“智慧政务”中的文本挖掘

摘　要

近年来，网络问政平台逐步成为政府听民声，解民忧，集民智，暖民心的重要渠道，各类社情民意、政策民生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巨大挑战。本文依据自然语言处理技术和文本挖掘方法，在此问政潮流的背景下构建出一套全面的“智慧政务系统”，用于解决群众留言分类、热点问题挖掘以及答复意见评价等多类民生问题，成为连通政府与百姓沟通的桥梁。

群众留言分类模块：此模块对5折分层交叉验证划分下的训练集和验证集进行one-hot编码，采用基于TF*IDF词袋向量特征工程的机器学习方法和基于卷积神经网络CNN、循环神经网络RNN以及长短记忆网络LSTM下的深度学习方法进行留言分类。经择优对比，本文最终采用在机器学习下基于词性级别的TF*IDF模型的支持向量机分类器进行分类，经此步骤后F1-Score在5折分层交叉验证下的均值达到0.9013，并输出了相应的预测标签数据集。(针对问题一)

热点问题挖掘模块：此模块基于留言详情语料库，分别构建LSI主题模型和TF*IDF模型的相似度矩阵，并采用阈值判定和谱聚类两种方法来挖掘热点问题。经合理性综合比较，本文最终确定使用在LSI主题模型下聚类数为500的谱聚类方法来对群众留言详情进行分类，并构建具有时效性、持续性和认同性的热度评价指标计算各类别留言的热度值。筛选热度值排名前五的留言分类，总结各分类留言的“地点/人群”和“问题描述”，最终输出热点问题表和热点问题留言明细表。(针对问题二)

答复意见评价模块：此模块在考虑留言答复意见的相关性、完整性、可解释性和及时性的特点下，设计出了一套合理的量化方案。然后通过熵权法的权重计算方式对各评价指标间的权重进行客观的衡量评价，得到每条留言意见回复的满意度得分，并输出留言意见评价得分汇总表和留言答复意见评价方案，从而提高政府部门的绩效管理水平。(针对问题三)

关键词：TF*IDF，神经网络，LSI主题模型，谱聚类，熵权法

Text mining in "Smart Government Affairs"

Abstract

In recent years, the online questioning platform has gradually become an important channel for the government to listen to the people, relieve the people’s worries, gather people’s wisdom, and warm the hearts of the people. The amount of text data of various social conditions and public opinions, policies and people’s livelihoods has continued to increase. The work of the relevant departments collated with hotspots has brought great challenges. Based on natural language processing technology and text mining methods, this paper builds a comprehensive "smart government system" under the background of this political trend, which is used to solve many kinds of people's livelihood problems such as the classification of mass messages, mining of hot issues, and evaluation of answer opinions , becoming a bridge between the government and the people.

Mass message classification module: This module performs one-hot coding on the training set and verification set under the 5-fold hierarchical cross-validation division. It uses a machine learning method based on TF*IDF word bag vector feature engineering and a CNN based on convolutional neural network. The deep learning method under the recurrent neural network RNN and the long and short memory network LSTM performs message classification. Based on the comparison of optimal selection, this paper finally uses the support vector machine classifier based on the TF*IDF model of the part-of-speech level under machine learning for classification. After this step, the average value of F1-Score under 5-fold hierarchical cross-validation reaches 0.9013 and outputs the corresponding prediction label data set. (For task 1)

Hotspot problem mining module: This module builds the similarity matrix of the LSI theme model and the TF*IDF model based on the message detail corpus, and uses threshold judgment and spectral clustering to mine hotspot problems. After a comprehensive comparison of rationality, this paper finally determined to use the spectral clustering method with a clustering number of 500 under the LSI theme model to classify the details of the mass message, and construct a heat evaluation index with timeliness, continuity and approval to calculate each The popularity value of category messages. Filter the top five message categories of the popularity value, summarize the "location / crowd" and "problem description" of the messages of each category, and finally output the hot question list and the hot question message list. (For task 2)

Reply comment evaluation module: This module has designed a set of reasonable quantization schemes considering the characteristics of relevance, completeness,

interpretability and timeliness of the reply comments. Then use the weight calculation method of the entropy weight method to objectively measure and evaluate the weights between the evaluation indicators, obtain the satisfaction score of each message opinion reply, and output a summary table of message opinion evaluation scores and a message reply opinion evaluation scheme to improve the performance management level of government departments.(For task 3)

Keywords: ***TF*IDF, neural network, LSI topic model, spectral clustering, entropy weight method***

目　录

摘　　要.....I
Abstract.....II
一、“智慧政务系统”构建思路.....1
1.1“智慧政务系统”背景.....1
1.2 数据对象与输出成果.....1
1.2.1 处理的数据对象.....1
1.2.2 要求的输出成果.....3
1.2.3“智慧政务系统”框架搭建思路.....3
1.3 系统构建过程中所使用到的编程语言和相关的程序包.....4
二、模型假设与理论工具.....4
2.1 相关的假设前提.....4
2.2 理论工具介绍.....5
2.2.1 分词、分词方式.....5
2.2.2 去停、停用词表.....5
2.2.3 TF*IDF权值.....6
2.2.4 文本向量空间模型.....6
三、群众留言分类.....7
3.1 群众留言分类背景.....7
3.2 文本分类结果评价指标F1-Score.....8
3.3 群众留言分类详细步骤.....9
3.3.1 基于机器学习所建立的一级标签分类模型.....10
3.3.2 基于深度学习所建立的一级标签分类模型.....15
3.4 机器学习与深度学习结果对比.....19
3.4.1 机器学习文本分类结果对比.....19
3.4.2 深度学习文本分类结果对比.....22
3.5 “智慧政务系统”留言分类最终框架.....27
四、热点问题挖掘.....28
4.1 热点问题挖掘的背景.....28
4.2 热点问题阐述.....28
4.3 热点问题挖掘结果形式.....29
4.4 有关热点评价指标的构建.....30
4.5 热点问题挖掘解决步骤.....31
4.5.1 热点问题人群及时间段取舍.....31

4.5.2 相似度分类规则及留言分类流程.....32
4.5.3 应用不同的聚类原则.....34
4.6 不同的模型和聚类原则的选择所形成的结果对比呈现.....35
4.7 热点问题挖掘最终框架.....38
五、答复意见评价.....39
5.1 答复意见评价背景.....39
5.2 定性指标量化以及满意度指标构建.....39
5.2.1 相关性、完整性、可解释性和及时性的量化原理和过程.....40
5.2.2 权重评价方法的确立.....43
5.3 相关留言答复意见评价的评分展示.....45
5.4 答复意见评价改进分析.....46
六、“智慧政务系统”最终框架.....47
七、 参考文献.....49

一、“智慧政务系统”构建思路

1.1“智慧政务系统”背景

近年来，随着移动网络和相关电子产品在群众里的进一步普及，以及相关大数据处理技术的进一步发展，针对政务问题的反馈和民生问题收集这一块的政府工作流程，开始出现了一些新的更便于群众反馈以及政府部门进行相关意见收集和管理的主流方式。以往的绝大多数群众是通过去相关行政单位现场面对面提及意见、申述以及递交相关意见书，绝大多数政府机关单位是通过走访、现场办理等途径来进行双方意见交流，如今各类网络问政平台的出现使得群众反馈意见和提出述求更为方便，政府机构部门收集和管理相关信息更具效率。

但与之而来的问题就是各类社情民意相关的文本数据量不断攀升，此时如果还是按照传统的方式来依靠人工去进行留言划分和热点整理等相关工作的话，从另外一个角度来看，政府部门工作的效率反而可能是降低了的，也就无法把网络问政这种新方式的优势发挥到最大。

因此，为了使网络问政这种新的政府工作方式的优势最大化，本文基于NLP（自然语言处理技术）和文本挖掘技术构建了一个旨在解决如下三方面问题的“智慧政务系统”：群众留言分类、热点问题挖掘以及答复意见评价,以达到提升政府管理水平和施政效率的目的。其中的答复意见评价模块更是有利于政府部门自身进行客观的绩效考核。

1.2 数据对象与输出成果

1.2.1 处理的数据对象

一个用于业务处理的系统必定有其所要依托的数据对象，具体到本文所构建的“智慧政务系统”当中，其所要解决的数据对象即为第八届泰迪杯-数据挖掘挑战赛C题所附带的4个附件的非结构化数据。与结构化数据相比，更难让计算机理解。如下为4个附件文本数据的的部分截图展示：

一级分类	二级分类	三级分类
城乡建设	安全生产	事故处理
党务政务	党的建设	制度建设
国土资源	海洋气象地震	气象
环境保护	环保管理	生态示范和模范城区创建
纪检监察	党政处分	党纪处分
交通运输	出租车管理	经营权转让
经济管理	安全生产	安全生产管理
科技与信息产业	电信	其他
民政	福利慈善	其他
农村农业	草原草场	退牧还草
商贸旅游	旅游管理	其他
卫生计生	公共卫生	突发公共卫生事件处理
政法	法律服务	律师
教育文体	教师队伍和待遇	企业教师
劳动和社会保障	城镇居民社会保险	其他

图1 附件1中三级分类标签体系的第一级标签展示

对于附件1所附带的三级分类标签体系，应留意其完备性和实务可缺性。

完备性：据统计，附件一所附带的三级分类标签体系涵盖了15个一级分类、103个二级分类以及517个三级分类，我们假设这三者间的互相结合可以把任何一个现实当中的政务留言问题均归在某一个交集分类当中，如留言A可划分在城乡建设→安全生产→事故处理的三级分类交集当中；留言B可划分在国土资源→海洋气象地震→气象的三级分类交集当中等等；

实务可缺性：在某一次所收集到的群众留言问题的数据集当中，不一定这15个一级分类、103个二级分类以及517个三级分类的所有交集分类都能涉及到。正如下图所展示的附件2所收集到的留言数据集一样，经过简单的筛选可以发现其所涉及到的第一级分类标签只有7个，而并非15个一级分类都涉及到。

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
24	A00074011	西湖建筑集团占道施工有安全	2020/1/6 12:09:38	A3区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工	城乡建设
4445	U0004752	坪塘大道施工及大车噪音严重	2018/10/11 10:49:	坪塘大道施工及大车噪音严重扰民近几个月，之前环评造假的高架桥持续施工，加上坪塘大道夜间货车非常之	环境保护
3757	U0005806	代倾城小区道路交通安全出	2018/12/25 16:15:	地处月亮岛街道时代倾城小区一二三期之间的公共道路两边的人行道常年停满了社会车辆严重影响行人通行，	交通运输
165	U0005008	物“文昌阁”寂寞地荒废了	2010/11/2 10:38:4	文物是国家不可再生的文化资源，是我市悠久历史文明的见证。市级重点文物“大成殿”，原名文昌阁，位于	教育文体
1599	U0008058	效业工作者的工资什么时候	2016/11/9 14:29:4	全国每年都讲我们基础畜牧工作人员的工资是最低，工作是辛苦的，为什么每年只讲，而没行动呢？请问厅长	劳动和社会保障
9018	U0001885	楚江世纪城融江苑有传销组	2017/4/7 14:51:30	我朋友被家人骗到楚江世纪城搞所谓的两型社会投资7万多，里面有很多传销组织，希望当地政府挨家盘查。	商贸旅游
2468	U0007473	层医院的机构臃肿容易滋生	2011/3/11 16:47:4	我们是L市安江纱厂职工，现在是忠实的患者，见证安纺医院过度到卫生服务中心的明白人。原来院长只一	卫生计生

图2 附件2中部分用户留言及其一级标签分类情况展示

强调标签体系具有完备性和实务可缺性的目的在于给所构建的“智慧政务系统”当中的群众留言分类模块进行一定的警示标志。即这次的程序自动划分效果可能只是针对本次所涉及到的7个一级标签分类的留言问题划分。这就要求我们要时常对本系统的群众留言分类模块进行一定的程序更新和维护。

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
188006	A000102948	一米阳光婚纱艺术摄影是否合法纳税	2019/2/28 11:25:05	金额就上百万，因为地处居民楼内部，而且有	0	0
188007	A00074795	道路命名规划初步成果公示和城乡	2019/2/14 20:00:00	村的门牌号10年都未曾更换过，什么时候会给	0	1
188031	A00040066	县春华镇金鼎村水泥路、自来水到户	2019/7/19 18:19:54	组都通路灯了，且天还没黑就开灯了，浪费资	0	1
188039	A00081379	兴路步行街大古道巷住户卫生间粪	2019/8/19 11:48:23	仅仅是表面进行清扫。没有解决根本问题。主	0	1

图3 附件3中部分用户留言及其所获得的点赞和反对数展示

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
2549	A00045581	区景蓉华苑物业管理有问	2019/4/25 9:32:09	公司却以交20万保证金，取停车管理费，在业主大会结束后业主		2019/5/10 14:56:53
2554	A00023583	楚南路洋湖段怎么还没	2019/4/24 16:03:40	面的生意带来很大影响，当整体换填，且换填后还有三趟雨污水		2019/5/9 9:49:10
2555	A00031618	提高A市民营幼儿园老师	2019/4/24 15:40:04	时更是加大了教师的工作儿园聘任教职工要依法签订劳动合同		2019/5/9 9:49:14
2557	A000110735	寓能享受人才新政购房	2019/4/24 15:07:30	户A市，想买套公寓，请问35周岁以下（含），首次购房后，可		2019/5/9 9:49:42

图4 附件4中部分用户留言及工作人员意见回复展示（包含回复时间）

上面所列举到的四个附件数据集分别反映了在网络问政工作中的三个迫需解决的问题：群众留言分类、热点问题挖掘以及答复意见评价。

1.2.2 要求的输出成果

附件1和2所涉及到的群众留言分类问题需要“智慧政务系统”可以根据附件1的三级标签分类体系来对附件2的群众留言数据集进行第一级标签程序自动划分，评价划分效果的好坏使用的是数值指标F1-Score，F1值越高，效果越好；

附件3所涉及到的热点问题挖掘方面需要“智慧政务系统”可以依据系统内初设定的热度评价公式来对附件3所收集到的群众留言进行聚类和排序，给出热度排名前5的热点问题，并保存为文件“热点问题表.xls”；给出相应热点问题所对应的具体留言信息，并保存为“热点问题留言明细表.xls”

附件4所涉及到的答复意见评价问题需要“智慧政务系统”来构建一个满意度评价的打分体系，从而能对数据集中的每一条留言所对应的答复意见均进行客观的工作效果打分。

1.2.3“智慧政务系统”框架搭建思路

三大模块当中，群众留言分类模块和热点问题挖掘模块的构建思路均为对比择取的选取思路，其中前者为择优对比选取，后者为合理性对比择取。

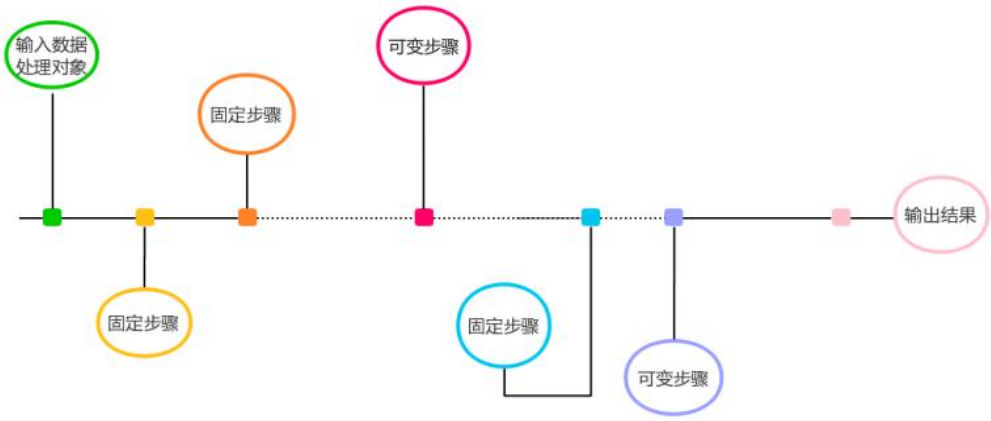

图5 群众留言分类模块和热点问题挖掘模块的构建思路步骤图

如图5所示，在群众留言分类模块和热点问题挖掘模块当中，程序思路的步骤解决均可以抽象成以流程图的形式来展示。从输入文本对象数据到最终结果输出的过程当中总有些步骤是固定下来的，而总有些过程是可以操控可变的。

群众留言分类模块的机器学习模型流程和热点问题挖掘模块流程：

- 文本数据的预处理→根据分层交叉验证的数据划分、去停、分词→**特征处理→选择模型对应的分类算法（分类器）**→得出F1-Score并输出相应的预测标签集。
- 文本清洗处理→**分词方式择取**→去停用词→形成词典→构建语料库→**选择算法模型**→余弦相似度计算→**聚类原则的选取**→将附件3的留言划分成不同的类→算出划分后每个类别的热度指数并按要求输出相应的排名前五的“热点问题表.xlsx”和“热点问题留言明细表.xlsx”。

在群众留言分类模块的机器学习模型流程当中，在主要的流程上除了特征处理里面的TF*IDF的分词级别以及分类器可以改变之外，其他步骤基本上是处于一个固定的状态，也即最终的F1-Score的取值是取决于TF*IDF的分词级别和分类器的不同搭配。在热点问题挖掘模块当中我们可以发现，在主要的流程上除了选择不同的算法模型和聚类原则外，其他步骤基本上也是处于一个固定的状态，也即最终的留言划分类别以及对应的热度指数取值是取决于算法模型以及聚类原则的不同搭配。因此，对于群众留言分类模块和热点问题挖掘模块的构建均可采取对比择选的思路。

在整个答复意见评价模块构建了一个确定性的流程。原因在于答复意见评价模块和前两个模块的目的任务不同，前两个模块的最终目的可以说是为了简化政府部门的相关工作而诞生的，它是有一个明确输出结果任务要求的；而答复意见评价模块是为了对政府部门自身进行相对客观的绩效管理而产生的，而判断一条留言所对应的留言意见回复到底好不好，有没有达到真正的行政办公要求，这一过程本身就充满了主观性。

因此，答复意见评价模块的构建思路为：用四个关键的性质来衡量一条留言意见回复的优劣，分别是：相关性、完整性、可解释性和及时性→把四个性质分别用一定的规则量化，从而得出每条留言意见回复的四个性质得分分别是多少→使用熵权法来确定四个性质指标分别所具有的权重→构建评价留言意见回复优劣的满意度指标S，其为每条留言意见回复的四个性质得分值与其对应的熵权相乘后再加总的结果，即如下公式：

$$S_i = \sum^{4}(W_j \times X_{ij}), i = 1,2,.....,n;\ j = 1,2,3,4。$$

其中W_j分别表示四个性质的权重，X_{ij}表示第i条留言在第j个性质上得分。

1.3 系统构建过程中所使用到的编程语言和相关的程序包

本系统中涉及的程序代码是在windows10中Python3.7的运行环境下进行的，采用到的程序包有：gensim、sklearn以及tensorflow下的keras包。此外，由于在群众留言分类和热点问题挖掘的程序运行中计算量较大，其对应部分是在Colab平台上运行的。

二、模型假设与理论工具

2.1 相关的假设前提

在构建“智慧政务系统”的过程当中，对于具体数据集的研究对象有必要先作出一个明确的界定。这里对研究对象的界定指的是对“留言”一词的界定，无论是群众留言分类模块还是热点问题挖掘模块，乃至最后的答复意见评价模块当中，我们均避免不了对“留言”所指代对象的区分。

在这情况下，我们作出对“留言”研究对象的界定，它为：留言详情。

作出该假设前提的原因是从包含信息量的多少来考虑的，因为留言主题只是对其详情的一个概括性描述，甚至很多留言的主题存在缺失。而对于留言主题我们并不是忽略不计，我们可把留言主题用于分类标签与留言间的匹配关系预测、后续对模块输出成果的经验预查看对比以及提取出相关的留言集关键词等方面。

2.2 理论工具介绍

以下介绍文本挖掘技术和自然语言处理的相关基础概念和方法体系。

2.2.1 分词、分词方式

所谓的分词简单来说就是把一段话（一段文本）根据某一个已经存在的词库按照一定的计算原理分成词语之间的组合形式。这种组合形式一般以特定的格式表现，较为常见的是词语之间以空格隔开，且一般分词处理之前会先进行去除标点符号、空格、换行等文本清洗程序。

如：征战四海只为今日一胜，我不会再败了。

征战四海只为今日一胜我不会再败了

[征战 四海 只 为 今日 一胜 我 不会 再败 了]

所分成的词是依据一个本确定存在的词典来进行划分处理的，不同的分词程序所依据的参考词典和分词计算原理会有所差异。

在“智慧政务系统”的构建过程中本文涉及到的分词方式有：jieba、thulac、snownlp分词，其均可在Python中调用相关程序包来实现。

2.2.2 去停、停用词表

本文所采用的停用词表为在提交的压缩包里所附上chinesestopwords.txt。

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。这些词极其普遍，记录这些词在每一个文档中的数量需要很大的磁盘空间，且由于它们的普遍性和功能，这些词很少单独表达文档相关程度的信息。如果在检索过程中考虑每一个词而不是短语，这些功能词基本没有什么帮助。如“啊”，“哎”，“的”，“地”等词；此外还有标点符号或别类的词也须作为停用词而被过滤。

停用词表就是相关在程序中我们所要过滤掉的词的集合，一般在自然语言的处理过程中我们会有一些已经成型的中英文停用词表可供我们选择。而去停当然就是把上述需要过滤掉的词在程序运行内存中删去的意思，一般这步骤是在文本清洗和分词处理后进行的。

如：不同分词方式去掉停用词后的结果：

Jieba:[征战 四海 今日 一胜 再败]

Snownlp:[征战 四海 今日 胜 再败]

2.2.3 TF*IDF权值

TF*IDF是一种用于信息检索与数据挖掘的常用加权技术。其中TF指的是词频，IDF指的是逆文本频率指数。在一段给定的文本里，TF指的是某一个给定的词语在该文本中出现的频率。这个数字是对词数的归一化，以防止它偏向长的文件。

对于在某一给定文本里的某个词语来说，它的TF公式为：

$$TF_{ij} = \frac{t_{ij}}{\sum_{j} t_{ij}}$$

其中t_{ij}表示在第i个文本中j词出现的次数，$i = 1,2,...,n$。n为数据集中所包含的文本的个数。随着j词在i文本出现的次数增加，在i文本的总词数不变的条件下，TF值增加。

IDF是一个词语普遍重要性的度量。某一特定词语的IDF，其可以由总文本数目n除以[包含该词语的文本的个数加一]后，再将得到的商取以10为底的对数得到下述公式，其中m为包含j的文本个数。

$$IDF_j = \log_{10}(\frac{\mathrm{n}}{m+1})$$

从定义的式子中可以看到，随着在所有文本所形成的语料库中包含j词的文本的个数增加，IDF值反而减少。此时某个词语的TF*IDF权值我们定义为：

$$[TF * IDF]_{ij} = TF_{ij} \times IDF_j$$

也就是说TF*IDF权值的主要思想是：字词的重要性随着它在文本中出现的次数成正比增加，但同时会随着它在全部文本所形成的语料库中出现的频率成反比下降。如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现（IDF也高），则认为此词或短语具有很好的类别区分能力，适合用来分类。

2.2.4 文本向量空间模型

文本类问题的数据大部分是字符串类型的数据，若可以把相关的文本段落等价转化到一个可以用数值来表示其内容的空间，即可便于后续计算机进行程序运算，也方便去度量计算各类型的指标。而文本向量空间即是这样的一个文本和数值向量一一对应的向量空间。

- **词典的构建**

在构建词典之前，一般来说我们先要完成去重（把有可能重复出现的文本去除掉多余的，仅保留一个作为后续建模对象）、去除换行、空格、标点符号以及乱码等的文本清洗处理程序，紧接着就是对文本进行分词处理。

而所谓的词典是根据经文本清洗、分词和去停处理后的文本来形成的一个包含不重复词语的集合，且在该集合中每个词语均有自己的独立编号，其格式为：

{“a”:0,“b”:1,“c”:2,.....};

注:在上述表述中，字母代表某个词，阿拉伯数字代表词语对应的序号。

如:两段文本[泰迪 杯 爱 棒]和[爱 篮球 棒]可共同组成如下格式的词典:

{“泰迪”：0，“杯”：1，“爱”：2，“棒”：3，“篮球”：4}

- **文本词向量（稀疏向量）的形成**

利用上述步骤所形成的词典，将文本对象词向量化。所谓的词向量化即是把文本依据词典一一投影为一数值型的向量，向量的维数和词典中包含的词数一致，在这里为了方便表达，设其为N。数值型向量每个位置对应的元素与词典中序号相同的字词一一对应，且该位置的取值为对应该词在文本中出现的次数。

即，设文本对象词向量化后所形成的文本词向量为：

$$\vec{V}=(v_1,v_2,.....,v_N)。$$

其中表v_j示在词典中序号为j的词在文本对象中出现的次数，$j=1,2,...,N$。

这样一来所有的文本对象均可以一一转化为其所对应的文本词向量形式。

- **文本向量空间（稀疏矩阵）的构建**

而在文本分类的问题当中，一个所要处理的数据集往往包括多个文本，我们设为n个。根据上述步骤同样可以形成一个依据这n个文本所共同建立的词典，根据这词典所包含的内容，我们可以把这n个文本均一一向量化为一个N维的稀疏向量：

$$\vec{V_i}=(v_{i1},v_{i2},v_{i3},.....,v_{iN}),i=1,2.....,n,$$

其表示第i个文本所对应的稀疏向量。

那么，由这n个文本所形成的数据集就可以用这n个对应的稀疏向量所形成的集合来表示，我们即称其为文本向量空间V。

即：

$$V=\{\vec{V_1},\vec{V_2},.....,\vec{V_n}\}$$

当然，向量空间也可以等价成一个矩阵的形式来表示研究：

$$V=V_{N\times n}=[\vec{V_1},\vec{V_2},.....,\vec{V_n}]=\begin{pmatrix} v_{11} & v_{21} & & v_{n1} \\ v_{12} & v_{22} & & v_{n2} \\ ... & ... & ... & ... \\ v_{1N} & v_{2N} & & v_{nN} \end{pmatrix},i=1,2...n。$$

这里的$\vec{V_i}$是以列向量的表达形式出现的，$i=1,2...,n$。这也是我们通常所说的稀疏矩阵。

三、群众留言分类

3.1 群众留言分类背景

为了便于面对面、点对点地有针对性处理相关群众的留言问题，在实际政务工作的过程中一般先由工作人员按照一定的标签划分体系对来自各类网络问政平台中所收集到的群众的留言进行分类划分，以便于后续把不同类型的留言分派到所属不同政府部门的工作人员手里，从而能实现专业的人干专业的事。但凭借人工经验处理来进行留言的分派划分会存在工作量大、效率低且差错率高等问题。因此，若能建立一个具备留言程序自动分类划分性的“智慧政务系统”，其就能极大的提高相关工作效率和准确性。当然，这个程序自动分类划分的模块要设置的合理，所建立的模型要适用于相关的政府工作实际，这样才能保证较高的分类准确性。

具体到本题所给的文本数据集当中，本文构建的“智慧政务系统”所要解决的问题是：建立留言内容的第一级标签分类模型，对附件2所收集到的群众留言数据集进行第一级标签的分类划分，把每条收集到的群众留言均“贴上”其特有的第一级标签。

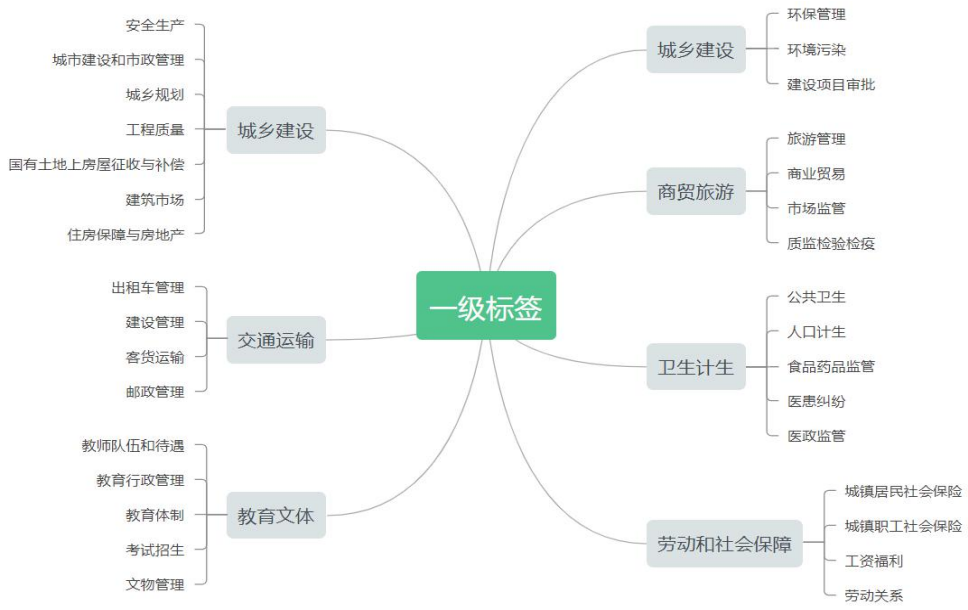

图6 附件二的群众留言所涉及到的一二级分类标签划分图

3.2 文本分类结果评价指标F1-Score

在自然语言处理NLP的文本分类问题中，分类模型均需要有一个具体的数量指标来评价其分类的准确性，而这样的一个指标通常指的就是我们所说的F-Score。

在“智慧政务系统”的分类效果评价当中，我们把F-Score进一步定义为F1-Score，其表达式为：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

$$P_i = \frac{True\ positive}{True\ positive + False\ positive}$$

$$R_i = \frac{True\ positive}{True\ positive + False\ negative}$$

其中P_i为第i类留言的查准率，R_i为第i类留言的查全率。

查准率和查全率的直观理解如下表：

表1 与查准率和查全率相关的标签阐释表

预测结果		1	0
1表示预测属于此类	1	True positive	False positive
0表示预测不属于此类	0	False negative	True negative

F值作为精确率与召回率的调和平均，其优势在于把每类留言的模型拟合率都考虑进其中，取的是一个综合平均的概念，使得评价的结果更具一般性和说服力。其取值越大即代表该模型的拟合预测效果越好。获取了评价指标之后，我们将进一步阐述所建立的一级标签分类模型，即群众留言分类的具体解决步骤。

3.3 群众留言分类详细步骤

在处理实际问题中，如果只建立一个模型，往往不够客观。因此，针对群众留言分类这一问题，我们采用对比择优的思路：即先建立多个关于解决此问题可能可行的模型，而每个模型中又可以使用不同的分类器算法，再分别计算出不同模型的F1-Score，最终取F1-Score值最大的模型及在其中所采用到的分类算法作为我们本次处理政务留言问题所择用的模型和算法。

在文本分类问题中，常用的分类方法包含机器学习方法和深度学习方法。在机器学习中通常使用监督学习对历史数据训练生成模型，用于预测文本的类别，常用的机器学习方法步骤包括使用词袋模型对文本进行one-hot编码，使用共线矩阵的上下文单词关系表示方法，使用TF*IDF的特征权重和特征提取方法，使用朴素贝叶斯分类器，支持向量机，线性分类器，随机森林等分类器方法；在深度学习中，为了处理高维度高稀疏的数据，常用的深度学习方法包括使用Word2vec的词向量嵌入方法，使用TEXTCNN的卷积神经网络的深度学习方法，使用TEXTRNN的循环神经网络的深度学习方法，使用LSTM的长短记忆模型的深度学习方法，使用FastText的快速文本分类方法以及使用Attension机制的NLP方法。

在此题中，文本分类类型为多分类问题，类别的一级标签数为7个，分别为“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”以及“卫生计生”。本文在词袋模型的基础上使用分层交叉验证的机器学习方法进行分类，使用的主要python包为sklearn;在TextCNN和TextRNN以及LSTM的深度学习方法基础上使用交叉验证进行分类，使用的主要

python包为gensim以及tensorflow下的keras包（TensorFlow的高级API）。

3.3.1 基于机器学习所建立的一级标签分类模型

机器学习一级标签分类问题的流程图如下：

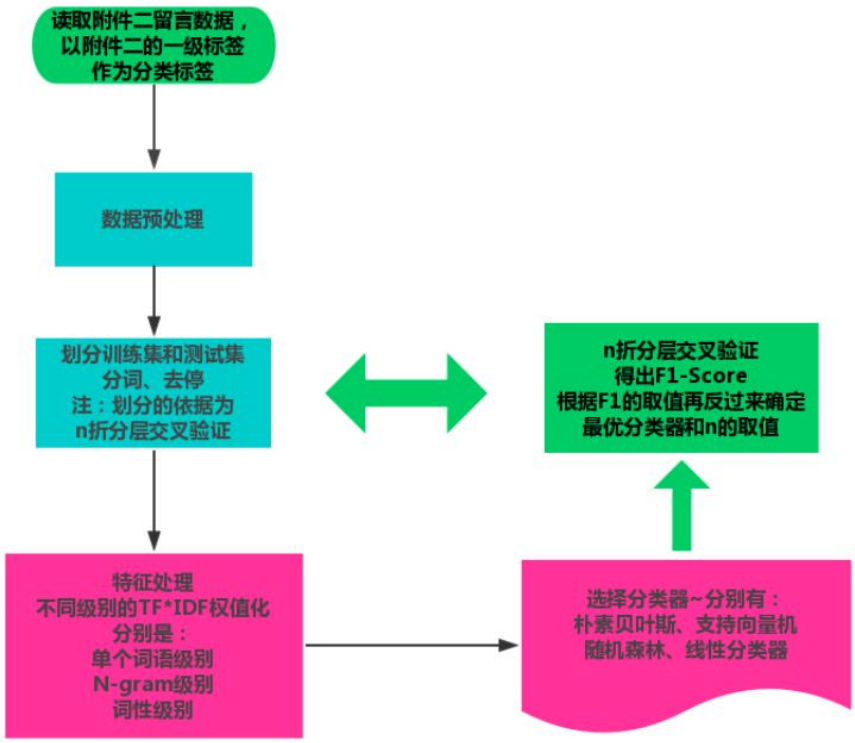

图7 基于机器学习所建立的一级标签分类模型的步骤流程图

Step1:留言数据的预处理

将附件2数据读取到“智慧政务系统”中，筛选留言详情和一级标签列，对每条留言详情进行去重、并使用正则表达式去除换行及空格等操作。重新合并每条留言详情及对应的标签，使其成为进一步待处理的数据集。

Step2:根据分层交叉验证进行数据划分

使用机器学习进行留言标签分类时，需要进行数据集的划分，使其分为训练集和验证集两大模块。其中训练集用来训练模型，验证集用来评估验证模型，模型通过合适的方法从训练集中学习，然后调用score方法在验证集上进行评估，打分，便可知模型当前的训练水平和F1-score。

如果只进行了一次划分，验证结果具有偶然性：在某次划分中，训练集里数据容易学习，验证集里数据复杂，这会导致最终的结果偏差。因此，考虑交叉验证以及分层交叉验证的方法来对数据集进行多次划分。

- **交叉验证原理：**

将合并后的留言和标签数据划分为n份，其中n-1份为训练集，余下一份为验证集，并进行score评分；每次数据的筛选条件满足已使用的验证集数据不再出现在之后的验证集中，并将去除验证集后的数据集作为训练集。以此类推，直至n份

数据集均成为验证集。其中n为数据集被划分的份数，当n=5时，为5折交叉验证。

在本文中，可以考虑使用划分成功的n份数据集进行n次score得分验算并取其平均值得到最终的F1值。因此，n折交叉验证与一次划分验证相比大大提升了对数据的使用效率，降低了偶然性数据出现的概率。

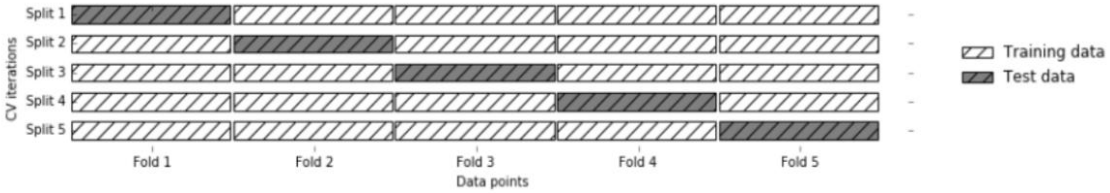

图8　5折交叉验证的直观图示

- **分层交叉验证原理：**

从上图可以看出，这种交叉验证方式，需要每次划分时对数据进行均分。根据本题数据可知：数据集存在7种类别，假设抽取出来的数据正好是按照类别划分的7类，即fold1全是0类，fold2全是1类，以此类推。这样的结果会导致模型训练时，没有学习到验证集中数据的特点，从而使得模型得分很低，甚至为0！因此，为了避免这种情况，本文中的“智慧政务系统”采用n折交叉验证的深化方法：分层n折交叉验证(Stratified n-fold cross validation）：其在每一折中都保持着原始数据中各个类别的比例关系，假设留言数据中7类的比例为1:2:2:3:2:2:1，采用5折分层交叉验证，则划分的5折中，每一折的数据类别都保持着1:2:2:3:2:2:1的比例，这样的验证结果更加可信。

至此，根据分层交叉验证的训练集和测试集的数据划分已经介绍完毕，具体划分的折数应先行自行确定，如取n=5。而最终折数确定将结合后面分类器的择取效果后，选择使得F1值取最大的折数为n(详细训练及验证集数据可见附件)。

Step3:分词、去除停用词

在划分好训练集与测试集之后，对二者进行分词和去除停用词。

在NLP中常见的用于对比分词效果的分词方法有：jieba、thulac、snownlp。在本文中，jieba、thulac、snownlp分词效果近似，但thulac中文分词包和snownlp分词包运行速率和效率较低，所占运行空间较大，因此对于文本分类，主要使用jieba分词包，使用jieba分词的精确模式，cut_all参数设置为False；

表2 运用jieba、thulac、snownlp的留言分词展示(**去掉了停用词**)

分词包	分词效果
原留言	想问下德群领导，医师注册需要哪些材料？
jieba	['想', '问', '下德群', '领导', '医师', '注册', '材料']
thulac	['想','问','德群','领导','医师','注册','需 ','材料']
snownlp	['想', '问', '德', '群', '领导', '医师', '注册', '材']

Step4:特征工程处理

- **初始化训练集向量空间模型**
- **计算训练集中每条留言详情中的每个词语的 TF*IDF 权值；**

此时文本向量空间模型的构建以及对应的留言向量TF*IDF权值化的过程均和分词的级别有关，其分别是：

<1>单个词语级别TF*IDF：稀疏矩阵代表了每个词语在不同留言详情中的TF*IDF分数，这时的分词形式是依所选用的对应分词方式的附带词库来进行的；

<2>N-gram级别TF*IDF：N-gram是多个词语在一起的组合，稀疏矩阵代表了N-gram的TF*IDF分数。所谓的N-gram是一种基于统计语言模型的算法。它的基本思想是将文本里面的内容按照字节进行大小为N的滑动窗口操作，形成了长度是N的字节片段序列；

如：以“留言分类”为例：

一元模型（unigram model）分为“留”、“言”、“分”、“类”；

二元模型（bigram model）分为“留言”、“言分”、“分类”；

三元模型（trigram model）分为“留言分”、“言分类”。

以此类推，分好词后，就可以像文本向量空间模型的处理方式那样，按照词库去比较句子中出现的次数。N-gram能够比较好的记录句子中词之间的联系，N越大句子的完整度越高，但是随之而来的是词的维度成指数级增长。所以一般取N=2或N=3。

<3>词性级别TF*IDF：稀疏矩阵代表了语料中多个词性的TF*IDF分数。所谓的词性是指名词、动词、形容词、副词、代词等词语属性，此时的分词形式是将每条留言详情按照词语属性来划分后再进行对应的权值向量化。

如上所说的在留言TF*IDF向量权值化的过程中所采用到的三个不同级别的分词形式将会是群众留言分类模块的构建当中一个重要的可变控制对比量。

- **将稀疏矩阵 TF*IDF 权值化，把稀疏矩阵中每个元素的取值改为该词语所对应的 TF*IDF 权值。**

$$V = V_{N\times \mathrm{n}} \Leftrightarrow M_{N\times \mathrm{n}} = [\overrightarrow{M_1}, \overrightarrow{M_2}, ..., \overrightarrow{M_{\mathrm{n}}}] = \begin{pmatrix} m_{11} & m_{21} & & & m_{n1} \\ m_{12} & m_{22} & & & m_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{1N} & m_{2N} & & & m_{nN} \end{pmatrix}$$

其中，这里的 $\overrightarrow{M_i}$ 是以列向量的表达形式出现，$i = 1,2...,n$。它为TF*IDF权值化后所形成的文本词向量，m_{ij} 是指在第 i 条留言中词典序号为 j 的词的TF*IDF权值，$j = 1,2...,N$；

- **根据权值化后的稀疏矩阵来获取特征词**

在本文中，首先通过留言详情中的词频画出词频图，可以看出在整个智慧政务系统的群众留言部分，“公司”，“学校”，“工作”等的词语出现频率较高，表明群众比较关注热切期望政府解决与这些关键词相关的问题。然后通过词语级别TF*IDF，N-gram级别TF*IDF以及词性级别TF*IDF获取权值化后的稀疏矩阵的特征词，词云图如下。其中“公司”，“政府”，“业主”，“学校”等出现的特征级别也较高，表明三种TF*IDF级别的关键词具有一定的相似性，相关问题为政府亟待解决的重点问题。

图9 按照留言词频建立的词云图

图10 词语级别TF*IDF的特征词提取

图11 N-gram级别TF*IDF的特征词提取

图12 词性级别TF*IDF的特征词提取

- **得验证集 TF*IDF 值并导入训练集，经 TF*IDF 权值化后形成文本向量空间。**

Step5：选择模型对应的分类算法（分类器），进行训练分类、预测，并计算在不同分类器的条件下所对应的 F1-Score；

<1>朴素贝叶斯

朴素贝叶斯，它是一种简单但极为强大的预测建模算法。称为朴素贝叶斯的原因是因为作出了每个输入变量之间是相互独立的假设。这个假设过于理想化，较难在实际问题当中得到满足（朴素），但是这项技术对于绝大部分的复杂问题仍然非常有效。

朴素贝叶斯模型由两种类型的概率组成：不同类别的概率P(Bj)和每个属性的条件概率P(Ai|Bj)。为了训练朴素贝叶斯模型，我们需要先给出训练数据②及其对应的分类标签。随后，如上所列举的两个概率，均可以从给出的训练数据中计算出来。一旦计算出来，概率模型就可以使用贝叶斯原理对新数据或测试集数据进行预测；

<2>支持向量机

支持向量机（support vector machines, SVM）是一种二分类模型，把其用于所构建的“智慧政务系统”当中实质上是把多分类问题转化为多个二分类问题来解决。它的基本模型是定义在特征空间上的间隔最大的线性分类器，其有别于后面所列举的感知机是因为其“间隔最大”的特性；SVM还包括核技巧，这令其具有非线性分类器的实质。其学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题⇔正则化的合页损失函数的最小化问题。它的学习算法就是求解凸二次规划的最优化算法。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。它的一个关键特性：训练完成后，大部分的训练样本都不需要保留，最终模型仅与支持向量有关；

<3>随机森林

集成学习中的随机森林算法的特质在于：“随机采样”。

随机采样(bootsrap)就是从我们的训练集里面采集固定个数的样本，但是每采集一个样本后，都将样本放回，也可以理解成有放回的抽样。对于我们的随机森林算法，一般会随机采集和训练集样本数m一样个数的样本，这样得到的采样集和训练集样本的个数相同，但是内容不同。若对M个样本训练集做T次的随机采样，则由于随机性，T个采样集各不相同。

对于一个样本，它在某一次含M个样本的训练集的随机采样中，每次被采集到的概率是$1/M$ 。不被采集到的概率为$1/(M-1)$。如果M次采样都没有被采集中的概率是$(1-1/M)^M$，当M→+∞时，$(1-1/M)^M \to 1/e \approx 0.368$ 。

即在随机森林的每轮随机采样中，训练集中大约有36.8%的数据没有被采样集采集。对于这部分大约36.8%的没有被采样到的数据，我们常常称之为袋外数据(Out Of Bag, 简称OOB)。OOB没有参与训练集模型的拟合，所以可以用来检测模型的泛化能力；

<4>线性分类器

线性分类器也就是通常所讲的感知机，用于区分数据应该属于哪个类的。感知机是二分类的线性模型，实例的特征向量作为其输入，事例的类别作为其输出，分别是+1和-1，属于判别模型。

假设训练数据集是线性可分的，求得一个能够将训练数据集正实例点和负实例点完全正确分开的分离超平面是感知机学习的目标，如若是非线性可分的数据，则最后将无法求解出超平面。感知机学习算法实质上是对所定义的损失函数进行极小化求解参数。

3.3.2 基于深度学习所建立的一级标签分类模型

基于深度学习所建立的一级标签分类模型主要由两大模块部分组合而成，分别是：文本预处理模块和模型的搭建模块。

- **文本预处理模块**

在我们所描述的深度学习的文本处理运用当中，文本预处理模块不像之前机器学习部分所描述的仅是包括去重、去空格、去除换行和标点符号以及进行分词和去停等基本文本处理工作那样简单。此时的文本预处理模块的过程在表现形式上类似于机器学习部分从数据集的输入直至形成文本向量空间（稀疏矩阵）的步骤。不过步骤中间具体所涉及的原理及最后所输出的数值型矩阵（文本向量空间）的形式是有较为明显的差异的。该文本预处理过程简述如下：

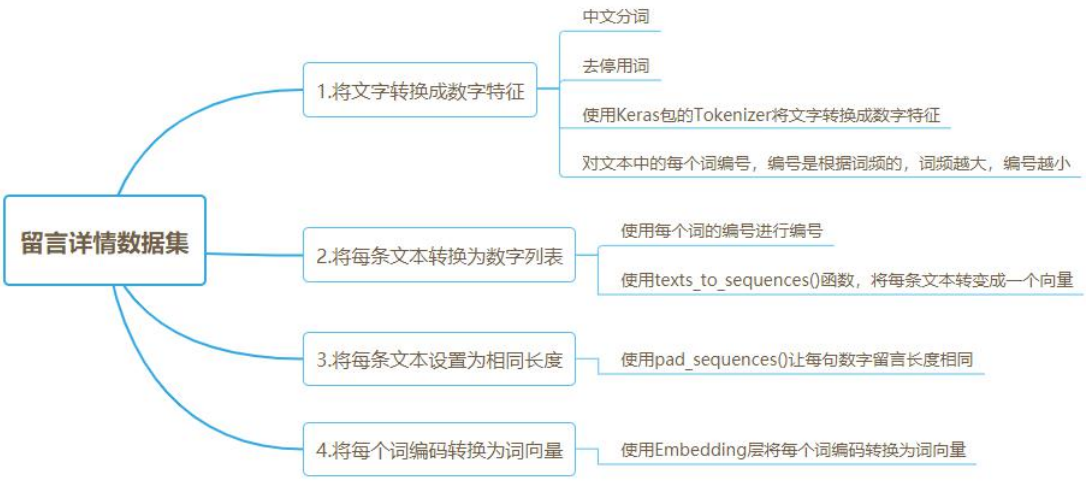

图13 keras留言详情文本预处理过程

- 读取数据集；
- 将文字转换成数字特征（分词方式选择：jieba、thulac、snownlp）；

具体通过使用Keras包的Tokenizer模块实现转换。当我们创建了一个Tokenizer对象后，使用该对象的fit_on_texts()函数，可以将输入的文本中的每个词编号，编号是根据词频的，词频越大，编号越小（使用word_index属性可以看到每次词对应的编码）。

- 将每条文本转换为数字列表

将数据集中的每条文本转换为数字列表，使用每个词的编号进行编号，使用该对象的texts_to_sequences（）函数，将每条文本转变成一个向量；

- 将每条文本设置为相同长度

使用pad_sequences()让每句数字影评长度相同，由于每句话的长度不唯一，需要将每句话的长度设置一个固定值。将超固定值的部分截掉，不足的在最前面用0填充；

- 将每个词编码转换为词向量

Embedding层基于上文所得的词编码，对每个词进行one-hot编码，每个词都会是一个vocabulary_size维的向量。通过神经网络的训练迭代更新得到一个合适的权重矩阵，行大小为vocabulary_size，列大小为词向量的维度，从而将本来以one-hot编码的词向量映射到低维空间，得到低维词向量。（在本文中，使用已经预训练好的词向量表示现有语料库中的词，嵌入层维度为300）

整个文本预处理后得到的嵌入矩阵如下：（矩阵维度为（79125，300））

$$\begin{matrix} -0.9654 & -0.6791 & -0.6597 & \cdots & -0.1350 & 0.7178 \\ -0.0177 & 0.5692 & 0.7086 & \cdots & -0.0129 & 0.6746 \\ 0.9571 & 0.2398 & 0.9052 & \cdots & -0.6236 & -0.9969 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -0.4803 & -0.1936 & -0.5468 & \cdots & 0.1511 & 0.5450 \\ -0.2598 & 0.3846 & 0.5135 & \cdots & -0.5949 & 0.1284 \end{matrix}$$

- **模型的搭建模块（文本分类算法的选取）**

本文主要参考对比的基于深度学习的文本分类算法有TextCNN,TextRNN,LSTM,以及CNN和RNN的组合（将CNN的输出接上RNN的输出或将CNN的输出和RNN的输出合并成一个输出）：

<1>TextCNN模型

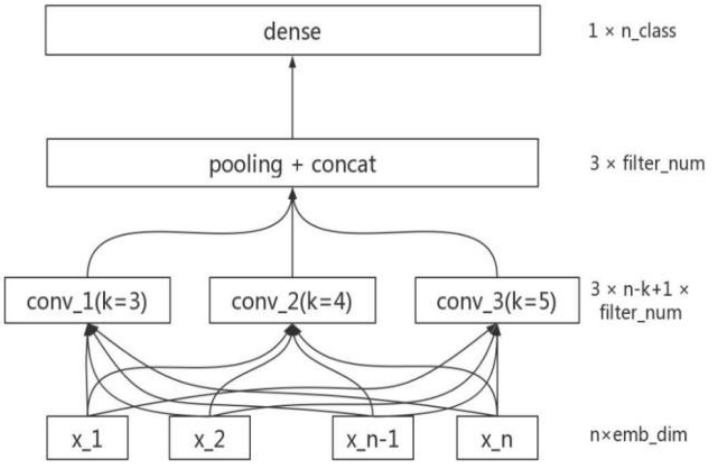

图14CNN模型的程序步骤分析图

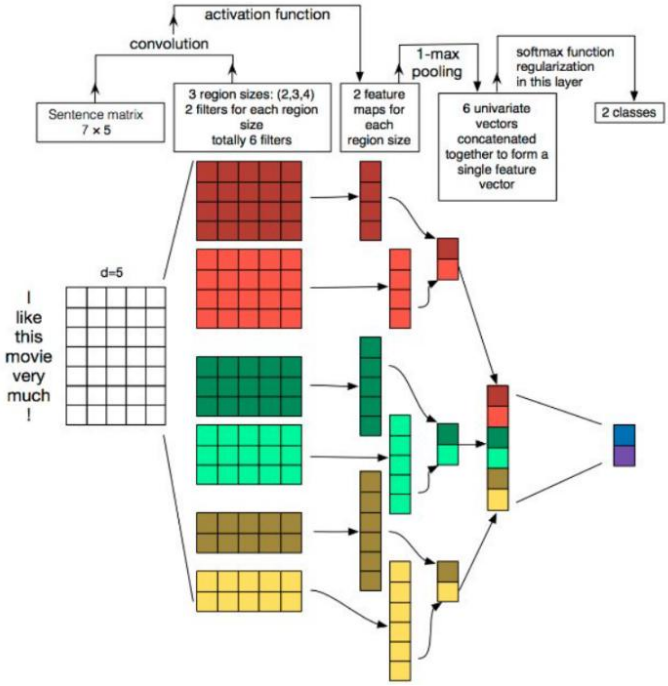

图15 CNN模型的详细原理讲解图

对应上述两个CNN模型的直观原理步骤图，可简单地概括TextCNN流程如下：

先将文本分词做Embeeding（词嵌入）得到词向量，将词向量经过一层卷积（Convolution）,一层Max-pooling，最后将输出外接softmax来做n分类。

Embedding：第一层是图中最左边的7乘5的句子矩阵，每行是词向量，维度=5，这个可以类比为图像中的原始像素点；

Convolution：然后经过kernel_sizes=(2,3,4)的一维卷积层，每个kernel_size有两个输出 channel；

Max-Polling：第三层是一个1-max-pooling层，这样不同长度句子经过pooling层之后都能变成定长的表示；

Softmax：最后接一层全连接的softmax层，输出每个类别的概率。

TextCNN的优势：模型简单，训练速度快，效果不错。

TextCNN的缺点：模型可解释性不强，在调优模型的时候，很难根据训练的结果去针对性的调整具体的特征，因为在TextCNN中没有类似gbdt模型中特征重要度的概念，所以很难去评估每个特征的重要度；

<2>TextRNN模型

利用RNN做文本分类也比较好理解，其实就是一个N vs 1模型。

对于英文都是基于词的；对于中文，首先要确定是基于字的还是基于词的。如果是基于词，要先对句子进行分词。之后，每个字/词对应RNN的一个时刻，隐层输出作为下一时刻的输入。最后时刻的隐层输出h为整个句子的抽象特征，再接一个softmax进行分类。如下是其流程的简述流程图：

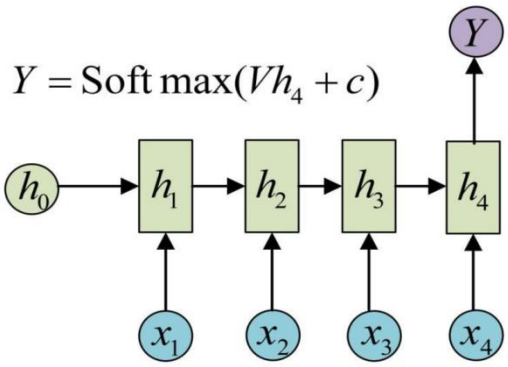

图16 RNN模型的直观步骤展示图

RNN与上述所提及到的利用CNN进行文本分类不同，CNN说到底还是利用卷积核寻找n-gram特征，卷积核的大小是超参。而RNN则可以处理时间序列，它通过前后时刻的输出链接保证了“记忆”的留存。但RNN模型的缺点在于循环机制过于简单，前后时刻的链接采用了最简单的f=activate(ws+b)。这样在梯度反向传播时出现了时间上的连乘操作，从而导致了梯度消失和梯度爆炸的问题；

<3>LSTM模型

简单来说LSTM为了解决长期依赖问题而生的，LSTM通过刻意的设计来避免长期依赖问题。记住长期的信息在实践中是 LSTM的默认行为，而非需要付出很大代价才能获得的能力！

上述提及到的RNN模型都具有一种重复神经网络模块的链式的形式。在标准的RNN中，这个重复的模块只有一个非常简单的结构，例如一个tanh层。

LSTM同样是这样的结构，但是重复的模块拥有一个不同的结构。不同于单一神经网络层，这里是有四个，以一种非常特殊的方式进行交互。

如下是LSTM模型的原理步骤展示：

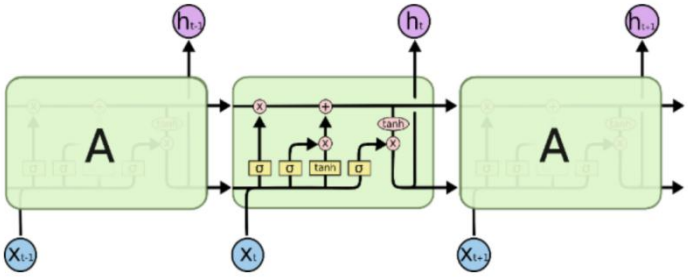

图17 LSTM模型的原理步骤展示图

在上面的图例中，每一条黑线传输着一整个向量，从一个节点的输出到其他节点的输入。粉色的圈代表 pointwise 的操作，诸如向量的和，而黄色的矩阵就是学习到的神经网络层。合在一起的线表示向量的连接，分开的线表示内容被复制，然后分发到不同的位置。

3.4 机器学习与深度学习结果对比

3.4.1 机器学习文本分类结果对比

经过随机选取部分数据进行程序的预运行验证后可发现当分词方式采用jieba分词、分层n折交叉验证采用的折数为5以及N-gram级别TF*IDF的N取（2,3）时，程序整体的运行效率和最终F1-Score的取值均较高且稳定，因此下述的结果呈现是在这几个小环节确定下来后继续展开讨论的。

表3 词语级别TF*IDF的不同分类器5折分层交叉验证的F1得分汇总

预测结果	Bayes	Linear	SVM	Random Forest
数据划分一	0.8697	0.8860	0.8914	0.8623
数据划分二	0.8583	0.8739	0.8837	0.8560
数据划分三	0.8627	0.8866	0.8967	0.8687
数据划分四	0.8680	0.8832	0.8870	0.8565
数据划分五	0.8584	0.8901	0.8873	0.8544
五折交叉验证均值	0.8634	0.8840	0.8892	0.8596

表4 N-gram级别TF*IDF的不同分类器5折分层交叉验证的F1得分汇总

预测结果	Bayes	Linear	SVM	Random Forest
数据划分一	0.7055	0.7136	0.6815	0.6864
数据划分二	0.7316	0.7134	0.6980	0.6938
数据划分三	0.7170	0.7216	0.6907	0.6781
数据划分四	0.7232	0.6925	0.6799	0.6730
数据划分五	0.7416	0.7260	0.6968	0.6828
五折交叉验证均值	0.7238	0.7134	0.6894	0.6828

表5 词性级别TF*IDF的不同分类器5折分层交叉验证的F1得分汇总

预测结果	Bayes	Linear	SVM	Random Forest
数据划分一	0.8714	0.8964	0.9096	0.8757
数据划分二	0.8646	0.8753	0.8869	0.8681
数据划分三	0.8673	0.9007	0.9108	0.8717
数据划分四	0.8529	0.8899	0.8964	0.8593
数据划分五	0.8540	0.8924	0.9027	0.8564
五折交叉验证均值	0.8620	0.8909	0.9013	0.8662

对比上述三个表格可知，机器学习模型输出的12个5折交叉验证均值中，当模型选择为机器学习模型，TF*IDF权值化时选用词性级别以及分类器算法采用SVM（支持向量机）的条件下，F1-Score达到最大值为：0.9013。其中在数据划分三验证集的验算下，F1-Score更是达到了：0.9108的高分。进一步，查看该模型下各个细项的报告：

表6 词性级别TFIDF、数据划分三、SVM细项评估报告

预测结果	Precision	Recall	F1-score	Supports
交通运输	0.90	0.75	0.82	61
劳动和社会保障	0.92	0.95	0.94	197
卫生计生	0.98	0.92	0.95	88
商贸旅游	0.88	0.83	0.85	121
城乡建设	0.84	0.94	0.88	201
教育文体	0.98	0.93	0.95	159
环境保护	0.94	0.94	0.94	94
accuracy			0.91	921
macro avg	0.92	0.89	0.90	921
weighted avg	0.91	0.91	0.91	921

从上述报告可知，分类效果最差的类为交通运输类，F1值仅为0.82，而分类效果最好的为教育文体和卫生卫计，F1值都为0.95。总体来说，该模型下分类效果是最优的，因此，根据群众留言分类模块构建所遵循的对比择优思路，理应选择上述使F1-Score值达到0.9108的步骤流程为机器学习最优的程序步骤。

为了更直观地去了解这最优程序步骤的运行结果，我们把其在5折分层交叉验证下的每次预测分类对比图均输出如下。根据下述图可以发现，分类效果最差的类依然为交通运输类，而分类效果最好的依然是教育文体和卫生卫计。

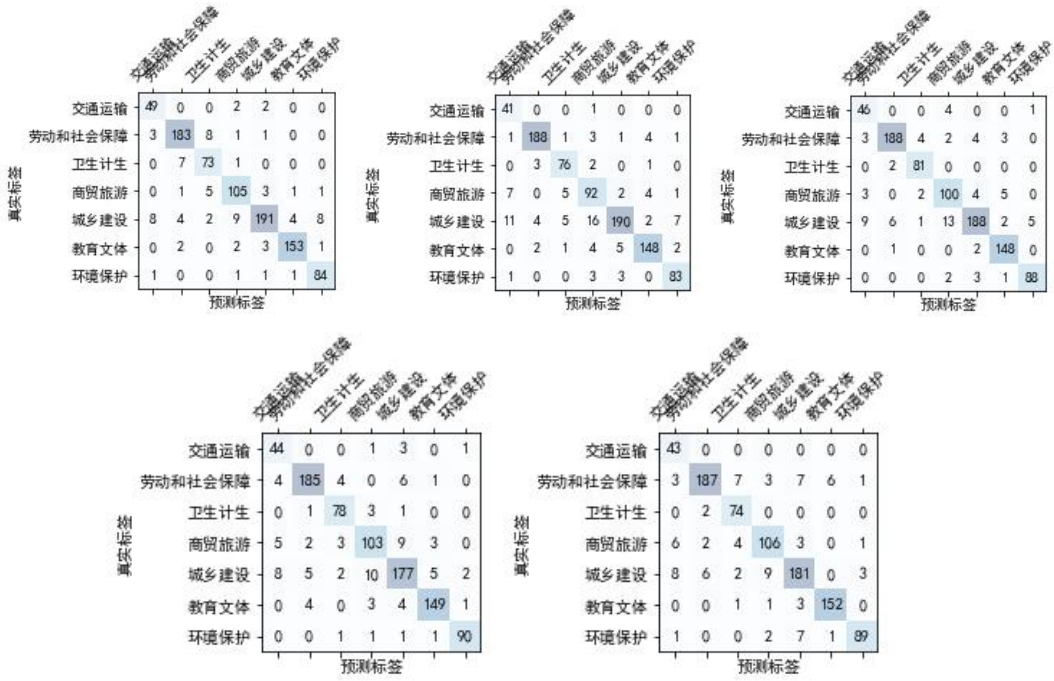

图18 在词性级的TF*IDF下，5折分层交叉数据验证集在SVM下的预测结果对比图

此外我们还可把上述最优程序步骤路径在5折分层交叉验证中的某一次测试集的全体标签预测结果输出为：predicted_label0.xlsx。并把其部分内容展示如下（完整结果predicted_label[0-5]可在压缩包中进行查看）：

	一级标签	预测一级标签
6855	劳动和社会保障	劳动和社会保障
5608	劳动和社会保障	劳动和社会保障
8846	卫生计生	卫生计生
1323	城乡建设	城乡建设
2242	环境保护	环境保护
7472	商贸旅游	商贸旅游
2737	环境保护	环境保护
5479	劳动和社会保障	劳动和社会保障
2782	环境保护	环境保护
8356	卫生计生	卫生计生
1137	城乡建设	城乡建设
6680	劳动和社会保障	劳动和社会保障
932	城乡建设	城乡建设
5925	劳动和社会保障	劳动和社会保障
1637	城乡建设	城乡建设
8102	商贸旅游	商贸旅游
4068	教育文体	教育文体
4998	教育文体	教育文体
7601	商贸旅游	城乡建设
3270	交通运输	交通运输
2462	环境保护	环境保护
3185	交通运输	交通运输
8998	卫生计生	卫生计生
8395	卫生计生	劳动和社会保障
6923	劳动和社会保障	劳动和社会保障
2711	环境保护	环境保护
3564	教育文体	教育文体
3338	交通运输	交通运输
7962	商贸旅游	商贸旅游

图19 最优程序步骤中在5折分层交叉验证下的某一次测试集的部分标签预测结果

在图中红色标注的部分为预测错误的分类留言，此结果形式的输出有助于我们有针对性地去寻找出分错类别的留言，去汇总它们所集中出错的分类从而达到归纳总结出出错的缘由并为后续程序的改进而服务的目的。

3.4.2 深度学习文本分类结果对比

- **TextCNN 神经网络运行结果：**

在本例中，TextCNN模型结构为词嵌入-卷积池化*3-拼接-全连接-dropout-全连接层（激活函数使用softmax层产出7分类结果）。其余参数compile过程：最优算法“adam”为梯度下降法，损失函数为“categorical_crossentropy”（多分类问题交叉熵），“acc”为精度；fit过程中设置epochs=20,batch_size=64。

TextCNN模型结构：

Model: "model_3"

Layer (type)	Output Shape	Param #	Connected to
input_5 (InputLayer)	(None, 100)	0	
embedding_15 (Embedding)	(None, 100, 300)	23737500	input_5[0][0]
conv1d_13 (Conv1D)	(None, 100, 256)	230656	embedding_15[0][0]
conv1d_14 (Conv1D)	(None, 100, 256)	307456	embedding_15[0][0]
conv1d_15 (Conv1D)	(None, 100, 256)	384256	embedding_15[0][0]
max_pooling1d_10 (MaxPooling1D)	(None, 25, 256)	0	conv1d_13[0][0]
max_pooling1d_11 (MaxPooling1D)	(None, 25, 256)	0	conv1d_14[0][0]
max_pooling1d_12 (MaxPooling1D)	(None, 25, 256)	0	conv1d_15[0][0]
concatenate_4(Concatenate)	(None, 25, 768)	0	max_pooling1d_10[0][0] max_pooling1d_11[0][0] max_pooling1d_12[0][0]
flatten_5 (Flatten)	(None, 19200)	0	concatenate_4[0][0]
dropout_20 (Dropout)	(None, 19200)	0	flatten_5[0][0]
dense_14 (Dense)	(None, 7)	134407	dropout_20[0][0]

Total params: 24,794,275
Trainable params: 1,056,775
Non-trainable params: 23,737,500

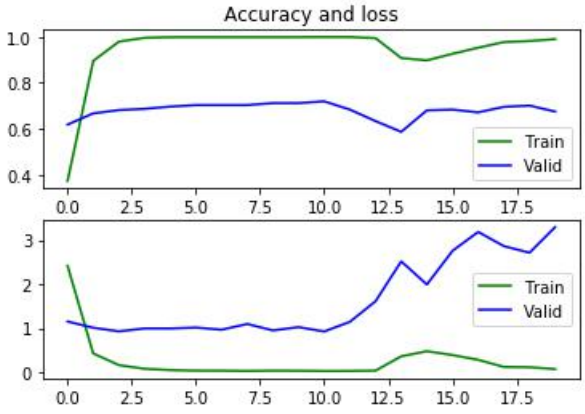

图20 TEXTCNN模型的精度和损失值

在上图中，第一行图为训练集和测试集的精度随epochs的增加而变动的情况；第二行图为损失函数产生的损失值随epochs的变动过程。在前十次的过程中，训练集的精度大致在第三次就达到了饱和，因此在这之后损失值和精度都没有太大的变化，运行效率较快，但是拟合的结果并不是太好；在后十次的过程

中，训练集和测试集的损失值反向延申，存在着过拟合的情况。因此选取第十次左右的验证集精度val_acc为0.7123,验证集的F1_score为0.6754；

- **TextRNN 神经网络运行结果：**

在本例中，模型结构为嵌入层+dropout层（权重0.5）+simpleRNN层+dropout层（权重0.5）+全连接层（激活函数使用softmax层产出7分类结果）。其余参数compile过程：最优算法“adam”为梯度下降法，“categorical_crossentropy”为多分类问题交叉熵,“acc”为精度；fit过程设置caepochs=20,batch_size=64。

TextRNN模型结构如下：

Model: "sequential_10"

Layer (type)	Output Shape	Param #
embedding_13 (Embedding)	(None, None, 300)	23737500
dropout_18 (Dropout)	(None, None, 300)	0
simple_rnn_6 (SimpleRNN)	(None, 16)	5072
dropout_19 (Dropout)	(None, 16)	0
dropout_6 (Dropout)	(None, 100)	0
dense_13 (Dense)	(None, 7)	119

Total params: 23,742,691
Trainable params: 5,191
Non-trainable params: 23,737,500

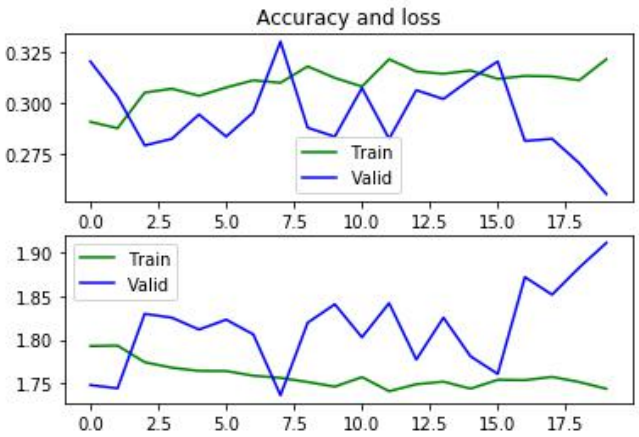

图21 TEXTRNN模型的精度和损失值

在上图中，第一行图为训练集和测试集的精度随epochs的增加而变动的情况；第二行图为损失函数产生的损失值随epochs的变动过程。由于train和valid杂乱无章，精度始终保持在很低的状态，表明拟合的结果较差，精度acc和F值也只有0.3左右，因此本例不建议使用RNN来进行文本分类；

- **LSTM 神经网络运行结果：**

在本例中，模型结构为嵌入层+dropout层（权重0.5）+LSTM层+全连接层（激活函数为relu函数）+dropout层（权重0.5）+全连接层（激活函数使用softmax层

产出7分类结果）。其余参数compile过程：最优算法“adam”为梯度下降法，“categorical_crossentropy”为多分类问题交叉熵，“acc”为精度；fit过程中设置epochs=20,batch_size=64。

LSTM模型结构如下：

Model: "sequential_3"

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, None, 300)	23737500
dropout_5 (Dropout)	(None, None, 300)	0
lstm_3 (LSTM)	(None, 300)	721200
dense_5 (Dense)	(None, 100)	30100
dropout_6 (Dropout)	(None, 100)	0
dense_6 (Dense)	(None, 7)	707

Total params: 24,489,507
Trainable params: 752,007
Non-trainable params: 23,737,500

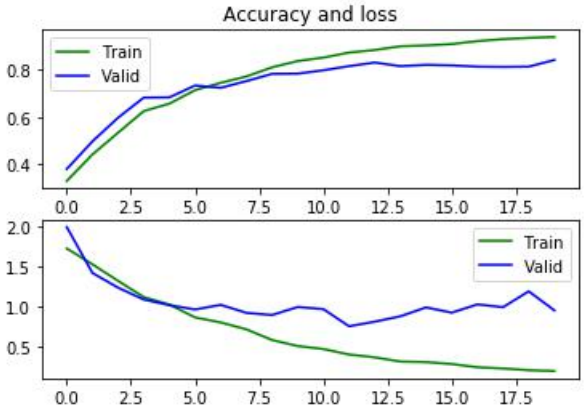

图22 LSTM模型的精度和损失值

在上图中，第一行图为训练集和测试集的精度随epochs的增加而变动的情况；第二行图为损失函数产生的损失值随epochs的变动过程。由于train和valid都是同步上升和同步下降，因此该模型不存在过拟合问题，拟合结果较好，验证集精度val_acc为0.8404,验证集的F1_score为0.8409：具体结果如下：

表7 LSTM长短记忆模型下的各项明细F值

预测结果	Precision	Recall	F1-score	Supports
交通运输	0.74	0.74	0.74	61
劳动和社会保障	0.85	0.87	0.86	197
卫生计生	0.90	0.81	0.85	88
商贸旅游	0.80	0.82	0.81	121
城乡建设	0.78	0.82	0.80	201
教育文体	0.92	0.88	0.90	159
环境保护	0.89	0.88	0.89	94

accuracy			0.84	921
macro avg	0.84	0.83	0.83	921
weighted avg	0.84	0.84	0.84	921

- **CNN+RNN1 神经网络运行结果：**

在本例中，使用CNN的输出直接拼接上RNN.

模型结构为：嵌入层+卷积层池化层+GRU层*2+全连接层（激活函数使用softmax层产出7分类结果）。其余参数compile过程：最优算法“adam”为梯度下降法，“categorical_crossentropy”为多分类问题交叉熵，“acc”为精度；fit过程设置epochs=20,batch_size=64。

CNN+RNN1模型结构如下：

Model: "sequential_2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, None, 300)	23737500
conv1d_1 (Conv1D)	(None, 100, 256)	230656
activation_1 (Activation)	(None, 100, 256)	0
max_pooling1d_1 (MaxPooling1	(None, 50, 256)	0
gru_1 (GRU)	(None, 50, 256)	393984
gru_2 (GRU)	(None, 256)	393984
dense_1 (Dense)	(None, 7)	1799

Total params: 24,757,923

Trainable params: 24,757,923

Non-trainable params: 0

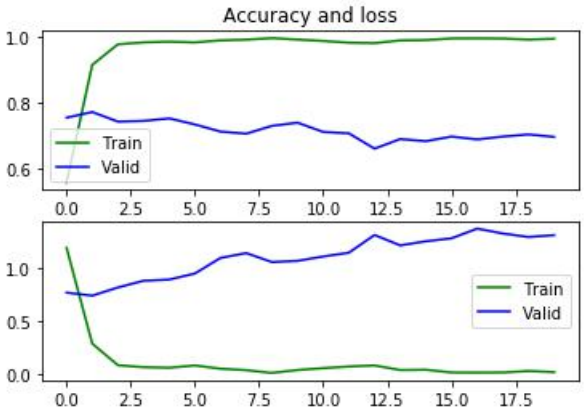

图23 CNN+RNN1神经网络的精度和损失值

在上图中，第一行图为训练集和测试集的精度随epochs的增加而变动的情况；第二行图为损失函数产生的损失值随epochs的变动过程。在epochs小于3时，模型便达到了精度为0.7112的结果，在epochs大于3的时候，训练集精度变化不大，而损失函数的损失值训练集和验证集有背向的趋势，验证集精度val_acc为0.7112,验证集的F1_score为0.7077。

- **CNN+RNN2 神经网络运行结果：**

在本例中，使用CNN的输出和RNN的输出合并成一个输出.

模型结构为：嵌入层+卷积层池化层+全连接层+拼接层+全连接层+双向GRU+全连接（激活函数使用softmax层产出7分类结果）。其余参数compile过程：最优算法“adam”为梯度下降法，“categorical_crossentropy”为多分类问题交叉熵，“acc”为精度；fit过程设置epochs=20,batch_size=64。

CNN+RNN2模型结构如下：

Model: "model_1"

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	(None, 100)	0	
embedding_4 (Embedding)	(None, 100, 300)	23737500	input_2[0][0]
conv1d_3 (Conv1D)	(None, 100, 256)	230656	embedding_4[0][0]
max_pooling1d_3 (MaxPooling1D)	(None, 25, 256)	0	conv1d_3[0][0]
flatten_2 (Flatten)	(None, 6400)	0	max_pooling1d_3[0][0]
bidirectional_1 (Bidirectional)	(None, 512)	855552	embedding_4[0][0]
dense_3 (Dense)	(None, 256)	1638656	flatten_2[0][0]
dense_4 (Dense)	(None, 256)	131328	bidirectional_1[0][0]
concatenate_1 (Concatenate)	(None, 512)	0	dense_3[0][0] dense_4[0][0]
dense_5 (Dense)	(None, 7)	3591	concatenate_1[0][0]

Total params: 26,597,283
Trainable params: 26,597,283
Non-trainable params: 0

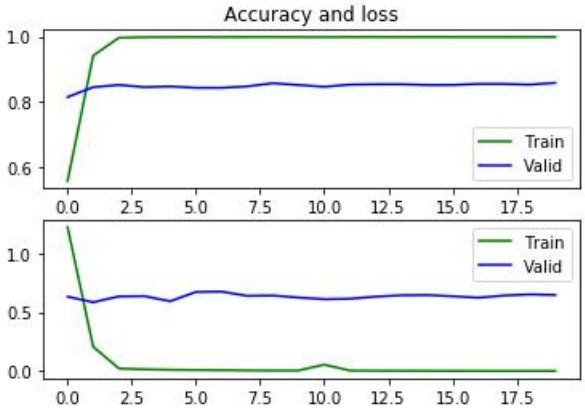

图24 CNN+RNN2神经网络的精度和损失值

在上图中，第一行图为训练集和测试集的精度随epochs的增加而变动的情况；第二行图为损失函数产生的损失值随epochs的变动过程。在运行过程中，epochs=3的时候结果便已经运行到最佳，此时的acc精度值为0.8588，效果最佳且训练集与验证集趋势相同，不存在过拟合的问题。

- **深度学习最终结果对比：**

表8 深度学习最终结果对比

深度学习模型	模型结构	Accuracy	F1
TextCNN	词嵌入-卷积池化*3-拼接-全连接-dropout-全连接层	0.7123	0.7023
TextRNN	嵌入层-dropout层-simpleRNN层-dropout层-全连接层	0.32	0.31
LSTM	嵌入层-dropout层-LSTM层-全连接层-dropout层-全连接层	0.8404	0.8409
CNN+RNN(1)	嵌入层-卷积层池化层-GRU层*2-全连接层	0.7112	0.7077
CNN+RNN(2)	嵌入层-卷积层池化层-全连接层-拼接层-全连接层-双向GRU-全连接	0.8588	0.8562

深度学习结果分析：在本文留言分类的深度学习中，由于accuracy和F1值的结果一般只相差0.01左右，因此可以使用accuracy近似等价于分类的结果。由表知，RNN在本例子的文本分类中不适用，而TEXTCNN和CNN+RNN(1)的结果一般，但是可能存在过拟合的情况，而使用LSTM和CNN+RNN(2)结果较好，尤其CNN+RNN(2)，在训练集上的精度趋近于99%的情况下，验证集上的精度可以达到0.8588还不存在着过拟合，因此深度学习选用CNN+RNN(2)模型来作为留言分类的结果。

3.5 “智慧政务系统”留言分类最终框架

对比机器学习和深度学习，机器学习选用词性级别的tfidf,在分类器选择支持向量机，得到的5折分层交叉验证F1值为**0.9013**;而使用深度学习CNN+RNN(2)的精确度为**0.8588**，因此，根据对比择优思路，本文选择**机器学习方法**建立一级标签分类模型。该模块的最终构建途径总结如下：

- 数据预处理：对附件2每条留言详情利用正则表达式去除换行及空格，重新合并每条留言详情和对应的标签，使其成为进一步待处理的数据集；
- 5折分层交叉验证的数据划分，并在划分好训练集和验证集后对两者均进行jieba分词和停用词表去停；
- 特征处理：初始化训练集向量空间模型，计算训练集中每条留言详情中的每个词语的TF*IDF权值，将稀疏矩阵TF*IDF权值化(采用词性级别TF*IDF)，根据权值化后的稀疏矩阵来获取特征词，计算验证集TF*IDF值并导入训练集，经TF*IDF权值化后形成文本向量空间；
- 选择SVM（支持向量机）作为分类算法（分类器）
- 计算出分层5折交叉验证的平均F1-Score值，并可从中输出某次测试集的标签

预测结果predicted_label.xlsx。

群众留言分类流程图：

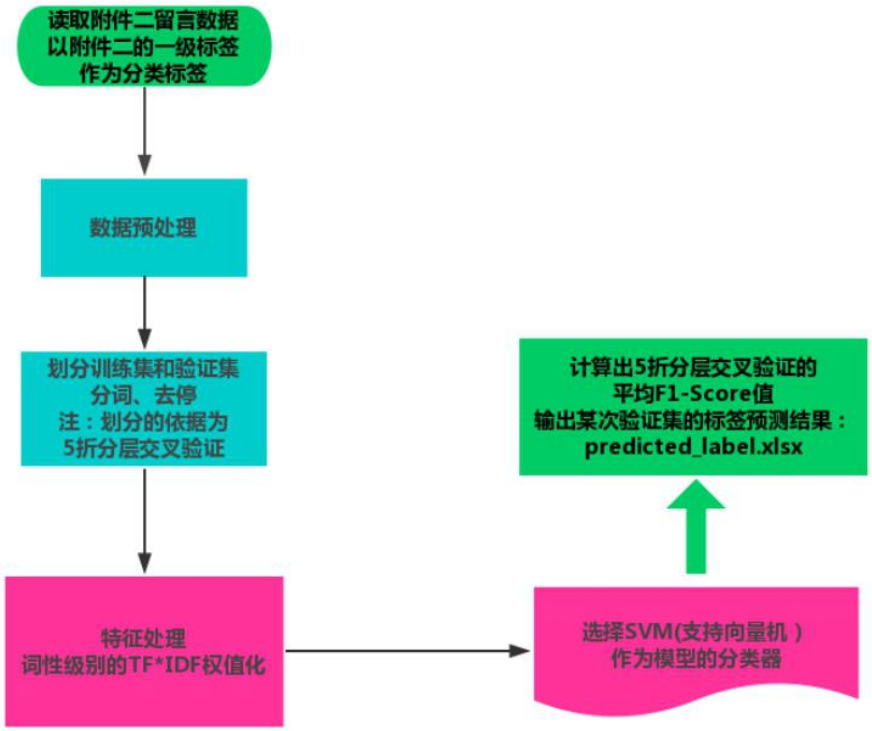

图25 群众留言分类模块的最终流程步骤框架思路图

上述最终框架搭建的步骤流程图是由基于机器学习所建立的一级标签分类模型的步骤流程在确定其中可控可变的步骤后对比择优而形成的。

四、热点问题挖掘

4.1 热点问题挖掘的背景

一个能对提升政府的管理水平和施政效率起极大推动作用的“智慧政务系统”，除了要解决群众留言分类问题外，还需进一步考虑热点问题挖掘的相关解决方案。

所谓的热点问题简单来说可理解成“某一时段内群众集中反映的某一问题”。“智慧政务系统”需重视热点问题的原因在于政府部门的工作重点之一就是急民之所急。而所谓的“急”就是“热”。

4.2 热点问题阐述

图26 热点问题特性

所谓的热点问题应该具有时效性、持续性和认同性

（1）时效性：一段相关的问题留言，如果距离政府部门意见收集的截止时间段越近，那么可以称其具有越强的热点时效性。原因在于群众的意见越是近段时间反映的，就证明其越是近期政府机构需要高度重视的“新出炉问题”。这类问题的解决速度往往反映了政府部门的工作效率如何。

（2）持续性：如果某方面的问题被某些人群在一段时间内持续性反映，则大概率这方面的问题一直未得到解决，群众一直被该类问题所困扰，因此这方面的问题也应该被看作成热点问题。在相关的留言收集中体现为某一类型的留言在好几个月，或者好几年都时不时有人反映，甚至是以经常的频率。

（3）认同性：一类问题是否能得到大多数群众的认同是确认该问题是热点问题的关键。只有民之所向了，才是众！这往往可以从该类型问题被反映的次数，以及每次（每条留言）被反映中得到多少人的赞同（点赞数）这两方面来考虑。

4.3 热点问题挖掘结果形式

热点问题挖掘模块的主要输出结果形式如下：

（1）对人群问题的留言进行归类，定义合理的热度评价指标，并给出对应不同分类问题的热度评价数据集。侧重点在于对留言所在的时间段以及隶属何种群体进行综合的考虑。

（2）依据定义的热度指标值的大小对分好的不同类型的留言进行类之间的排序，排序后，我们取热度前五类型的具体所有留言以一定的表格归纳形式进行输出管理。

其输出的表格具有如下的表头形式：

表9 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

表10 热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数

其分别反映了对分类留言的热度评价以及具体类别留言的提取管理。表9是对热点问题的提炼和总结，有利于政府策略制定者较快的了解问题所在。而表10则

是对表9具有辅助作用，在具体实施政府改进措施时，可以明确实施对象。

4.4 有关热点评价指标的构建

在进行热点评价指标的构建过程中，当然也涉及关于时效性、持续性和认同性的量化，如下就是相关的性质量化和评价指标的构建过程：

$$R_i = C_i \times \sum_{j=1}^{n_i} [P_{ij} \times (1 + T_{ij} - \mathrm{F}_{ij})],$$

其中：

$$P_{ij} = [1 + \frac{Q_{ij} - \text{数据集中最早记录的时间}}{\text{数据集中记录的截止时间} - \text{数据集中最早记录的时间}}]$$

$$C_i = [1 + \frac{\text{第}i\text{类留言中最晚记录的时间} - \text{第}i\text{类留言中最早记录的时间}}{\text{数据集中记录的截止时间} - \text{数据集中最早记录的时间}}]$$

Q_{ij} 表示相似性分类后第 i 类留言中的第 j 条留言的时间

T_{ij} 表示第 i 类留言中的第 j 条留言所获得的点赞数

F_{ij} 表示第 i 类留言中的第 j 条留言所获得的反对数

n_i 表示第 i 类留言所包含的留言数目

以下分析热点评价指标公式中各变量的具体含义（也即性质量化的原理）：

①有关 C_i，表示的是相似性分类后第 i 类留言所具有的经过数值修正后的热度持续性。其中，

$$\frac{\text{第}i\text{类留言中最晚记录的时间} - \text{第}i\text{类留言中最早记录的时间}}{\text{数据集中记录的截止时间} - \text{数据集中最早记录的时间}}$$

这一整体代表了该留言的原定热度持续性，其代表了第 i 类留言的时间跨度，取值在[0,1]之间，而把1加上这个分式整体的目的在于把分式整体的取值从[0,1]转移到[1,2]，而[1,2]的取值区间可以理解如下：

当原分式的整体取值为0时，C_i 取值为1，此时代表这类留言完全没有时间跨度，只在某一时间点上反映，而这类留言我们也要考虑，只不过其相对于其他类来说没有持续性而导致相对热度不足而已，但我们并不能忽视（即取值为0），因为该留言是确实存在的；当原分式整体取值为1时，C_i 取值为2，此时代表这类留言具有完全的时间跨度，即在有留言记录的这一段时间间隔内，均有群众留言反映该类型问题，因此其所代表的热度持续性最足，我们应“加倍重视”（即取值为2）；

②有关 P_{ij}，其表示相似性分类后第类留言中第 j 条留言所具有的经过数值修正后的时效性。其中，

$$\frac{Q_{ij} - \text{数据集中最早记录的时间}}{\text{数据集中记录的截止时间} - \text{数据集中最早记录的时间}}$$

这一整体取值为1代表了该条留言的原定时效性，留言时间越接近于截止时间

意味着时效性越强，理应得到更大的权重，而把1加上这个分式整体的目的在于把整体从[0,1]的取值转移到[1,2]，而[1,2]的取值区间可以理解如下：

代表无论留言时间多早，只要群众反映了该问题，政府部门就得有所重视，而不能忽视（即为0）；取值2进一步代表时效性，越接近于截止时间的留言越有可能是最近关注的热点，相对于较早的留言来说我们应“加倍重视”（也即取值为2），这样就可以避免由于0取值的出现而导致某些年份过早的留言在某种程度上被理解成无效的情况；

③有关$1+T_{ij}-F_{ij}$这一整体其表示的是第i类留言中第j条留言的单次被认同度，而要在$[T_{ij}-F_{ij}]$这一整体的基础上加上1是因为即使一条留言无人点赞也无人反对，但至少也有留言者自己一人表达了反馈意见；

④$\sum_{j=1}^{n_i}[]$则表示第i类留言中每条留言的时效性和认同度的累加，也可以认为是热度持续性的一种体现。

因此，将把上述R_i定义成第i类留言的热度评价指标能反映该类留言的持续性、时效性和认同性。(在上述性质的量化过程中本文涉及到了时间跨度的衡量，而在实际的程序编写过程当中本文把时间跨度转化为以秒数来计算的。)所以，R_i的定义是合理的。

4.5 热点问题挖掘解决步骤

4.5.1 热点问题人群及时间段取舍

本文所构建的“智慧政务系统”采取的是逆向思维的思考方式来解决问题，即不从正面去考虑把时间地点和人群均一一确定下来之后，再去讨论这一部分的限定人群中，哪方面的问题能称之为最“热”。因为这种思考方式会导致限定人群变化过于复杂，难以界定。只要时间，地点，人群三者中有一发生了改变，这一限定人群也即不是原来的人群了。

此时，逆向思维的方式就可以解决上面限定人群变化过于复杂，难以界定热点所在范围的困难了。本文直接把所有的留言进行相似性分类，只要相似性分类的方法选择的好，留言的区分类别效果肯定良好。这时，再从每一类的留言中去查看其最早留言的时间和最晚留言的时间来作为该类留言的时间跨度，从该类留言的人群特点直接察看总结出其所在的地点和所属的人群，并进行最终的问题概括描述。

在留言已经根据相关相似性分类规则来区分为不同类别的条件下，就可以使用4.4所定义的热点评价指标来计算出每类留言的热度，从而即可对不同类别的留言进行热度排序并提取出前五热度类型留言的所有单条结果。

因此，所构建的“智慧政务系统”解决热点问题挖掘的关键就在于能否采用

合适、正确率高且具有效率的相似性分类规则来对留言集合进行分类。

4.5.2 相似度分类规则及留言分类流程

留言相似性分类具有不同的模型和方法，每种模型和方法又都具有各自的优点和缺点。我们所构建的“智慧政务系统”在选用何种相似性分类规则的问题上遵循的是合理性对比择取的原则，即对全体留言运用不同的模型来进行处理，根据处理后所输出的分类结果的合理性来综合判断到底何种模型更适合于进行群众留言的热点聚类问题。

在整个留言相似性分类的过程中，除了使用的训练模型和最终运用的聚类原则不同之外（这也是合理性对比择取所重点研究的两方面），有很多步骤是具有相似和本质雷同之处的，因此可以把整个留言相似性判别的流程总结如下：

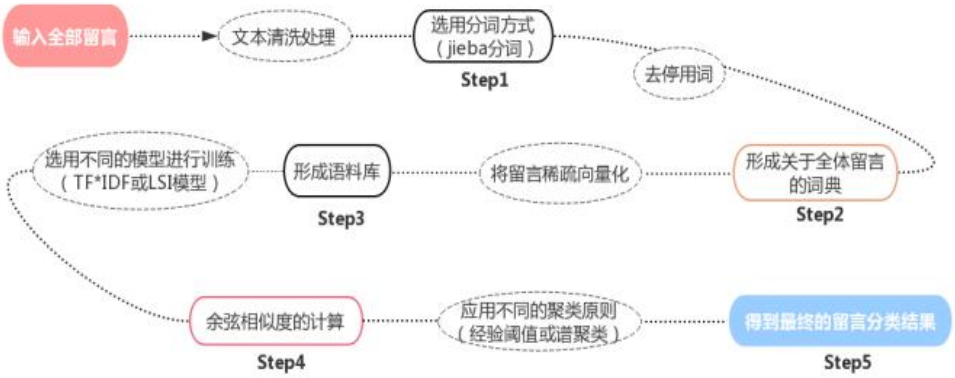

图27 留言相似性判断的流程步骤图

对从全部留言的输入到Step3形成语料库的步骤都在2.2中进行了详细的介绍，把文本留言转化成了文本向量的形式，这样的好处在于可以把文本度量数值化，以便于后续输入模型中训练以及进行相关的相似度计算。

在构建“智慧政务系统”的热点挖掘模块的过程中，主要参考对比了如下两个模型：

- **TF*IDF算法:⇔将稀疏矩阵TF*IDF权值化**

  即把稀疏矩阵中每个元素的取值改为该词语所对应的TF*IDF权值

  也即：

$$V = V_{N\times \mathrm{n}} \Leftrightarrow M_{N\times \mathrm{n}} = [\overrightarrow{M_1}, \overrightarrow{M_2}, ..., \overrightarrow{M_{\mathrm{n}}}] = \begin{pmatrix} m_{11} & m_{21} & & m_{n1} \\ m_{12} & m_{22} & & m_{n2} \\ \cdots & \cdots & \cdots\ \cdots & \cdots \\ m_{1N} & m_{2N} & & m_{nN} \end{pmatrix}$$

其中这里的 $\overrightarrow{M_i}$ 是以列向量的表达形式出现的， $i = 1,2...,n$。它为TF*IDF权值化后所形成的文本词向量。 m_{ij} 表示的是第 i 条留言中词典序号为 j 的词的TF*IDF权值， $j = 1,2...,N$。

TF*IDF算法的优点在于其给稀疏向量进一步权重化，TF*IDF值与原来词频相比更好的体现了该词的重要性，且该模型算法便于理解，程序实现起来效率高；TF*IDF算法的缺点在于它是基于单词的出现与否及TF*IDF等信息来检索，但是说了和写了哪些词和真正要表达的意思之间有很大的区别，其中两个最主要的阻碍是单词的多义性和同义性。简单来说就是TF*IDF算法并没有考虑到语义和词义。

多义性指的是一个单词可能有多个意思，比如苹果，既可以指的是水果苹果，也可以指的是苹果公司；同义性指的是多个不同的词可能表示同样的意思，比如“强制”和“强迫”。

另外一个缺点在于该算法把代表每条留言的稀疏向量直接用于模型TF*IDF权值化训练和之后的相似度计算，在这种情况下如果留言长度一旦增加，有可能会使得语料库中所包括的词语数激增，从而根据语料库所形成的稀疏向量或者文本向量就有可能会产生维数灾难的现象。

- **LSI（LSA）模型算法：该模型也叫隐语义模型**

其原理步骤如下：依据Step3语料库所构建的稀疏矩阵：

$$V = V_{N\times \mathrm{n}} = [\overrightarrow{V_1}, \overrightarrow{V_2}, ..., \overrightarrow{V_{\mathrm{n}}}] = \begin{pmatrix} v_{11} & v_{21} & & v_{n1} \\ v_{12} & v_{22} & & v_{n2} \\ \cdots & \cdots & \cdots \ \cdots & \cdots \\ v_{1N} & v_{2N} & & v_{nN} \end{pmatrix}, i = 1,2...n。$$

这里的$\overrightarrow{V_i}$是以列向量的表达形式出现的，$i = 1,2...,n$。

或者也可先把稀疏矩阵经TF*IDF模型训练后形成一权值矩阵：

$$V = V_{N\times \mathrm{n}} \Leftrightarrow M_{N\times \mathrm{n}} = [\overrightarrow{M_1}, \overrightarrow{M_2}, ..., \overrightarrow{M_{\mathrm{n}}}] = \begin{pmatrix} m_{11} & m_{21} & & m_{n1} \\ m_{12} & m_{22} & & m_{n2} \\ \cdots & \cdots & \cdots \ \cdots & \cdots \\ m_{1N} & m_{2N} & & m_{nN} \end{pmatrix}$$

其中这里的$\overrightarrow{M_i}$是以列向量的表达形式出现的，$i = 1,2...,n$。它为TF*IDF权值化后所形成的文本词向量。m_{ij}表示的是第i条留言中词典序号为j的词的TF*IDF权值，$j = 1,2...,N$。

设X是N×n维的矩阵，代表的是上述两个矩阵中的其一，利用SVD（奇异值分解）对单词文档矩阵进行分解：

$$\mathrm{X} = \mathrm{T} \times \mathrm{A} \times \mathrm{D}^{\mathrm{T}}$$

其中T为N×M维矩阵，A为M×M维对角矩阵，D为n×M维矩阵。A中每个对角线元素值称为奇异值，T中的每一列称为左奇异向量，D中每一列称为右奇异向量。保存A中最大的K个奇异值，以及T和D中对应的K个奇异向量，K个奇异值构成新的对角矩阵A_K，K个左奇异和右奇异向量构成新的矩阵T_K和D_K。此时：

$$\mathrm{X} \approx \mathrm{T_K} \times \mathrm{A_K} \times D_{\mathrm{K}}^{\mathrm{T}}$$

该步骤的目的在于从单词-文档矩阵中发现不相关的索引因子，将原来的数据

降维映射到相关的主题语义空间中。因为在单词-文档矩阵中不相似的两个文档，可能在降维后的主题语义空间内比较相似。$D_{\mathrm{K}}^{\mathrm{T}}=[\overrightarrow{d_{K1}},\overrightarrow{d_{K2}},.....,\overrightarrow{d_{Kn}}]$中的每一个列向量，$\overrightarrow{d_{Ki}},i=1,2,.....,n$，此时代表每一条留言在主题语义空间中的投影，即将文本向量从N维空间映射到了K维空间，在此过程中去掉了原本文档矩阵X中的很多噪音。这也意味着经过LSI模型训练后所输出的结果也是向量的形式，是将原本N维的文本向量降维到K维，此后便可用$\overrightarrow{d_{Ki}},i=1,2,.....,n$来进行留言之间的相似度计算。

LSI算法的优点在于：一次奇异值分解就可以得到主题模型，同时通过降维的方式能够捕获到单词之间的相关性从而解决TF-IDF模型中没有解决到的语义和词义问题。缺点在于：SVD计算非常的耗时，尤其是我们的文本处理，词和文本数都是非常大的，对于这样高维度矩阵做奇异值分解是非常难的。同时主题值选取对结果的影响非常大，较难选择合适的k值。它得到的不是一个概率模型，缺乏统计基础，结果难以直观的解释等。

相似度的计算规则有很多，其中当属向量之间的余弦相似度计算最为简单，且其具有明确的几何意义。在经过TF*IDF模型或者LSI模型训练后，原本数据集当中的全体留言集合均转化成了数值型向量的矩阵形式，即：

$$\vec{V}=V_{N\times \mathrm{n}}=[\overrightarrow{V_1},\overrightarrow{V_2},.....,\overrightarrow{V_{\mathrm{n}}}]（TF*IDF\ 模型）$$

$$D_{\mathrm{K}}^{\mathrm{T}}=D_{k\times \mathrm{n}}=[\overrightarrow{d_{K1}},\overrightarrow{d_{K2}},.....,\overrightarrow{d_{Kn}}]（LSI\ 模型）$$

此时任意两条留言之间的相似度计算均可由向量之间的余弦值计算获得，即：

$$\cos<i,j>=\frac{\overrightarrow{V_i}\bullet\overrightarrow{V_j}}{\left|\overrightarrow{V_i}\right|\times\left|\overrightarrow{V_j}\right|}=\frac{\sum_{\mathrm{t}=1}^{N}(v_{it}\times v_{jt})}{\sqrt{\sum_{t=1}^{N}v_{it}^2\times\sum_{t=1}^{N}v_{jt}^2}}=\frac{\overrightarrow{d_{Ki}}\bullet\overrightarrow{d_{Kj}}}{\left|\overrightarrow{d_{Ki}}\right|\times\left|\overrightarrow{d_{Kj}}\right|}=\frac{\sum_{\mathrm{t}=1}^{N}(\overrightarrow{d_{Kit}}\times\overrightarrow{d_{Kjt}})}{\sqrt{\sum_{t=1}^{N}d_{Kit}^2\times\sum_{t=1}^{N}d_{Kjt}^2}},\text{其中}i,j=1,2,.....,n$$

在用余弦相似度来度量任意两条留言的相似性后，我们可以根据留言间所形成的相似度矩阵simM来进行最终的留言分类；在留言间相似度矩阵的计算上，为了有效的减少计算量，计算过程中可只计算上三角矩阵。n条留言详情所形成的相似度矩阵应为一n×n的对称矩阵，在这前提下，只计算对应的上三角矩阵后再把其对称或直接把该上三角矩阵用于后续的聚类计算的两种做法均可避免对留言间都进行两两的相似度计算，从而可极大地提高程序的运行效率。

4.5.3 应用不同的聚类原则

聚类原则同样有多种可运用的法则，本文主要采用如下两种聚类方式来进行最终的留言分类比较：

- **实例数据支持验证下的阈值选择分类**

在得到每条留言两两之间的相似度余弦值之后，形成了一个留言相似度矩阵simM。阈值选择分类指的是若两留言之间的余弦相似度低于某个数值，我们即把该两条留言划分到同一类别的留言当中，因此，只要我们把相似度矩阵simM和阈值相比较就能得到最终的留言分类。

阈值选择分类的难点在于阈值的选取，我们需要尝试选择不同的阈值来对相似度矩阵simM进行分类，最后通过实际查看所分类到同一类的留言的实际详情来判断具体的留言分类是否符合常理。

- **谱聚类**

谱聚类是指把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

它与阈值判断方法一样也涉及到经验判断的步骤：聚类数的选择。其算法流程概述如下：

图28 谱聚类算法流程图

谱聚类只需要数据之间的相似度矩阵，并使用了降维，因此对于处理稀疏数据的聚类很有效，且处理高维数据聚类时的复杂度相比于传统聚类算法来说更好。但如果最终聚类的维度非常高，谱聚类的运行速度和最后的聚类效果会不好；不同的相似矩阵得到的最终聚类效果也会不同，在本文中所使用的是余弦相似度计算准则。

4.6 不同的模型和聚类原则的选择所形成的结果对比呈现

无论是阈值的选取还是聚类数的确定依靠的都是对留言聚类后的结果进行人为的经验合理性判断来互相对比的。而在热点问题挖掘模块的构建过程当中，我们使用了如下的训练模型和相关的阈值或聚类数进行搭配尝试并把与上述搭配相

关的热点问题表和热点问题留言明细表都输出附在压缩包上。

表11 LSI模型与其搭配的阈值或谱聚类数汇总表

训练模型	LSI模型							
阈值	0.45	0.49	0.50	0.51	0.52	0.53	0.54	0.55
谱聚类数	450	480	490	500	502	510	520	

表12 TF*IDF模型与其搭配的阈值或谱聚类数汇总表

训练模型	TF*IDF模型			
阈值	0.05	0.06	0.07	0.08
谱聚类数	500	600		

经过对这总共21种的搭配结果输出表的查看对比后，我们发现当采用LSI模型以及聚类数为500的谱聚类方法的搭配时，留言的聚类合理性相对最好，此时我们把此搭配所对应的两个结果输出表进行进一步的地点/人群的确定和相关问题描述的汇总，从而得出热点问题挖掘模块所要求输出的“热点问题表.xls”和“热点问题留言明细表.xls”。并将两个表格的内容亦呈现如下：

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	1	5237.46	2019/01/31 至 2020/01/17	A市多个不同小区的业主	小区物业管理混乱，在网络、消防、水电以及环境污染等方面均出现不同程度的问题
2	2	4146.40	2019/01/04 至 2019/12/10	A市多个不同小区	小区所附带的学位分配均出现不均衡的现象
3	3	3078.96	2019/01/08 至 2019/05/28	西地省A市A4区	恶性车贷现象久未解决
4	4	1643.98	2019/01/08 至 2019/09/06	A4区绿地海外滩小区	与高铁轨道的距离太近了，噪音扰民现象严重
5	5	1635.90	2019/02/12 至 2019/12/25	A市A6区晟通城金桥国际	请求增设地铁站点

表14 热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	赞成数	反对数
1	208636	A0007717 1	A市A5区汇金路五矿万境K9县存在一系列问题	2019/08/19 11:34:04	我是A市A5区汇金路五矿万境K9县24栋的一名业主，我们小区一开始的定位是一个高端别墅小区，实行人车分流管理，物业也标榜37度五星级服务，到目前为止小区群租房泛滥成灾……	2097	0

1	270070	A909094	A市优山美地物业不管事	2019/3/20 08:25:16	A市A5区体育新城优山美地物业,门禁系统及小区监控系统形同虚设，楼上飞线给电动车充电不计其数，消防通道每天拥堵严重，物业园林无人照顾，小区水景如同臭水沟。请相关部门予以追究责任。	0	0
…	…	…	…	…	…	…	…
2	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	书记先生：您好！我是梅溪湖金毛湾的一名业主，和其他业主一样因为当初金毛湾的承若学校都是金毛建的，果断买了梅溪湖金毛湾。楼盘当时承若小学配套周南小学或实验三小，中学配套周南中学或西雅中学。2018年4月16号学区划分没有金毛湾……	1762	5
2	221996	A00080850	A市博才长房云时代小学迟迟不开学，上万业主心急如焚	2019/1/4 11:36:26	尊敬的市领导：我是长房云时代的一名业主，当初为小孩能就近接受义务教育，免除家长和小孩奔波之苦，在长房云时代开发商和高新区教育局的肯定2018年能小学开学的答复下，购买了……	28	0
…	…	…	…	…	…	…	…
3	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	尊敬的胡书记：您好!A4区p2p公司58车贷，非法经营近四年。在受害人要求下，于去年8.20立案侦察，至今已6个月整。未发一字立案公告和案件进展财产处置通报……	821	0
3	217032	A00056543	严惩A市58车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	胡市长：您好！西地省展星投资有限公司设立58车贷https://baidu.com/亿。2018年8月6日，58车贷爆雷，其法定代表人、大股东苏纳和董事长邢明向（化名邢ze）（夫妻关系）外逃美国……	790	0
…	…	…	…	…	…	…	…
4	263672	A00041448	A4区绿地海外滩小区距长赣高铁最近只有30米不到，合理吗？	2019/9/5 13:06:55	您好，近日看到了渝长厦高铁最新的红线征地范围以及走向经过，其经过北三环的地方紧挨着绿地海外滩小区二期，我测算了一下距离，最近的位置只有30米不到，于1988年颁布的国家标准gb8702-88《电磁辐射防护规定》……	669	0

4	216316	A00097196	A4区绿地海外滩二期业主被噪音扰得快烦死了	2019/9/6 10:16:27	我们是A市A4区绿地海外滩二期居民，2019年8月看到新闻，A市至赣州高铁可研评审会已在A市成功召开并初步确定线路方案（下图所示）,高铁线路将会紧挨着现有的石长铁路线跨过月亮岛。按照当前的高铁规划线路，北三环北边附近的受影响的楼盘有……	669	0
…	…	…	…	…	…	…	…
5	232217	A909094	希望A市地铁能在A6区晟通城或者金桥国际设个站点	2019/9/151 7:11:27	随着A6区逐渐开发与A市的对接，金桥国际和晟通城附近小区越来越多，居住的人和工作的人也越来越多，希望有关领导能考虑下将地铁能延长至附近正好至A市高铁西终点站……	0	0
5	214147	A00053191	A市地铁能否在A6区晟通城或者金桥国际设站？	2019/9/11 17:06:37	随着A6区逐渐开发与A市的对接，金桥国际和晟通城附近小区越来越多，居住的人和工作的人也越来越多，希望有关领导能考虑下将地铁能延长至附近正好至A市高铁西终点站……	0	0
…	…	…	…	…	…	…	…

4.7 热点问题挖掘最终框架

热点问题挖掘模块的整体构建遵循的是一个把逆向思维以及合理性对比择取的准则综合考虑的思路，其详细的构建步骤如下：

- 输入附件3的留言数据集作为文本处理对象
- 文本清洗处理：对所有留言都进行去重、去除空格、标点符号、乱码等的清洗处理；选用jieba分词方式进行分词及对应的去停；
- 形成关于全体留言详情的词典，并将每条留言详情均转化为稀疏向量形式，从而可形成一留言文档矩阵X；
- 通过LSI模型算法将上述的留言文档矩阵降维后形成一新的矩阵 D_K^T；
- 对 D_K^T 中任意两个列向量进行空间余弦值的计算即可衡量每两条留言间的相似度是多少，也即可以形成全体留言间的相似度矩阵simM（存为simM.xlsx）；
- 确定聚类原则为聚类数是500的谱聚类，从而将全体留言进行分类聚集；
- 定义如下公式的热度值评价指标（详细各变量的定义可查阅上文）：

$$R_i = C_i \times \sum_{j=1}^{n_i} [P_{ij} \times (1 + T_{ij} - \mathrm{F}_{ij})]；$$

- 算出各分类别的热度指数，并将其顺序排序；
- 取排名前五的热点问题进行对应的汇总输出，形成热点问题表.xlsx和热点问题留言明细表.xlsx。

五、答复意见评价

5.1 答复意见评价背景

在4.1中我们提及到一个合格的“智慧政务系统”是可以对政府的管理水平起极大推动作用的。而政府的管理我们可以认为是一个双向的过程，除对外考虑处理民生事情之外，我们还应考虑对内进行的绩效考核，或者说满意度评价。

面对群众通过各种各样方式所反映的各类问题，对相关负责部门的办公人员处理群众问题的效果进行较为客观的评价，这是政府内部部门需要密切关注的一个问题，如若发现某些部门处理问题的效率偏低，群众的反响不好，则应及时让其进行内在整顿；当然，如若在相关考核中部门人员或其整体表现优异，那可以对其进行一定的奖励。针对平台系统所收集的群众留言及其对应的答复意见，本文将从以下几个方面来评价：

（1）答复的相关性

答非所问是粗政办公的一个典型。一段合格的意见回复必须是和问题相关，直戳重点的；

（2）答复的完整性

也可称之为规范性。在官方的意见答复中，理应体现出行政办公人员的专业素养。用词是否规范、答复步骤的逻辑是否清晰是一个很好的评判标准；

（3）答复的可解释性

也可称之为合理性。既尝试去分析政府部门所提出的相关解决方案是否可行、是否有效，也包括去商讨政府部门针对目前所存在问题而进行的原因分析是否合理、是否能让人民群众信服；

（4）答复的及时性

群众反映的社会问题对于老百姓本身来说往往是迫需解决或是沉积已久仍未解决的问题。作为政府机构理应急民之所“急”，解民之所“想”。因此，相关的问题解疑的答复以及解决方案的提出必须尽可能的迅速。从提出问题到得到相关部门的答复所跨的时长如何，这是一个很关键的参考标准。

5.2 定性指标量化以及满意度指标构建

为了较为公正地去判断政府部门工作人员的留言答复工作的质量，一个对整个部门的工作人员来说均较为相对客观公正的评价指标的建立是必不可少的，有了它之后我们才能将某个工作人员的月度、季度、半年度及年度工作的质量效果量化，从而能初步对工作人员之间的工作效能进行优劣排序，为后续的评奖评优以及判断部门人员素质是否需要进行内在整顿提供直观的数值参考标准。此外，把答复意见的“质量”量化后还方便政府部门之间进行相关的部门对比，达到一

个相互监督的效果。

在我们所构建的“智慧政务系统”里，我们把这样的一个数值型量化指标称为满意度评价指标S。而在构建满意度评价指标时，我们所采取的思路是先将每条留言答复意见的相关性、完整性、可解释性和及时性进行定量化，然后选用一定的权重评价方法来对这四个数量化后的性质进行综合衡量，从而得出每条留言答复意见评价的满意度得分S。从而也就能把属于某位工作人员或某个政府部门的满意度得分值进行一定的加总，为后续的评优、划及（及格）、和对比提供数值参考标准的客观支持。

5.2.1 相关性、完整性、可解释性和及时性的量化原理和过程

留言用户	答复意见	答复时间
A00045581	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于2019年4月10日至4月27日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5月5日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。**2019年5月9日**	2019/5/10 14:56
A00023583	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋湖街道高度重视，立即组织精干力量调查处理，现回复如下：您反映的为潇楚大道西线道路工程项目，该项目位处于坪塘老集镇，目前正在进行土方及排水施工。因该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，且换填后还有三趟雨污水管道施工，施工难度较大，周期较长。加之坪塘集镇原有管线、排水渠道较多，需先处理管线和渠道才能进行道路施工，且因近期持续雨天，为保证道路施工质量，需在晴好天气才能正常施工。目前该项目已完成75土方及50排水，预计今年8月底将完工通车。感谢您对我们工作的关心、监督与支持。**2019年4月29日**	2019/5/9 9:49:10
A00110735	网友“A00110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反映的问题交由市房屋交易管理中心办理。现将相关情况回复如下：按照《A市人才购房及购房补贴实施办法（试行）》第七条规定：新落户并在A市域内工作的全日制博士、硕士毕业生（不含机关事业单位在编人员），年龄35周岁以下（含），首次购房后，可分别申请6万元、3万元的购房补贴。“首次购房”是指在A市限购区域内首次购买商品住房（含住宅类公寓）。因此，如购买商业性质公寓（非商品住房），则不可申领购房补贴。以上情况，望您知晓和理解。如您还有疑问，建议可拨打市房屋交易管理中心咨询电话0000-00000000详询。特此回复！**2019年4月30日**	2019/5/9 9:49:42

图29 给予部分用户留言的答复意见及时间的示例图

在对四个关键性质进行相应的量化之前，首先对附件4进行简单的浏览，以了解其文本格式即留言回复的一些细节。

如上图所示，我们通过对附件4所附带的全体留言答复进行粗略的查看后会发现：在大多数答复意见的最后工作人员都会书写上相应留言答复意见的答复时间，而这个答复时间又往往与真正的附件4所自带的留言答复意见时间一列不同。在大多数留言答复意见的最后都出现了工作人员的答复时间，这说明了工作人员在进行相关的留言答复时原本就可能遵循着某种规范，而这种规范与上所述的完整性（规范性）是密切相关的。因此，这个位于“答复意见”尾部的留言时间在完整性的量化过程中是需要考虑的；同时，它又往往与真正的附件4所自带的留言答复意见时间一列不同（通常来说比数据集中真正所记录的留言时间早）。其原因可能是因为留言工作的批量回复性所导致的，也有可能是因为相关的转办审批流程而造成的。

在进行相关性、可解释性以及及时性的量化过程中，把留言答复意见的最后会出现的留言时间“隐藏”起来，不去考虑，因为这个时间一般来说是回复意见的工作人员自己填写的，而真正将答复意见呈现给留言者的时间仍以附件4中答复

时间一列为准。而多出一个时间的文字性表述并不能对一个问题起到什么的解释作用，还会影响到字符数，也即会对可解释性的度量产生负效应的影响；但在进行完整性的量化过程中我们却要把其重新考虑进来，此时，将把这个时间看作是回复留言的一个必要部分。利用正则表达式，将这类时间信息从答复意见中分离出来。以下是对四个关键性的性质进行相应的量化：

- **相关性**

在“智慧政务系统”里，为了去描绘留言详情和答复意见之间是否“相关”，我们所采用的量化指标是4.5.2中所介绍的LSI模型处理后的每条留言详情和其对应的答复意见之间的相似度。在经过LSI模型进行留言详情和答复意见的相似度计算后，相关性衡量的量化也就等价于相似度的取值，其在[0,1]之间；此时相似度计算纯粹是答复意见主要内容与留言详情的匹配，时间信息被删除也就去掉了与答复内容没有明显作用的信息，使相似度的计算更准确。

这样做的好处在于一方面可利用留言和答复之间是否相似来直观判断该答复是否“戳到点”上，而采用LSI模型的目的在于对留言详情和答复之间进行“潜语义”分析来判断其是否在同一语言主题，这样也更有理由相信答复意见和对应的留言之间是问答相关的；另一方面的好处在于可使得整个“智慧政务系统”具有整体性，同时提高系统的运行效率。

- **完整性**

在有关于完整性的描述当中我们也把其称为“规范性”，用词是否规范、答复步骤的逻辑是否清晰是一个很好的评判标准。本文采用统一的用词词典来规范每位工作人员的答复用语，用固定的理应具备的答复流程来判断该留言答复是否符合逻辑。本文考虑的五个主要用语部分有：礼貌、收到、解释、解决以及时间提及。也就是说，一个合格且规范的留言答复步骤理应要具备礼貌用语的提及、表示工作人员已经收到并查看了群众所留下的意见留言、有在解释问题所存在的原因、提出最后处理所反映问题的方案以及留言意见所回复的时间。

先规定涉及到五个小部分的关键用词分别是什么，形成一个关键词的词典，通过检索一段留言答复中是否具备该小部分的用词来判断该留言是否能得到该小部分的得分（每个小部分都是用1分来衡量）。考虑到每位工作人员的语言表达习惯不同，且尽量使得用词规范统一化，我们尽量的把相关步骤的用词扩展，形成了如下的规范用词词典：

表15 规范用词词典

	表示礼貌	表达收到	表达留言解释	表达方案解决
规范用词	您好 谢谢（您） 你好 感谢（您）	收到 收悉 查看	澄清 反映 因为 调查 核实	回复 答复 可以

而在表达时间提及方面正如前所说的我们是通过检索留言答复意见的最后是否会出现留言回复时间来作为该段留言回复意见的规范结束程序步骤来判定这1分能否取得。为了与相关性得分的取值相对应，本文在计算完整性指标的得分之后会进行规范化，使完整性分数介于0到1之间。

经过上述所定义的完整性指标量化的原理部分，我们可成功地量化其在[0,1]之间。这量化部分的缺点在于一方面可能逻辑步骤思考的不够完善，可能不只有现考虑的5个小部分能反映其逻辑性，还应包括更多的部分；另一方面在于工作人员的语言表达具有多样性，可能他（她）已经反映了某个所要求的小部分，但其用词不在规范用词里面。但这样做的完整性指标量化当然也有其所具备的优点，政府人员可以参与制定和补充规范用词。

- **可解释性**

在可解释性的阐述当中也可以把它理解成是合理性，即尝试去分析政府部门所提出的相关解决方案是否可行、是否有效。但往往解决方案的这些合理性判断需要的是专业人士的评审，一个解决方案在某些情况下它是合理可行的，某些情况下它却是对应不上的，考虑难以周全。因此，在这情况下我们采用另外一个较为简单但又具备合理性的评判标准：答复意见的字符数。

我们所采用的可解释性衡量的数值取值是一个0-1变量，即答复评价的字数超过某一临界值我们就把其可解释性的取值定义为1，如果低于该临界值我们就把该段留言答复的可解释性取值定义为0。经过对总体留言的了解，本文设定这个答复意见的字符数应至少不低于22个字符才能成为可解释性的最低基准。

- **及时性**

及时性的衡量是一个客观的标准，本文利用附件4所给出的留言时间一列和答复时间一列为依据，来进行及时性指标的计算。为了体现及时性量化步骤的客观性和有根据性，本文根据相关政府信访部门的网站所给出的信访工作的留言时间回复要求，决定采用如下规则：“一般性投诉问题在5个工作日内办结回复，重要投诉问题在10个工作日内回复；情况特殊的，经领导同意后可延长到20个工作日不能按期回复的，责任单位应及时向区政府办说明情况，并在网上公开回复说明。对逾期未办，又未说明情况的，将予以通报。”

留言时间与答复时间之间越接近越好，也就是越相近的时间跨度得到的权重及时性得分理应越高。具体的及时性量化步骤如下：

<1>若以秒数来衡量的答复时间间隔在五天之内，答复的分数将定义在0.6-1.0之间。也即以该条留言发表时间为参考对比点，在留言发出的该秒立刻回复即得满分1分；如若留言回复的时间与留言发表的时间以秒数来衡量其间隔刚好为5天，那么其得分即为0.6分；在0—5天之内的秒数回复的话分值依照0.6分到1分来进行平均插值处理；

<2>若以秒数来衡量的答复时间间隔在五到十天之内，答复的分数将定义在0.3-0.6之间，分数平摊的原理和<1>一样；

<3>若以秒数来衡量的答复时间间隔在十天到二十天之内，答复的分数将定义在0.1-0.3之间，分数平摊的原理和<1>一样；

<4>若以秒数来衡量的答复时间间隔在二十天到三十天之内，答复的分数将定义在0-0.1之间，分数平摊的原理和<1>一样；

<5>若以秒数来衡量的答复时间间隔超过三十天，我们将判定该留言回复为滞后留言回复，直接判定该条留言的及时性得分为0。

上述及时性得分的量化步骤能很好的体现出不同的时间间隔理应得到不同权重的及时性得分奖励的这一特点。这种得分间隔处理也和信访回复章程的规范相对应，如果一段回复越到后面，越拖沓才回复，该段留言理应被倾向于判断为较差的留言回复，也即该留言在及时性量化得分上越迟回复则越难得高分。

5.2.2 权重评价方法的确立

- **量化得分矩阵X**

根据以上分析将每一条留言意见答复的相关性、完整性、可解释性以及及时性均进行了数值的量化，并成功地把其取值都限定在[0,1]之间，这样做的好处在于从一开始就尽量减少性质之间因为量纲不同而导致的影响。当然，为了进一步把量纲之间的差异消除，我们在计算完全体留言意见的四个性质得分后，均统一对四大部分的得分进行标准化处理，并形成了如下的得分矩阵X。

X	留言编号	相似性X1	完整性X2	可解释性X3	及时性X4
	2549	0.380394012	1	1	0.20451
	2554	0.029462781	0.4	1	0.194799
	2555	0.411958337	0.4	1	0.195127
	2557	0.316862136	1	1	0.195586
	2574	0.571428478	0.6	1	0.214003
	2759	0.06099375	0.4	1	0
	2849	0.312764049	0.4	1	0
	3681	0.252798557	0.8	1	0.085215
	3683	0.181554496	0.6	1	0.224649
	3684	0.337009311	0.6	1	0.224793
	3685	0.115534268	0.6	1	0
	3692	0.480847031	0.6	1	0
	3700	0.2154921	0.6	1	0.222214
	3704	0.05610545	0.8	1	0.351849
	3713	0.121045507	0.6	1	0.245554
	3720	0.431268424	0.6	1	0
	3727	0	0.8	1	0.45031
	3733	0.44883281	0.6	1	0.278069
	3747	0.150170222	0.8	1	0.177279
	3755	0.087959915	0.6	1	0.598539
	3756	0.286100596	0.6	1	0.598748
	3760	0.316567421	0.6	1	0.181358

图30 得分矩阵X的部分截图（黄色部分为矩阵的主体，其余部分为文字的说明标注）

量化得分矩阵X的维数为**2816×4**，行数2816代表了附件4包括的全体留言意见回复共拥有2816条，列数4代表着关键的量化性质有四个。完整的得分量化矩阵可查看附件中的“量化表.xlsx”。其中，用 X_{ij} 来表示第 i 条留言意见回复的第 j 个性质得分。

- **熵权法**

为了客观的确定四个性质的权重，本文采用熵权法作为其权重确定方法。熵权法是一种客观的赋权方法，其基本思路是根据指标变异性的大小来确定客观权重。在具体的使用过程中，熵权法根据各指标值的变异程度，利用信息熵计算出各指标的熵权，再通过熵权对各指标的权重进行修正，从而得出较为客观的权重。

它与层次分析法以及模糊综合评价等权重确定方法相比，前者是从数据本身所具备的差异性出发，是一个客观的权立过程；而后者则依赖于相关专家的层打分，那是一个主观性很强的过程，在没有可靠的专家团体的支持下，本系统理应选取可客观确立指标权重的方法。熵权法的使用步骤分为两步：

①确立指标的熵 E_j

$$E_j = -\frac{1}{\ln(n)}\sum_{i=1}^{n}\left[\frac{X_{ij}}{\sum_{i=1}^{n}X_{ij}}\ln\left(\frac{X_{ij}}{\sum_{i=1}^{n}X_{ij}}\right)\right], j = 1,2,3,4 。$$

原则上说应先把矩阵X所对应的元素先经标准化处理后再使用，但我们在量化矩阵X的介绍部分已经提及到X矩阵原本就是已经经过标准化处理后的矩阵，因此在这里可以直接使用相应的 X_{ij} 来进行计算；

②由指标的信息熵E_j来确定其对应的权W_j

$$W_j=\frac{1-E_j}{\sum_{j=1}^{4}(1-E_j)},j=1,2,3,4。$$

经计算,得到权值向量W：

$$W=[0.26366774\quad 0.16986976\quad 0.05150751\quad 0.51495499]$$

在得出有关四个量化性质的权数后我们即可把每条留言意见的量化指标值与其对应的熵权相乘后再累加，从而得出每条留言意见的满意度评价指标值S，即：

$$S_i=\sum_{j=1}^{4}(W_j\times X_{ij}),i=1,2,.....,2816$$

其中S_i是第 i 条留言答复意见的满意度得分值。

5.3 相关留言答复意见评价的评分展示

由于全体留言详情及其意见的篇幅过长，在此我们仅列举满意度得分值在倒数第一及顺数第一的留言来进行简单的评分呈现分析，详细的每条留言意见的满意度及四个性质的量化得分情况将汇总在压缩包中所附带的留言意见评价得分汇总表.xlsx中。

答复意见	满意度S	相似性X1	完整性X2	可解释性X3	及时性X4
感谢您对我们工作的关心、监督与支持。	0.033973951	0	0.2	0	0
网友：您好，您反映的情况已收悉，现将有关情况回复如下：1.物业办和业主代表共同查看物业公司签订的聘用合同，实物业公司是受原业主委员会（现在业委会已经解散）委托聘请，同业主委员会签订合同。2.对业主反应有物业管理人员服务态度差及保洁不到位的情况，我物业办责令物业公司立即进行整改，加强规范管理，完善员工工作考核机制，提升工作人员的服务素质。金盛物业公司表示接到业主投诉和反映后，已经更换了服务态度不好的工作人员及保洁不到位的工作人员，并保证以后会加强规范管理，提升服务质量。3.对于反应的水泥块坠落砸坏业主车辆情况属实，未出现伤人情况，现场调查金盛物业公司已经在事发后的相关地点设置了安全警示标语。物业办督促物业公司要加强巡查，一旦发现外墙有老化剥落松动现象要及时将外墙水泥块敲落。物业公司表示此次砸坏车辆的水泥块经现场调查后发现并非外墙脱落物，已向公安机关报案。4.安装电梯门禁卡属实。物业公司表示2楼是洗脚城，3楼是ktv，进出人员复杂，有偷盗事件发生，装电梯门禁卡的目的是为了保障小区业主的安全，为预防偷盗等治安问题。物业办告知物业公司如果装电梯门禁卡是为了制约不交物业费的业主，这种行为欠妥。物业公司承诺，如果大部分业主反对，将取消电梯门禁卡。5.电梯频繁故障导致业主多次被卡在电梯内，物业办责令物业公司要加强巡查，五方对讲必须保持畅通，并加强对电梯的检修养护，将业主的安全放在第一位，一旦发现业主被卡的情况，必须尽快采取相应的措施，同时建议业主动用维修资金对电梯进行大修，彻底消除安全隐患。6.小区外墙开裂渗水需要维修及防水，因房屋早已过维保期，建议业主动用维修资金，物业公司无权反对业主动用维修资金对外墙进行维修，物业公司有义务协助处理。7.物业办督促物业公司采取措施加强车位的规范管理，限制外面车辆随意进入小区。物业公司表示一直在尽力协调停车问题，也将继续加大力度对停车问题进行规范管理。现场调查有业主反映停车问题的确比以前有所改善，例如以前小门口停满外面的车辆，影响小区内业主出行，现在在物业公司的管理协调下得到了很大改善。8. 对于2楼洗脚城3楼金色豪门ktv声音嘈杂影响业主的正常生活，物业办告知，物业公司虽无执法权，但有协调处理的责任，如果物业公司协调无果，噪音扰民建议拨打12369向环保部门进行投诉。物业公司表示多次进行协调，并要求ktv加装隔音材料。物业办建议物业公司协调时带上业主代表。据调查，现ktv已经停业。9.因小区原业委会已经解散，建议小区业主尽快申请重新筹备选举出业委会，对小区的日常事务，公共收入等进行接管，并对物业公司的日常工作进行监督。2019年11月8日网友您好：您反映的问题已收悉并交有关单位调查核处，如有情况将及时回复！2019年11月6日	0.87043821	0.52946	1	1	0.98933

图31 满意度得分值在顺数第一及倒数第一的留言意见列举图

对于满意度得分值位于倒数第一的留言例子[满意度得分S仅为：0.034]，我们发现其留言意见答复极为简短(可解释性得分为0)且形式客套，单纯表达了对群众的感谢之意（完整性得分为0.2）而对实际的问题没有做出任何有针对性的回应(相关性得分为0)不仅如此，该留言意见还是一时间跨度超过一个月的滞后回复(及时性得分为0)；

而对于满意度得分值位于顺数第一的留言例子[满意度得分S达到：0.870]，我们发现其留言意见答复相当详细(可解释性得分为1)且逻辑步骤清晰（完整性得分为1）。此外，通过深究该留言意见的内容，我们会发现其中所涉及到的一些原

因分析以及方案提议还是相对在点上的，即有一定的针对性（相关性得分为0.529），且该回复在5天的理想时间间隔跨度之内（及时性得分为0.989）。

因此，通过上述两个例子的查看分析，本文所构建的答复意见评价模块具有一定的合理可靠性的。

5.4 答复意见评价改进分析

答复意见评价模块的核心处理步骤有两个，分别是四大性质的量化以及量化后的权重评价方法选择，如下为这几个方面有可能存在的改进思路的分别列举：

<1>相关性以及可解释性的量化改进：

在本模块中，对于相关性的量化我们所采用的方法是在热点问题挖掘模块中所提及到的LSI模型来测度留言详情和对应留言意见的相似度，并以该相似度来作为留言意见的相关性得分，但相似并不一定代表着相关。如当留言详情和其对应的留言意见一模一样时，此时两者间完全相似，留言意见的相关性被定义为1，但此时也可能工作人员只是把群众的留言问题复述了一遍，本质上根本没有对问题进行任何的意见提出，此时的留言意见回复是完全失职的。而在可解释性的量化方面我们只是以字符数的多少来判断其得分情况如何，但字数多并不代表着该意见回复就是一个可行、有效或者合理的建议。它有可能是文不对题，避重就轻的。

因此针对上述存在的问题，这里所提出的建议思路是尝试建立一个有标签的监督学习模型，并把相关性及可解释性的取值离散化来作为对应的二级标签体系并进行有监督学习。

<2>完整性的量化改进：

可把系统所规定的逻辑回答部分：礼貌用语、收到、解释、解决、时间提及，及其所对应的关键词词典均作拓展处理，而拓展的过程可邀请不同部门的工作人员共同参与，以达到集思广益的效果；

<3>及时性的量化改进：

可改变不同时间段中分值分配的方式，从原本系统所采用的平均插值处理改为一个用以往该时间段的得分数据所拟合出来的函数f(x)去衡量计算的方式。当然这个f(x)也可由自己根据相关信访章程所强调的不同时间段的重要性来选用一些符合该章程特征的函数；

<4>权重评价方法的改进：

在系统的模块构建过程当中，由于缺乏相关专家团体的意见支持，我们采用了从数据本身特征所出发的熵权法来确定指标间的权重占比。也即是说当我们可以得到相关专家团体的支持时，像层次分析法、模糊综合评价方式等主观性权重测定方法也可以被考虑进来，此时的权重分配也因专业人员参与而更具说服力。

六、“智慧政务系统”最终框架

本文所构建的“智慧政务系统”主要由三大模块组成，分别是：群众留言分类模块、热点问题挖掘模块以及答复意见评价模块，三大模块的目的作用分别是：

群众留言分类模块：可把从相关网络问政平台中收集到的群众留言（附件2）按照一个确定已有的标签体系（附件1）来进行程序的标签自动划分类，并输出相关验证集的标签预测结果：“predicted_label”。该模块的实现使得大量庞杂的留言问题可以分门别类的投入到相关的部门进行处理，不仅减轻了工作量，让专业人员处理更加专业的事务，同时也提高了政府部门的办事效率，将计算机技术应用到实际中，加快创建服务型政府；

热点问题挖掘模块：可对从相关网络问政平台所收集到的群众留言（附件3）进行自动程序热点挖掘（热度指数及相应的性质量化指标均由系统本身自定义），并输出：“热点问题表.xlsx”和“热点问题留言明细表.xlsx”。该模块的实现为政府了解群众事务提供了更加快速的途径，使其可以更加及时地解决群众问题，而不是后知后觉；

答复意见评价模块：对于从相关网络问政平台中所收集到的群众留言，政府部门工作人员一般都会在一定时间内尽快地对这些留言给予相应答复（附件4），而本模块能达到的效果是可根据系统本身所自确定的性质量化规则及权重评价方法来对工作人员的留言意见进行满意度S的打分，并输出：“留言意见评价得分汇总表.xlsx”和“留言答复意见评价方案.pdf”,从而可进一步把得分数据用于政府部门日常的绩效考核中去。

系统的框架搭建即分别由三个模块的程序步骤所组合而成，即如下图的展示：

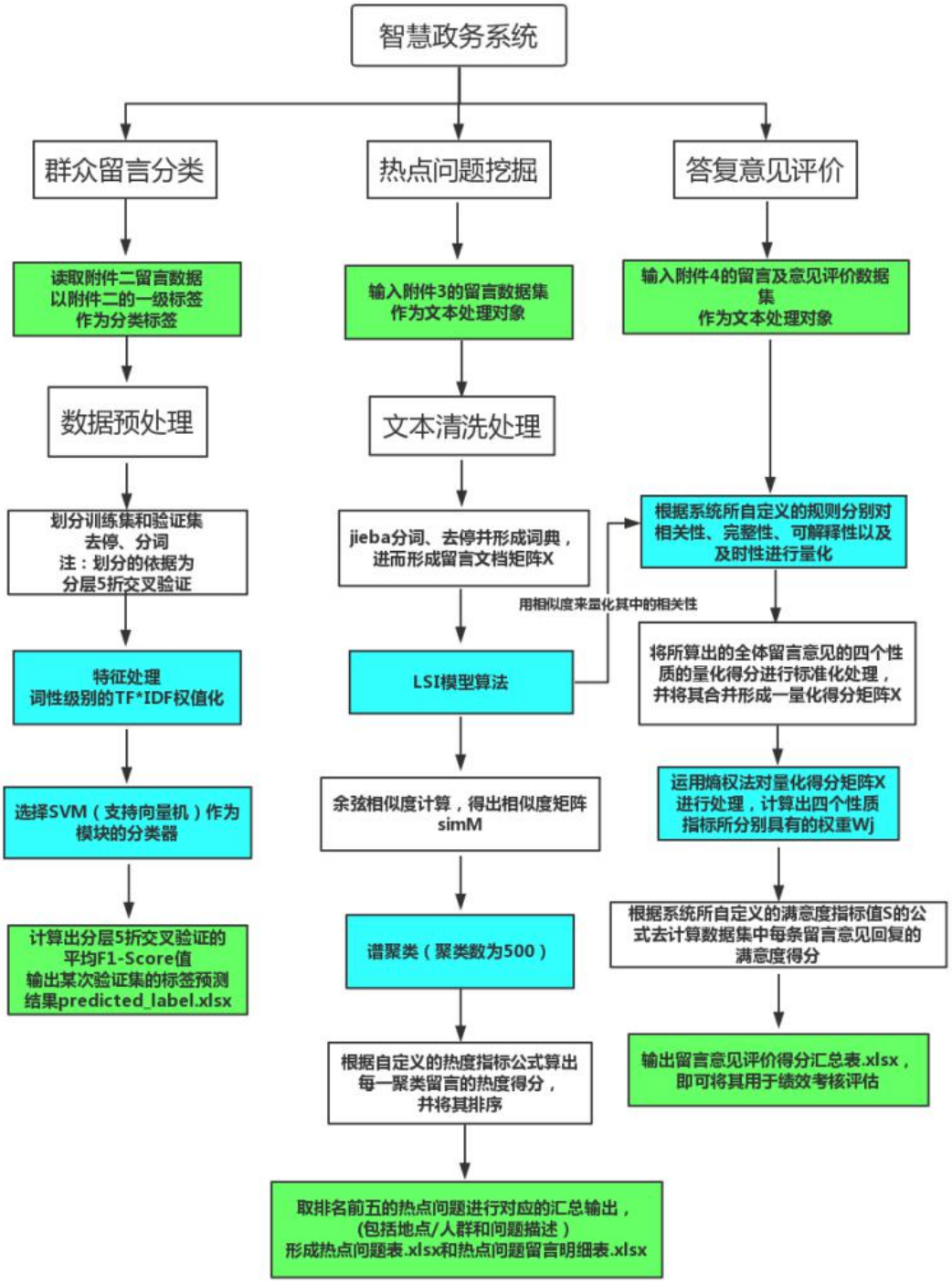

图32 “智慧政务系统”的最终框架搭建流程图

七、 参考文献

[1]Yoon Kim.Convolutional Neural Networks for Sentence Classification[J].2014
[2]张振豪, 过弋, 韩美琪, 王吉祥. 基于关键词相似度的短文本分类方法研究[J/OL]. 2020, 37(1).[2018-11-05].http://www.arocmag.com/article/02-2020-01-001.html
[3]Ye Zhang,Byron Wallace.A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.[J].2015
[4]Nal Kalchbrenner,Edward Grefenstette,Phil Blunsom.A Convolutional Neural Network for Modelling Sentences.[J].2014
[5]Chunting Zhou,Chonglin Sun,Zhiyuan Liu,Francis C.M.Lau.A C-LSTM Neural Network for Text Classification.[J].2015
[6]殷风景.面向网络舆情监控的热点话题发现技术研究 [D].国防科学技术大学研究生院.2020.11.
[7]文本分类的数据预处理[转][OL]
https://www.iteye.com/blog/forever1220-2097568.
[8]Python 中文分词工具大合集：安装、使用和测试[OL]
http://www.52nlp.cn/python.
[9]基于 TFIDF 实现文本分类，并比较词集模型与词袋模型的分类效果[OL]
https://blog.csdn.net/weixin_43629813/article/details/103846646.
[10]手把手教你在 Python 中实现文本分类（附代码、数据集）[OL]
https://blog.csdn.net/sinat_38682860/article/details/80421697.
[11]朴素贝叶斯分类:原理[OL]
https://blog.csdn.net/qiu_zhi_liao/article/details/90671932.
[12]支持向量机（SVM）——原理篇[OL]
https://zhuanlan.zhihu.com/p/31886934.
[13]随机森林算法原理[OL]
https://blog.csdn.net/hello_zybwl/article/details/87826073.
[14]CS231n——机器学习算法——线性分类（上: 线性分类器）[OL]
https://blog.csdn.net/weixin_38278334/article/details/82831541.
[15]几种常用交叉验证（cross validation）方式的比较[OL]
https://blog.csdn.net/qq_37466121/article/details/87916870.
[16]几种交叉验证（cross validation）方式的比较[OL]
https://www.cnblogs.com/ysugyl/p/8707887.html.
[17]文本分类算法综述[OL]
https://zhuanlan.zhihu.com/p/76003775.
[18]TextCNN 文本分类（keras 实现）[OL]
https://blog.csdn.net/asialee_bird/article/details/88813385.
[19]使用 Keras 进行深度学习:（六）LSTM 和双向 LSTM 讲解及实践[OL]
http://www.tensorflownews.com/2018/05/04/keras_lstm/.
[20]Keras with R (RNN)[OL]
https://blog.csdn.net/qq_24834541/article/details/81087093.
[21]文本相似度计算-文本向量化[OL]

https://www.cnblogs.com/huangyc/p/9785420.html#_label1_1.
[22]Python+gensim-文本相似度分析（小白进）[OL]
https://blog.csdn.net/Yellow_python/article/details/81021142.
[23]文本相似度算法[OL]
https://www.cnblogs.com/liangxiaxu/archive/2012/05/05/2484972.html.
[24]奇异值分解（SVD）详解[OL]
https://blog.csdn.net/wangzhiqing3/article/details/7446444#comments).
[25]LSA(LSI)算法简介[OL]
https://www.cnblogs.com/mlfighting/archive/2013/04/21/3034337.html.
[26]LSI/LSA 算法原理与实践 Demo[OL]
https://blog.csdn.net/qq_16633405/article/details/80577851.
[27]scikit-learn 中的无监督聚类算法[OL]
https://www.cnblogs.com/xc-family/p/11006525.html.
[28]谱聚类（spectral clustering）原理总结[OL]
https://www.cnblogs.com/pinard/p/6221564.html.
[29]政府网站网上留言办理管理办法 [OL]
https://wenku.baidu.com/view/fbcf73ed6294dd88d0d26b26.html.
[30]指标权重确定方法之熵权法[OL]
https://blog.csdn.net/qq_32942549/article/details/80019005.
[31]熵权法——指标权重确定 [OL]
https://wenku.baidu.com/view/010703d4240c844769eaee6f.html.