基于文本挖掘的“智慧政务”系统

摘 要

随着网络问政平台的逐步发展，各类民意文本数量不断升高，这将成为政府了解民义、汇聚民智的重要渠道。然而由于人工划分留言与整理热点问题使得工作难度大、耗费时间长、准确率难以保证，政府相关部门将面临的负担加重。因此针对大量文本留言分类与热点问题提取，本文利用已给的数据，基于 python 数据分析对以上问题进行分析与处理，得出相应结果，从而为政府工作提供绵薄之力。

首先，本文对群众留言的数据进行详细分析。对“留言详情”进行预处理后，提取文本特征信息。接着用 TF-IDF 模型进行文本特征向量化，对多个多分类模型进行 F-Score 评价，选择效果最好的 SVM 模型对留言内容进行预测，预测出留言内容对应的一级分类标签，从而达到分类效果。分析预测结果得知：城乡建设与环境保护是群众民生关注的一个重点；其次随着人民生活质量水平的提高，商贸旅游方面变得越来越受人们重视。

其次，为了获取热点问题，本文采用 BOW 词袋模型和 TF-IDF 模型对文本进行处理，再通过 similarities 函数对“留言主题”进行相似度分析。将相似度大于 0.2 的文本视为同一类问题。对各个问题留言的频率热度、浏览热度、时间跨度进行定义并归一化处理，赋予一定的权值，得到相应问题的热度指数，从而获取热度指数排名前五的热门话题。详细结果数据具体请参考“热点问题表”与“热点问题明细表”。

最后，就相关部门对留言的答复意见，我们从相关性、完整性、及时性三个角度对答复意见的质量进行评价。根据 similarities 函数和余弦相似度对答复的相关性进行分析，将相似度大于 0.3 的视为答复相关；根据答复前后文是否有敬词（如：您好、尊敬、感谢等），将均有敬词的视为答复完整；根据留言与答复的相隔时间，将相隔小于 5 个工作日的视为答复及时，由此得出一套答复意见的质量评价方案。

关键字：文本挖掘、 TF-IDF 模型、BOW 词袋、文本相似度、支持向量机、热度指数

Abstract

With the gradual development of online political platform, the number of various public opinion texts keeps increasing, which will become an important channel for the government to understand the people's justice and gather their wisdom. However, due to manual division of messages and sorting out hot issues, the work is difficult, time-consuming, and the accuracy is difficult to ensure, these problems make the government face an increased burden. Therefore, aiming at the classification of a large number of text messages and the extraction of hot issues, this paper analyzes and deals with the above issues based on the given data and the python data analysis, and obtains the corresponding results, thus providing a useful force for the government's work.

First of all, this article carries on the detailed analysis to the populace message data. After preprocessing "message details", the text feature information was extracted. Then, TF-DIF model was used for text feature vector-quantization, f-score evaluation was conducted on multiple multi-classification models, and the best SVM model was selected to predict the message content, so as to predict the corresponding first-level classification label of the message content, so as to achieve the classification effect. The results show that: urban and rural construction and environmental protection is a focus of people's livelihood; Secondly, with the improvement of people's living quality, business tourism has become more and more important.

Secondly, in order to get the hot issues, this article USES the BOW bag model and TF - IDF model to deal with the text, then by similarities function similarity analysis was carried out on the theme "message". Treat text with a similarity greater than 0.2 as the same kind of problem. The frequency heat, browsing heat and time span of each question were defined and normalized, and certain weight was given to obtain the heat index of the corresponding question, so as to obtain the top five hot topics of the heat index. For detailed results, please refer to the "hot issues table" and "hot issues list".

Finally, we evaluate the quality of the comments from three perspectives: relevance, completeness and timeliness. According to the similarities function

and cosine similarity analysis of the correlation of reply, the similarity is greater than 0.3 as a reply; According to whether there are honorific words before and after the reply (such as: hello, respect, thank you, etc.), will be regarded as a complete reply; According to the time interval between the message and the reply, the reply which is less than 5 working days is regarded as timely, and a quality evaluation scheme of the reply is obtained.

Keywords: Text mining, TF-IDF model, BOW, text similarity, SVM, heat index

目　录

表 录

图 录

第一章 问题描述

1.1 问题描述

近年来，互联网技术的不断发展催生出越来越多新型的交互渠道，特别是诸如微信、微博、市长信箱、阳光热线等网络问政平台的开发使得政府更直接更贴近地了解民意、汇聚民智、凝聚民气，但同时，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。

如今，大数据、云计算、人工智能技术的发展使得建立基于自然语言处理技术的智慧政务系统成为社会治理创新发展的新趋势，这可以极大地减少不必要的人员劳动并且提高政府工作的效率。而我们的任务，便是利用收集好的自互联网公开来源的群众问政留言记录和部门答复意见建立相关模型，解决群众留言分类、热点问题挖掘和答复意见评价的问题。

首先，我们需要对附录中的数据进行结构分析，以便规划详细的数据预处理方法，接着我们便根据设计好的方法进行数据清洗和文本处理；其次，我们要就已提取好的文本特征建立分类模型，经过多模型的检验评估，选择最优的一者；第三，我们利用相似度模型进行文本聚类，根据留言的频率热度、浏览热度、时间跨度三个因子以较为合理的权重提出热度指数，从而得出排名前五的热门话题；最后，从相关性、完整性、及时性三个角度出发建立合理的答复评价指标模型和体系，对相关部门对留言的答复意见的质量进行评价。

1.2 论文结构安排

本文共分为五章，各章内容安排如下：

第一章，对论文要解决的问题的相关背景和内容进行描述，并简单介绍解决的思路。

第二章，多角度对给出的文本数据进行结构分析，给出相应的数据预处理方案并实行。

第三章，留言分类模型的建立，通过合理的检验标准进行评估和选择，并对模型进行优化。

第四章，热度评价指标体系的提出。

第五章，综合考虑留言答复的相关性、完整性和可解释性来建立合理的答复意见质量评价方案，并呈现评价结果。

第二章 数据分析及预处理

2.1 数据分析

2.1.1 数据结构分析

由附件 Excel 表格可知，附件 2、3、4 分别有 9210、4326、2816 条数据。附件 2 的每条留言数据记录了“留言编号”、“留言用户”、“留言主题”、“留言详情”、“一级标签”5 个属性，附件 3 较附件 2 多了“留言时间”、“点赞数”、“反对数”3 个属性，少了属性“一级标签”，附件 4 较附件 2 多了“留言时间”、“答复意见”、“答复时间”3 个属性，少了属性“一级标签”。

首先对“留言详情”一栏进行重复值统计，发现附件 2、3、4 分别有重复数据 158、101、33 条，因此存在数据重复的情况。其次，根据对附件 3、4“留言时间”“留言内容”一栏文本数据的分析，发现存在噪声数据，例如留言时间较为久远、留言内容不规范等问题。这些数据对于后续的文本分析工作是无意义的，甚至可能会影响文本处理结果，给我们的模型建立和优化带来一系列不必要的麻烦，故而对这些数据有必要进行一系列的清洗。下面将具有代表性的数据质量问题列于表 1：

表 1 数据质量问题及原因

问题	脏数据例子	原因
重复数据	重复数据	同一用户重复留言等
文本格式	“2019-6-8” “2019/12/5”	/
无意义值	留言内容为“()”	用户留言错误提交等

2.1.2 中文文本分析

不同于英文语言本身的空格使句子分割成一个个具有独立意义的词，中文以字为基本单元进行书写，无法自动完成对词的切分辨识，因而需要我们进行分词工作。同时，即使是切分后的中文文本，也存在大量的标点等特殊符号、停用词、重复词、数字等，对于特殊符号、停用词和部分数字来说，这不是我们想要提取的文本特征，因而可以将它们去除，对于重复词来说，这在一定程度上表明了该词在文本内容上的重要性，因而我们可以根据词语出现的频率来提取文本特征。另外，分类标签若为具体

的文本名称，无法进行数值型运算，不利于后续分类工作的开展，故而有必要将“一级标签”数值化处理。

2.2 数据预处理

根据上述数据分析结果，我们提出了数据处理的主要流程，如图 1 所示：

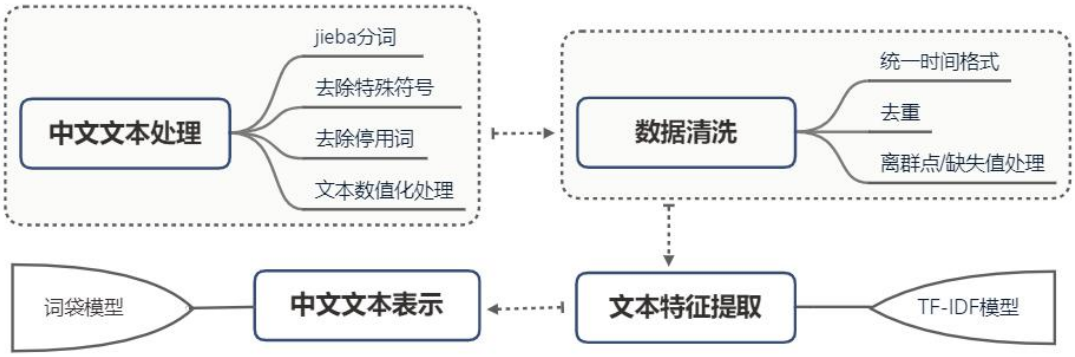

图 1 数据预处理流程图

2.2.1 中文文本处理

（一）分词

本文使用 jieba 库来对中文文本进行切分。

jieba 分词属于概率语言模型分词，支持三种分词模式[1]：Ⅰ.全模式，把句子中所有可以成词的词语扫描出来，但是存在词语歧义问题；Ⅱ.精确模式，将句子最精确地切分开来，适合应用于文本分析；Ⅲ.搜索引擎模式，在精确模式产生的长词的基础上，将长词再次切分，广泛应用于搜索引擎分词。本文则使用精确模式来进行分词。

jieba 分词基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)，动态规划查找最大概率路径，找出基于词频的最大切分组合，从而达到分词的目的[2]。但由于 jieba 分词 HMM 识别新词在时效性上是不足的，且命名实体识别效果不够好，因而需要我们添加一些类如日期、地名、专有名词等词语到 jieba 词库中去。本文使用正则表达式将文本数据中的“x 区”、“x 县”、“x 市”、“x 镇”、“x 小区”、“x 社区”、“x 苑”、“x 年”、“x 号”、“x 月”、“x 日”以及类如“五矿万镜”、“金星北片”等命名实体的词提取出来并加入到 jieba 词库中，再用 jieba 库进行分词得到较为精准的分词效果。

（二）去除特殊符号

本文将中文字符“【{}】”加入到英文字符库 punctuation 中，自定义函数将中文文本数据中的标点等特殊符号去除。

（三）去除停用词

为提高关键词密度和提高计算效率，我们需要将文本数据中的停用词如语气助词、副词、介词、连接词等无明确意义的词去除，本文利用一个存储了大量常见停用词的 stopword.txt 文件，通过建立函数将文本数据中的停用词去除。

（四）文本数值化处理

为了方便后续计算工作，本文利用 factorize 函数[3]将分类的一级标签通过函数映射到一组数值上，其中，相同的标称映射为相同的数字，从而达到用数值代替标称型数据的效果。例如，将六个分类标签“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”分别用 0、1、2、3、4、5、6 来表示，如下图 2 所示：

留言详情	一级标签	category_id
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...	城乡建设	0
位于书院路主干道的在水一方大厦一楼至四楼人为拆除水、电等设施后，烂尾多年，用护栏围着，不但占...	城乡建设	0
尊敬的领导：A1区苑小区位于A1区火炬路，小区物业A市程明物业管理有限公司，未经小区业主同意...	城乡建设	0
... ...		...
领导你好，我们属于未婚生子，但是在2013年已经接受处罚，小孩已上户...	卫生计生	6

图 2 一级标签数值化

2.2.2 数据清洗

（一）统一时间格式

由于文本数据中的时间数据格式不为一致，如“年-月-日”和 “年/月/日”，会给后续热点指数的计算带来不便，故本文将时间数据格式统一为“年/月/日”。

（二）去重

同一个用户可能进行了重复多次的留言，这些数据大部分属性是相同的，不同的只是留言时间这样的相对不重要的属性，并且重复的情况会影响热度指数的计算，在此我们用 drop_duplicates()函数去除重复数据，默认保留重复数据中第一次出现的数据。

（三）离群值、缺失值处理

留言内容为特殊符号的（如“（）”），经过去除特殊符号的处理，将变为空缺值，由于这样的数据数量较少，且没有一个较为标准的内容进行填充，本文考虑将该数据去除。同时由于少数离群值的存在，如 4326 条数据中，仅有一条数据的留言时间在 2017 年，与其他数据的留言时间相隔太久，因此我们考虑将其去除。

2.2.3 文本特征提取

特征选择就是在相关特征中选出最优子集，用于代表非结构化文本，同时降低维度，提高数据处理的效率。首先需要将分词后的文本向量化，可以得到词汇表中每个词在各个文本中形成的词向量，向量中的数值代表词频。但有可能每一条文本中某个“不重要”的词的词频大于某个更为特殊、更具有代表性的词的词频，这将导致特征提取的精确度下降[4]，于是我们采取 TF-IDF 模型来提取文本特征。

TF（x）指词 x 在当前文本中的词频，IDF（x）指词 x 在所有文本中出现的频率，N 代表语料库中文本的总数，N（x）代表语料库中包含词 x 的文本总数，平滑后的 IDF 公式[5]表现为：

$$\boldsymbol{IDF(x) = log\frac{N+1}{N(x)+1}+1}$$

TF-IDF 值为：

$$TF-IDF(x)=TF(x)*IDF(x)$$

TF-IDF 值越大，表示该词的特征越关键。

2.2.4 中文文本表示

中文的文本表示模型是用于生成文本处理算法的处理目标，是将非结构化文本转化成结构化数据的最终实现，本文采用词袋模型。一个语料库中包含了若干个文本语料，将所有的文本预料切割而成的词语装进一个“袋子”，不考虑其词法和语序的问题，即每个词语都是独立的，由此形成词袋。因此对于任何一个文本，都可以从词袋中找到对应的词语来组合，即用向量的形式来表现文本。

第三章：分类模型建立与优化

面对网络问政平台的群众留言，工作人员建立了三级标签体系，但由于目前大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低且差错率高等问题，故本组通过数据挖掘建模的方法，建立相关一级标签模型。

3.1 主要思路与框架描述

如下图 3 所示：

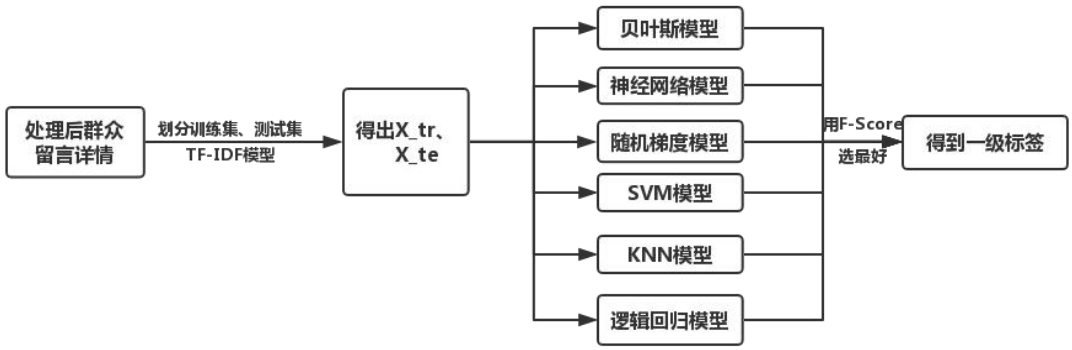

图 3 分类模型建立主要思路图

先用 train_test_split 对处理完的群众留言详情进行划分训练集、测试集，本文采用训练集占 80%，计 7368 条；测试集占 20%，计 1842 条。再利用 TF-IDF 模型对训练集进行训练。通过多个多分类模型对测试集进行预测，通过 f1 值选择结果最好的模型来得到一级标签。

3.2 F-Score 分类评价

F1 分数(F1 Score)，又称平衡 F 分数(blanced F Score),是统计学中用来衡量二分类模型精确度的一种指标，其可以看成是模型精确率和召回率的一种调和平均。

表 2 混淆矩阵

实际类别	预测类别			
		Yes	NO	总计
	Yes	TP	FN	P(实际为 Yes)
	NO	FP	TN	N(实际为 NO)
	总计	P’(被分类为 Yes)	N’(被分类为 NO)	P+N

TP：将正类预测为正类数。

FN：将负类预测为正类数误报。

FP：将负类预测为正类数误报。

TN：将负类预测为负类数。

准确率(precision)：

$$P = \frac{TP}{TP + FP}$$

召回率：

$$R = \frac{TP}{TP + FN}$$

F1 分数：

$$F_1 = \frac{2 * P * R}{P + R}$$

因为本题目是一个多分类问题，故将 F1 的公示更新为：

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

3.3 模型分析

从题意得，这是一个多分类问题。我们要将问题分成 7 大类，因此我们选择贝叶斯模式、神经网络模型、随机梯度模型、SVM 模型、KNN 模型、逻辑回归模型。数据完成预处理后，得到 7223 条数据与 72283 个特征。特征数量会随着训练集、测试集的不

同而改变，我们设置了 random_state 以确保每次的特征不变，我们将数据带入模型进行分析。

此外，为了更好的选择模型，我们导入了 time 库去计算每个模型预测所花的时间。

3.3.1 贝叶斯模型

贝叶斯模型通过概率统计的原理实现。通过朴素贝叶斯模型 MultinomialNB，参数 alpha=1.0，得出最终的运行时间为 0.065 秒，评价指标 f1 值为 0.753.

由于朴素贝叶斯模型假设的特诊之间时相互条件独立的，因此这个假设往往很难达到，特别是在文本当中。很多词语之间存在着相关关系，因此在此使用贝叶斯模型可能效果并不理想。

3.3.2 神经网络模型

人工神经网络是输入值在加权后通过激活函数输出的一个模型。它在模仿人类大脑神经网络处理、记忆信息的方式的基础上，由大量的、简单的处理单元(称为神经元)广泛地相互连接而形成的复杂网络系统，具有高度的非线性性，能够进行复杂的逻辑操作和非线性关系实现。

神经网络中的隐藏层个数 hidden_layer_sizes 我们设置为 10，学习率 learning_rate_init 设置为 0.1，最大迭代次数 max_iter 设置为 800。由于存在 10 个隐藏层和 72283 个输入特征，神经网络的运行速度相对其他模型较慢，最终运行时间为 51.555 秒，评价指标 f1 值为 0.47。其中预测值与真实值之间仍存在 0.415 的误差。

3.3.3 支持向量机 SVM 模型

SVM 是一种广泛使用的二分类模型。根据输入模型的数据特征，支持向量机可简单分为可分线性支持向量机、线性支持向量机和非线性支持向量机。SVM 的目的是通过找到最大边缘超平面来对样本进行分割，最终转化为一个凸二次规划问题来求解。

根据过往的经验，我们将初始惩罚系数设定为 1，loss 函数设定为“hinge”，最大迭代次数为 500，然后将分好词且已经成功转化为词向量的训练集数据输入模型训练，再使用模型去预测测试集里同样已经转化为词向量的数据，根据最终输出与实际标签算出的 f1 得分为 0.912，最终运行时间为 2.845 秒。

3.3.4 其他分类模型

我们还分别使用了随机梯度模型、KNN 模型、逻辑回归模型。得到结果，随机梯度模型运行时间为 0.206 秒，评价指标 f1 值为 0.879。KNN 模型运行时间为 2.333 秒，评价指标 f1 值为 0.818。逻辑回归模型运行时间为 7.631 秒，评价指标 f1 值为 0.881。

3.4 模型选择

对以上模型 F1 评价结果做可视化，如下图 4 所示：

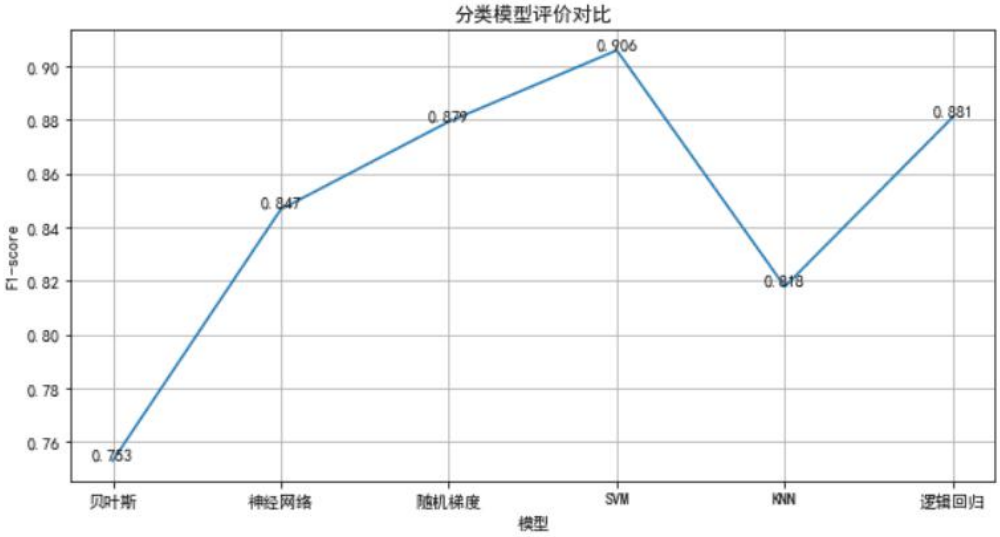

图 4 F1 评价结果对比图

对以上模型运行速度时间做可视化分析，如下图 5 所示：

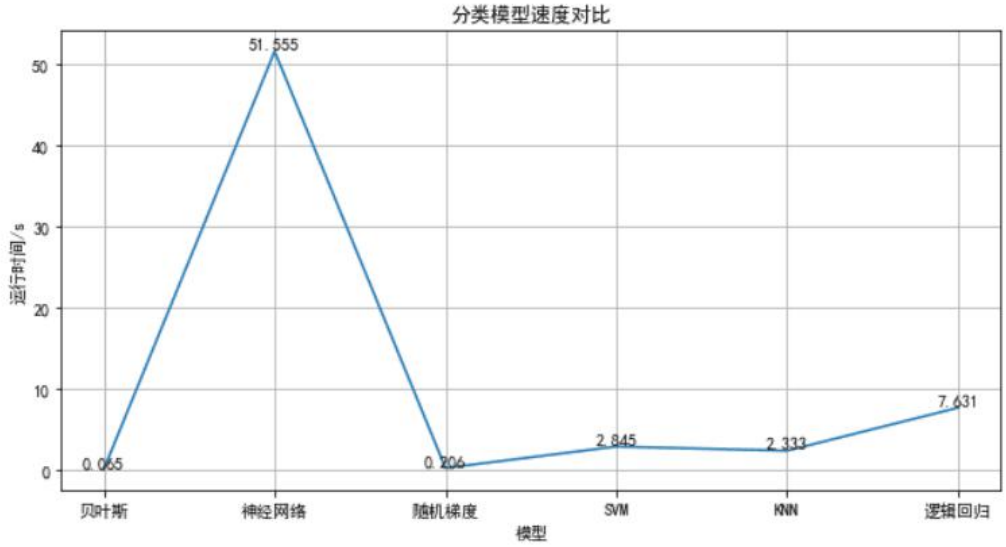

图 5 运行时间对比图

综合两图得出表格，如下表 3 所示：

表 3 各分类模型的 F1 得分及运行时间

模型	F1 得分	耗时(s)
贝叶斯模型	0.753	0.065
人工神经网络	0.847	51.555
随机梯度模型	0.879	0.206
SVM 模型	0.906	2.845
KNN 模型	0.818	2.333
逻辑回归模型	0.881	7.631

从上表可以很直观的得出，SVM 的评价指标 F1 为 0.906，模型运行速度为 2.845 秒，不管是运行速度还是 F1 评价指标都是十分可观的。因此我们选择 SVM 模型作文本题的最佳模型。

3.5 模型参数优化

本组选择 SVM 作为最终的模型，其重要参数有惩罚参数 C 和损失函数 loss。针对该两项参数，本组设计了简单的试验去验证该两种参数的不同组合在验证集上的表现状况，具体情况如下图 6：

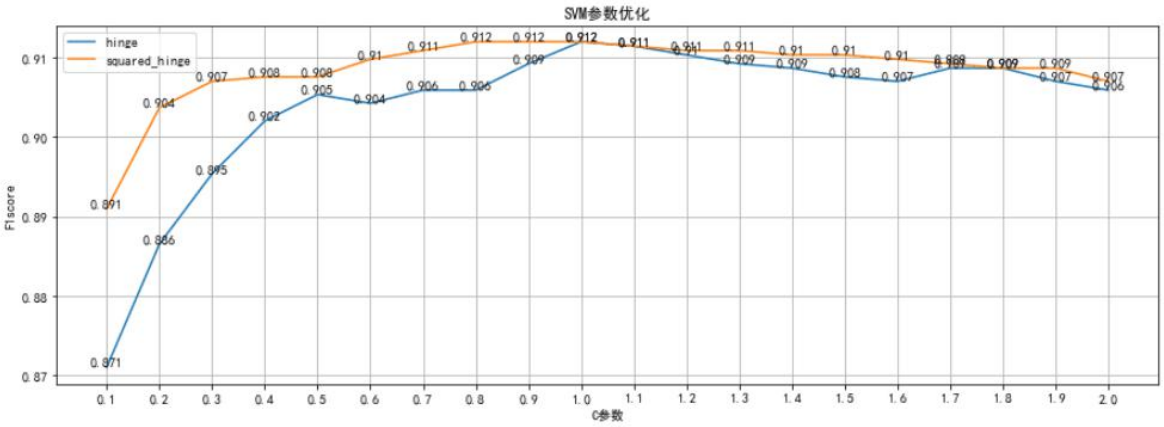

图 6 SVM 模型参数优化

从图中可见，当 loss 函数选择“hinge”且惩罚参数 C 为 0.8、0.9、1.0 时，模型在验证集上的 F1_score 达到峰值 0.912，这里我们将惩罚函数 C 取为 1.0。

3.6 误差项分析

为了进一步分析分类情况，我们引入了 confusion_matrix 函数，通过混淆矩阵更好的看出模型预测的情况。得到的混淆矩阵与对应的热力图如下图 7 所示：

```
[[384   8   5   6   3  12   0]
 [  4 167   0   1   0   1   0]
 [ 15   3  99   2   4   7   1]
 [  4   0   0 293   9   6   1]
 [  4   0   0   3 371   1  12]
 [ 14   2   2   4   3 184   3]
 [  2   1   0   0   8   8 149]]
```

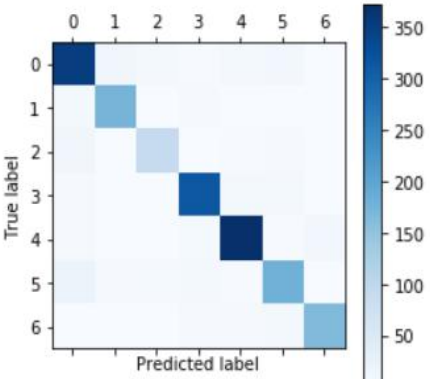

图 7 混淆矩阵及其热力图

通过混淆矩阵的定义，按行相加我们可以发现在 1806 条测试数据中，城乡建设 418 条，劳动和社会保障 173 条，教育文体 131 条，商贸旅游 313 条，环境保护 391 条，卫生计生 212 条，交通运输 168 条。而预测正确的，城乡建设 384 条，预测正确率为 91.9%；劳动和社会保障 167 条，预测正确率为 96.5%；教育文体 99 条，预测正确率为 75.6%；商贸旅游 293 条，预测正确率为 93.6%；环境保护 371 条，预测正确率为 94.9%；卫生计生 184 条，预测正确率为 86.8%；交通运输 149 条，预测正确率为 88.7%。

从混淆矩阵的数值来看，将教育文体、卫生计生错误判断为城乡建设以及将交通运输错误判断为环境保护的数量相对较多。这是由于分词效果不够好和文本未能联系上下文语义而造成的，后续我们将对此进行改进与优化。

3.7 总结

通过多个多分类模型比较，我们最终选择 SVM 支持向量机对群众留言进行一级标签分类，其中参数 loss 函数选择“hinge”，惩罚参数 C 选择 1.0。通过原文本与混淆矩阵，我们可以得出：城乡建设与环境保护是群众民生关注的一个重点；其次随着人民生活质量水平的提高，商贸旅游方面变得越来越受人们重视。政府有关部门应当在这些方面加强管理与发展，共同努力为人民谋幸福。

第四章：热度评价指标体系

热点问题指的是一段时间内群众集中反映的某一问题。在提取和获取热点话题时，一些信息的提取是必要的，其中包括特定的时间、特定的地点或人物。合理的热度评价指标是热点话题提取的关键。本章将主要就上述几个方面的问题，对群众留言主题进行分析和探索，获取热点问题。

4.1 主要思路与框架描述

主要思路框架如下图 8 所示：

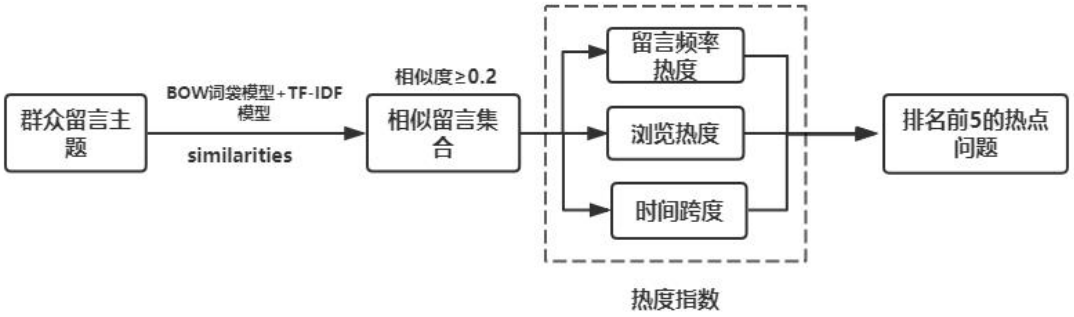

图 8 热度评价指标体系建立主要思路图

首先利用 BOW 对处理后的群众留言主题文本建立词袋模型，得到其二元组的向量表示，在此基础上使用 TF-IDF 模型进行建模，以此再利用函数 similarities[6]计算各文本之间的相似度，保留相似度大于 0.2 的文本，形成相似留言集合。热度指数的三个影响因子：浏览热度、留言频率、时间跨度。通过热度指数评价分数前五的作为热点问题。

4.2 获取相似留言集合

本题运用所有群众分词后的留言主题制作词料库，采用 BOW 模型建立词袋，利用 TF-IDF 模型计算 TF-IDF 值。在 TF-IDF 的基础上利用 SparseMatrixSimilarity 对稀疏矩阵进行相似度检索。将附件 3 中的 4326 条留言主题作为训练语料，把每一条留言主题作为测试语料分别带入模型计算。保留文本之间相似度大于 0.2 的索引，因此一个索引对应着一类问题。

由于文本数据量大，程序运行速度较慢，于是对程序进行优化改进。由于问题索引也对应着文本索引，因此将前面做过相似度计算的文本清空，从而在后面不需要再

重复运算一遍。例如问题索引 1，对应的相似文本索引 1i=[1,442]，那么我们就清空索引为 442 的文本，这样后面就无需重复。可以看到问题索引 2、4 等等，它们对应的相似文本索引列表都非常多，所以这样的优化操作大大减小了程序运行时间。同时我们只打印相似文本索引列表长度大于 1 的值，去除了仅与自己文本相似的文本内容。这也是下图表 1 中出现空值的原因。

得到的相似文本的索引如表 4 所示：

表 4 相似文本索引表

问题索引	相似文本的索引（空表示没有）
0	
1	1,442
2	2,3354,184,4014,2155,3841,3447,17
3	
4	4,1164,126,1352,2621,1902,2255,1955...1051
……	……
4325	

4.3 热度指数计算

热度指数的三个影响因子：留言频率、浏览热度、时间跨度。根据表 4 得到的数据对以上三个影响因子进行提取与处理。

在热度指数中浏览热度、各主题留言频率与问题热度成正比，作为分子；时间跨度与问题热成反比，作为分母。其中浏览热度与各主题留言频率分别归一化处理后进行加权，时间跨度以月为单位取整后同样进行归一化处理。这样既考虑了一些只进行投票而没有进行留言的人，以及留言了但是未对留言投票的人。归一化处理也消除了不同量纲和量纲单位带来的影响，解决了数据指标之间的可比性。

4.3.1 提出频率热度

不难得出，每类问题的留言条数反映着热点问题的一部分。某一类问题的留言条数越多，意味着这类问题是人们反映越多，政府就更应该关注与重视。

计算每一个索引对应的相似文本个数，在此基础上分别除以留言总数，得出各类问题的留言条数占总留言条数的比例。结果如表 5 所示：

表 5 各类留言问题频率表

问题索引	所占比例
0	0.0
1	0.00046
2	0.00185
3	0.0
4	0.00462
……	……
4325	0.0

4.3.2 提出浏览热度

浏览热度，这里我们定义为每一类问题点赞反对数之和占总点赞反对数之和的比例。先结合点赞数和反对数，计算得出投票总数。投票数目越高，意味着越多的人关注、浏览过此话题并且对此留言表态。因此浏览热度也应当作为热度指数的一部分。再对每一类问题的点赞反对数求和，在此基础上分别除以总投票数，得出各类问题的投票数占总投票数的比例。结果如表 6 所示：

表 6 各类留言问题浏览热度表

问题索引	所占比例
0	0.0
1	0.00108
2	0.00131
3	0.0
4	0.00077
……	……
4325	0.0

同时，我们考虑到每一条留言的点赞数与反对数的比值，反映着群众对某一问题的赞同度，这里我们将每一条留言的赞同率定义为：

$$赞同率 = \frac{点赞数}{点赞数 + 反对数}$$

于是我们得出每一条留言的赞同率，如下图 9 所示：

	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	投票总数	点赞率	赞同率
0	188006	A000102948	A3区一米阳光婚纱艺术摄影是否合法纳税了?	2019/2/28 11:25:05	座落在A市A3区联丰路米兰春天G2栋320，一家名叫一米阳光婚纱艺术摄影的影楼，据说年单这一...	0	0	0	0.000	0.000
1	188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	A市A6区道路命名规划已经初步成果公示文件，什么时候能转化成为正式的成果，希望能加快完成的路...	0	1	1	1.000	1.000
2	188031	A00040066	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	本人系春华镇金鼎村七里组村民，不知是否有相关水泥路到户政策和自来水到户政策，如政府主导投资村...	0	1	1	1.000	1.000
3	188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	靠近黄兴路步行街，城南路街道、大古道巷、一步两搭桥小区（停车场东面围墙外），第一单元一住户卫...	0	1	1	1.000	1.000

图 9 留言赞同率

赞同率为 1，说明无反对数，赞同率为 0，说明无点赞数，而经过计算得出，赞同率为 1 和 0 的留言条数占总留言条数的 96.37%，此时赞同率对总体热度指数来说意义不大，因此我们忽略赞同率这个影响因子，仅考虑浏览热度。

4.3.3 提出时间跨度

每个留言都有其对应的时间，而一个热点问题的出现总是在特定的时间。我们想要做的是，在大范围下，时间跨度越小的一类问题，就越热门，政府就更应该重视。因此应当是时间跨度越小，热度指数越大，则它们呈反比的关系，时间跨度应作为分母进行计算。同时，我们也需要排除一类问题中留言条数过少而引起的时间跨度小，从而使得热度指数变高的问题。因此我们对时间跨度以月为单位进行处理，不满 1 个月的按 1 个月计算。

由于给定的留言时间格式不一致，有 datetime.datetim 和 str 类型的数据，有以‘/’为分隔和以‘-’为分隔的数据,那么首先我们用 strftime 和 split 方法对时间进行处理，将留言内容都转换成 datetim.datetime 形式，并将每类问题的最早时间和最迟时间输出，部分数据如图 10 所示：

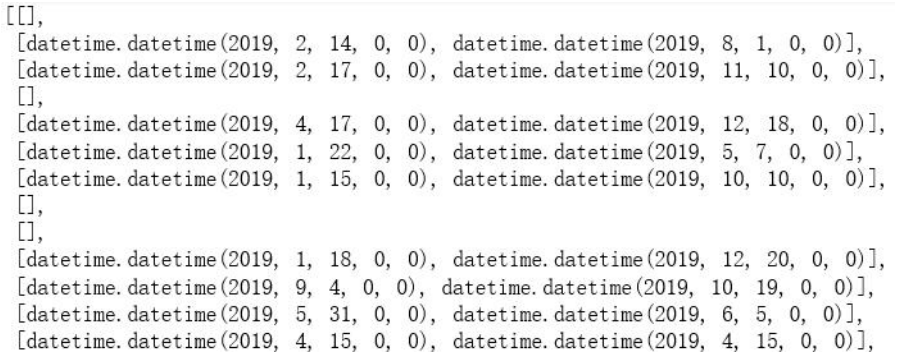

图 10 各类留言问题的最早和最近留言时间

再根据得出的列表，计算每一类问题相距的月份数，不满 1 个月的按 1 个月处理。利用（整除天数+1）最终得到一个对应各类问题的时间跨度列表（以月为单位），部分数据如下图 11 所示：

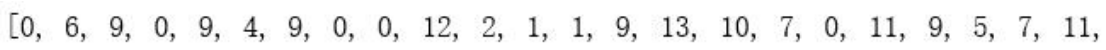

图 11 各类留言时间的时间跨度

4.3.4 归一化与加权

由于三个影响因子的量纲不同，我们不能单纯的把三个因子简单相加减。为此自定义一个函数，使得列表中的值映射到数值 0-1 之间，这样就很好的解决了这样一个量纲单位不一的问题。再分别把上述得到的频率热度、浏览热度、时间跨度的值带入函数，就可得到归一化后各因子下各类问题的对应数值。如下表 7 所示：

表 7 频率热度、浏览热度、时间跨度的归一化处理

问题索引	频率热度	浏览热度	时间跨度
0	0.0	0.0	0.0
1	0.04236	0.006	0.3
2	0.17035	0.00728	0.45
3	0.0	0.0	0.0
4	0.42541	0.00428	0.45
……	……	……	……
4325	0.0	0.0	0.0

三个因子的数值归一化处理好之后，分别对频率热度和浏览热度的权重取 0.5，0.5。得出的热度指数公式如下所示：

$$热度指数 = \frac{0.5 * 归一化后的频率热度 + 0.5 * 归一化后的浏览热度}{归一化后的时间跨度}$$

4.3.5 最终热度指数评价指标

计算出各类问题的热度指数后，对热度指数进行排序，取出排名前 5 的热度指数，并通过索引找到这 5 个热点问题对应的文本。如下表 8 所示:

表 8 热度前五的留言问题及文本索引

问题索引	热度指数	相似文本索引
0	5.212	249,1383,1212,1507
1	1.685	1483,2842
2	1.672	88,2345,2936,318,1482,1777.....1055,2351
3	1.546	818,1150,843,3684,237
4	1.492	511,1588,3546,3122

根据热度指数的公式，我们可以推断得出的 5 个热点问题中有这么三类：

1. 文本数量不少并且浏览量高。2. 文本数量少，但是浏览量极高。3. 浏览量不高，但文本数量极多。

4.4 获取热点问题关键字

取出各个热点问题的文本索引，用索引取出分词后的文本。利用 TfidfVectorizer 函数对各个热点问题进行计算求 TF-IDF 值，并且设置最小词频数 min_df，小于这个数便不记录这个词。为了避免关键词数量太多使得关键词获取不准确，则当模型的 get_feature_names()属性个数大于 25 时，即一个问题的关键字个数大于 25 时，则最小词频数+1，再利用模型获取关键词。输出 5 个热点问题对应的关键词，如下表 9 所示:

表 9 热度前五留言问题的关键词

问题 ID	关键词
1	58 车贷 A4 区 A 市 严惩 书记 保护伞 全国 公告 关注 典型 创造 承办 案件 特大 留言 立案 西地省 警官 诈骗案 跟进 集资
2	A 市 业主 伊景园滨河苑 商品房 定向 广铁集团 开发商 强行 投诉 捆绑 新城 权益 武广 维权 职工 购房 车位 违法 违规 小手 限价 项目
3	A 市 A5 区 K9 县 一楼 一系列 二期 五矿万境 建房 开发商 房屋 施工方 桃花苑 汇金路 消防安全 缩水 质量 隐患 面积
4	110kv A6 区 A 市 几个建议 投诉 月亮岛路 架设 沿线 现状 规划 金星北片 高压线
5	A1 区 A 市 商户 地方 广场 招商 消防 美食 营业 虚假

对上述得出的热点问题关键字与词频统计，分别绘制词云图，如下图 12 所示：

图 12 热度前五的留言问题词云图

4.5 结果展示

对以上所得出的结果，绘制数据框，提取的热点问题结果如图 13 所示。表格结果详见数据附件“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	5.212	2019/01/11至2019/03/01	A市A4区西地省	58车贷特大欺诈案
2	2	1.685	2019/07/18至2020/01/01	A市伊景园滨河苑的广铁集团职工	开放商把定向商品房捆绑车位违法销售
3	3	1.672	2019/03/22至2019/09/12	A市A5区五矿万境K9县	房屋存在质量问题和安全隐患
4	4	1.546	2019/03/26至2019/04/15	A市金星北片月亮岛路	110kv高压线的现状和规划
5	5	1.492	2019/12/15至2020/01/03	A市A1区老地方美食广场	美食广场虚假招商，无消防手续营业

图 13 热度前五的留言问题具体信息

同时根据相应热点问题的索引（见表 8）取出对应的留言信息，留言信息详情的部分结果如图 14 所示。表格结果详见附件“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	反对数	点赞数
1	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	胡书记：您好！58车贷案发，引发受害人举报投诉，也引起市领导的重视，公布了受害人的留言，便受...	0	733
1	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...	0	821
1	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。20...	0	790
1	223787	A00034861	西地省58车贷案件创造全国典型诈骗案，立案至今无公告	2019/1/11 21:12:34	原帖链接：西地省展星投资有限公司涉嫌诈骗58车贷案件从案发至今已经有五个月，西地省警方表现有...	0	0

图 14 热度前五问题的留言详情

第五章：答复意见评价方案

群众对问题进行留言，提出自己的意见与观点，那么政府相关部门就会对留言进行回复。本章将从回复的及时性、完整性、相关性进行分析。主要思路如下图 15 所示：

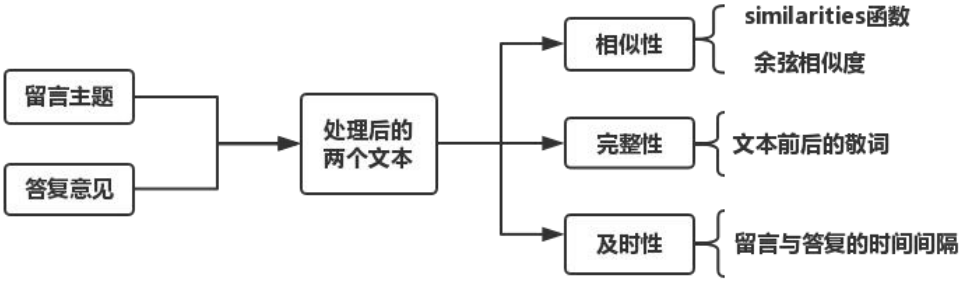

图 15 答复意见评价方案主要思路图

5.1 获取数据

取出数据中留言主题和答复意见，分别进行分词和去停用词等操作。分词时使用正则表达式找出特定的一些词，添加至 jieba 库中的词料库中，再经过 jieba 分词从而提高分词效果，接着利用停用词表去除一些无用的词。但是结果仍可以发现一些非英文的符号，为此定义一个去除各种标点的自定义函数。在分词之前，先导入 punctuation，通过自定义函数去除标点符号。得到数据处理后的 2816 条留言主题如下表 10 所示：

表 10 数据预处理后的留言主题

索引	数据清洗后的分词
0	A2 区 景荣华苑 物业管理
1	A3 区 潇楚南路 洋湖 段 修好
2	请 加快 提高 A 市 民营 幼儿园 老师 待遇
3	A 市 买 公寓 享受 人才 新政 购房 补贴
4	A 市 公交站点 名称 变更 建议
……	……
2815	呼吁 宁朱公路 拓宽 提质 改造

得到数据处理后的 2816 条答复意见如下表 11 所示：

表 11 数据预处理后的答复意见

索引	数据清洗后的分词
0	现将 网友 平台 问政 西地省 栏目 胡华衡 书记……
1	网友 A00023583 您好 A3 区 潇楚南路 洋湖……
2	市民 同志 您好 请 加快 提高 民营 幼儿园 教师……
3	网友 A000110735 您好 平台 问政 西地省 留言 ……
4	网友 A000100804 您好 A3 区 首席 现将 具体内容 ……
……	……
2815	UU008363 您好 留言 我厅 领导 高度重视……

5.2 相关性分析

对于留言与答复，两个文本的相关性十分重要。答复要有针对性，不可答非所问，这不仅耽误双方的时间，也使得政府在群众心中的满意度大大降低。

5.2.1 用 similarities 做文本相似度分类

把所有留言主题利用 corpora 形成词典，再用 doc2bow 函数制作语料库，每一条答复意见作为测试文本用 doc2bow 先转换成二元组的向量，再用 TF-IDF 模型计算其 TF-IDF 值。在此基础上用 similarities.SparseMatrixSimilarity 进行相似度分析，并取出与自己文本索引相对的数据。例如：答复意见索引 0 的文本与留言主题索引 0 的文本相似度值。部分结果如下表 12 所示：

表 12 答复意见与对应留言主题的文本相似度（similarities）

文本索引	相似度
0	0.19652611
1	0.27196157
2	0.37485312
3	0.32485312
4	0.22555285
……	……
2815	0.09982832

由于答复意见相较于留言主题的文本较大，分词后的词语数量较多，从而造成相似度的值较小。

得到留言主题和答复意见的相似度，就要进行划分答复的相关性。根据实际数据我们将相似度大于 0.3 的列为相关，相似度大于 0.1 而小于 0.3 的列为一般，相似度小于 0.1 的列为较差，同时也输出每个类别的文本个数。结果如下表 13 所示：

表 13 答复意见与留言相关度（similarities）

类别	对应文本索引与相似度	类别个数
相关	(2,0.3708773),(3,0.3268606)……	904
一般	(0,0.13008454),(1,0.29235035)……	1025
较差	(10,0.09271512),(16,0.020307498)……	887

5.2.2 用余弦相似度对文本做相似度分类

设置自定义函数来计算数据处理完以后的留言主题和答复意见的余弦相似度[7]。得到结果如下表 14 所示：

表 14 答复意见与对应留言主题的文本相似度（余弦相似度）

文本索引	相似度
0	0.182
1	0.246
2	0.439
3	0.41
4	0.19
……	……
2815	0.099

同样的，我们根据实际数据进行相关性划分。在此我们将相似度大于 0.25 的列为相关，相似度大于 0.1 而小于 0.25 的列为一般，相似度小于 0.1 的列为较差，同时也输出每个类别的文本个数。结果如下表 15 所示：

表 15 答复意见与留言相关度（余弦相似度）

类别	对应文本索引与相似度	类别个数
相关	(2,0.439),(3,0.41),(5,0.319),(6,0.455)……	671
一般	(0,0.182),(1,0.246),(4,0.19),(7,0.154)……	1296
较差	(10,0.043),(14,0.093),(19,0.052)……	849

5.2.3 评价两种相似度方法

从相似度数值上看来，两者的差距并不大。同时通过对比计算得出：在 2816 个文本中，两种方法判断相同的概率大约为 0.81。因此可推断在相当大的样本下，用 similarities 求文本相似度与用余弦相似度存在差异，但不显著。因此这两种方法我们均采用。

5.3 及时性分析

留言有留言的具体时间，答复也有答复的具体时间。倘若留言问题再热门，没有答复也让这个热门话题变得毫无作用。快节奏的时代人们追求速度，答复意见如果不够及时，人们的热情就会降低，同时也说明着相关人员的工作不到位。因此，答复意见的及时性也应当受到重视。

对于时间数据的处理，首先要先判断时间的字段类型，若时间字段为 object，那么就可以用 split 对字段进行处理；若时间字段为 datetime.datetime，那么就需要先用 strftime 转换为字符串类型。经过处理后的时间数据结果如下表 16 所示：

表 16 留言时间与对应答复时间表

留言时间	答复时间
2019/4/25	2019/5/10
2019/4/24	2019/5/9
2019/4/24	2019/5/9
2019/4/24	2019/5/9
2019/4/23	2019/5/9
……	……
2011/10/3	2012/2/28

我们再利用得出的留言时间与答复时间，计算出中间相隔的天数。在此我们将天数小于 5 天的，即在一周工作日时间内回复的，列为及时；天数大于 5 天但小于 15 天的，即在三周工作日时间内回复的，列为一般；天数大于 15 天，即超过三周工作日时间答复的，列为不及时，并且输出及时、一般、不及时三种情况所对应的答复条数。结果如下表 17 所示：

表 17 留言意见答复及时性

及时性	对应的文本索引	类别个数
及时	88,127,159,205,230……	796
一般	0,1,2,3,13,16,18,19……	959
不及时	4,,4,5,6,7,8,9,10,11……	1061

分析结果看来，相关工作人员在留言答复及时性这方面还有待加强。避免让群众认为自己的留言石沉大海，避免让问题的热度消散造成答复的滞后，相关工作人员都应该让留言的回复更加及时。

5.4 完整性分析

越来越注重生活品质的人们，同样也提高了对服务的要求。对于答复意见的完整性，我们认为可以从答复的开头与结尾是否有相关的敬词来判断，例如：开头的您好、尊敬等，结尾的感谢、谢谢、支持、理解等。

若开头和结尾都有敬词，那么在此我们就认为答复是完整的；若开头有敬词或者结尾有敬词，我们就认为答复的完整性为一般；若开头与结尾都没有敬词，我们就认为答复是不完整的。经计算后结果如下表 18 所示：

表 18 留言答复完整性

类别	对应文本索引与相似度	类别个数
完整	0,1,2,3,4,5,6,7,,8,9,10,11……	1573
一般	111,123,138,141,148,382……	979
不完整	163,241,1005,1009,1072……	264

从结果可以看出，在答复的完整性这一方面，相关工作人员做的还是相对比较到位。

5.5 评价标准总结

根据以上我们从答复意见的相关性、及时性、完整性所做的相应分析，绘制相应的数据框。用 random 函数随机抽取 5 位群众的留言信息，我们抽取到留言编号为 46740、97540、24673、68462、158341 的群众留言信息，如下图 16 所示：

留言编号	留言	留言主题	留言时间	留言详情	答复意见	答复时间
24673	00081	关于对阳光大道道路拓宽改进	2019/5/28 17:13:42	对道路时间影响性较大，特别是对造成影响，目前暂无相关改造计划。2、为保证		2019/6/4 10:41:24
46740	00086	中沙镇也举行“欢乐潇楚”	2013/5/22 17:02:36	把了最大的文艺活动，还举办 参加“毕业嘉园”环静众广场健身舞比赛，共取		2013/5/28 17:21:38
68462	00084	关于市区快修建小区配套公办	2019/9/5 15:56:23	享受优质教育资源，随获教育规划时，应当充分考虑中小学校、幼儿园规划建		2019/9/30 9:26:22
97540	00088	医院处理不当，导致患者错过	2019/7/27 22:40:09	为必须马上手术，把相关风险告诉请外院专家来院会诊以进一步确定是否需要做		2019/10/21 17:21:57
158341	00089	关于市黔城镇山门新村林木补偿	2018/11/3 22:37:01	过线路工程已经搞完了这么久直接向当地林业主管部门咨询，感谢你对西地省		2018/11/16 15:45:56

图 16 随机抽取的五位群众留言信息

对应取出依据答复评价方案所得的答复评价，如下图 17 所示：

	留言编号	回复完整性	tfidf回复相关性	cos回复相关性	回复时差 (天)	回复速度
1535	46740	完整	相关	相关	6	一般
2052	97540	完整	相关	一般	86	慢
1027	24673	一般	相关	相关	7	一般
1679	68462	一般	一般	一般	25	慢
2711	158341	完整	较差	较差	13	一般

图 17 随机抽取的五位群众留言的答复评价

下面我们拿留言编号为 24673 的群众信息来验证。

留言时间：2019/5/28　17:13:42

留言主题：“咨询 A7 县向阳路光大通道涵洞改进的问题”。

答复时间：2019/6/4　10:41:24

答复意见：“先生、女士：您好，对于您在信中提及‘向阳路光大通道涵洞的改进’的建议，我局高度重视，现回复如下：1、向阳路光大通道涵洞原设计即为车行通道，不适宜行人通行，且光大通道涵洞下穿京港澳高速，改造将对上方京港澳高速交通造成影响，目前暂无相关改造计划。2、为保证行人安全，建议市民走北面金茂路人行通道及南面漓楚路跨线桥过京港澳高速。感谢您对城市管理工作的关注，今后如您再次遇到城市管理相关问题，欢迎致电城管服务热线 0000-00000000 或 12319，我局将热忱为您服务。2018 年 5 月 30 日”。

相关性：可以看出文本的相关性还是比较强烈，相关部门回答的也很详细。

完整性：开头有相应称呼和“您好”的敬词，结尾有“感谢关注”的敬语，可以看出相关部门答复的完整性较好。

及时性：留言与答复时间相隔 7 天，答复及时性一般。

依次对照原数据与我们得出的答复评价，可以发现随机抽取的 5 条数据答复评价都较为准确，因此可以验证得出我们的答复评价方案较为合理。

总 结

面对近年来各种民意获取渠道增多带来的各类社情民意相关文本数据量激增，难以依靠人工实现准确分类整理的问题，我们小组给出了文本分类模型、热度指标体系和答复意见的质量评价方案。

对于文本分类模型，我们利用 python 对数据进行预处理，利用多个多分类模型对数据进行训练和预测。根据相应的 F1-Score 和训练耗时均衡考虑，最终选择支持向量机 SVM 作为最终模型。并发现城乡建设与环境保护是群众民生关注的一个重点，随着人民生活质量水平的提高，商贸旅游方面变得越来越受人们重视，政府有关部门应当在这些方面加强管理与发展，共同努力为人民谋幸福。

对于热度指标体系，我们根据浏览热度、留言频率、时间跨度三个因子得出热度指数计算公式，找到排名前五的热点问题，相关部门应该提高对其的关注度，精准地了解民意。

对于答复意见的质量评价方案，我们从答复的相关性、完整性、及时性三个角度分别建立了划分标准，形成了较为合理的答复意见评价方案。

但是，本文仍存在不足，主要是对文本类型的数据处理不够完善，模型的选择与搭建仍需要改进。因此在后续的学习中，考虑扩展机器学习与深度学习相关知识，运用概率统计、统计学、优化理论等方法优化程序，使得模型效果得到更多的提升。同时我们也希望，政务系统能够在大数据环境下得到更好地发展，使得民意得到更好的汇聚和反馈，促使我们的社会更加和谐美好。

参考文献

[1]邢彪,根绒切机多吉.基于jieba分词搜索与SSM框架的电子商城购物系统[J].信息与电脑(理论版),2018(07):104-105+108.

[2] miner_zhu. NLP 之 jieba 分词原理简析[EB/OL].https://blog.csdn.net/miner_zhu/article/details/83246153,2018-10-21

[3] Revolver. 使用 scikit-learn 解决文本多分类问题(附 python 演练)[EB/OL].https://www.cnblogs.com/panchuangai/p/12568198.html, 2018-08-25

[4]邓天浪. 呼叫中心大数据文本挖掘分析与实现[D].北京邮电大学,2015.

[5] 刘建平 Pinard . 文本挖掘预处理之 TF-IDF[EB/OL].https://www.cnblogs.com/pinard/p/6693230.html, 2017-04-11

[6] 番番要吃肉. 用 Python 进行简单的文本相似度分析[EB/OL].https://blog.csdn.net/xiexf189/article/details/79092629, 2018-01-18

[7] HankTown. 计算 2 篇文本的文本相似度(python 实现)[EB/OL].https://www.cnblogs.com/HankTown/p/12757832.html, 2020-04-22