基于文本挖掘的“智慧政务”模型

摘要

为解决当前人工处理问政留言所带来的耗费人力大、智能化水平低等缺陷，本文建立基于自然语言处理技术的智慧政务系统。主要解决包括留言分类、热点问题挖掘、答复意见评价三个问题。

针对问题一，首先进行中文数据挖掘的数据清洗与预处理，包括分词和词性、去停用词等，得到初步文本素材。继而采用基于词频统计的特征选取模型筛选分类特征。最后，基于多种分类算法建立不同文本分类模型，包括朴素贝叶斯、逻辑回归、支持向量机等五种分类算法，进行模型性能比较。在测试阶段，提取75%的文本作为训练集训练模型，并预测剩余数据，得到相对应分类模型的混淆矩阵。之后采用F-Score评价算法对分类模型进行评价。经测试，基于朴素贝叶斯分类的模型效果最好，F=0.876。

针对问题二，我们采用基于改进的文档主题生成模型（PLDA）进行分类,之后经多算法测试,得出随机森林算法能较大的提高准确率，并最终得到10个主题的分类结果。另根据热点问题的定义，进行热度评价指标的选取。继而采用因子分析法计算指标方差贡献率，在验证热度指标因子可靠度的结果下给出热度影响力的加权值。最终计算出热点问题的热度指数，并顺序公布出排名前五的热点问题 ，其中A市伊景园滨河苑捆绑销售车位问题热度指数达到95.67位列第一。

针对问题三，首先选取答复的速度，相关度，规范性，完整度为

评价指标，根据附件4中文本信息，归纳出可以用于评价回复质量的信息，进而提取指标。对相关指标进行量化处理，最终建立综合评价模型。根据评价结果，如若政府能在保证有效规范的回答结果前提上提高回答的时间效率，积极调整政务服务工作方案，回复质量将会得到进一步的提升。

关键词：朴素贝叶斯分类;改进的文档主题生成模型（PLDA）；热度评价、自然语言处理

"Intelligent government affairs" model based on text mining

Abstract

In order to solve the defects of high cost of manpower and low level of intelligence caused by manual processing of political messages, this paper establishes an intelligent government system based on natural language processing technology. The main solutions include the classification of comments, hot issues mining, response to comments evaluation of three issues.

Aiming at the first problem， firstly, data cleaning and preprocessing of Chinese data mining were carried out, including word segmentation and part of speech, stop and stop words, etc., and preliminary text materials were obtained. Then a feature selection model based on word frequency statistics is used to screen the classification features. Finally, different text classification models were established based on various classification algorithms, including naive bayes, logistic regression, and support vector machines, to compare the performance of the models. In the test phase, 75% of the text is extracted as the training model of the training set, and the remaining data is predicted to obtain the confusion matrix of the relative classification model. Then the classification model was evaluated by the F-score algorithm.After testing, the model based on

naive bayesian classification has the best effect, F=0.876.

For the second problem, we use the improved document topic generation model (PLDA) to classify. After the multi-algorithm test, it is concluded that the random forest algorithm can greatly improve the accuracy, and finally get the classification results of 10 topics. In addition, according to the definition of hot issues, the heat evaluation index is selected. Then factor analysis was used to calculate the variance contribution rate of the index, and the weighted value of the heat influence was given based on the results of verifying the reliability of the heat index factor. Finally, the heat index of hot issues was calculated, and the top five hot issues were published in order. Among them, binhe garden, yijingyuan, A city, ranked first in the heat index of bundled parking Spaces, which reached 95.67.

For the third problem, the speed, relevance, standardization and integrity of the response were selected as the evaluation indexes. According to the Chinese text in annex 4, the information that can be used to evaluate the quality of reply is summarized, and then the indicators are extracted. The relevant indicators are quantified and the comprehensive evaluation model is finally established. According to the evaluation results, if the government can improve the response time efficiency on the premise of ensuring effective and standardized response results, and actively adjust the work plan of government service, the

quality of response will be further improved.

Key words: Naive bayesian classification algorithm; Improve-Latent Dirichlet Allocation; Evaluation of heat; natural language processing

目录

一、问题重述

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。

1.2 问题重述

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。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

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容

的一级标签分类模型。

通常使用 F-Score 对分类方法进行评价：

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

其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

（2）热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，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。请根据附件 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，以及对应的留言信息。

（3）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。

二、问题分析

2.1 问题一的分析

针对问题一，题目要求我们根据附件 2 建立关于留言内容的一级标签分类模型。

由附件 1 可知，附件 2 里的文本内容全部按照一级分类进行标签，因此，我们只需要对附件 2 中的文本数据进行自然语言处理，进而对

该文本数据建立分类模型。鉴于目前文本分类模型所涉及的分类算法颇多，我们采用几种常用的分类算法分别进行训练，比较准确率，最终采用准确率最高，能够满足实际需要的模型，并对其进行评价。

2.2 问题二的分析

针对问题二，题目要求我们根据附件 3 进行热点问题挖掘。

首先，我们定义热点问题的关键词为：一段时间集中爆发的问题、多人反映同一问题。因此，需要将附件 3 中相似留言进行挖掘，继而将某一段时间内反映特定地点或特定人群问题的留言进行归类。最后，定义合理的热度评价指标以及计算方法，得到评价结果。

2.3 问题三的分析

针对问题三，题目要求我们对附件 4 相关部门的答复意见进行评价。

对答复意见进行评价，首先需要设定评价指标，从答复的相关性、完整性、可解释性等角度考虑，必须要考虑留言内容。根据留言内容与答复意见进行分析对比，从而得出评价结果。

三、文本预处理

文本预处理就是进行数据清洗，对特殊字符进行处理，得到干净的文本素材。然后利用分词工具对句子进行分析，对于无意义的词进行去停用词，从而过滤，得到中心词。继而用数学方法选取最具分类

信息的特征。

3.1 数据清洗

为了去除一些特殊字符及标点符号，使文本数据更加干净，便于后续工作的开展，我们利用 Python 进行编码，进行实现。例如：

表格 1 数据清洗

清洗之前	清洗之后
蔡处长说是坚决不能搞科室承包的。。。	蔡处长说是坚决不能搞科室承包的
引起安全隐患的汽车检测站为何可以施工至今？！	引起安全隐患的汽车检测站为何可以施工至今

3.2 中文分词

为了将中文文本转换为计算机能识别的语言，需要对文本进行向量化。在此之前，便要进行中文分词。我们借助于 Python 的分词工具完成任务。

Jieba 分词是最简单的 Python 分词工具。结巴分词支持三种分词模式：精确模式，可以将句子精确分开，适合文本分析；全模式，将句子中所有可以生成词的词语进行扫描，但不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再切分，适用于搜索引擎分词。

在 Jieba 分词的基础上，我们还可以自定义词典，即使 Jieba 词库中词语有限，我们同样可以根据需要，自行添加新词，以保证更高的使用效率。

我们对分词结果进行示例如下：

表格 2 分词结果

留言主题	留言详情
A2 区远鑫逸园西北边堆放垃圾长期不清理	尊敬的胡书记远鑫逸园西北边堆放垃圾长期不清理影响周边居民的生活质量影响周边的环境卫生虽然通过一些物业管理渠道反映了但是每次反映就清理一点总是不能根除垃圾既对附近居民生活造成影响也影响了市容市貌请督促他们根治此垃圾场建议改建为绿化带美化环境
K 市乡村全科执助可以直接报考执业医师吗	尊敬的卫健委领导您们好我于 2017 年以中专文凭报考取得乡村全科执业助理医师资格证请问满足年限以后可以直接报考执业医师吗还是取得执业助理医师资格证后再报考执业医师呢
M1 区妇幼保健院结核门诊不开门	2019 年 12 月 4 日下午 14.15 到区妇幼保健院看病等了很久始终不见开门请问这样的工作态度随意性太随便了请上级领导部门加以纠正狠刹 2 作作风纪律

3.3 去停用词

我们分析中文分词之后的结果，发现存在很多对于文本分析没有用处的词语，比如“的”、“了”、“既”、“为”、“了”、“于”、“太”等。我们在使用停用词表的基础上，为了更好的满足我们的需要，增加“反映”、“周边”、“通过”、“一些”等大量词。结果示例如下：

表格 3 去停用词

留言主题	留言详情
M　1 区妇幼保健院结核门诊不开门	2019 年 12 月 4 日下午 14.15 妇幼保健院看病等久不开门请问工作态度随意性随便请上级领导部门纠正刹 2 作作风纪律

3.4 特征向量化

3.4.1 向量空间模型

特征表示是指利用一定的特征项来表示文本。文本的表示方法有多种，目前较为常用的是向量空间模型。

向量空间模型是在对中文文本内容实现分词、去停用词等处理之后，用向量来表示文本。[1]文本空间被看作由一组正交词条向量所张成的空间，每一个词条T_i称为一个特征项，每一个文本D则表示空间内的一个向量或空间点，用向量来表示文本：$(W(T_1),W(T_2),\cdots W(T_n))$，其中，$n$为文本空间的维数，当一个文本被表示成文本空间的一个向量时，函数$W(T_i)$表示计算第i个特征项

的权重。

3.4.2 特征选择

中文文本具有高维特性，即特征向量的噪声较多且出现频率不均匀。如此，一方面导致接下来分类算法代价过高，另一方面影响文本类别信息的准确提取。为了解决该问题，采用特征选择[2]的方法进行改善。特征选择是除去不能表示文本主题信息的词，以提高实现效率、减少计算复杂度。

我们采用文献[3]所研究的*CHI*统计量公式计算。特征选择算法的描述：

输入：原始特征集合；

输出：经过特征选择后的特征集合；

具体：

Step1：经过文本数据处理之后，建立自行添加词库、分词、去停用词后所包含的词作为原始特征集合；

Step2：对于每个词，计算词和类别的x^2值：

$$\log(N/N_{ft})\frac{N\times(A\times D-C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \qquad (1)$$

其中，A表示属于某一特定类型c类且包含特定的特征项t的文本频数，B表示不属于c类但包含t的文本频数，C表示属于c类但是不包含t的文本频数，D是既不属于c也不包含t的文本频数，N表示训练集的文本总数。

Step3：对训练集中各个类计算$\sum_{i=1}^{m} p(C_i)x^2(t,C_i)$。对于该类中所有词，根据 Step2 的 *CHI* 统计量进行排序；

Step4：选择一定数量特征词，具体选择多少，根据实验结果确定最佳值。

Step5：将各个类中所有的训练文本，根据特征项，进行向量维数压缩。

四、问题一的求解

4.1 整体构思

基于文本预处理的步骤，我们进行文本分类模型的建立。

首先,根据附件 2 中所有数据,可以清楚地看出,共有城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生共 7 项一级分类标签。

为了更好地完成文本分类，达到精确效果，我们将 7 项一级分类分别进行分类模型预测，验证模型的准确率。

鉴于常用的分类算法种类颇多，因此，为了找出最优模型，我们分别利用贝叶斯分类算法、逻辑回归、支持向量机、随机森林、决策树等算法分别进行模型预测。

最终，比较多种模型的准确率与一致性，找出最优模型，进行评价。使用 F-Score 对分类方法进行评价：

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

其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

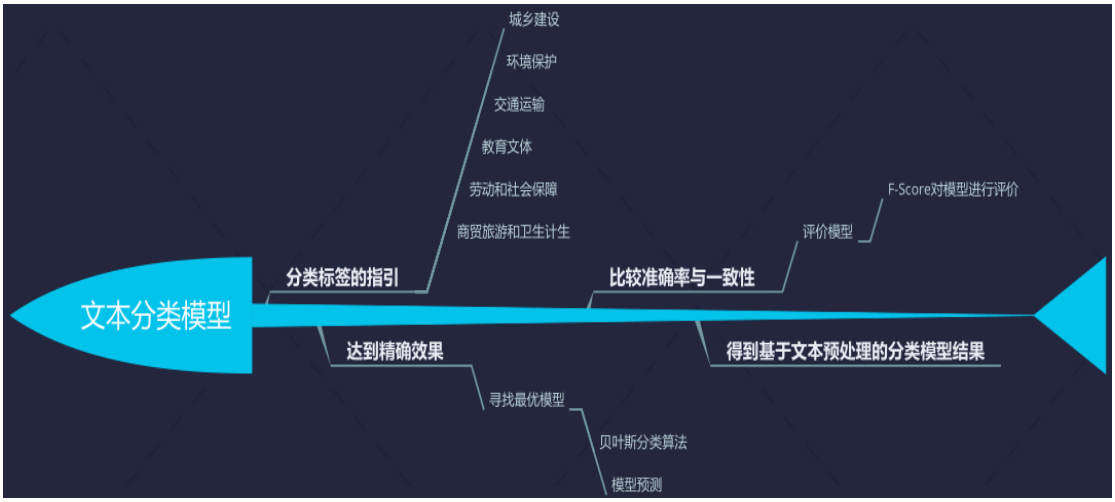

图 1 框架图

4.2 分类算法

基于文本分类的分类方法目前种类繁多，常用的分类方法有：

表格 4 分类方法

贝叶斯分类	决策树分类	支持向量机	神经网络分类
简单向量距离分类	K-最近邻分类	粗糙集分类	逻辑回归算法

根据我们的需要，主要对个别分类算法进行试验。

4.2.1 朴素贝叶斯分类算法

朴素贝叶斯算法是通过训练集样本输入特征 X 和输出结果 Y 的联合分布 $P(X,Y)$ 计算 $P(Y=C_k \mid X=X^{(test)})$，得到测试集最大条件概率对应的类别[4]。

贝叶斯公式为：

$$P(Y_k|X)=\frac{P(X|Y_k)P(Y_k)}{\sum_k P(X|Y=Y_k)P(Y_k)} \quad (2)$$

假设有 m 个样本，n 维特征，k 个类别，联合分布为：

$$P(X,Y=C_k)=P(Y=C_k)P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|Y=C_k)$$

要求得 $P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|Y=C_k)$，贝叶斯模型提出了独立假设：X 的 n 个维度是相互独立的，等价于：

$$P(X_1=x_1|Y=C_k)P(X_2=x_2|Y=C_k)\cdots P(X_n=x_n|Y=C_k)$$

求解使 $P(Y=C_k|X=X^{(test)})$ 最大的类别，根据独立假设，公式如下：

$$C_{result}=\underbrace{\arg\max}_{C_k} P(Y=C_k)\prod_{j=1}^{n}P(X_j=X_j^{(test)}|Y=C_k) \quad (3)$$

为求解公式（3），需要考虑输入变量的先验分布，如下：

①若 X_j 为离散值，假设 X_j 是多项分布式，则：

$$P(X_j=x_{jt}|Y=C_k)=\frac{m_{kjl}+\lambda}{m_k+S_j\lambda}$$

其中，λ 为大于 0 的常熟，O_j 为维度 j 的个数。

②若 X_j 是稀疏二项离散值，则：

$$P(X_j=x_{jt}|Y=C_k)=P(j|Y=C_k)x_{jt}+(1-P(P(j|Y=C_k)(1-x_{jt})$$

③若 X_j 是连续值，假设 X_j 是正态分布，则：

$$P(X_j=X_j|Y=C_k)=\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp(-\frac{(X_j-\mu_k)^2}{2\sigma_k^2})$$

其中，μ_k 为 C_k 类别中，X_j 的均值，σ_k^2 为 C_k 类别中，X_j 的方差。

紧接着，对于测试集样本 $X^{(test)}$，计算

$$P(Y=C_k)\prod_{j=1}^{n}P(X_j=X_j^{(test)}|Y=C_k)$$

对于样本 $X^{(test)}$ 对应的类别，就是使上式值最大对应的类别。

朴素贝叶斯算法稳定，且对于样本量少的数据可以有很好的分类效果，对缺失数据不敏感，但是，朴素贝叶斯有特征独立的假设，对于特征变量之间有相关性的情况，分类效果较差，且通过先验知识和数据计算后验概率，会存在误差。

4.2.2 决策树分类算法

决策树算法是一种归纳分类算法[5]，通过对训练集的学习，挖掘出有用的规则，用于对新集进行预测。在其生成过程中，分割时属性选择度量指标是关键。决策树算法是判断数生成过程的最优路径，决策树的学习过程包括特征选择、树生成及树剪枝。

决策树算法包括 ID3,C4.5 以及 CART 算法。

表格 5 决策树算法对比

决策树算法	支持模型	树结构	特征选择	连续值处理	缺失值	剪枝
ID3	分类问题	多叉树	信息增益	不支持	不支持	不支持
C4.5	分类问题	多叉树	信息增益比	支持	支持	支持
CART	分类、回归问题	二叉树	基尼系数、均方差	支持	支持	支持

决策树算法基本思想：

◆输入：

数据记录 D,包含类标的训练数据集

属性列表 attributeList,候选属性集，用于在内部结点中作判断的属性。

属性选择方法 AttributeSelectionMethod(),选择最佳分类属性的方法。

◆输出：一棵决策树

◆过程：

Step1：构造一个节点 N;

Step2:如果数据记录 D 中的所有记录的类标都相同(记为 C 类)，则将节点 N 作为叶子节点标记为 C，并返回结点 N;

Step3：如果属性列表为空：则将节点 N 作为叶子结点标记为 D 中类标最多的类，并返回结点 N;

Step4：调用 AttributeSelectionMethod(D,attributeList)选择最佳的分裂准则 splitCriterion;

Step5：将节点 N 标记为最佳分裂准则 splitCriterion;

Step6：如果分裂属性取值是离散的，并且允许决策树进行多叉分裂：从属性列表中减去分裂属性，attributeLsit -= splitAttribute;

Step7：对分裂属性的每一个取值 j:

①记 D 中满足 j 的记录集合为 D_j;

②如果 D_j 为空：则新建一个叶子结点 F，标记为 D 中类标最多的类，并且把结点 F 挂在 N 下;

③否则：递归调用 GenerateDecisionTree(Dj,attributeList)得到子树结点 Nj，将 Nj 挂在 N 下；

Step8：返回结点 N；

4.2.3 支持向量机算法

支持向量机是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机。SVM 的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题[6]。

SVM 学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，$\omega x+b=0$即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个(即感知机)，但是几何间隔最大的分离超平面却是唯一的。

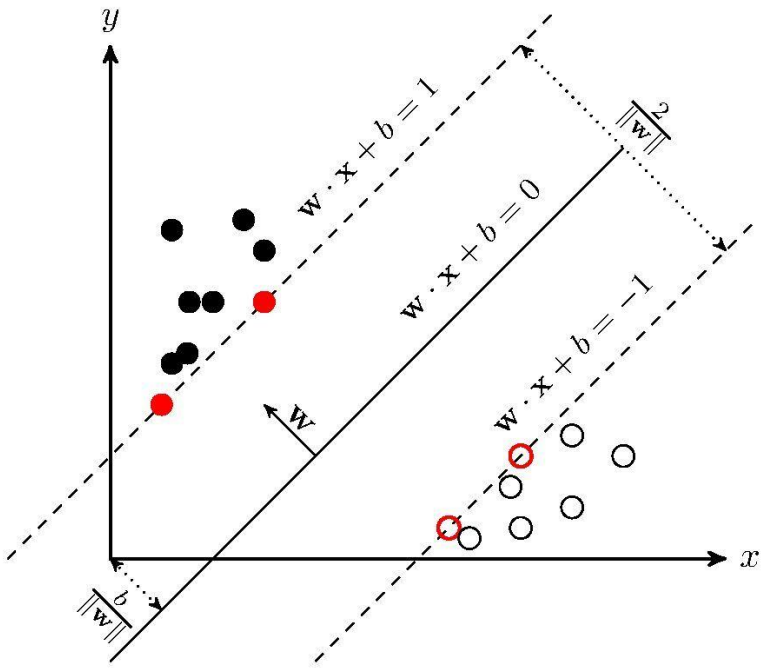

图 2 超平面

基本算法步骤：

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，

其中，$x_i\in R^n$，$y_i\in\{+1,-1\},i=1,2,\cdots,n$。

输出：分离超平面和分类决策函数

(1)选择惩罚参数 $C>0$，构造并求解凸二次规划问题:

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

$$s.t.\quad\begin{cases}\sum_{i=1}^{N}\alpha_i y_i=0\\ 0\le\alpha_i\le C,i=1,2,\cdots,n\end{cases}$$

得到最优解 $\alpha^*=(\alpha_1^*,\ \alpha_2^*,\cdots,\ \alpha_n^*)^T$

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

选择 α^* 的一个分量 α_j^* 满足条件 $0<\alpha_j^*<C$,

计算 $b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\bullet x_j)$

(3）求分离超平面: $\omega^*\bullet x+b^*=0$;

分类决策函数： $f(x)=sign(\omega^*\bullet x+b^*)$

4.3 分类模型预测

4.3.1 基于朴素贝叶斯分类算法的分类模型

运用 Python 的 NLTK 中的贝叶斯类别的库 NaiveBayesClassifier 对文本进行训练，将各一级分类标签按照 3：1 比例随机将数据分为训练集和测试集。附件 2 中文本数据如下：

表格 6 附件 2 各标签数量

标签	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	卫生计生
数量	2009	938	613	1589	1969	1215	877

基于文本预处理的步骤，我们根据该问题的需要，对我们所要分类的标签进行特征提取。将特征提取词作为变量，训练贝叶斯分类模型，并在测试集上进行预测，将测试集上预测结果与实际类别进行对比。

表格 7 预测结果

标签	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	卫生计生
样本数	494	239	143	413	516	298	200
预测准确率	0.817813765	0.966527197	0.776223776	0.888619855	0.914728682	0.802013423	0.875

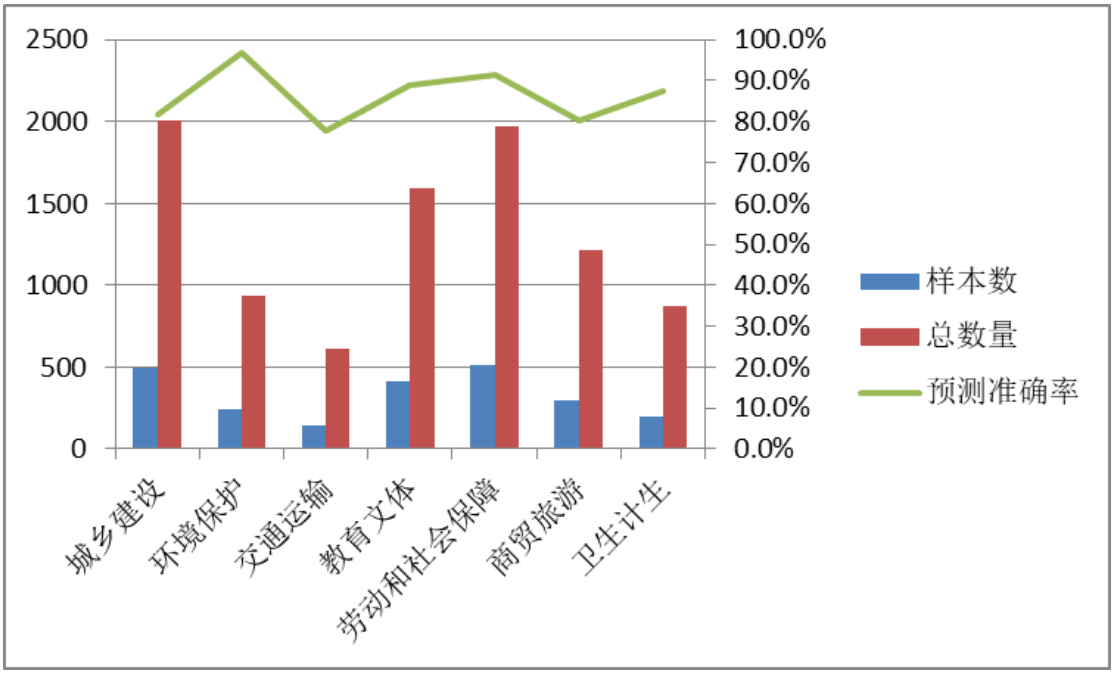

图 3 结果预测图

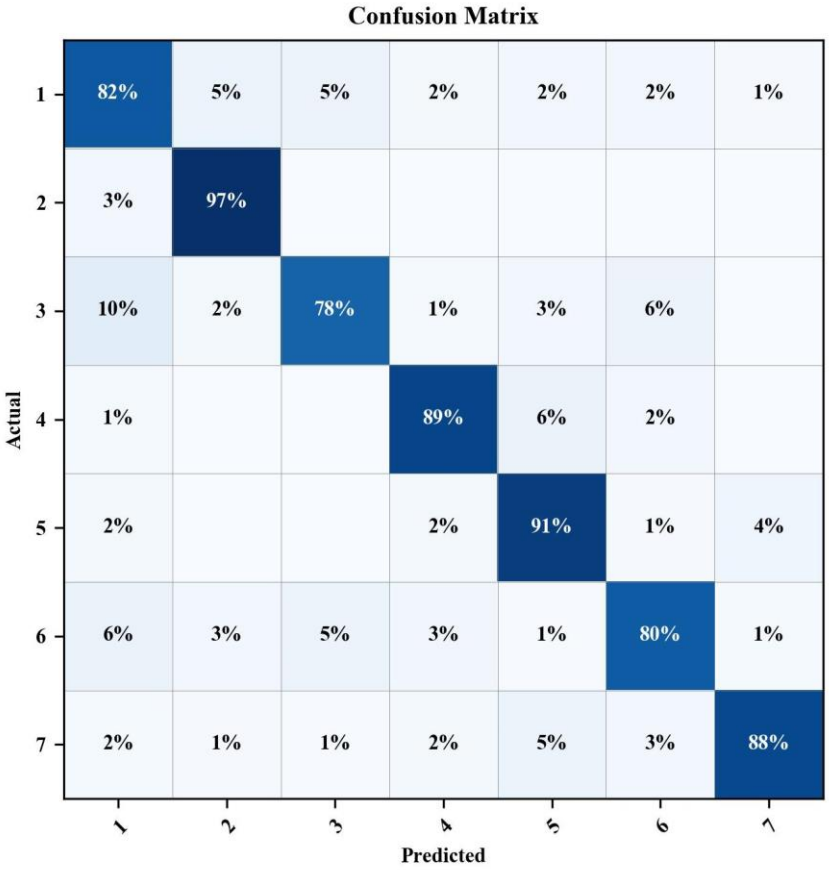

图 4 基于朴素贝叶斯分类模型预测

通过模型训练，我们可以直观的看出环境保护的准确率高达

96.6652%,最终结果中准确率依次:

0.966527197>0.914728682>0.888619855>0.875>0.817813765>0.802013423>0.776223776

总的来说，该模型分类效果平均已经达到了 86.2989528%的效果。

4.3.2 不同算法模型比较

我们利用 Python 根据不同算法对分类模型进行预测，得到在不同分类算法支撑下，模型的准确率如下图所示:

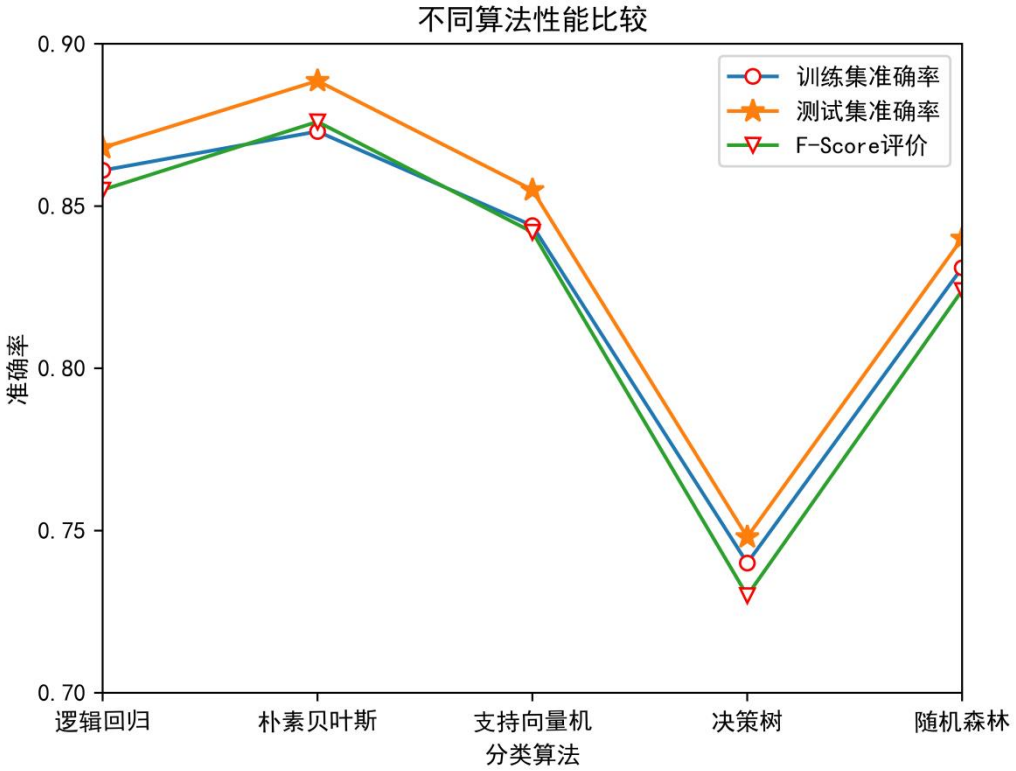

图 5 不同算法比较

我们可以看出朴素贝叶斯算法更适用于该分类模型，而决策树分类算法应用于本模型的准确率低。

4.4 模型评价

经过 F-Score 对分类方法进行评价：

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

得出：

表格 8 评价结果

分类算法	朴素贝叶斯	逻辑回归	支持向量机	随机森林	决策树
F	0.876	0.855	0.842	0.824	0.73

因此，我们最终选择应用基于朴素贝叶斯算法的分类模型，其评价得分最高为：0.876。

利用最优模型对群众留言分类，将留言及时下发至相应政府部门进行处理，提高政务工作的效率。

五、问题二的求解

5.1 整体构思

首先，我们根据题目要求进行热点问题挖掘。因此，需要对热点问题进行定义。题目中“某一时段内群众集中反映的某一问题可称为热点问题”，我们选取关键词为：一段时间集中爆发的问题、多人反映同一问题。

其次，我们基于 PLDA 主题模型进行文本分类。将特定地点或人群的数据归类，即相似留言归为同类。

最后，定义热度评价指标的定义，并基于因子分析方法进行热度

分析，对指标进行排名，得出热点问题，给出结果。

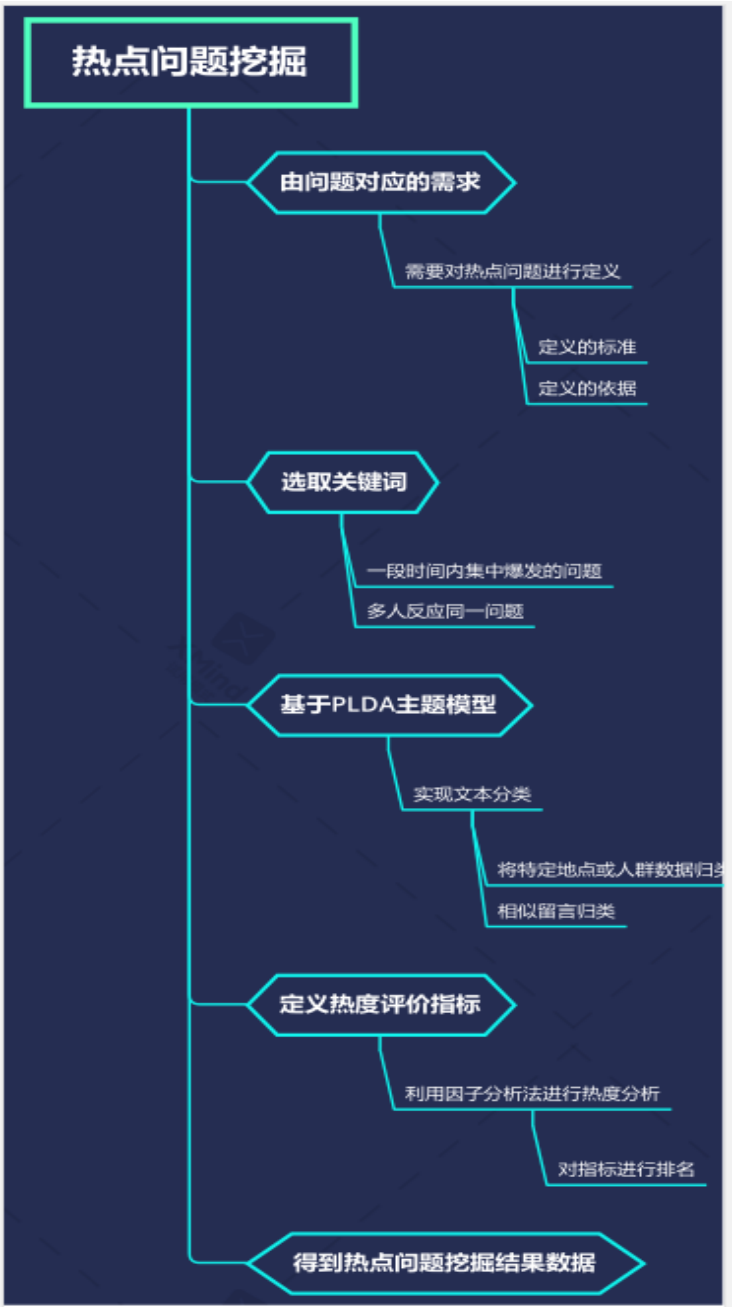

图 6 框架图

5.2PLDA 主题模型

由 3.4 特征向量化可知，文本的高维度会对文本处理工作带来难度。在问题一中，我们采用了 *CHI* 统计量进行维度压缩。为了准确快速的根据热点问题定义提取热点问题，我们采用 PLDA 主题模型[7]

对文本特征进行处理，提高文本处理的效能。

5.2.1PLDA 主题模型理论

PLDA 主题模型基于贝叶斯算法，因此该理论可以总结为：先验分布+数据=后验分布。

假设有 M 个文本，对应的第 d 个文本之中含有 n_d 个词，表示为：

$$\begin{aligned} d\text{oc1} &= W_{11}, W_{12}, \cdots, W_{1n} \\ d\text{oc2} &= W_{21}, W_{22}, \cdots, W_{2n} \\ &\vdots \\ d\text{ocm} &= W_{m1}, W_{m2}, \cdots, W_{mn} \end{aligned}$$

在 PLDA 主题模型中，首先假设一个主题数 K,所有分布都基于 K 个主题展开，PLDA 主题模型假设文本主题和词的先验分布为 Dirichlet 分布，对于某个文本 d，主题分布 $\theta_d = Dirichlet(\vec{\alpha})$，$\alpha$ 是分布的超参数。对于任一主题词 K，词分布 β_k 为：$\beta_k = Dirichlet(\vec{\eta})$，$\eta$ 是分布的超参数，是一个 V 维向量。V 代表词表里词的个数。

PLDA 主题模型中，有 M 个主题的 Dirichlet 分布，文本是 M 个主题的多项分布，这样 $(\alpha \to \theta_d \to \vec{z}_d)$ 就形成了 Dirichlet-muti 共轭，可以使用贝叶斯推断得到基于 Dirichlet 分布文本主题的后验分布。假设在第 d 个文本中，第 k 个主题的词个数为：$n_d^{(k)}$，相对应的多项分布可以表示为元 $\vec{n}_d = (n_d^{(1)}, n_d^{(2)}, \cdots, n_d^{(k)})$,运用 Dirichlet-multi 共轭,得到 θ_d 的后验分布为:$Dirichlet(\theta_d \,|\, \vec{\alpha} + \vec{n}_d)$。对于主题与词的分布，有 K 个主题和词的 Dirichlet 分布，对应的文本有 K 个主题编号的多项分布，则 $(\eta \to \beta_k \to \vec{w}_{(k)})$ (7➞ β ➞)形成了 Dirchletmuti 共轭，使用贝叶斯方法得到 Dirichlet 分布的主题词的后验分布。

在第 k 个主题中，第 v 个词的个数为: $n_k^{(v)}$，则多项分布可以表示为:

$$\overrightarrow{n_k} = (n_k^{(1)}, n_k^{(2)}, \cdots, n_k^{(v)})$$

利用 Dirichlet-multi 共轭，得到 β_k 的后验分布为:

$$Dirichlet(\beta_k | \vec{\eta} + \overrightarrow{n_k})$$

5.2.2 主题模型优点

LDA 是一种非监督机器学习技术[8]，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋的方法，却又不同于传统的词袋模型，可以发掘文本语义之间的关联从而将文本信息转化为了易于建模的数字信息。可将文本的词袋模型进行很好的降维处理且主题模型结果可以很好的进行文本的信息表达，方便进行文本的分类处理。

5.3 基于 PLDA 主题模型的文本分类

5.3.1 主题数的确定

PLDA 主题模型是无监督的算法，需要确定主题数，按照主题模型中的困惑度指标进行主题数的确定，图 7 反映了文本在选择不同主题数的复杂程度。由图 7 可以看出，30 处于复杂度下降突变，因此，我们选择 30 个主题关键词数足以表示留言主题和留言内容文本的语义信息，并且计算复杂度降低，可以简化模型，方便矩阵计算。

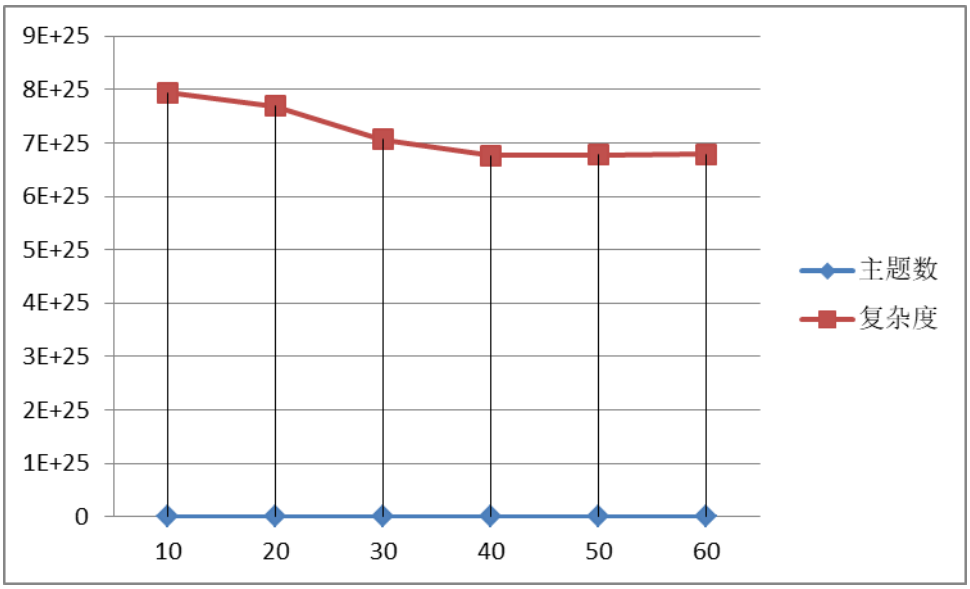

图 7 复杂度随主题数的变化趋势

首先，我们通过 Python 对附件 3 进行基于 PLDA 主题模型的主题确定。结果如下：

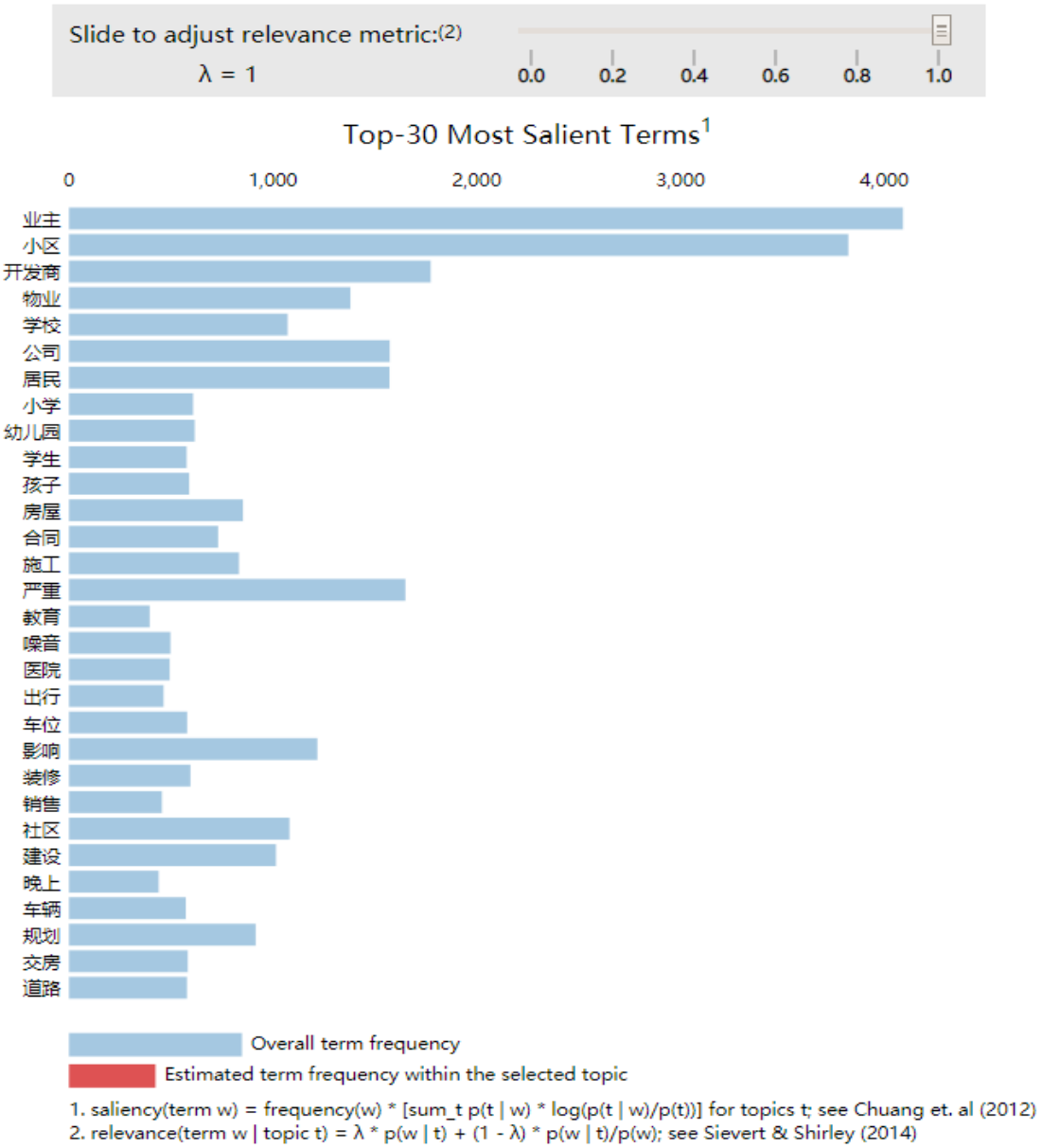

图 8 主题关键词

从图中，我们可以看出部分主题关键词语义明显，能够表达文本主题意义，与类别具有较大的相关性。如：物业、医院、交通、学校、出行、教育、建设词。但是也有部分语义不明显，对于类别归类相关性较小。如：影响、严重、晚上、规划等。

鉴于，存在部分模糊主题词，且高频词汇之间语义会存在交叉，所以，不能够从主题模型高频词中进行直接分类，为了将相似的、关

联性大的主题词进行归并，我们接下来基于 PLDA 主题模型的算法进行自动分类。

5.3.2 分类算法选择

鉴于 5.2 中分类算法的介绍，同样可以利用贝叶斯分类算法、决策树分类算法、支持向量机算法等分类算法进行分类。由于我们基于主题模型的主题作为特征变量，需要训练特征变量。因此 5.2 中所述分类算法在此问题当中效果较差。由问题需要，我们引入随机森林算法[9]。

随机森林算法的执行步骤主要如下：

Step1：创建大量决策树，每棵树之间都不一样，基于观察点和变量的不同子集。

Step2：为每棵树用自助法（bootstrap）来采样观察数据集（用置换法从原始数据采样）。相同的观察点可以在相同的数据集出现多次

Step3：为每棵树随机选择并仅适用一部分变量。

Step4：适用由采样所排除的管擦点来估计每棵树的性能。

Step5：在全部数据被拟合和预测后，获取最终预测，即为回归估计的平均值或用于预测的最频繁类。

6.3.3 分类效果展示

利用 Python 工具中的 Scikit-Learn 包对文本进行训练，同样将文

本数据按照 3：1 的比例分为训练集和测试集。对输入的训练集的 PLDA 主题模型进行训练，主题模型中的主题作为特征变量，训练模型，并在测试集上进行预测。得到如下分类结果：

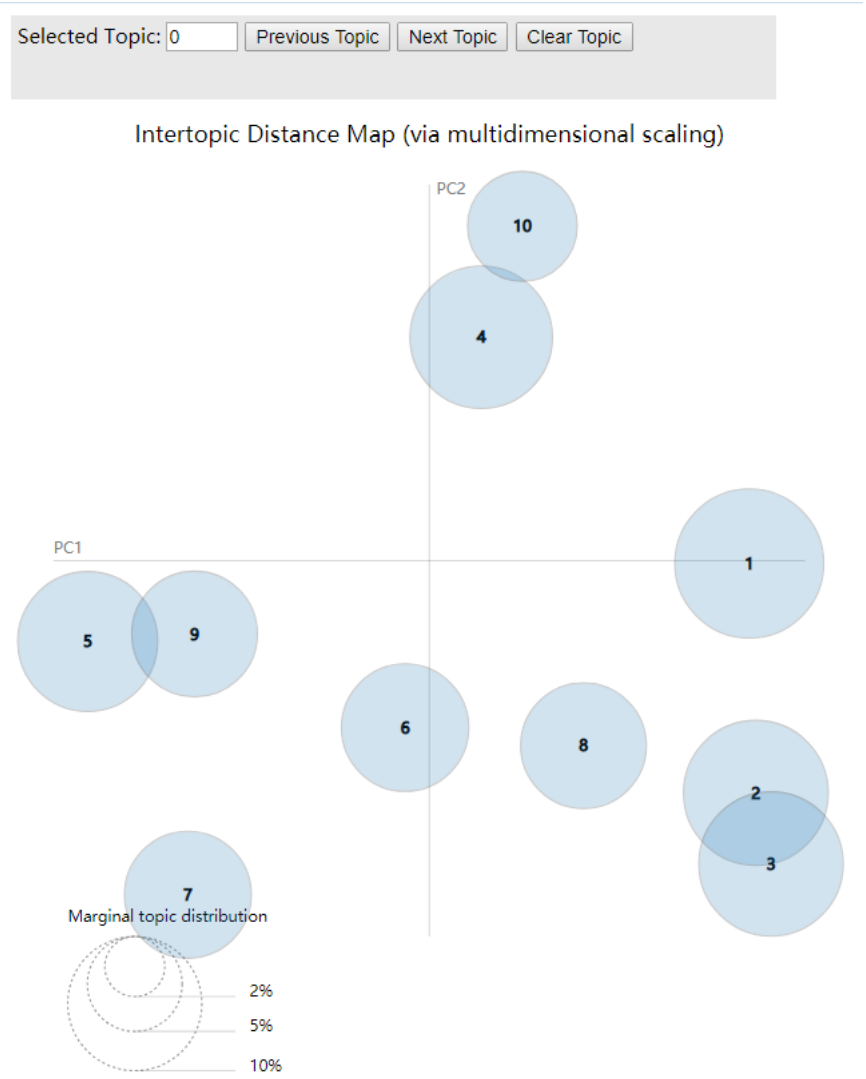

图 9 多维度距离图

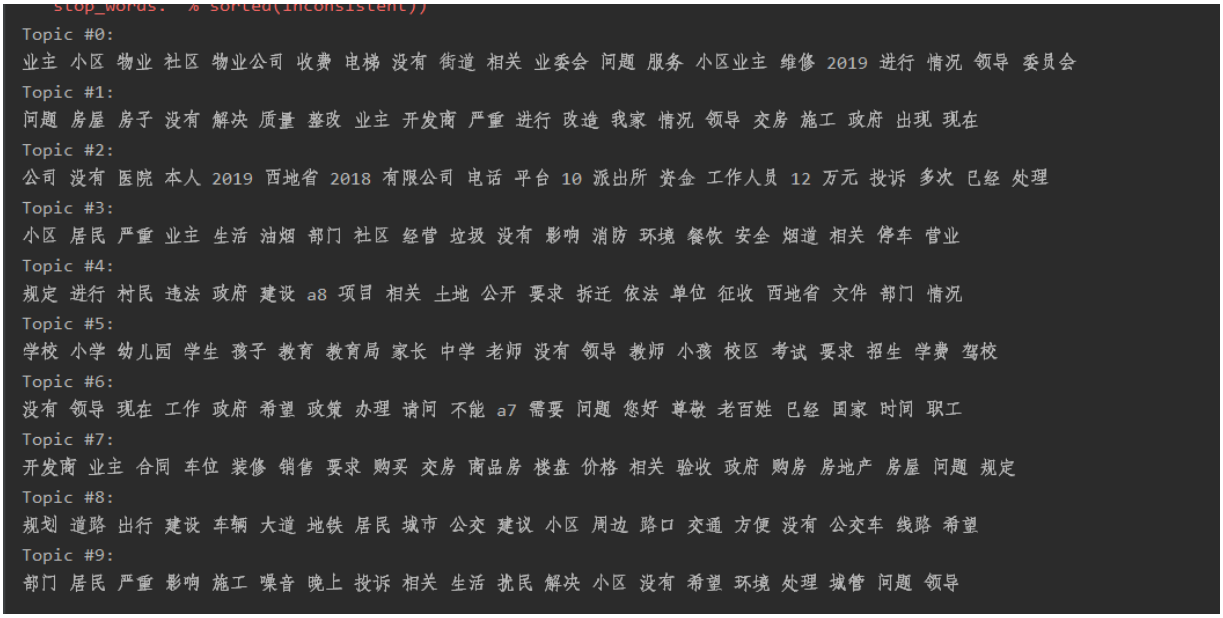

图 10 Topic 10

由运行结果可看出，针对附件 3 的文本数据，基于 PLDA 主题模型的随机森林分类结果为 Topic 0-Topic 9,共十个主题分类。

5.4 热点问题挖掘

5.4.1 热度评价指标

我们通过将某一时段内反映特定地点或特定的人群问题的留言进行归类，定义了如下表所示的热度评价指标集。

表格 9 热度评价指标集

一级指标	二级指标	指标含义
传受众特征热度影响力	反对数	反映问题认可性
	点赞数	
内容特征热度影响力	词量充实度	反映问题待解决迫切性
	集中出现频率	
	出现时长	

5.4.2 因子分析法

在综合考虑模型的使用性之后，我们采用因子分析法来进行热点指标的评价[10]。它通过使用少数几个“抽象”的变量来表示其基本的数据结构。这几个抽象的变量被称作“因子”，能反映原来众多变量的主要信息。

步骤一：主因子的选取

因子分析法对主因子进行选取，将指标元素视为研究对象，并将其设为n个样本，利用p个指标对样本进行观测，其中n大于p，分别将p个指标设为$U_1,U_2,...,U_p$,则因子原始矩阵如下:

$$U=\begin{bmatrix} U_{11} & ... & U_{1p} \\ ... & ... & ... \\ U_{n1} & ... & U_{np} \end{bmatrix}=(U_1,U_2,...,U_p) \qquad (4)$$

其中，U代表因子原始矩阵。

利用因子原始矩阵表示p个指标的对应随机变量，则对应随机变量的协方差与均值分别为:

$$\begin{cases} \sum = Var(U) \\ \mu = E(U) \end{cases},$$

其中,$\sum$代表对应随机变量的协方差;μ代表对应随机变量的均值。对原始矩阵U中p个指标的对应随机变量实施线性变换，得到一个新的综合变量V，则综合变量的具体表达方式为:

$$\begin{cases} V_1=\alpha_{11}U_1+\alpha_{12}U_2+...+\alpha_{1p}U_p \\ V_2=\alpha_{21}U_1+\alpha_{22}U_2+...+\alpha_{2p}U_p \\ ...V_p=\alpha_{p1}U_1+\alpha_{p2}U_2+...+\alpha_{np}U_p \end{cases},$$

其中，α代表综合变量的变化系数；V_1指p个指标的线性组合里

具备最大方差的值；V_2指V_2与没有线性关系的p个指标的全部线性组合里具备最大方差的变量；V_p指与V_1，V_2，…，V_p-1无关的p个指标的所有线性组合里具备最大方差的变量。当综合变量完全满足上述条件时，V_1，V_2，…，V_p即为原始变量的1至p个主成分，由于各个主成分占据的总体方差比例是逐渐缩小的，为了降低评价维数，所以选择方差较大的主成分作为评价的主因子。

步骤二：构建因子分析模型

完成主因子的选取后，需要对主因子进行计算，并根据计算结果构建因子分析模型。首先需要对求解矩阵进行选择，获得更加标准的数据。

根据主因子的约束条件可知第一主因子的求解公式如下:

$$\begin{cases} \max Var(V_i) = Var(Ux_1) = x_1 \sum x_1 \\ x_1' x_1 = 1 \end{cases}. \tag{5}$$

其中，λ代表Σ的特征值，当Σ满秩，即存在p个正特征根，其中最大的记为λ_i，λ_i的对应特征向量为x_i。通过拉格朗日极值法求出评价的第一主因子，第一主因子为:

$$V_1 = Ux_1,$$

同理求出其他主因子。

根据主因子的取值结果与个数对因子分析模型，构建

$$\begin{cases} X_1 = V_{11}F_1 + V_{12}F_2 + ... + V_{1m}F_m + \xi_1 \\ X_2 = V_{21}F_1 + V_{22}F_2 + ... + V_{2m}F_m + \xi_2 \\ \cdots \\ X_p = V_{p1}F_1 + V_{p2}F_2 + ... + V_{pm}F_m + \xi_p \end{cases},$$

其中，$F_1, F_2, ..., F_m$代表实施标准化操作后的公共因子，可以用于

所有原始观测变量;X_p代表向量分量;ξ_p代表向量分量的相应独有因子。

步骤三：实现热度指标评价

利用因子分析模型对热度指标特征进行分析，以实现对信息热度的评价。分析步骤为：首先需要对因子荷载矩阵A进行确定，然后是旋转因子，最后计算因子得分。

下面我们给出具体的方程加以说明：

先考虑两个因子的平面正交旋转。对A按行计算共同度，考虑到各个变量的共同度之间的差异所造成的不平衡，需对A中的元素进行规格化处理，即每行的元素用每行的共同度除之。规格化后的矩阵，为了方便仍记为A，施行方差最大正交旋转（C为正交阵）：

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \cdots & \cdots \\ a_{p1} & a_{p2} \end{bmatrix}$$

$$C=\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

$$B=AC=\begin{bmatrix} a_{11}\cos\phi+a_{12}\sin\phi & -a_{11}\sin\phi+a_{12}\cos\phi \\ \cdots & \cdots \\ a_{p1}\cos\phi+a_{p2}\sin\phi & -a_{p1}\sin\phi+a_{p2}\cos\phi \end{bmatrix}$$

$$=\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ \cdots & \cdots \\ b_{p1} & b_{p2} \end{bmatrix}$$

此时，我们引入因子得分函数，即$F_j=b_{j1}x_1+b_{j2}x_2+...+b_{jp}x_p, j=1,2,...,m$

为了合理的估计因子得分函数，我们采用汤姆森回归法进行估计：

假设公共因子可以对 p 个变量做回归，由于假设变量及公共因子都已经标准化了，所以常数项为 0，即回归方程为：

$$\hat{F}_j = b_{j1}x_1 + b_{j2}x_2 + ... + b_{jp}x_p, j = 1,2,...,m$$

我们现在仅知道由样本值可得因子载荷阵 A，由因子载荷的意义知；

$$\alpha_{ij} = \gamma_{x_i F_j} = E(X_i F_j) = E\left[X_i(b_{j1}X_1 + ... + b_{jp}X_p)\right]$$
$$= b_{j1}\gamma_{i1} + ... + b_{jp}\gamma_{ip} = \left[\gamma_{i1}.....\gamma_{ip}\right]\left[b_{j1}.....b_{jp}\right]^T$$

则，我们有以下的方程组：

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \cdots & \cdots & & \cdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix} \begin{bmatrix} b_{j1} \\ b_{j2} \\ \cdots \\ b_{jp} \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \cdots \\ a_{pj} \end{bmatrix} \quad \text{j} = 1,2,...,m \qquad (6)$$

$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \cdots & \cdots & & \cdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix}$ 为原始变量的相关系数矩阵

$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \cdots \\ a_{pj} \end{bmatrix}$ 为载荷矩阵的第 j 列 $\begin{bmatrix} b_{j1} \\ b_{j2} \\ \cdots \\ b_{jp} \end{bmatrix}$ 为第 j 个因子得分函数的系数，记为 B。

于是 $F = BX$ 就是估计因子得分的计算公式。

5.4.3 主题热度分析

我们基于 5.3 中分类出来的主题结果，结合热度评价指标进行关

键词提取，进而挖掘热点问题。利用 Python 进行，以第一个主题分析为例，

得到分析如图所示：

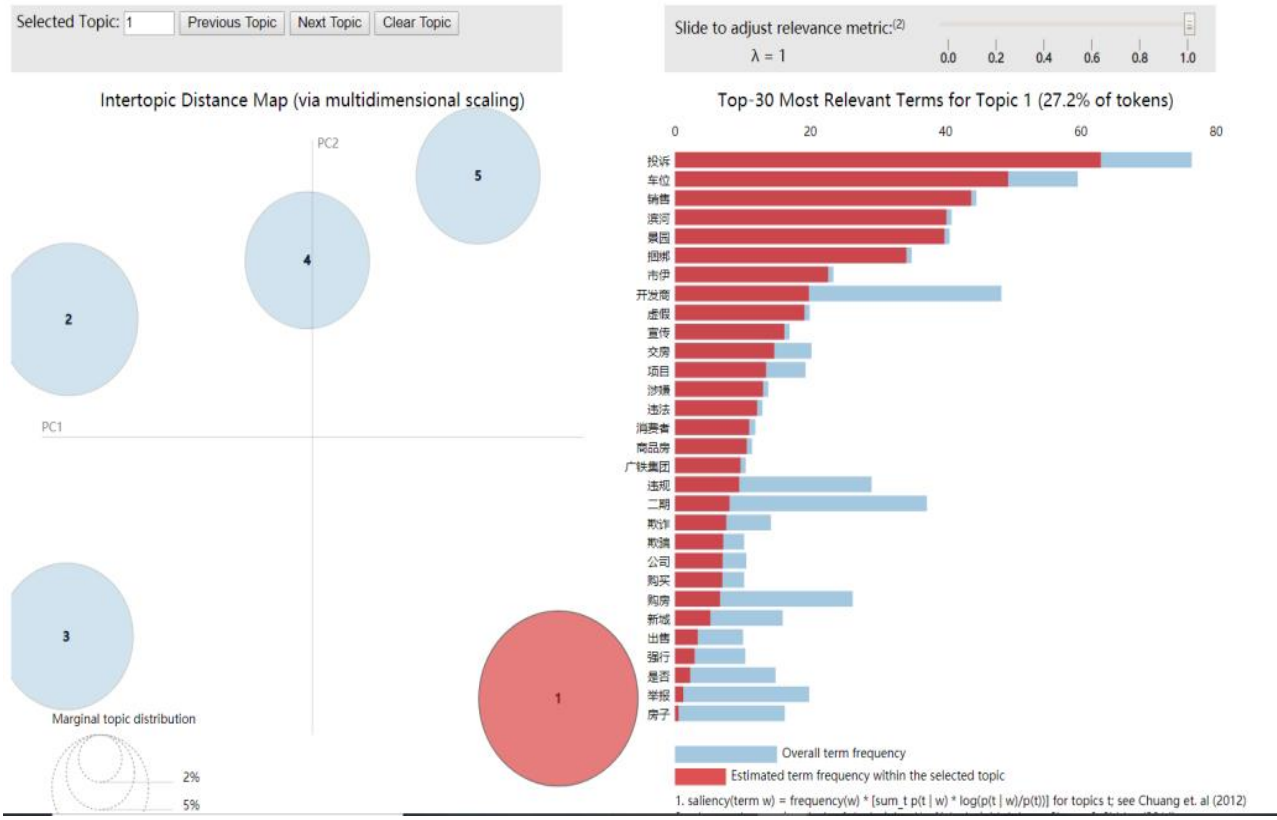

图 11 热点词分析

5.5 结果展示

根据上述的操作，我们就会得到各个热度指标的综合得分情况，通过得分情况对各个热度指标进行先后次序的排名，排名结果即为所求。

以上步骤分析处理，我们最终给出排名前五的热点问题：

表格 10 热点问题

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	95.67	2019/7/7 至 2019/9/1	A 市伊景园滨河苑户主	伊景园滨河苑捆绑销售车位
2	2	93.35	2019/6/13 至 2019/11/22	A3 区幼儿园	A3 区幼儿园上学情况忧心
3	3	92.79	2019/1/3 至 2020/1/7	A 市地铁	A 市地铁优化调整建议
4	4	90.26	2019/7/21 至 2019/09/25	A5 区魅力之城小区	魅力之城小区商贩扰民
5	5	89.21	2018/11/15 至 2019/12/2	A 市人才	A 市人才购房补贴问题

根据热点挖掘出来的结果，即使针对性的完成热点问题处理，提高政务处理工作效率，同时能够极大的满足群众需求，提高人民满意度。

六、问题三的求解

6.1 整体构思

首先，我们定义评价的指标，可以从相关性、完整性、可解释性等角度进行考虑；其次，对指标进行量化处理，便于接下来的计算；最终，建立评价模型，对答复意见进行评价。

6.2 评价指标的建立

6.2.1 指标的选取

针对附件 4，我们可以看出文本可利用内容包括以下 7 项内容：

表格 11 附件 4 内容结构

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
2549	A0004558 1	A2 区景蓉华苑物业管理有问题	2019/4/25 9:32:09	2019 年 4 月以来，位于 A 市...	现将网友在平台《问政西地省》...	2019/5/10 14:56:53

其中，用户信息与答复意见信息分别为：

◆用户信息：留言编号、留言用户、留言主题、留言时间、留言详情；

◆答复意见信息：答复意见、答复时间。

根据以上信息，我们提取重要信息作为评价指标的依据。

用户：

表格 12 用户相关信息

留言主题	留言时间	留言详情
A2 区景蓉华苑物业管理有问题	2019/4/25 9:32:09	2019 年 4 月以来，位于 A 市...

相关部门：

表格 13 相关部门答复信息

答复意见	答复时间
现将网友在平台《问政西地省》...	2019/5/10 14:56:53

基于此，我们可定义评价指标：

表格 14 评价指标

答复速度	答复相关度	答复规范性	答复完整度

6.2.2 指标的量化

答复速度 Δt

对于答复速度的量化，我们可以通过对比用户的留言时间与相关部门的答复时间进行量化，用时间间隔代表回复速度。因为留言时间与答复时间间隔大部分在 10 天以上，因此，我们可以忽略秒的间隔。

例如：

表格 15 答复速度

留言时间	答复时间	答复速度
2019/4/25 9:32:09	2019/5/10 14:56:53	15（天）

其中，时间间隔代表速度，我们利用 Python 进行编码完成。

答复相关度 $\cos\theta$

根据前两个问题的经验，我们已经可以对留言详情进行主题词提取，同样也可用于答复意见当中进行主题词提取。基于此，我们可以进行留言详情与答复意见的相似度计算，用于量化该指标。

我们引进相似度模型：

基于词频-逆向文件频率模型（TF-IDF）的主要思想[11]，我们构建了一个信息相似度模型：

$$TF(\text{词频}) = \frac{\text{某个词在留言中出现次数}}{\text{在答复中次数最多的词语次数}}$$

$$\text{IDF(逆文档频率)} = \log\left(\frac{\text{语料库中主题词总数}}{\text{回复中包含该词的总数}}\right)$$

计算得出：

$$\text{TF-IDF} = \text{TF} * \text{IDF} \qquad (7)$$

首先，把留言详情与答复内容的主题词表示成向量空间模型。为了方便阐述，做以下符号定义：

留言详情D中出现所有主题词的集合标记为$W=(W_1,W_2,...,W_M)$。通过TF-IDF算法，可以得到答复内容d中每个词语TF-IDF值的向量，记做$t=(t_1,t_2,...,t_M)$，其中t_1表示w_1在d中的TF-IDF值。

于是可以将要比较的问题d_1与答复d_2表示为TF-IDF值的向量：

$$d_1=(t_{11},t_{12},...,t_{1M})$$

$$d_2=(t_{21},t_{22},...,t_{2M})$$

最后利用余弦定理计算相似度：

$$\cos\theta=\frac{\sum_{i=1}^{n}t_{1i}*t_{2i}}{\sqrt{\sum_{i=1}^{n}(t_{1i})^2}*\sqrt{\sum_{i=1}^{n}(t_{2i})^2}} \quad (8)$$

当余弦值越接近1时，表明问题d_1与答复d_2越相似。

答复规范性 P

通过对附件 4 以上步骤的处理，我们发现答复意见中存在某种格式：

您好：...反映问题...引起重视...对...回复如下：...经过核查（据查）...得...因为...目前...正在...感谢...理解、监督。

我们根据以上格式提取关键词作为主题词，与答复意见中的文本内容进行比较，计算主题词在答复意见中匹配成功率 P。

答复完整性 N

我们根据答复意见内容的丰富度代表完整性，而丰富度利用答复意见当中的文本词量来表示。基于文本预处理的步骤对答复意见文本进行去停用词后词量统计。

6.3 评价指标模型建立

6.3.1 综合评价模型建立

$$z_i = \sum_{i=1}^{m}\sum_{j=1}^{4} x_{ij} y_j$$
$$\text{其中,} \sum_{j=1}^{4} y_j = 1 \qquad (9)$$

z_i 是第 i 个评价对象的评价结果，x_{ij} 是第 i 个评价对象的第 j 项指标值，y_j 是第 j 个评价指标的权重。

6.3.2 指标无量纲化

由于我们所选取的指标中答复速度用时间间隔来表示，是一个极小型指标，而其他三个指标均是极大型指标，因此，我们需要对指标进行一致化。首先我们利用 Matlab 将评价指标一致化，得到一致化之后的指标。

评价指标的无量纲化方法主要有：

极差化量化法、极小值量化法、极大值量化法、中心化量化方法、均值量化方法。

我们借助于斯皮尔曼等级相关系数法筛选最优量化方法。[12]根据留言用户需求，我们对各项指标赋权，其中：

$$y_1 = 0.2, y_2 = 0.35, y_2 = 0.15, y_2 = 0.3$$

我们选取部分附件 4 的文本数据进行实证，发现均值量化对于此问题是最优方法。

假设以上指标 $X = (x_1, x_2, x_3, x_4)$，则进行指标无量纲化后：

$$x_{ij} = \frac{x_{ij}}{\tilde{x}_{ij}} \tag{10}$$

其中，$\tilde{x}_{ij}$ 是样本平均值。

6.4 评价展示

我们选取附件 4 的部分数据进行答复意见评价，结果如下：

表格 16 评价结果示例

问题 ID	留言编号	留言用户	留言主题	留言时间	答复时间	答复意见	答复时间	评价结果
1	2549	A00045581	A2 区景蓉华苑物业管理有问题	2019/4/25 9:32	2019/5/10 14:56	现将网友在....	2019/5/10 14:56	0.837
1	2759	A00077538	A3 区含浦镇马路卫生很差	2019/4/08 08:37:20	2019/5/9 10:02:08	网友“A00077538”: 您好	2019/5/9 10:02:08	0.801
1	2849	A00010080 4	A3 区教师村小区盼望早日安装电梯	2019/3/29 11:53:23	2019/5/9 10:18:58	网友“A00010080 4”	2019/5/9 10:02:08	0.745

我们仅截取了了热点问题 1 中的几个问题进行评价展示，评价结果满分为 1，越接近 1 代表该回复质量越高。根据评价结果，根据实际需求，可以按照得分区间划分：{优、良、中、差}，以便及时改善

调整政务服务工作方案。

七、参考文献

[1]B.C William, M.TJohn, N-Gram-Based Text Categorization [C]Proceedings of SDAIR 3ndAnnual Symposium on Document Analysis and Information Retrieval, 1994:161-175.

[2]王雅琊，基于朴素贝叶斯和 BP 神经网络的中文文本分类问题研究，云南师范大学硕士论文，2008.

[3]王绪峰,基于支持向量机的中文网页多类分类问题研究及实现[D].昆明:云南师范大学，2007.

[4]https://blog.csdn.net/u014741673/article/details/62886975/ 分类算法之朴素贝叶斯分类

[5]https://blog.csdn.net/qq_41587243/article/details/88314403/ 决策树分类算法三种方式

[6] https://zhuanlan.zhihu.com/p/31886934/支持向量机 SVM 篇

[7]廉素洁,基于文本分类和情感评分的电信投诉文本挖掘研究,z浙江工商大学硕士论文，2018.

[8] https://www.jianshu.com/p/39a8372d7ccd/ LDA 文档主题生成模型入门

[9] https://blog.csdn.net/java_fresh_man/article/details/84862039/随机森林详解

[10]谭晶，基于文本挖掘的电子商务网站评价指标体系建立,《知

识经济》，2018.

[11]王海明,基于 TF-IDF 改进计算模型的实时大数据处理系统设计与实现[D].

[12]张卫华，赵铭军指标无量纲化方法、对综合评价结果可靠性的影响及其实证分析[J],《统计与信息论坛》，2005.