关于“智能政务”中通过数据挖掘分析热点政务

摘要

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，网络留言成为关键渠道，但伴随着的是，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。因而采取大数据方式进行分析分类，对不同留言进行精准答复，并对答复的相关性、完整性、可解释性等角度对 答复意见进行评价，成了我们所需要完善的内容。

针对问题一，我们建立贝叶斯模型，在我们已知的基础上先验概率与条件概率的情况下进行模式分类，待分样本的分类结果取决于各类域中样本的全体,但由于总体的概率分布和各类样本的概率分布函数是不确定的。为了解决上述问题，提出了一种基于 SVM-EM 算法的贝叶斯算法,首先利用非线性变换和结构风险最小化原则将流量分类问题转化为二次寻优问题,然后要求 EM 算法对贝叶斯算法要求条件独立性假设进行填补，利用贝叶斯算法进行群众留言分类,提高了分类的准确性和稳定性。最后评判标准 F-score 公式，建立评价标准模型进行评价。

针对问题二，我们先用 Anaconda 中的 Notebook 工具对附件 3 的内容进行 Jieba 分词操作，并对其进行留言主题的高频词筛选，将筛选出的高频词汇记录并分析整理成表。将高频词汇表导入 excel 中对附件三主题内容进行查询检索排序，并结合点赞数构成用列表。在用例表中将主题数与点赞数结合构建热度指数动态方程，建立模型进行热度指数计算排名，整理数据构成热度问题值表，从热度排名排列热点问题留言明细表。

针对问题三，针对问题三，我们使用 LDA 主题模型进行模型建立，获得“主题”和“主题-词项”两个分布结果。根据主题模型特色和结果计算出留言主题代表词与答复意见的相似程度，并根据相似度大小和相应的选择规则选择主题词。随后，提出契合度概念作为对于留言标签和答复意见之间的匹配性和相似性的衡量，该概念的本质为文本相似度的衡量，构建思路上参考多种相似度计算方法，最终契合度以加权主题词相似度占比来体现，以实现从答复的相关性、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。

关键词：贝叶斯模型，SVM – EM 算法，最小化原则，F-score 公式，Jieba 分词操作，热度指数动态方程，主题模型，契合度

Abstract

In recent years, with WeChat, such as weibo, mayor mailbox, sun hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, network become the key message channel, but with all kinds of public opinion related text data quantity rising, leave a message to past mainly depends on artificial to divide and hot spots of relevant departments work has brought great challenge. Therefore, it has become necessary for us to take the big data approach to analyze and classify, give accurate answers to different messages, and evaluate the replies from the perspectives of relevance, completeness and interpretability.

In view of problem 1, we establish a bayesian model and conduct pattern classification on the basis of our known prior probability and conditional —probability. The classification result of samples to be divided depends on all samples in all kinds of domains, but the probability distribution function of the population and all kinds of samples is uncertain. In order to solve the above problems, this paper proposes a bayesian algorithm based on SVM - EM algorithm, first of all, using nonlinear transformation and structural risk minimization principle will flow classification problem into a quadratic optimization problem, and then demands the EM algorithm for bayesian algorithm to fill conditional independence assumption, the mass message classification using bayesian algorithm, improves the classification accuracy and stability. Finally, the evaluation standard f-score formula is used to establish the evaluation standard model for evaluation.

For question 2, we first use the Notebook tool in Anaconda to perform Jieba word segmentation operation on the contents of attachment 3, and then filter the high-frequency words of the message topic, record and analyze the filtered

high-frequency words into a table. The high frequency vocabulary was imported into excel to search and sort the subject content of attachment 3, and the list was composed with the number of likes. In the use case table, the topic number and the thumb up number are combined to construct the dynamic equation of heat index, the model is established to calculate the heat index ranking, the data is arranged to form the heat problem value table, and the message list of hot issues is arranged from the heat ranking.

For problem 3, we use LDA theme model to build the model and obtain two distribution results of "theme" and "theme-word item". According to the characteristics and results of the theme model, the similarity degree between the representative words of the message subject and the comments of the reply is calculated, and the subject words are selected according to the similarity degree and the corresponding selection rules. Then, put forward the concept of "fit" reply to a message label and opinions match between sex and similarity measure, the essence of the concept of text similarity measure, to construct reference on a variety of similarity calculation method, the final fit with weighted keywords similarity ratios reflect, in order to realize the reply from the Angle of relevance, integrity and interpretability reply opinion the quality of a set of evaluation scheme is given.

Key words: bayes model, svm – em algorithm, minimization principle, f-score formula, Jieba word segmentation, heat index dynamic， quation, LDA theme model, Integrating degree

目录

一、问题描述

1.1 问题背景

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。 因此通过大数据进行挖掘留言信息，成为减轻政府工作负担，方便答复民众的重要措施手段。

1.2 问题分析

针对问题一，处理群众留言的划分体系问题，由于传统人工分类工作量大，我们引入了我们建立贝叶斯模型，在我们已知的基础上先验概率与条件概率的情况下进行模式分类，待分样本的分类结果取决于各类域中样本的全体，但由于总体的概率分布和各类样本的概率分布函数是不确定的。为了解决上述问题，提出了一种基于 SVM-EM 算法的贝叶斯算法,首先利用非线性变换和结构风险最小化原则将流量分类问题转化为二次寻优问题,然后要求 EM 算法对贝叶斯算法要求条件独立性假设进行填补，利用贝叶斯算法进行群众

留言分类,提高了分类的准确性和稳定性。此外，需要对分类的数据进行评价，我们使用评判标准 F-score 公式，建立评价标准模型进行评价。

针对问题二，我们对热点数据挖掘时，考虑到不同主题留言时可能会有一定的联系性，因此我们不仅筛选出不同类型主题更要从留言详情中分析两者的相关性；除此之外，我们在分析附件 3 时发现留言主题个数多的主题点赞数不一定多，而反映较少的主题却有可能很多人关注到，因此对热点排名不应该仅仅从留言主题个数来决定，更应该与点赞数想结合看待。

针对问题三，我们应该对政府相关部门答复意见进行规范，并从从答复的相关性、完整性，依照一定准则对其评价，提出契合度概念作为对于留言标签和答复意见之间的匹配性和相似性的衡量，最终契合度以加权主题词相似度占比来体现。

二、模型假设

假设一：留言内容为认真填写，内容以中文字符为主

假设二：测试数据少无实义词，确保 jieba 分词后将无实义词筛选剔除

假设三：留言主题和答复之间存在契合度可挖掘

三、符号说明

表 1 建模过程中符号说明

公式符号	符号说明
T	测试数据
h_{ML}	极大似然值
V_{MAP}	目标值
T_i	约束条件对应 Lagrange 乘子
b	分类阈值
K	核函数
$\mathbf{P_i}$	第 i 类的查准率
$\mathbf{R_i}$	第 i 类的查全率
S（t）	留言热度
α	传播系数
S’（t）	热度指数
D	文档
W	主题代表词

四、模型建立与数据处理

4.1 问题一：

4.1.1贝叶斯模型概述[1]：

朴素贝叶斯分类技术以贝叶斯定理为基础，通过数据的先验概率，利用贝叶斯公式计算出其后验概率，并选择具有最大后验概率的类作为该对象所属的类。

在给定训练数据 T 时，确定假设空间 λ 中的最佳假设。贝叶斯算法提供了从先验概率 $P(\lambda)$以及 $P(T)$和 $P(T|\lambda)$计算后验概率 $P(\lambda|T)$的方法，其公式为：

$$P(\lambda|T) = \frac{P(T|\lambda)P(\lambda)}{P(T)}$$

其中，$P(\lambda)$表示 λ 的先验概率，$P(T)$表示待观察训练数据 T 的先验概率，$P(T|\lambda)$表示给定 λ 时观察数据 T 对应的概率，$P(\lambda|T)$表示 λ 的后验概率，即给定训练数据 T 时 λ 成立的概率。假设集合 λ 并寻找观察数据 T 在其中的概率最大的假设 h∈ λ，称为极大后验（MAP）假设。利用贝叶斯公式计算每个待选假设的后验概率，以此来确定 MAP 假设，即

$$h_{MAP} = \operatorname{argmax} P(\lambda|T) = \operatorname{argmax} \frac{P(T|\lambda)P(\lambda)}{P(T)} = \operatorname{argmax} P(T|\lambda)P(\lambda), \quad h \in H$$

$P(T)$是不依赖于 λ 的常量。在某些情况下，可假定 λ 中每个假设有相同的先验概率（即对 H 中任意λ_i和λ_j，$P(\lambda_i) = P(\lambda_i)$）。进一步简化，只需考虑 $P(T|\lambda)$来寻找极大可能假设。$P(T|\lambda)$常称为给定 λ 时数据 T 的似然度，而使 $P(T|\lambda)$最大的假设被称为极大似然h_{ML}，即

$$h_{ML} = \operatorname{argmax} P(T|\lambda) \quad h \in H$$

朴素贝叶斯分类器应用的学习任务中，每个实例 x 可由属性值的合取描述，而目标函数 f(x)从某有限集合 V 中取值。 贝叶斯方法的新实例分类

目标是在给定描述实例的属性值 $\{a_1, a_2, \cdots, a_n\}$下 ，得到最可能的目标值 v_{MAP}，即

$$v_{MAP} = \operatorname{argmax} \mathrm{P}\left(v_j \middle| a_1, a_2, \cdots, a_n\right) \text{ ,v} \in \mathrm{V}$$

基于贝叶斯公式重写为

$v_{MAP} = argmax \frac{P(a_1, a_2, \cdots, a_n | v_j) P(v_j)}{P(a_1, a_2, \cdots, a_n)} = argmax\, P\left(a_1, a_2, \cdots, a_n \middle| v_j\right) P\left(v_j\right)$ ，$v_j \in \mathrm{V}$

朴素贝叶斯 N B 分类器算法基于一个简单的假定：估算目标值时属性 $\{a_1, a_2, \cdots, a_n\}$之间条件是相互独立的 ，则观察到$a_1, a_2, \cdots, a_n$的联合概率正好是对每个单独属性的概率乘积 ，为

$$\mathrm{P}(a_1, a_2, \cdots, a_n | v_j) = \prod_i \mathrm{P}(a_i | v_j)$$

将其代入式 (2)中 ，可得到 NB 分类器所使用的方法为

$$v_{\mathrm{NB}} = \operatorname{argmax} P\left(v_j\right) \prod_i P(a_i | v_j)$$

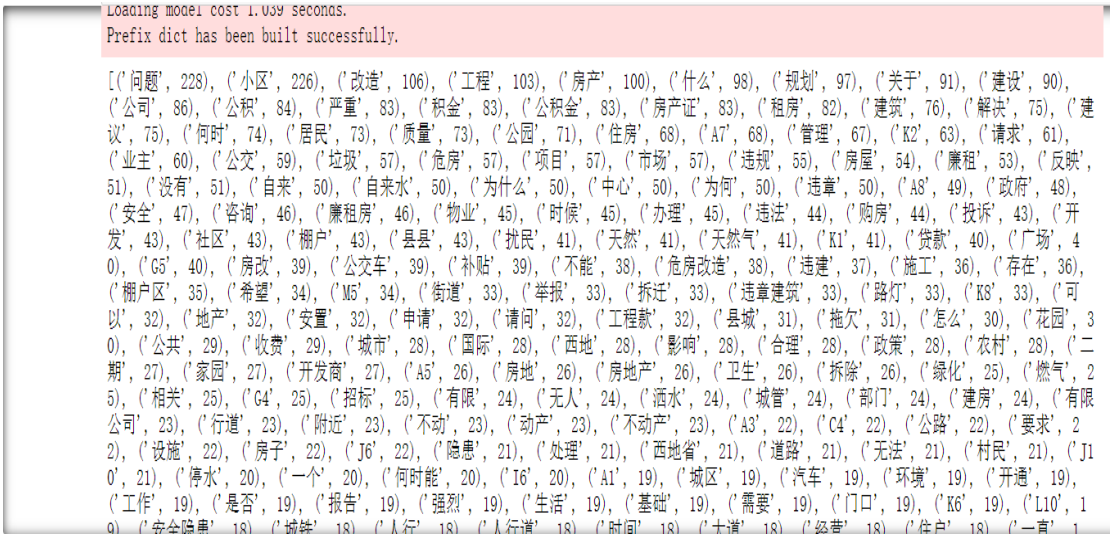
当所需的条件独立性能够被满足时 ，朴素贝叶斯分类器输出的（v_{NB}）等于 MAP 分类[4]。

4.1.2 朴素贝叶斯群众留言分类流程和步骤

朴素贝叶斯分类算法利用贝叶斯定理的优势 ，在网络留言分类中有广泛应用 ，是文本分类最为精确的技术。 在智能文本分类技术中 ，通过贝叶斯分类器的“自我学习”智能技术 ，能有效对群众留言进行分类。 朴素贝叶斯分类分为以下 3 个阶段[2]。

第 1 阶段：准备工作阶段。收集群众留言作为样本 ，确定每个一级分类属性 ，并对每个特征属性进行适当划分 ，然后通过 Python 中的 jieba 库提取留言样本中主题和留言中的字符串（如下） ， 通过对词语的使用频率排序，去除无关与无实意的词语，获取到每个一级分类标签中常用词语，建立对应的数据库分类 ，输出特征属性和训练样本。

图 1 据词库



第 2 阶段：分类器训练阶段。计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计 ，创建贝叶斯概率库统计出每个字串在群众留言中出现的概率以及在正常留言中出现的概率 ，然后根据公式计算出留言中含某字串则为一级分类标签留言的概率。

第 3 阶段：应用阶段。使用训练好的贝叶斯分类器对分类项进行分类 ，其输入是分类器和待分类项 ，输出是待分类项与类别的映射关系。通过对群众留言样本的训练 ， 贝叶斯分类器可以自动获取群众留言的特征 ，并根据特征的变化 ，对群众留言进行有效分类。

4.1.3贝叶斯模型评测[3]：

优点

朴素贝叶斯算法假设了数据集属性之间是相互独立的，因此算法的逻辑性十分简单，并且算法较为稳定，当数据呈现不同的特点时，朴素贝叶斯的分类性能不会有太大的差异。换句话说就是朴素贝叶斯算法的健壮性比较好，对于不同类型的数据集不会呈现出太大的差异性。当数据集属性之间的关系相对比较独立时，朴素贝叶斯分类算法会有较好的效果。

缺点

属性独立性的条件同时也是朴素贝叶斯分类器的不足之处。数据集属性的独立性在很多情况下是很难满足的，因为数据集的属性之间往往都存在着相互关联，如果在分类过程中出现这种问题，会导致分类的效果大大降低。

4.1.4优化 – 基于 SVM – EM的朴素贝叶斯分类算法[2]

朴素贝叶斯算法是一种简单而高效的分类算法 ,但是它的条件独立性假设极大影响了分类性能 ,因此 ,我们引入了一种基于支持向量机 SVM、EM 算法融合的改进朴素贝叶斯分类算法: SVM-EM-NB 算法 ,该算法通过贝叶斯统计方法对群众留言进行分类 ,挖掘有效的数据信息 ,并结合支持 SVM 技术预测群众留言概率 ,贝叶斯前验分布和后验分布用来估计 SVM 中的参数。

4.1.5支持向量机 SVM 训练 [3]

支持向量机 SVM 是目前一种新兴的技术 ,在文本分类方面越来越受到重视。 支持向量机 SVM 的提出有很深的理论背景 ,其训练的本质是解决一个二次规划问题 ,得到的是全局最优解。

SVM 训练的基本思想概括为: 1、对线性可分情况进行分析 ,通过使用核函数与非线性转换算法 ,将低维输入空间线性不可分的样本变化为高维特

征空间并使其线性可分，从而使得高维特征空间可以采用线性算法对样本的非线性特征进行线性分析。 2、于结构风险最小化理论在特征空间中构建最优分割超平面，使得学习器得到全局最优化。 SVM 分类器的优点在于通用性较好，可以提高泛化性能，解决非线性问题，且分类精度高，分类速度与训练样本个数无关，其和朴素贝叶斯分类算法融合将大大提高查准率和查全率。

设分类线性方程为 x·w+ b=0,对它进行归一化，使得对线性可分的样本集 $(x_i, y_i)\ (i = 1, \dots, n, n \in Rd, y \in \{+1, -1\})$满足约束条件:

yi [(w· xi)+ b]- 1≥ 0(i= 1,⋯ , n)。

利用 Lagrange 优化方法将最优分类面问题转化为对偶形式，即: 在约束条件 $\sum_{i=1}^{n} y_i T_i = 0$和$T_i \geq 0$，$(i = 1, \dots, n)$ 下，对 求解下列函数的最大值

$$\mathrm{MaxQ}(T) = \sum_{i=1}^{n} T_i - \frac{1}{2} \sum_{i,j=1}^{n} T_i T_j y_i y_j (x_i \bullet x_j)$$

其中，T_i为原问题中与每个约束条件相对应的 Lagrange 乘子。 这样就转换为一个不等式约束下二次函数寻优的问题，即存在惟一解。 对应的样本是支持向量，采用线性分类解上述问题后得到的最优分类函数为

$$f(x) = sgn\{(w \bullet x) + b\} = sgn\{\sum_{i=1}^{n} T_i y_i (x_i \bullet x) + b\}$$

其中，b 表示分类阈值，可以用任一个支持向量求得。

对非线性问题，通过非线性映射 H 把输入空间的样本转化为某个高维特征空间，核函数 $K(x_i, x_j) = H(x_i) \cdot H(x_j)$，在高维空间进行内积运算时，无需了解变换 H 的形式。这样不仅能实现非线性变换后的线性分类，而且还没有增加时间复杂度，此时对偶形式变为

$$MaxQ(T) = \sum_{i=1}^{n} T_i - \frac{1}{2}\sum_{i,j=1}^{n} T_i T_j y_i y_j K(x_i \bullet x_j)$$

而相应的分类函数变为

$$f(x) = sgn\left[\sum_{i=1}^{n} T_i y_i k(x_i, x) + b\right]$$

SVM 将大的分类工作量放在输入空间而不是高维特征空间中完成 ,避免算法可能导致的“维数灾难” ,快速解决二次规划的问题 ,具有较高的训练准确度

4.1.6 EM 算法[1]

期望最大化算法 EM 是一种迭代方法 ,主要用来计算后验分布的极大似然估计和用于不完整数据的填补应用。其基本步骤为: 1)在已知变量和当前参数的情况下估计缺失属性的期望 ,即 E 步。 2)重估参数的期望 ,保证参数值是在 E 步中补充完整数据的极大似然值 ,即 M 步最大化问题 ,最后就可以保证能得到局部收敛最大值。

定义 2 个样本空间 X 和 Y ,其中 X 是完整数据空间 ,Y 是观察数据(即 incomplete data) ,令 Z表示添加数据 ,那么 $X = (Y,Z)$,h 表示参数集合 , $g(y|h)$表示观察后验概率密度函数 , $f(x|h)$表示添加数据Z 后得到的后验密度函数 , $k(x|y,h)$表示给定数据 h 和观察数据 y下 x 的条件密度函数。 则

$$k(x|y,h) = f(x|h) / (y|h)$$

定义似然函数 $L(h) = log(g(y|h))$,得出

$$L(\hbar) = log(f(x|\hbar) - log\, k\,(x|y,\hbar)$$

定义似然函数 $L(\hbar) = log\,(g\,(y|\hbar))$，得出

$$L(\hbar) = log(f(x|\hbar) - log\, k\,(x|y,\hbar)$$

定义函数

$$Q(\hbar'|\hbar) = E(\log(f(x|\hbar')|y,\hbar)$$

和

$$H(\hbar'|\hbar) = E(log\,(k\,(x|y,\hbar'))|y,\hbar)\ ,$$

最后得到

$$Q(\hbar'|\hbar) = L(\hbar') + H(\hbar'|\hbar)。$$

记$\hbar_i$为第 i+1 次迭代开始时参数的估计值，则第 i+1 次迭代的步骤为:

步骤 1: 计算每个隐藏变量z_{ij}的期望值$E[z_{ij}]$

假定当前假设 $\hbar = <\_1,\_2>$ 成立。

步骤 2: 计算一个新的极大似然假设 $\hbar' = <\_1',\_'2>$，假定由每个隐藏变量 z_{ij}所取的值为步骤 1 中得到的期望值$E[z_{ij}]$, 将假设 $\hbar = <\_1,\_2>$ 替换为新的假设 $\hbar' = <\_'1,\_'2>$，然后循环计算每个z_{ij}的期望值。 此$E[z_{ij}]$正是实例x_i由第 j 个正态分布生成的概率，为

$$E[z_{ij}] = \frac{p(x = x_{i|\_} = \_j)}{\sum_{n=1} p(x = x_i|\_ = \_n)} = \frac{e^{-\frac{1}{2e^2}(x_i - \_j)^2}}{\sum_{n=1}^{2} e^{-\frac{1}{2e^2}(x_i - \_n)^2}}$$

第 1 步可将当前值 <_ 1 ， _ 2> 和已知的 xi 代入到式 （ 3)中实现。在第 2 步，使用第 1 步中得到的 E [zij]来导出一个新的极大似然假设 h′ =〈_ ′ 1 ， _ ′ 2〉 。 这时的极大似然假设为

$$\_j \leftarrow \frac{\sum_{i=1}^{m} E[z_{ij}]x_i}{\sum_{i=1}^{m} E[z_{ij}]}。$$

新的表达式只是对 _ j 的加权样本均值 ,每个实例的权重为其由第 j 个正态分布产生的期望值。

4.1.7 SV M − EM − NB算法[2]

首先用优化训练的 SVM 训练留言集解决一个二次规划问题 ,使学习器得到一个全局最优解 ,然后把数据集分成完整集和缺失集 ,计算缺失属性的数据项与完整属性数据项的相关度 ,取相关度最大的数据项对应的属性作为缺失属性的一个估计值 ,此估计值作为 EM 算法的初始值 ,然后执行 EM 算法的两步 ,完成极大似然估计 ,用最后估计的值来完成缺失属性的填补 ,最后用朴素贝叶斯分类算法对完整数据集进行分类。

输入: $T = \{X\,1\,,X\,2\,,\ldots\,,Xn\}$, 其中 ,$X_1\,,X_2\,,\ldots\,,X_n$为原始属性集 ,$\lambda = \{t1\,,t2\,,\ldots\,,tm\}$为类别属性。

输出: 样本 X 的类别。

算法主要步骤为:

步骤 1 把数据集 T 分为 2 个数据子集 T_i和T_j。T_i中的记录全部为完整记录 ,任何属性不含缺失值; T_j中的记录为不完整记录 ,即属性中含有一个及以上的缺失值。

步骤 2 调用 EM 算法 ,完成缺失数据填补。

步骤 3 随机选择 4 /5 的样本作为训练集 ,剩余 1 /5 的样本作为测试集 ,计算训练集样本的先验概率 P (λ)。

步骤 4 在假设类条件独立的情况下 ,根据贝叶斯公式计算条件 概 率 P$(Xi|\lambda$〉

步骤 5 根据式 (1)计算后验概率 P $(\lambda|X_i\,)$,输出类别 ,求出分类准确率。

改进的朴素贝叶斯复合智能算法 ,不仅能够快速得到最优分类特征子集 ,而且大大提高了其学习和分类的效率和准确率 ,降低了分类器的错误率 ,具有较高的实用价值。在训练集达到 5 000 之后 ,查全率达到了 95%以上 ,远高于朴素贝叶斯算法 ,这也正说明了其相对朴素贝叶斯算法的优越性。适当增加训练集样本数 ,改进的贝叶斯复合智能算法将在查全率和误报率方面有更好的表现。

4.1.8 分类的评判标准 F-score

目的：建立评价标准进行评价

现已有分类好的标准数据，现假设有测试数据，需对测试数据进行评价，现采用 F-score 算法进行评价。

F-score 原理：

简单的二分法可以判断 True(1)或 Not(0)，在已知结果的时候我们可以判断原先的预测结果为 positive 或 negative

首先建立基于混淆矩阵（Confusion Matrix）的一级分类标准

表 2

混淆矩阵		预测值	
		positive	negative
真实值	positive	TP	FN(error)
	negative	FP(error)	TN

在混淆矩阵的基础上建立两个二级分类标准 P、R

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

在二级分类标准的基础上建立三级分类标准 F-score

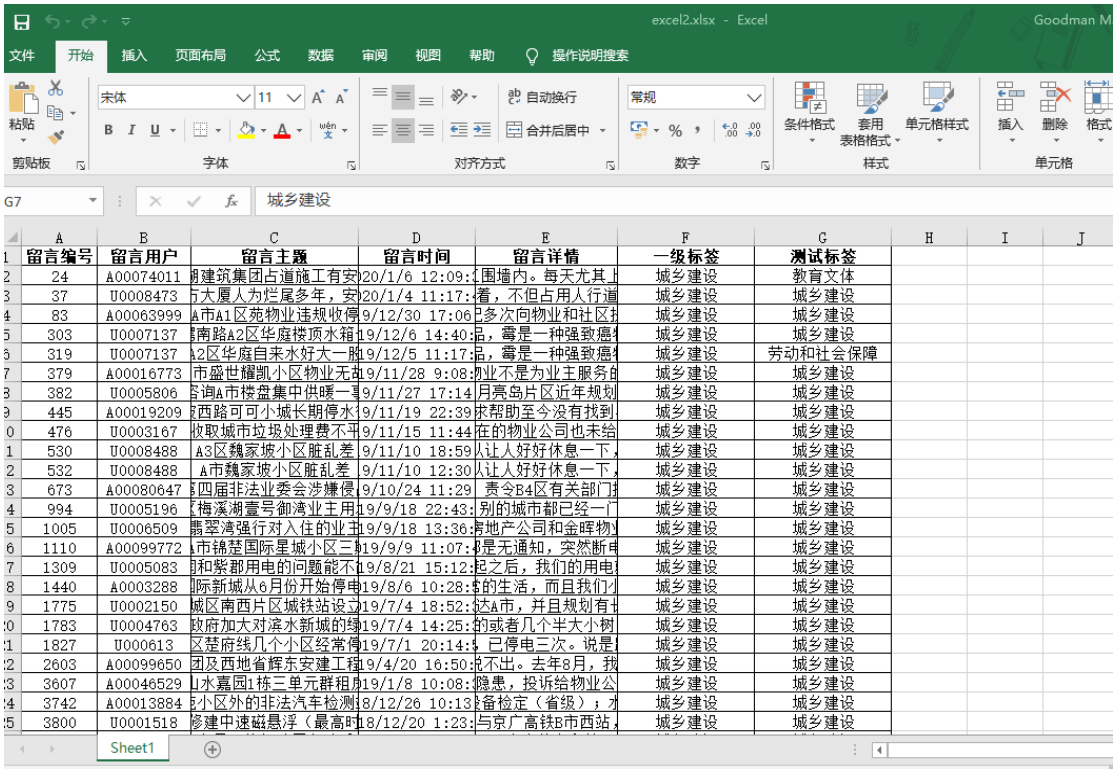
$$\mathrm{F_1}=\frac{1}{n}\sum_{i=0}^{n}\frac{2P_iR_i}{p_i+R_i} \quad (0<F_1<1)$$

2

程序中使用了 numpy、pandas、sklearn 库，利用 sklearn 中的 f1_score 函数可以轻松得到F_1(越接近 1 表示分类越准确)

假设我们已经做好了分类（测试分类），且知道正确的分类，将测试分类作为一列放在附件 2 中的一级分类右边如图

图 2



留言编号	留言用户	留言主题	留言时间	留言详情	一级标签	测试标签
24	A00074011	朗建筑集团占道施工有安	20/1/6 12:09:	围墙内。每天尤其上	城乡建设	教育文体
37	U0008473	古大厦人为烂尾多年，安	20/1/4 11:17:	着，不但占用人行道	城乡建设	城乡建设
83	A00063999	A市A1区苑物业违规收停	9/12/30 17:06	多次向物业和社区投	城乡建设	城乡建设
303	U0007137	南路A2区华庭楼顶水箱	19/12/6 14:40:	品，霉是一种强致癌	城乡建设	城乡建设
319	U0007137	A2区华庭自来水好大一股	9/12/5 11:17:	品，霉是一种强致癌	城乡建设	劳动和社会保障
379	A00016773	市盛世耀凯小区物业无故	19/11/28 9:08:	物业不是为业主服务的	城乡建设	城乡建设
382	U0005806	咨询A市楼盘集中供暖一事	9/11/27 17:14	月亮岛片区近年规划	城乡建设	城乡建设
445	A00019209	西路可可小城长期停水	9/11/19 22:39	求帮助至今没有找到	城乡建设	城乡建设
476	U0003167	收取城市垃圾处理费不平	9/11/15 11:44	在的物业公司也未给	城乡建设	城乡建设
530	U0008488	A3区魏家坡小区脏乱差	9/11/10 18:59	人让人好好休息一下，	城乡建设	城乡建设
532	U0008488	A市魏家坡小区脏乱差	9/11/10 12:30	人让人好好休息一下，	城乡建设	城乡建设
673	A00080647	四届非法业委会涉嫌侵	9/10/24 11:29	责令B4区有关部门按	城乡建设	城乡建设
994	U0005196	梅溪湖壹号御湾业主用	19/9/18 22:43:	别的城市都已经一门	城乡建设	城乡建设
1005	U0006509	翡翠湾强行对入住的业主	19/9/18 13:36:	房地产公司和金晖物业	城乡建设	城乡建设
1110	A00099772	A市锦楚国际星城小区三	19/9/9 11:07:	都是无通知，突然断电	城乡建设	城乡建设
1309	U0005083	和紫鄙用电的问题能不	19/8/21 15:12:	起之后，我们的用电	城乡建设	城乡建设
1440	A0003288	际新城从6月份开始停电	19/8/6 10:28:	的生活，而且我们小	城乡建设	城乡建设
1775	U0002150	城区南西片区城铁站设立	19/7/4 18:52:	达A市，并且规划有七	城乡建设	城乡建设
1783	U0004763	政府加大对滨水新城的绿	19/7/4 14:25:	的或者几个半大小树	城乡建设	城乡建设
1827	U000613	区楚府线几个小区经常停	19/7/1 20:14:	已停电三次。说是	城乡建设	城乡建设
2603	A00099650	团及西地省辉东安建工程	19/4/20 16:50:	说不出。去年8月，我	城乡建设	城乡建设
3607	A00046529	山水嘉园1栋三单元群租房	19/1/8 10:08:	隐患，投诉给物业公	城乡建设	城乡建设
3742	A00013884	5小区外的非法汽车检测	8/12/26 10:13	备检定（省级）；水	城乡建设	城乡建设
3800	U0001518	修建中速磁悬浮（最高时	18/12/20 1:23:	与京广高铁B市西站，	城乡建设	城乡建设

最右边新增的一列为测试标签

测试标签相对于一级标签只修改了两个单元格，相似度极高，预计将接近 1。

经代码测试（代码详情见附件第一问程序）后结果为

0.9998476508803523

(接近 1 表明两列文本相似度极高,符合预期)

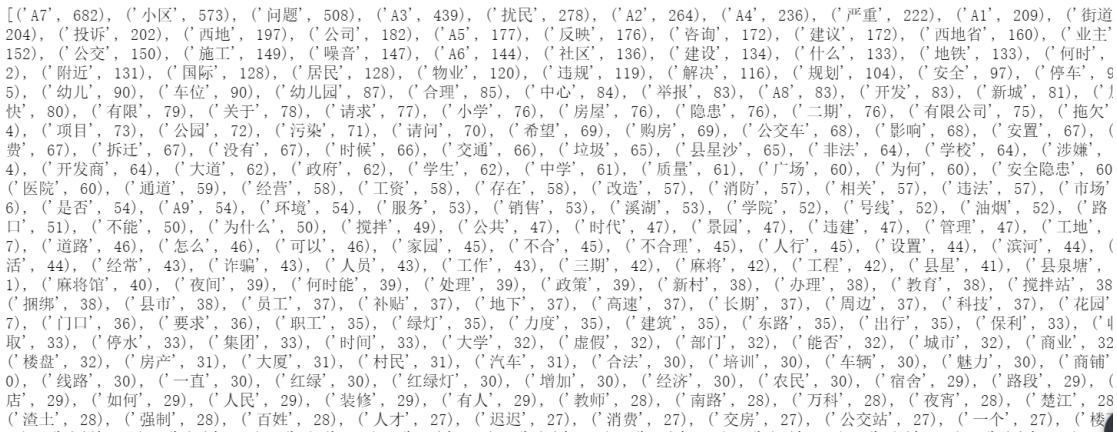
4.2、问题二：

4.2.1 数据挖掘和处理：

在此我们针对问题二中附件三数据进行挖掘和筛选：我们采用对关键词出现频率优先，而后查看点赞数的方法从而评选热度的方法进行排名先后顺序。

利用 Python 中 jieba 组件对从 excel 表格中转化的留言主题进行挖掘，查看高频词汇（代码详情见附件问题二程序代码），如下图所示：

图 3



在这里我们可以通过代码分析出附件 3 中出现相对高的词语频率（除去像“问题、严重”这类无实义的词语以及部分重复意思相近的词语外）得出群众对政府意见高频热词整理如下表：

表 3 高频热词频数

词语	频数
A7	682

小区	573
A3	439
扰民	278
A2	264
公司	182
西地省	160
业主	152
公交	150
施工	149
噪音	147
社区	136
地铁	133
幼儿	90
车位	90
新城	81
小学	76

4.2.2 数据分析

通过表一的高频词对附件三留言主题进行如下检索图所示（注：以下筛选图仅列举部分）：

筛选图 1

留言主题	留言时间	留言详情	反对数	点赞数
丽南路丽发新城居民区附近搅	2019/11/19 18:07:54	米处建搅拌站，运渣	0	1
区丽发新城附近建搅拌站噪	2019-11-13 11:20:21	方建搅拌站。可想而	0	0
丽发新城小区旁边建搅拌站	2019-12-21 15:11:29	影响几千名学生的健	0	1
城违建搅拌站，彻夜施工扰	2019/12/26 13:55:15	量；3、搅拌站几百米	0	0
市A2区丽发新城道路坑坑洼洼	2019/7/3 12:03:51	道路坑坑洼洼，下雨	0	1
富绿物业丽发新城强行断业主	2019/6/19 23:28:27	提供地摊上买的收据	0	242
新城附近修建搅拌厂，严重	2019/11/25 10:17:56	北还得了疾病住院，	0	0
侧面建设混泥土搅拌站，粉尘	2019/11/19 14:51:53	大的粉尘，严重影响	0	2
城附近修建搅拌站，污染环境	2020-01-02 00:00:00	量和声环境质量急剧	0	4
发新城小区附近违建搅拌站噪	2019-12-10 12:34:21	搅拌站。该搅拌站的	0	0
城附近违规乱建混凝土搅拌站	2019-12-27 23:34:32	强烈呼吁政府和有关	0	0
城小区附近搅拌站噪音扰民和	2020-01-25 09:07:21	大吵，烦死了不仅吵	0	0
新城小区旁边的搅拌厂是否	2019-12-06 12:21:32	了噪音和灰尘。这给	0	0
A2区丽发新城小区附近太吵了	2020-01-26 19:47:11	搅拌厂是怎么回事！	0	0

筛选图 2

留言主题	留言时间	留言详情	反对数	点赞数
景园滨河苑捆绑销售车位的维	2019-08-23 12:22:00	发文件，强制要求耶	0	0
滨河苑协商要求购房同时必约	2019-08-16 09:21:33	作为一名退休职工，	1	12
路职工定向商品房伊景园滨	2019-08-10 18:15:16	未给出首付款详细的	0	0
市伊景园滨河苑捆绑车位销	2019/8/7 19:52:14	令五申禁止捆绑车位	0	0
园滨河苑捆绑车位销售合法吗	2019-08-14 09:28:31	购买12万一个的车位	0	1
对伊景园滨河苑强制捆绑销	2019-08-03 10:03:10	的违法捆绑销售！这	0	2
园滨河苑强行捆绑车位销售给	2019-08-23 12:16:03	位，不买车位就取消	0	0
河苑项目绑定车位出售是否	2019-08-28 19:32:11	车位。而单个车位高	0	0
景园滨河苑定向限价商品房	2019-07-28 13:09:08	，无视法律法规，	0	0
伊景园滨河苑销售若干问题	2019-08-22 00:00:00	价值12万的成本价	0	3
市伊景园滨河苑捆绑销售车位	2019/8/1 22:42:21	规购房合同，强制收	0	1
市伊景园·滨河苑欺诈消费者	2019-08-24 00:00:00	交付车位定金，还不	0	0
城伊景园滨河苑为广铁集团的	2019-08-12 12:37:28	不起，要贷款，何来	0	0
河苑定向限价商品房项目违规	2019-08-28 10:06:03	资格相要挟，逼迫耶	0	0
市A市伊景园滨河苑捆绑销售车	2019/7/23 17:06:03	捆绑。2.希望这个	0	0
苑车位捆绑销售！广铁集团	2019-07-30 14:20:08	，说什么预购不用！	0	0
市A市伊景园滨河苑捆绑车位销	2019-08-19 10:22:44	动骨的，购房者中有	0	0
权益的A市伊景园滨河苑车位捆	2019-08-20 12:34:20	买房子的同时一对一	0	0
背景下的A市伊景园滨河苑车位	2019-08-30 16:32:12	还是和谐社会么？政	0	0
权益取消伊景园滨河苑捆绑销	2019-09-01 10:03:10	在还要求一户一车位	0	1
伊景园滨河苑捆绑销售车位的	2019-08-24 18:23:12	有，但属于城镇公共	0	0
伊景园滨河苑定向限价商品房	2019/7/28 10:36:05	购款18.5万，违法强	0	0
景园滨河苑开发商强买强卖	2019-08-21 19:05:34	项目，但现在却要求	0	2

筛选图 3

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
189381	A000109815	之城商铺无排烟管道，小区	2019/12/4 16:25:06	若罔闻，几年了都没	0	0
195095	A00039089	之城小区临街门面油烟直排	2019/09/05 12:29:01	24小时都是烟。请	0	3
198084	A00022429	魅力之城小区近百户楼板开裂	2019/10/23 15:01:30	入住，还要每月背负	0	0
205168	A00022429	魅力之城近百户房屋楼板、墙	2019/10/24 10:22:16	板开裂，墙面开裂，	0	1
232892	A00015335	城开发商未通知业主就进行车	2019/1/8 9:54:00	位开盘，开发商未通	0	1
233338	A00022429	力之城楼板和墙面开裂，请政	2019/11/13 10:58:25	验收没有质量问题。	0	0
236303	A00022429	科魅力之城有的房屋楼板严重	2019/10/29 16:57:02	一年多房屋就会出现	0	0
236798	A00039089	劳动东路魅力之城小区油烟	2019/07/28 12:49:18	有。每天油烟直排	0	4
240330	A00087099	魅力之城大降价折损前购买业	2019/3/21 9:03:00	，损失十几万，投诉	0	0
242792	A909115	小区一楼被搞成商业门面，	2019/08/26 08:33:03	我们的晚年生活。架	0	1
245136	A909117	城小区底层门店深夜经营，各	2019/09/04 21:00:18	声、拼酒声、炒菜烧	0	0
246362	A909114	小区底层商铺营业到凌晨，各	2019/08/26 01:50:38	严重影响健康。大家	0	0

筛选图 4

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
195917	A909119	经济学院组织学生外出打工	2019/11/05 10:31:38	时以上，（晚班时间	0	1
211395	A00050903	省财政经济学院校园宽带被	2019/9/14 17:57:34	手机卡都有免费赠	0	0
211800	A00046925	院食堂只开放十余个窗口，	2019/2/25 23:24:54	也因为卖完了而吃不	0	0
233759	A909118	市涉外经济学院强制学生实	2019/04/28 17:32:51	生必须去学校安排的	0	0
235521	A0006920	林三路涉外经济学院外街理发	2019/10/15 18:59:08	从早上8.30左右至	0	0
240721	A00050903	地省财政经济学院涉嫌宽带垄	2019/9/14 17:56:17	手机卡都有免费赠	0	0
242062	A00028889	外经济学院变相强制学生“社	2019/11/27 23:14:33	实践，起码是从学	0	0
264084	A00074365	济学院以报名名额已满拒绝让	2019/3/19 23:11:44	涉及想要参加考试	0	0
266368	A00038920	济学院寒假过年期间组织学生	2019/11/22 14:42:14	！虽说不是强制性的	0	0
360110	A110021	学院寒假过年期间组织学生去	2019-11-22 14:42:14	虽说不是强制性的	0	0
360111	A1204455	济学院组织学生外出打工合	2019-11-05 10:31:38	时以上，（晚班时间	1	0
360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	必须去学校安排的几	0	0
360113	A3352352	市经济学院强制学生外出实	2018-05-17 08:32:04	道！学校很小但是这	3	0
360114	A0182491	经济学院体育学院变相强制	2017-06-08 17:31:20	了合同，并且公司也	9	0

筛选图 5

	B	C	D	E	F	G
编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
388	A000111518	高速公路2018年招聘收费员工	2019/6/23 12:17:31	导都说以前的招聘么	0	1
576	A0006565	剥夺西地省高速基层员工的事	2019/4/24 14:50:16	法真是令人寒心到极	0	0
093	A00032672	高速建设开发总公司工资制	2019/4/24 12:43:25	自己定的，为何基层	0	0
516	A00052581	高考学生少数民加分名单何时	2019/5/26 19:59:58	布？加分审核有几	0	0
459	A00041751	省高速公路新能源汽车充电桩	2019/8/11 19:43:59	在城际间使用；西地	0	2
386	A00061821	公路建设开发总公司薪酬改革	2019/1/30 4:28:53	有法可依，有章可循	0	2
720	A000108426	地省高速收费员不熟悉政策，	2019/7/9 0:14:03	市出口（三一大道）	0	0
018	A00033700	高速一线员工年终奖金发放	2019/1/30 18:42:09	倍的事情还是多操了	0	5
408	A00091838	西地省高速为何同工不同酬？	2019/4/4 2:53:21	工，为什么工资能这样	0	0

经过对高频词进行以上模式的检索，可得出排名靠前的热点留言主题，结合点赞数统计构成用例表：

表 4

留言主题	留言个数	点赞数
A2 区丽发新城附近修建搅拌厂扰民严重	42	15

A 市伊景园车位捆绑销售行为	26	18
魅力之城商业铺噪音油烟扰民	20	18
广铁员工被捆绑车位销售买房	8	9
西堤省高速公路员工编制问题	6	10
A 市地铁 7 号线的建议	6	56
A 市经济学院强制学生实习	8	1
西堤省高速公路员工编制问题	6	10
A3 区梅溪湖看云路油烟扰民	5	5
A3 区郝家坪小学何时扩建？	3	57

由表二可挖掘到信息：留言主题个数多的主题点赞数不一定多，而反映较少的主题却有可能很多人关注到（如 A 市地铁 7 号线的建议、A3 区郝家坪小学何时扩建？）的问题，以及在留言中我们发现上表中广铁员工被捆绑车位销售买房事件与伊景园车位捆绑销售关系密切因而我们将两事件结合看待。

因此我们不能仅仅看留言主题个例的多少来决定热点问题，应该将两者相结合。

4.2.3 建立模型

在通过 excel 筛选后对以上高频词进行查询后，结合点赞数进行构建热度指数方程：

以 S（t）表示留言热度（即为留言主题数），α表示该主题对应的点赞数即为传播系数，点赞数越高网民对该事件政府处理的关注度越高，可得：

$$S'（t）= \alpha S(t)^{[5]}$$

通过计算我们得出留言主题相应的热度指数表（表三）：

表 5

留言主题	热度指数
A2 区丽发新城附近修建搅拌厂扰民严重	31.5
A 市伊景园车位捆绑销售行为	23.4
魅力之城商业铺噪音油烟扰民	20.5
广铁员工被捆绑车位销售买房	4.6
西堤省高速公路员工编制问题	3.2
A 市地铁 7 号线的建议	16.8
A 市经济学院强制学生实习	1.9
西堤省高速公路员工编制问题	3.2

A3 区梅溪湖看云路油烟扰民	1.35
A3 区郝家坪小学何时扩建？	6.34

由此我们可根据热度指数进行对热点问题整理后，进行排列成热点问题值表，并以热点问题整合热点问题留言明细表（见附件热点问题值表、热点问题留言明细表）

4.3、问题三：

我们对问题三中提出的针对附件 4 相关部门对留言的答复意见，进行分析评价，采用 LDA 主题模型对其相关性、完整性、可解释性研究，做出规范化。

4.3.1LDA 主题模型概述[8]：

LDA（Latent Dirichlet Allocation）是一种概率主题模型， 在 2003 年由 Blei, David M 等人提出。 LDA 是一种典型的词袋模型， 即一篇文档由许多组的词组成，词与词之间没有顺序以及先后的关系。 核心思想是先确定文档的多个主题， 然后在主题下选择与主题相关的词汇， 再把词汇组合到一起， 从而形成完整的文档。

LDA 主题模型，是一种三层贝叶斯主题模型，目的在于使用无监督的学习方法发现文本中所隐含的主题信息。

4.3.2 评测过程：

第 1 阶段：准备工作阶段。收集群众留言回复作为样本,然后通过 Python 中的 jieba 库提取留言样本中主题和留言中的字符串（如图 1），通过对词语的使用频率排序,去除无关与无实意的词语，包括使用 jieba 分词工具进行分词、词性标注、去停用词等过程。获取到每个一级分类标签中常用词语，建立对应的数据库分类 ,输出特征属性和训练样本。

第 2 阶段：主题词提取。但是 LDA 主题模型输出的“主题”是隐含的，所以我们要将其进行具体化，即主题词提取。根据“主题-词语”分布，我们将按照概率选择 top M 来代表主题，为减少信息损失，不使用概率最大的单个词语代表主题，而是将 M 定为 4。

第 3 阶段：留言主题和内容与留言回复相似度计算。输入群众留言包括主题即内容和留言得回复，通过上述相似度的计算方法，计算相似度

每个词与文档的相似程度定义如下

$$P(w|D)=\sum_{k=1}^{K} P(w|Z_k)(Z_k|D)^{[9]}$$

其中 D 为文档，w 为主题代表词，Z_k 表示第 k 题。

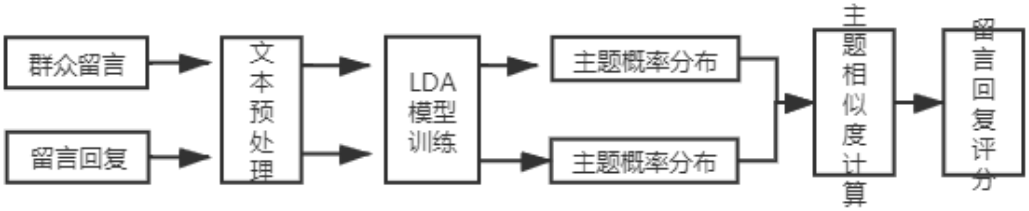
根据公式（1）可计算出所有主题代表词与文档的相似度，根据相似度高低和适当的数量来确定主题词。虽然有的词能进入主题代表词，但是其相似度是非常低的，所以我们要从多个主题代表词中继续筛选获得主题词。首先主题词的相似度应不小于代表词的平均值，且需大于阈值 Q，符合上述标准的便可以评测该留言回复的相关性满足条件。

4.3.3 对留言答复进行评测：

图 4 群众留言回复频率排序

[('物业', 7), ('小区', 6), ('投票', 6), ('公司', 5), ('物业公司', 5), ('业主', 5), ('A2', 3), ('水电', 3), ('公安', 2), ('高昂', 2), ('委会', 2), ('业委会', 2), ('收费', 2), ('制定', 2), ('采用', 2), ('方式', 2), ('这种', 2), ('区景蓉华苑', 1), ('管理', 1), ('物业管理', 1), ('问题', 1), ('2019', 1), ('以来', 1), ('位于', 1), ('桂花', 1), ('街道', 1), ('安分', 1), ('分局', 1), ('公安分局', 1), ('宿舍', 1), ('宿舍区', 1), ('景蓉华苑', 1), ('出现', 1), ('一番', 1), ('乱象', 1), ('美顺', 1), ('扬言', 1), ('退出', 1), ('因为', 1), ('改造', 1), ('造成', 1), ('电费', 1), ('水电费', 1), ('收取', 1), ('不了', 1), ('4.23', 1), ('一吨', 1), ('0.64', 1), ('一度', 1), ('所以', 1), ('通过', 1), ('征收', 1), ('停车', 1), ('车费', 1), ('停车费', 1), ('增加', 1), ('加收', 1), ('收入', 1), ('增加收入', 1), ('不知', 1), ('处于', 1), ('何种', 1), ('理由', 1), ('一再', 1), ('挽留', 1), ('提出', 1), ('应聘', 1), ('以交', 1), ('20', 1), ('保证', 1), ('保证金', 1), ('不能', 1), ('提高', 1), ('苛刻', 1), ('条件', 1), ('之门', 1), ('门外', 1), ('拒之门外', 1), ('召开', 1), ('全体', 1), ('大会', 1), ('业主大会', 1), ('情况', 1), ('方案', 1), ('票箱', 1), ('投票箱', 1), ('表格', 1), ('人员', 1), ('这一', 1), ('利害', 1), ('关系', 1), ('利害关系', 1), ('机构', 1), ('负责', 1), ('组织', 1), ('隐私', 1), ('隐私权', 1), ('没有', 1), ('任何', 1), ('保护', 1), ('反对', 1), ('反对票', 1), ('领导', 1), ('工作', 1), ('要求', 1), ('改变', 1), ('同意', 1), ('何来', 1), ('公平', 1), ('公正', 1), ('公开', 1), ('面对', 1), ('干警', 1), ('公安干警', 1), ('合法', 1), ('合法性', 1)]

图 5 群众留言回复评测流程



五、模型评价与优化

5.1 模型优点：

5.1.1： 利用贝叶斯算法进行群众留言分类,提高了分类的准确性和稳定性。

5.1.2:对于现实生活中政府工资人员对于网友众多的网络留言可以较为简洁进行分类，并根据热度排名优先处理

5.1.3：在评价答复意见时采用了 LDA 主题模型评测，数据更加客观

5.2 模型改进与推广之处：

问题一当中的贝叶斯模型实现计算量复杂，新手操作难度大，因此我们需要在对此说明如何操作，对问题二所建立的模型仍然存在不够简洁的操作，依旧需要一定人工进行排查排序。此后将对数据处理改进，使其智能化

在问题一、二、三所建立的模型中采用了多种公式结合 Notebool 中 jieba 库分词操作，使得所得数据具有信服力，模型客观分析评测答复意见，减轻了政府压力。简化了工作程序，使得政府办公效率加强。

六、参考文献

[1] 朱明 ,王俊普 .一种最优特征集的选择算法 [J].计算机研究与发展 , 2006, 35(9): 803-805

[2] 巩知乐 ,张德贤 ,胡明明.一种改进的支持向量机的文本分类算法 [J].计算机仿真 , 2009, 26(7): 165-168.

[3]荀雪莲,王晓宁.基于中文摘要关键词的毕业论文质量评价系统[J].廊坊师范学院学报(自然科学版),2019,19(04):30-32.

[4] 邓乃扬 ,田英杰 .数据挖掘中的新方法: 支持向量机 [M].北京: 科学出版社 , 2004.

[5]王长宁,陈维勤,许浩.对微博舆情热度监测及预警的指标体系的研究[J].计算机与现代化,2013(01):126-129.

[6] 李明杰, 刘小飞. 基于 Hadoop 框架的图书数据管理系统 [J] . 科学技术创新, 2019, (30) : 72-73.

[7] 李垚周, 李光明. 分布式数据清洗系统设计 [J] .网络安全技术与应用, 2020, (2) : 60-62.

[8]王振振.基于主题模型的中文文本分类相关技术研究[D].北京工业大学,2014.

[9] 基于 LDA 的一体化智能评分系统设计与实现 曹捷 1,李梦瑶 2,陈大卫 湖南师范大学教育科学学院,长沙 410081;2. 湖南师范大学信息科学与工程学院,长沙 410081