”	“抢救”
[41]“无效”	“死亡”	“医院”	“承认”	“责任”	“西地省”	“楚”	“雅”
[49]“司法”	“鉴定”	“中心”	“2019 年”	“11 月”	“28 日”	“出具”	“尸”
[57]“检”	“报告”	“显示”	“儿子”	“洗胃”	“过程”	“中”	“发生”
[65] “肺气肿”	“支气管痉挛”	“导致”	“呼吸”	“障碍”	“死亡”	“洗胃”	“操作”
[73] “不当”	“胃”	“插入”	“气管”	“中”	“导致”	“死亡”	“死亡”
[81]“原因”	“证实”	“M”	“2”	“县卫生局”	“M2 县”	“人民医院”	“说法”
[89]“M2 县”	“人民医院”	“直至”	“今日”	“未”	“正式”	“协商”	“回应”
[97] “群众”	“亡”	“父亲”	“深深”	“感受”	“到了”	“医院”	“强势”
[105]“维权”	“不易”	“恳请”	“领导”	“督促”	“相关”	“部门”	“予以”
[113] “解决”	“早日”	“家人”	“痛苦”	“中”	“走”	“感激不尽”	

图 3.5　停用词过滤结果

由图 3.5 可以看出，“的”、“直至”、“因为”等不含有效信息的词语很好的从分词结果中被剔除出去，保证了分词的准确性和有效性。

四、问题求解

4.1 问题一——文本分类

4.1.1 基本流程图

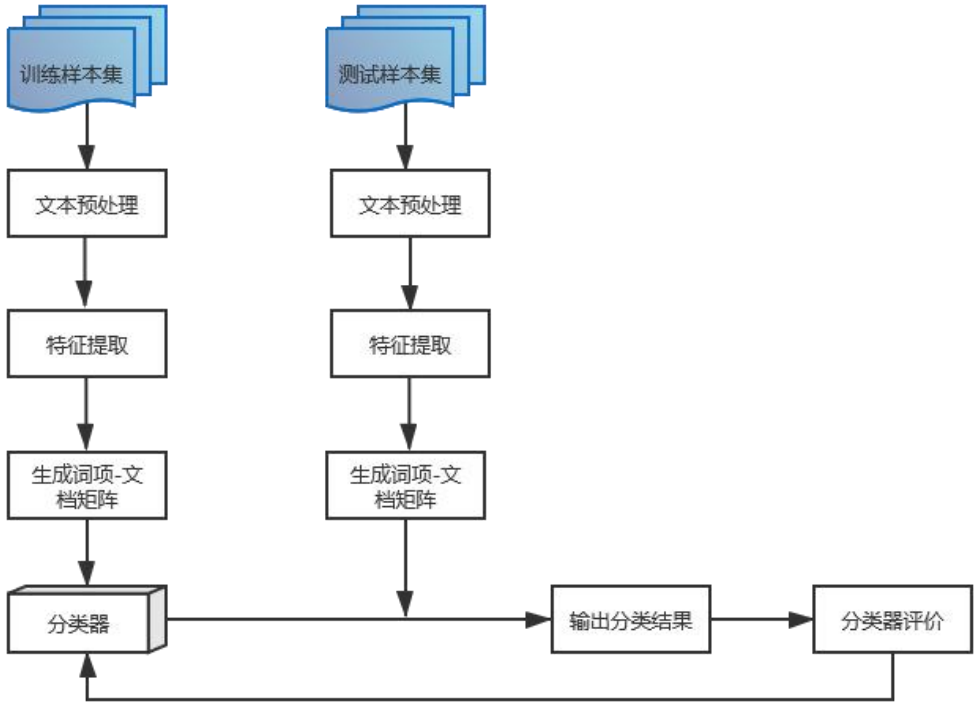

图 4.1　基本流程图

4.1.2 TF-IDF 特征提取

在使用向量空间模型来描述文本向量时，如果直接用分词算法和词频统计方法得到的特征项来表示文本向量中的各个维，所得到的向量维度将非常庞大。这种未经处理的文本矢量不仅给后续工作带来巨大的计算开销，使整个处理过程的效率非常低下，而且会损害分类、聚类算法的精确性，从而使所得到的结果很难令人满意。因此，必须对文本向量做进一步净化处理，在保证原文含义的基础上，找出对文本特征类别最具代表性的文本特征。为了解决这个问题，最有效的办法就是通过特征选择来降维。常用的特征选择方法主要有词频法，文档频次法、信息增益法、期望交叉熵、χ^2统计等。本文使用的 TF-IDF 算法是文本挖掘中广泛使用的一种特征向量化方法，它反映了一个词在语料库中对文档的重要性。

$$TFIDF(t,d,D) = TF(t,d) \cdot log\frac{|D|+1}{DF(t,D)+1} \qquad 式（4.1）$$

其中，词频$TF(t,d)$表示词语 t 在文档 d 中出现的次数，文档频率$DF(t,D)$是包含词语 t 的文档数量，$|D|$表示语料库中文档的总数。

使用此算法进行特征提取，最终得到维度为 9052*102943 的特征矩阵，为了降低维度以提高分类效率，我们引入稀疏度作为我们降维的工具，稀疏度表示一个名词在文档中的相对频率

$$S = (1 - \alpha) \cdot dim(x) \qquad \text{式（4.2）}$$

其中α为稀疏度阈值，$dim(x)$为特征矩阵 x 的行维度。

我们初步将稀疏度阈值设为 0.99，即(1- 0.99)*9052= 181, 因此，对于任何一个词语，如果包含它的文档少于 181 个，它就会被删除，据此规则我们最终得到维度为 9052*704 的特征矩阵，可以看到，通过这种方式，特征矩阵的维度有了显著降低，对于后续分类效率的提升有着很大帮助。

4.1.3 分类模型选取

特征矩阵提取后，我们便可以进行文本分类器的构建工作，目前，在处理分类问题时，常用的训练算法有 KNN、贝叶斯、SVM、随机森林、决策树以及神经网络等，为了选择最适合的模型，我们随机选取 70%的样本作为训练集，其余 30%作为预测集，使用上述算法分别构建了分类器，由于问题一是一个多分类问题，所以我们首先选择适用于多分类问题的 KNN 算法，该算法的核心思想是：在给定新文本后, 在训练文本找到与该新文本距离最近(最相似)的 K 篇文本, 其中 K 通常是不大于 20 的整数，根据这 K 篇文本大多所属的类别判定新文本所属的类别，简单举例说明：

如下图（图 4.2）所示，绿色圆圈代表一个新的待分类文本，而红色三角形和蓝色正方形代表训练文本集中不同的类别。此时绿色圆圈要被分到红色三角形还是蓝色正方形？当 K=3 时，距离绿色圆圈最近的 K 个样本中，红色三角形比例为 2/3，绿色圆圈将被分到红色三角形那个类别，而 K=5 时，由于蓝色正方形比例为 3/5，绿色圆圈被分到蓝色正方形所在的类别。

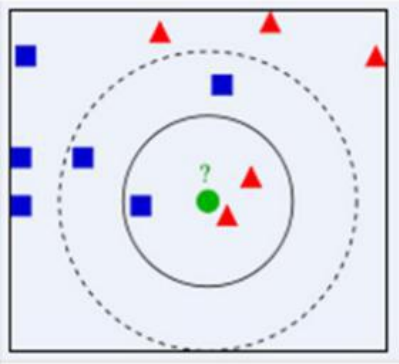

图 4.2

其中，文本间的距离计算方法有：欧氏距离法和夹角余弦法。而对于文本分类问题，用夹角余弦法来度量文本间的相似度更合适。

具体算法步骤如下：

（1）对于一个测试文本，根据特征词形成测试文本向量 x。

（2）计算该测试文本与训练集中每一个文本的距离，即文本相似度。我们在文本相似度算法中选择了更适合 KNN 的夹角余弦法，计算公式为

$$Sim\left(x,d_j\right)=\frac{\sum_{k=1}^{M}W_k\times W_{jk}}{\sqrt{\left(\sum_{k=1}^{M}W_k^2\right)\left(\sum_{k=1}^{M}W_{jk}^2\right)}} \qquad \text{式（4.3）}$$

其中，d_j为训练集中的某个文本 j，W_k和W_{jk}分别表示两个文本向量 x 和d_j中第 k 个特征项的权重值，M 表示特征向量的维数。

（3）按照文本相似度，在训练文本集中选出与测试文本最相似的 K 个文本，其中最佳的 K 值只能通过取不同的 K 值来测试，最后根据测试分类效果来确定合适的 K 值。

（4）在测试文本的 K 个近邻中，依次计算每类的权重，计算公式为

$$p\left(x,C_j\right)=\sum_{d_i\in KNN}Sim(x,d_i)\times y(d_i,C_j) \qquad \text{式（4.4）}$$

其中，$y(d_i,C_j)$为类别属性函数，如果d_i属于类C_j，则函数值为 1，反之为 0。

（5）比较类的权重，将文本分到权重最大的那个类别中。

据此算法，我们使用 6301 条训练集数据对 KNN 分类器进行训练，并进一步使用此分类器对剩余的 2751 条测试集进行预测，预测结果如下（表 4.1）：

表 4.1　KNN 算法预测结果

真实值	预测值						
	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生
城乡建设	262	83	39	51	39	96	31
环境保护	34	161	13	14	7	27	10
交通运输	45	16	58	15	10	35	10
教育文体	25	13	6	363	26	43	11
劳动和社会保障	47	17	18	37	407	44	46
商贸旅游	59	15	27	41	18	137	26
卫生计生	28	10	4	23	19	30	155

计算得到，预测结果中共有 1935 条数据预测正确。

为了对模型进行评价，我们选取了准确率（Accuracy）和 F-Score 作为评价指标。准确率表示所有预测正确的样本数占总样本数的比重。但当感兴趣的主类是稀少的，即数据集分布反映负类显著占多数，而正类占少数，数据类显然是不平衡的。例如：一个地区有 10 万人，其中 10 人患有白血病，我们感兴趣的类是“患病”（正类），其出现远不及负类“不患病”频繁。如果一个预测人们是否患病的模型把所有人都归为“不患病”，其准确率达到 99.99%，这显然不是我们想要的结果。

$$\text{Accuracy}=\frac{\text{所有预测正确的样本数}}{\text{总样本数}} \qquad \text{式（4.5）}$$

因此，为了更好地评估模型，我们还使用了 F-Score 来评价模型的好坏。F-Score 是精度（Precision，也称为查准率）和召回率（Recall，也称为查全率）的调和函数。

$$F=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \qquad \text{式（4.6）}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

可以算出，KNN 算法的准确率为 0.70，F-Score 为 0.68。同理，我们使用其他 5 种算法分别构建分类器，由于其分类思路大同小异，本文不再进行赘述。下面给出各分类器的分类结果：

表 4.2　各分类算法的准确率和 F-Score 值

分类算法	预测集数目	预测正确数	准确率	F-Score 值
KNN	2751	1935	0.70	0.68
SVM	2751	2479	0.90	0.88
贝叶斯	2751	1392	0.51	0.48
决策树	2751	1598	0.59	0.54
随机森林	2751	2345	0.85	0.83
神经网络	2751	2021	0.73	0.74

由表 4.2 可以看出，基于 SVM 和随机森林这两种算法得到的分类模型效果较好，准确率和 F-Score 值均达到 80%以上，可以认为很好的完成了分类目标。

4.1.4 模型优化

在进行多次实验后，我们发现特征维度对于模型准确率有着很重要的影响，为此，我们设定了一系列稀疏度阈值α，得出 SVM 的α~F-Score 曲线图如下：

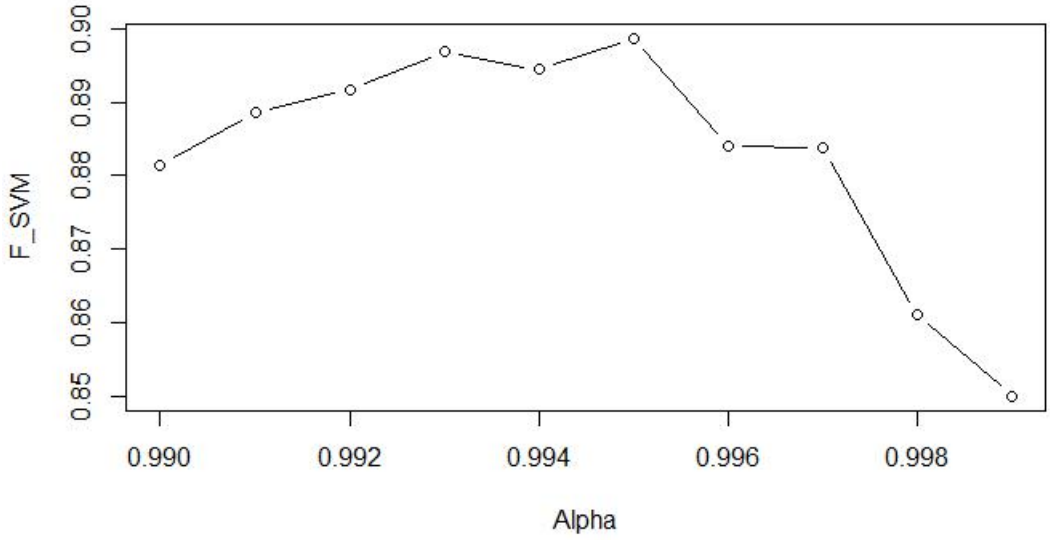

图 4.3　SVM 的α~F-Score 曲线图

由图 4.3 可以看出，F-Score 值在一定范围内随着稀疏度阈值α增大而增加，但在α达到某一点后便逐渐趋于稳定接着出现下降趋势，SVM 的 F-Score 在α_{svm}=0.995 处达到最大，最大值为 0.898，这是因为文本权重小的特征词对分类有干扰作用，特征词的抽取并不是越多越好，每一类需要的最合适特征项数目，需要利用大量的语料进行试验才能确定。

4.1.5 分类评价

根据上述结果，我们最终选取 SVM 作为我们的分类器算法，优化模型的输出结果如下：

表 4.3　SVM 算法输出结果

标签类别	预测集数目	预测结果	预测正确数	准确率	召回率	F-Score 值
城乡建设	601	852	556	0.65	0.93	0.77
环境保护	266	234	215	0.92	0.81	0.86
交通运输	189	80	75	0.94	0.40	0.56
教育文体	487	433	413	0.95	0.85	0.90
劳动和社会保障	616	648	569	0.88	0.92	0.90

续表 4.3　SVM 算法输出结果

商务旅游	323	306	235	0.77	0.73	0.75
卫生计生	269	198	186	0.94	0.69	0.80

由表 4.3 可以看出，该模型具有较高的标签分类性能，对提升政府的管理水平和施政效率具有极大的推动作用。

4.2 问题二——热点问题挖掘

4.2.1 文本聚类

4.2.1.1 基于 K-means 算法的文本聚类

由于不同人的表述方式存在不同，同一问题的的文本形式存在较大差异，通过文本聚类我们可以对同主题文本进行冗余消除、信息融合、文本生成等处理，从而生成一篇简明扼要的摘要文档，对于热点问题的挖掘整合具有重要作用。本文采用基于 K-Means 算法实现文本聚类，K-Means 是常用的聚类算法，与其他聚类算法相比，其时间复杂度低，且聚类的效果良好，其基本步骤如下:

（1）从 n 个数据对象任意选择 K 个对象作为初始聚类中心。

（2）根据每个聚类对象的均值（中心点），计算每个对象与这些中心点的相似度（距离），并根据最小距离重新对相应对象进行划分，在这里，我们采用欧氏距离作为相似度的度量。

（3）重新计算每个（有变化）聚类的均值（中心对象）。

（4）计算准则函数，当满足一定条件，如函数收敛时，则算法终止，如果条件不满足则回到步骤（2），准则函数定义为最小化数据对象到其簇中心的距离的平方和，即 $\min \sum_{i}^{k} \sum \operatorname{dist}\left(c_i, x\right)^2$，其中，k 是簇的个数，$C_i$ 是第 i 个簇的中心点，dist(C_i,x)为 x 到 C_i 的距离。

由于 K-means 算法需要事先指定中心点的个数 K，且 K 值的选取对于最后的聚类结果有着很大影响，为了得到最合适的 K 值，我们按递增的顺序尝试不同的 K 值，同时画出其对应的误差值，通过寻求拐点来确定 K 值。图 4.4 是中心点的个数从 2 到 200 对应的误差值的曲线：

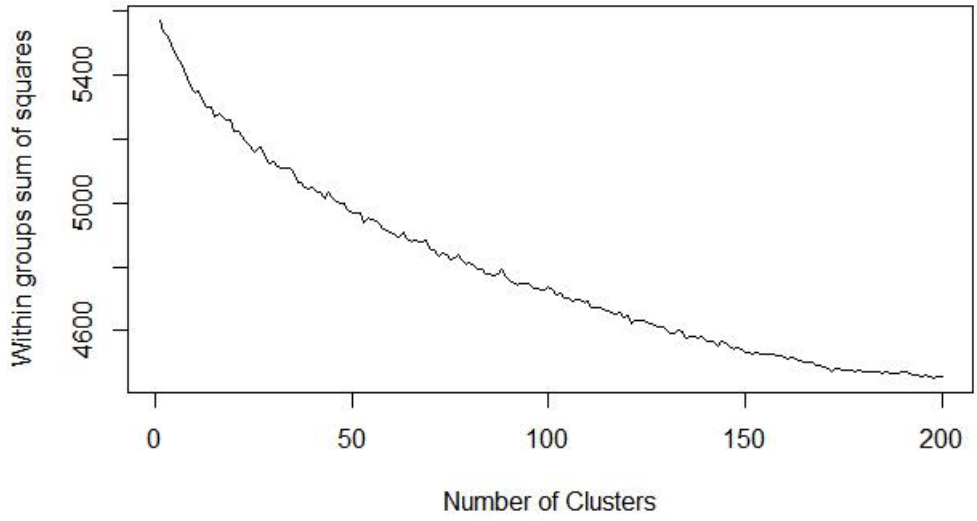

图 4.4　误差值曲线

可以看出，曲线在 K=150 附近的时候出现拐点，由此确定的簇个数为 150 个，确定 K 值后，我们利用 R 语言对文本特征矩阵进行迭代处理，得到各个簇文档个数结果如下可以看出，曲线在 K=150 附近的时候出现拐点，由此确定的簇个数为 150 个，确定 K 值后，我们利用 R 语言对文本特征矩阵进行迭代处理，得到各文档所对应的簇标签,为了简化后续工作，我们从中选取文档个数排名前 10 的簇并得到其对应文档个数如下：

表 4.4

簇标签	簇 1	簇 2	簇 3	簇 4	簇 5	簇 6	簇 7	簇 8	簇 9	簇 10
文档个数	1896	366	317	304	271	266	231	229	175	157

4.2.1.2 主题词抽取

聚类完成后，我们需要通过一些标签来描述各簇信息，以此得到热点问题的具体描述，为此我们对每个簇分别进行了主题词抽取，主题词的抽取工作主要分为词性标注和命名实体识别两部分。

（1）词性标注

词性标注就是在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程。一个完整的问题描述由四个基本要素构成：时间、地点、人物和事件。通过词性标注，我们可以得到地名、人名、动词、时间词等多个词性标签，这对于我们对各类簇信息进行整合具有重要意义。

在进行词性标注时，我们引用了《PFR 人民日报标注语料库》作为标注训练材料，该语料库除 26 个基本词类标记外，还从语料库应用的角度及语言学角度

增加了专有名词等标记，可以很好的满足本文所需词性标签的要求。利用 jiebaR 包中的 vector_tag 函数完成词性标注任务后的部分结果如表 4.5 所示：

表 4.5　词性标注结果

id	word	nature
28	业主	n
28	信访	v
28	世景	x
28	小学	n
28	配套	a
28	垃圾站	n
28	建在	v
28	南	ns
28	门口	s
28	距离	n
28	业主	n
28	家门	n
28	5	x
28	米	q

表 4.6　词频统计部分结果

	Var1	Freq
1	垃圾	270
2	生活	85
3	小区	80
4	A市	77
5	业主	71
6	居民	70
7	城市	67
8	领导	65
9	焚烧	61
10	环境	57
11	影响	56
12	工程	50
13	路	49
14	社区	48

（2）词频统计

为了从各簇文本中提取出信息量大的主题词，以此完成冗余消除和信息融合等工作，得到最终的热点问题描述，我们采取词频作为文本信息量的度量方式，对各簇的文本分别进行词频统计，按词频大小降序排列得到（簇 1）部分结果如表 4.6:

在完成主题词抽取工作后，我们便可以对各簇文本进行信息整合完成热点问题的描述工作，首先，根据词性标注结果，我们得到了问题主体及发生所在地，进一步根据词频统计结果结合留言完整内容完成事件描述。以簇 1 为例，词性标签中，ns 表示地名，nr 表示人名，nt 为组织机构名，通过检索这三个标签我们可以得到簇 1 中含有的地名，人名，组织机构名如下表：

表 4. 7

ns	nr	nt
滨江	胡书记	城管局
梅家田	居民	建设部
大桥	村民	镇政府

同时我们观察簇 1 的词频统计结果，可以看到“垃圾”、“焚烧”、“生活”词频最高，结合完整留言内容我们可以得出簇 1 对应的热点问题为“A 市 A6 区梅家田社区垃圾处理费收取不公平”。

4.2.2 热度评价指标体系设计

4.2.2.1 热度影响要素分析

（1）时间特征

某一热点问题的时间特征为问题持续的时间区间，反映了这一问题亟待解决的急迫性，能否及时发现并解决问题反映了政府的信息管理能力，对于加强城乡建设具有重要意义，问题存在时间越长，政府管理解决的必要性越大，因此，时间特征的统计度量对于挖掘发现热点问题是必不可少的。

（2）聚集特征

不同的地点和人群类别，问题的发生频率具有较大的差异性，发生频率反映了某一地点或人群问题易发程度，掌握热点问题的聚集特征对于提高问题的处理效率，对问题频发地区或人群进行针对性整改具有很大帮助。

（3）内容特征

本文将内容特征分为三部分：情感极性、问题严重程度、问题处理难度。情感极性表示了文本所表达的一种情感倾向程度，通过分析留言的情感极性，我们可以得到留言者对于问题的不满程度及诉求强烈程度，而问题严重程度和处理难度也是政府在处理问题前所必需考虑的两个方面。

（4）关注特征

点赞率作为关注特征的基本指标，反映了人民群众对于留言问题的态度和看法，对于政府处理问题的效率提升具有很大的推动力，若同一问题的点赞率越高，说明民众对处理问题的诉求越强烈，政府处理问题的必要性也越大。

综合对上述热度影响要素的分析，我们初步完成了热度评价指标体系的建立，如表 4.8：

表 4.8　热度综合评价指标体系

一级指标	二级指标	指标内容
时间特征热度影响力	时间范围	问题持续反映的时间差
聚集特征热度影响力	地点集中率	地点的问题发生频率
	人群集中率	人群的问题发生频率
内容特征热度影响力	情感极性	留言文本的情感极性得分

续表 4.8　热度综合评价指标体系

内容特征热度影响力	问题严重程度	问题给社会带来负面影响程度
	问题处理难度	处理问题所需花费成本大小
关注特征热度影响力	点赞率	问题的点赞率

4.2.2.2 相关指标度量

为了更好的利用各指标信息完成模型构建，我们需要对指标内容进行量化及标准化处理，结合留言文本中的内容，我们对相关指标度量的具体工作如下：

（1）时间范围度量

由于二级指标中的时间范围是文本数据，无法直接运算，因此在对时间范围进行量化处理时，我们首先需要将日期转化为数字，为此，我们定义了这样一个转化规则：以 2017/6/8 为基数，并将这一天定义为整数 1，之后每过一天就加 1，以此类推。例如，对于日期 2019/10/29，相对于基数而言增加了 873 天，据此规则得到的数值为 874，进一步以 2020/1/9 作为其上限，对数据进行 0-1 标准化处理，得到最终分数为 0.92，由此时间范围就变成了可度量数据。

（2）地点及人群集中率度量

对于地点和人群集中率，我们使用问题发生频率作为其度量的依据，即某一地点或人群出现在全部留言文本中的频率，由于同一类问题通常出现在某一特点地点，因此统计时我们不考虑地市县这类行政区地名，同时为避免基数过大导致总体频率过低，在进行比较时存在困难，我们引入调整系数α对最终结果进行放大处理，在这里，我们将α定为 100，如“春花镇”这个地名共在 13 条文本数据中出现，对应的文本频率为 1000×13/9052=0.14。

（3）情感极性度量

情感极性分析是对带有感情色彩的主观性文本进行分析、处理、归纳和推理的过程，前常见的情感极性分析方法主要是两种：基于情感词典的方法和基于机器学习的方法。本文采取的方法为基于词典的方法，词典来源于 BosonNLP 数据下载的情感词典，其基本原理是根据词典计算每条文本中的正负词数，并进行分数计算得出该文本的情感得分。

（4）问题严重程度及处理难度度量

对于严重程度的度量，我们参考了预警级别将等级划分为四类：I 级（特别严重）、II 级（严重）、III 级（较重）、IV 级（一般），并对应赋予 4 至 1 分

的分数，同样的，我们对处理难度也进行了相应等级划分和赋值，由此完成了这两个指标的量化度量。

4.2.3 综合评价模型

4.2.3.1 指标权重确定

我们希望根据各指标的信息，构建一个综合评价模型计算出每个问题的综合分数，以更好地对热点问题进行热度排序。在前文的叙述，我们已经构造出组成综合评价体系的各个指标，包括 4 个一级指标和 7 个二级指标，下一步我们需要给各个指标赋予相应的权重，以反映该评价指标对整体的贡献程度。为了减少人为因素的偏差和避免客观评价对指标重要程度的忽视，我们团队基于层次分析法和熵值法提出两步定权法，即先运用层次分析法确定二级指标权重，再运用熵值法确定一级指标的权重，以期获得更加准确的权重分配。

（1）层次分析法

以内容特征为例，首先我们通过背对背小组打分的方式得到情感极性、问题严重程度以及问题处理难度三者之间的两两比较关系，并将得到的数据构成如下的判断矩阵。

表 4.9　判断矩阵

	情感极性	问题处理难度	问题严重程度
情感极性	1	1/4	1/3
问题处理难度	4	1	1/2
问题严重程度	3	2	1

其中，矩阵元素 a_{ij} 表示第 i 个因素相对于第 j 个元素的重要程度。

进一步我们通过公式 1 对判断矩阵的各行向量进行几何平均和归一化处理，得到各指标权重向量 W

$$w_i = \bar{w}_i \ / \ \sum^{n} \bar{w}_i \qquad \text{式（4.7）}$$

W{情感极性，问题处理难度，问题严重程度} = {0.12, 0.36, 0.52}

同理，我们可以得到聚集特征中各指标权重如下：

W{地点集中率，人群集中率}={0.46，0.54}

（2）熵值法

熵值法是一种客观赋权法，其原理是根据各项指标观测值的离散程度大小判断该指标对综合评价的影响，离散程度越大，熵值越大，所占的权重越小。由于四个特征热度影响力均为正向指标，我们首先采用以下公式对数据进行标准化处理

$$x'_{ij} = \frac{x_{ij} - \min\left\{x_{ij}, \cdots, x_{nj}\right\}}{\max\left\{x_{1j}, \cdots, x_{nj}\right\} - \min\left\{x_{1j}, \cdots, x_{nj}\right\}} \quad \text{式（4.8）}$$

接着计算第 j 项指标下第 i 条观测占该指标的比重：

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} (i = 1, \cdots, n, j = 1, \cdots, m) \quad \text{式（4.9）}$$

进一步计算第 j 个指标的熵值

$$e_j = \begin{cases} -\frac{1}{\ln(n)} \sum_{i=1}^{n} P_{ij} \ln\left(P_{ij}\right) & \text{如果 } P_{ij} \neq 0 \\ 0 & \text{如果 } P_{ij} = 0 \end{cases} \quad \text{式（4.10）}$$

最终得到第 j 个指标的权重

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \quad \text{式（4.11）}$$

其中，$d_j = 1 - e_j$，表示信息熵冗余度。

由此，我们计算出四个一级指标权重如下：

W{时间特征，聚集特征，内容特征，关注特征}={0.23，0.18，0.32，0.27}

4.2.3.2 模型构建

以上步骤我们完成了综合评价体系所包含的各评价指标对应的权重计算，即我们所说的贡献程度，将其代入综合分计算公式，即可计算出每类问题的综合分数，此分数即为该问题的热度指数。

S=0.23*时间范围+0.18*（0.46*地点集中率+0.54*人群集中率）+0.32*（0.12*情感极性+0.36*问题处理难度+0.52*问题严重程度）+0.27*点赞率

以簇 1 对应的热点问题为例，其各指标的得分如表 4.10：

表 4.10　热度指标得分

二级指标	指标得分
时间范围	0.94
地点集中率	0.34

续表 4.10　热度指标得分

人群集中率	0.25
情感极性	0.8
问题严重程度	3
问题处理难度	2
点赞率	0.98

由此，我们可以计算得到其热度指数为：

S=0.23*0.94+0.18*（0.46*0.34+0.54*0.25）+0.32*（0.12*0.8+0.36*3+0.52*2）+0.27*0.98=1.24

4.3 问题三——评价模型的构建

4.3.1 答复相关性

在度量相关性的指标中，我们选取文本相似度来衡量相关性。在计算文本相似度上有三个方法，一是基于关键词匹配的传统方法；二是将文本映射到向量空间，再结合余弦相似度、欧式距离等方法；三是深度学习的方法，如基于用户点击数据的深度学习语义匹配模型 DSSM，以及目前 state-of-art 的 Siamese LSTM 等方法。通过 TF-IDF 提取出文本的特征项矩阵，结合相似度计算方法，可以得出两个文本间的相似度，其中采用 TF-IDF 的前提是文本的词语重要性与词语在文章中出现的位置无关。

相似度就是比较两个事物的相似性。一般可以通过计算事物的特征之间的距离来度量，如果距离小，说明相似度大；如果距离大，则说明相似度小。相似度的计算方法有欧氏距离、切比雪夫距离、马氏距离、夹角余弦、汉明距离等等。

我们采用欧式距离进行计算，下图为部分计算结果：

	V1
1102	0.7181403
1363	0.6301153
1173	0.6212607
242	0.6208764
1074	0.6005730
1366	0.6000277
2089	0.5858136
1899	0.5846681
1861	0.5451809
1352	0.5264700
1446	0.5224800
2656	0.5070226
2209	0.4797530
1413	0.4673121

	V1
768	0.02384366
421	0.03078408
79	0.03992461
389	0.04172910
1125	0.04257808
522	0.04292280
168	0.04445372
455	0.04470614
1897	0.04787538
2266	0.04823204
1129	0.04945661
428	0.05138559

图 4.5

由图 4.5 可以看出，第 1102 条答复与其留言的距离较远，相似度较弱，说明该条留言相关性较差；而第 768 条答复与其留言的距离较近，相似度较高，说明该条留言相关性较强。

4.3.2 答复完整性及有效性

我们认为对文本向量化后的答复意见进行结构完整性的判断，可以通过答复意见中的文明用语、答复时间、解决方案等方面来构建指标，结构越完整说明该条答复质量较高。此外，我们还将答复时长作为衡量答复及时性的指标，答复时长越长，答复的有效性也随之降低。

五、总结

本文的主要目的在于利用自然语言处理技术和文本挖掘算法建立能对群众留言进行分类，提取热点问题，以及评估相关部门对群众留言答复意见的质量的模型。

我们对数据进行文本分类，在数据预处理后，使用 KNN、贝叶斯、SVM、随机森林、决策树、神经网络等算法分别构建文本分类器，之后比较 F-Score 的大小来评估模型的优劣，最终我们得出分类效果最好的是基于 SVM 的分类器算法，因此建立关于留言内容的一级标签分类模型。然后我们对留言情况进行文本聚类，通过词性标注、词频统计、文本压缩抽取主题词，将特定时间反映特定地点或者特定人群问题的留言进行归类，其次定义热度评价指标体系，分析热度影响因素，接着对相关指标进行量化及标准化处理，最后利用层次分析法和熵值法确认指标权重，构建模型确认评价结果

综上所述，我们较完善地解决了问题一及问题二，而对于问题三，我们将继续对答复完整性及可解释性进行指标的构建及量化，进一步利用深度学习及 NLP 方面的知识去构建一个系统的质量评价体系，继续发展智能政务系统，为政府部门及公民提供更多的便利。

六、参考文献

[1]苏毅娟, 邓振云, 程德波, 宗鸣. 大数据下的快速 KNN 分类算法[J]. 计算机应用研究, 2016,33(04):1003-1006+1023.

[2] 耿丽娟, 李星毅. 用于大数据分类的 KNN 算法研究[J]. 计算机应用研究, 2014,31(05):1342-1344+1373.

[3]李荣陆, 胡运发. 基于密度的 kNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展,2004(04):539-545.

[4]孙国英. 文本挖掘技术研究及应用[D]. 南开大学硕士学位论文, 2001.

[5]黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011,34(5):856-864.

[6]张天宇, 谌志群, 黄孝喜, 王荣波. 基于改进 CFSFDP 算法的电信投诉文本聚类方法[J]. 电子科技, 2017,30(10):93-96.

[7]张宁, 贾自艳, 史忠植. 使用 KNN 算法的文本分类[J]. 计算机工程, 2005(08):171-172+185.

[8]向晓雯, 史晓东, 曾华琳. 一个统计与规则相结合的中文命名实体识别系统［J］. 计算机应用, 2005,25(10):2404-2406.