“智慧政务”中的文本挖掘应用

摘要：

目前，大部分电子政务系统主要依靠人工进行事务处理，存在工作量大、效率低，且差错率高等问题，因此难以高效解决民众的诉求。随着互联网技术的发展，运用自然语言处理技术（NLP）建立智慧的政务系统已成为社会治理创新发展的新趋势，其对于提升政府的管理水平和施政效率具有极大的推动作用。针对上述问题，本文展开了如下研究：

对于问题一，群众留言分类问题。通过对数据可视化，我们发现该问题属于不均衡文本数据的多分类问题。首先运用 JIEBA 分词工具对中文进行分词，并标注了词性，应用 word2vec 作为词向量计算工具，通过 TF-IDF 值作为文本特征选择；然后在分类器的选择问题上，通过比较不同的分类模型，选取了最优的基于支持向量机的模型；最后运用 F1-score 作为评价指标，绘制了混淆矩阵的热力图，得出了正确率高于 80.6%的模型。

对于问题二，热点问题挖掘。给出基于文本聚类与 LDA 模型的热度评价方法。首先运用 JIEBA 分词工具对中文进行分词，并标注词性，应用 word2vec 作为词向量计算工具对文本进行聚类；然后通过 LDA 模型对于每个簇进行建模和主题提取，并从词频、词跨度、词长三方面计算每个主题中关键词的权值，利用 Gibbs Sampling 抽样计算模型参数，从而获取留言热点及相应的热点词语。

对于问题三，答复意见的质量评价。设计了一套基于多变量的质量评价规则。首先定义衡量质量好坏的三大指标：相关性、完整性和可解释性；然后运用 Excel API 函数库中的 GetMatchingDegree 函数，计算答复与留言的相关度百分比判断相似性，根据答复文本能提取的主题数判断其完整性，依据答复意见句子语言层的嵌套情况判定其可解释性；最后定义了一个评价函数，综合相似性、完整性、可解释性三大变量来评价答复意见的质量。

关键词：文本分类、热点问题挖掘、支持向量机、K-Means 聚类、LDA 模型

Abstract

At present, most e-government systems mainly rely on manual transaction processing, which has the problems of heavy workload, low efficiency, and high error rate. Therefore, it is difficult to efficiently solve the demands of the people. With the development of Internet technology, the use of natural language processing technology (NLP) to establish a smart government system has become a new trend in the development of social governance innovation, which has greatly promoted the government's management level and governance efficiency. In response to the above problems, this article has carried out the following research:

For question one, the question of the classification of the masses. By visualizing the data, we found that the problem belongs to the multi-classification problem of unbalanced text data. First, use the JIEBA word segmentation tool to segment the Chinese word, and mark the part of speech, use word2vec as a word vector calculation tool, and select the text feature through the TF-IDF value; then on the selection of the classifier, by comparing different classification models The optimal model based on support vector machine was developed. Finally, F1-score was used as the evaluation index, and the thermal map of the confusion matrix was drawn, and the model with the accuracy rate higher than 80.6% was obtained.

For problem two, hot spot mining. A heat evaluation method based on text clustering and LDA model is given. First, use the JIEBA word segmentation tool to segment the Chinese word and mark the part of speech, and use word2vec as a word vector calculation tool to cluster the text; then use the LDA model to model and extract each cluster, and from the word frequency, word span, word In three aspects, the weights of the keywords in each topic are calculated, and Gibbs Sampling is used to calculate the model parameters, so as to obtain the message hotspot and the corresponding hot words.

For question three, the quality evaluation of the comments. A set of multi-variable quality evaluation rules is designed. First define the three major indicators

that measure quality: relevance, completeness, and interpretability; then use the GetMatchingDegree function in the Excel API function library to calculate the percentage of relevance between the reply and the message to determine the similarity. The number of topics judges its completeness, and its interpretability is judged according to the nested situation of the language level of the reply opinion sentence. Finally, an evaluation function is defined, which combines the three variables of similarity, completeness and interpretability to evaluate the quality of the reply opinion.

Keywords: Text Classification, Hotspot Problem Mining, Support Vector Machine, K-Means Clustering, LDA Model

目 录

1.挖掘目标

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越来越大，且其本身不具有结构化特征，给以往主要依靠人工这种效率相对较低的方式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，随着现代互联网技术的发展，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，对网上的留言，通过分类、聚类等算法对其进行数据挖掘可以在很大程度上减少相关部门工作人员的工作量，从而对提升政府的管理水平和施政效率具有极大的帮助。

2.总体流程

因为各类社情民意相关文本的数据量巨大及其非结构化特征，普通的文本数据处理工具实现起来效果并不理想，所以本文使用 Python 语言配合 PyCharm 平台上 Anaconda 的环境对现有问题进行编程实现。本文的用例可分为三个步骤：

1) 步骤一：数据预处理，对网络上的留言进行预处理，去除数据中的噪声以及脏数据，保证后续步骤不受数据源本身的影响，为模型提供可靠的数据；

2) 步骤二：模型建立，对处理后的模型进行特征提取，特征分类等操作，建立留言数据的模型；

3) 步骤三：模型应用，将建立的模型，应用到实际数据处理中，实现对留言数据的挖掘。

本文的总体架构及思路如图 1 所示。

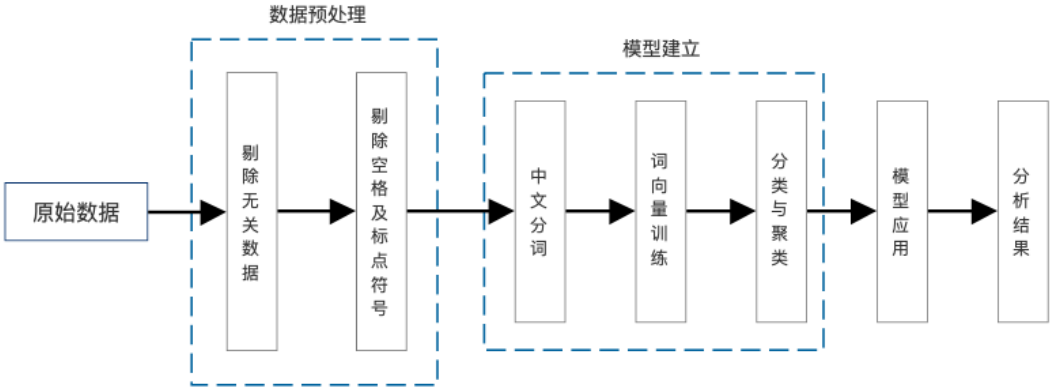

图 1 总流程图

3.问题解决

3.1.群众留言分类

在处理网络问政平台的群众留言时，工作人员需要按照一定的划分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进行处理。

我们选择采用附件 1 的一级标签分类，将附件 2 的 9210 条留言数据按照城乡建设、劳动和社会保障、教育问题、商贸旅游、环境保护、卫生计生、交通运输划分成七个大类，且每条留言只能属于七类中的一类。这是一个文本的多分类问题。

目前，中文文本的自动分类模型包括文本预处理、特征抽取、特征选择、利用分类算法提取分类模型、对分类模型进行质量评估五个方面。如图 1 所示。

3.1.1.数据探索

首先我们查看一下每个类别数据的条数，如图 2 所示。

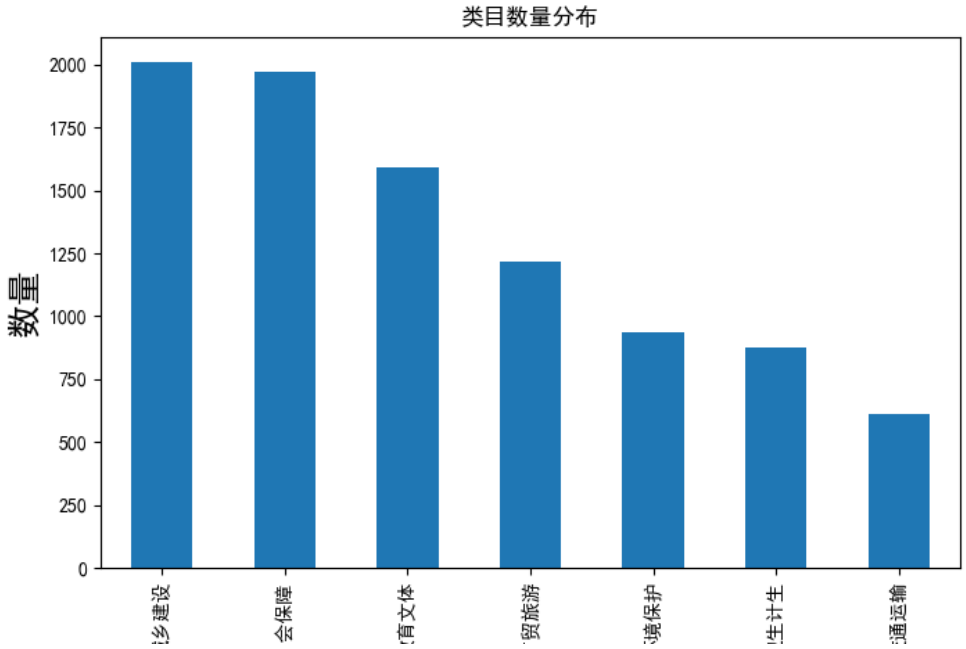

图 2 类目数量分布图

根据各个类别分布的图形化结果，我们看到各个类别的数据量不一致，城乡建设、劳动和社会保障、教育文体和商贸旅游五大类的数据量都在 1 千条以上，环境保护、卫生计生和交通运输三大类的数据量在 600-1000 条，存在一定的数据不均衡问题。

3.1.2.数据预处理

虽然现下文本分类的研究已经非常成熟，各种分类器都显示出了各自不同的性能，然而在此之前不可忽略的是，文本的预处理也是影响文本分类精确度的关键因素之一。其中中文分词、去除停用词都是预处理中的关键部分。

3.1.2.1.数据清洗

首先，我们抽取留言数据中的两个字段，其中 cat 字段表示类别，review 字段表示用户的留言主题信息。数据总量为 9210 条，留言内容全部为中文，且不存在空数据，因此不需要进行空值清洗。

考虑到以中文来命名各类型的名字会在后续分类模型的训练时引起不方便，我们将 cat 的七个类名分别转换成由数字序列 0-6 对应的 cat_id，如表 1 所示。

表 1　cat-cat_id 表

留言主题	编号
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

3.1.2.2.分词及去除停用词

其次，我们需要删除文本中除中文、字母、数字以外的所有标点符号、特殊符号，删除一些无意义的常用词(stopword)。这些词和符号对系统分析预测文本的内容没有任何帮助，反而会增加计算的复杂度和增加系统开销，因此在使用这些文本数据之前必须要将它们清理干净。

停用词表中包含了中文日常使用频率很高的常用词，如：吧，吗，呢，啥等感叹词，这些高频常用词无法反应出文本的主要意思，所以要被过滤掉。除此之外，我们还向停用词表中加入了留言信息中高频出现但对类别划分无太大实际帮助的词汇，如国际、街道、投诉、社区、解决等。

去除停用词后生成了新的字段 clean_review。最后，我们要在 clean_review 的基础上进行分词，使用 Jieba 库把每个评论内容分成由空格隔开的一个一个单独的词语，得到新的字段 cut_review，如表 2 所示。

表 2　预处理结果

一级标签	去除停用词后留言主题	分词结果
城乡建设	A 市西湖建筑集团占道施工有安全隐患	A 市 西湖 建筑 集团 占道 安全隐患
城乡建设	A 市在水一方大厦人为烂尾多年，安全隐患严重	A 市 在水一方 大厦 人为烂尾 多年 安全隐患
城乡建设	投诉 A 市 A1 区苑物业违规收停车费	A 市 区苑 物业 收 停车费
城乡建设	A1 区蔡锷南路 A2 区华庭楼顶水箱长年不洗	区 蔡锷 南路 区华庭 楼顶 水箱 长年 不洗
城乡建设	A1 区 A2 区华庭自来水好大一股霉味	区 区华庭 自来水 好大一股 霉味

3.1.2.3.文本特征选择

目前，已经提出的文本分类特征选择方法比较多，常用的方法有：文档频率（Document Frequency，DF）、信息增益（Information Gain，IG）、卡方（χ^2）检验（CHI）和互信息（Mutual Information，MI）等方法。

我们采用经典的 TF-IDF 方法进行加权处理。

TF-IDF（term frequency–inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。TF 意思是词频(Term Frequency)，IDF 意思是逆文本频率指数(Inverse Document Frequency)。TF-IDF 是在单词计数的基础上，降低了常用高频词的权重，增加罕见词的权重。因为罕见词更能表达文章的主题思想，比如在用户留言中出现的“A 市”和“在水一方”两个词,那么后者将更能体现文章的主题思想，而前者是常见的高频词，它不能表达文章的主题思想。所以“在水一方”的 TF-IDF 值要高于“A 市”的 TF-IDF 值。

我们调用机器学习的 sklearn 库，使用

sklearn.feature_extraction.text.TfidfVectorizer 方法来抽取文本的 TF-IDF 的特征值。其中我们使用了参数 ngram_range=(1,2)，这表示我们除了抽取评论中的每个词语外，还要抽取每个词相邻的词并组成一个“词语对”，例如：词 1，词 2，词 3，词 4，（词 1，词 2），（词 2，词 3），（词 3，词 4）。这样就扩展了我们特征集的数量，有了丰富的特征集才有可能提高我们分类文本的准确度。如图 3 所示。

```
(9210, 53908)
-----------------------------
  (0, 52266)    0.4426325236876198
  (0, 27818)    0.4426325236876198
  (0, 47006)    0.4426325236876198
  (0, 22616)    0.2916226513877324
  (0, 52256)    0.302091175500458
  (0, 27796)    0.29761980758687295
  (0, 47005)    0.3838605839834035
  (1, 20936)    0.33061857028391556
  (1, 39078)    0.33061857028391556
  (1, 8740) 0.33061857028391556
  (1, 21076)    0.33061857028391556
  (1, 19676)    0.33061857028391556
  (1, 20925)    0.2303222153137964
  (1, 39076)    0.28671963914361454
  (1, 8737) 0.3063295219046831
```

图 3 特征集

我们得到 features 的维度是(9210，53908)，其中 9210 表示我们总共有 9210 条留言数据，53908 表示我们的特征数量这包括全部评论中的所有词语数+词语对（相邻两个单词的组合）的总数。下面我们要使用卡方检验的方法来找出每个分类中关联度最大的两个词语和两个词语对。

卡方检验（Chi-square test/Chi-Square Goodness-of-Fit Test），是一种用途很广的计数资料的假设检验方法，主要用于比较两个及两个以上样本率（构成比）以及两个分类变量的关联性分析。其根本思想就是在于比较理论频数和实际频数的吻合程度或拟合优度问题。

```
# '交通运输':
  . Most correlated unigrams:
       . 的士
       . 出租车
  . Most correlated bigrams:
       . 出租车 管理
       . 滴滴 出行
```

图 4 卡方检验

如图 4 所示，经过卡方检验后，我们找出了每个分类中关联度最强的两个词和两个词语对。这些词和词语对能很好的反映出分类的主题。

3.1.3.分类器的选择

我们尝试不同的机器学习模型，并评估它们的准确率。

（1） 随机森林（Random Forest）

随机森林是一种集成算法（Ensemble Learning），它属于 Bagging 类型，通过组合多个弱分类器，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度和泛化性能。其可以取得不错成绩，主要归功于“随机”和“森林”，一个使它具有抗过拟合能力，一个使它更加精准。

（2） 线性支持向量机（Linear Support Vector Machine）

线性可分 SVM 模型输入是线性可分的 m 个样本(x1,y1),(x2,y2),...,(xm,ym),(x1,y1),(x2,y2),...,(xm,ym)，其中 x 为 n 维特征向量，y 为二元输出，值为 1 或者-1。输出是分离超平面的参数 w*和 b*w*和 b*和分类决策函数。

算法过程如下：

1. 构造约束优化问题

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

$$s.t. \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\alpha_i \geq 0 \;\; i = 1,2,\dots m$$

2. 用 SMO 算法求出上式最小时对应的α向量的值α^*向量
3. 计算$\omega^* = \sum_{i=1}^{m} \alpha_i^* y_i x_i$
4. 找出所有的 S 个支持向量，即满足$\alpha_s > 0$对应的样本(x_s, y_s)，通过$y_s(\sum_{i=1}^{m} \alpha_i x_i x_i^T x_s + b) = 1$，计算出每个支持向量$(x_s, y_s)$对应的$b_s^*$，计算出最终的$b^*$

这样最终的分类超平面为

$$\omega^* \cdot x + b^* = 0,$$

最终的分类决策函数为

$$f(x) = sign(\omega^* \cdot x + b^*) \tag{2}$$

（3） 多项式朴素贝叶斯（Multinomial Naive Bayes）

朴素贝叶斯法利用贝叶斯定理首先求出联合概率分布，再求出条件概率分布。这里的朴素是指在计算似然估计时假定了条件独立。基本原理可以用下面的公式给出：

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \tag{3}$$

其中，$P(X|Y) = P(X_1, X_2, \dots X_n|Y) = P(X_1|Y)P(X_2|Y) \dots P(X_n|Y)$，$P(Y|X)$叫做后验概率，$P(Y)$叫做先验概率，$P(X|Y)$叫做似然概率，$P(X)$叫做证据。

而多项式朴素贝叶斯的原理由下面的公式给出，其中条件概率$P(\omega_i|c)$表示的是词 ω_i 在类别 c 中的权重

先验概率

$$P(C = c) = \frac{\textit{属于类c的文档数}}{\textit{训练集文档总数}} \tag{4}$$

条件概率

$$P(\omega_i|c) = \frac{\textit{词}\omega_i\textit{在属于类c的所有文档中出现次数}}{\textit{属于类c的所有文档中的词语总数}} \tag{5}$$

朴素贝叶斯分类器最适合用于基于词频的高维数据分类器，最典型的应用如垃圾邮件分类器等。

（4） 逻辑回归（Logistic Regression）

我们知道 Logistic 回归本身只能解决二分类问题，不过一般所有的二分类算

法都可以通过改进使之支持多分类，改进办法一般有 OvR（One vs Rest）和 OvO（One vs One）两种。sklearn 库中单独封装了 OvR 和 OvO，这样一来对于我们书写的二分类代码，在符合 sklearn 的调用逻辑的基础上，就可以通过 OvR 和 OvO 的方式应用到多分类任务中去了。

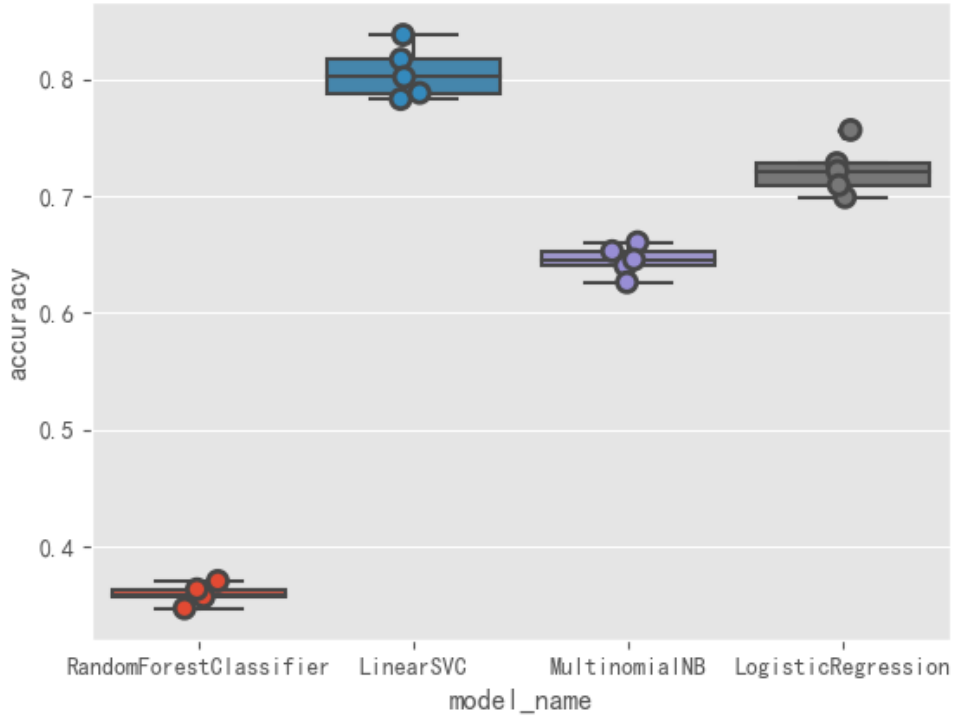

图 5　准确度箱体图

从图 5 可以看出随机森林分类器的准确率是最低的，这是因为随机森林属于由若干个子分类器组合而成的集成分类器，一般来说集成分类器不适合处理高维数据（如文本数据），因为文本数据有太多的特征值，会使得集成分类器难以应付。而另外三个分类器的平均准确率都在 80%以上，其中线性支持向量机的准确率最高，如表 3 所示。

表 3　模型准确率

model_name	accuracy
LinearSVC	0.805753
LogisticRegression	0.723020
MultinomialNB	0.645159
RandomForestClassifier	0.359287

我们看到线性支持向量机的平均准确率达到了 80.6%，其次是逻辑回归和朴素贝叶斯。

3.1.4.模型评估

我们采用平均准确率最高的 LinearSVC 模型，并针对其查看**混淆矩阵**，同时显示预测标签和实际标签之间的差异，如图 6 所示。

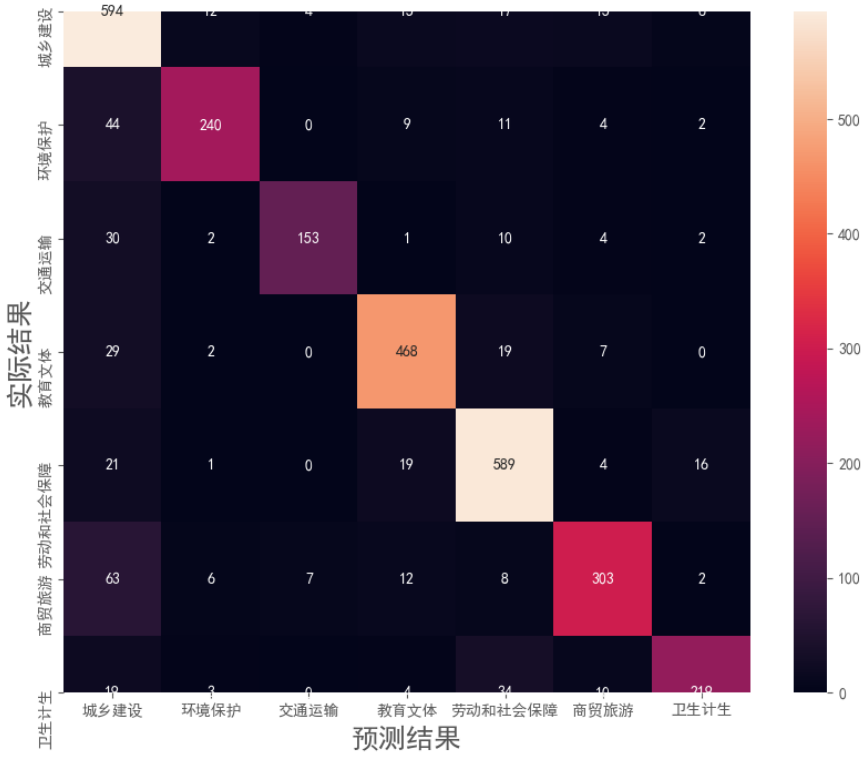

图 6　混淆矩阵

混淆矩阵的主对角线表示预测正确的数量,除主对角线外其余都是预测错误的数量。从上面的混淆矩阵可以看出，"教育文体"类预测最准确，“城乡建设”和“商贸旅游”类预测的错误数量教多。

由于当训练数据不平衡（有的类数据很多,有的类数据很少）时，准确率（accuracy）不能反映出模型的实际预测精度，因此多分类模型一般不使用 accuracy 来评估模型的质量。我们使用 F-Score 对分类的方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{6}$$

其中P_i为第i类的查准率，R_i为第i类的查全率。下面我们将查看各个类的 F1 分数，如表 4 所示。

表 4　F1-Score 表

	precision	recall	f1-score	support
城乡建设	0.74	0.90	0.81	663
环境保护	0.90	0.77	0.83	310
交通运输	0.93	0.76	0.84	202
教育文体	0.89	0.89	0.89	525
劳动和社会保障	0.86	0.91	0.88	650
商贸旅游	0.87	0.76	0.81	401
卫生计生	0.89	0.76	0.82	289
accuracy			0.84	3040
macro avg	0.87	0.82	0.84	3040
weighted avg	0.85	0.84	0.84	3040

从以上 F1 分数上看，“教育文体”类的 F1 分数最大，“城乡建设”和“商贸旅游”类的 F1 分数最低，为 81%。

3.2.热点问题挖掘

与新闻报道相比较, 投诉文本的结构更加复杂且短小, 这加大了提取话题的难度。本文针对群众留言投诉文本，应用“热点问题识别”的相关知识，从中识别投诉文本中的热点话题。

3.2.1.热度指标体系构建

热点问题的本质是群众集中反映的问题和某一问题被反映出得到群众的认可和反对程度，如何定量的来描述投诉问题的热度，目前学术界还没有标准的指标体系。参考武汉大学张敏对于电子政务研究热点的文章，结合我们具体问题，我们提出了三个方面对于热点问题的评价：关键词抽取与词频统计，反对数，点

赞数。

（1）关键词抽取与词频统计

关键词能高度概括文章主题，对于热点问题的反应有着重要的影响，统计一个词语在投诉问题出现的次数，可以反映出困扰很多群众的问题。

（2）反对数和点赞数

可以看出群众对于某个问题的讨论激烈程度，和对于这个问题是否自己也有同样的困扰或者对于提出的问题有着不同的意见。

3.2.2.K-Means 聚类

由于群众留言跟新闻报道不一样，它的形式简短，单条涵盖的内容信息很少，而且语义很难分析。为了更好的提取话题，首先将文本进行聚类，这样每一类中的投诉文本不仅存在着共性，而且内容比较充实，LDA 模型抽取话题表达效果就会更好，针对性更强。

本文采用 k-means 进行聚类，k-means 是经典划分聚类算法。这种方法简单快速，在对文档进行聚类,前需要通过 k 值来确定簇数量。主要过程是从含文本的文档集中随机选择 k 个文本作为初始的聚类中心，并通过计算得到其他文本到每个簇中心点的距离,将文档划分到离它最近的簇中，用迭代的方式不断重复上述过程，直到满足准则函数或划分过程中相邻簇的中心不再发生变化为止。通过不断的迭代过程增加簇内的紧凑性，降低簇间的相似性。图 7 为本文聚类的流程。

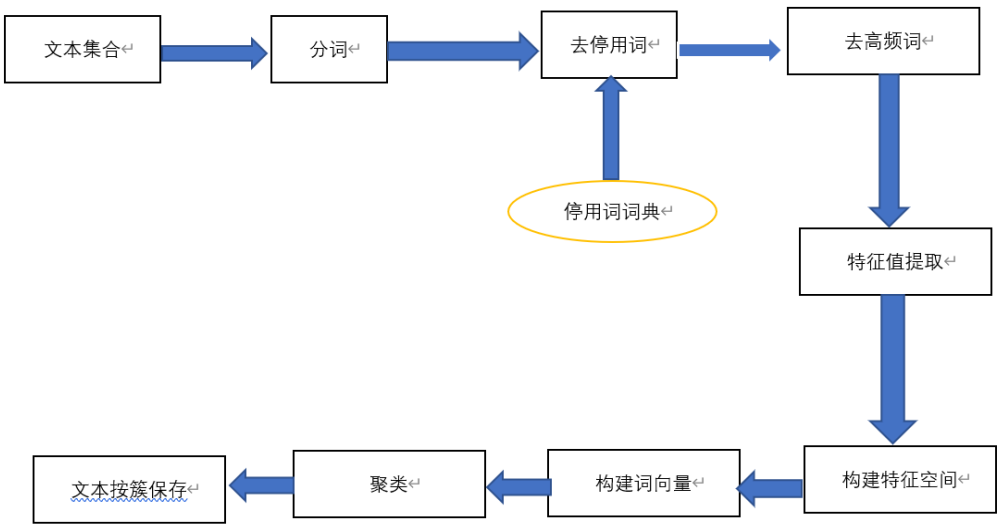

图 7　文本聚类流程

3.2.2.1.k-means 聚类基本思想

Kmeans 聚类算法算法解决的是将含有 n 个数据点(实体)的集合 X={x1,x2,...,xn}划分为 Zj 个类簇的初始簇中心，集合中每个数据点被划分到与其距离最近的簇中心所在的类簇之中，形成了 k 个聚类的初始分布。对分配完的每一个类簇计算新的簇中心，然后继续进行数据分配过程，这样迭代若干次后，若簇中心不再发生变化，则说明数据对象全部分配到自己所在的类簇中，聚类准则函数收敛，否则继续进行迭代过程，直至收敛。这里的聚类准则函数一般采用聚类误差平方和准则函数。本算法的一个特点就是在每一次的迭代过程中都要对全体数据点的分配进行调整，然后重新计算簇中心，进入下一次的迭代过程，若在某一次迭代过程中，所有数据点的位置没有变化，相应的簇中心也没有变化，此时标志着聚类准则函数已经收敛，算法结束。

3.2.2.2.k-means 聚类算法流程

（1）随机选择 K 个中心点
（2）把每个数据点分配到离它最近的中心点；
（3）重新计算每类中的点到该类中心点距离的平均值

（4）分配每个数据到它最近的中心点；

（5）重复步骤 3 和 4，直到所有的观测值不再被分配或是达到最大的迭代次数；

3.2.2.3.k 值的选取

按递增的顺序尝试不同的 k 值，同时画出其对应的误差值，通过寻求拐点来找到一个较好的 k 值

使用 TF-IDF 进行特征词的选取，下图是中心点的个数从 0 到 80 对应的误差值的曲线，如图 8 所示。

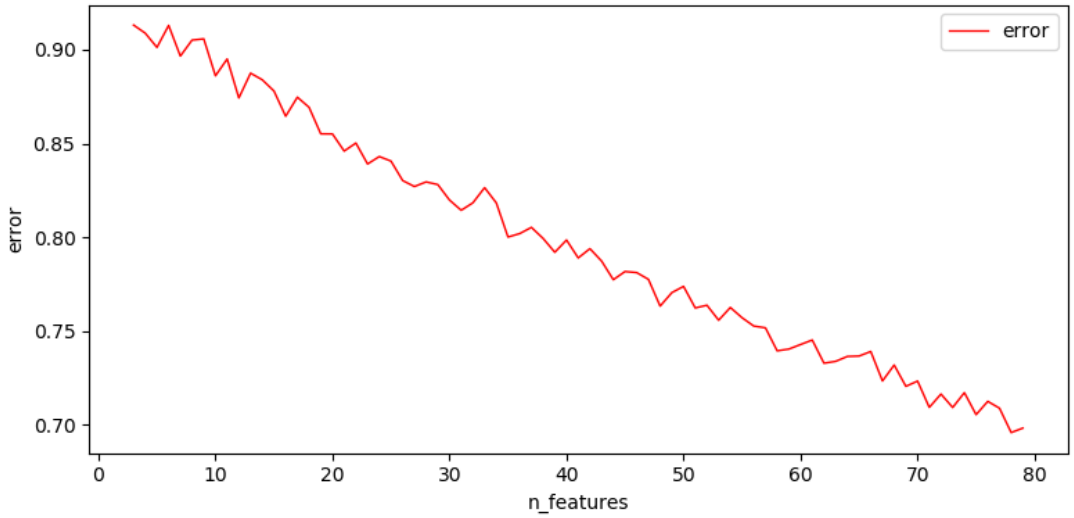

图 8 误差值曲线

从图中可以看出在 k=14 左右明显也有一个拐点，因此选取 k=14 作为中心点的个数，下面是 14 个簇数据集的分布情况，如表 5 所示。

表 5　簇数据集分布

簇的编号	簇的数目
0	16
1	21
2	3774
3	40
4	27
5	66
6	59
7	15
8	23
9	71
10	57
11	26
12	59
13	73

3.2.2.4.簇标签生成

聚类完成后，我们需要一些标签来描述簇，聚类完后，相当于每个类都用一个类标，这时候用 TF-IDF 方法来选取特征词作为标签。如表 6 所示。

表 6　簇标签表

簇	特征词
Cluster 0	加快 力度 步伐 刻不容缓 国家 城市 中心 覆盖 整治 发展
Cluster 1	公交车 a8 有限公司 幼儿园 开发商 公司 物业 举报 大道 小学
Cluster 2	江山 帝景 安全隐患 脏乱差 导致 幼儿园 楼盘 买房 学士 a8
Cluster 3	电梯 小区 安全隐患 嘉园 经常 三期 存在 庭院 安全 覆盖
Cluster 4	影响 生活 休息 小区 周边 出行 新城 正常 环境 交通
Cluster 5	违建 小区 县星沙 私自 拆除 现象 碧桂园 规划局 举报 保利
Cluster 6	项目 拖欠 滨河 景园 拆迁 是否 开发 合法 夜间 安置
Cluster 7	管理 加强 服务中心 部门 有限公司 需要 希望 企业 无人 车位
Cluster 8	市经 开区 泉星 项目 以南 优化 公园 大厦 新安 开放
Cluster 9	中学 补课 实验 学生 收费 工地 市长 深业 门口 区长
Cluster 10	住户 用水 蓝湾 金域 万科 商铺 物业公司 一个 小区 天下
Cluster 11	小区 新城 搅拌站 油烟 物业 安置 门面 魅力 停水 夜宵
Cluster 12	房屋 质量 小区 漏水 外墙 请求 开裂 存在 安全 村民
Cluster 13	改造 提质 旧城 小区 县星沙 请求 进行 工程 县星 启动

由于是无监督学习，每次结果都不一样，我们最后做出来的聚类效果达到了 0.86-0.88 之间，运行结果如图 9 所示。

```
score= 0.8741955454995932
```

图 9　聚类评分

聚类过后按簇分 14 类保存到 xlsx 文档中。

3.2.3.LDA 模型抽取

3.2.3.1.数据来源和语料预处理

我们对每个簇进行 LDA 主题模型抽取。在分词上我们选用 JIEBA 分词工具，停用词典使用哈尔滨工业大学的词典。由于现有的词典无法完全识别群众留言业

务的专业术语和业务词，为了提高分词效果，我们在原有的词典上，增加了许多词语。(1)我们使用正则表达式去除了留言文本中的特有的短语，电话号码，"领导"。（2）引入自定义词典，使用结巴工具进行分词，保留名词和动词等重要词语。并去除停用词(3)去除无关的高频词，如"谢谢"，"问题 ","举报 ","加快 "，预处理流程如图 10 所示。

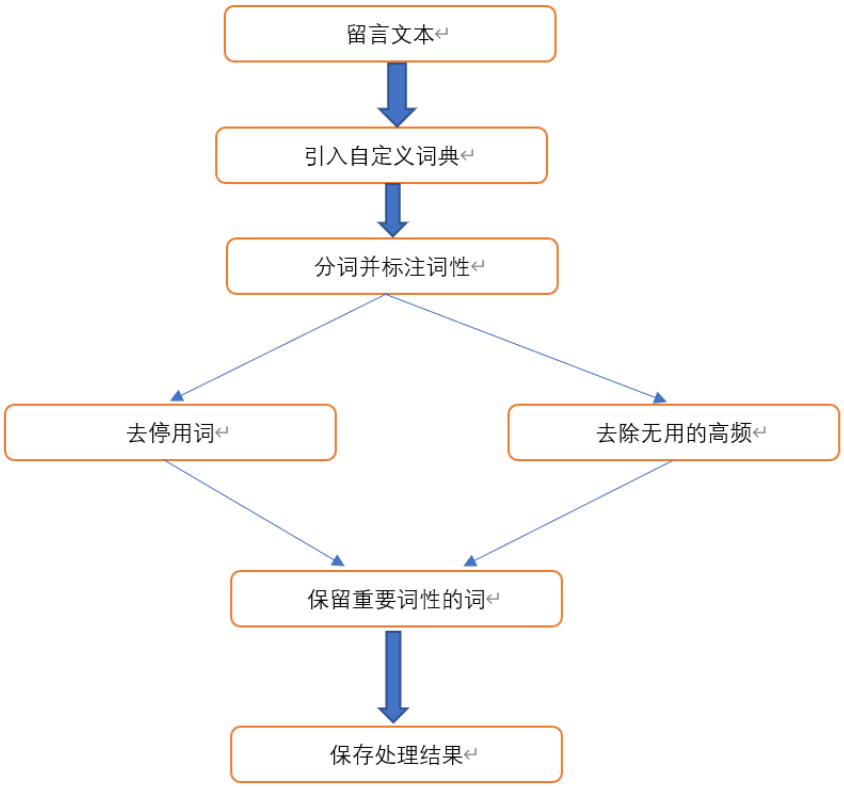

图 10　预处理流程图

LDA 模型中对话题的定义为: 一组语义上相关的词及这些词在该话题上的分布概率。由于无法对 LDA 模型的未知参数进行求解，在这里使用 GibbsSampling 的方法近似求解, Gibbs Sampling 通过迭代采样达到逼近真实结果的效果，其关键点在于对当前单词采样概率的求解，如公式(7)所示。

$$P\left(z_i^d\right)\left|z_{-i}^d, w\right) = \infty \frac{C_{dj}^{DK} + \alpha}{\prod_{K=1}^{K} C_{dk}^{DK} + K\alpha} \frac{C_{ij}^{VK} + \beta}{\prod_{K=1}^{K} C_{ki}^{VK} + V\beta} \tag{7}$$

其中各符号含义如表 7 所示。

表 7　符号含义

符号	符号含义
w	词表个数
K	话题数目
C_{ij}^{VK}	计数矩阵中的第ij项，表示第j个话题中的i个词出现的个数
C_{dj}^{DK}	计数矩阵中的第dj项，表示第d篇文章中的j词包含词的数目

通过 GibbsSampling 方法，可以得到 θ 和 φ 的后验值，如公式(8)和公式(9)所示：

$$\theta_j^d = \frac{C_{dj}^{DK} + \alpha}{\prod_{K=1}^{K} C_{dk}^{DK} + K\alpha} \tag{8}$$

$$\varphi_j^i \frac{C_{ij}^{VK} + \beta}{\prod_{K=1}^{K} C_{kj}^{VK} + V\beta} \tag{9}$$

在推导参数之前，需要预先将话题的数目 K 设置好，数值越大则话题越多，话题的颗粒度越小，反之 亦然。K 的取值对 LDA 模型文本提取和拟合性能影响较大，其最佳的确定可以通过两种方法：一种是词汇被选中的概率 p(w|T)，另一种是困惑度(perplexity)。

本文用困惑度确定 K，困惑度越小，话题的拟合性就越好。困惑度计算如公式(10)所示。

$$perlexity = \exp\left(-\frac{\sum_{i=1}^{M} log(p(d_i))}{\sum_{i=1}^{M} N_i}\right) \tag{10}$$

其中M为文本数，N_i为文本d_i的长度(即单词个数)，$p(d_i)$为 LDA 模型产生文本d_i的概率。

3.2.4.热点问题识别

使用 Gibbs Sampling 抽样可以得到“话题–词语” 和“文档–话题”的概率分布。对于“话题–词语”分布，每个话题 z 下分布着词语 w 和它在此话题中的概率 p(w|z)，话题 z={(w1,p(w1|z)),...,(wi,p(wi|z)),..(wn,p(wn|z))}对于“文档–话题”分布，

每个文档 d 下分布着 k 个话题的概率分布，形如 D={P(z1|d),...,P(zi|d),...,P(zk|d)} 。

使用 GibbsSampling 抽取的话题数量会比较多，而且有些话题可能表达的意思十分接近，有些话题几乎不能表达文档的意思，所以要进行话题选取。话题的选取就要用到上面的“话题–词语”和“文档–话题”的概率分布。经过话题选取之后，确定了文本的全局话题，然后从全局话题中发现热点话题。

3.2.4.1.选取标签词

文本经过聚类，得到了 H 个类，每个类使用 GibbsSampling 得到了若干个隐含的话题，每个话题下分布着 n 个话题相关的词，对每个话题中的词计算其在该话题所在类文本中的词频(count)、词跨度(cover)和词的长度(length)，则该词的权值(weight)计算公式如(11)所示。

$$weight = count + length + cover \tag{11}$$

计算完话题中词的权值后，选出权值最大的词作为该话题的标签词。

3.2.4.2.计算话题的文档概率分布均值

通过 Gibbs Sampling 对每个类抽样后，各自得到一个“文档–话题”概率分布矩阵，如公式(12)所示

$$\begin{array}{c|cccccc} & z_1 & \cdots & z_i & \cdots & z_k \\ d_1 & p(z_1|d_1) & \cdots & p(z_i|d_1) & \cdots & p(z_k|d_1) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ d_i & p(z_1|d_i) & \cdots & p(z_i|d_i) & \cdots & p(z_k|d_i) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ d_m & p(z_1|d_m) & \cdots & p(z_i|d_m) & \cdots & p(z_k|d_m) \end{array} \tag{12}$$

矩阵中有 k 个话题和 m 条文档中的分布概率。通过上面的矩阵概率分布就可以得出每个话题的分布概率均值，具体计算如公式(13)所示。

$$AVG(Z_i) = \frac{\sum_{j=1}^{m} p(Z_i|d_j)}{m} \tag{13}$$

3.2.4.3.话题提取

得到了话题的标签词和话题的文档概率分布均值后，构建话题矩阵如公式(14)所示。

$$\begin{matrix} topic_1: \\ \vdots \\ topic_i: \\ \vdots \\ topic_n: \end{matrix}\begin{bmatrix} topic_1\_tag & \cdots & avg(topic_1) & \cdots & H_I \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ topic_i\_tag & \cdots & avg(topic_i) & \cdots & H_J \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ topic_n\_tag & \cdots & avg(topic_n) & \cdots & H_K \end{bmatrix} \tag{14}$$

矩阵中一共有 n 个话题(topic), topici_tag 为 topici 的标签词, avg(topici)为 topicj 的文档概率分布均值，Hi、Hj 和 Hk 属于文本类集合 H。由于话题标签词存在相同的情况，所以先以话题标签词分组。认为同一组中的话题表达的意思相近，如果一组中有多个话题选取其中分布概率均值最大的话题，将其删除。接下来按每个话题的均值排序，去除均值极小的话题，因为均值小的话题不能很好地表达文档的意思，剩下的话题就是文档的全局话题。

3.2.4.4.热点问题识别

根据 LDA 模型的原理，每篇文档都是由数个不同的话题按照一定的比例生成的。这里假设一条经过预处理的投诉文本中有不少于话题 z 中百分之几的词，则认为这条投诉文本是话题 z 的支持文档。之后使用徐佳俊等的方法计算文档话题支持率，如公式(15)所示。如果在一个时间段内，话题的支持文档的数量或者文档话题支持率超过一个设定的阈值，那么这个话题就是热点话题。

$$S(z,t) = \frac{|D_i^t|}{|D^t|} \tag{15}$$

其中，z 表示话题，t 表示时间段，$|D_i^t|$为时间段 t 内话题 z 的所有支持文档数，$|D^t|$为时间段 t 内所有文档数目。

3.2.5.LDA 实验结果与分析

将每篇评论中概率值排在前两位的潜在主题看作是该篇评论的主题，统计所有评论的主题，然后按照主题出现次数的多少判定该主题是否为评论热点。本

文选取出现次数排在前 9 位的主题作为评论热点，特征词语在这 9 个主题上的概率分布情况，如表 8 所示（作为实例只显示了 4 个主题）。

表 8　话题-词语表

主题	词语即频率
Topic1	丽发 0.234 搅拌机 0.192 噪音 0.046 A 市 0.092 扰民 0.078 严重 0.004 显示 0.015 社区 0.006 栋 0.005 开发商 0.004
Topic2	景园 0.14729 滨河 0.14729 销售 0.08624 捆绑 0.04750 市伊 0.04224 商品房 0.04224 广铁集团 车位 0.018 开发商 0.005 购买 0.021
Topic3	东路 0.007 劳动 0.004 魅力 0.143 小区 0.034 油烟 0.121 夜宵 0.092 临街 0.0083 污染 0.022 开裂 0.0061 餐饮 0.009
Topic 4	相关 0.013 部门 0.008 发展 0.008 政策 0.009 购房 0.009 人才 0.110 国道 107 同意 0.007 我家 0.005 开发商 0.004

从表 8 中 Topic1,Topic2,Topic3,Topic4 代表留言集里面的四个潜在主题，就是热点问题 每个主题下面是该主题在各个特征词上的概率分布情况，这里将主题在各个特征词上的概率分布值从大到小进行排列，列出排在前 10 位的主题特征词。

在 Topic1 中丽发（0.234）搅拌机 (0.192)，噪音(0.046)，A 市(0.092)，扰民(0.078)概率比较高，可以判断该主题对应的热点问题，丽发新城附近施工，搅拌机扰民。在 topic2 中景园（0.14729),滨河(0.14729),销售(0.08624) 这些特征词出现概率比较高，可以推断它们与热点问题伊景园捆绑销售有关。在 topic3 中魅力（0.143） 小区(0.034)油烟(0.121)夜宵(0.092) 临街 (0.0083) ,在 excel 表格中查找到了这个属于热点问题“魅力之城小区临街门面油烟直排扰民”在 topic4 中发展（0.008） 政策(0.009)，购房(0.009),人才(0.110) 这些特征值出现概率比较高，得出对应的热点问题是 A 市自选人材住房补贴

表 9 列出了实验获得的 9 个热点问题以及它们对应的前 10 个特征词（通过热点词语，通过筛选得出对应的热点问题）。

表 9　热点词语

热点问题	热点词语 (前 10 个)
A 市丽发新城违建搅拌站，彻夜施工扰民污染环境	领导 小区 公平 举报 丽发 县泉塘 中学 希望 新城 私自
关于伊景园滨河苑捆绑销售车位的维权投诉	景园 滨河 销售 车位 捆绑 市伊 商品房 广铁集团 项目 开发商
魅力之城小区临街门面油烟直排扰民	东路 劳动 魅力 一楼 小区 门面 油烟 夜宵 临街 污染
咨询 A 市人才购房补助发放问题	相关 部门 发展 政策 购房 人才 国道 107 同意 我家
A 市涉外经济学院强制学生实习	学生 学院 职业 技术 强制 经济 学费 组织 补课 中学
A7 县诺亚山林小区门口设置医院	医院 反对 门口 人民 山林 职工 坚决 小区 设置 不能
A 市黄花机场噪音扰民	每天 a9 深夜 超市 三期 货车 机场 凌晨 有人 门口
A 市中海国际小区附件噪音扰民	中海国际 噪音 3 期 社区 一直 扰民 小孩 门口太吵
A 市经开区规划需优化	经开区 规划 优化 A 市 项目 不合理 有关一步 建议

从表 9 中可以看出,前十个热点词语的内容很好的反映了政府留言中的主要热点问题，“噪音扰民”，"捆绑销售"，“咨询人才补贴”,"强制实习"，“夜宵油烟污染环境”等等与群众所生活的息息相关,体现了大部分群众所关心的问题，希望政府可以尽快解决，因此成了人们的热点留言问题也理所当然。

3.2.6.点赞数与反对数

首先我们对点赞数和反对数筛选出来排名前十的问题，然后对于点赞数和反对数进行归一化处理，我们定义比例系数大于 0.1 定义为热点问题，如表 10 所

示，我们可以看到在点赞数中，“A 市 A5 区汇金路五矿万境 K9 县存在一系列问题”，“反映 A 市金毛湾配套入学的问题”反应的比例最大，这是一个热点问题，在反对数中，“咨询 A9 市高铁站选址的问题”问题反对比例最大，这也是一个热点问题。

表 10　点赞数-反对数归一化

问题	点赞数(归一化)
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	0.166560763
反映 A 市金毛湾配套入学的问题	0.139952343
请书记关注 A 市 A4 区 58 车贷案	0.065210485
严惩 A 市 58 车贷特大集资诈骗案保护伞	0.062748213
承办 A 市 58 车贷案警官应跟进关注留言	0.05822081
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？	0.053137411
A 市富绿物业丽发新城强行断业主家水	0.019221604
建议西地省尽快外迁京港澳高速城区段至远郊	0.006354249
请问 A 市为什么把和包支付作为任务而不让市场正当竞争？	0.006195393
问题	**反对数（归一化）**
咨询 A9 市高铁站选址的问题	0.099437
A7 县未来漫城物业不作为，还能评为五星级	0.028143
A7 县东一路公交首末站是临时的？	0.026266
A7 县泉塘中学领导对待老师不公平	0.020638
A7 县泉塘中学领导对待老师不公平	0.020638
咨询 A7 县道路规划的问题	0.016886
A3 区奥克斯缔壹城小区开放式入户通道存在严重安全隐患	0.016886
A 市经济学院体育学院变相强制实习	0.016886
A7 县广圣大酒店 4 楼瑞生堂足道涉黄	0.013133

3.2.7.热点问题总结

我们需要筛选出排名前五的热点问题，我们结合关键词和和词语频率和点赞数反对数指标,对于关键词和词语赋予 0.4 的比重(关键词按出现的概率计算)，反对数和赞成数各赋予 0.3 的比重(表格中空白代表数目太小取 0)对于关键词和词语频率和反对数赞成数筛选出的热点问题进行排名，筛选出以下排名前五热点问题，如表 11 所示。

表 11　热点问题表

问题	关键词与词频	反对数	赞成数	得分	排名
A 市 A2 区丽发新城小区附近搅拌站扰民污染环境	0.234			0.0936	1
A 市伊景园滨河购房捆绑销售车位	0.147			0.0588	2
A 市万科魅力之城小区店铺油烟噪音扰民	0.143			0.0572	3
A 市人才新政的购房补贴相关咨询	0.11			0.044	6
A 市经济学院强制学生学习	0.09			0.036	7
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题			0.166560763	0.049968229	4
反映 A 市金毛湾配套入学的问题			0.139952343	0.041985703	5
咨询 A9 市高铁站选址的问题		0.09944		0.029811	8

3.3.答复意见的评价

3.3.1.评价指标的说明

附件 4 中，相关部门对群众的留言做出了答复意见，我们需要从相关性、完整性和可解释性等角度对答复意见的质量做出评价。现根据留言与答复的实际内容，可将答复意见评价的三大指标按如下方式进行定义：

（1） 相关性：

答复意见需要与相对应的留言内容具有很高的相关性，同时与不对应的留言主题具有较低的相关性。

（2） 完整性

答复意见的完整程度应取决于对群众留言中提出问题的包含程度。较高的完整性意味着答复不仅对当前留言的问题做出了回答，还附带了对该问题的解释、说明、评价及结论等。

（3） 可解释性

一个合理的答复意见需要具备足够的可解释性。有些答复看似对问题做出了大量的解释，但实际上并没有解决该问题的实质性举措，这样的答复所具有的可解释性应当是较低的。评价一个答复可解释性的高低就在于能从中挖掘出具体有价值的信息的多少。

3.3.2.答复意见的相关性评价

首先，判断两者之间的相关程度应先看文字的重叠程度，根据留言中关键词在答复中出现的频率考核答复的相关性。其次，还需要考虑语义上的相似性。

我们使用 Excel API 中的 GetMatchingDegree 函数，其算法原理是比对两个文本字符串中每一个字或单词在长文本中出现的次数，不分顺序，然后返回百分比，百分比越高相似程度就越高。选择匹配留言详情和答复意见两段文本的相关度，可以得到一个相似度百分比R_i，当相似度高于 10%时我们判定该答复意见与对应留言内容相关，反之判定为不相关。

3.3.3.答复意见的完整性评价

我们可以通过对每条答复意见单独进行挖掘，提取出相应的主题。每条答复意见可以包含多个方面的内容，关键词就能较好地反映一句话的逻辑思路。“经调查”、“目前”等词语后面的内容多为某一问题现状的陈述，“将”、“争取”等词语则表明了政府接下来针对问题可以采取的措施等，“可以”、“即可”等词语往往含有政府对个人解决问题的指导性建议。通过对单一答复语义的深度挖掘，总结其包含的语义门类，再结合留言的关键词可以初步计算该答复的复杂程度，得到一个完整性系数 I_i ，作为答复意见的完整性评价指标。

3.3.4.答复意见的可解释性评价

在考虑一段答复的可解释性时，我们需要主要一大段文字所包含的语义主题往往不止一个，可能有两三个甚至更多。因此我们只要能把大段文字更精确的分开成多个并没有什么关联的主题，那么模型的区分度就会更高，答复内容更清晰，其可解释性也就越高。

除此之外，一段文字也有可能由多个语言层嵌套而成。要识别这样一句话的真正含义也会更加复杂。因此当句子的逻辑层次越简单时，答复的可解释性也就越高。可以通过词语组合、概率统计等方法进行处理。定义一个清晰度C_i来反映答复意见的可解释性。

3.3.5.答复意见的质量评价

通过上述三个方面的考量，我们初步得到答复意见质量的评价标准：

$$Q_i = \lambda(R_i, I_i, C_i) \tag{16}$$

其中质量Q_i与三个指标R_i、I_i、C_i均成正相关关系。

4.参考文献

[1] 李英. 基于词性选择的文本预处理方法研究[J]. 情报科学, 2009, 27(5):717-719.

[2] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[3] 徐佳俊, 杨飏, 姚天昉, 等. 基于 LDA 模型的论坛热点话题识别和追踪[J]. 中文信息学报, 2016, 30(1): 43-50.

[4] 余传明, 张小青, 陈雷. 基于 LDA 模型的评论热点挖掘:原理与实现%Mining Hot Topics of User Comment Based on LDA Model: Principle & Approach[J]. 情报理论与实践, 2010, 033(005):103-106.

[5] 伍万坤, 吴清烈, 顾锦江. 基于 EM-LDA 综合模型的电商微博热点话题发现[J]. 现代图书情报技术 , 2015(11):33-40.

[6] 张敏 吴郁松. 我国电子政务的研究热点与研究趋势分析[J].ChinaXiv 合作期刊,2017,2(2):1-27.

[7] 张小青 陈雷. 基于 LDA 模型的评论热点挖掘:原理与实现[J]. 知网,2010,2(2):1-26.

[8] 哈尔滨工业大学停用词词典 [OL]. [2016-11-23]. http://more.datatang.com/data/13281. (Stop Word Dictionary by Harbin Institute of Technology [OL]. [2016-11-23]. http://more.datatang.com/data/13281.)

[9] jieba [CP/OL]. [2016-11-23]. http://www.oschina.net/p/jieba.