基于自然语言处理技术的智慧问政系统研究

摘　要：近年来，网络问政平台的涌现带来了政府与公众互动模式的革新，促进了公众与政府互动交流的无缝连接，也逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道。但是，随着各类社情民意相关的文本数据量不断攀升，给留言划分和热点整理的工作带来了极大挑战。为了解决这个问题，本文将建立基于自然语言处理技术的智慧问政系统，实现对海量文本的自动化处理。

首先，对文本数据进行预处理操作，包括文本清洗、停用词去除、歧义消解、分词和文本向量化等步骤，为后面的分析提供所需要的数据。

对于第一题，由于数据各个类别的数量差距过大，不够均衡，因此我们首先利用**回译法**对数据进行增强，得到较为均衡的样本。然后我们分别建立了**朴素贝叶斯、XGBoost 和 LightGBM 模型**，效果均较好。为了得到更好的效果，我们将三种模型进行融合，最终得到的融合模型准确率达 93.3%，F1 为 0.932。

对于第二题，首先建立文本的 VSM 向量，采用 **Single-Pass 算法**对留言进行聚类，并采用轮廓系数衡量聚类效果，然后根据定义的热度指标筛选出排名前五的热点问题；针对聚类后的每一类问题，提取它们的标题，并利用 **TextRank 算法**，对标题的重要程度进行排序，用重要性最高的标题来描述该处热点的问题，最终以热点描述表和词云图展现。

对于第三题，我们通过特征提取，从回复**是否有理有据、内容是否充实、回复是否及时和是否套用模板**四个方面对回复质量进行量化描述，通过权重系数加权求和得到综合评分，以此建立回复质量评价模型。实验结果表明，该模型可以很好地评价一个回复的质量，具有良好的区分度和合理性，能够促进优质和高效的问题回复，帮助问政系统提供更好的用户体验。

最后，通过对问政文本数据的分析与研究，我们对所做的工作进行总结，并进行系统化的应用，采用 Streamlit 技术将算法**快速部署**，实现了网页上的**交互式操作**，降低使用门槛，以帮助相关政府部门进行对问政数据的分析，提高问政平台的效率和水平。

关键词：智慧问政系统　LightGBM　Single-Pass　TextRank　评价模型

Abstract

In recent years, the emergence of web-based question-and-answer platforms has brought about an innovation in the mode of interaction between the government and the public, promoting a seamless connection between the public and the government, and has gradually become an important channel for the government to understand public opinion, gather public wisdom and gather public sentiment. However, as the volume of text data related to various social and public opinions continues to rise, it has brought great challenges to the classification of messages and the collation of hotspots. To solve this problem, this paper will build a smart government system based on natural language processing technology to automate the processing of large volumes of text.
First, pre-processing operations are performed on the text data, including steps such as text cleaning, deactivation word removal, ambiguity decomposition, participle and text vectorization to provide the data needed for later analysis.

For the first question, since the number of categories is too different and not balanced enough, we first enhance the data by using the translation method to get a more balanced sample. We then built the Park Bayesian, XGBoost and LightXGB models, respectively, with better results. To get better results, we fused the three models and got the final fusion model accuracy as high as 93.3% and F1 as 0.932.

For the second question, the data were first cleaned, the messages were clustered using the Single pass algorithm, and the clustering effect was measured using the profile coefficient. The top five hotspot issues are screened according to pre-determined evaluation metrics; for each category of issues after clustering, their headings are extracted, and the importance of the headings is ranked using the TextRank algorithm to describe the hotspot issues in the location with the most important headings.

For the third question, the quality of the responses was quantified by extracting the characteristics, describing the quality of the responses in terms of whether it were justified, whether the content was adequate, whether the responses were timely and whether the templates were used. The experimental results show that the model can evaluate the quality of a response well, with a good degree of differentiation and reasonableness, and can promote quality and efficient responses to questions and help smart government systems to provide a better user experience.

Finally, through the analysis and study of the textual data of grievances, we summarize the work done and apply it systematically, using Streamlit technology to rapidly deploy the algorithm to realize the interactive operation on the webpage and lower the threshold of use, so as to help the relevant government departments to analyze the data of grievances and improve the efficiency and level of the platform.

Key Words: LightXGB, Single-Pass, TextRank, Evaluation model

目　录

1. 引言

1.1 背景

近年来，互联网已经发展成为人们获取信息、关注热点事件、了解国情乃至了解世界的重要媒介。目前，我们已经步入了“互联网+”的生活时代。社会各方面高速发展，科技发达，信息泉涌，人们与政府之间的交流越来越密切，网络问政平台逐渐成为公众与政府互动交流最便捷的方式，也逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道。

但是，随着各类社情民意相关的文本数据量不断攀升，给留言划分和热点整理的工作带来了极大挑战。因此，引入自然语言处理及其相关技术来实现对海量文本的自动化处理变得尤为重要。同时，如何高效、合理地评估政府回复的质量成为问政平台发展中亟需解决的问题。

1.2 研究内容

针对电子政务的特点，本文利用政务问答数据和自然语言处理技术，主要研究以下四个方面的问题：

（1）基于群众留言的文本和分类标签，构建文本分类模型，能够对新文本进行标签分类；

（2）基于海量留言数据，挖掘出其中的热门事件，并且给出事件的相关描述和热度指数；

（3）对政府回复从多角度进行量化分析和评估，给出对答复意见质量的评价方案；

（4）构建综合智慧问政平台，实现对问政数据的一站式处理。

2. 数据预处理

本文的数据预处理流程如图 2-1 所示。

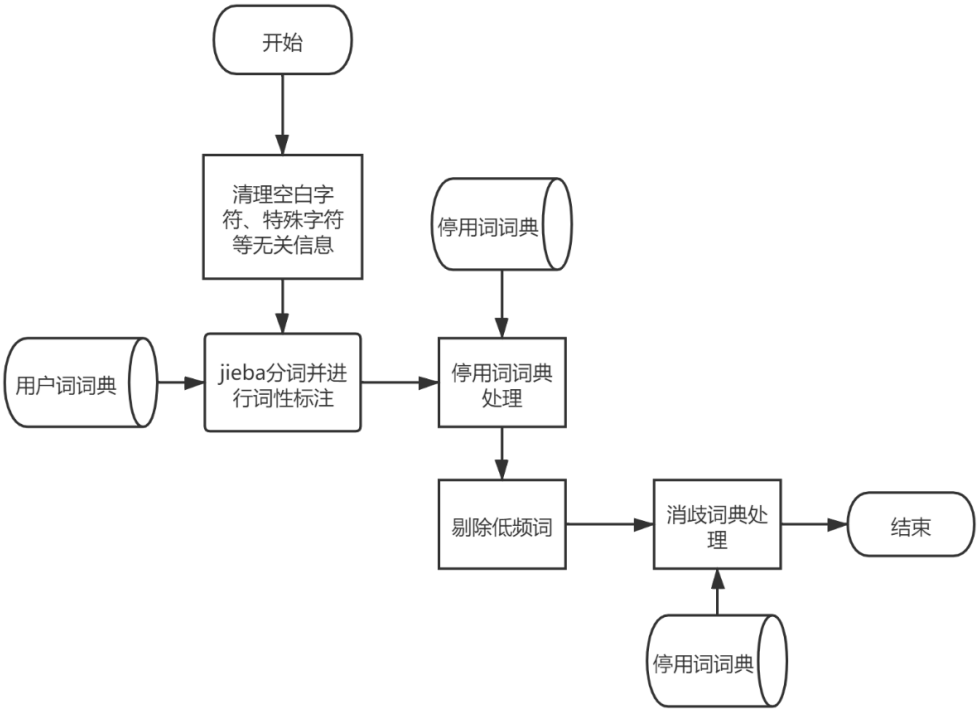

图 2-1 文本预处理流程图

2.1 去除无关字符

通过对文本的分析，我们发现文本中存在许多无用无意义的字符或连接，如空白字符、“baidu.cm”等，这些无意义的字符给模型训练和评估带来负面的影响， 因此，我们首先在文本中将这些无关字符去除。

2.2 去除停用词

由于并不是每一个词都能表征这篇文章的内容，如果保留，输入特征会很大，影响训练效果，因此有一些形如“这样”、“1.2.3.4”的词就应该被删除掉，本文从网络上寻找一份中文的停用词表作为参考。将文章中的词与停用词表中的词作比较，如果在表中出现该词，就将其删除，如果没有出现，就跳过。

2.3 分词和词性标注

常见的分词方法有：基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法和基于规则的分词方法，每种方法下面对应许多具体的算法。对于此问题而言，本文采用 Jieba 分词库。

首先由 Jieba 分词工具进行中文分词，然后对分词进行词性标注，最后通过过滤词性保留有意义的词。这里过滤掉的词性有：标点符号、连词、助词、副词、介词、时语素、数词、代词，分词的部分处理结果如图 2-2 所示。

	title	time	content	up	down	text	content_cut
0	A3区一米阳光婚纱艺术摄影是否合法纳税了?	2019/2/28 11:25:05	\n\t\t\t\t\t\n\t\t\t\t\t座落在A市A3区联丰路米兰春天G2栋320，...	0	0	A3区一米阳光婚纱艺术摄影是否合法纳税了? A3区一米阳光婚纱艺术摄影是否合法纳税了? A3区一...	[一米阳光, 婚纱, 艺术摄影, 合法, 纳税, 一米阳光, 婚纱, 艺术摄影, 合法, 纳...
1	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	\n\t\t\t\t\t\n\t\t\t\t\tA市A6区道路命名规划已经初步成果公示文件，...	1	0	咨询A6区道路命名规划初步成果公示和城乡门牌问题咨询A6区道路命名规划初步成果公示和城乡门牌...	[咨询, 道路, 命名, 规划, 成果, 公示, 城乡, 门牌, 咨询, 道路, 命名, 规...
2	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	\n\t\t\t\t\t\n\t\t\t\t\t本人系春华镇金鼎村七里组村民，不知是否有相关...	1	0	反映A7县春华镇金鼎村水泥路、自来水到户的问题反映A7县春华镇金鼎村水泥路、自来水到户的问题...	[春华, 金鼎村, 水泥路, 自来水, 到户, 春华, 金鼎村, 水泥路, 自来水, 到户,...
3	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	\n\t\t\t\t\t\n\t\t\t\t\t靠近黄兴路步行街，城南路街道、大古道巷、一步...	1	0	A2区黄兴路步行街大古道巷住户卫生间粪便外排A2区黄兴路步行街大古道巷住户卫生间粪便外排A2...	[黄兴路, 步行街, 古道, 住户, 卫生间, 粪便, 外排, 黄兴路, 步行街, 古道, ...
4	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	\n\t\t\t\t\t\n\t\t\t\t\tA市A3区中海国际社区三期四期中间，即蓝天璞...	0	0	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民A市A3区中海国际社区三期与四期中间...	[市区, 中海, 国际, 社区, 四期, 空地, 施工, 噪音, 扰民, 市区, 中海, 国...

图 2-2 部分处理结果

2.4 TF-IDF

定义一个词的权重通常采用的是 TF-IDF 的方法[10]。在信息检索理论中，TF-IDF 是 Term Frequency - Inverse Document Frequency 的简写，TF 是词频，IDF 是逆文档频率，用于反映一个词对于语料中某篇文档的重要性。在信息检索和文本挖掘领域，它经常用于因子加权。TF-IDF 的主要思想就是：如果某个词在一篇文档中出现的频率高，也即 TF 高；并且在语料库中其他文档中很少出现，即 DF 低，也即 IDF 高，则认为这个词具有很好的类别区分能力。

Cornell SMART 系统的词频的计算公式如下：

$$TF(d,t)=\begin{cases}0, \text{ if } freq(d,t)=0\\ 1+\log(1+\log(freq(d,t))), \text{ others}\end{cases}$$

其中，$freq(d,t)$ 是词在文档中出现的频率。

IDF 的计算公式如下：

$$IDF(t) = \log \frac{1+|d|}{|d_t|}$$

其中，d 是文档的集合，d_t 是包含词的文档的集合。

TF-IDF 的计算公式为:

$$TF \times IDF = TF(d,t) \times IDF(t)$$

本文先将语料转化为词袋向量，根据词袋向量统计 TF-IDF，此数据集文本序列长度分布如图 2-3 所示，通过计算发现 98.6%的样本文本序列长度都小于 2500。为了简化计算，做出每 2500 词划分一次的调整，长度大于 2500 的进行切分，小于 2500 的进行填充。最终可以得到文本的 TF-IDF 表示。

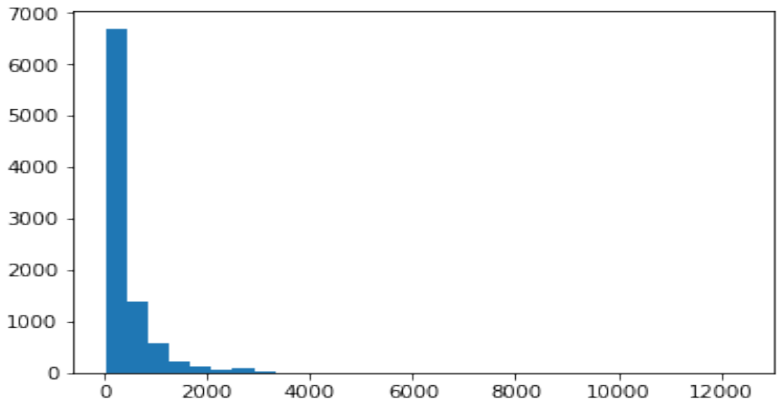

图 2-3 文本长度分布图

3. 问题一：多标签分类

3.1 数据增强

首先，观察数据集，数据分布如图 3-1 所示，共有七个分类，且数据具有不平衡的特性，最多的类别“城乡建设”的数量是最少类别“交通运输”的三倍多。由于多数类和少数类在数量上的倾斜，以总体分类精度最大为目标会使得分类

模型偏向于多数类而忽略少数类，导致少数类被判断为多数类的概率大大增加，造成少数类的分类精度较低。

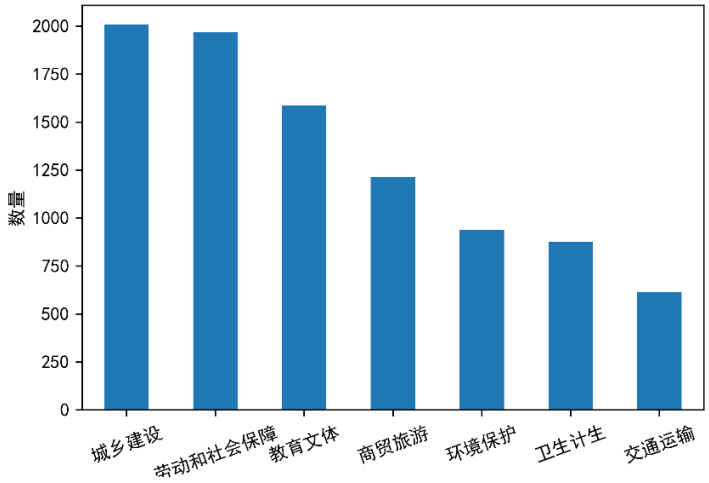

图 3-1 类别分布图

对于数据不平衡问题，可以从三个角度予以解决，分别是数据角度、评价指标角度和算法角度。此处采用从数据角度来着手，对原始数据进行回译以达到增强数据的效果，即将目标数据进行多次不同语种翻译，再翻译回来的一种方法，将其它六类扩充到与城乡建设数量相同。扩充后数据集数量达到 14063 条。数据增强示意图如图 3-2 所示。

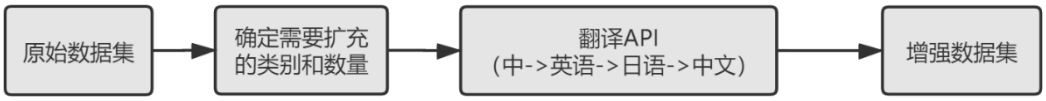

图 3-2 数据增强示意图

数据增强示例：

原文本：尊敬的各位领导，我是来自 K8 县的一个公民，我投诉的内容是 K8 县一中高二老师宋建海在外私自办培训机构“聚能教育”，望上级领导重视。

增强文本：杰出的领导人，我是 K8 县的公民。我的抱怨是，宋建海是 K8 县第一中学的一名高中教师，经营着一所私立教育机构“聚能教育”。

3.2 朴素贝叶斯模型

3.2.1 算法原理

朴素贝叶斯是经典的机器学习算法之一，通过考虑特征概率来预测分类，是为数不多的基于概率统计学的分类算法。朴素贝叶斯的核心是贝叶斯定理，而贝叶斯定理的公式本质上是条件概率。

$$p(c_i|x,y) = \frac{p(x,y|c_i)p(c_i)}{p(x,y)}$$

这里的 C 表示类别，输入待判断数据，式子给出要求解的某一类的概率。我们的最终目的是比较各类别的概率值大小，而上面式子的分母是不变的，因此只要计算分子即可。

3.2.2 模型效果

将 14063 条数据 80% 划分为训练集，20% 划分为测试集。运用朴素贝叶斯模型进行预测，在测试集上准确率达到了 91.68%，F1-Score 达到了 0.9161，预测结果如图 3-3 所示。

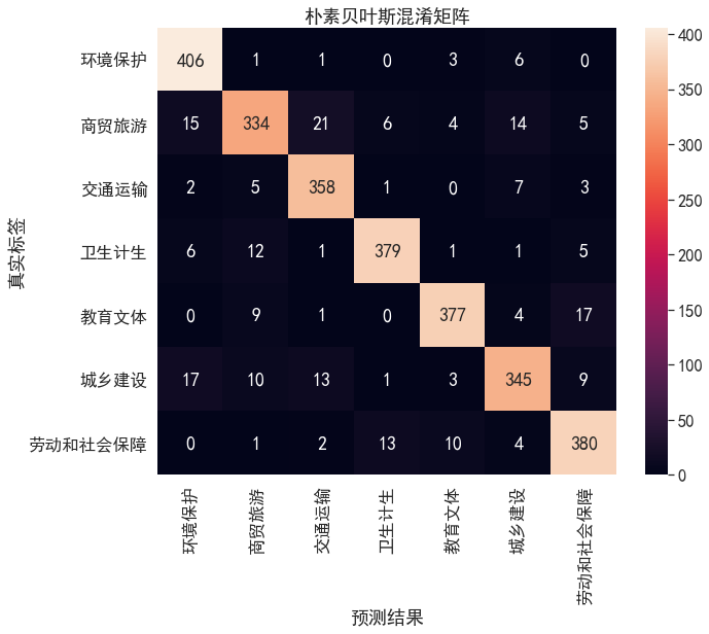

图 3-3 朴素贝叶斯预测结果

3. 3 XGBoost 模型

3.3.1 算法原理

XGBoost（eXtreme Gradient Boosting）全名叫极端梯度提升，它是一种 tree boosting 的可扩展机器学习系统。XGBoost 的核心算法思想是：

（1）不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数 $f(x)$，去拟合上次预测的残差；

（2）当我们训练完成得到 k 棵树，我们要预测一个样本的分数，其实就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数；

（3）最后只需要将每棵树对应的分数加起来就是该样本的预测值。

3.3.2 算法效果

将 14063 条数据 80% 划分为训练集，20%划分为测试集。运用 XGBoost 模型进行预测，在测试集上准确率达到了 92.78%，F-Score 达到了 0.9277，预测结果如图 3-4 所示。

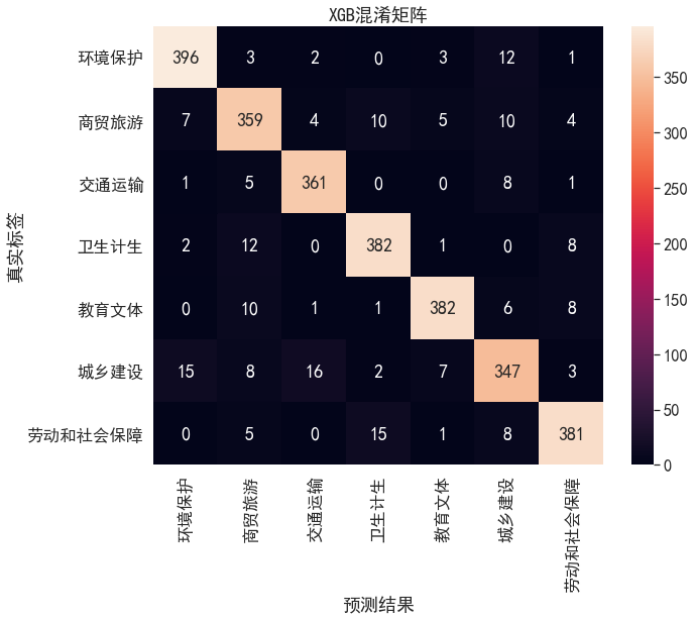

图 3-4 XGBoost 预测结果

3. 4 LightGBM 模型

3.4.1 算法原理

LigthtGBM 是 boosting 集合模型中的新进成员，由微软提供，它和 XGBoost 一样是对 GBDT 的高效实现，原理上它和 GBDT 及 XGBoost 类似，都采用损失函数的负梯度作为当前决策树的残差近似值，去拟合新的决策树。而 LGB 的优化点有：

1. 基于 Histogram（直方图）的决策树算法，代替 pre-sorted 所构建的数据结构，利用 histogram 后，会有很多有用的 tricks。例如 histogram 做差，提高了 cache 命中率（主要是因为使用了带深度限制的 leaf-wise 的叶子生长策略）；

2. 在机器学习当中，面对大数据量时候都会使用采样的方式（根据样本权值）来提高训练速度。又或者在训练的时候赋予样本权值来关于于某一类样本（如 Adaboost）。LightGBM 利用了 GOSS（基于梯度的 one-side 采样）GOSS 来做采样算法；

3. 由于 histogram 算法对稀疏数据的处理时间复杂度没有 pre-sorted 好。因为 histogram 并不管特征值是否为 0。因此采用 EFB（互斥的特征捆绑）来预处理稀疏数据。

3.4.2 算法效果

将 14063 条数据 80% 划分为训练集，20%划分为测试集。用 LightGBM 模型进行预测，在测试集上准确率达到了 92.74%，F-Score 达到了 0.9271。预测结果如图 3-5 所示。

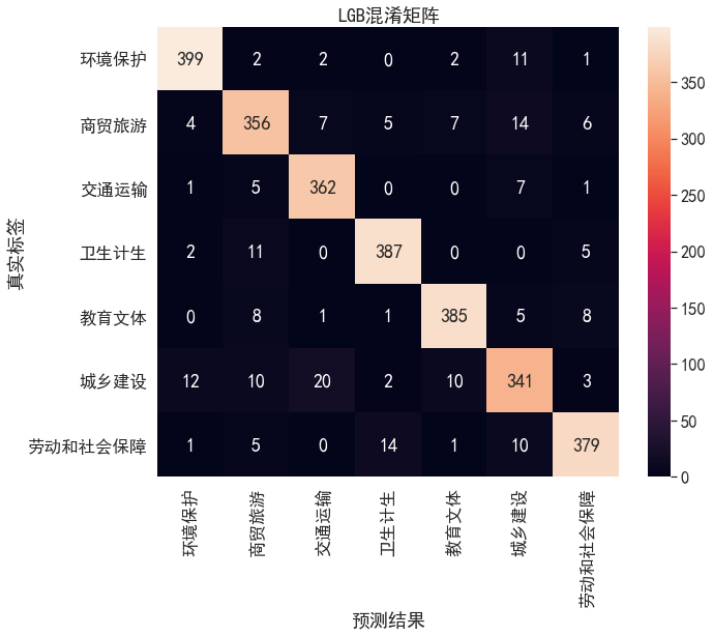

图 3-5 LightGBM 测试效果

3.5 模型集成

机器学习中的分类模型在训练结束后，我们希望训练出一个在各种指标下的表现都十分良好的模型，但是真实的情况往往不是如此，一个模型在某些评价指标上表现良好，在另外的评价指标上的表现可能就很差。通常我们只能得到在某几个指标下表现良好的多个单一的分类模型。Ensemble learing 的主要思想就是将多个单一的分类模型的结果综合起来考虑，来获得最后的分类结果。在这种情况下，模型对某几个模型产生的错误就会具有一定的容错性。

3.5.1 投票融合

我们采用上述朴素贝叶斯模型、XGBoost 模型、LightGBM 模型三种模型获得不同的结果，然后采用投票的方式，选取出最高的票的类别作为最终分类结果。关键实现代码如下：

```
predict = np.zeros((3, 2813))
predict[0], predict[1], predict[1] = xgb_pred, lgb_pred, NB_pred
merge_vote = np.zeros(2813)
```

```
4.
5. for i in range(2813):
6.     (values, counts) = np.unique(predict[:, i], return_counts=True)
7.     idx = np.argmax(counts)
8.     merge_vote[i] = values[idx]  评估结果
```

但是效果不增反降，在上述三个模型的预测基础上采用投票的方法进行集成，在测试集上准确率只有 88.80%，F-Score 下降到了 0.8941。

3.5.2 算术平均融合

我们利用上述三个模型计算出数据样本在各个分类上的概率，通过求三个结果的算术平均值来作为判断的依据，选取概率最大的类别作为最终分类结果。关键实现代码如下：

```
1. pred = (xgb_pred + lgb_pred + NB_pred) / 3
2. predict = np.argmax(pred, axis=1)
```

其结果如图 3-6 所示。此方法较前面的几种方法效果有所提高，在测试集上准确率达到了 93.31%，F-Score 提升到 0.9329。

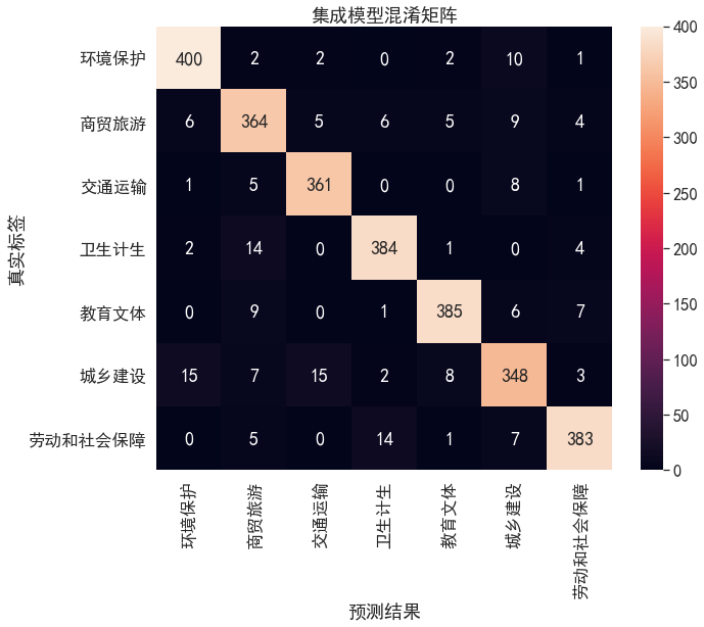

图 3-6 集成模型测试效果

3.6 问题一总结

通过多次的模型调整与融合，我们最终的集成模型 F1 高达 0.933，能够出色完成多文本分类任务。

表 3-1 模型结果

模型	准确率	$F1$
朴素贝叶斯	0.9168	0.9161
XGBoost	0.9278	0.9277
LightGBM	0.9274	0.9271
投票融合	0.8880	0.8941
算术平均融合	0.9331	0.9329

集成模型详细评测数据如图 3-7 所示

	precision	recall	f1-score	support
环境保护	0.94	0.96	0.95	417
商贸旅游	0.90	0.91	0.90	399
交通运输	0.94	0.96	0.95	376
卫生计生	0.94	0.95	0.95	405
教育文体	0.96	0.94	0.95	408
城乡建设	0.90	0.87	0.89	398
劳动和社会保障	0.95	0.93	0.94	410
accuracy			0.93	2813
macro avg	0.93	0.93	0.93	2813
weighted avg	0.93	0.93	0.93	2813

图 3-7 集成模型效果

4. 问题二:热点事件挖掘

面对纷繁复杂的各类留言，人工寻找最需解决的问题即热点问题会耗费很多的时间和精力.因此，利用自然语言处理和文本挖掘的方法在众多的留言中挖掘热点问题很有必要，政府可以了解民众对某个领域的关切程度和社会需要解决的

问题。了解当前的舆论焦点和民意，有助于相关部门进行有针对性的处理，提升服务效率。

4.1 国内外研究现状

热点事件提取是话题检测与跟踪（Topic detection and tracking,简写 TDT）的一个分支[1], TDT 是信息检索的研究领域之一, 包含五种任务, 如表 4-1 所示。:

表 4-1 TDT 任务

任务	描述
报道切分	将原始数据流切分成具有完整结构和统一话题的报道
新故事检测	从一系列新闻报道流中识别出新的话题或者报道。
关联故事检测	判断两个报道或者文档是否是谈到同一个话题。
聚类检测	将涉及到相关话题的报道或者文档聚成同一个类。
话题跟踪	从新闻报道数据流中追踪已经获取到话题

TDT 以事件信息组织为特征，以事件为目标。因此，大部分的研究将 TDT 的技术应用于事件检测中（event detection）。相对于传统的信息检索，TDT 更倾向于处理动态非确定的概念，类别和基于内容的话题。然而传统的信息检索技术如文本分类，话题模型（topic model），LSA 等适合文档分类和索引的方法，但却不太适合于事件和新事件发现，因而 Allan, Laverenko 等人提出基于传统全文相似度的上界来检测[2]。

事件检测有许多种方法，主流的方式是采用聚类，这样检测事件可以借鉴传统的信息检索的方法聚类对文档进行聚类，将关联的文档按照一定的度量方法，通常是相似度，将相关的文档放到同一个类中。每一个类代表一个单独的事件或者话题。聚类可以分为两类，一类是层次聚类（Hierarchical clustering）建立关系，将关系相近的簇聚成一类，分层进行聚类，另一类平聚类（flat clustering）并不利用层次关系聚类[5]。

Single-Pass 聚类的类别生成是通过比较新的数据与先前所有簇的相似程度。如果足够相似，那么新的数据将被聚到先前的簇中；如果不够相似，则作为新簇。Single-Pass 被广泛用于 TDT 中的新事件检测的任务中，Ron Papka 提出了将 Single-Pass 聚类用于在线的新事件检测中[6]，每一个类用簇的平均值作为质心向量来表示。Hila Beckert 提出一种集成方法，从文档结构中提取标题词，描述词，地点和时间作为特征来进行 Single-Pass 的聚类，侧重相似度的度量，提出了归一化互信息方法进行评价[7]。

4.2 主要步骤和任务

热点问题挖掘主要包括文本预处理，文本聚类，聚类评估和热点问题排行四个模块

图 4-1 总体模块流程

4.3 基于 Single-Pass 聚类算法的留言聚类分析

我们首先将文本进行预处理，并建立 VSM（Vector Space Model）向量，它能够以空间上的相似度表达语义的相似度，直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。然后运用 Single-Pass 算法对文本进行聚类分析。

Single pass 算法的基本流程是，首先将第一篇到达的文本设为种子文件，然后后面依次到达的文本与已有的文本进行相似度的比较，得到与先前文本最大的相似度，如果这个相似度大于给定的阀值，则将这个文本分配到与其相似度最大的文本所在的话题类中去；如果将此文本与所有的已存在文本比较，其相似度都

小于给定的阈值，则以此文本建立一个新的聚类[11]，该算法流程如图 4-2 所示

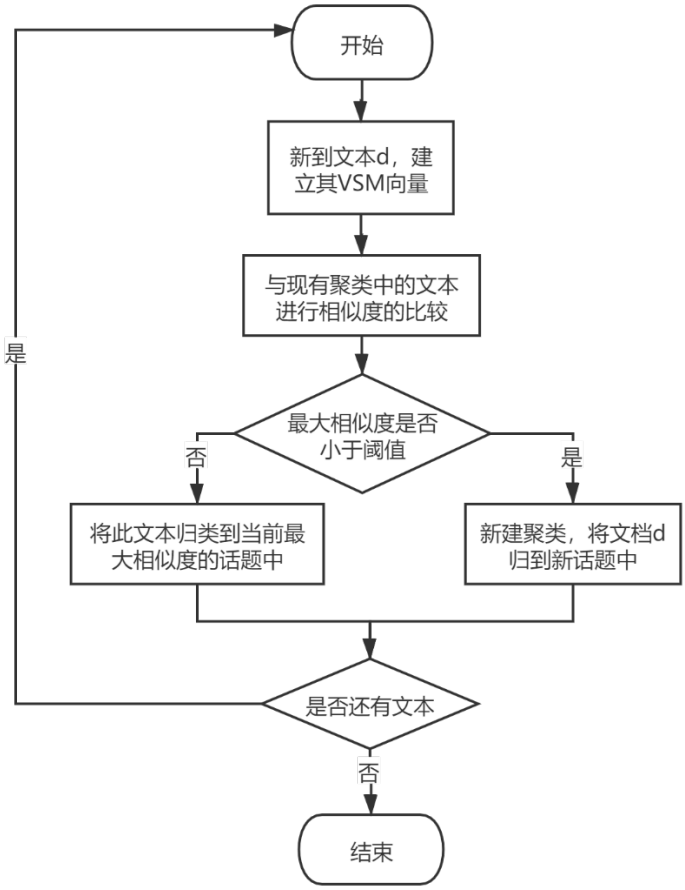

图 4-2 Single-Pass 聚类算法流程图

通过聚类，我们可以获得每条回复所属的话题（label 列），部分聚类结果如图 4-3 所示：

	title	time	content	up	down	text	content_cut	label
0	A3区一米阳光婚纱艺术摄影是否合法纳税了?	2019/2/28 11:25:05	\r\n\t\t\t\t\t\t\r\n\t\t\t\t\t\t座落在A市A3区联丰路米兰春天G2栋...	0	0	A3区一米阳光婚纱艺术摄影是否合法纳税了? A3区一米阳光婚纱艺术摄影是否合法纳税了? A3区一...	['一米阳光', '婚纱', '艺术摄影', '合法', '纳税', '一米阳光', '婚纱...	0
1	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	\r\n\t\t\t\t\t\t\r\n\t\t\t\t\t\tA市A6区道路命名规划已经初步成果公...	1	0	咨询A6区道路命名规划初步成果公示和城乡门牌问题咨询A6区道路命名规划初步成果公示和城乡门牌...	['咨询', '道路', '命名', '规划', '成果', '公示', '城乡', '门牌...	1
2	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	\r\n\t\t\t\t\t\t\r\n\t\t\t\t\t\t本人系春华镇金鼎村七里组村民，不知是...	1	0	反映A7县春华镇金鼎村水泥路、自来水到户的问题反映A7县春华镇金鼎村水泥路、自来水到户的问题...	['春华', '金鼎村', '水泥路', '自来水', '到户', '春华', '金鼎村',...	2
3	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	\r\n\t\t\t\t\t\t\r\n\t\t\t\t\t\t靠近黄兴路步行街，城南路街道、大古道...	1	0	A2区黄兴路步行街大古道巷住户卫生间粪便外排A2区黄兴路步行街大古道巷住户卫生间粪便外排A2...	['黄兴路', '步行街', '古道', '住户', '卫生间', '粪便', '外排', ...	3
4	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	\r\n\t\t\t\t\t\t\r\n\t\t\t\t\t\tA市A3区中海国际社区三期四期中间，...	0	0	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民A市A3区中海国际社区三期与四期中间...	['市区', '中海', '国际', '社区', '四期', '空地', '施工', '噪音...	4
...	...	...	...	...	...	...	...	...
4321	A市经济学院寒假过年期间组织学生去工厂工作	2019-11-22 14:42:14	\r\n\t\t\t\t\t\t\t\t\t关于西地省A市经济学院寒假过年期间组织学生去工厂...	0	0	A市经济学院寒假过年期间组织学生去工厂工作A市经济学院寒假过年期间组织学生去工厂工作A市经济...	['经济', '学院', '寒假', '期间', '组织', '学生', '工厂', '工作...	254
4322	A市经济学院组织学生外出打工合理吗?	2019-11-05 10:31:38	\r\n\t\t\t\t\t\t\t\t\t一名中职院校的学生,学校组织我们学生在外边打...	0	1	A市经济学院组织学生外出打工合理吗? A市经济学院组织学生外出打工合理吗? A市经济学院组织学生...	['经济', '学院', '组织', '学生', '外出', '打工', '经济', '学院...	254

图 4-3 部分聚类结果

4.4 基于轮廓系数的聚类效果评估

轮廓系数（Silhouette Coefficient），是聚类效果好坏的一种评价方式。最早由 Peter J. Rousseeuw 在 1986 提出。它结合内聚度和分离度两种因素。可以用来在相同原始数据的基础上评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。评估方法如下：

（1）计算样本i到同簇其他样本的平均距离a_i。a_i越小，说明样本 i 越应该被聚类到该簇。将a_i 称为样本i的簇内不相似度。簇 C 中所有样本的a_i均值称为簇 C 的簇不相似度；

（2）计算样本i到其他某簇C_{ij} 的所有样本的平均距离b_{ij}，称为样本i与簇C_j的不相似度。定义为样本i的簇间不相似度：$b_i = min\{b_{i1}, b_{i2}, b_{ik}\}$，$b_i$越大，说明样本$i$越不属于其他簇；

（3）根据样本i的簇内不相似度a_i 和簇间不相似度b_i ，定义样本i的轮廓系数：

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases}$$

（4）判断：s_i接近 1，则说明样本i聚类合理；s_i接近-1，则说明样本i更应该分类到另外的簇；若 s_i 近似为 0，则说明样本i在两个簇的边界上。

我们运用轮廓系数法对聚类结果进行评估，使用 sklearn 的 silhouette_score 方法实现，结果如图 4-4 所示：

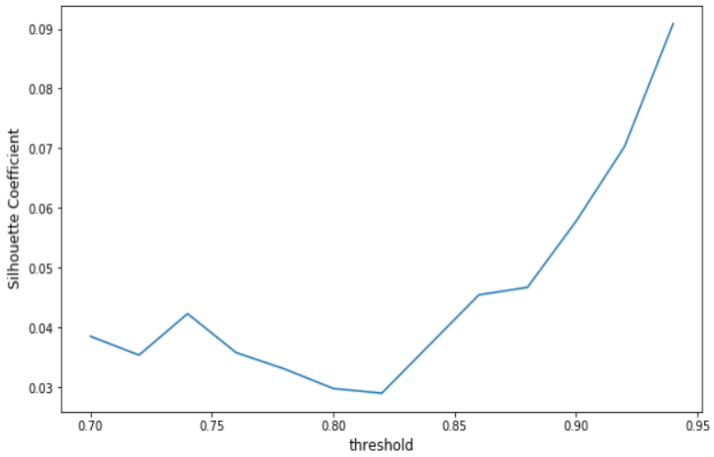

图 4-4 聚类评估结果图

由图 4-4 知，随着阈值的提高，轮廓系数也相应的上升，聚类结果也有着大幅提升，因此我们重新选择较大的阈值来获取所需聚类结果。

4.5 基于 TextRank 的自动文摘算法

TextRank 是一种基于图的用于文本的排序算法，基本思想来自于 Google 的 PageRank 算法[12]。类似于网页的排名，对于词语可得到词语的排名，对于句子也可得到句子的排名，所以 TextRank 可以进行关键词提取，也可以进行自动文摘。其用于自动文摘时的思想是：将每个句子看成 PageRank 图中的一个节点，若两个句子之间的相似度大于设定的阈值，则认为这两个句子之间有相似联系，对应的这两个节点之间便有一条无向有权边，边的权值是相似度，接着利用 PageRank 算法即可得到句子的得分，把得分较高的句子作为文章的摘要。TextRank 算法的主要步骤如下：

（1）预处理：分割原文本中的句子得到一个句子集合，然后对句子进行分词以及去停用词处理，筛选出候选关键词集；

（2）计算句子间的相似度：在原论文中采用如下公式计算句子 1 和句子 2 的相似度：

$$句子相似度=\frac{两个句子都出现的数目}{\log(句子1中词的数目)+\log(句子2中词的数目)}$$

对于两个句子之间的相似度大于设定的阈值的两个句子节点用边连接起来，设置其边的权重为两个句子的相似度；

（3）计算句子权重：

$$W_1=(1-系数)+系数\times\sum_{Join(1)}\frac{Similar(s1,s2)\times W_2}{\sum_{Join(2)}W}$$

其中，W_i表示第i个句子的权重，$Join(i)$代表与第i个句子相连的全部句子集合，$Similar(a,b)$表示a与b的相似度。由该公式多次迭代计算直至收敛稳定之后，可得各句子的权重得分；

（4）形成文摘：将句子按照句子得分进行倒序排序，抽取得分排序最前的几个句子作为候选文摘句，再依据字数或句子数量要求筛选出符合条件的句子组成文摘，效果如图 4-5 所示：

关键词	关键短语	摘要
业主 开发商 装修 交房 质量 房屋 施工 小区 宣传 设计 政府 建设 整改 现场 楼盘 ...	业主业主 开发商业主 质量业主 开发商开发商 质量 开发商 业主开发商 装修业主 装修质量 交...	反映A市万科金域蓝湾开发商相关违规问题 再次反映A市时代年华项目装修价质不符问题 反映A市万...
教育局 业主 小区 教育 学校 孩子 小学 入学 小孩 幼儿园 政府 开发商 学位 义务教育...	学校教育 学校学校 小区小区 学校小区	请解决A2区中建A1区嘉苑的业主小孩入学问题 请A3区协调解决旭辉御府业主子女入学问题 A2...
地铁 居民 泉塘 小区 出行 领导 公交线路 大道 号线 设置 线路 增加 开通 出口 建议...	居民出行 出行泉塘 地铁出行 小区小区 地铁泉塘 泉塘居民 小区居民 出行泉塘居民 泉塘地铁	反映A市地铁3号线松雅西地省站西北方向10万民众安全问题 反映A市地铁3号线松雅西地省站地下...
受害人 出借 立案 办案 诈骗 案件 派出所 平台 公安局 受害者 西地省 经侦 资金 有限...	立案立案	请A市公安局A2区分局依法调查余易贷诈骗一案 请A市A5区公安分局和候家塘派出所认真对待聚利...
业主 小区 社区 物业 小区业主 业委会 街道 开发商 投票 领导 街道社区 建局 成立 志...	业主业主 物业物业 业主社区物业 小区业主物业 业主小区 社区小区 业主社区 物业小区 社区...	A3区银盆岭街道办事处银双社区违规插手金色山庄小区内部事务 A3区银盆岭街道办事处、银双社区...
开发商 业主 交房 房产证 房屋 办理 办证 小区 房子 楼盘 政府 公司 产权证 购房 补...	业主房产证 业主业主 开发商业主 开发商房屋 交房 房产证 业主开发商 交房业主 开发商交房 ...	在A市江滨家园华庭苑买房八年才交房，还没有房产证 A市南郡明珠业主8年了都没有房产证 A市云...
车位 捆绑 职工 销售 广铁集团 购买 领导 铁路职工 认购 商品房 景园 滨河 强制 定向...	捆绑车位销售 销售车位 捆绑销售 车位捆绑销售 车位捆绑销售广铁集团 捆绑销售车位 销售车位...	伊景园滨河苑车位捆绑销售 投诉A市伊景园滨河苑捆绑车位销售 关于伊景园滨河苑捆绑销售车位的维权投诉
村民 土地 农民 户口 政府 尖山 社员 安置 市区 社区 征收 集体 政策 村干部 集资 ...	土地农民 土地土地 农民土地 村民土地	A市A4区沙坪街道汉回村支书纵容华颖实业集团侵占农民山林 A4区捞刀河街道高源村农民的土地被...
小区 医院 业主 居民 环境 变电站 影响 小区 业主 医疗 用地 医疗机构 眼科医院 建设 ...	医院小区医院 业主业主 医院小区 业主小区 小区居民 环境医院 小区医院	坚决反对在A7县诺亚山林小区门口设置医院 坚决反对在A7县诺亚山林小区门口设置医院 坚决反对...
业主 开发商 楼盘 政府 房地产 公司 建委 公寓 法律 规划 商业 企业 领导 消费者 购...	开发商业主	A市保利大都汇开发商虚假宣传，修改规划 投诉A市保利大都汇虚假宣传 质疑A市保利大都汇学校用...

图 4-5 摘要结果示例

4.6 热点问题排行

4.6.1 基于统计的总体问题的热点排行

热点问题(hot problem)的热门程度依赖于两个因素，一个是热门词语出现在留言中的次数，另一个是该词在多少个留言中出现。热点问题不可能一直热门，会随时间热门程度衰减，新的热点问题会出现。 因此，我们将热点问题定义为一个问题在一段时间内频繁出现。

根据对热点问题的定义，我们认为一个事件在一段时间内发生的次数越多，则其热度越高，即热度高的事件，留言的出现的频率也越高，同时，热度受到持续时间的影响也不应过大。因此，我们结合上述分析给出热度指数的计算公式:

$$Hot = \frac{n}{N \times \log_{10}(T)}$$

其中 Hot 为事件的热度指数， n 为该事件在留言中出现的条数， N 为总留言数，T 为事件的持续时间，当持续事件小于 10 天时，我们取 $T = 10$ 。

4.6.2 基于 TextRank 自动文摘算法的热点内部话题排行

利用 TextRank 算法进行热点内部话题排行的过程如图 4-6 所示:

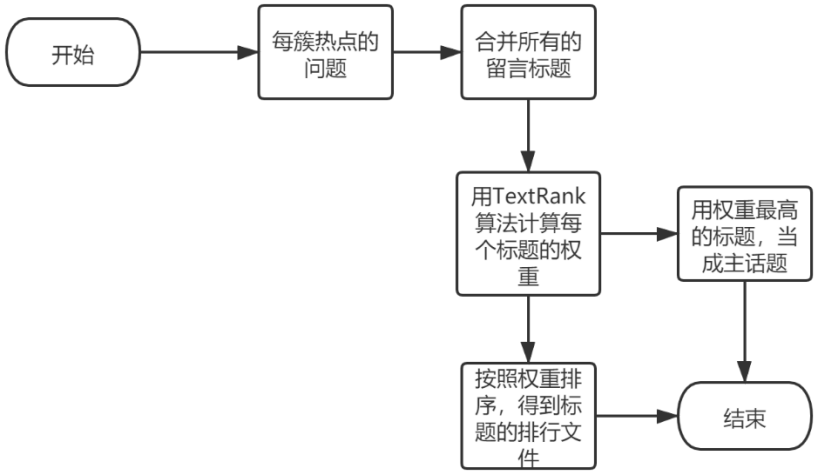

图 4-6 标题重要性排序

4.6.3 热点排行结果

热点问题排行结果如图 4-7 所示。

	hot_rate	topic_num	关键词	关键短语	摘要
0	0.004623	248	业主 开发商 装修 交房 质量 房屋 施工 小区 宣传 设计 政府 建设 整改 现场 楼盘 ...	业主业主 开发商业主 质量业主 开发商开发商 质量 开发商 业主开发商 装修业主 装修质量 交...	反映A市万科金域蓝湾开发商相关违规问题 再次反映A市时代年华项目装修价质不符问题 反映A市万...
0	0.003467	53	教育局 业主 小区 教育 学校 孩子 小学 入学 小孩 幼儿园 政府 开发商 学位 义务教育...	学校教育 学校学校 小区小区 学校小区	请解决A2区中建A1区嘉苑的业主小孩入学问题 请A3区协调解决旭辉御府业主子女入学问题 A2...
0	0.003467	13	地铁 居民 泉塘 小区 出行 领导 公交线路 大道 号线 设置 线路 增加 开通 出口 建议...	居民出行 出行泉塘 地铁出行 小区小区 地铁泉塘 泉塘居民 小区居民 出行泉塘居民 泉塘地铁	反映A市地铁3号线松雅西地省站西北方向10万民众安全问题 反映A市地铁3号线松雅西地省站地下...
0	0.003236	23	受害人 出借 立案 办案 诈骗 案件 派出所 平台 公安局 受害者 西地省 经侦 资金 有限...	立案立案	请A市公安局A2区分局依法调查余易贷诈骗一案 请A市A5区公安分局和候家塘派出所认真对待聚利...
0	0.003005	141	业主 小区 社区 物业 小区业主 业委会 街道 开发商 投票 领导 街道社区 建局 成立 志...	业主业主 物业物业 业主社区物业 小区业主物业 业主小区 社区小区 业主社区 物业小区 社区...	A3区银盆岭街道办事处银双社区违规插手金色山庄小区内部事务 A3区银盆岭街道办事处、银双社区...
0	0.003005	65	开发商 业主 交房 房产证 房屋 办理 办证 小区 房子 楼盘 政府 公司 产权证 购房 补...	业主房产证 业主业主 开发商业主 开发商房屋 交房 房产证 业主开发商 交房业主 开发商交房 ...	在A市江滨家园华庭苑买房八年才交房，还没有房产证 A市南郡明珠业主8年了都没有房产证 A市云...
0	0.002774	82	车位 捆绑 职工 销售 广铁集团 购买 领导 铁路职工 认购 商品房 景园 滨河 强制 定向...	捆绑车位销售 销售车位 捆绑销售 车位捆绑销售 车位捆绑销售广铁集团 捆绑销售车位 销售车位...	伊景园滨河苑车位捆绑销售 投诉A市伊景园滨河苑捆绑车位销售 关于伊景园滨河苑捆绑销售车位的维权投诉
0	0.002774	84	村民 土地 农民 户口 政府 尖山 社员 安置 市区 社区 征收 集体 政策 村干部 集资 ...	土地农民 土地土地 农民土地 村民土地	A市A4区沙坪街道汉回村支书纵容华颖实业集团侵占农民山林 A4区捞刀河街道高源村农民的土地被...
0	0.002774	92	小区 医院 业主 居民 环境 变电站 影响 小区业主 医疗 用地 医疗机构 眼科医院 建设 ...	医院小区医院 业主业主 医院小区 业主小区 小区居民 环境医院 小区医院	坚决反对在A7县诺亚山林小区门口设置医院 坚决反对在A7县诺亚山林小区门口设置医院 坚决反对...
0	0.002774	223	业主 开发商 楼盘 政府 房地产 公司 建委 公寓 法律 规划 商业 企业 领导 消费者 购...	开发商业主	A市保利大都汇开发商虚假宣传，修改规划 投诉A市保利大都汇虚假宣传 质疑A市保利大都汇学校用...

图 4-7 热点问题排行结果图

每条热点都通过热度指数大小进行排序，通过增加一 列“hot_rate”来代表问题的热度，该值越大，代表该类问题的热度越高，图 4-8 是对前四个热点问题所作的词云展示

图 4-8 热度最高的四个热点词云

5. 问题三：多角度回复评价模型

5.1 答复意见优劣的特征

对答复意见的评价是电子政务平台发展亟需解决的重要问题。回复的质量由很多种因素决定，必须从不同角度对其进行描述。回复质量的描述方法的优劣直接影响着回复评价的准确度。下面，首先对附件中回复数据进行预处理，剔除有问题的数据；然后从 4 个方面对回复质量进行量化描述。我们对附件中回复质量优劣的特征进行归纳，得到以下优劣评判特征，如表 5-1 所示。

表 5-1 参考特征

良好回复的参考特征	较差回复的参考特征
回复内容充实，有法律依据	内容空洞，没有逻辑
回复内容充实，信息量多	回复内容少，套用模板
处理高效，回复及时	处理效率低，回复间隔长

我们将基于这些特征利从不同角度对回复的质量进行量化描述，从而建立评价模型，实现对答复意见的智能评价。

5.2 回复质量的量化描述

① 回复内容是否有理有据

电子政务回复的内容需要有严密的逻辑和准确的表达，即回复信息是否有理有据。我们将有理有据分为运用法律法规和经过实际考证两个角度。首先，利用正则表达式将引用法律条文时出现的关键词和句式（如“根据 XXX 法第 XX 条”）和回复中的文本进行语义匹配。若在回复的文本中匹配到相应的内容，则认为该条回复引用了法律条文进行论述，记为 $E_{law}=1$。同样的方法，我们可以得到回复是否是经过实际考证和研究，若经过实际考证，则记为 $E_{act}=1$。据此，我们就可以得到“是否有理有据”的评价项，记为 F_1

$$F_1 = E_{law} + E_{act}$$

其中 E_{law} 和 E_{act} 初始值为 0。

② 回复内容是否充实

回复内容的详细程度与回复的文本长度有直接的关系。简短内容的回复信息量一般不够，评分应该较低；同时，较长文本的回复评分不应该过高。因此，可以考虑使用对数函数来量化回复的文本长度与评分的关系，建立“回复内容是否充实”的评价项 F_2

$$F_2 = \log_m L$$

其中 L 为该问题回复的文本长度， m 为常数。

③ 回复是否及时

对问题的回复速度反映了政府工作人员的工作效率与工作积极性，一般来说，回复和提问的时间间隔越长，评分应该较低；同时，较短时间的回复评分不应该过高。因此，可以考虑使用对数函数量化回复长度与评分的关系。通过对数据进行分析，我们得到平均回复时间约为 20 天，中位数为 11 天，据此可以认为回复天数小于 10 天的回复是较为及时的，其余回复应当给予相应的惩罚项。通过以上分析，建立“回复是否及时”的评价项 F_3

$$F_3 = \begin{cases} -\ln(\log_{10} D) & D > 10 \\ 0 & D \le 0 \end{cases}$$

其中 D 为该问题回复的时间间隔，函数图像如图 5-1 所示。

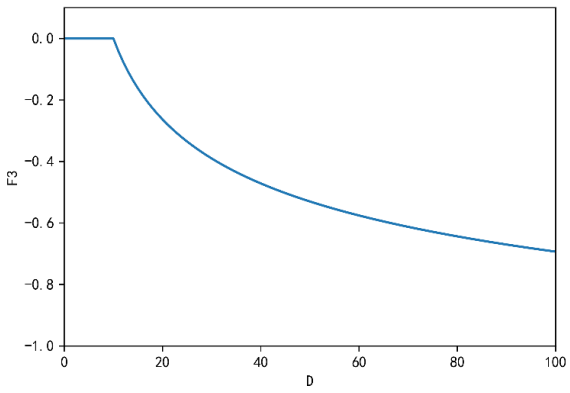

图 5-1 F_3 的函数图像

④ 回复是否相似度过高

好的回复应该针对性强，若对于多个不同的问题，回复的相似度过高，可以认为其采用了相同的模板或者复制之前的回复，表明该回复问题的质量不高，表达的信息量也少。为此，我们采用 DBSCAN 聚类算法，对回复内容进行聚类，找出相似度极高的回复内容，将其判定为套用模板。

DBSCAN 算法是一个基于密度的聚类算法，不需要提前确定聚类的数量，只需设定聚类的距离和每类的最小数量即可完成聚类任务。因此，相对于其他聚类算法，DBSCAN 非常适合用来实现对模板回答的聚类，避免了提前确定模板回答种数的困难，并且能够通过调整参数较好的实现对少量高相似度文本的聚类。

为了找出相似度高的模板回答，我们只将回复数据中涉及的用户名称、时间等特殊字段去除，其余内容不再进行停用词的去除和消歧操作。同时，为了防止将正常回复误判为模板回答，我们将 DBSCAN 的聚类距离设置为 0.3，每类最小数量设置为 8，保证聚类结果的合理性。聚类密度图如图 5-2 所示。

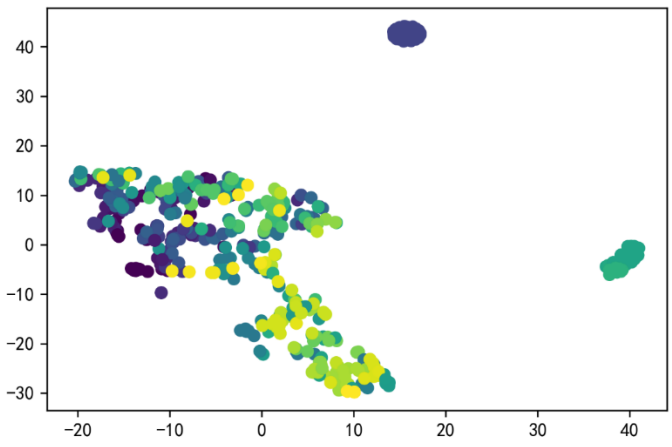

图 5-2 聚类密度图

从聚类密度图中我们可以看到聚类的效果较好，各个模板点的距离都很相近，符合模板回复的特点。其中一组聚类的文本数据如表 5-2 所示。

表 5-2 聚类结果展示

ID	回复时间	回复内容
3980	2019/1/4	您的留言已收悉。现将有关情况回复如下：目前，市规划局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，将充分论证研究……
4197	2018/11/27	您的留言已收悉。现将有关情况回复如下：对于您的意见和建议，市城乡规划局将充分论证研究……
4295	2018/11/14	您的留言已收悉。现将有关情况回复如下：目前，市城乡规划局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，将充分论证研究……
4504	2018/11/5	您的留言已收悉。现将有关情况回复如下：目前，市城乡规划局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，将充分论证研究……
6154	2018/5/16	您的留言已收悉。现将有关情况回复如下：目前，市规划局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，将充分论证研究……
7518	2017/11/13	您的留言已收悉。现将有关情况回复如下：你好，目前我局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，我局将充分论证研究……
8096	2017/9/1	您的留言已收悉。现将有关情况回复如下：首先，非常感谢您对 A 市规划建设的关心与支持，我局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，将充分论证研究……

有表 5-2 知，从 17 年到 19 年针对轨道交通的问题，回复均采用的同一套模板，没有实质性的信息，也没有能够对提问者产生帮助，此类回复评分应当较低。因此，我们将成功聚类的回复认为是模板回复，其余的作为正常回复，建立“回复是否及时”的评价项 F_4

$$F_4=\begin{cases}1 & \text{离散点}\\ 0 & \text{聚类点}\end{cases}$$

至此，我们完成了对四项评价指标的量化。

5.3 回复评价模型

下面对 4 项量化指标进行整合，以计算回复信息的综合得分情况，建立回复信息的质量评价函数 F，即构造如下回复 K 的最终回复质量的评价函数 $F(K)$：

$$\begin{cases} F(K) = M_K \cdot \lambda^{\mathrm{T}} \\ \lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4), M_K = (F_1, F_2, F_3, F_4) \end{cases}$$

其中，向量 λ 和 M_K 为反映回复信息不同侧面的权重向量和得分向量，λ^T 为向量 λ 的转置向量。

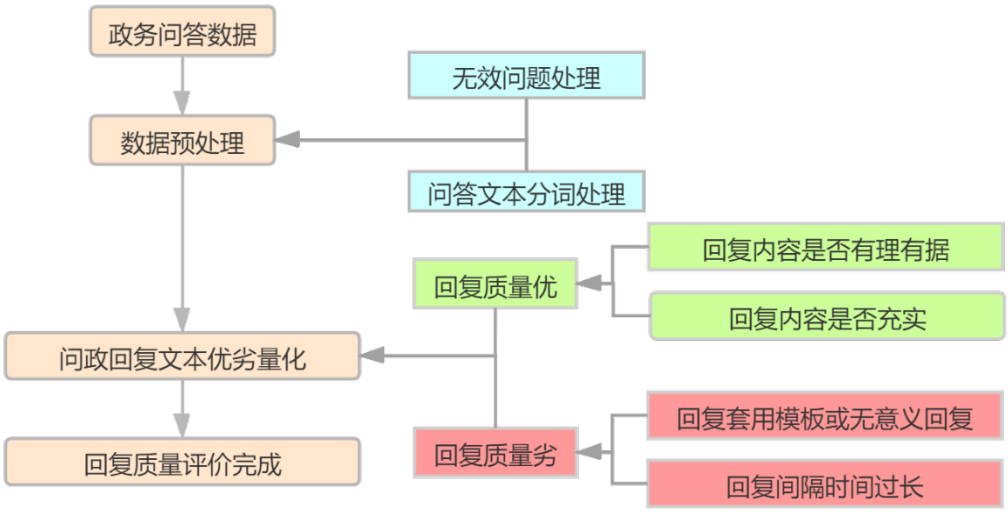

图 5-3 评价模型流程图

5.4 模型结果与分析

首先遍历每一个回复，计算 $M_K = (F_1, F_2, F_3, F_4)$，将权重向量设置为 $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$，根据有理有据、没有套用模板、回复及时、内容充实这 4 项判断回复质量的评分，而这 4 项对于评估一个回复质量的重要性依次减弱，因此本文给予这 4 项的权重分别为 0.35，0.3，0.2，0.15，这样分配的权重不仅满足权重之和为 1，也符合实际的权重分布。如此将回复质量权重取定后，就可以得到律师的回复质量评价函数 $F(K)$：

$$F(K) = 0.35 \times (E_{law} + E_{act}) + 0.2 \times F_4 + 0.2 \times F_3 + 0.15 \times \log_m L$$

把所有回复质量评价函数的 F 值映射到 $[0,1]$ 后得结果如表 5-3 所示。

表 5-3 部分回复的评价函数值

排名	留言编号	得分
1	99213	1
2	133336	0.971666
3	20726	0.944154
4	177915	0.942694
5	139949	0.937358
……	……	……
2813	74739	0.02723
2814	12451	0.025079
2815	12452	0.0244
2816	74851	0

图 5-4 给出了回复质量评分的分布，其中横坐标为回复质量函数值区间，纵坐标表示回复评分的概率密度。从图中可以看出，评分总体服从正态分布的形态，体现了本文基于网络回复的律师评价方法的合理性。

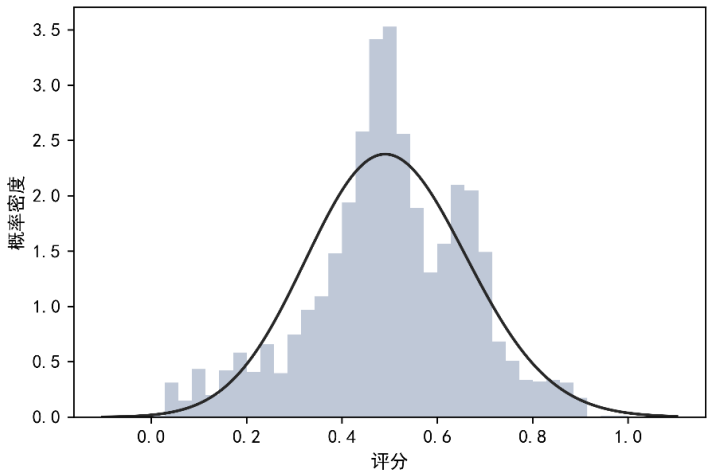

图 5-4 基于评分的回复分布图

表 5-4 和表 5-5 分别列出了部分评分排名靠前和评分排名靠后的回复，以验证模型的正确性。回复质量排名靠前的的回复对问题的回复有理有据，很好地解决了市民提出的问题，满足之前提出的各项标准。而排名靠后的回复对于

不同的问题均采取同一种无意义的回答，完全没有解答用户提出的问题。从这些回复中可以看出，对电子政务回复质量的评价符合之前制定的标准，也符合对回复质量的一般认识，验证了评价方法的准确性和合理性。

表 5-4 部分高评分回复内容展示

回复编号	回复内容
177915	留言标题：反映 G 市滨江中心物业管收费太贵等问题 2019 年 6 月 20 日 回复内容：您好！经调查处理，现将有关情况回复如下： 1、关于物业管理费 1.8 元/平方米。根据《西地省物业服务收费管理办法》第十条、十三条明文规定…… 2、关于物业企业收取 3000 元/户装修管理保证金及其他装修费用。根据《西地省物业服务收费管理办法》第二十三条…… 3、关于业主要求自己交水费问题……物业公司从未代收代缴。 4、关于小区绿化等配套建设。滨江中心城小区开发商根据规划批准，严格按照设计图纸进行施工建设。 2019 年 6 月 26 日
12452	留言标题：投诉 A7 县茶叶市场 B2 栋 155 门面通宵噪音扰民 2017 年 10 月 21 日 回复内容：您好！……收到投诉后，我局直属分局星沙中队执法队员即对投诉所称 B2 茶叶市场门面通宵营业扰民问题进行了现场调查……日常无店外经营行为。执法人员上户核查，进一步调查得知……证实投诉人反映的噪音问题是该店晚上机动车取货送货所产生的社会噪音。根据《中华人民共和国环境噪声污染防治法》第 58 条规定……执法人员告知该店负责人要求接货车辆晚间行驶时减速慢行，减少鸣笛，货物轻拿轻放 …… 2017 年 10 月 30 日

表 5-5 部分低评分回复内容展示

回复编号	回复内容
3980	留言标题：建议 A 市地铁 2 号线西延二期暂缓修建 2018 年 12 月 1 日 回复内容：您好！您的留言已收悉。现将有关情况回复如下：目前，市规划局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，将充分论证研究。感谢您对我们工作的支持、理解与监督！ 2018 年 12 月 29 日
4197	留言标题：关于 A 市 BRT 公交规划的咨询 2018 年 11 月 17 日 回复内容：您好！您的留言已收悉。现将有关情况回复如下：对于您的意见和建议，市城乡规划局将充分论证研究。感谢您对我们工作的支持、理解与监督！ 2018 年 11 月 21 日
118451	留言标题：K6 县君泰家园违规收取水电开户费费！请求处理 2019 年 12 月 30 日 回复内容：您好！您所反映的问题已收悉。就您所反映的问题我办已转交给相关部门研究处理。 在此，感谢您对我们工作的关心与支持。 2019 年 12 月 31 日
12451	请求易书记增加 A 市交通辅警工资或待遇 2014 年 3 月 11 日 网友：您好！留言已收悉 2014 年 4 月 28 日

实验结果表明，通过特征提取、量化描述和加权综合得到回复质量评价模型可以很好地评价一个回复的质量。该模型研究也有助于建设智能政务的网络

平台，促进优质和高效的问题回复，帮助政务网站提供更好的用户体验。

6. 基于自然语言处理的智慧问政分析平台搭建

6.1 系统简介

经过对问政文本分类、挖掘算法和评价模型的研究后，可以看到本文提出的算法有效可行。因此，为了帮助相关政府部门进行对问政数据的分析，提高问政平台的效率和水平，我们将本文研究的算法进行系统化的运用，搭建了基于自然语言处理的智慧问政分析平台。

6.2 系统框架和展示

本系统主要基于 B/S 架构，开发语言为 Python，采用了最新的 Streamlit 技术，实现对算法和模型的快速部署。系统主要分为文本评价、文本分类和热点挖掘三个功能区，图 6-1 所示是在本地搭建的系统 Demo。

图 6-1 回复评价界面示例

评价系统能够对输入的数据按照本文的预处理处理算法进行一键式预处理，并且自动计算处理后数据的各个指标的权值，最终呈现加权的总评分。

图 6-2 文本分类界面示例

文本分类系统与评价系统相似，此处不再赘述。通过提供数据分析系统，能够让对算法和代码完全不了解的工作人员实现对文本的快速处理，降低了使用门槛，提升了实用性。

7. 总结与展望

本文通过对问政文本数据进行分析，研究了多标签分类、热点事件挖掘和回复评价三方面的内容，并对所作的研究进行系统化的应用，实现了网页上的交互式操作，以帮助相关政府部门进行对问政数据的分析，提高问政平台的效率和水平。

随着电子问政的进一步发展，可供算法使用的数据量也会进一步增长。对于更高量级的数据可能会存在更出色的算法和模型，这就需要我们不断探索和进一步的研究。

参考文献

[1] J. Allan;J. Carbonell;G. Doddington et al. J. Allan. Topic detection and tracking pilot study: Final report. In Proceedings of Broadcast News Transcription and Understanding Workshop. Lansdowne, VA: NIST, 1998: 94-218

[2] J. Allan;V. Laverenko. First story detection in TDT is hard. Arvin Agah. In Proceedings of the 9th international conference on Information and Knowledge Management(CIKM). New York, NY, USA: ACM press, 2000: 374-381

[3] David M. Blei;Ng, Andrew Y. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022 [4] Xuerui Wang;Natasha Mohanty;Andrew McCallum. Group and topic discovery from relations and text. Jafar Adibi. Link KDD-2005 Proceedings of the 3rd international workshop on Link discover. New York, USA: ACM, 2005: 28-25

[5] C. D. Manning;P. Raghavan;H. Schiutze. Introduction to information retrieval. New York, USA: Cambridge University Press, 2008: 108-200

[6] Ron Papka, James Allan. On-line new event detection using single pass clustering. Technical Report: UM-CS-1998-021. MA, USA: University of Massachusetts Amherst, 1998: 20-30

[7] Hila Becker;Mor Naaman;Luis Gravano. Event identification in social media. Twelfth International Workshop on the Web and Databases (WebDB 2009), Providence, Rhode Island, USA: Association for the advancement artificial intelligence, 2009: 110-104

[8] J. Fiscus;G. Doddington;G. Kuhn. Topic detection and tracking evaluation overview. Topic Detection and Tracking - Event-based Information Organization. Norvell, MA, USA: Kluwer Academic Publisher, 2002: 17-31

[9]R. C. Swan and J. Allan. Automatic generation of overview timelines. Emmanuel
Yannakoudakis. In Proceedings of the 23rd ACM SIGIR International Conference on Research and Dvelopment in Information Retrieval. New York, NY, USA: ACM, 2000: 49-56

[10]G. Salton and C. Buckley. Term-Weighting Approached in Atuomatic Text Retrieval.
In: Information Processing and Management. Cornell University Ithaca, NY, USA:
Cornell University Ithaca, 1987: 513-523

[11] 基于改进 Single-Pass 算法的热点话题发现系统的设计与实现[D].张培伟 华中师范大学, 2015：28-32

[12] 蒲梅,周枫,周晶晶,严馨,周兰江.基于加权 TextRank 的新闻关键事件主题句提 取[J].计算机工程,2017,43(08):219-224.

[13] 杨开平，李明奇，覃思义. 基于网络回复的律师评价方法[J]. Computer Science, 2018, 45(9):237-242.

[14] 于书鳗. 网络问政平台的“回应性陷阱”[D].吉林大学,2019.

[15] 王文琳. 使用层次聚类和 N-gram 模型的新闻热事件检测研究[D].华中科技大学,2011.