摘要

近些年，随着网络问政平台逐渐完善，平台已成为政府了解民意、汇聚民智、凝聚民气的重要渠道。本文利用文本挖掘和舆情热点技术，对群众留言分类、热点问题挖掘、答复意见评价做出了分析。

针对问题一，对留言内容进行了去重、去空、数据平衡、jieba 分词、TF-IDF 权重赋值方面的预处理，并绘制了词云图。对于群众留言分类，进行了模型的比选，涉及到的分类模型有**高斯朴素贝叶斯**、**多项式朴素贝叶斯**、**逻辑回归**、**线性支持向量机**、**随机森林**五种模型。基于上述 5 个模型，采用了机器学习工具包 Scikit-learn 封装算法分类器进行分类器的训练，结合 F-Score 分析，得到线性支持向量机（Linear SVC）效果最佳，其 F-Score 值为 0.8890。故在本题分类模型中，选用基于 TF-IDF 算法的**线性向量机**，差错率最低。

针对问题二，建立了**双阈值 Single-Pass 模型**，实现了热度问题的文本聚类，并且定义了与关注度、时间有关的热度指数，完成了热度指数排序。本题采用双阈值 Single-Pass 模型，降低了数据顺序对模型的影响。通过计算**余弦距离相似度**，选定与话题簇之间的关系，并将最大文本相似度与两阈值进行比较，判断放入话题簇或是簇缓冲区，更新话题向量，簇缓冲区数据重新计算分类，直至所有数据遍历完成并且簇缓冲区数据为 0 为止。

针对问题三，根据答复的相关性、完整性、可解释性、时效性角度对答复内容进行分析，做出不合格回复、合格回复、优质回复的评价。对于相似度，在留言内容和回复内容之间计算**余弦相似度**，若相似度等于 0，即可说明为不合格回复，不等于 0，为合格回复；对于完整性和可解释性，建立两个**词典**，同时满足词典要求的评论，即为优质回复；对于时效性，是在优质回复中完成的，答复时间越快，及时解决问题的可能性越高，该回复越好。

关键词：jieba 分词 TF-IDF 算法 线性向量机 双阈值 Single-Pass 余弦相似度

Abstract

In recent years, with the gradual improvement of online political platform, the platform has become an important channel for the government to understand public opinion, pool people's wisdom and pool people's spirit. This paper uses text mining and public opinion hot spot technology to analyze the classification of public comments, hot spot issues mining, reply comments evaluation.

According to question one, the message content was preprocessed in terms of get rid of duplicate data, delete blank data, data balance, jieba participle and TF-IDF weight assignment, and the word cloud maps were drawn. For the classification of crowd comments, the comparison and selection of models are carried out. The classification models involved include **Gaussian NB**, **Multinomial NB**, **Logistic Regression**, **Linear SVC**, and **Random Forest Classifier**. Based on the above five models, the machine learning kit, Scikit-learn package algorithm classifier, was used to train the classifier, combined with F-Score analysis, **Linear SVC** has the best effect, and its F-Score is 0.8890. Therefore, **Linear SVC** based on TD-IDF algorithm has the lowest error rate in the classification model.

For the second problem, a **double threshold Single-pass model** was established to realize the text clustering of the heat problem, and the heat index related to the attention time was defined to complete the heat index ranking. Subject adopts double threshold model of Single - Pass, reduces the impact of data order on the model. By calculating the **cosine distance similarity**, the relationship between the message and the topic cluster was selected, and the maximum text similarity was compared with the two thresholds to determine whether the date was put into the topic cluster or the cluster buffer. And then the topic vector was updated, and the cluster buffer data vector was recalculated and classified until all the data traversed was completed and the cluster buffer data number was 0.

In view of question three, the content of the reply was analyzed from the perspective of relevance completeness, interpretability and timeliness of the reply, and the evaluation of **unqualified reply**, **qualified reply** and **good reply** was made. For similarity, calculate cosine similarity between the content of the message and the content of the reply. If the similarity is equal to 0, it can be interpreted as **unqualified reply**. If it is not equal to 0, it is **qualified reply**. For completeness and interpretability, the establishment of two dictionaries, while meeting the dictionary requirements of the comments, that is **good reply**. For timeliness, it is completed in good reply. The faster the reply time is, the higher the possibility of timely solving the problem is, and the better the reply is.

Key word: jieba participle; TD-IDF algorithm; Linear SVC; double threshold single-pass; cosine distance similarity

目录

1、挖掘目标

本次建模目标是利用网络问政平台上的群众问政留言记录，对留言内容进行去重，去空处理，利用 jieba 中文分词工具对群众的留言内容进行分词，利用 TF-IDF 算法对文本数据向量化，Single-pass Clustering 单遍聚类算法，余弦相似度检验等，达到以下三个目标：

1）数据处理后，利用 TF-IDF 算法对文本数据向量化，然后运用多个模型对群众留言记录进行一级标签分类，找出最好的分类模型结果以便交由相关部门处理。

2)根据所给的群众留言数据集，定义合理的热度评价指标，找出某时段群众集中反映的热点问题排序。

3)根据相关部门所给的留言答复，针对其相关性，完整性，可解释性，及时性等方面，进行检测，对答复内容的质量作出评价。

2、分析方法与过程

总体流程图：

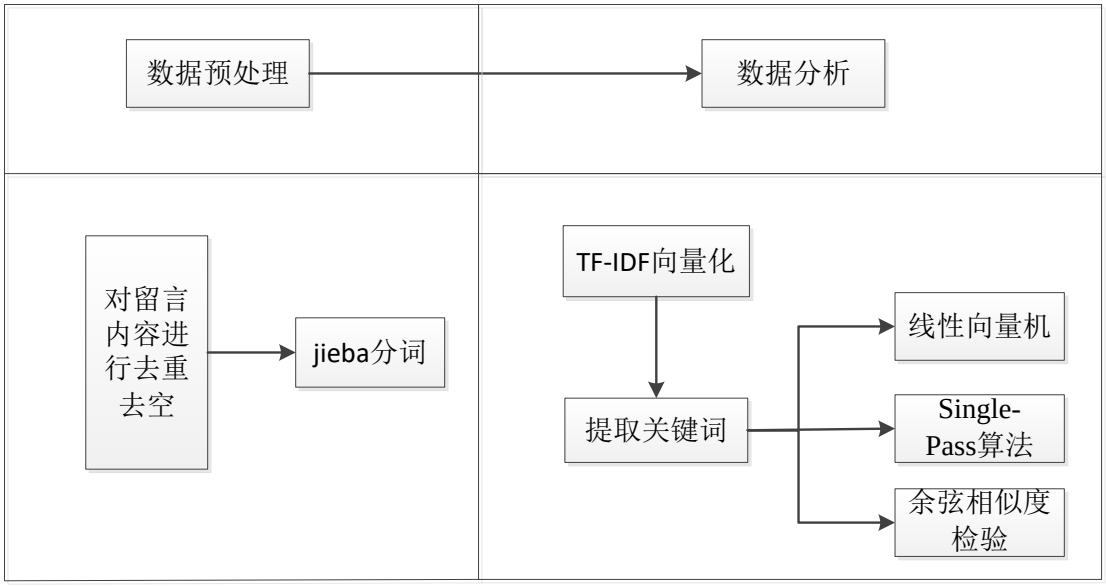


图 1：总体流程图

本用例主要包括如下步骤：

步骤一：数据预处理，在所给留言数据中，对留言详情中的缺失值和重复值进行去重，去空处理，各类别的数据平衡后，进行中文分词，同时去停用词，去特殊符号。

步骤二：数据分析，采用 TF-IDF 算法，找出每条留言详情的关键词，把留言详情中的文本信息转换为权重向量，运用线性向量机进行留言分类；运用 Single-Pass 算法进行聚类分析，挖掘热点问题；运用余弦相似度检验留言回复的质量。

2.1 问题 1 分析方法与过程

2.1.1 问题 1 分析

问题 1 主要解决网络问政平台的群众留言分类问题，我们的目标是建立合适的模型对群众留言进行分类，即如果来了一条新的群众留言，我们希望将其分配到附件 1 所提供的的一级分类标签中的一个，且假设每条新留言都分配给一个且仅一个类别。这是文本多分类问题。为了获得更加准确的分类结果，我们尝试了逻辑回归、（多项式）朴素贝叶斯、高斯朴素贝叶斯、线性支持向量机、随机森林五种模型。

2.1.2 流程图

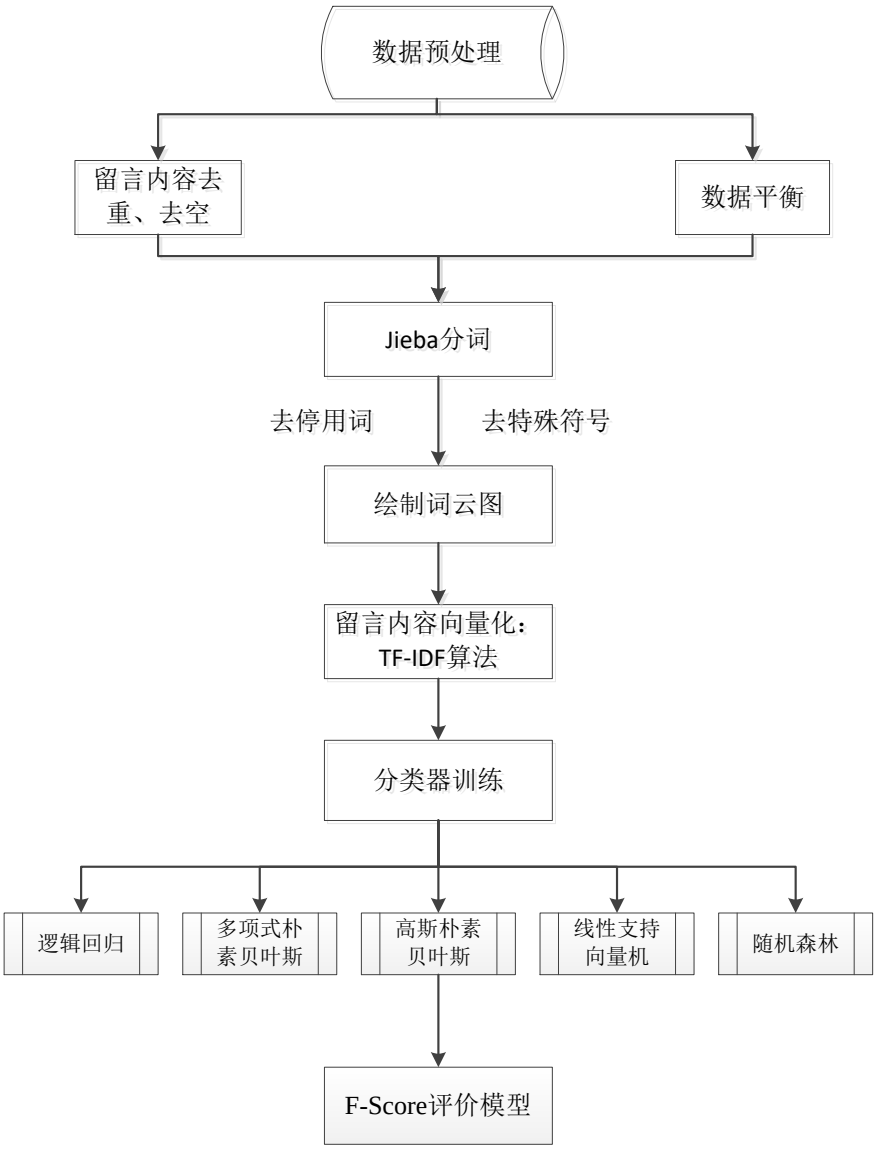


图 2：问题一流程图

2.1.3 数据预处理

2.1.3.1 留言内容去重、去空

对于问题 1，我们提取了附件 2 中的“留言详情”和“一级标签”两列数据共 9210 条留言进行建模。并删除“留言详情”这列中含缺失值和重复值的记录，修改列名称为“category”和“content”。为了方便建模，我们将“一级标签”进行了量化，添加一列将“category”编码为整数的列，列名称为“label”并创建了“category_to_id”、“id_to_category”两个字典对象保存类标签“label”和“category”的映射关系，供将来使用。最终得到了列名称为“category”、“content”、“label”共 9210 条数据，将其命名为 data。

category	content	label
城乡建设	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…	0
城乡建设	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…	0
城乡建设	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…	0
城乡建设	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…	0
城乡建设	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…	0

图 3：data 数据集

category_to_id ={'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}

id_to_category ={0: '城乡建设', 1: '环境保护', 2: '交通运输', 3: '教育文体', 4: '劳动和社会保障', 5: '商贸旅游', 6: '卫生计生'}

2.1.3.2 数据平衡

通过统计各类别留言的条数，如图 2 所示，我们发现样本数据不平衡，为了解决此问题我们采用了欠抽样法在各类别数据中随机抽取了 600 条数据，得到了新的数据集 data_new。

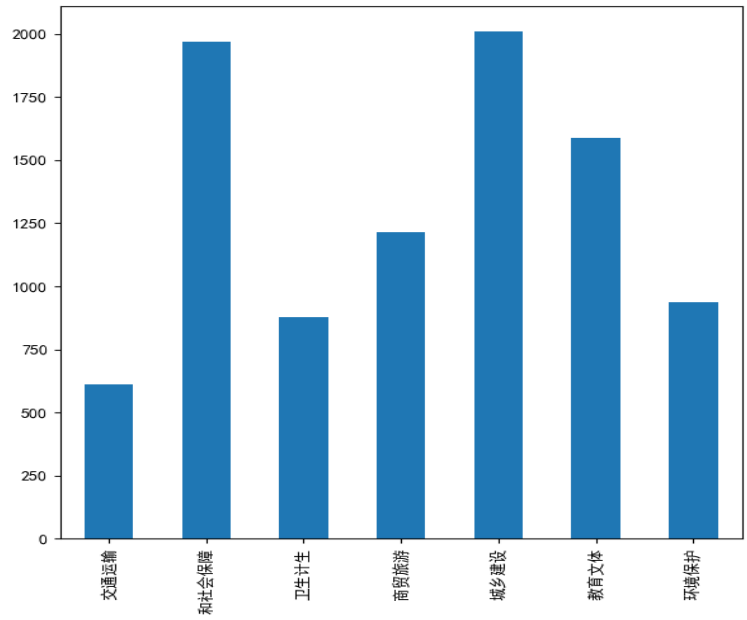


图 4：各类别留言条数统计图

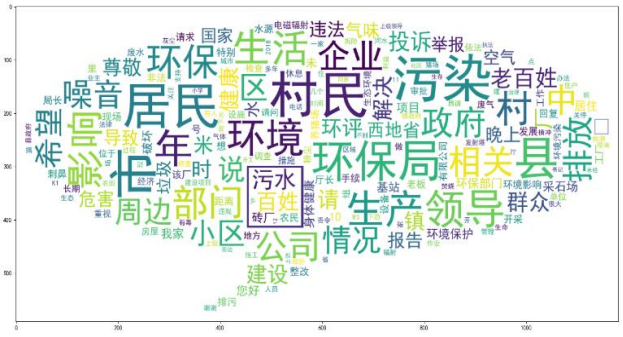
2.1.3.3 对留言内容进行中文分词

分类器和学习算法没办法对文本的原始形式做直接处理，因为它们期望的输入是长度固定且为数值型的特征向量，而不是具有可变长度的原始文本。因此，在预处理阶段，文本需要被转换为更易于操作的表示形式。为此，我们首先采用了 Python 自带的 jieba 分词对留言内容进行分词，jieba 分词中，首先通过对照词典生成句子的有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取，对于未登陆词（不在词典中的词）使用 HMM 进行新词发现，能达到很好的中文分词效果。接着我们对分词后的数据去除了特殊字符和停用词，这里采用网上下载的停用词表，得到了较为干净的数据 adata。

2.1.4 绘制词云图

词云图，也叫文字云，能够对文本中出现频率较高的“关键词”予以视觉化的展现，词云图过滤掉大量的低频低质的文本信息，使得浏览者只要一眼扫过文本就可领略文本的主旨。对于本题来说，通过分词后的数据集可以画出不同分类的留言的词云图，如图 3 所示：

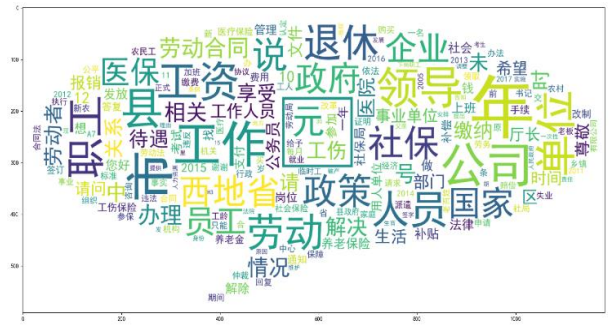
城乡建



环境保

交通运

教育文



劳动和社会保

商贸旅游

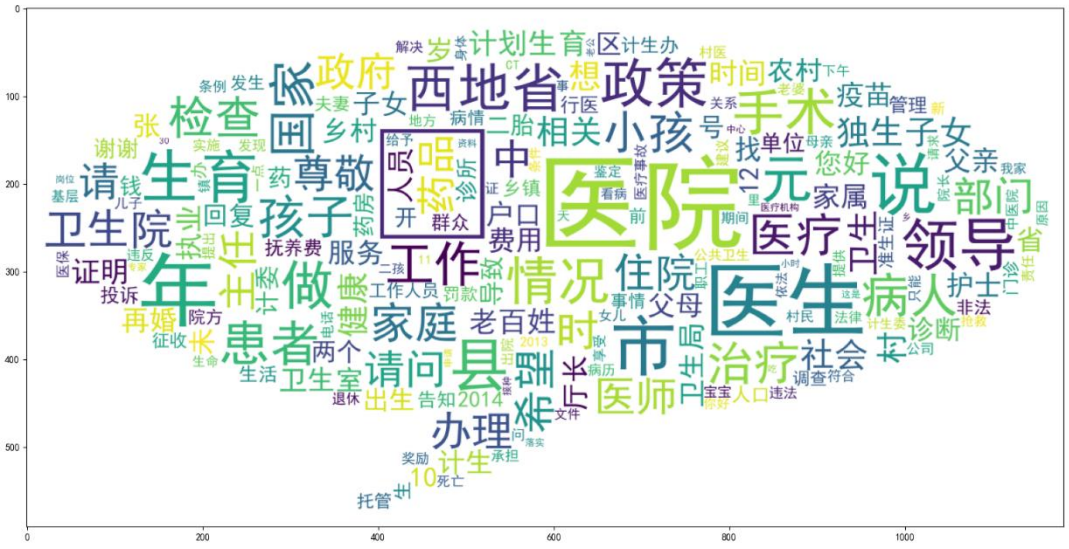


图 5：词云图

通过词云图可以明显看出，在城乡建设类的留言中，“小区”、“业主”等词出现的频率最高；

2.1.5 TF-IDF 算法

对留言内容进行分词等操作后，需要将其进行向量化，以便计算机能方便处理。这里我们采用了 TF-IDF 算法。具体来说，对于我们数据集中的每个单词，我们将计算它的词频和逆文档频率，简称 tf-idf。我们将使用 sklearn.feature_extraction.text.TfidfVectorizer 来计算每个群众留言的向量的 tf-idf 向量：

TF-IDF 算法的思想：特征单词在某特定文本中出现的频数越大，其对于该文本的分类作用越大，特征单词在大多数文档中出现的频数越大，对于文本的分类作用越小，其结合了词频与反文档频率两者的优点。采用 TF-IFD 方法，将词频和反文档频率结合作为特征的权重，分别表示为：

$$TF(t_k)=\frac{n(t_k,m_i)}{\sum_{k=1}^{K} n(t_k,m_i)} \tag{1}$$

式（1）中，$n(t_k,m_i)$ 为特征 t_k 在文本 m_i 中出现的次数，$\sum_{k=1}^{K} n(t_k,m_i)$ 为文本 m_i 出现特征 t_k 的总个数。

$$IDF(t_k)=\log\frac{M}{n(t_k)} \tag{2}$$

式（2）中 M 表示训练集中文本的总数，$n(t_k)$ 表示出现特征 t_k 的文本数。

2.1.6 分类模型

在获得留言内容的 tf-idf 值及分类所对应的标签后便可以开始模型的构建。

为了获得效果较好的分类器，我们尝试了高斯朴素贝叶斯、多项式朴素贝叶斯、逻辑回归、线性支持向量机、随机森林五种模型。

2.1.6.1 多项式朴素贝叶斯

多项式朴素贝叶斯模型的权重为：

$$W_k = \frac{TF(t_k) \times IDF(t_k)}{\sqrt{\sum_{i=1}^{M} (TF(t_k) \times IDF(t_k))^2}} \tag{3}$$

式（3）中，表示为文本 $m_i \quad (i = 1, 2, \cdots, M)$ 。

多项式朴素贝叶斯模型建立过程如下：

在计算条件概率时，该模型需要统计单词出现的频率，引入 TF-IDF 权重，设文本类别为 $C = \{C_1, C_2, \cdots, C_j\}$，$j = 1, 2, 3, \cdots, 15$，设 D_i 为任意一篇文档，其包含的 k 个特征词为 $D_i = \{t_1, t_2, \cdots, t_k\}$，其对应的最大的后验概率的类别即为文档 D_i 的所属的类别，后验概率公式可表示为：

$$P(C_j \mid D_i) = \frac{P(D_i \mid C_j) P(C_j)}{P(D_i)} \tag{4}$$

式（4）中，$P(C_j)$ 表示类别 C_j 出现的概率，$P(D_i \mid C_j)$ 表示文本 D_i 属于类别 C_j 的条件概率，$P(D_i) = P(t_1 t_2 \cdots t_k)$ 表示所有特征的联合概率。

贝叶斯的过程就是求解 $P(C_j \mid D_i)$ 最大值的过程，得到文本 D_i 属于的类别，显然对于给定的训练文本 $P(D_i)$ 是常数。化简后表达式为：

$$C_{map} = \max P(C_j \mid D_i) = \max_{C_j \in C} P(C_j) P(D_i \mid C_j) \tag{5}$$

式（5）中，C_{map} 表示 D_i 属于类别 C 的分类结果。

引入 TF-IDF 权重，最终的判别方式为下式：

$$C_{map} = \max_{C_j \in C} P(C_j \mid D_i) = \max_{C_j \in C} [\ln P(C_j) + \sum_{k=1}^{K} \ln P(t_k \mid C_j) \times W_k \times TF(t_k)] \tag{6}$$

式（6）中，K 表示文本 D_i 中特征单词数，$t_k (k = 1, 2, \cdots, K)$ 表示文本 D_i 中的第 k 个特征词，先验概率 $P(C_j)$ 和条件概率 $P(t_k \mid C_j)$ 的计算公式如下：

$$P(C_j) = \frac{\sum_{i=1}^{M} \delta(C_i, C_j) + 1}{M + 15} \tag{7}$$

$$P(t_k \mid C_j) = \frac{\sum_{i=1}^{M} TF(t_k) \delta(C_i, C_j) + 1}{\sum_{k=1}^{K} \sum_{i=1}^{M} TF(t_k) \delta(C_i, C_j) + K} \tag{8}$$

式（7）（8）中，M 表示总的训练文档数，C_i 表示第 i 个训练文本的类别，$TF(t_k)$ 表示特征值 t_k 在文本 D_i 中出现频数，$\delta(\bullet)$ 表示二值函数。

通过上述模型可以确定在 15 个一级标签中文本 D_i 的类别。

2.1.6.2 高斯朴素贝叶斯

高斯朴素贝叶斯模型是假设条件概率 $P\left(X=\mathrm{x}\middle|Y=c_k\right)$ 是多元高斯分布，另一方面，由特征的条件独立性假设，我们就可以通过对每个特征的条件概率建模，每个特征的条件概率 $N\left(\mu_t,\sigma_t^2\right)$ 也服从高斯分布.

在 c 类下第 i 个词对应的高斯分布为：

$$g(x_i;\mu_{i,c},\sigma_{i,c})=\frac{1}{\sigma_{i,c}\sqrt{2\pi}}\exp\{-\frac{(x_i-\mu_{i,c})^2}{2\sigma_{i,c}{}^2}\}$$

式中，$\mu_{i,c}$，$\sigma_{i,c}$ 表示 c 类下第 i 个特征的均值和方差。

根据特征之间的独立性假设，我们可以得到条件概率：

$$P(X=x\,|\,Y=c)=\prod_{i=1}^{d}g(x_i;\mu_{i,c},\sigma_{i,c})$$

式中，d 表示特征的个数。

高斯朴素贝叶斯化简为：

$$P(Y=c_k\mid X=x)=\frac{P(Y=c_k)P(X=x\,|\,Y=c_k)}{\sum_k P(X=x\,|\,Y=c_k)P(Y=c_k)}\propto P(Y=c_k)P(X=x\,|\,Y=c_k)$$

$$y=\arg\max_{ck}P(Y=c_k)P(X=x\,|\,Y=c_k)$$

2.1.6.3 逻辑回归

逻辑回归（Logistic Regression）是一种用于解决二分类或多分类问题的机器学习方法。logisitc 回归的结果并非数学定义中的概率值，不可以直接当做概率值来用。该结果往往用于和其他特征值加权求和，而非直接相乘。Logistic 回归虽然名字里带“回归”，但是它实际上是一种分类方法，其利用了 Logistic 函数（或称为 Sigmoid 函数），函数形式为：

$$\mathrm{g}(z)=\frac{1}{1+e^{-z}}$$

2.1.6.4 线性支持向量机

支持向量机(SVM)是在解决小样本、非线性、高维的分类和回归问题时具有特有优势的机器学习方法，在SVM基础上发展的线性支持向量机（Linear SVW）已成为处理文本分类等海量高维稀疏数据的一种有效机器学习方法。[1]

给定一组训练样本 $\{(x_i,x_i)\}_{i=1}^{l}$，$x_i\in R^n$，$y_i\in\{1,-1\}$，*LinearSVM* 可以表示为求解下式的无约束优化问题：

$$\min_w f(w)=\frac{1}{2}w^Tw+C\sum_{i=1}^{l}\xi(w,x_i,y_i)$$

式中，$\xi(w,x_i,y_i)$ 是损失函数，$C>0$ 是惩罚因子，适用的损失函数有

$$\xi(w,x_i,y_i)=\begin{cases}\max(0,1-y_iw^Tx_i) & L_1-SVM\\ \max(0,1-y_iw^Tx_i)^2 & L_2-SVM\\ \log(1+\exp(-y_iw^Tx_i)) & LR\end{cases}$$

2.1.6.5 随机森林

随机森林属于模式识别中有监督的分类中的一种方法。它的原理是以决策树为基本分类器的一个集成学习模型，它包含多个由 Bagging 集成学习技术训练得到的决策树，当输入待分类的样本时，最终的分类结果由决策树的输出结果的众数决定。其基本原理如下：

1.从原始数据m*n维数据中有放回的抽取样本容量与原数据相同的数据样本m*n，样本数量为ntree；

2.对每一个数据样本应用决策树的计算方法（但并不全部应用），即从数据的n维特征中无放回的随机抽取mtry维特征。以随机色林的分类功能为例，通过mtry维特征中，通过计算信息增益的方式找到分类效果最好的一维特征k，及其阀值th，小于th的样本划分到左节点，其余的划分到右节点，继续训练其他节点。

3.重复训练所有的数据样本，得到ntree个决策树

4.随机森林便是ntree个决策树集合起来的森林，当预测结果时，所有的决策树对预测集——进行分类，得到其分类结果，统计票数得到结果。

2.1.7 训练分类器

基于上述五个模型，采用 python 的开源机器学习工具包 Scikit-learn 封装的 GaussianNB 、 MultinomialNB 、 LinearSVC 、 RandomForestClassifier 、 LogisticRegression 算法进行分类器的训练。并 使用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

训练结果如下：

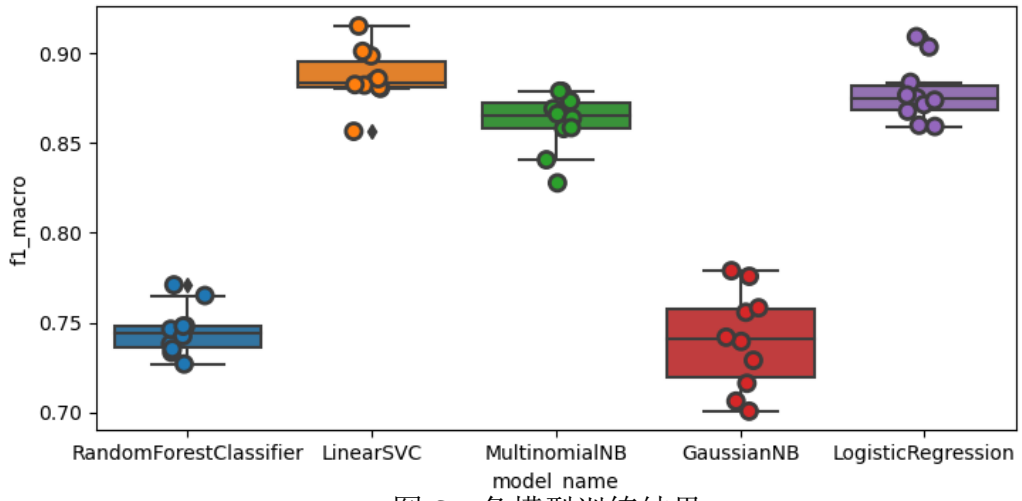


图 6：各模型训练结果

利用 F-Score 评价模型算出的结果见下表。

表 1：各模型训练结果 F-Score 表

Model	MultinomialNB	GaussianNB	LinearSVC	RandomForestClassifier	LogisticRegression
F-Score	0.8604	0.7371	0.8890	0.7544	0.8794

由表 1 可知，上述五个模型中线性支持向量机（LinearSVC）效果最佳。其 F-Score 为 0.8890。

2.2 问题 2 分析方法与过程

2.2.1 问题 2 分析

第二问是文本聚类问题，需要挖掘出热点问题，常用的文本聚类方法有划分法中的 K-means 算法，但是关于评论问题的类别个数是不确定的，对于这种情况，可以选用 Hierarchical Clustering 层次聚类算法、Mean shift 算法和 Single-pass Clustering 单遍聚类算法，而对于文本挖掘来说，最适用的是 Single-pass Clustering 单遍聚类算法。

2.2.2 问题 2 流程图

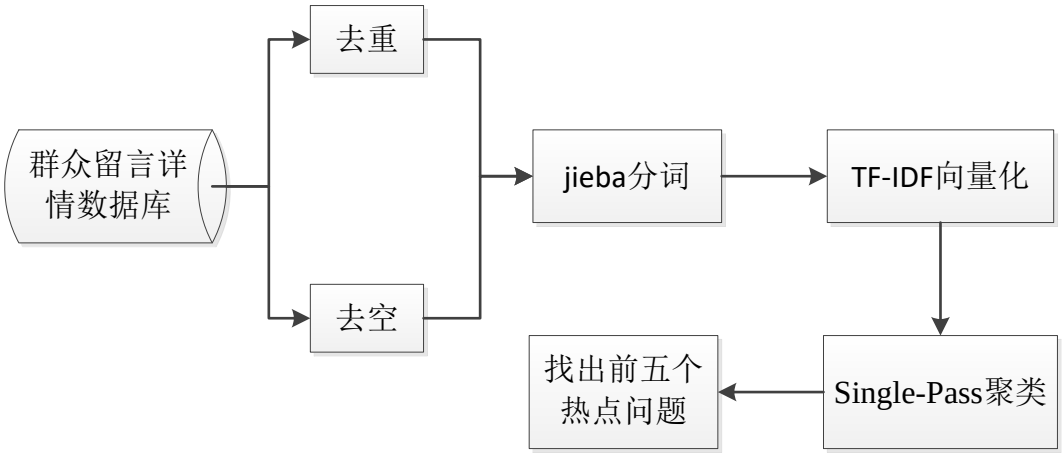


图 7：问题 2 流程图

2.2.3 数据预处理

2.2.3.1 留言内容去重、去空

对于问题 2，我们提取了附件 3 中的“留言详情”这一列数据。并删除“留言详情”这列中含缺失值和重复值的记录。在留言详情中，可能出现具体描述为空的情况，干扰了问题的分析，采取直接滤过的方法，从文本中删除，最后保留去重，去空后的文本数据。

2.2.3.2 对留言内容进行中文分词

分类器和学习算法没办法对文本的原始形式做直接处理，因为它们期望的输入是长度固定且为数值型的特征向量，而不是具有可变长度的原始文本。因此，

在预处理阶段，文本需要被转换为更易于操作的表示形式。为此，我们首先采用了 Python 自带的 jieba 分词对留言内容进行分词，jieba 分词中，首先通过对照词典生成句子的有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取，对于未登陆词（不在词典中的词）使用 HMM 进行新词发现，能达到很好的中文分词效果。接着我们对分词后的数据去除了特殊字符和停用词，这里采用网上下载的停用词表，得到了较为干净的数据 adata。

2.2.3.3 TF-IDF 算法

在对群众留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。采用 TF-IDF 算法，把留言详情信息转换为权重向量。

TF-IDF 算法的思想：特征单词在某特定文本中出现的频数越大，其对于该文本的分类作用越大，特征单词在大多数文档中出现的频数越大，对于文本的分类作用越小，其结合了词频与反文档频率两者的优点。采用 TF-IFD 方法，将词频和反文档频率结合作为特征的权重，分别表示为：

$$TF(t_k)=\frac{n(t_k,m_i)}{\sum_{k=1}^{K} n(t_k,m_i)} \tag{1}$$

式（1）中，$n(t_k,m_i)$ 为特征 t_k 在文本 m_i 中出现的次数，$\sum_{k=1}^{K} n(t_k,m_i)$ 为文本 m_i 出现特征 t_k 的总个数。

$$IDF(t_k)=\log\frac{M}{n(t_k)} \tag{2}$$

式（2）中 M 表示训练集中文本的总数，$n(t_k)$ 表示出现特征 t_k 的文本数。

2.2.4 基于双阈值的 Single-Pass 聚类

Single-Pass 聚类是一种增量式聚类，比较适合处理增量数据，但 Single-Pass 算法对于相同数据集，输入数据的顺序对其分类结果会造成很大的影响。Single-Pass 算法中，每次处理一个新数据都需要将其归属于某个话题或者变成一个新话题，所以话题簇集的中心向量变动较为频繁，相同一个数据在不同的时间段被处理也许归属的话题会有所不同。因此对数据顺序敏感严重影响了 Single-Pass 算法的聚类效果。在这里引用双阈值 Single-Pass 算法，模型建立过程具体如下：

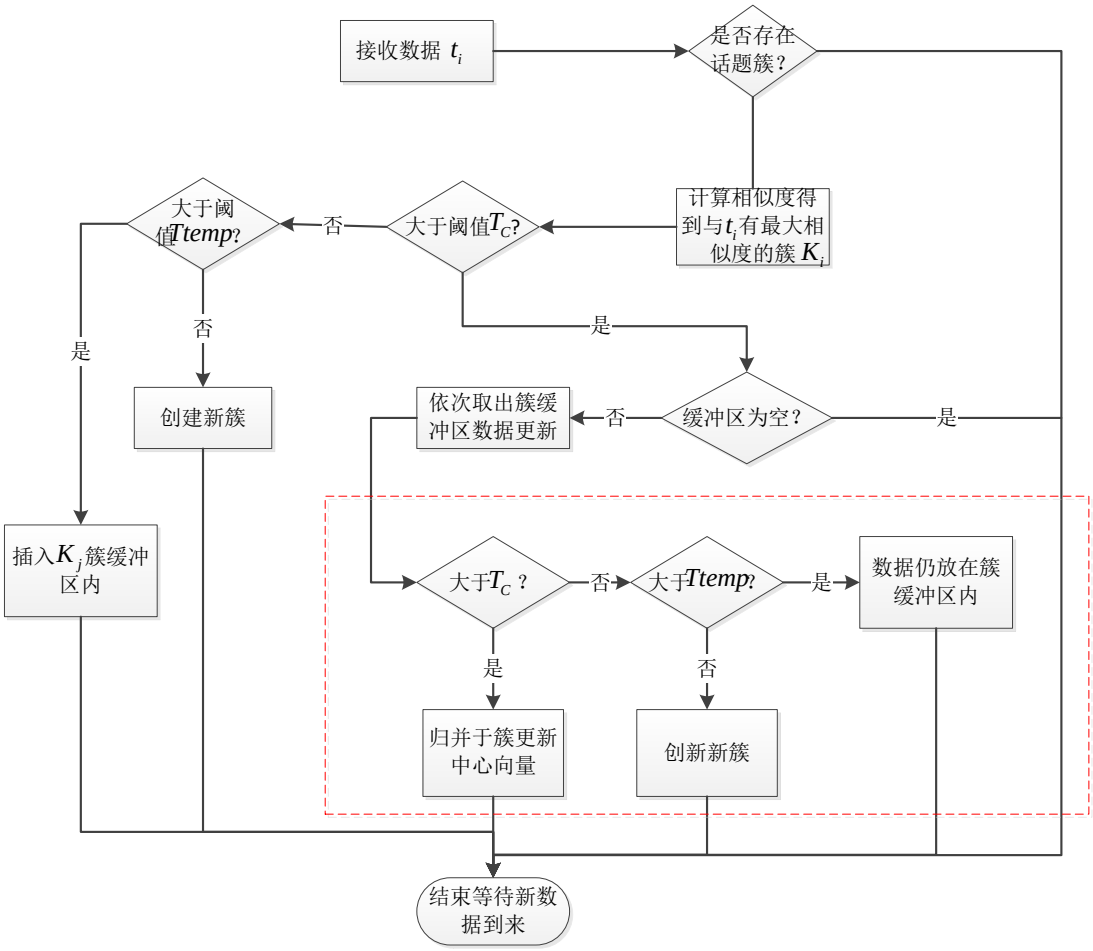


图 8：基于双阈值的 Single-Pass 聚类算法流程图

步骤一：根据问题一的模型，将第二问的留言内容进行一级标签的分类，可视为相同的留言事件在同一一级标签下，在每一一级标签下进行 Single-Pass 模型计算，分析归纳热点问题。

步骤二：通过数据分析将文本内容特征值进行了空间内容向量化，利用 TF-IDF 算法，对特征值进行赋权。

步骤三：通常在每一个一级标签的第一个数据输入时，自成一话题簇。第二个输入的数据与其他所有已知簇的中心文本向量分别进行相似度计算，得到最大相似度T,利用余弦距离公式计算如下：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

式中，n表示数据空间向量的维度。

步骤四：确定两个阈值：T_c和T_{mid}（其中$T_c>T_{mid}>0$）。若$T>T_c$，将输入数据带入到对应的话题簇中去，同时更新话题簇向量（认为是在一个话题簇内的每一个文本空间向量每一维度平均得到）；若$T_{mid}<T<T_c$，将输入数据带入到对应的话题簇的缓冲堆中，等待下一个数据输入，给出定义“簇缓冲区”为每个话题簇都有一个对应的容器，可用来存放数据，该容器被称为该簇的缓冲区；若

$T < T_{mid}$，将该输入数据放入新建的话题簇中。

步骤五：处理话题簇的缓冲堆的数据向量。在下一个数据输入后，并成功更新了话题向量，将缓冲堆里的数据向量按照步骤三的方法重新比对，有放入话题簇、放入话题簇缓冲堆、新建话题簇三种情况。直至所有数据全部被输入，分类完成即结束算法。

当所有数据分类完成后，进行话题热度指标定义计算。已知话题的热度与话题参与讨论人数和持续的时间有关，将同一话题的留言整理出来，设话题持续的时间为 t 天，话题参与讨论人数为 m，包括每一个话题留言的点赞或反对人数，热度指数 $Hear$ 定义公式如下：

$$Hear = \frac{m}{t}$$

当 $Hear$ 越高时，则能说明在短时间内有很多人都反应了这个问题，该问题是急需要解决的，表面该话题热度高；反之，能说明该话题热度低。

2.3 问题 3 分析方法与过程

2.3.1 问题 3 分析

针对第三问，需要对相关部门对留言的答复意见进行评价，从答复的相关性、完整性、可解释性、时效性等角度对答复内容进行检验，看答复内容是否合格。针对答复内容的相关性，主要是对群众的留言内容和平台的答复内容做出相关性检验；针对答复的完整性和可解释性，可以先构建一个合格词典，若出现该词典中的词语，则为优质回复。即可构建一个完整的针对答复内容的质量评价体系。

2.3.2 问题 3 流程图

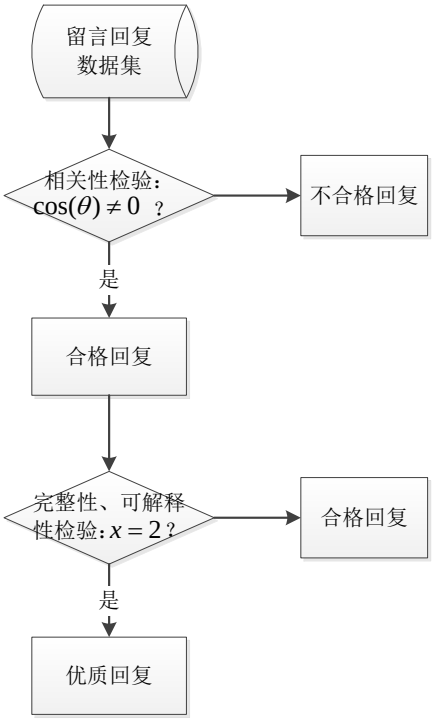


图 9：问题 3 流程图

2.3.3 答复质量评价模型

- 相关性：需要考察答复意见内容是否与问题相关，排除答非所问的情况。

判别方法为根据留言内容提取出的特征向量，与答复内容提取出的特征向量进行相似度计算，利用余弦计算公式计算：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

若 $\cos(\theta)=0$，即可说明两者毫不相关，该答复内容是不合格回复；若 $\cos(\theta)\neq 0$，可视为答复意见是相关的，该答复内容是合格回复。

- 完整性和可解释性：须在合格回复的基础上，进行计算。

完整性是指答复内容中有没有用政策法规解释，以规范为依据；可解释性强调的是因果关系，判别方法是答复内容中有无包含因果关系的副词。据此需要构造完整性词典和可解释性词典，具体说明如下，设初始 $X=0$。

完整性词典：由文件、通知、规范、调令、协议、合同、规定、政策、工作方案、条例、细则、法规、实施办法构成词典。答复内容中包含词典中的任意一词，对于该回复 $X=X+1$。

可解释性词典：由因为、因此、因、为（了）、故、所以、下一步、解释、最终构成词典。答复内容中包含词典中的任意一词，对于该回复 $X=X+1$。

若最终 $X=2$ 时，可视为该答复为优质回复；若 $X\neq 2$ 时，该答复仍为合格回复。

- 时效性：须在优质回复的基础上，进一步对答复进行排序评价。判别方法为留言评价时间与答复时间的时间段的长短。具体表示方法如下：

$$T=T_r-T_q\quad（天）$$

式中，T_r 是指答复时间；T_q 是指留言评论时间。

时间段 T 越短，即可说明该留言问题解决的越好，具有很强的时效性；反之，可以说明该问题没能在较短的时间完成，解决落实不到位。

综上所述，该评价方案是根据答复的相关性、完整性、可解释性、时效性等角度来对答复内容进行评价，将答复内容分为不合格答复、合格答复、优质答复三类，并在优质答复中进行时效排序，排序较高的答复留言能保证解决问题的及时性。

3、结果分析

3.1 问题 1 结果分析

在对留言内容进行数据处理后，我们采用了五个模型对数据进行训练，最后评价得出最好的分类模型为线性向量机模型。根据留言的一级标签分类结果，对比实际的一级标签，我们可以得到如下的稀疏矩阵。

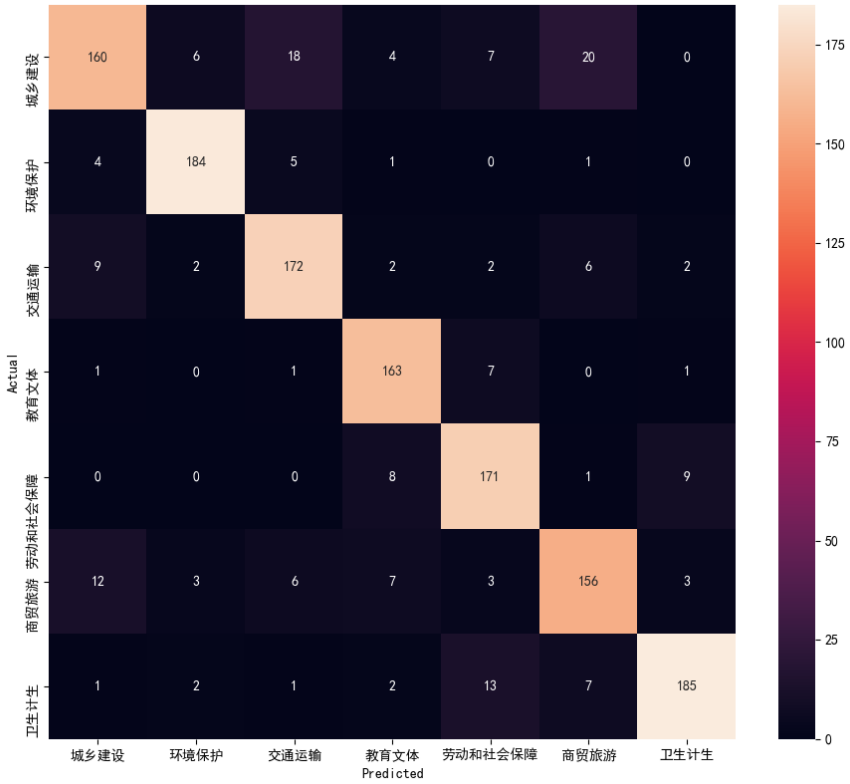


图 10：问题一稀疏矩阵

由图可知，预测结果的绝大多数都位于对角线上（预测标签=实际标签），也就是我们希望它们会落到的地方。只有少数不在对角线上，而这些留言提取出来的关键词包含多个主题，导致预测不准确。因此，可以得出，我们的分类模型的分类结果较好，可以使用该模型对大多数留言信息进行较为准确的分类。

3.2 问题 2 结果分析

在对留言信息去重去空后，进行中文分词，对中文评论进行向量化，然后对留言进行双阈值的 Single-Pass 聚类，挖掘出一定时间内反映较多的热点问题，通过文本挖掘，最后我们可以得出最热的五个热点问题如下表。

3.3 问题 3 结果分析

基于答复的相关性、完整性、可解释性、时效性角度，建立的答复评价模型，可以较全面的、多角度地判断该答复是否完善。在附件 4 中可以看出，达到“优质回复”的答复还只是少数，其中能达到在较短的时间内回复的评论，经筛选是留言编号为 32914，对应的评论回复。该回复在留言问题提出 4 小时 5 分后，给予回复，具体回复内容如下：

“‘UU0082355’您好！您通过平台《问政西地省》的留言收悉，已转有关

部门答复反馈，谢谢！2019 年 9 月 12 日尊敬的“心殤”网友：您好！根据西地省教育厅 2016 年 1 月 26 日所发的《西地省义务教育学校办学标准[政府发文]4 号文件精神，目前我省对中小学校教室是否安装空调没有明确要求；近几年来随着 B4 区城区的急速发展，我区每年新建一至二所学校，教育投入巨大，在全市的城区中排名第一。目前区政府暂时没有足够资金安排教室安装空调。另外，在 B4 区区委、区政府的关心下，自 2017 年 9 月开始，B4 区所有公办中小学校及幼儿园安装了直饮水设备，费用由政府财政支付；由于供水设备的耗电量非常大，大大增加了学校的运营成本，如果在教室全面安装空调，一方面学校势必电力增容，改造费用较高；同时空调使用的电费也是一笔很大的开支，目前的生均经费远远无法满足这些需求。 2019 年 9 月 17 日”

上述评论，满足了相关性、完整性、可解释性的要求，且在最短的时间内完成回复，属于优质回复中的范本。

4、结论

对网络问政平台的留言信息进行一级标签分类，了解群众留言的热点问题，对于了解民声，帮助解决民众的难题有着重大意义。

由于较为传统的人工进行留言分类操作不仅效率低，而且准确率不高，所以本文采用线性向量机对留言进行分类，得到了较好的结果。运用基于双阈值的 Single-Pass 聚类算法对于留言进行热点挖掘，有利于发现民众的主要问题并且及时解决。最后运用留言回复评价模型对留言回复进行质量评价，可知留言回复是否合格，运用该系统可以及时反馈相关部门的群众留言回复质量，并且提升和改进。

5、参考文献

[1]石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.

[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]许甜华,吴明礼.一种基于 TF-IDF 的朴素贝叶斯算法改进[J].计算机技术与发展,2020,30(02):75-79.

[4]姜天宇,王苏,徐伟.基于朴素贝叶斯的中文文本分类[J].电脑知识与技术,2019,15(23):253-254+263.

[5]李华. 酒店回复及消费者评论文本特征对酒店销量的影响研究[D].华侨大学,2018.

[6]张培伟. 基于改进 Single-Pass 算法的热点话题发现系统的设计与实现[D].华中师范大学,2015.

[7]张蕾. 基于机器学习的网络舆情采集技术研究与设计[D].电子科技大学,2014.

[8]周炎涛,唐剑波,吴正国.基于向量空间模型的多主题 Web 文本分类方法[J].计算机应用研究,2008(01):142-144.