基于“智慧政务”的文本挖掘分析

摘要：

数据时代，“智慧政务”[1]成为服务高效化、数据实时化、响应及时化的主流电子政务工具，为解决基层百姓各类问题提供便捷式途径。于政府而言，可以借助大数据分析技术实现智慧感知，全面、精准、及时了解公众的多样化需求，并作出针对性响应，实现良性互动，以有效决策。但目前，主要依靠人工根据经验处理的电子政务系统存在工作量大、效率低且差错率高等问题，与社情民意需求不断攀升之间产生突出的矛盾。

本研究基于“智慧政务”的文本信息，主要运用了线性支持向量机、LDA 主题模型及逻辑回归数据挖掘技术，围绕群众问政留言分类及其若干关键模型与相关部门答复意见进行一系列研究。本文建立线性支持向量机模型实现了对群众留言的精准分类，以便政府及时且准确地分派各类社情民意给相应的职能部门处理；并通过正则表达式及 if 条件语句+for 循环使用 LDA 主题模型提取出前 5 个热点问题，帮助相关部门能够有针对地处理，提升服务效率，从而有效推动政府的管理水平和施政效率的提升。此外，本文还运用逻辑回归模型从答复的相关性、完整性及可解释性 3 个角度来评价相关部门答复意见的质量。

关键词：TF-IDF 法；线性支持向量机；正则表达式；LDA；逻辑回归

目录

1. 挖掘目标

近年来，微信、微博、市长信箱、阳光热线等网络平台逐渐成为群众问政及政府施政的重要渠道，社情民意的种类多、群众问政的数据量庞大，而对于主要依靠人工进行处理的政府来说不仅工作量大、效率低，且差错率高。因此，需要解决以下 3 个问题，提高智慧政务的施政效率及管理水平：

①问题 1：对电子政务系统上的群众留言进行精准分类，以便及时分派给相应的职能部门处理；②问题 2：提取群众留言中的前 5 个热点问题，帮助相关部门进行有针对地处理，提升服务效率；③问题 3：建立一套方案，从相关部门答复意见的相关性、完整性及可解释性角度来评价答复意见的质量。

✧ 针对问题 1，我们首先对群众留言进行了数据的清洗、类的转换、剔除特殊符号及停用词等数据预处理，得到高质量的数据后，通过中文分词、绘制云此图及 TF-IDF 法得到能够很好反映出主题群众留言的关键词和词语对，进而比对逻辑回归、多项式朴素贝叶斯、线性支持向量机及随机森林四种模型的分类效果，选取准确率最高的线性支持向量机作为群众留言的分类器，通过 F_1-score 评估结果可知道，线性支持向量机能够精准地将群众留言分类成一级标签。

✧ 针对问题 2，经过分词、自定义词及去停用词的数据预处理后，我们先用正则表达式匹配出区县（A2 区、A7 县等），接着用 if 条件语句+for 循环将出现次数最多区县的留言明细提取出来，然后对区县使用 LDA 主题模型，每个区找出一个主题，根据主题关键词用 excel 表格搜索、筛选，最后结合群众留言条数、点赞数、反对数计算出热度指数，选出排名前五的热点问题。

✧ 针对问题 3，我们主要利用附件 4 中相关部门的答复意见，先后通过定义质量指标、人工评分得到训练样本、提取文本特征及建立逻辑回归分类模型，预测测试集的答复意见是否存在相关性 correlation、完整性 integrality 及可解释性 interpretability，最终实现从答复的相关性、完整性及可解释性对答复意见的质量进行评价。

2. 总体流程

由于文本数据的非结构化特征，同时要处理的文本数据量巨大，本文将主要运用 Python 平台编程并结合 excel 技术实现对文本数据的处理，最终得到对群众留言的一级标签精准分类，挖掘出了群众留言中的前 5 个热点问题，并建立了一套评价相关部门答复意见的方案。本文的总体架构及思路如下：

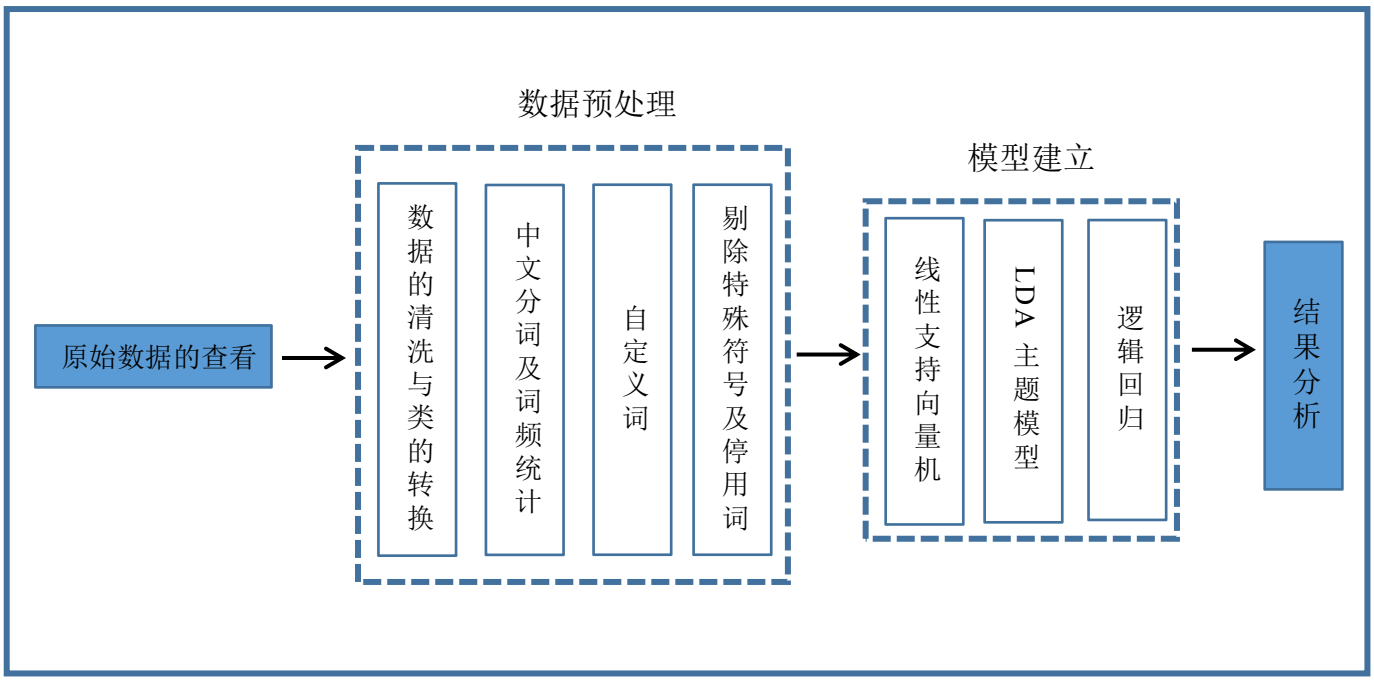

图 2-1　总流程图

本研究主要包括以下几个步骤：

步骤一：数据查看：查看原始数据的数据总量、分布情况及缺失值等，初步掌握原始数据集的情况。

步骤二：模型准备：对数据进行数据预处理，去除数据中的噪声，以免对后续数据的处理，及数据处理质量产生不良影响；根据实际情况灵活对数据进行类的转换、文本特征提取等过程，为建模提供有效数据。

步骤三：模型建立：探究 3 个问题的要求，将复杂问题分析转换成文本数据的识别主题和分类挖掘问题，从而选取最恰当的模型。

步骤四：模型应用：对建立好的模型进行评估分析并应用到实际数据处理中，实现对群众留言及相关部门答复意见数据的挖掘，分析出大数据中的有效信息。

3. 相关理论

3.1 TF-IDF

TF 即指词频(Term Frequency)，IDF 是逆文本频率指数(Inverse Document Frequency)，TF-IDF（term frequency－inverse document frequency）则一种用于信息检索与数据挖掘的常用加权技术，它通过在单词计数的基础上降低常用高频词的权重，增加罕见词的权重，使得我们更能通过罕见词的表达看出主题思想。例如，一则文章中出现“武汉”和“抗击新型冠状病毒”两个词，容易看出“抗击新型冠状病毒”更能表达文章的主题思想，而“武汉”是常见的高频词，它不能准确体现文章的主题思想。所以“抗击新型冠状病毒”的 TF-IDF 值要高于“武汉”的 TF-IDF 值。

3.2 F_1-score

F_1-score 即 F_1 分数，又称为平衡分数（balanced F Score），它被定义为精确率和召回率的调和平均数，它的最大值是 1，最小值是 0。F_1 分数被广泛运用在自然语言处理领域，比如命名实体识别、分词等，以用来衡量算法系统的性能。

通常使用 F1-score 对分类方法进行评价，计算公式为：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

3.3 正则表达式

正则表达式（Regular Expression）又称规则表达式，它是对字符串（包括普通字符（例如，a 到 z 之间的字母）和特殊字符（称为“元字符”））操作的一种逻辑公式，即用事先定义好的一些特定字符及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。

简单地说，正则表达式是用于搜索、替换和解析字符串等文本匹配的工具，它在源字符串中查找与给定的正则表达式相匹配的部分。一个正则表达式是由字幕、数字和特殊字符（括号、星号、问号等）组成。正则表达式中有许多特殊的字符，这些特殊字符是构成正则表达式的要素。常见的符号及描述有：

表 3-1　常见的正则表达式特殊符号

符号	描述
^	正则表达式的开始字符
$	正则表达式的结束字符
\w	匹配字母、数字、下划线
\W	匹配不是字母、数字、下划线的字符
\s	匹配空白字符
\S	匹配不是空白的字符
\d	匹配数字
\D	匹配非数字的字符
\b	匹配单词的开始和结束
\B	匹配不是单词开始和结束的位置
.	匹配任意字符，包括汉字
[m]	匹配单个字符串
[m1m2...n]	匹配多个字符串
[m-n]	匹配 m 到 n 区间内的数字、字母
[^m]	匹配除 m 意外的字符串
()	对正则表达式进行分组，一对圆括号表示一组
*?	匹配零次或多次，且最短匹配
+?	匹配一次或多次，且最短匹配
??	匹配零次或一次，且最短匹配
{m,n}?	重复 m 次，且最短匹配
(?#...)	正则表达式中的注释
(?P<name>...)	给分组命名，name 表示分组的名称
(?P=name)	使用名为 name 的分组

3.4 LDA

3.4.1 LDA 的定义

LDA（Latent Dirichlet Allocation）是一种非监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。LDA 主题模型将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析文档抽取出主题（分布）后，便可根据主题（分布）进行主题聚类或文本分类。

3.4.2 LDA 的主体思想

在 LDA 模型中，一篇文档生成的方式如下：

（1）从狄利克雷分布 α 中取样生成文档 i 的主题分布 θ_i；

（2）从主题的多项式分布 θ_i 中取样生成文档 i 的第 j 个词的主题 $z_{i,j}$；

（3）从狄利克雷分布 β 中取样生成主题 $z_{i,j}$ 对应的词语分布 $\phi_{z_{i,j}}$；

（4）从词语的多项式分布 $\phi_{z_{i,j}}$ 中采样最终生成词语 $\omega_{\mathrm{i,j}}$。

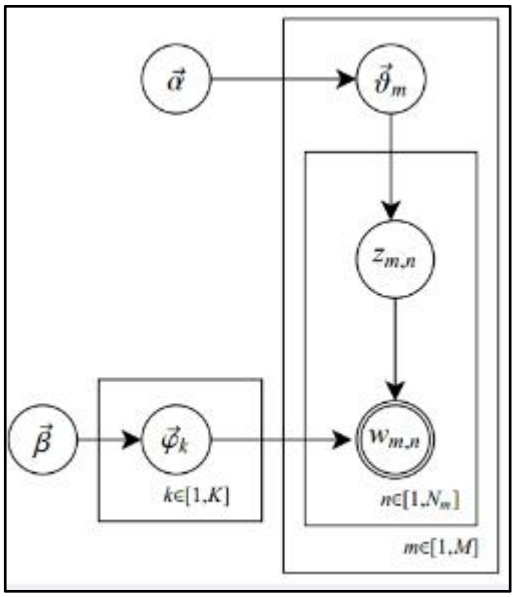

图 3-1 LDA 模型结构图

3.5 Linear SVM(线性支持向量机)

线性支持向量机[2]是一种分类学习算法，可以解决二分类、多分类问题。线性支持向量机的算法为：

输入：线性可分训练集 T={(x1,y1),(x2,y2),...,(xn,yn)}，

其中，$x_i \in \chi = R^n, y \in Y = \{1,-1\}, i = 1,2,...,N$ 。

输出：分离超平面 ω^* 和分类决策函数 f(x)。

（1）选择惩罚系数 C>0，构造并求解凸二次规划问题

$$\min_\alpha \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j y_i y_j (x_i x_j) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$C \geq \alpha_i \geq 0,\ i=1,2,...,N$$

求得最优解 $\alpha^* = (\alpha^*_1, \alpha^*_2, ..., \alpha^*_N)^T$ 。

（2）计算 $\omega^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i$

选择 α^* 的一个分量 α^*_j 满足条件 $0 < \alpha^*_j < C$，计算 $b^* = y_i - \sum_{i=1}^{N}\alpha_i^* y_i (x_i \cdot x_j)$ 。

注：由于求得的 b 不唯一，故可以取所有满足条件的样本点的平均值。

（3）求得分离超平面：$\omega^* \cdot x + b^* = 0$，

分离决策函数为：$f(x) = sign(\omega^* \cdot x + b^*)$

3.6 Logistic Regression(逻辑回归)

逻辑回归[3]不是回归，而是一种分类学习方法，它的因变量可以是二分类的，也可以是多分类的。逻辑回归的主要用途有以下寻找危险因素、预测及判别三种：

（1）寻找危险因素：寻找某一疾病的危险因素等；

（2）预测：根据模型预测在不同的自变量情况下，发生某种情况的概率有多大；

（3）判别：实际上跟预测有些类似，根据模型判断属于某种情况的概率有多大。

逻辑回归解决问题的常规步骤为：

（1）构造预测函数: $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$，表示结果取 1 的概率。

（2）构造损失函数：

$$J(\theta) = \frac{1}{m} Cost(h_\theta(x^{(i)}, y^{(i)}) = -\frac{1}{m}\sum_{i=1}^{n}(y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)})))$$

其中，$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)), y=1 \\ -\log(1-h_\theta(x)), y=0 \end{cases}$ 。

（3）求在损失函数 $J(\theta)$ 最小时的逻辑回归参数θ。

4. 挖掘过程

4.1 问题 1 — 群众留言的分类

4.1.1 原始数据的查看

为了能够对原始数据掌握简单的总体认识及了解，以便于后续想要把不同留言数据分到不同的分类中且每条数据只能对应一级分类中的一个类等的数据处理，对原始数据进行查看是很有必要的。经过数据的系列相关查看，我们得到以下信息：

首先，查看群众留言数据总量。我们的数据有两个字段，其中 cat 字段表示留言内容标签类别，review 表示用户的留言信息，数据总量为 9210，且留言内容全部为中文。随机抽取 10 条数据查看如下图：

数据总量： 9210 .

	cat	review
8146	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t\t 最近跟机动车检测站的朋友一起吃饭，聊起了...
4811	教育文体	\n \n B市教育局：您好，本人小孩在美的城小区小哈佛幼儿园上学，2017...
6045	劳动和社会保障	\n \n 我儿子出生在农村，随我们夫妇转正而变成非农户口。 因其不是农...
636	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t\t E市南站 景秀天源后面100米有三个烧烤...
3769	教育文体	\n\n局长： 你好！现在教育局是否鼓励中学教师脱产读研究生，和学校签订协议。另外教师为何...
2245	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t\t尊敬的领导： 你们好！我现在举报简家龙乡建...
3365	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t\t 2014年10月29日从深圳市宝安区福永...
407	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t\t2011年，A市城内集中城中村改居后，我们这...
1789	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t\t 关于请求拆除G市东紫大酒店（私自占用消防...
8051	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t\t 家住茂华国际楚A7栋。该小区电梯一直都是...

图 4-1　数据总量及随机查看

然后，统计各个类别的数据量。可以看到，各个类别的数据量不一致，城乡建设的数据量最多有 2009 条，教育文体、商贸旅游均有 1000 多条，环境保护、卫生计生次之分别有 938、877 条，而交通运输的数据量最少只有 613 条，各类别的数据量分布很不均匀。

	cat	count
0	城乡建设	2009
1	劳动和社会保障	1969
2	教育文体	1589
3	商贸旅游	1215
4	环境保护	938
5	卫生计生	877
6	交通运输	613

图 4-2　各类别的数据量

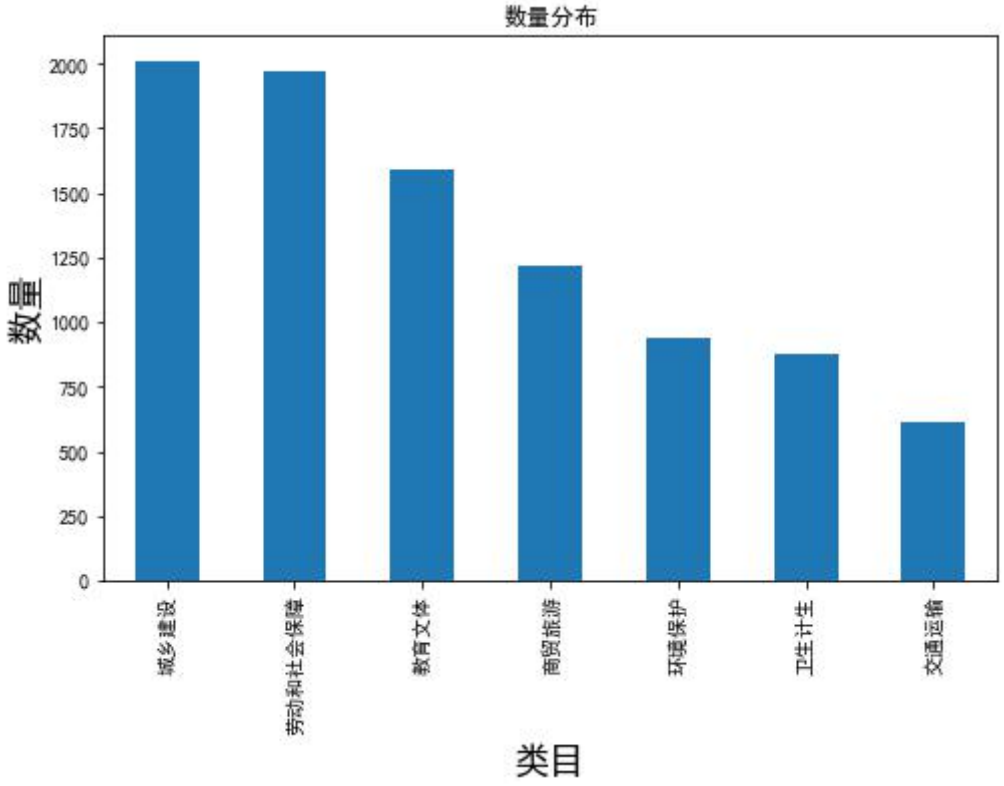

图 4-3　类目数量分布图

4.1.2 数据预处理

原始数据集中往往潜藏缺失值、异常值等“脏”数据，影响高质量的数据挖掘结果。为了得到高质量的数据，我们对群众留言进行了数据的清洗、类的转换、剔除特殊符号及停用词等数据预处理，以识别不正确、不完整、不相关、不准确等“脏”数据，检测和纠正数据集，实现建立有效的机器学习模型的第一步。

（1）数据的清洗

在数据的清洗过程中，我们主要清洗原始数据中的缺失值，检测数据的完整性。通过运行可知，原始数据中不存在缺失值，可见数据是非常完整的：

```
print("在 cat 列中总共有 %d 个空值." % df['cat'].isnull().sum())
print("在 review 列中总共有 %d 个空值." % df['review'].isnull().sum())
df[df.isnull().values==True]
df = df[pd.notnull(df['review'])]
#查看是否有缺失值
```

```
在 cat 列中总共有 0 个空值.
在 review 列中总共有 0 个空值.
```

图 4-4　数据的清洗

（2）类的转换

接下来，我们将 cat 类转换成 id（0 到 6），这样可便于分类模型的训练。

	cat	cat_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

图 4-5　类的转换

（3）剔除标点符号、特殊符号及停用词

由于数据集的留言内容都是中文，文本中的标点符号、特殊符号及无意义的常用词（stopword）等这些词和符号对系统分析预测文本的内容没有任何帮助，反而会增加计算的复杂和增加系统开销。在该处理中，我们定义删除了除字母、数字、汉字以外的所有符号，并过滤掉中文停用词中的“吧、吗、呢”感叹词等无法反应文本主要意思的高频常用词，最终生成了新的字段 clean_review。

	cat	review	cat_id	clean_review
7453	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 作为国家级风景名胜区的所在地，南湖风景区...	5	作为国家级风景名胜区的所在地南湖风景区是市委市政府对外接待的重要窗口我感觉环境卫生还有待加强...
2609	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t2019年10月22日，L5县葛竹坪镇过镇河...	1	2019年10月22日L5县葛竹坪镇过镇河河水被严重污染呈奶白色望有关部门严查此事还百姓一片...
2394	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t 关于强烈要求拆除K9县东侧村灌下洞违规基...	1	关于强烈要求拆除K9县东侧村灌下洞违规基站的申请报告尊敬县环保局领导我们K9县东侧村灌下洞村...
68	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 书记您好！ 我是刚需购房群体中的一员...	0	书记您好我是刚需购房群体中的一员A市房价暴涨后年轻人购房压力倍增感恩政府出台有力限购限贷措施...
4133	教育文体	\n \n 尊敬的王局长您好： 我是屈原管理区的学生家长，向您好反映市教育...	3	尊敬的王局长您好我是屈原管理区的学生家长向您好反映市教育局开展的英语竞赛乱收费的事F市教育局...
7450	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t F市南湖区集聚大量的外地人员，采用租住房...	5	F市南湖区集聚大量的外地人员采用租住房屋使用联通不记名SIM卡通信干着资本运作传销活动组织达...
4113	教育文体	\n\n A3区代课教师，工资拖欠是平常事，现在还没发五个月工资，工资1500，扣除保险1...	3	A3区代课教师工资拖欠是平常事现在还没发五个月工资工资1500扣除保险1200寒暑假没有工资...
7818	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t前阵子无限极倒了，权健倒了，都有损消费者权益...	5	前阵子无限极倒了权健倒了都有损消费者权益M市体育馆竟然还存在一个广州香爵生活馆一查百度三级分...
5484	劳动和社会保障	\n \n 3月23日中午12点，烟草网上报名截止，但网上缴费入口同时也...	4	3月23日中午12点烟草网上报名截止但网上缴费入口同时也不能进入点击出现不在时间范围内网上缴...
6882	劳动和社会保障	\n \n 尊敬的领导：我是通过省考进入公务员队伍，然后通过公开遴选从公务员...	4	尊敬的领导我是通过省考进入公务员队伍然后通过公开遴选从公务员单位进入A市正处级参公单位并进行...

图 4-6　新的字段 clean_review

4.1.3 中文分词

中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。本研究中，我们在 clean_review 的基础上进行分词，把每个留言内容分成由空格隔开的一个一个单独的词语。经过分词以后我们生成了 cut_review 字段，在 cut_review 中每个词语中间都是由空格隔开。

	cat	review	cat_id	clean_review	cut_review
0	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	0	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...	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑...
1	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	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用人行道...	位于 书院 路 主干道 在水 一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏...
2	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...	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利用业...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公...
3	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水霉...
4	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水霉...

图 4-7　中文分词

4.1.4 词频统计

简单来说，词频统计即指对文中出现的词语频次进行统计分析。在本研究中，我们首先画出群众留言的各个类别前 100 个高频词的词云，然后计算生成字段 cut_review 的 TF-IDF 特征值，再运用卡方(chi2)检验的方法来找出每个分类中关联度最大的两个词语和两个词语对，以此来反映出分类的主题。

（1）云词图

我们在 cut_review 的基础上生成每个分类的词云，在每个分类中罗列前 100

个高频词，这些高频词的词云如下：

图 4-8　各类别前 100 个高频词的云词图

（2）计算及抽取 TF-IDF 特征值 festures

本研究中，我们使用 sklearn.feature_extraction.text.TfidfVectorizer 方法来抽取群众留言文本的 TF-IDF 的特征值。其中参数 ngram_range=(1,2)，表示不仅抽取留言详情中的每个词语，还抽取每个词相邻的词并组成一个“词语对”(相邻两个单词的组合)，如：词 1，词 2，词 3，词 4，(词 1,词 2)，(词 2,词 3)，(词 3,词 4)。通过这样，可以扩展特征集的数量，提高分类文本准确度。其中参数 norm='12'，是一种数据标准划处理的方式，可以将数据限制在一点的范围内。

通过运算可知 features 的维度是(9210,695514)，则表示总共有 9210 条留言数据，有 695514 个特征数量（其中包括全部留言中的所有词语数与词语对的总数）。

```
(9210, 695514)
------------------------------
  (0, 24464)    0.09048238686512874
  (0, 256660)   0.0894235514720539
  (0, 601105)   0.15622076750032005
  (0, 117964)   0.15622076750032005
  (0, 436627)   0.15622076750032005
  (0, 634791)   0.09493620623431935
  (0, 170960)   0.11252439478269673
  (0, 635042)   0.19286549306096207
  (0, 100002)   0.09783281128157213
  (0, 174531)   0.0727674081013982
  (0, 635141)   0.09665802304744943
  (0, 600975)   0.11984007087877187
  (0, 325476)   0.07434015537017552
  (0, 677100)   0.08440577861421532
  (0, 493306)   0.13803041918954595
  (0, 275431)   0.0792415337313895
  (0, 681238)   0.06427725835949805
  (0, 413376)   0.07410612223604766
  (0, 232889)   0.1025961269249894
  (0, 49238)    0.10930964325866781
  (0, 435286)   0.06625154279011952
  (0, 634607)   0.08775246365365537
  (0, 99514)    0.11578143896976714
  (0, 637781)   0.11809555691530425
  (0, 274248)   0.07506278213290167
```

```
:     :
(9209, 231250)        0.10561913612050072
(9209, 565139)        0.10561913612050072
(9209, 20363) 0.10561913612050072
(9209, 370922)        0.10561913612050072
(9209, 357006)        0.10561913612050072
(9209, 220877)        0.10561913612050072
(9209, 92094) 0.10561913612050072
(9209, 291404)        0.10561913612050072
(9209, 524866)        0.10561913612050072
(9209, 432653)        0.10561913612050072
(9209, 92607) 0.10561913612050072
(9209, 482650)        0.10561913612050072
(9209, 357074)        0.10561913612050072
(9209, 522124)        0.10561913612050072
(9209, 249998)        0.10561913612050072
(9209, 147650)        0.10561913612050072
(9209, 370905)        0.10561913612050072
(9209, 149156)        0.10561913612050072
(9209, 460185)        0.10561913612050072
(9209, 555351)        0.10561913612050072
(9209, 44381) 0.10561913612050072
(9209, 218211)        0.10561913612050072
(9209, 147648)        0.10561913612050072
(9209, 44441) 0.10561913612050072
(9209, 500085)        0.10561913612050072
```

图 4-9　features 维度及 TF-IDF 特征值

（3）卡方(chi2)检验

运用卡方检验的方法，我们找出每个分类中关联度最大的两个词语和两个词语对，这些词和词语对能够较准确地反映出分类的主题，如下面所示：

'交通运输':

. Most correlated unigrams:

. 快递

. 出租车

. Most correlated bigrams:

. 的士 司机

. 出租车 司机

'劳动和社会保障':

. Most correlated unigrams:

. 退休

. 社保

. Most correlated bigrams:

. 劳动 关系

. 退休 人员

'卫生计生':

. Most correlated unigrams:

. 医生

. 医院

. Most correlated bigrams:

. 社会 抚养费

. 乡村 医生

'商贸旅游':

. Most correlated unigrams:

. 传销

. 电梯

. Most correlated bigrams:

. 小区 电梯

. 传销 组织

'城乡建设':

. Most correlated unigrams:

. 小区

. 业主

. Most correlated bigrams:

. 住房 公积金

. 公积金 贷款

'教育文体':

. Most correlated unigrams:

. 学生

. 学校

. Most correlated bigrams:

. 教育局 领导

. 培训 机构

'环境保护':

. Most correlated unigrams:

. 环保局

. 污染

. Most correlated bigrams:

. 周边 居民

. 环保局 领导

4.1.5 留言分类模型的建立

在对留言分类中，我们分别使用 Logistic Regression(逻辑回归)、(Multinomial) Naive Bayes(多项式朴素贝叶斯)、Linear Support Vector Machine(线性支持向量机)及 Random Forest(随机森林)四种不同的机器学习模型，并评估比较这四种模型的准确率[4]。

我们均对四种模型进行了 5 次预测和准确率计算，得到它们的平均准确率分别为：逻辑回归模型为 80.7%，多项式朴素贝叶斯模型为 65.2%，线性支持向量机模型最高为 87.3%，随机森林模型最低为 39.5%。又由各个方法准确率的箱体图可已看出，随机森林分类器的准确率是最低的，这是由于随机森林属于集成分类器(有若干个子分类器组合而成)，不适合处理高维数据(如文本数据)，而文本数据有太多的特征值，使得集成分类器难以应付；此外，可以看出线性支持向量机模型的准确率是最高的。因此，我们最终选取准确率最高的线性支持向量机作为留言分类的模型。

```
model_name
LinearSVC                0.872863
LogisticRegression       0.807385
MultinomialNB            0.652005
RandomForestClassifier   0.394889
Name: accuracy, dtype: float64
```

图 4-10　分类器的平均准确率

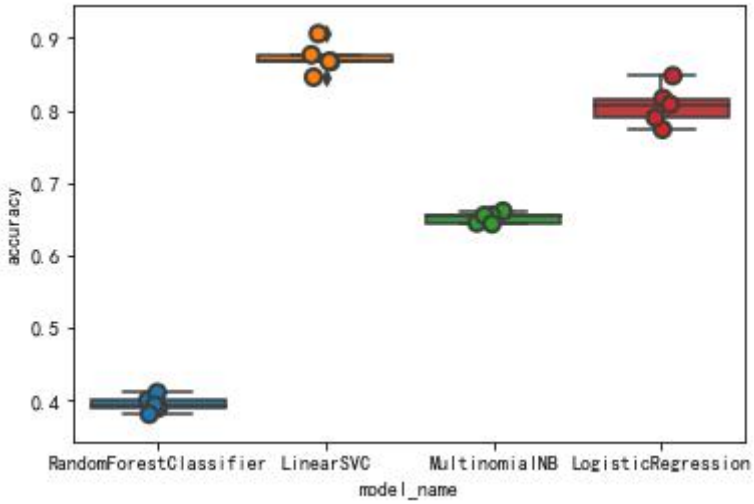

图 4-11　分类器准确率的箱体图

4.2 问题 2 — 热点问题的挖掘

本研究中，我们针对群众留言主题文本数据进行分词、自定义词及去停用词的预处理后，用正则表达式匹配出区县（A2 区、A7 县等），接着用 if 条件语句+for 循环语句提取出现次数最多区县的留言明细，然后对区县使用 LDA 主题模型，每个区找出一个主题，根据主题关键词用 excel 表格搜索、筛选，最后结合群众的留言条数、点赞数、反对数计算选出排名前五的热点问题。

4.2.1 数据预处理

（1）分词

进行文本数据挖掘时，应对文本中文分词，将连续的字序列按照一定的规范即分词处理重新组合成词序列，文本按照粒度可分为词语、句子、段落，所以词语是分词最终所得的能进行一定描述意义的最小单位， python 中的“jieba”库可以对文本中的留言主题进行中文分词，有利于去停用词、格式转换等后续操作过程。因此，我们调用 jieba 工具对群众留言的主题进行 jieba 分词并过滤词性。

（2）自定义词

在用分词的时候发现有些未登录词使得实体名分得不规则，因此我们需要使用自定义词典，建立自己需要用到的不需要分开的词语。如：

A4 区：直接分词→A 4 区，自定义词典后→A4 区。

（3）去停用词

我们对群众留言主题进行去停用词处理，这一步骤能够帮助我们去除掉一些冗余不需要的字词，得到高质量的数据，有助于建立有效的机器学习模型。

4.2.2 正则表达式匹配

借助正则表达式可以搜索、替换和解析字符串等文本匹配，以便在源字符串中查找与给定的正则表达式相匹配的部分的作用，我们对经过与处理后的数据进行正则表达式，把 A1 区，A2 区等区县提取出来。

4.2.3 if 条件语句提取

在正则表达式从留言主题中提取到区县后，我们通过编程 if 条件语句+for 循环代码，提取出现次数最多的 A7 县、A1 区、A2 区、A3 区、A4 区、A 市、经济学院及魅力之城这 8 个地区的留言明细，具体结果如下所示：

	留言编号	留言用户	留言主题	留言时间	\
0	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	
1	188251	A00013092	A7县特立路与东四路口晚高峰太堵，建议调整信号灯配时	2019/10/19 11:02:40	
2	188451	A00013004	A7县春华镇石塘铺村有党员家开麻将馆	2019/4/11 17:54:25	
3	188535	A00061775	A7县时代星城4幢有非法经营的家庭旅馆	2019/6/13 15:28:44	
4	188780	A00094754	请依法解决A7县黄花镇梁坪村黄泥岭山地建房问题	2019/10/20 14:59:13	
..	...	...	...	...	
676	289520	A00034066	请还A7县国泰九龙湾2000多住户一个公道	2019/3/5 10:49:54	
677	289727	A00049650	咨询A7县北横线拆迁标准	2019/3/18 17:19:07	
678	289771	A0004202	A7县行政执法局榔梨分局公务员以吞路货为由拒不还款	2019/5/23 17:56:08	
679	289783	A00063935	请问A7县中茂城商圈规划在哪个位置	2019/10/15 10:04:24	
680	289834	A00051906	A7县文体中心乒羽中心何时能建成开放？	2019/10/22 15:28:17	

	留言详情	点赞数	反对数
0	\n\t\t\t\t\t\t\n\t\t\t\t\t\t本人系春华镇金鼎村七里组村民，不知是否有相关...	1	0
1	\n\t\t\t\t\t\t\n\t\t\t\t\t\t近来，下午晚高峰五点半左右，经过特立路与东四...	0	0
2	\n\t\t\t\t\t\t\n\t\t\t\t\t\t我是春华镇一名村民，最近接到政府的通知，02...	2	0
3	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的各位领导A7县时代星城小区4幢1321...	0	0
4	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的县领导：您好！2015年我组的村民要在...	0	0
..		...	...
676	\n\t\t\t\t\t\t\n\t\t\t\t\t\tA7县国泰九龙湾业主委员会在老物业合同到期后...	1	0
677	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的领导，您好！请问目前A7县北横线的拆迁...	0	0
678	\n\t\t\t\t\t\t\n\t\t\t\t\t\t作为优秀的我们国家员，蒋清本人亲自带头接触p...	0	0
679	\n\t\t\t\t\t\t\n\t\t\t\t\t\t领导，请问一下中茂城金科时代、新长海商圈规划...	3	0
680	\n\t\t\t\t\t\t\n\t\t\t\t\t\t请问A7县文体中心乒羽中心何时能正式对外开放...	8	1

[681 rows x 7 columns]

图 4-12　A7 县留言明细

	留言编号	留言用户	留言主题	留言时间	\
0	188922	A00039738	A市A1区纬二路上有两种处罚标准？	2019/6/20 13:13:17	
1	189635	A00029819	A1区桐阴里小区一直夜间施工	2019/7/15 21:27:02	
2	190090	A000108593	严重举报A1区华海3c朝阳菜市场附近夜宵摊门店口摆满桌椅	2019/3/26 11:00:05	
3	191052	A00094257	A2区建中A1区嘉苑偶寓22栋违规搭建玻璃阳光房	2019/8/1 17:39:01	
4	191062	A00027934	请公开A1区东成大厦的相关规划文件	2019/5/13 16:10:18	
..	...	...	...	...	
200	286999	A00052140	A1区苏家巷社区复建房屋严重影响我家房屋采光！	2019/2/16 1:13:18	
201	287527	A00019121	A1区东盈商业广场开发商欺骗、恶弄全体业主	2019/3/14 17:36:10	
202	287897	A00063590	A市渣土办在A1区新安村倾倒渣土，有审批手续没	2019/6/4 0:30:26	
203	289808	A000112212	A1区朝阳街道解放东路二里牌向韶村马路市场死灰复燃	2019/3/20 17:42:38	
204	289930	A000100792	A1区辉煌国际城二期居民楼下商铺违法开饭店，维权近三月没有用	2019/1/27 12:18:43	

	留言详情	点赞数	反对数
0	\n\t\t\t\t\t\t\n\t\t\t\t\t\tA市A1区纬二路是一条东西方向的马路，马路两...	0	0
1	\n\t\t\t\t\t\t\n\t\t\t\t\t\t桐阴里夜间施工扰民，桐阴里小区长期以来一直夜...	0	0
2	\n\t\t\t\t\t\t\n\t\t\t\t\t\t还我一个安静的夜晚！我要严重举报A1区华海3...	0	0
3	\n\t\t\t\t\t\t\n\t\t\t\t\t\tA2区建中A1区嘉苑偶寓22栋有07、09、...	0	0
4	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的冯意刚局长并A市自然资源和规划局:我会...	0	0
..		...	...
200	\n\t\t\t\t\t\t\n\t\t\t\t\t\t我叫陈元春，今年82岁，系A市A1区苏家巷根...	0	0
201	\n\t\t\t\t\t\t\n\t\t\t\t\t\t我们是A1区东盈商业广场的全体业主！该楼盘位...	0	0
202	\n\t\t\t\t\t\t\n\t\t\t\t\t\t我是A市A1区东岸乡新安村八组村民，请问A市...	0	0
203	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的A市政府胡维忠市长：您好！我们是A1区...	1	0
204	\n\t\t\t\t\t\t\n\t\t\t\t\t\tA1区远大一路辉煌国际城二期物业公司（北京东...	0	0

[205 rows x 7 columns]

图 4-13　A1 区留言明细

	留言编号	留言用户	留言主题	留言时间	\
0	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	
1	188074	A909092	A2区富绿新村房产的性质是什么？	2019/1/31 20:17:32	
2	188546	A0006817	A2区佳兆业水新都小区垃圾无人处理	2019/1/23 13:09:19	
3	188774	A00048792	A2区政府东门至万美路段经常有改装车飙车，真的很扰民	2019/6/18 23:03:31	
4	188829	A00026141	A2区先锋派出所办个签证，拒收现金	2019/2/15 13:17:06	
..	...	...	...	...	
259	287631	A00040634	请政府为A2区银杏嘉园居民拿回车坪所有权、使用权	2019/7/29 15:34:03	
260	287644	A000108212	请求解决A2区西文庙坪粮食局宿舍房屋漏水问题	2019/12/10 16:34:51	
261	288603	A0006857	A2区hpv九价疫苗一针难求	2019/5/16 11:46:27	
262	288619	A00017807	反映A2区育园中信小学1904班班主任安排座位不合理等问题	2019/11/15 13:59:35	
263	289832	A00096296	A2区竹塘西路—新姚路施工项目夜间严重扰民	2019/12/18 2:40:00	

	留言详情	点赞数	反对数
0	\n\t\t\t\t\t\t\n\t\t\t\t\t\t靠近黄兴路步行街，城南路街道、大古道巷、一步...	1	0
1	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\t"二高一部"发出关于针对非法集资的打击的通知...	0	0
2	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\t敬爱的领导：你好，我是佳兆业水新都一期小区业...	0	0
3	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\t多年来A2区迎新路区政府东门至万美路段经常有...	0	0
4	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\t为什么办个签证，港澳台通行证现在还拒收现金？...	2	0
..	...	...	...
259	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\tA2区银杏嘉园9栋南面有一大块场坪，应该是小...	0	0
260	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\t各位领导你们好！我们是西文庙坪粮食局宿舍的居...	0	0
261	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\tHPV九价疫自2018年10月登录A市，一苗...	0	0
262	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\t1.反映老师让孩子一直坐在靠窗户孩子刚上一年...	0	0
263	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\t你好！我是A市A2区莲花小区的居民，A2区竹...	0	0

[264 rows x 7 columns]

图 4-14　A2 区留言明细

	留言编号	留言用户	留言主题	留言时间	\
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	
1	188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	
2	188073	A909164	A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质	2019/3/11 11:40:42	
3	188249	A00084085	A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民	2019/9/17 4:25:00	
4	188260	A00053484	A3区育育家园小区乐果果零食炒货公共通道播放空调扰民	2019/5/31 17:06:13	
..	...	...	...	...	
427	288790	A00080913	A3区麓山国际梅溪湖校区已经停工很久了，还能开课吗	2019/8/4 1:21:16	
428	289130	A00054443	A3区雷峰大道A3区段沿线违停严重	2019/11/22 8:17:48	
429	289321	A00030338	A3区佳兆业云顶梅溪湖二期承重墙被砸了没有恢复原样	2019/12/16 15:21:47	
430	289772	A00041363	A3区A3区街道惟盛园安置小区收费不合理	2019/3/19 0:21:45	
431	289813	A00092873	A3区博才学士校区大门前地下通道急需建设	2019/12/24 9:42:43	

	留言详情	点赞数	反对数
0	\n\t\t\t\t\t\t\n\t\t\t\t\t\t座落在A市A3区联丰路米兰春天G2栋320，...	0	0
1	\n\t\t\t\t\t\t\n\t\t\t\t\t\tA市A3区中海国际社区三期四期中间，即蓝天璞...	0	0
2	\n\t\t\t\t\t\t\n\t\t\t\t\t\t作为麓泉社区麓谷明珠小区6栋居民，我们近期感...	0	0
3	\n\t\t\t\t\t\t\n\t\t\t\t\t\t保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2...	0	0
4	\n\t\t\t\t\t\t\n\t\t\t\t\t\t还我宁静我要复习迎考，大半年底商空调/冰柜外...	0	0
..	...	...	...
427	\n\t\t\t\t\t\t\n\t\t\t\t\t\t目前麓山国际梅溪湖校区已经停工很久了，周边建...	0	0
428	\n\t\t\t\t\t\t\n\t\t\t\t\t\t雷峰大道从杜鹃路口到尖山路口沿线违停严重！尤...	0	0
429	\n\t\t\t\t\t\t\n\t\t\t\t\t\t砸了承重墙可以不恢复原样吗我是佳兆业云顶梅溪...	0	0
430	\n\t\t\t\t\t\t\n\t\t\t\t\t\t1、征收承诺拿出5平米面积做商铺收益，永久免...	5	0
431	\n\t\t\t\t\t\t\n\t\t\t\t\t\t博才学士校区大门前地下通道急需建设，每天学生...	0	0

[432 rows x 7 columns]

图 4-15　A3 区留言明细

	留言编号	留言用户	留言主题	留言时间	\
0	188414	A00096844	A4区北辰小区非法住改商问题何时能解决?	2019/8/1 7:20:31	
1	189029	A00080952	A4区楚江北路能走大货车吗	2019/4/3 12:23:04	
2	189345	A00077163	A4区东风街道蚌塘社区自来水管网改造设计存在严重缺陷	2019/7/31 6:52:29	
3	191068	A00010562	A4区华章路路灯光源污染严重影响新领地小区居民生活!	2019/2/12 17:49:08	
4	191150	A000104272	A4区大道黄泥扩段南北约1500米的路灯不亮	2019/6/25 16:21:30	
..	...	...	...	...	
229	287424	A00063210	A4区大塘村陈家冲组集体用地被强占建黑油站	2019/12/5 19:53:53	
230	288373	A000113017	A4区恒大雅苑五期商业路摆摊现象严重	2019/7/11 17:52:36	
231	288773	A00012717	A4区沙坪街道中营社区大明村民的征地征收补偿款一分钱都没拿到	2019/6/19 15:25:45	
232	289103	A0009368	A4区双湾国际小区"饮于故事"餐饮店将油烟废气直接排入下水道	2019/3/12 10:05:01	
233	289178	A00079835	A4区沙坪绣坊街到底是怎么回事	2019/8/14 17:50:42	

	留言详情	点赞数	反对数
0	\n\t\t\t\t\t\n\t\t\t\t\t您好！我是北辰D1区一名业主，近年来，我小区...	0	0
1	\n\t\t\t\t\t\n\t\t\t\t\t从2018年底开始，政府对楚江北路二环至巍凝...	0	0
2	\n\t\t\t\t\t\n\t\t\t\t\t我是1996年购的安居房，在东风街道蚌塘社区...	0	0
3	\n\t\t\t\t\t\n\t\t\t\t\t从昨天2月11号晚上7点开始，华章路的路灯可...	0	0
4	\n\t\t\t\t\t\n\t\t\t\t\tA市A4区大道黄泥扩段南北约1500米，市政...	0	0
..	...	...	...
229	\n\t\t\t\t\t\n\t\t\t\t\各级领导大家好，我是A市A4区秀峰街道大塘村...	0	0
230	\n\t\t\t\t\t\n\t\t\t\t\每天早上7点至10点，下午4点至7点，A市A...	0	0
231	\n\t\t\t\t\t\n\t\t\t\t\2005年到现在大明村多次征地，由于历任现任...	0	0
232	\n\t\t\t\t\t\n\t\t\t\t\tA市A4区双湾国际小区2栋一楼门面"饮于故事...	0	0
233	\n\t\t\t\t\t\n\t\t\t\t\本人如2015年8月购买沙坪绣坊街门面一个5...	0	0

[234 rows x 7 columns]

图 4-16　A4 区留言明细

	留言编号	留言用户	留言主题	留言时间	\
0	188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	
1	188119	A00035029	对A市地铁造规用工问题的质疑	2019/5/27 16:04:44	
2	188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	
3	188399	A00097934	A市利保壹号公馆项目夜间噪声扰民	2019/7/3 6:23:25	
4	188409	A0003274	A市地铁3号线星沙大道站地铁出入口设置极不合理!	2019/6/19 10:14:39	
...	...	...	...	...	
1867	360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作	2019-11-22 14:42:14	
1868	360111	A1204455	A市经济学院组织学生外出打工合理吗?	2019-11-05 10:31:38	
1869	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	
1870	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	
1871	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	

	留言详情	点赞数	反对数
0	\n\t\t\t\t\t\n\t\t\t\t\tA市A3区中海国际社区三期四期中间，即蓝天璞...	0	0
1	\n\t\t\t\t\t\n\t\t\t\t\我是一名在A市某地铁站上班的安检员，我是由中...	0	0
2	\n\t\t\t\t\t\n\t\t\t\t\12月21日下午17时52分许，6路公交车（...	0	0
3	\n\t\t\t\t\t\n\t\t\t\t\您好，我想举报 A市利保壹号公馆项目 夜间噪...	0	0
4	\n\t\t\t\t\t\n\t\t\t\t\尊敬的领导您好，我是A7县星沙街道的一个有两...	4	0
...		...	...
1867	\n\t\t\t\t\t\t\t\t\关于西地省A市经济学院寒假过年期间组织学生去工厂工作...	0	0
1868	\n\t\t\t\t\t\t\t\t\一名中职院校的学生,学校组织我们学生在外边打工,...	0	1
1869	\n\t\t\t\t\t\t\t\t\各位领导干部大家好，我是A市经济学院的一名学生，...	0	0
1870	\n\t\t\t\t\t\t\t\tA市经济学院强制16届电子商务跟企业物流专业实习，其中我...	0	3
1871	\n\t\t\t\t\t\n\t\t\t\t\t　书记您好，我是来自西地省经济学院体育学院...	0	9

[1872 rows x 7 columns]

图 4-17　A 市留言明细

	留言编号	留言用户	留言主题	留言时间	\
0	195917	A909119	A市涉外经济学院组织学生外出打工合理吗？	2019/11/05 10:31:38	
1	211395	A00050903	西地省财政经济学院校园宽带被垄断	2019/9/14 17:57:34	
2	211800	A00046925	西地省财政经济学院食堂只开放十余个窗口，有些学生吃不上饭	2019/2/25 23:24:54	
3	233759	A909118	A市涉外经济学院强制学生实习	2019/04/28 17:32:51	
4	235521	A0006920	A3区枫林三路涉外经济学院外街理发店扰民	2019/10/15 18:59:08	
5	240721	A00050903	西地省财政经济学院涉嫌宽带垄断	2019/9/14 17:56:17	
6	242062	A00028889	西地省涉外经济学院变相强制学生“社会实践”	2019/11/27 23:14:33	
7	264084	A00074365	西地省财政经济学院以报名名额已满拒绝让学生报名cet-4	2019/3/19 23:11:44	
8	266368	A00038920	A市涉外经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	
9	360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作	2019-11-22 14:42:14	
10	360111	A1204455	A市经济学院组织学生外出打工合理吗？	2019-11-05 10:31:38	
11	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	
12	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	
13	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	

	留言详情	点赞数	反对数
0	\n\t\t\t\t\t\n\t\t\t\t\t一名中职院校的学生，学校组织我们学生在外边打...	1	0
1	\n\t\t\t\t\t\n\t\t\t\t\t财政经济学院宽带被电信垄断，引发大量学生不满...	0	0
2	\n\t\t\t\t\t\n\t\t\t\t\t开学的时候学校竟然只开放十余个窗口，但我们学...	0	0
3	\n\t\t\t\t\t\n\t\t\t\t\t各位领导干部大家好，我是A市涉外经济学院的一...	0	0
4	\n\t\t\t\t\t\n\t\t\t\t\tA市A3区，枫林三路，西地省涉外经济学院外街...	0	0
5	\n\t\t\t\t\t\n\t\t\t\t\t财政经济学院宽带被电信垄断，引发大量学生不满...	0	0
6	\n\t\t\t\t\t\n\t\t\t\t\t请制止和修改西地省涉外经济学院变相强制学生进...	0	0
7	\n\t\t\t\t\t\n\t\t\t\t\t本人是西地省财政经济学院的一名在读全日制本科...	0	0
8	\n\t\t\t\t\t\n\t\t\t\t\t关于西地省A市涉外经济学院寒假过年期间组织学...	0	0
9	\n\t\t\t\t\t\t\t\t\t关于西地省A市经济学院寒假过年期间组织学生去工厂工作...	0	0
10	\n\t\t\t\t\t\t\t\t\t一名中职院校的学生，学校组织我们学生在外边打工，...	0	1
11	\n\t\t\t\t\t\t\t\t\t各位领导干部大家好，我是A市经济学院的一名学生，...	0	0
12	\n\t\t\t\t\t\t\t\tA市经济学院强制16届电子商务跟企业物流专业实习。其中我...	0	3
13	\n\t\t\t\t\t\n\t\t\t\t\t书记您好，我是来自西地省经济学院体育学院...	0	9

图 4-18　经济学院留言明细

	留言编号	留言用户	留言主题	\
0	189381	A000109815	A市万科魅力之城商铺无排烟管道，小区内到处油烟味	
1	195095	A00039089	魅力之城小区临街门面油烟直排扰民	
2	198084	A00022429	A市万科魅力之城小区近百户楼板开裂墙面开裂	
3	205168	A00022429	A市万科魅力之城近百户房屋楼板、墙面开裂！	
4	232892	A00015335	A市万科魅力之城开发商未通知业主就进行车位开盘销售活动	
5	233338	A00022429	A市万科魅力之城楼板和墙面开裂，请政府管一管吧	
6	236303	A00022429	A市万科魅力之城有的房屋楼板严重开裂	
7	236798	A00039089	A5区劳动东路魅力之城小区油烟扰民	
8	240330	A00087099	A市万科魅力之城大降价折损前购买业主利益	
9	242792	A909115	A5区魅力之城小区一楼被搞成商业门面，噪音扰民严重	
10	245136	A909117	万科魅力之城小区底层门店深夜经营，各种噪音扰民	
11	246362	A909114	A市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	
12	246598	A00054842	A5区劳动东路魅力之城小区临街门面烧烤夜宵摊	
13	253314	A00089954	反映A5区万科魅力之城小孩入学问题	
14	263498	A00015643	A市万科魅力之城未交房业主孩子不能上小区配套小学	
15	268914	A0006238	A5区劳动东路魅力之城小区底层餐馆油烟扰民	
16	272122	A909113	A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理！	
17	284147	A909113	A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	
18	287386	A909116	A市万科魅力之城商铺无排烟管道，小区内到处油烟味	
19	360100	A324156	魅力之城小区临街门面油烟直排扰民	
20	360101	A324156	A5区劳动东路魅力之城小区油烟扰民	
21	360102	A1234140	A5区劳动东路魅力之城小区底层餐馆油烟扰民	
22	360103	A0012425	A5区劳动东路魅力之城小区临街门面烧烤夜宵摊	
23	360104	A012417	A市魅力之城商铺无排烟管道，小区内到处油烟味	
24	360105	A120356	A5区魅力之城小区一楼被搞成商业门面，噪音扰民严重	
25	360106	A235367	A市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	
26	360107	A0283523	A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	
27	360108	A0283523	A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理！	
28	360109	A0080252	万科魅力之城小区底层门店深夜经营，各种噪音扰民	

图 4-19　魅力之城留言明细（1）

	留言时间	留言详情 \
0	2019/12/4 16:25:06	\n\t\t\t\t\t\n\t\t\t\t\tA市万科魅力之城自打交房入住后，底层商铺无排...
1	2019/09/05 12:29:01	魅力之城小区楼下烧烤摊、快餐店无证经营，长期油烟烧烤熏死人。一天24小时都是烟。请政府关闭处...
2	2019/10/23 15:01:30	\n\t\t\t\t\t\n\t\t\t\t\t本人花近180万购买万科魅力之城的房子一套，...
3	2019/10/24 10:22:16	\n\t\t\t\t\t\n\t\t\t\t\tA市万科魅力之城从交房到现在一年半，近百户房...
4	2019/1/8 9:54:00	\n\t\t\t\t\t\n\t\t\t\t\t您好！我是57栋业主，2018年12月25日...
5	2019/11/13 10:58:25	\n\t\t\t\t\t\n\t\t\t\t\t我是万科魅力之城4.1期65栋1003房的业...
6	2019/10/29 16:57:02	\n\t\t\t\t\t\n\t\t\t\t\t我是万科魅力之城的业主，房屋2018年3月底...
7	2019/07/28 12:49:18	尊敬的政府：A5区劳动东路魅力之城小区临街门面长期油烟直排。长期投诉无果。由于各部门无权力强...
8	2019/3/21 9:03:00	\n\t\t\t\t\t\n\t\t\t\t\t本人2018年底购买https://baid...
9	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，噪声...
10	2019/09/04 21:00:18	您好：\n我是万科魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还...
11	2019/08/26 01:50:38	2019年5月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了...
12	2019/09/25 00:31:33	A5区劳动东路魅力之城小区临街夜宵摊、烧烤摊24小时经营，油烟扰民。违法经营。本小区临街居民...
13	2019/4/17 23:59:08	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的市委书记：您好！首先介绍下我的基本情况...
14	2019/4/17 14:21:00	\n\t\t\t\t\t\n\t\t\t\t\t市长：您好！我是A5区万科魅力之城准业主，于...
15	2019/09/10 06:13:27	A5区劳动东路魅力之城小区，底层有几家餐管，油烟严重影响我们居民的生活，(衣服晾晒外面有油烟...
16	2019/08/01 16:20:02	局长：　　你好，A5区劳动东路魅力之城小区的一楼，开了几家夜宵快餐店，厨房内多个灶台油烟随意...
17	2019/07/21 10:29:36	局长：　　你好，A5区劳动东路魅力之城小区的一楼，开了几家夜宵快餐店，厨房内多个灶台油烟随意...
18	2019/08/18 14:44:00	A市万科魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天...
19	2019-09-05 12:29:01	魅力之城小区楼下烧烤摊、快餐店无证经营，长期油烟烧烤熏死人。一天24小时都是烟。请政府关闭处...
20	2019-07-28 12:49:18	尊敬的政府：A5区劳动东路魅力之城小区临街门面长期油烟直排。长期投诉无果。由于各部门无权力强...
21	2019-09-10 06:13:27	A5区劳动东路魅力之城小区，底层有几家餐管，油烟严重影响我们居民的生活，(衣服晾晒外面有油烟...
22	2019-09-25 00:31:33	A5区劳动东路魅力之城小区临街夜宵摊、烧烤摊24小时经营，油烟扰民。违法经营。本小区临街居民...
23	2019-08-18 14:44:00	A市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天进出...
24	2019-08-26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，噪声...
25	2019-08-26 01:50:38	2019年5月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了...
26	2019-07-21 10:29:36	局长：　　你好，A5区劳动东路魅力之城小区的一楼，开了几家夜宵快餐店，厨房内多个灶台油烟随意...
27	2019-08-01 16:20:02	局长：　　你好，A5区劳动东路魅力之城小区的一楼，开了几家夜宵快餐店，厨房内多个灶台油烟随意...
28	2019-09-04 21:00:18	您好：\n我是万科魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还...

图 4-20　魅力之城留言明细（2）

	点赞数	反对数
0	0	0
1	3	0
2	0	0
3	1	0
4	1	0
5	0	0
6	0	0
7	4	0
8	0	0
9	1	0
10	0	0
11	0	0
12	1	0
13	0	0
14	0	0
15	0	0
16	6	0
17	3	0
18	0	0
19	0	3
20	0	4
21	0	0
22	0	1
23	0	0
24	0	1
25	0	0
26	0	3
27	0	6
28	0	0

图 4-21　魅力之城留言明细（3）

4.2.4 建立 LDA 主题模型

用 if 条件语句+for 循环提取出现次数最多区县的留言明细 excel 表后，经过分析，由于只需要按照地点/人群划分排名前五的热点问题，因此我们选取了其中的 A7 县、A1 区、A2 区、A4 区及 A 市这 5 个区县建立 LDA 主题模型[5]，从每个区县中找出使混淆度最低时的一个一个主题，以便可根据主题关键词进行热点问题查找。以下是 LDA 主题模型提取出的主题及它的 20 个关键词：

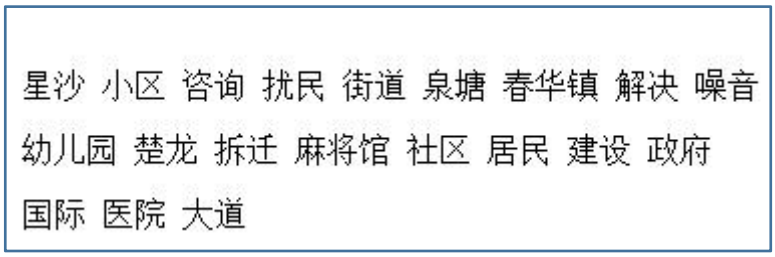

图 4-22　A7 县 topic

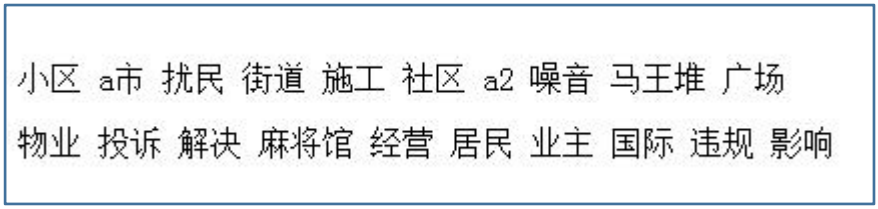

图 4-23　A1 区 topic

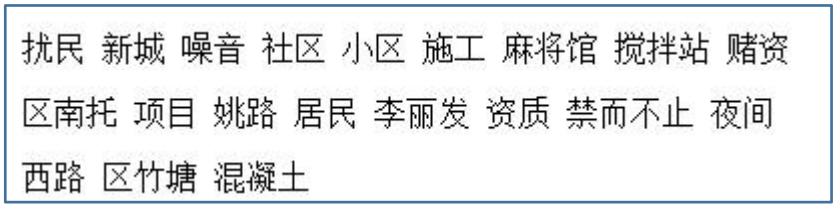

图 4-24　A2 区 topic

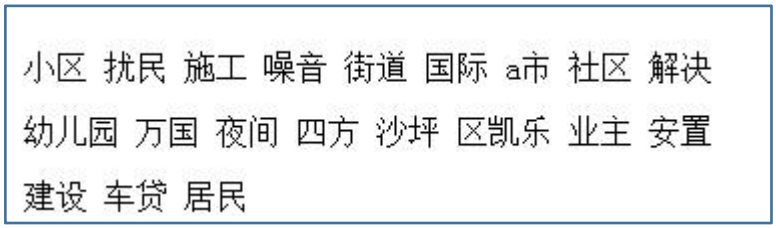

图 4-25　A4 区 topic

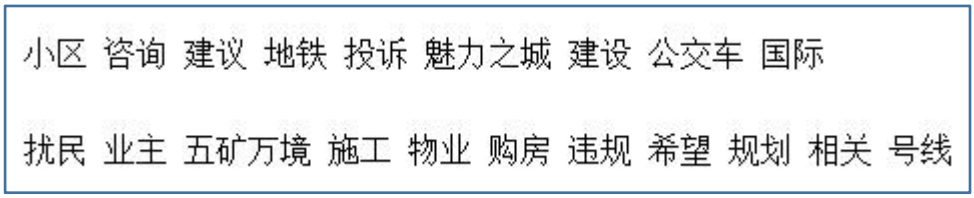

图 4-26　A 市 topic

4.2.5 提取热点问题

根据主题关键词，我们用 excel 表格进行相关搜索，然后结合群众的留言条数、点赞数、反对数按照 1:1:1 的权重以（权重 1*留言条数+权重 2*点赞数-权重 3*反对数）加权计算出各问题的热度指数，选出排名前五的热点问题，最后运用 excel 的排序和筛选功能选出时间范围并把地点/人群、问题描述概括出来如下图：

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	2110	2019/06/20至2019/09/19	A市五矿万境K9县群租房	房屋出现一系列问题急需解决
2	2	828	2019/01/14至2019/02/21	A4区公安派出所	派出所没给58车贷案进展通报
3	3	32	2019/02/15至2020/01/07	A7县春华镇多地区	麻将馆违规营业深夜噪音扰民
4	4	31	2019/07/21至2019/09/25	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民
5	5	24	2019/11/13至2020/01/15	A2区丽发新城附近	小区违规乱建搅拌站噪音扰民

图 4-27　热点问题表

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
1	208636	A00077171	A市A5区汇金路五矿万境	2019/8/19 11:34:04	我是A市A5区汇金路五矿万境K9县24栋的一名	2097	0
1	234086	A00099869	A市五矿万境K9县房子的	2019/6/20 9:30:44	五矿万境K9县的房子又出问题了，又是墙壁	6	0
1	215507	A00010323(	A市五矿万境K9县存在严	2019/9/12 14:48:07	预交房23栋没有通往负一楼的楼梯，存在严	1	0
1	252650	A00010531	A市五矿万境K9县交房后	2019/9/11 15:16:02	尊敬的相关部门，本人家庭于2018年购置A市	0	0
1	262599	A000100428	A市五矿万境K9县房屋出	2019/9/19 17:14:49	我是西地省A市五矿万境K9县的业主，2016年	0	0
1	275491	A00061339	A市五矿万境K9县负一楼	2019/9/10 9:10:22	关于五矿万境·K9县负一楼面积缩水的反馈，	0	0
2	220711	A00031682	请书记关注A市A4区58车	2019/2/21 18:45:14	尊敬的胡书记：您好!A4区p2p公司58车贷，	821	0
2	214238	A00061787	请问A4区公安派出所对5	2019/1/20 22:28:40	标题：恳请市委胡书记督促A4区经侦大队和	2	1
2	272413	A000106062	西地省A市58车贷恶性退	2019/1/14 20:23:57	西地省58车贷邪ze恶性退出，潜逃员国已半	2	0
2	264119	A00084445	58车贷立案五个月过去，	2019/1/19 9:47:23	58车贷立案五个月过去，A4区公安分局未公	0	0
3	239650	A00043724	A7县春华镇麻将馆好多，	2019/2/17 19:44:47	A7县春华镇春华垃圾站隔壁胡细国麻将馆打	6	0
3	237409	A00050277	A7县M9县城三期小区麻	2019/2/15 23:40:54	我是A7县M9县城三期的小区18栋的业主，我	4	1
3	268997	A00039048	A7县星沙丁家岭安置区5	2019/11/7 23:15:02	星沙丁家岭安置区5栋沙县小吃对面的吉员高	3	0
3	188451	A00013004	A7县春华镇石塘铺村有5	2019/4/11 17:54:25	我是春华镇一名村民，最近接到政府的通知，	2	0
3	288526	A00088728	何时解决A7县奥镇麻将馆	2019/5/13 11:28:08	张县长，您好，我是A7县金井镇居民，在我	2	0
3	218587	A00055713	A7县原灰埠大市场小区4	2019/9/6 0:15:15	原灰埠大市场小区47栋一楼麻将馆基本上每	1	0
3	191153	A909097	A7县泉塘小区麻将馆扰	2019/12/15 18:43:57	我们是西地省A市A7县泉塘小区A区26—28栋	0	0
3	198493	A00011876	A7县棚梨领东汇小区一	2019/7/14 7:36:50	A7县棚梨领东汇小区3栋2单元5楼505室非法	0	0
3	206197	A000102986	A7县星沙恒大翡丽华庭	2020/1/7 12:11:48	您好，我是恒大翡丽华庭的业主，作为A7县	0	0
3	211453	A00011876	A7县领东汇小区的非法	2019/9/16 17:29:44	A7县棚梨街道高峰社区领东汇小区3栋2单元	0	0
3	225880	A00011876	A7县棚梨领东汇小区非	2019/7/8 13:28:12	A7县棚梨领东汇小区3栋2单元5楼505室非法	0	0
3	240110	A00035029	A7县泉塘街道蔻天雅郡	2019/7/30 15:12:25	尊敬的县领导：我是A7县星沙镇泉塘街道蔻	0	0
3	243626	A00056669	A7县好望谷云邸2栋一楼	2019/9/20 10:14:19	您好，政府领导，我是住在好望谷云邸3栋的	0	0
3	272567	A00011876	A7县棚梨领东汇小区非	2019/5/22 14:56:18	A7县棚梨领东汇小区3栋2单元5楼505室非法	0	0
3	285102	A00046691	A7县蔻绣城麻将馆夜夜	2019/6/26 8:53:00	A7县开元路蔻绣城d05栋3单元一楼麻将馆上	0	0
4	284147	A909113	A5区劳动东路魅力之城	2019/07/21 10:29:36	局长：　　你好，A5区劳动东路魅力之城小	3	0
4	236798	A00039089	A5区劳动东路魅力之城	2019/07/28 12:49:18	尊敬的政府：A5区劳动东路魅力之城小区临	4	0
4	272122	A909113	A5区劳动东路魅力之城	2019/08/01 16:20:02	局长：　　你好，A5区劳动东路魅力之城小	6	0
4	287386	A909116	A市万科魅力之城商铺无	2019/08/18 14:44:00	A市万科魅力之城小区自打交房入住后，底层	0	0
4	246362	A909114	A市魅力之城小区底层商	2019/08/26 01:50:38	2019年5月起，小区楼下商铺越发猖张，不仅	0	0
4	242792	A909115	A5区魅力之城小区一楼	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两	1	0
4	245136	A909117	万科魅力之城小区底层	2019/09/04 21:00:18	您好：我是万科魅力之城小区的业主，小区	0	0
4	195095	A00039089	魅力之城小区临街门面	2019/09/05 12:29:01	魅力之城小区楼下烧烤摊、快餐店无证经营，	3	0
4	268914	A0006238	A5区劳动东路魅力之城	2019/09/10 06:13:27	A5区劳动东路魅力之城小区，底层有几家餐	0	0
4	246598	A00054842	A5区劳动东路魅力之城	2019/09/25 00:31:33	A5区劳动东路魅力之城小区临街夜宵摊、烧	1	0
4	232892	A00015335	A市万科魅力之城开发商	2019/1/8 9:54:00	您好！我是57栋业主，2018年12月25日万科	1	0
4	198084	A00022429	A市万科魅力之城小区近	2019/10/23 15:01:30	本人花近180万购买万科魅力之城的房子一套	0	0

图 4-27　热点问题留言明细

4.3 问题 3 — 答复意见的评价方案

本研究针对附件 4 中相关部门的答复意见，先后通过定义质量指标、人工评分得到训练样本、提取文本特征及建立逻辑回归分类模型，预测测试集的答复意见是否存在相关性 correlation、完整性 integrality 及可解释性 interpretability，最终实现从答复的相关性、完整性及可解释性对答复意见的质量进行评价。

4.3.1 答复意见质量的相关指标

我们对答复意见质量的相关性 correlation、完整性 integrality 及可解释性 interpretability 三个性质定义指标，若相关部门的答复有相关性则为 1，否则为 0；若答复是完整的则为 1，否则为 0；若答复具有可解释性则为 1，否则为 0。

表 4-1　答复意见的质量指标

	correlation	integrality	interpretability
有	1	1	1
无	0	0	0

4.3.2 训练样本的建立

本研究中，我们从附件 4 的答复意见中随机抽取全部数据的 0.25 得到 704 条答复意见的数据，对其进行人工评分，将评分后的数据作为训练样本 train（详细训练样本请见附件），如下图所示：

	A	B	C	D	E
1	留言详情	答复意见	correlation	integrality	interpretability
2	A7县安沙十地	网友“UU008144	1	1	1
3	书记你好，问	网友“UU008551	1	1	1
4	黎书记您好，	“UU008864”：	1	1	1
5	尊敬的领导：	网友“UU008394	1	0	0
6	我是西地省M2县20	网友：您好，留	0	0	0
7	易书记，您好	网友“UU008106	1	0	0
8	关于西地省B9市均	网友：您好！收	1	1	1
9	本人经常往返D1区	网友：您好！请	0	0	0
10	1、农贸市场摊位	网友：您好！您	1	1	1
11	我是一名消费者，	网友“UU008191	1	1	1
12	L10县县是否	亲爱的网友：	0	0	0
13	尊敬的市委市	网友：您好！留	0	0	0
14	C4市大部分县	尊敬的网友：	1	1	1
15	1.希望村里出	网友“UU008238	0	0	0
16	书记您好！我	网友“UU008226	1	1	1
17	尊敬的领导	网友“UU008130	1	0	0
18	尊敬的市长您好！	网友：您好！您	1	1	1
19	书记您好！	网友“UU008553	1	1	1
20	M市公路局改建32	网友：您好，留	0	0	0
21	尊敬的曾书记：曾	网友“UU008904	1	1	1

图 4-28　训练样本 train

4.3.3 CountVectorizer 提取文本特征

CountVectorizer 是通过 fit_transform 函数将文本中的词语转换为词频矩阵，矩阵元素 a[i][j] 表示 j 词在第 i 个文本下的词频，即各个词语出现的次数。get_feature_names()可看到所有文本的关键字；ocabulary_可看到所有文本的关键字和其位置；toarray()可看到词频矩阵的结果。本研究中，我们通过使用该技术可以提取到训练样本与测试数据的文本特征。

（1）CountVectorizer(ngram_range=(1, 2), min_df=3, max_df=0.9, max_features=100000)参数说明：

① ngram_range：词组切分的长度范围。

② max_df：可设置为范围在[0.0，1.0]的 float，也可设置为没有范围限制的 int，默认为 1.0。这个参数的作用是作为一个阈值，当构造语料库的关键词集的时候，如果某个词的 document frequence 大于 max_df，这个词将不会被当作关键词；如果这个参数是 float，则表示词出现的次数与语料库文档数的百分比；如果是 int，则表示词出现的次数。

③ min_df：类似于 max_df，不同之处在于如果某个词的 document frequence 小于 min_df，则这个词不会被当作关键词。

④ max_features：默认为 None，可设为 int，对所有关键词的 term frequency 进行降序排序，只取前 max_features 个作为关键词集。

4.3.4 答复意见质量评价方案的建立

在依照质量指标进行人工评分之后，通过 CountVectorizer 提取到文本特征，可以将答复意见的评价归为分类预测的挖掘过程。针对问题 3，我们选取了 Logistic Regression 逻辑回归模型[6]对相关部门的答复意见质量进行分类，最终实现评价附件 4 中相关部门答复意见的质量（于附件 result 可见）：

	A	B	C	D
1	答复意见	correlation	integrality	interpretability
2	现将网友在平台《问政西地省	1	1	1
3	网友“A00023583”：您好！我	1	1	1
4	市民同志：你好！您反映的“	1	1	1
5	网友“A000110735”：您好！	1	1	1
6	网友“A0009233”，您好，您	1	1	1
7	网友“A00077538”：您好！我	1	1	1
8	网友“A000100804”：您好！	1	1	1
9	网友“WU00812”您好！您的信	1	1	1
10	网友“WU008792”您好！您的	1	1	1

图 4-29　答复意见的评价

5. 结果分析

5.1 群众留言分类的模型评估与分析

5.1.1 混淆矩阵评估

选取平均准确率最高的 Linear Support Vector Machine(线性支持向量机)[7]作为群众留言分类模型后，我们针对该模型查看混淆矩阵，检测预测标签和实际标签之间的差异。在混淆矩阵中，主对角线表示预测正确的数量，除主对角线外其余都是预测错误的数量。

从混淆矩阵可以看出，"环境保护"类预测错误的数量最少、预测得最准确，“城乡建设”和“劳动和社会保障”预测的错误数量较多，线性支持向量机模型的预测准确率 accuracy 为 90.16%。

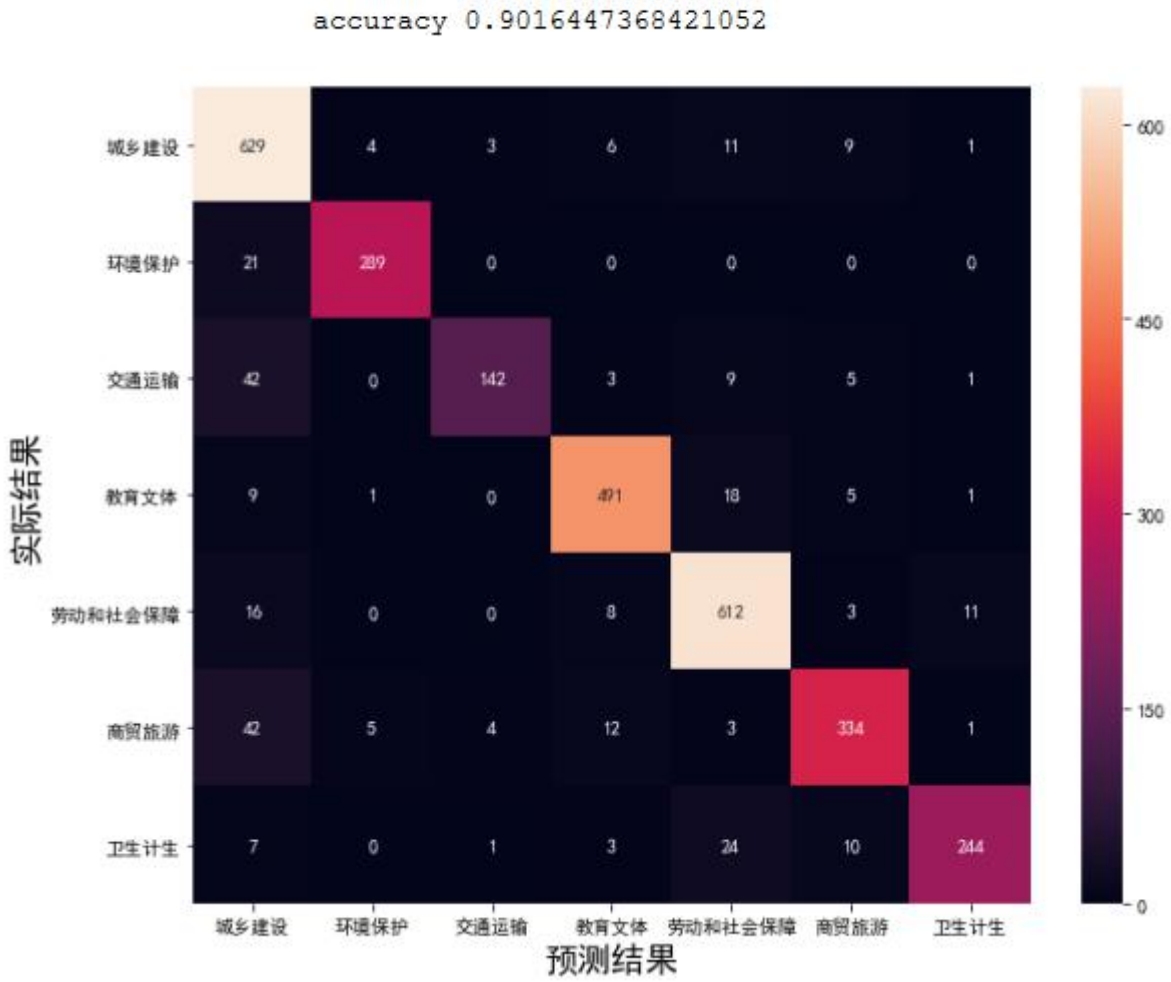

图 5-1　　混淆矩阵

5.1.2 F_1-score 评估

在对原始数据集进行查看过程中，我们已经知道数据的分布是很不均匀的。由于多分类模型训练数据的不均匀，准确率不能反映出模型的实际预测精度，因而通常需要借助 F_1 分数、ROC 等指标来评估模型。其中，F_1 分数是统计学中用来衡量二分类模型精确度的一种指标；它同时兼顾了分类模型的准确率和召回率。F_1 分数可以看作是模型准确率和召回率的一种加权平均，它的最大值是 1，最小值是 0，越接近 1 说明模型的准确率越高。

本研究中，我们将查看各个类的 F_1 分数。如下图所示，在“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”及“卫生计生”7 个类别中，“环境保护”的 F_1 分数最大为 95%，预测效果很准确；“交通运输”的 F_1 分数最小为 81%，预测效果准确。可见，7 个类别的 F_1 分数均已超过 80%，十分接近 1，说明使用线性支持向量机的分类留言的效果很好。

	precision	recall	f1-score	support
城乡建设	0.82	0.95	0.88	663
环境保护	0.97	0.93	0.95	310
交通运输	0.95	0.70	0.81	202
教育文体	0.94	0.94	0.94	525
劳动和社会保障	0.90	0.94	0.92	650
商贸旅游	0.91	0.83	0.87	401
卫生计生	0.94	0.84	0.89	289
avg / total	0.91	0.90	0.90	3040

图 5-2　F_1 分数

5.1.3 预测错误分析

由混淆矩阵可知，绝大多数预测结果都在对角线上（预测标签=实际标签），但是存在一些错误分类，我们发现这是由于以下预测错误的情况所造成的：

环境保护 预测为 城乡建设 ： 21 例.

	cat	review
2160	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t 我是一名在省会A市工作D市人之一，我和很...
2195	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的龚书记，你好！本人家住E市E2区宝庆西...
2071	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t2019年9月10日，B市生态环境局在《关于...
2041	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t在含浦镇开心农场匝道附近，天气一放晴就大量烧...
2064	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的黎局长： 您好！ A3区施家港社区...
2337	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t J市二医院家属区内，有人饲养家禽，臭气熏...
2067	环境保护	\n\t\t\t\t\t\n\t\t\t\t\tA市A4区雅居乐花园小区一楼商铺在环评报告明...
2664	环境保护	\n\t\t\t\t\t\n\t\t\t\t\tA市A4区大安路北多年，垃圾化工，渣土堆放，...
2045	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t 家住在A3区后湖小区，该小区成立好多没有...
2091	环境保护	\n\t\t\t\t\t\n\t\t\t\t\tB9市浦建混凝土有限公司每天工作到两三点，甚...
2507	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t K10县鼎丰宾馆的2吨无任何吸尘设备的烟...
2056	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t局长： 你好，我是星沙经济开发区丁家岭安置...

图 5-3 环境保护 预测为 城乡建设

交通运输 预测为 城乡建设 ： 42 例.

	cat	review
3220	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 沿河路在水一方处到L9县路连接的一段，也...
2947	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t地处月亮岛街道时代倾城小区一二三期之间的公共...
3460	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的谭书记：您好！ 我认为I4县交通令人...
3067	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 近来，每当夜幕降临，老百姓出门慢步休闲在...
3351	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 我们因为工作的原因经常要去中国邮政局A6...
3541	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 1.隧道技术已成熟。建成合武高速那样连续...
3148	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t我是F8市弼时镇湄江村村民彭元中，我家是建档...
3502	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t地铁1号线从南湖路站-马厂站偏离人流量极大的...
3024	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t C市金薮追远乡镇街道下雨污水严重，路面狭...
2984	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t周县长： 您好！我是朱良桥乡权梓桥村草坝子...
2950	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t楚府路快速化改造在2016年启动，说是201...
3231	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t L12市岩垅乡江丘村三组没有通水泥路，何时...

图 5-4 交通运输 预测为 城乡建设

教育文体 预测为 城乡建设 ： 9 例.

	cat	review
3817	教育文体	\n \n 虽然某些学院的卫生比较好，但在食堂了看到某些菜放在水里进下就...
4283	教育文体	\n\t\t\t\t\t\n\t\t\t\t\t M3县体育馆应走发展公共体育事业的正路 ...
4143	教育文体	\n \n 尊敬叶书记： 一直以来，M9县政府及旅游局都在立志把M9县...
4449	教育文体	\n \n 请求政府及时接收"恒大城"小区配套幼儿园并尽快开办公立园报告...
3674	教育文体	\n \n 对H市教育方面的几点提议 1 青少年宫建设问题 2 ...
5081	教育文体	\n \n 尊敬的领导：我是F市一个鑫科.惠风园小区的业主。买房住进来3年发...
4320	教育文体	\n\t\t\t\t\t\n\t\t\t\t\t本人想了解下M14县新修的体育馆是否已免费向...
5011	教育文体	\n \n 做为家乡人，看了看到杭州良渚文化遗址申请世界遗址成功，我就想到我...
4899	教育文体	\n \n 前几天，本人有幸去B市动力谷园区里面的B市B4区公益图书馆看书，...

图 5-5　教育文体 预测为 城乡建设

劳动和社会保障 预测为 城乡建设 ： 16 例.

	cat	review
5657	劳动和社会保障	\n \n 尊敬的书记： 你好!我在工作好几年，随着年龄的增长，渴望有自己...
5325	劳动和社会保障	\n \n J市自来水原是家有60余年历史的国有企业，位于体育路2号，J...
6703	劳动和社会保障	\n \n 我是人才落户A市，如今已经在A市工作和买了社保，本想在A市买房安...
7094	劳动和社会保障	\n\t\t\t\t\t\n\t\t\t\t\t 彭主任： 您好，我是西地省人，一直在A...
6001	劳动和社会保障	\n \n 我是A4区捞刀河街道罗汉庄村安置小区社员，2005年被A市苗...
6440	劳动和社会保障	\n \n 我是K2区花桥街镇石塘村铁里冲，我爷爷吴因存今年80岁了贫困...
7044	劳动和社会保障	\n \n 您好！本人于明年硕士毕业，目前户籍在学校，计划毕业回长工作。想咨...
5573	劳动和社会保障	\n \n 领导： 你好！我是去年八月份拿的驾照，可这补贴快一年了，去A9...
5814	劳动和社会保障	\n \n 我本人的工作地在A市,国家规定的各项社保和公积金都是缴纳在A...

图 5-6　劳动和社会保障 预测为 城乡建设

商贸旅游 预测为 城乡建设 ： 42 例.

	cat	review
8009	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t尊重的领导：你好！对于大汉二期翡翠湾房屋质量...
7279	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t市长先生： 你好！ 在B9市一个小区买了...
7229	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：我于2018年5月5日在A市A5...
8287	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t前几天看新闻说天然气会要涨价，所以特地关注了...
7657	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\tL5县交通两大工具：1，火车，以前县城火车很...
8146	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 最近跟机动车检测站的朋友一起吃饭，聊起了...
7742	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 各位政府领导好，苍冲村到兰里街上不到20...
8086	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 西地省A9市华盛彩虹城，至少百分之40的...
8043	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 华嘉地产公司汇金城二期项目于2018年1...
7313	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 书院路长征新村和鑫源花苑入口处，有一小摊...
7945	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\tA市安置房存在质量问题电梯，发生多次掉层事故...
7660	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t请问L5县至低庄的公交车为啥又涨价了？请问有...
7599	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t近来得知消息,K市的公交车全部实行一波涨价政...
8101	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 我购买的A市世茂房产，开发商于0000-...
7420	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t我想知道E市夏季特色旅游线路是什么？想带小孩...

图 5-7　商贸旅游 预测为 城乡建设

卫生计生 预测为 城乡建设 ： 7 例.

	cat	review
8345	卫生计生	\n \n 局长您好 我是A8县药店的，我们是3月份申请过的GSP，5...
8401	卫生计生	\n \n 甲亢即碘多.我儿甲亢.食非碘盐时病情正常.是否意味是碘盐导致...
9201	卫生计生	\n \n 我开的一个小型餐馆为什么到食品药品监督局去办个卫生许可证要3...
9139	卫生计生	\n \n 我是金沙湾业主，有关宁永健康体检未取得相关合法手续在金沙湾北门转...
9196	卫生计生	\n \n 领导您好：我要咨询这个民生问题也携及到一部分群众的心声，就是农村...
8952	卫生计生	\n \n 本人在K8县中医院做透视，为什么以前不用交钱，现在要我们交...
8469	卫生计生	\n\t\t\t\t\t\n\t\t\t\t\t 尊敬的政府领导： 自C4市（居）“两委...

图 5-8　卫生计生 预测为 城乡建设

城乡建设 预测为 环境保护 ： 4 例.

	cat	review
1925	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t 今年春节回家探亲，经过走访千山红镇部分居民...
499	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\tC3县城管局：雪松北路烧烤摊已存在有两年之久...
232	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t 我是大成桥乡玉新村一村民，下面有一事想像...
797	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\tJ6县城北华谊驾校，往荞麦塘村200米方向，...

图 5-9 城乡建设 预测为 环境保护

商贸旅游 预测为 环境保护 ： 5 例.

	cat	review
7831	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t 德润广场广告电子屏经常通宵播放广告，夜里...
7752	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的县政府领导，你好，我是L5县城南一居民...
7268	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的胡书记：我是来龙门街道丁家坊社区中旦组...
7376	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\tD8县白果镇街道及花果山景区垃圾没有经过任何...
8129	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t 成立该公司的前身是关闭了几年的I市达宏水...

图 5-10 商贸旅游 预测为 环境保护

商贸旅游 预测为 交通运输 ： 4 例.

	cat	review
7623	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t周局长：我于2018年7月在上汽大众K市卫高...
7812	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t出租车，车用天然气二个月内共提价3次，共涨价...
8177	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t E市出租车加气站的加气设备为什么没有检验...
7766	商贸旅游	\n\t\t\t\t\t\t\n\t\t\t\t\t\t出租车，车用天然气二个月内共提价3次，共涨价...

图 5-11 商贸旅游 预测为 交通运输

城乡建设 预测为 教育文体 ： 6 例.

	cat	review
1419	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t我是L5县葛竹坪镇中心小学的一名教师。儿童节...
1609	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t 2014年12月3日，我在平台百姓呼声发...
697	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的领导：您好！我是一名桃林寺镇中心小学的...
696	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t李市长：您好,华荣的2路公共汽车,每天下午6...
1732	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t蒋厅长： 你好！我在网上看到你是F...
169	城乡建设	\n\t\t\t\t\t\t\n\t\t\t\t\t\t 星沙城北新区近几年在政府的主导下发展迅速...

图 5-12 城乡建设 预测为 教育文体

劳动和社会保障 预测为 教育文体 ： 8 例.

	cat		review
5433	劳动和社会保障	\n \n	现在孩子都是被爷爷奶奶宠大的，不管在哪只要受一点点的委屈...
6450	劳动和社会保障	\n \n	贺厅长，您好！我是今年5月份参加A市心理咨询师二级考试的考生...
5287	劳动和社会保障	\n \n	你好，彭厅长！我是J市技师学院的一名学生，2010年就读于J...
7027	劳动和社会保障	\n \n 我是于2005年在外省毕业中医学类的中专学生，因为这么多年都没有...	
6592	劳动和社会保障	\n \n	I6市市2017年教师绩效工资百分之三十奖励部分已经到位（其...
6297	劳动和社会保障	\n \n	A8县教育局在给教职工生育保险缴费方面，一直欠缴，整个教育系...
6021	劳动和社会保障	\n \n	我们是A8县电影发行放映公司（以下简称电影公司）聘用的12名...
6134	劳动和社会保障	\n \n	敬爱的领导，你们好。我是15年毕业并参加护士资格考试的一名护...

图 5-13　劳动和社会保障 预测为 教育文体

商贸旅游 预测为 教育文体 ： 12 例.

	cat		review
7714	商贸旅游	\n \n	L市通K6县塘头古侗寨必须尽快加强保护　何彬　近几年来，...
7406	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的郭书记：　一天一杯奶，是国家"学生饮...	
8255	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t	浮邱山位于I市I4县浮邱山乡和高桥乡境内...
7907	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t首先，该公司涉及虚假宣传。该公司用"市场缺口...	
7989	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t今天我到淮阳中学接我家孩子放学。居然发现学校...	
8082	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t	A3区第十五幼儿园与2017年6月份发给...
8251	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t	L市通K6县塘头古侗寨必须尽快加强保护　...
7423	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的唐书记：　你好！E9县...	
7231	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t2019年12月11日稻田中学午餐红辣椒炒肉...	
7496	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t	党的十八大五中全会，提出了"绿色"新理念...

图 5-14　商贸旅游 预测为 教育文体

城乡建设 预测为 劳动和社会保障 ： 11 例.

	cat	review
1771	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t厅长先生： 我的住房是1979年所建单位房...
1780	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 高厅长您好，现在国家提出要统筹城乡发展，...
275	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 请求延长311路至洋湖，以便西地省建筑设...
1455	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导： 我叫肖华立2013年9月2日...
606	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 我们是西地省E市E7县的新就业人员，首先...
1638	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t1:拨打12329，直接响应"请拨号"，也没...
1338	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的县委书记： 你好！我想问一下我们县新...
1450	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t M2县走马街镇中心卫生院的国债项目职工集...
1417	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 二凉亭园艺场马上要进行农垦区危房改造，这...
1810	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t局长大人： \r\n 您好！\r\n 我...

图 5-15 城乡建设 预测为 劳动和社会保障

交通运输 预测为 劳动和社会保障 ： 9 例.

	cat	review
3367	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 邮政局是百货超市吗？作为储蓄营业员为什么...
3529	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 感谢西地省交通运输厅下文明令西地省高速公...
3178	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 经过候车室要收钱，只有一句话：那是我们全...
3288	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的李市长： 目前M7市城里安装了交通违...
3535	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t省交通运输厅： 由于前任领导官僚主义严重，...
3458	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 今天去银行修改一下号码，由于填错了一张纸...
3308	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t但凡过年过节，只要有东西要卖就打出各种旗号，...
3298	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 您好!我想请问您一下有关如何出口货物的问...
3425	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t我的社保卡，去年在下发到人时因我外出有事，未...

图 5-16 交通运输 预测为 劳动和社会保障

教育文体 预测为 劳动和社会保障 ： 18 例.

	cat	review
4126	教育文体	\n\n尊敬的市长： 你好！百忙之中打扰一下。我是西地省J10 县渡口乡石云村万选组陈德啟...
4242	教育文体	\n \n 一直以来，我们给学生缴纳医保有两个选择，学校或社区，且缴纳标...
4559	教育文体	\n \n 尊敬的李书记： 向你汇报K6县祁剧团的基本情况以及现状： ...

图 5-17 教育文体 预测为 劳动和社会保障

卫生计生 预测为 劳动和社会保障 ： 24 例.

	cat	review
8611	卫生计生	\n \n 我妻子是省医保的缴费者，今天到A市口腔医院补牙（属于省医保的...
9191	卫生计生	\n \n 您好！向您咨询个问题，请在百忙中回复为盼!根据国家卫生健康委员会...
8576	卫生计生	\n \n 2014年卫生系统高级职称报名时间：5月5日-5月11日24...
9158	卫生计生	\n \n 2019.9.12日，西地省卫健委职改办发布了一份“假冒期刊”，...
8660	卫生计生	\n \n 詹鸣书记： 请问2014年主治医师考试西地省内乡镇基层分数线何...
8780	卫生计生	\n \n 正在楚雅附一医院感染科52床进行紧急救治的E市E7县六都寨镇...

图 5-18 卫生计生 预测为 劳动和社会保障

城乡建设 预测为 商贸旅游 ： 9 例.

	cat	review
947	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t从2016年东山景区开园以来，东山大小的古樟...
782	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t戴局长： 您好！现在在G市进行两工地设备检...
419	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t建设中路阳光新城小区是由发城集团进行开发的，...
972	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t K市江南幸福里第三期没有安装燃气，请政府...
1346	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tL2县的天然气管道已经安装好几年了，请问什么...
1598	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 我是一驻“西地省电京物业有限公司”提供物...
645	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t上周末，偶尔发现昭阳公园儿童游乐场天价收费：...
1747	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的蒋厅长： 您好，我是一名在外...
1797	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t I2区城乡建设局近日发布保障性住房电梯公...

图 5-19 城乡建设 预测为 商贸旅游

交通运输 预测为 商贸旅游 ： 5 例.

	cat	review
3301	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 每天晚间在楚江A市段楚江一桥至二桥之间都...
3019	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t作为新建的客运站，无论是硬件还是软件应该说是...
3177	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 邮政储蓄银行K市分行顶风违反中央八项规定...
3236	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 各位L3县领导，我是一名H市的钓友，因借...
3187	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 尊敬的领导，您好！ 我想向您反应一个细...

图 5-20 交通运输 预测为 商贸旅游

教育文体 预测为 商贸旅游 ： 5 例.

	cat	review
4239	教育文体	\n \n I市，是一片山清水秀，人杰地灵的美丽热土，这里是中华人文始祖...
4724	教育文体	\n \n 金花村位于K1区古城西南约四十公里的梳子铺乡三近田洞中，村中古院...
4149	教育文体	\n \n 各位乡亲挚友我回来了！我从大陆内地「H市」旅游考察平安的...
3844	教育文体	\n \n D5区要走向国际化，还有很长的路要走：\n\n 一是要深入...
4925	教育文体	\n \n 2017年3月我经高中老师介绍以八千元的价格在河西仁和会计培训机...

图 5-21　教育文体 预测为 商贸旅游

卫生计生 预测为 商贸旅游 ： 10 例.

	cat	review
8674	卫生计生	\n \n F7县仁康堂中药饮片有限责任公司，是全国最典型严重违法违规企...
8969	卫生计生	\n \n 国家食品药品监督管理总局令第17号第三章16条包含：申请豫包...
8744	卫生计生	\n \n 我爱人于本月中旬在长株潭超市门口购买白板病死猪肉，事后我和爱...
8681	卫生计生	\n \n 肖局长： 你好！我是K市K10县人，我想问一下，我们那里那么多...
8368	卫生计生	\n \n 局长好，https://baidu.com/房以火疗为名大量...

图 5-22　卫生计生 预测为 商贸旅游

劳动和社会保障 预测为 卫生计生 ： 11 例.

	cat	review
5763	劳动和社会保障	\n \n E4县退休职工的独生子女费的手续已经办了四年了，但是为什么到...
5514	劳动和社会保障	\n \n M4市医保局关心病人不错，但是采取承包制的办法，现在对于慢性...
6924	劳动和社会保障	\n \n 近来我侄女曹赵媛因患再生障碍性贫血（B型RH阴性熊猫血型），从J...
5548	劳动和社会保障	\n \n 2013年6月接到举报信，反映我县江桥街道东塔社区卫生室文红...
5752	劳动和社会保障	\n \n 本人杨华荣，男，苗族，现年57岁，身份证号码是*******...

图 5-23　劳动和社会保障 预测为 卫生计生

6. 挖掘结论

本文基于群众留言与相关部门答复意见的文本数据，围绕群众问政留言分类及其若干关键模型与相关部门答复意见进行了一系列文本挖掘，实现了对留言及答复文本数据的分析，得到了具有一定价值的结果，这些结果对于政府实施智慧政务具有一定性的指导意义。

本文通过原始数据的清洗、分词、词频统计等数据预处理，建立了线性支持向量机模型并通过模型评估分析得出该模型能够实现对群众留言的精准分类，该结果说明政府可以利用线性支持向量机模型对电子政务系统中的群众留言进行准确地分类，以便及时将各类社情民意分派给相应的职能部门处理。本文还针对留言主题数据，通过正则表达式及 if 条件语句+for 循环语句使用 LDA 主题模型，结合群众的留言条数、点赞数、反对数提取出了前 5 个热点问题，以此促使相关部门能够有针对地处理，提升政府的服务效率，从而有效推动政府的管理水平和施政效率的提升。此外，本文运用了逻辑回归模型从答复的相关性、完整性及可解释性 3 个角度来评价相关部门答复意见的质量，利用该评价方案，政府可以检测职能部门的回复质量，起到上级监督及自我调节的作用。

7. 参考文献

[1]林崇责,黄炜.基于云计算电子政务一体化服务平台建设探索[J].计算机时代,2016（8）：63-64.

[2] 张梦笑.基于 LDA 模型的观点聚类研究[D].山西：山西大学，2012 年 6 月.

[3] 李新福，赵蕾蕾，何海斌，李芳.使用 Logistic 回归模型进行中文文本分类[J].《计算机工程与应用，2009 年 14 期.

[4] 徐爱华.面向文本分类中文文本挖掘技术研究及发现[D].武汉理工大学，2004 年 5 月.

[5] 石晶,范猛,李万龙.基于 LDA 模型的主题分析[J].自动化学报,2009,35(12),1587-1592.

[6] 李森.基于线性支持向量机的语言特征描述研究[J].《现代语文：下旬．语言研究》，2012 年，第 5 期 125-128 页.

[7] 徐华.基于支持向量机的 Web 文本挖掘研究[D].合肥：合肥工业大学，2004 年.