“智慧政务”中的文本挖掘应用

摘要

随着网络问政平台的兴起，各种社情民意相关文本的数量不断攀升，此时需要运用人工智能和自然语言处理技术对文本资料进行快速处理。本文采用机器学习的方法，对某问政平台的大量文本资料进行建模，进而解决文本分类，热点挖掘等常见问题。

首先，为了解决文本分类任务，本文运用 python 的 jieba 库对文本进行分词、删除停用词等操作。在文本清洗完成之后，选择词频最高的 4000 词构建特征矩阵，并分别采用“朴素贝叶斯”和“逻辑回归”建立模型。逻辑回归的 F1 均分较高，达到了 0.85，并且不存在过拟合情况，模型有较好的泛化能力。其次，运用 Kmeans 对不同地域的问题留言进行聚类，找出可能成为热点的事件，进一步根据建立的热度指标，挑选出了 5 个热度值最高的事件。最后，本文选用文本长度和文本余弦相似度对答复意见的完整性和相关性进行量化，再根据两个指标数值的分布情况设立合适的判断条件。将判断条件整合，形成一个可以给答复意见质量分为“优”、“中”，“差”的一个评价体系。

关键词： 自然语言处理、朴素贝叶斯、逻辑回归、Kmeans、答复质量评价

目　　录

一、引言

（一）问题背景和重述

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，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。同时，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。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。我们需要利用利用自然语言处理和文本挖掘的方法解决下面的问题：

（1）群众留言分类

针对相关部门人工手动对留言进行分类的工作量大、低效率问题，利用自然语言处理、文本挖掘方法结合机器学习理论和应用，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）热点问题挖掘

及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。根据所给内容，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。

（3）答复意见评价

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，并尝试实现。

（二）研究思路和方法

（1）为解决第一问，首先需要对群众的留言详情进行初步的预处理，处理完成后，再对每一条群众留言进行分词，将群众留言详情这一特征的取值转化为分词后的结果，随后，去除分词结果中没有意义的停词。紧接着，我们再从分词、去停词后的结果中提取出文本特征并生成特征稀疏矩阵，最后根据提取出来的文本特征来建立分类模型并给出分类方法的评价。

（2）对于热点问题的挖掘，首先利用 jieba 库分词和正则表达式将各个地区分类，并且统计留言频数，得到了留言出现频率最高的 10 个地区。其次本文基于 k 均值聚类算法将各个地区的主题分类，更加直观的分辨出各个问题的留言人数和评价人数。最后我们通过定义热度评价指标来筛选热点问题及其留言明细。

（3）针对第三问，为了从不同角度对答复意见进行评价，首先，本文分别选取“文本长度”，和“文本余弦相似度”（留言主题和答复意见）对答复意见的完整性和相关性进行量化。其次，根据两个指标的数值分布情况，选择合适的判断条件作为评价依据。最后将不同的条件判断整合，形成一个完整的评价体系。

二、模型建立

（一）群众留言分类

1. 数据清洗，分词以及去停词

将数据以数据框的形式读入到 python 中，对一级标签进行标签编码，用 0-6 来代表这七个不同的一级标签，具体为：城乡建设：0，环境保护：1，交通运输：2，教育文体：3，劳动和社会保障：4，商贸旅游：5，卫生计生：6。然后去除留言详情中的换行符、制表符、特殊符号和多余空格，为后续的分词工作做准备。

本题使用 jieba 库来进行分词工作，针对每一条留言详情进行分词，生成一个包含分词结果的列表，最后的总体分词结果为一个 list of list 形式，并用这个 list of list 代替留言详情。部分结果如图 2.1 所示：

	theme	label
0	[A3, 区, 大道, 西行, 便, 道, ，, ，未管, 所, 路口, 至, 加油站, 路段...	0
1	[位于, 书院, 路, 主干道, 的, 在水一方, 大厦, 一楼, 至, 四楼, 人为, 拆...	0
2	[尊敬, 的, 领导, ：, ，A1, 区苑, 小区, 位于, A1, 区, 火炬, 路, ，, ...	0
3	[A1, 区, A2, 区华庭, 小区, 高层, 为, 二次, 供水, ，, ，楼顶, 水箱,...	0
4	[A1, 区, A2, 区华庭, 小区, 高层, 为, 二次, 供水, ，, ，楼顶, 水箱,...	0

图 2.1 分词结果

接下来，我们进行去停词工作，将停词表读入到 python 中，删除停词表中包含的分词，值得一提的是，针对此问题，我们将地区编号（A1、B1 等）、宽泛的名词（领导、政府、部门、区、企业等）、礼貌用语（您好、尊敬、希望、请求等）也当作停用词进行处理。去停词后的部分结果如图 2.2 所示：

	theme	label
0	[大道, 西行, 道, 未管, 路口, 加油站, 路段, 人行道, 包括, 路灯, 杆, 圈...	0
1	[位于, 书院, 路, 主干道, 在水一方, 大厦, 一楼, 四楼, 人为, 拆除, 水, ...	0
2	[区苑, 小区, 位于, 火炬, 路, 小区, 物业, 市程明, 物业管理, 有限公司, 未...	0
3	[小区, 高层, 二次, 供水, 楼顶, 水箱, 长年, 不洗, 自来水, 龙头, 水, 霉...	0
4	[小区, 高层, 二次, 供水, 楼顶, 水箱, 长年, 不洗, 自来水, 龙头, 水, 霉...	0

图 2.2 去停词结果

此外，在对文本预处理后，通过词云图可以大致了解每个类别的常用词语，例如图 2.3 给出的是“教育文体”的词云图。

图 2.3 文体教育词云图

2. 文本的特征选择

我们先使用留出法划分训练集和测试集，训练集的大小为原始数据的 80%，测试集为 20%。根据训练集中的分词结果来提取文本特征，测试集中所使用的文本特征与训练集中提取出来的特征一致。在这里我们使用 Python 的 Sklearn 包中 CountVectorizer 函数来实现文本特征的提取。首先统一文本中字母的大小写，然后将特征提取规则设置为对所有关键词的词频进行降序排列，取前 4000 个作为关键词集来生成特征。

通过以上的工作，我们将文本性质的群众留言详情转化为了可以用于建模的各个文本特征，这些文本特征均为出现在各类群众留言中的高频词（关键词），如果某一条群众留言中包含 n 个关键词 a，那么他在关键词 a 这一特征上的取值就为 n。

3. 模型的构建与验证调优

（1）朴素贝叶斯建模

这里我们尝试建立两种分类模型，分别为朴素贝叶斯建模和 Logistic Regression 模型。

首先运用朴素贝叶斯建模。贝叶斯分类器是基于贝叶斯决策论来进行决策的一种分类模型，贝叶斯决策论考虑如何基于这些概率和误判损失来选择最优的类别标记。首先，我们定义在样本 x 上的条件风险为：

$$R(c_i|x) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|x) \qquad \text{式（1）}$$

其中 c 为类别标记，我们的任务是寻找一个判定准则 h 以最小化总体风险：

$$R(h) = E_x[R(h(x)|x)] \quad \text{式（2）}$$

对每个样本 x，若 h 能最小化条件风险，则总体风险也将被最小化，这就产生了贝叶斯判定准则：为最小化总体风险，只需要在每个样本上选择那个能使风险$R(c|x)$最小的类别标记。即：

$$h^*(x) = argminR(c|x) \quad \text{式（3）}$$

此时，h^*称为贝叶斯最优分类器。若目标是最小化分类错误率则贝叶斯最优分类器为：

$$h^*(x) = argminP(c|x) \quad \text{式（4）}$$

即对于每个样本 x，选择能使后验概率 $P(c|x)$最大的类别标记。

不难发现，基于贝叶斯公式来估计后验概率的困难在于类条件概率 $P(c|x)$是所有属性上的联合概率，难以从有限的训练样本中直接估计而得。为了避开这个障碍朴素贝叶斯分类器采用了属性条件独立性假设，对已知的类别假设所有属性相互独立。每个属性独立的对分类结果发生影响。

$$P(c|x) = \frac{P(c)}{P(x)}\prod_{i=1}^{d} P(x_i|c) \quad \text{式（5）}$$

我们使用 Sklearn 包中的 MultinomialNB 函数来建立朴素贝叶斯模型，根据训练集上混淆矩阵得出的分类结果报告如表 2.1 所示：

表 2.1 朴素贝叶斯模型训练集上的分类结果报告

类别	精确度	召回率	F1-score	真实样本数
0	0.92	0.80	0.85	1793
1	0.84	0.99	0.91	848
2	0.71	0.90	0.79	545
3	0.91	0.93	0.92	1437
4	0.93	0.91	0.92	1779
5	0.89	0.83	0.86	1096
6	0.88	0.91	0.89	791
准确率			0.88	8289

从报告结果中我们可以看出，各类别的F_1值都非常的理想，模型在训练集上的表现十分不错。但想要拥有理想的泛化性能，我们更需要观察模型在测试集上的表现。接下来，我们得出测试集上的混淆矩阵分类结果报告，并画出学习曲线，结果如表 2.2 和图 2.4 所示：

表 2.2 朴素贝叶斯模型测试集上的分类结果报告

类别	精确度	召回率	F1-score	真实样本数
0	0.90	0.78	0.84	216
1	0.77	0.98	0.86	90
2	0.63	0.78	0.70	68
3	0.88	0.94	0.91	152
4	0.91	0.87	0.89	190
5	0.83	0.70	0.76	119
6	0.82	0.86	0.84	86
准确率			0.84	921

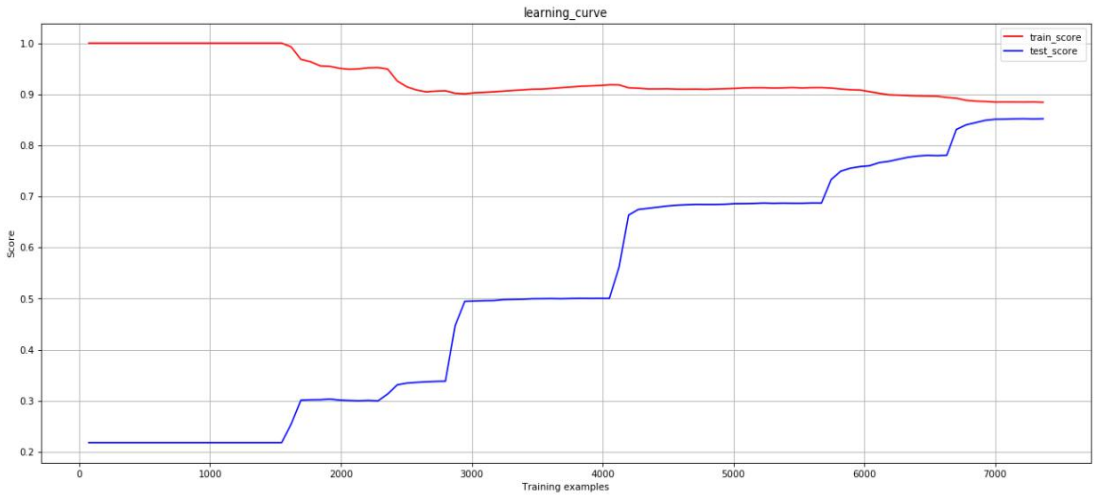

图 2.4 朴素贝叶斯分类模型的学习曲线

结果显示，测试集上分类报告的各类别F_1得分也都十分理想，从学习曲线中我们也可以看出，模型既不存在高偏差也不存在高方差，有较好的泛化性能。

（2）Logistic Regression 建模

逻辑回归模型（Logistic Regression，LR）是由线性回归演化而来，属于广义线性回归，是一种概率模型。逻辑回归是典型的二分类算法，也可以用作多分类。

逻辑回归算法在解决分类问题时流程如下：

①构造假设函数（预测函数）

预测函数是我们输入自变量来预测因变量结果的函数，逻辑回归是将 Sigmoid 函数和线性回归模型结合，来给出因变量发生的概率值，其线性边界函数如公式（6）所示：

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i = \theta^T x \qquad \text{式（6）}$$

根据 sigmoid 函数进一步构造假设函数，具体表示形式如公式（7）所示：

$$h_\theta(\mathrm{x}) = \mathrm{g}(\theta^T x) = \frac{1}{1+\exp(-\theta^T x)} \quad \text{式（7）}$$

根据θ值带入某样本的特征到假设函数进行计算，若输出结果大于等于 0.5，则判别其为类别 1，若小于 0.5 则将其归为类别 0。

②构造损失函数 J(θ)

根据公式（7）可知对于输出结果分类得到类别 1 和 0 的概率分别如公式：

$$\mathrm{p}(\mathrm{y}=1|\mathrm{x};\theta) = h_\theta(\mathrm{x}) \quad \text{式（8）}$$

$$\mathrm{p}(\mathrm{y}=0|\mathrm{x};\theta) = 1 - h_\theta(\mathrm{x}) \quad \text{式（9）}$$

进一步通过最大似然算法和对数变换可得

$$\mathrm{J}(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{n} y_i log h_\theta(x_i) + (1-y_i)\log(1 - log h_\theta(x_i))\right] \quad \text{式（10）}$$

③梯度下降来获得回归参数θ

想要得到最小化的 J(θ)，通过梯度下降法，可以得到θ的更新优化公式（11）：

$$\theta_j{:}\theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)x_i^j \quad \text{式（11）}$$

我们使用 Sklearn 包中的 LogisticRegression 函数来建立逻辑回归模型，同样的先使用留出法划分训练集和测试集，训练集的大小为原始数据的 20%。在数据清理的过程中，我们发现数据存在样本类别不平衡的问题，特别是第六类：卫生计生和第二类：交通运输的样本明显的少于城乡建设和环境保护等样本量很多的类别，类别的不平衡会对学习过程造成一定的困扰，因此需要在模型中设置类别权重参数，这里，我们将 class_weight 参数设为“balanced”，这样类库会根据训练样本量来计算权重。某种类型样本量越多，则权重越低，样本量越少，则权重越高。此外，我们对参数 C（正则化系数λ的倒数）也进行了选择，我们在训练集上使用五折交叉验证法，设置 C 从 0.1 到 1.1（增加值为 0.1）的 10 个值，通过图形来从这十个参数 C 不同的模型中选择预测效果最好的模型。如图 2.5 所示：

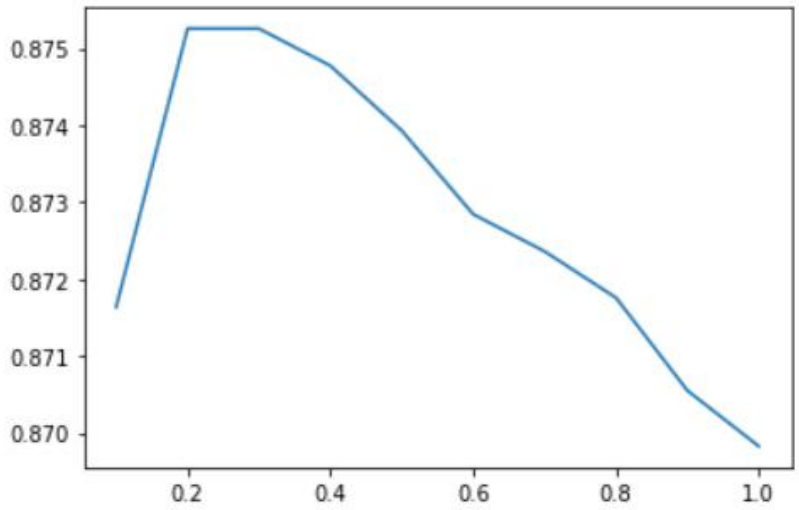

图 2.5 各正则化参数的预测效果图

从图中我们可以看出，参数 C 取 0.3 时，预测效果最好，因此我们设置参数 C 为 0.3。完成参数的选择后，我们开始在训练集上训练模型，根据训练好的模型来对测试集的数据进行分类预测，分类结果报告和学习曲线如表 2.3 和图 2.6 所示：

表 2.3 逻辑回归模型测试集上的分类结果报告

类别	精确度	召回率	F1-score	真实样本数
0	0.81	0.85	0.83	219
1	0.91	0.87	0.89	78
2	0.79	0.79	0.79	73
3	0.95	0.91	0.93	140
4	0.93	0.90	0.91	209
5	0.73	0.74	0.73	115
6	0.86	0.89	0.87	87
准确率			0.86	921

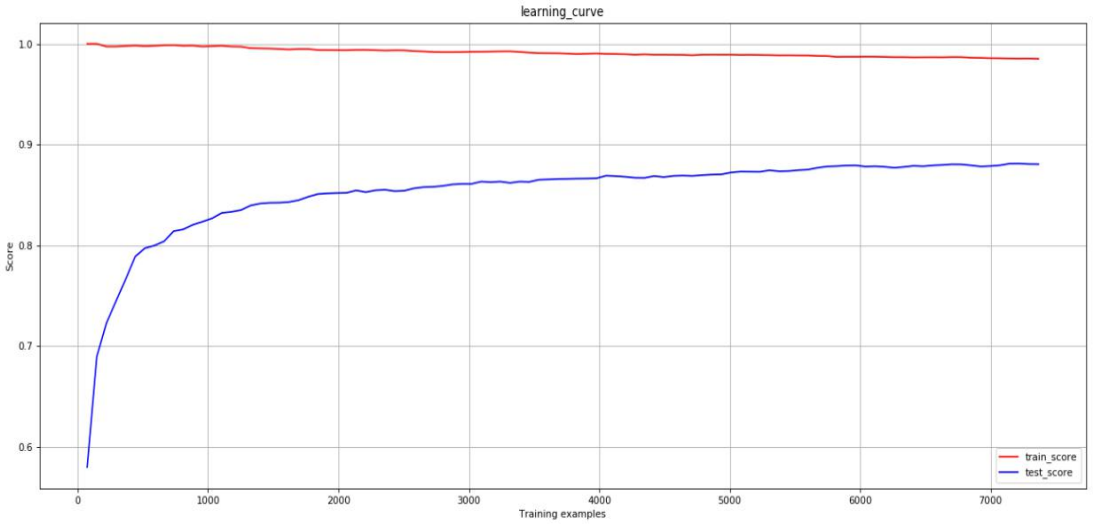

图 2.6 逻辑回归模型的学习曲线

从测试集上的分类结果报告中我们可以看出，用逻辑回归模型来进行分类,效果也很理想，各类别的F_1得分普遍很高，平均达到 0.85，从学习曲线中我们也可以看出，分类结果没有很大的偏差以及过拟合现象。

（二）热点问题挖掘

1．模型介绍

k 均值聚类算法（k-means clustering algorithm）是一种无监督学习算法。在给定样本集 D $=\{x_1,x_2,...,x_m\}$，“k 均值”算法针对聚类所得簇划分 C $=\{C_1,C_2,...,C_k\}$ 最小化平方误差

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}||x-\mu_i||_2^2 \qquad \text{式（12）}$$

其中$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$是簇$C_i$的均值向量。直观来看，式（12）在一定程度上刻画了簇内样本围绕簇均值向量的紧密程度，E 值越小则簇内样本的相似度越高。

K 均值聚类算法流程如下：

（1）从 N 个样本数据中选取 K 个对象作为初始的聚类中心；

（2）分别计算每个样本到各个聚类中心的距离，并将对象分配到最近的聚类中；

（3）所有对象完成分配后，重新计算 K 个聚类中心；

（4）与前一次的聚类中心比较，如果聚类中心发生变化，转 2），否则转 5）；

（5）当质心不再发生变化时停止并输出聚类结果。

度量样本之间的相似性最常用的是欧几里得距离、曼哈顿距离和闵可夫斯基距离；样本与簇之间的距离可以用样本到簇中心的距离 $\mathrm{d}(C_i,\mathrm{x})$；簇与簇之间的距离可以用簇中心的距离 $\mathrm{d}(C_i,C_j)$。

用 p 个属性来表示 n 个样本的数据矩阵如下：

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

欧几里得距离：

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+...+(x_{ip}-x_{jp})^2} \quad \text{式（13）}$$

曼哈顿距离：

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}| \quad \text{式（14）}$$

闵可夫斯基距离：

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|)^q+(|x_{i2}-x_{j2}|)^q+...+(|x_{ip}-x_{jp}|)^q} \quad \text{式（15）}$$

2．建立模型

本节要解决的问题是：将附件 3 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。最终给出排名前五的热点问题以及相应热点问题对应的留言信息。

（1）文本预处理

读取附件 3 的全部数据内容。首先利用正则表达式的方法编写函数，获取各留言主题的所在特定地区，当然也存在部分留言主题内未说明地点，我们将其记作“None_area”。

此外，虽然 A 市可以分成各个区和县，但是部分问题只说明了发生地点位于 A 市，未指明具体的区或县，也没有与指明区县的留言主题存在重复关系。所以我们将这些留言单独归为一类，记作“A 市”。

我们提取留言出现频率最高的 10 个地区，留言频数统计如表所示

表 2.4 留言频数统计

地区	频数
A 市	1714
A7 县	678

A3 区	432
None_area	318
A2 区	259
A4 区	234
A1 区	184
A5 区	175
A6 区	143
A8 县	83
A9 市	44

其次，我们需要对留言主题进行中文分词，在分词之前，需要采用正则表达式的方式删除文本中空格、回车以及特殊符号。我们采用 Python 中的 jieba 库对留言主题进行中文分词。

除此之外，我们希望在数据庞大的语料库中删除那些对分析无用的词语，在中文的语句中，这些词往往也不存在特殊的意义，只是用来连接句子，使句子逻辑更加通顺，但这些词语又是非常常见的，例如“一个”、“因为”、“所以”等词语。由于 Python 中并没有自带的中文停用词语料库，所以本文引用人大经济论坛提供的中文停用词表。同时，考虑到留言主题中存在的一些无用词，如“政府”，“企业”，“管理”等，我们也将其补充到停用词表中。利用 Python 将其读取，从而构成停用词列表，最后构造函数将语料库中的停用词删除。现在查看前四个留言主题的分词结果：

表 2.5 分词结果

分词前	分词后
'A3 区一米阳光婚纱艺术摄影是否合法纳税了？'	'一米阳光', '婚纱', '艺术摄影', '合法', '纳税'
'咨询 A6 区道路命名规划初步成果公示和城乡门牌问题'	'咨询', '道路', '命名', '规划', '初步', '成果', '公示', '城乡', '门牌'
'反映 A7 县春华镇金鼎村水泥路、自来水到户的问题'	'春华', '镇金鼎村', '水泥路', '自来水', '到户'
'A2 区黄兴路步行街大古道巷住户卫生间粪便	'黄兴路', '步行街', '古道', '巷', '住户', '卫生间', '

外排'	粪便', '外排'

最后，我们需要提取各个地区数据，生成计数向量，将用到 sklearn 中的 CountVectorizer 创建词袋数据结构，从而把文本中的词语转换为词频矩阵。我们对所有关键词的 term frequency 进行降序排序，只取前 4000 个作为关键词集。

（2）k 均值聚类

对于分类簇数的确定，我们需要引入轮廓系数的概念，它是聚类效果好坏的一种评价方式。结合内聚度和分离度两种因素。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。

假设我们已经通过一定算法，将待分类数据进行了聚类。常用的比如使用 K-means ，将待分类数据分为了 k 个簇 。对于簇中的每个向量。分别计算它们的轮廓系数。

对于其中的一个点 i 来说：

计算 a(i) = average(i 向量到所有它属于的簇中其它点的距离)

计算 b(i) = min (i 向量到与它相邻最近的一簇内的所有点的平均距离)

那么 i 向量轮廓系数就为：

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad \text{式（16）}$$

可见轮廓系数的值是介于 [-1,1] ，越趋近于 1 代表内聚度和分离度都相对较优。

接下来我们以 A 市聚类为例，查看聚类结果，其余地区的操作步骤和算法都类似。通过 A 市的特征矩阵，进行 k 均值聚类，其中 k 值由轮廓系数图来确定。A 市的轮廓系数图如下：

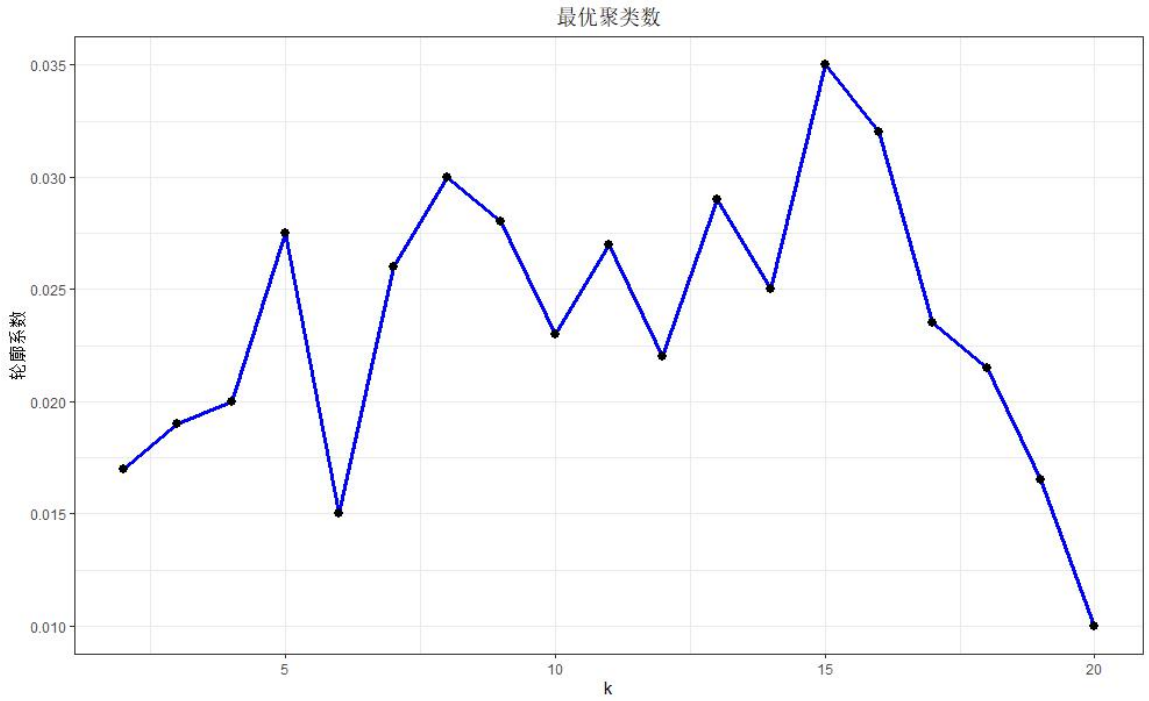

图 2.7 A 市轮廓系数

从图中可以看出，当 k=15 时，轮廓系数达到最大值，所以我们设簇的个数为 15。

此外，还存在一些属于 None_area 的留言，我们也同样对 None_area 的特征矩阵进行聚类，并且分出 None_area 的热点事件，将其与应该归属的地区合并。例如通过之前的聚类可以发现“魅力之城”属于 A5 区，但是某些留言中存在“魅力之城”的关键词，但是并未提及 A5 区，甚至未提到 A 市，我们就可以将 None_area 中有关“魅力之城”的留言与 A5 区的留言合并成一类。

3．热点问题评价结果

我们将 10 个地区的事件汇总，可以得出 7 个热点问题，如下表所示：

2.6 热点问题初次筛选

问题 ID	时间范围	地点/人群	问题描述
1	2018/11/15-2019/9/27	A 市/人才或工作人群	A 市长时间有人询问关于 A 市人才补贴问题
2	2019/7/18-2019/8/31	A 市/伊景园滨河苑居民	A 市伊景园滨河苑协商要求购房同时必须

			购买车位
3	2019/2/14-2019/12/28	A7 县/星沙旧城居民	A7 县星沙旧城改造项目问题
4	2019/11/2-2020/1/25	A2 区/丽发新城居民	A2 区丽发新城附近建搅拌站噪音扰民
5	2019/3/26-2019/4/9	A6 区/群众	A6 区月亮岛路沿线架设 110kv 高压电线杆存在问题
6	2019/7/21-2019/09/25	A5 区/魅力之城居民	A5 区劳动东路魅力之城小区附近存在扰民现象
7	2019/1/14-2019/3/1	A4 区/58 车贷诈骗案受害者	A4 区 58 车贷案

根据题目要求，我们需要从 7 个问题中选出排名前 5 的热点问题，这里就需要用到热度评价指标。

（1）热度评价指标定义

热度即人们对事件的关注程度，在这里我们需要从两种人的角度去考虑：一是点赞和反对的人数，不管是支持还是反对，评价人数越多，可以说明该事件的热度越高。但是，不同地区的具体人数也会不同，所以我们可以用结构相对指标来定义评价的热度；二是留言人数，即使评价的热度不高，我们也不能忽略问题发现者的反映情况，这也是定义问题热度的重要指标。所以对于热度评价指标，我们将采取如下的方式来定义：假设 Q 事件发生在 A 市，事件支持数和反对数的总和为事件的评价人数，则

$$\text{Q 事件热度} = \frac{Q\ 事件评价人数}{A\ 市所有事件评价人数} * Q\ 事件留言人数 \qquad 式（17）$$

我们提取出 7 个热点问题的所在地区的数据，并从中再度筛选出热点问题的反对数、支持数以及留言人数。经计算，7 个问题的热度排名如下

表 2.7 热点问题排名

问题 ID	地点/人群	问题描述	问题评价人数	所有事件评价人数	问题留言人数	热度	排名
1	A 市/人才或工作人群	人才补贴问题	29	5981	18	0.087276375	7
2	A 市/伊景园滨河苑居民	伊景园滨河苑协商要求购房同时必须购买车位	26	5981	52	0.226049156	5
3	A7 县/星沙旧城居民	星沙旧城改造项目问题	32	1882	12	0.204038257	6
4	A2 区/丽发新城居民	丽发新城附近建搅拌站噪音扰民	48	157	44	13.4522293	1
5	A6 区/群众	月亮岛路沿线架设 110kv 高压电线杆存在问题	175	449	7	2.728285078	3
6	A5 区/魅力之城居民	劳动东路魅力之城小区附近存在扰民现象	36	2290	18	0.282969432	4

7	A4 区/58 车贷诈骗案受害者	58 车贷案	3928	5623	9	6.28703539	2

从上表的排名可以看出，排名前五的热点问题分别是“A2 区丽发新城附近建搅拌站噪音扰民”、“A4 区 58 车贷案”、“A6 区月亮岛路沿线架设 110kv 高压电线杆存在问题”、“A5 区劳动东路魅力之城小区附近存在扰民现象”和“A 市伊景园滨河苑协商要求购房同时必须购买车位”

（三）答复意见的评价

1．建模思路

根据第三问的要求，需要从答复的相关性，完整性，可解释性等角度对答复的质量给出一套完整的评价方案。

本文首先挑选了 2 个指标来分别度量答复的相关性和完整性。这两个指标分别为：“文本的长度”，“本文的余弦相似度（留言主题和答复详情）”。其次了解各指标的数值分布情况，选择合适的判断条件，然后进行评价。最终评价模型给出的是三个类别：1 表示答复意见为优，2 表示答复意见为中，3 表示答复意见为差。

2．建立模型

（1）答复完整性指标的建立

本文运用文本长度去量化文本完整性。图 2.8 给出的是文本长度的密度曲线，可以大致了解文本长度的分布情况。从图 2.8 可以看出本文长度大部分集中在 200 左右，此外还有一些异常数据，文本长度为 5 以内。

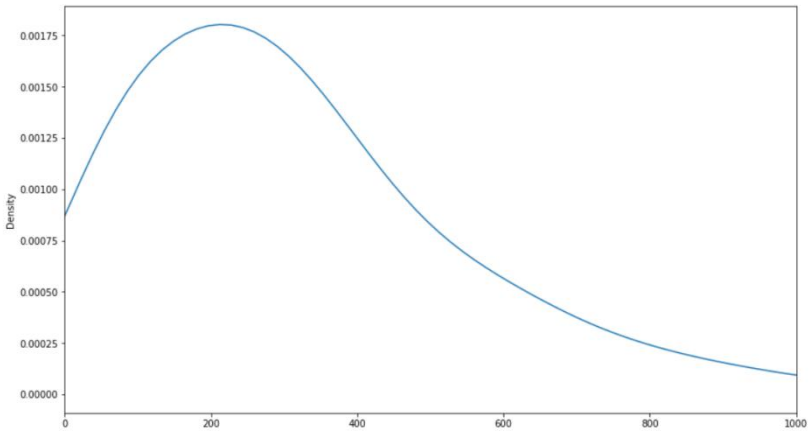

图 2.8 文本长度的密度曲线

为了确定具体的判断条件，表 2.8 给出了文本长度的各个分位数，其中最小值为 3，即那些异常的样本。同样的，存在一些留言答复字数上千。

表 2.8 文本长度分位数

count	2816
mean	360.6
std	429.9
min	3
25%	145
50%	276
75%	444
max	7883

结合以上考虑，本文暂且定文本长度小于其下四分位数的答复质量为“差”，如果文本长度大于上四分位数的，认为其答复质量为“优”。

（2）答复相关性指标的建立

由于上文提出的完整性评价指标仅仅对部分样本的质量进行评价的，还有一部分的样本没有给出评价，所以本文运用“留言主题”和“答复详情”的词计数向量计算余弦相似度，进而对答复的相关性进行量化。

①余弦相似度介绍

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0 度角的余弦值是 1，而其他任何角度的余弦值都不大于 1；并且其最小值是-1。从而两

个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90°时，余弦相似度的值为 0。

这对任何维度的向量空间中都适用，而且余弦相似性最常用于高维正空间。例如在信息检索中，每个词项被赋予不同的维度，而一个维度由一个向量表示，其各个维度上的值对应于该词项在文档中出现的频率。余弦相似度因此可以给出两篇文档的相似度。

②余弦相似度计算步骤

a. 分词

句子 A：我/喜欢/看/电视，不/喜欢/看/电影。

句子 B：我/不/喜欢/看/电视，也/不/喜欢/看/电影。

b. 计算词频

句子 A：我 1，喜欢 2，看 2，电视 1，电影 1，不 1，也 0。

句子 B：我 1，喜欢 2，看 2，电视 1，电影 1，不 2，也 1。

c. 转化为词计数向量

句子 A：[1, 2, 2, 1, 1, 1, 0]

句子 B：[1, 2, 2, 1, 1, 2, 1]

d. 计算余弦相似度：

$$\cos\theta = \frac{1\times1+2\times2+2\times2+1\times1+1\times1+1\times2+0\times1}{\sqrt{1^2+2^2+2^2+1^2+1^2+1^2+0^2}\times\sqrt{1^2+2^2+2^2+1^2+1^2+2^2+1^2}} = 0.938 \qquad \text{式（18）}$$

所以这两个文本的余弦相似度为 0.938

余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫"余弦相似性"。所以，上面的句子 A 和句子 B 是很相似的，事实上它们的夹角大约为 20.3 度。

③计算结果

下面给出了两个样本的例子 ：

表 2.9 留言余弦相似度

留言主题	答复意见	余弦相似度
建议 G 市地方乡镇政府能够倡导民众建立建设绿色垃圾回收通道	网友：您好，留言已收悉，已转交相关单位办理。 2019 年 2 月 1 日	0

希望政府能加大对公租房的管理和建设投入	我市的公租房建设力度正在逐年加大，到 2013 年底城区约有 2000 套公租房投入使用，可以有效地减轻公租房供需矛盾，急需公租房的住户有望在今年年底得到解决。 J11 市房产管理局	0.55

表 2.10 给出的是留言主题和答复详情的相似度各个分位数，从表 2 可大致了解到相似度的分布情况，基于此进一步选择出合适的决策条件。

表 2.10 留言主题和答复详情相似度分位数

count	2789
mean	0.201879
std	0.166308
min	0
25%	0.063888
50%	0.178174
75%	0.310045
max	0.878315

结合以上考虑，本文暂且定相似度小于其下四分位数的答复质量为“差”，如果相似度大于上四分位数的，认为其答复质量为“优”，其余定义为“中”。

3．构建评价模型

总结上文，将两个角度的判断结合，形成了一个对于“答复意见”的评价模型。该评价模型的判断步骤是：1.判断文本长度是否很少，或者文本长度是否很长；2.当文本长度适中的时候，判断“答复意见”和“留言主题”的余弦相似度。具体流程见图 2.9。

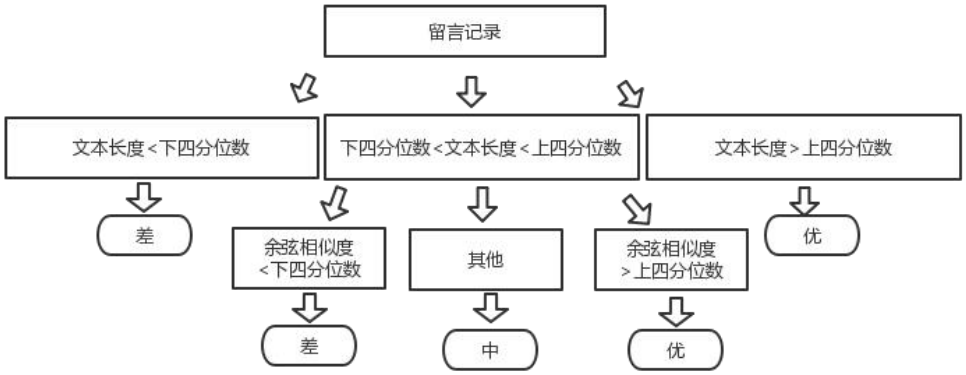

图 2.9 评级模型判断步骤

其中的判断条件，上四分位数和下四分位数自然可以修改，调整成一个合适的判断范围。

表 2.11 给出的是评价结果，每一个类别的样本数量，具体结果见附件。从中可以看出样本数据中很多的答复意见相对还是比较好的，大约有 1/5 的答复意见相对较差。

表 2.11 评价结果

优	2097
中	209
差	510

部分评价为差的结果如下：

表 2.12 评价结果为差

留言编号	留言用户	留言主题	答复意见	评价
6556	UU0081320	咨询打狂犬役苗报销比例是多少	已收悉	差
10924	UU0082420	投诉石长铁路火车夜晚鸣笛扰民	网友：您好！留言已收悉	差
37482	A00062705	建议 B9 市规划一个校车接送计划	2018 年 12 月 12 日	差

4．模型局限性

由于本文建立的评价体系比较简单，可能无法考虑到是否有答复人员有意将文字写的过长，从而躲避系统检测的行为。此外，对“答复意见”的可解释性没有建立评价指标，这一点可以根据“惠民平台”的具体要求来设定，本文在此就没有继续赘述。

三、结论与不足

（一）结论

互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，本文利用文本挖掘技术对“智慧政务”中的群众留言进行分类，从中解析热点问题，最后对于问题的答复意见按自己定义的指标进行评价。最终得出了如下结论：

1. 针对第一题，我们通过对数据的清洗、分词、去停词、提取文本特征，提取出了用于建模的文本特征，随后构建了两种不同的分类模型来对留言详情进行了分类预测，均取得了理想的效果，朴素贝叶斯分类器中各类别的F_1得分平均值为 0.83，逻辑回归模型中各类别的F_1得分平均值为 0.85。

2. 本文第二步使用了 jieba 库分词以及正则表达式得到了各个地区的留言的文本特征矩阵，随后通过 k 均值聚类算法将各个地区的留言主题进行分类。根据分类结果，我们从挑选了 7 个最热的留言主题，最后根据定义的热度评价指标筛选出了排名前五的热点问题，即“A2 区丽发新城附近建搅拌站噪音扰民”、“A4 区 58 车贷案”、“A6 区月亮岛路沿线架设 110kv 高压电线杆存在问题”、“A5 区劳动东路魅力之城小区附近存在扰民现象”和“A 市伊景园滨河苑协商要求购房同时必须购买车位”。最后，我们也从各个地区的分类中获取了它们的留言明细。

3. 第三问本文首先选取了本文长度来衡量答复意见的完整性，选择文本余弦相似度来衡量答复意见的相关性。其次根据两个指标数值的分布情况设立合适的判断条件。最后将判断条件整合，形成一个完整的评价体系。该评价体系会根据答复质量将答复意见分为三类：“优”、“中”、“差”。对于样本数据，该评价体系判断大多数的答复意见是“优”，还有大约 20%的答复质量为“差”。

（二）不足

针对第三问的评价体系还有可以改进的方面，对“答复意见”的可解释性本文没有建立评价指标，这一点可以根据“惠民平台”的具体要求来设定。

参考文献

[1]梁学芳. 基于逻辑回归模型的汽车评论挖掘研究[D].天津商业大学,2019.

[2]何伟. 基于朴素贝叶斯的文本分类算法研究[D].南京邮电大学,2018.

[3]邱智学. 面向 B2C 电商平台的短文本挖掘研究[D].浙江工商大学,2020.

[4]刘家岐.利用 Python 对自然语言进行简单处理[J].现代商贸工业,2019,40(07):159-160.

[5]赵谦益.k-means 算法中文文献聚类的 Python 实现[J].软件,2019,40(08):89-94.

[6]吴军.数学之美.第 2 版[M].人民邮电出版社,2014.

[7]刘艺彬.基于分词频的特征选择算法在文本分类中的研究[D].西安理工大学,2018.

[8]李春林,冯志骥.基于文本挖掘的新能源汽车用户评论研究[J].特区经济,2020(04):148-151.

[9]艾楚涵,姜迪,吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术,2020,44(04):65-70.

[9]秦俊忠. 热点事件挖掘系统的设计与实现[D].北京邮电大学,2016.

[10]罗引. 互联网舆情发现与观点挖掘技术研究[D].电子科技大学,2010.

[11]Vallejo-Medina Pablo,Correa Juan C,Gómez-Lugo Mayra,Saavedra-Roa Diego Alejandro,García-Montaño Eileen,Pérez-Pedraza Diana,Niebles-Charris Janivys,García-Roncallo Paola,Abello-Luque Daniella,Espada José Pedro,Morales Alexandra. A text mining approach for adapting a school-based sexual health promotion program in Colombia.[J]. Preventive medicine reports,2020,18.

[12]Syaamantak Das,Shyamal Kumar Das Mandal,Anupam Basu. Mining multiple informational text structure from text data[J]. Procedia Computer Science,2020,167.