所选题目：“智慧政务”中的文本挖掘应用

综合评定成绩：_____

评委评语：

评委签名：

“智慧政务”中的文本挖掘应用

摘要：大数据、智慧城市、国家宏观政策为电子政务向智慧政务的转型提供了一个难得的历史机遇。智慧政务是电子政务发展的新阶段，“新”主要指政府在智慧的办公平台上，通过智慧的决策，为公众提供智慧的服务。但随着网络问政平台文本数据量的不断攀升，通过自然语言处理和文本挖掘的方法来解决相关问题变得越来越重要。

该论文基于留言内容分类三级标签体系以及留言主题、留言详情、答复意见等数据，利用逻辑回归、文本相似度等模型方法，一共解决了三个问题：1、建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价；2、将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，同时给出排名前5的热点问题以及相应热点问题对应的留言信息；3、针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

对于这三个问题，首先进行数据预处理和数据分析；然后通过相应的模型进行训练测试；最后画出实验流程图以及对结果进行分析。

对于问题1的群众留言分类，首先，对附件2文本进行预处理，通过数据增强来平衡样本数据，将长文本转化为短文本，利用 jieba 中文分词工具对留言主题进行分词，去除停用词，并通过 TF-IDF 算法提取每个留言主题的前 5 个关键词。然后，将多分类问题转化为二分类问题，建立留言内容的一级标签分类模型逻辑回归模型对数据进行训练和测试。最后，通过评价指标F-score对分类方法进行评价。

对于问题2的热点问题挖掘，首先，对附件3中的文本进行预处理，经过分词及去除停用词，从留言主题中提取有关地点和人群的关键词；然后，根据地点和人群的关键词，通过相似度整理出相似留言，结合相应的反对数和点赞数，定义热度评价指标；最后，将相似留言整理成“热点问题留言明细表”，进而根据热度评价指标，生成排名前5的“热点问题表”。

对于问题3的答复意见的评价，首先，对附件4答复意见文本进行预处理，经过分词及去除停用词提取关键词，通过判断答复意见与留言主题关键词的相似性，从而判断出答复的相关性。然后，通过对答复意见的观察，整理出一个模板，对比答复意见与模板的匹配度来判断答复的完整性。最后，判断答复意见与留言详

情关键词的相似性，从而判断出答复的相关性可解释性。

关键词：智慧政务；文本挖掘；逻辑回归；文本相似性；TF-IDF算法

目录

1. 问题的研究背景及问题分析

1.1 研究背景

随着时代的发展，科技的进步，将智慧政务运用到社会治理方法中已经是社会治理创新发展的新趋势，智慧政务是政务发展的高级阶段，其实质是将高效处置信息这一功能引入到现在的公共部门政务服务流程当中，进而实现政务个性化、大数据化与智慧化[3]。

信息技术的不断发展和应用，深刻地改变了人们的生产生活方式，并逐步渗透到城市治理的方方面面。随着城市治理理念与方式的转变，传统的电子政务模式已然无法跟上信息时代的步伐，智慧政务悄然兴起。在实践中，许多地方政府如北京、广州、南京等地都在积极应对信息技术的挑战，利用信息技术，促进电子政务向智慧政务的转变[4]。在学术研究中，大多认为，智慧政务是电子政务发展的新阶段[1]。

近年来，我国陆续发布了一系列推进信息化的文件，例如云计算、大数据、”互联网 +”等等，通过互联网的创新，来使之与国家各大领域进行融合，从而使得效率的改善和技术的提升得以体现，与此同时，也确定了一系列便民的”互联网 +”系统。在当今的社会背景之下，互联网的创新使得智慧政务的发展势在必行[1]。这也就意味着我们要具体问题具体分析，针对政府职能的转变，来强化”互联网 +”的创新思维，开始新的政务服务模式，秉持一切为了群众，从群众来到群众中去的态度，来推进社会治理的智能化。

根据文本数据挖掘的现状，文本挖掘已经成为解决各领域实践问题的有效手段之一，但相比国外而言，国内研究在智慧政务方面应用仍不广泛，研究数据源与方法较为局限，对规划实践的指导作用不强，仍有较大的发展空间。

随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

1.2 问题重述

问题一：群众留言分类

根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。

问题二：热点问题挖掘

根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”

问题三：答复意见的评价

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

1.3 问题分析

针对问题1，需要建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。通过观察发现，附件2中包含留言编号、留言用户、留言主题、留言时间、留言详情以及一级标签等数据，具体的求解步骤如下：

（1）对附件2文本进行预处理，通过数据增强来平衡样本数据，将长文本转化为短文本，利用 jieba 中文分词工具对留言主题进行分词，去除停用词，并通过 TF-IDF 算法提取每类留言主题的前 5 个关键词；

（2）将多分类问题转化为二分类问题，建立留言内容的一级标签分类模型逻辑回归模型对数据进行训练和测试；

（3）通过评价指标F-score对分类方法进行评价。

针对问题2，需提交的结果为“热点问题留言明细表”以及“热点问题表”。通过观察发现，附件3在附件2的基础上增加了反对数和点赞数两种数据，具体的求解步骤如下：

（1）对附件3文本进行预处理，经过分词及去除停用词，从留言主题中提取有关地点和人群的关键词；

（2）根据地点和人群的关键词，通过相似度整理出相似留言；

（3）综合考虑反对数、点赞数、相关留言数以及问题持续时间，定义热度评价指标，来衡量每个热点问题相关留言每天受到的关注度；

（4）将相似留言整理成“热点问题留言明细表”，进而根据热度评价指标，生成排名前5的“热点问题表”。

针对问题3，需要从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。通过观察发现，附件4在附件2的基础上增加了答复意见与答复时间这两种数据，具体的求解步骤如下：

(1)对附件4答复意见文本进行预处理，经过分词及去除停用词提取关键词；

(2)通过判断答复意见与留言主题关键词的相似性，从而判断出答复的相关性；

(3)通过对答复意见的观察，整理出一个模板，对比答复意见与模板的匹配度来判断答复的完整性；

(4)判断答复意见与留言详情关键词的相似性，从而判断出答复的可解释性。

2. 符号说明

为了方便模型建立和求解，我们定义本文的符号说明如表1所示：

表1 符号说明表

符号	定义
$Index_{Heat}$	热度指数
Sup_{num}	点赞数
Opp_{num}	反对数
Mes_{num}	相关留言数
Pro_{time}	问题持续时间
θ	相似度阈值

本文中用到的其它具体符号会在首次使用时加以说明。

3. 分析方法与过程

本次数据挖掘挑战赛举办方提供了四份数据，其中附件1为内容分类三级标签体系，附件2到附件4为有关留言内容与答复内容的信息。*.xlsx 格式为数据常规格式，本团队队员主要以Python进行数据处理与分析，故对数据格式无需进行额外的处理。

针对题目中所给的原始数据，通过数据分析发现，存在着长文本的无意义表达多、数据不平衡、关键词表达多样化等问题，这将会严重的影响数据挖掘和建模的执行效率，甚至可能会导致数据挖掘的结果出现偏差，因此，必须先对给定的原始数据进行文本预处理操作，提高数据集的质量，并使得数据能够更好地适应我们的挖掘算法和数学模型。

在经过数据分析与数据预处理操作后，便可建立相应的模型来进行训练和测试，进而通过评测指标来观察模型效果。

3.1 问题1分析方法与过程

3.1.1 数据分析

通过对所给的数据进行深入分析，并挖掘它们之间的关系，以便进一步地构建模型。为了更好地建立关于群众留言内容的一级标签分类模型，本文对附件1和附件2中的数据进行了分析。对于附件1提供的内容分类三级标签体系，一级分类包括城乡建设、党政服务等15种，在每个一级分类下依次有二级分类和三级分类，对留言内容进行了更进一步的细化。附件2包含留言编号、留言用户、留言主题、留言时间、留言详情以及一级标签等数据，共计9210记录，重点关注留言主题以及留言详情这两部分文本数据。

3.1.2 数据预处理

首先，通过数据增强来平衡样本数据。通过观察附件2的数据，可以发现存在样本不均衡的问题，不利于后续的操作，因此通过数据增强来平衡样本数据。

接着，将长文本转化为短文本。附件2中的文本数据存在大量的长文本，主要集中在留言详情部分。长文本存在许多无意义表达，因此要将其转化为短文本或者关键词，一方面可以减少后续的工作量，另一方面可以提高模型的精确度。

然后，对留言主题进行分词，去除停用词。

在对留言信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件2的表中，以中文文本的方式给出了数据。为了便于转换，先要对这些留言信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，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)，同时采用了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，对于未登录词，采用了基于汉字成词能力的 HMM模型，使得能更好的实现中文分词效果。

在分词的同时，采用了 TF-IDF 算法，抽取每个留言信息中的前 5 个关键词,这里采用 jieba 自带的语义库。

最后，通过 TF-IDF 算法提取每类留言主题的前 5 个关键词，生成 TF-IDF 向量。

在对留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把留言信息转换为权重向量。 TF-IDF 算法的具体原理如下：第一步，计算词频，即 TF 权重（ Term Frequency）。

$$词频(TF)=某个词在文本中出现的次数 \tag{1}$$

考虑到留言信息有长短之分，为了便于不同留言的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$词频(TF)=\frac{某个词在文本中的出现次数}{文本的总词数} \tag{2}$$

或

$$词频(TF)=\frac{某个词在文本中的出现次数}{该文本出现次数最多的词的出现次数} \tag{3}$$

第二步，计算 IDF 权重，即逆文档频率（ Inverse Document Frequency），需要建立一个语料库（ corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$逆文档频率(IDF)=\log(\frac{语料库的文本总数}{包含该词的文本数+1}) \tag{4}$$

第三步，计算 TF-IDF 值（ Term Frequency Document Frequency）。

$$TF-IDF = 词频(TF) \times 逆文档频率(IDF) \quad (5)$$

实际分析得出 TF-IDF 值与一个词在留言信息中文本出现的次数成正比， 某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的留言信息中文本的关键词。

生成 TF-IDF 向量的具体步骤如下：

(1)使用 TF-IDF 算法，找出每类留言信息的前 5 个关键词；

(2)对每类留言信息提取的 5 个关键词，合并成一个集合，计算每类留言信息对于这个集合中词的词频，如果没有则记为 0；

(3)生成每类留言信息的 TF-IDF 权重向量， 计算公式如下：

$$TF-IDF = 词频(TF) \times 逆文档频率(IDF) \quad (6)$$

3.1.3 逻辑回归模型

针对问题1，提出逻辑回归分类模型[5]。逻辑回归也就是逻辑回归的分析，一般情况是在流行性疾病学中应用比较多。逻辑回归是处理分类问题的经典算法，尤其针对多种属性分类，逻辑回归算法能以较快的速度迭代出最优解，下面首先了解一下逻辑回归。

逻辑回归在某些书中也被称为对数几率回归，逻辑回归用了和回归类似的方法来解决了分类问题，是常用地分类问题模型。

Sigmoid 函数是单调可微的连续函数，是回归方程的函数模型。函数如下：

$$\phi(z) = \frac{1}{1+e^{-z}} \quad (7)$$

如图 1 所示：

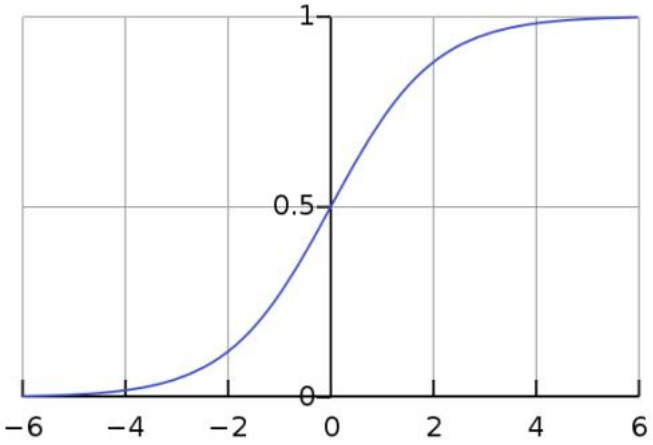

图1　Sigmoid函数模型

有了 Sigmoid 函数之后，由于其取值在[0,1]，我们就可以将其视为类 1 的后验概率估计 p(y=1|x)。简单的说，就是如果有了一个测试点 x，那么就可以用 Sigmoid 函数算出来的结果来当作该点 x 属于类别 1 的概率大小。

于是，非常自然地，我们把 Sigmoid 函数计算得到的值大于等于 0.5 的归为类别 1，小于 0.5 的归为类别 0。

$$\hat{y}=\begin{cases}1, if \phi(z)\geq 0.5\\ 0, otherwise\end{cases} \tag{8}$$

同时逻辑回归于自适应线性网络非常相似，两者的区别在于逻辑回归的激活函数是 Sigmoid 函数，而自适应线性网络的激活函数是 y=x。

接下来要做的就是根据给定的训练集，把参数 w 给求出来了。要找参数 w，首先就是得把代价函数(cost function)给定义出来，也就是目标函数。我们有：

$$p(y=1|x;w)=\varphi(w^T x+b)=\varphi(z) \tag{9}$$

$$p(y=0|x;w)=1-\varphi(z) \tag{10}$$

其中，$p(y=1|x,w)$ 表示给定 w，那么 x 点 y=1 的概率大小。上面两式可以写成一般形式：

$$p(y|x;w)=\varphi(z)^y(1-\varphi(z))^{1-y} \tag{11}$$

接下来我们就要用极大似然估计来根据给定的训练集估计出参数 w。

$$L(w)=\prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};w)=\prod_{i=1}^{n}(\varphi(z^{(i)}))^{y^{(i)}}(1-\varphi(z^{(i)}))^{1-y^{(i)}} \tag{12}$$

为了简化运算，我们对上面这个等式的两边都取一个对数。

$$l(w)=\ln L(W)=\sum_{i=1}^{n} y^{(i)}\ln(\varphi(z^{(i)}))+(1-y^{(i)})\ln(1-\varphi(z^{(i)})) \tag{13}$$

现在要求的是使得 l(w)最大的 w。只需要在l(w)前面加个负号就变成最小了，也就是代价函数。

$$J(w)=-\ln L(W)=-\sum_{i=1}^{n} y^{(i)}\ln(\varphi(z^{(i)}))+(1-y^{(i)})\ln(1-\varphi(z^{(i)})) \tag{14}$$

最后便是利用梯度下降法求参数，在开始梯度下降之前，了解到有一个很好的性质就是：

$$\varphi'(z)=\varphi(z)(1-\varphi(z)) \tag{15}$$

正因为这个性质，梯度的负方向就是代价函数下降最快的方向。借助泰特展开，得到：

$$f(x+\partial)-f(x)\approx f'(x)\cdot\partial \tag{16}$$

其中，$f'(x)$和∂为向量，那么这两者的内积就等于：

$$f'(x)\cdot\partial=\| f'(x)\|\cdot\|\partial\|\cdot\cos\theta \tag{17}$$

当$\theta=\pi$时，也就是∂在$f'(x)$的负方向上时，取得最小值，也就是下降的最快的方向。沿着负方向梯度下降：

$$w_j:=w_j+\Delta w_j,\Delta w_j=-\eta\nabla \mathrm{J(w)} \tag{18}$$

其中，w_j表示第 j 个特征的权重；η为学习率，用来控制步长。

在使用梯度下降法更新权重时，只要根据下式即可。

$$w_j:w_j+\eta\sum_{i=1}^{n}(y^{(i)}-\varphi(z^{(i)})x_j^{(i)} \tag{19}$$

此式与线性回归时更新权重用的式子极为相似，这也是叫逻辑回归的原因。当然，在样本量极大的时候，每次更新权重会非常耗费时间，这时可以采用随机梯度下降法，这时每次迭代时需要将样本重新打乱，然后用下式不断更新权重。

$$w_j = w_j + \eta\sum_{i=1}^{n}(y^{(i)} - \varphi(z^{(i)})x_j^{(i)} \quad (20)$$

也就是去掉了求和，而是针对每个样本点都进行更新。

3.1.4 实验流程图

数据集可以分为训练数据集和测试数据集，通有以下方法：自动定义，Bootstrap以及K倍交叉混合矩阵。本文选取K倍交叉混合矩阵，将数据集平分为K个大小相似的数据集，然后计算K次，通常使用一个数据集作为测试数据集，其他数据集用来建模。

对分类模型效果的评估是非常重要的一部分，通常给定一个数据集R，将其分为训练数据集 R_{train} 和测试数据集 R_{test}，而且 $R_{train} \cup R_{test} = R$。用训练数据集来建立评价模型，用测试数据集来检测评价模型。问题1使用 F-Score 对分类方法进行评价。

前面的工作中已经完成了数据的处理与准备，训练模型的制定以及检测的标准。问题1的实验流程图如图 2 所示：

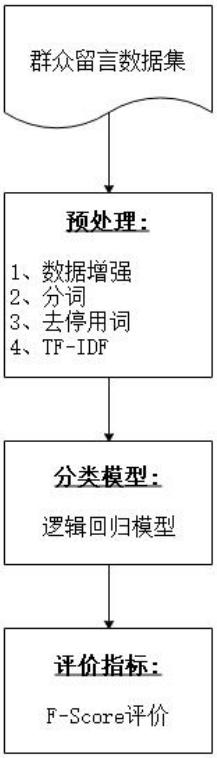

图2　问题1实验流程图

3.2 问题2分析方法与过程

3.2.1 数据分析

为了更好地进行热点问题挖掘，将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，本文对附件3中的数据进行了分析。附件3在附件2的基础上增加了反对数和点赞数两种数据，共计4326条记录，重点关注留言主题、留言详情、反对数以及点赞数这些数据。

3.2.2 数据预处理

对于问题2的数据预处理部分，与问题1类似，依次通过数据增强、长文本转短文本、分词以及去除停用词，最终得到有关特定地点和特定人群问题的关键词。

3.2.3 文本相似度计算

问题2需要对特定地点或特定人群问题的留言进行归类，在实验了许多聚类算法，如k-means聚类、层次聚类、FCM模糊聚类等算法后，实验效果均不理想。我们更换思路后采用Python中的Gensim库来对本文进行相似度计算，相似度大于阈值 θ 则将其归为相同类别。

具体分类步骤如下：

(1)对附件3中的所有数据进行预处理后，提取留言主题中的地点关键词(若留言主题中没有，则从留言详情中提取)；

(2)按照留言顺序，对两两留言关键词进行相似度计算，大于阈值θ，则视为相同类别，即属于同一热点问题；

(3)重复步骤(2)，已经归类的留言不需进行计算比较，直到所有留言都完成相似度计算。

经过实验和数据分析，阈值 θ 取值为0.3时实验效果最好，即相似度大于0.3的可视为同一热点问题。分类结果如表2所示：

表2 热点问题分类结果

问题	相似留言数	点赞数	反对数	时间跨度
A市房云时代建筑问题	4	56	5	2019/2/25至2019/11/16
A3区西湖街道茶场村五组拆迁问题	11	0	0	2019/1/6至2019/9/12
A4区绿地海外滩小区高铁扰民	6	691	0	2019/8/23至2019/9/6
A6区月亮岛路高压线	4	135	2	2019/3/26至2019/12/19
A市经济学院强制实习	9	1	13	2017/6/8至2019/11/27
A市梅溪湖教育问题	6	1765	5	2019/2/1至2019/11/22
A市三一大道交通问题	9	78	0	2019/2/20至2019/11/15
A市万家丽南路丽发新城搅拌场扰民	3	2	0	2019/11/13至2019/12/21
A市五矿万境K9县安全问题	9	2019	0	2019/1/15至2019/11/11
魅力之城小区扰民	21	18	18	2019/7/21至2019/12/4
伊景园滨河捆绑车位	40	23	1	2019/7/7至2019/9/1
西地省A市58车贷	13	2383	0	2019/1/11至2019/7/8

3.2.4 热度评价指标

对于热度评价指标的定义，本文主要从以下几个方面来考虑，如表3所示：

表3 热点问题综合评价指标体系表

	一级指标	二级指标	指标内涵
热点问题综合评价指标体系	问题特征热度影响力	相关留言数	热点问题相关留言数量
	内容特征热度影响力	热点问题及时性	热点问题相关留言第一条到最后一条的跨度时间
	受众特征热度影响力	点赞数	相关热点问题留言点赞总数
		反对数	相关热点问题留言反对总数

最终形成的热度评价指标的公式为：

$$Index_{Heat} = \frac{Sup_{num}+Opp_{num}}{Mes_{num}} * \frac{1}{Pro_{time}} \quad (21)$$

其中，$Index_{Heat}$表示热度指数即每个热点问题相关留言每天受到的关注度，Sup_{num}表示点赞数，Opp_{num}表示反对数，Mes_{num}表示相关留言数，$\mathrm{Pro_{time}}$表示问题持续时间。计算后，排名前5的热点问题分别为：A4区绿地海外滩小区高铁扰民、西地省A市58车贷案件、A市梅溪湖教育问题、A市五矿万境K9县安全问题和A6区月亮岛路高压线路设计问题。排名前5的热点问题见附件中“热点问题表.xls”，相应热点问题对应的留言信息，见附件中“热点问题留言明细表.xls”

3.3 问题3分析方法与过程

3.3.1 数据分析

为了更好地从答复的相关性、完整性、可解释性等角度对答复意见进行评价，并给出一套评价方案，本文对附件4中的数据进行了分析。附件4在附件2的基础上增加了答复意见与答复时间这两种数据，共计2816条记录，重点关注留言主题、留言详情以及答复意见这些数据。

3.3.2 数据预处理

对于问题3的数据预处理部分，与问题1类似，依次通过数据增强、长文本转短文本、分词以及去除停用词，最终得到有关留言主题、留言详情以及答复意见的关键词。

3.3.3 答复意见评价方案

问题3要求从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。本文对于这三个评价角度的定义如下：

相关性(*Pertinence*)：答复意见的内容是否与问题相关。这里使用留言主题与答复意见进行文本相似度匹配，判断答复意见是否与留言主题相同，匹配结果作为相关性评价指标。

完整性(*Integrality*)：是否满足某种规范。经过数据处理筛选后得出，答复格式应该为“称谓+您好/你好+意见+感谢+时间”的形式，缺少部分则认为答复不完整。

可解释性(*Interpretability*)：答复意见中内容是否对问题进行相关解释。这里使用留言详情与答复意见进行文本相似度匹配，判断留言详情中的问题在答复意见中是否都得到解释，匹配结果作为可解释性评价指标 。

通过计算可以得到相关性、完整性、可解释性各自的均值，然后将是否大于各自的均值作为判断标准，从而作为对于答复意见质量的评价方案。

4. 结果分析

4.1 问题1结果分析

问题1所需提交的结果为：建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。

本文建立的关于留言内容的一级标签分类模型为逻辑回归模型，通过对文本的分析得到各个一级分类对应的准确率和召回率，并计算出七个类别对应的F-Score的平均值作为最终的评价指标值。评价指标统计表如表4所示。

表4 评价指标统计表

一级分类	召回率	准确率
城乡建设	0.79167	0.84071
环境保护	0.94400	0.94400
交通运输	0.84426	0.90351
教育文体	0.92086	0.87671
劳动和社会保障	0.85217	0.89091
商务旅游	0.84956	0.80672
卫生计生	0.96491	0.90909
F_1	0.88078	

通过观察可以发现，问题1的F-Score值约为0.88。

4.2 问题2结果分析

问题2所需提交的结果为：将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，同时给出排名前5的热点问题以及相应热点问题对应的留言信息，并分别保存为文件“热点问题表.xls”和“热点问题留言明细表.xls”。

对于问题2，主要从相似性角度对特定地点或特定人群问题的留言进行归类。综合考虑反对数、点赞数、相关留言数以及问题持续时间，定义热度评价指标，来衡量每个热点问题相关留言每天受到的关注度。热点问题分类结果以及计算出的热度指数如表5所示：

表5 热点问题分类结果表

问题	相似留言数	点赞数	反对数	时间跨度	天数	热度指数
A市房云时代建筑问题	4	56	5	2019/2/25至2019/11/16	264	0.057765
A3区西湖街道茶场村五组拆迁问题	11	0	0	2019/1/6至2019/9/12	249	0
A4区绿地海外滩小区高铁扰民	6	691	0	2019/8/23至2019/9/6	14	8.22619
A6区月亮岛路高压线	4	135	2	2019/3/26至2019/12/19	268	0.127799
A市经济学院强制实习	9	1	13	2017/6/8至2019/11/27	902	0.001725
A市梅溪湖教育问题	6	1765	5	2019/2/1至2019/11/22	294	1.003401
A市三一大道交通问题	9	78	0	2019/2/20至2019/11/15	268	0.032338
A市万家丽南路丽发新城搅拌场扰民	3	2	0	2019/11/13至2019/12/21	38	0.017544
A市五矿万境K9县安全问题	9	2019	0	2019/1/15至2019/11/11	300	0.747778
魅力之城小区扰民	21	18	18	2019/7/21至2019/12/4	136	0.012605
伊景园滨河捆绑车位	40	23	1	2019/7/7至2019/9/1	56	0.010714
西地省A市58车贷	13	2383	0	2019/1/11至2019/7/8	178	1.029818

表5中分别统计出了相似留言数、点赞数、反对数、时间跨度、天数以及最终的热度指数。

生成的“热点问题表.xls”和“热点问题留言明细表.xls”的部分内容如表6和表7所示，全部内容见附件。

表6 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	8.22619	2019/8/23至2019/9/6	A4区绿地海外滩小区	小区外高铁扰民
2	2	1.029818	2019/1/11至2019/7/8	西地省A市58车贷用户	58车贷恶性退出
3	3	1.003401	2019/2/1至2019/11/22	A3区梅溪湖周边学生	梅溪湖学生教育问题
4	4	0.747778	2019/1/15至2019/11/11	A市五矿万境K9县	居民区存在安全问题
5	5	0.127799	2019.3.26至2019.12.19	A6区月亮刀岛路	人口密集区域建设高压线路

表7 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
1	191951	A00041448	A4区绿地海外滩小区距渝	2019/8/23 14:21	尊敬的领导：你好，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，这严重不符合我国于1988年颁布的国家标准GB8702-88《电磁辐射防护规定》。按照设计要求，”铁路两侧30米内严禁新建住宅，学校，医院等噪声敏感建筑物，距铁路中心30米外，200米内的区域不宜临路建学校，医院，敬老院和集中住宅区籍噪声敏感建筑物。现在情况是我们小区已经建完了，业主即将入住，你们要在这小区旁边建高铁铁路是何其扰民，甚至对居民的健康产生影响，这不是轻轨也不是火车，这是高铁，每天以350公里每小时的时速经过家边，再加上绕城高速和另一条火车铁路。我恳求政府三思而定，要不就把整个小区征收了吧，拆迁也比让我们住在这种环境下要好。谢谢领导，敬礼！希望能回复到我上面两个问题：1、关于高铁，如果从这里过，是否考虑将整个海外滩小区二期进行拆迁加咱？因按技术条件此小区和高铁太近，已经不适用于居住，无法消除噪声问题。此小时还未入住，可较好的操作。2.开始买房时开发商也没有提到会有这条高铁的问题，而快交房了出现了这条高铁，这块已经构成与开始商谈条件完全不符合的情况。请求如有高铁过的话，回收此房。	1	0
1	202575	A00092007	咨询A市绿地海外滩二期5	2019/9/4 18:32	我们是A市A4区绿地海外滩二期5栋居民，该小区合同约定2019年9月30日之前交房，目前暂未交房。我在2019年9月4日，同小区业主拉我进群我才知道，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，高铁线路将会紧挨着现有的石长铁路线跨过月亮岛。如此，我们将饱受高铁350公里带来的噪音困扰。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绿地海外滩、恒大御景天下、新城和樾、雅居乐花园、德峰小区、万境潇楚、楚江一号、钱隆世家等，尤其高铁线路紧挨着新建的绿地海外滩二期楼盘和未开建的绿地海外滩三期项目。据测算，高铁规划线路距离绿地海外滩二期最近的位置只有30米不到，距离小区的幼儿园100米不到，这严重不符合我国于1988年颁布的国家标准gb8702-88《电磁辐射防护规定》。按照设计要求，“铁路两侧30米内严禁新建住宅，学校，医院等噪声敏感建筑物，距铁路中心30米外，200米内的区域不宜临路建学校，医院，敬老院和集中住宅区籍噪声敏感建筑物。”如果线路属实，其将对小区的居民生活和小孩的学习成才造成严重的影响和伤害。同时，铁路距离A4区实验小学也不过500米左右。现在情况是小区已经建完了，合同约定9月底交房入住。在城镇化高歌猛进的时代、即便房价高企、但已经结婚生子的我们又不得不买房安家落户，于是大伙掏光了家里的“六个钱包”还抵押了下半辈子从银行贷了款才买了现在的这套房子。在等即将交付新房的时候却先等来了最近处鉴黏房屋不足28米距离，来了一个时速350公里的高铁这样晴天霹雳的消息。虽然这是个信息大爆炸的时代，但是处于信息链最底层的依然是平头百姓。在第一次听到高铁就在楼下的时候，整个小区的业主的心都凉透透了，因为一旦高铁建成我们这个小区就真的废了，这房子以后还怎么住人？我们纷纷在埋怨自己，有想退房的、想拒绝收房的、有想维权的、有想买第二套房，还有恳请地方政府拆迁的。这样的局面、我们也是很无助、很无可奈何的。绿地海外滩二期这个地方本身的噪音已经非常大，您打开百度卫星地图就可以清晰的看到距离小区大约四十几米处有一条石长铁路，再旁边就是一条绕城高速，在楼下不足30米处再加一条高铁对于小区来讲真的是灭顶之灾。声波向高处扩散，现如今再来了一条东西大动脉、时速350公里的高速铁路，实在无法去想象这样的环境，将来我们的孩子要在这样的环境下成长。因为长赣高铁有利于地方经济发展和交通改善，A市人民当然全力支持长赣高铁的规划与建设，不过我们也恳请政府及铁路总公司是否可以考虑A市北三环沿线的已有小区的情况，如果按照现行的规划进行	17	0

通过观察“热点问题表”，可以发现排名第一的是小区外高铁扰民，依次是：58车贷恶性退出、梅溪湖学生教育问题、居民区存在安全问题以及人口密集区域建设高压线路。

4.3 问题3结果分析

问题3所需提交的结果为：从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

量化后的评价部分结果如表8所示，完整结果见附件“问题3评价过程表.xls”

表8 问题3评价过程表

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	相关性	完整性	可解释性
2549	A00045581	蓉华苑物业管理	2019/4/25 9:32:09	司却以交20万保证	金，在业主为	2019/5/10 14:56:53	53.77%	100.00%	51.95%
2554	A00023583	南路洋湖段怎么还	2019/4/24 16:03:4	生意带来很大影响	且换填后还有	2019/5/9 9:49:10	63.98%	80.00%	27.70%
2555	A00031618	高A市民营幼儿园老	2019/4/24 15:40:0	更是加大了教师的	职工要依法	2019/5/9 9:49:14	41.44%	80.00%	48.29%
2557	A000110735	能享受人才新政吗	2019/4/24 15:07:3	市，想买套公寓，(	含），首	2019/5/9 9:49:42	55.16%	60.00%	51.44%
2574	A0009233	公交站点名称变更	2019/4/23 17:03:1	坡岭小学”，原“马	的问题。公交	2019/5/9 9:51:30	26.73%	80.00%	93.93%
2759	A00077538	区含浦镇马路卫生	2019/4/8 8:37	尼巴冲到右边，越是	明卫生较差	2019/5/9 10:02:08	33.56%	80.00%	32.88%
2849	A000100804	市村小区盼望早日	2019/3/29 11:53:2	社区惠民装电梯的	街办公室下发	2019/5/9 10:18:58	55.38%	80.00%	100.00%
3681	UU00812	澜湾社区居民的集	2018/12/31 22:21:5	，天寒地冻的跑好	施设备采购	2019/1/29 10:53:00	41.05%	80.00%	47.75%
3683	UU008792	光住宅楼无故停工	2018/12/31 9:55:0	得到相关准确开工	信查后，西地	2019/1/16 15:29:43	39.72%	80.00%	45.29%

表8在附件4的基础上添加了相关性、完整性以及可解释性三列数据，以百分数的形式展示。

接下来分别绘制相关性以及可解释性相关性分布散点图。

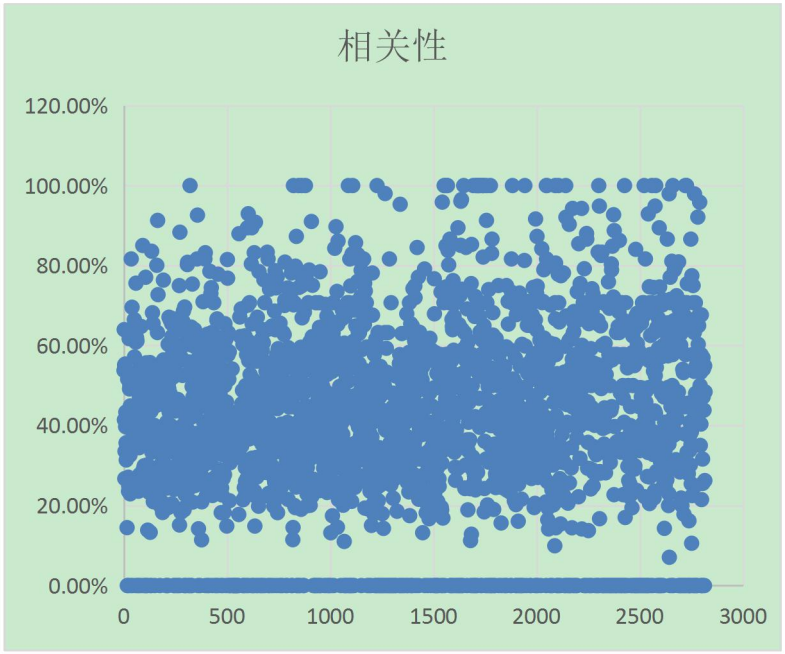

图3 相关性分布散点图

根据结果绘制相关性分布散点图，可以得知，大部分答复意见与留言相关性在20%至60%之间，也有部分意见与留言完全没有关联。

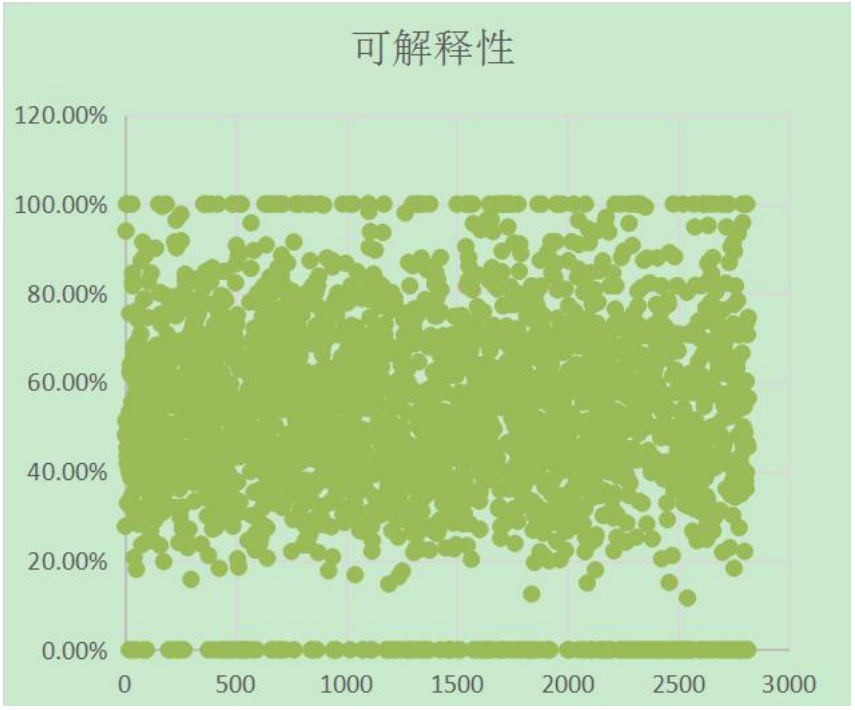

图4 可解释性分布散点图

根据结果绘制可解释性分布散点图，可以得知，大部分答复意见与留言可解释性在30%至80%之间，也有部分意见与留言完全没有关联。

对上述结果进行分析，共计2817条答复意见，答复相关性均值为38.59%，完整性均值为60.84%，可解释性均值为49.74%。现从答复的相关性、完整性、可解释性角度对答复意见给出评价方案，具体如下：

表9 答复意见评价方案

答复意见质量评价	评价指标
意见完整，与留言相关性高，可解释性强	$Pertinence \geq 38.59\%$ $Integrality \geq 60.84\%$ $Interpretability \geq 49.74\%$
意见完整，与留言相关性高，但可解释性不强	$Pertinence \geq 38.59\%$ $Integrality \geq 60.84\%$ $Interpretability \leq 49.74\%$
意见完整，可解释性强，但与留言相关性不高	$Pertinence \geq 38.59\%$ $Integrality \leq 60.84\%$ $Interpretability \geq 49.74\%$
意见完整，但与留言相关性不高，可解释性不强	$Pertinence \geq 38.59\%$ $Integrality \leq 60.84\%$ $Interpretability \leq 49.74\%$
意见不完整，但与留言相关性高，但可解释性强	$Pertinence \leq 38.59\%$ $Integrality \geq 60.84\%$ $Interpretability \geq 49.74\%$
意见不完整，可解释性不强，但与留言相关性高	$Pertinence \leq 38.59\%$ $Integrality \geq 60.84\%$ $Interpretability \leq 49.74\%$
意见不完整，与留言相关性不高，但可解释性强	$Pertinence \leq 38.59\%$ $Integrality \leq 60.84\%$ $Interpretability \geq 49.74\%$
意见不完整，与留言相关性不高，可解释性不强	$Pertinence \leq 38.59\%$ $Integrality \leq 60.84\%$ $Interpretability \leq 49.74\%$

5. 总结与展望

通过对“智慧政务”中的文本挖掘进行分析研究，对群众留言进行分类，进而对热点问题进行挖掘，并对相关部门的答复意见做出评价，这对于广大群众和政府部门来说都具有重大意义，同时也是文本分析的一个课题、 一个难题。

本文首先对附件中的数据进行了分析及预处理操作，然后建立相应的模型来进行训练和测试，进而通过评测指标来观察模型效果，最后对模型结果进行分析。

由分析结果可以看出，对于群众留言分类问题，建立逻辑回归分类模型，采用F-Score评价指标得到的最终结果为0.88。对于热点问题挖掘，综合考虑反对数、点赞数、相关留言数以及问题持续时间，定义热度评价指标，进而得到“热点问题表.xls”和“热点问题留言明细表.xls”。通过观察“热点问题表”，可以发现排名第一的是小区外高铁扰民，依次是：58车贷恶性退出、梅溪湖学生教育问题、居民区存在安全问题以及人口密集区域建设高压线路。对于答复意见的评价，通过计算出相关性、完整性、可解释性的各自的均值，从而作为阈值来对答复意见做出评价。

由于本小组成员能力有限，在这次实验过程中，也有许多不足之处有待改善：

1)对于文本语义带来的词语交叉，做了简化处理；

2)对于答复意见的评价指标没能很好地统一相关性、完整性、可解释性三个方面。

这次的“泰迪杯”数据挖掘挑战赛，我们小组成员分工明确，一人负责建模，一人负责代码，一人负责文档，遇到问题大家会一起讨论，协商解决。这次比赛虽然我们做的不是非常的完美，但是我们尽我们的所能去完成，从中学到了很多知识。困难总是在所难免的，但是只要大家团结一心，坚持不懈，最终总能实现我们的目标。

参考文献

[1] 吴钧鸿.基于智慧政务的社会治理创新方法[J].智库时代,2019(42):294-295.

[2] 龚言浩. 基于文本挖掘的智慧城市建设的热点与城市差异研究[D].南京大学,2018.

[3] 赵银红.智慧政务:大数据时代电子政务发展的新方向[J].办公自动化,2014(22):51-54+14.

[4] 李斌,刘际昕.中国电子政务发展的动力机制分析[J].电子政务,2012(11):8-12.

[5] 张红,甘利人,薛春香.基于标签聚类的电子商务网站分类目录改善研究[J].现代情报,2012,32(01):3-7.

[6] 梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.