基于文本挖掘的网络问政平台留言分析

摘要

随着互联网的迅速普及和移动互联网的兴起，网民数量和网站数量都在急剧增长，网络的社会影响在日趋扩大。其中，网络问政平台已成为政府汇集社情民意和民众发声谏言的重要渠道。面对海量的政务留言信息，单凭人工的力量很难进行全面的采集和整理。因此，如何通过计算机去自动采集、整理和分析海量的政务留言数据便具有重要的研究意义。

通过各类网络平台采集到的民众问政内容数据，蕴涵着大量值得挖掘的信息，能够对政府了解社情民意、聚焦热点民生问题以及民众监督政府服务工作提供有力的数据支撑与帮助，从而推动建设民主和谐中国。但是，由于民众问政留言内容的特殊性，采集到的数据往往存在着价值密度低、蕴含大量噪声和异常等问题。所以，本文将基于自然语言处理与文本挖掘技术对来自互联网公开来源的群众问政留言记录以及相关部门对部分群众留言的答复意见等近两万条政务留言内容数据进行严谨的预处理，提取我们需要进行分析的部分进行深度挖掘与分析。

针对题目中的具体问题，本文的解题思路如下:

首先, 本文通过长文本分析，jieba 分词，去停用词以及 pycorrector 错别字处理，对群众问政留言文本进行预处理，之后对题中三个问题进行进一步处理。

对于问题一群众留言的分类，首先采用 char 粒度将文本分字，然后利用 Tensorflow 创建字符嵌入，生成字向量以便做进一步深度监督训练。接下来通过 CNN 模型提取（留言文本向量）的卷积特征，在池化层连接所有特征后使用 Softmax 分类器输出留言对象对每个类别标签的概率，由此判断留言类别。最后，我们采用 F-Score 系数指标对模型进行评价。

对于问题二热度问题的挖掘，在利用 Word2Vec 模型完成对文本的词向量化后，结合命名实体识别算法 (NER) 和 TF-IDF 算法提取每个群众留言的关键词，然后通过余弦距离对文本信息进行相似度计算，再采用 DBSCAN 算法对文本进行聚类，选择最佳聚类簇数得到聚类结果。根据聚类得到的所有热点问题的相关特征，采用熵值法定量计算出问题热度并排序得到前五热点问题。

对于问题三中对答复意见的评价，首先构建答复意见的评价体系，根据数据集本身以及题目要求，我们设定了答复时限、内容相关性、形式完整性和详情可解释性四项评价指标。采用层次分析法 (AHP) 对指标的权重进行赋值，同时借鉴李克特表 (Likert) 的形式设置每项指标的评分标准。然后通过关键词匹配确定每条回复的所有指标值，最后计算综合评分，得出留言回复质量的评价结果。

关键字：jieba 分词; char-CNN; Word2Vec 词向量模型; DBSCAN 聚类; 层次分析法

Abstract

With the rapid spread of the Internet and the rise of the mobile Internet, the number of Internet users and the number of websites are growing exponentially, and the social impact of the Internet is growing by the day. Among them, the online questionnaire platform has become an important channel for the government to gather public opinion and the public to express their opinions. In the face of the huge amount of government message information, it is difficult to collect and collate a full range of information by human power alone. Therefore, how to automatically collect, organize and analyze the huge amount of government message data through the computer has important research significance.

The data collected through various online platforms contains a great deal of information that is worth digging into, and can provide strong data support and assistance to the government in understanding social and public opinion, focusing on topical livelihood issues, and monitoring government services by the public, thereby promoting the building of a democratic and harmonious China. However, due to the special nature of the content of the message, the data collected often has a low density of value, contains a lot of noise and anomalies and other problems. Therefore, this article will be based on natural language processing and text mining technology on the Internet public sources of the masses from the message records and related departments to some of the responses to the masses of the message of nearly 20,000 government message content data for rigorous pre-processing, to extract the parts of the analysis we need to carry out in-depth mining and analysis.In response to the specific questions in the topic, the solution of this paper is as follows:

First of all, we pre-processed the text of the message of the masses by long text analysis, jieba participle, de-deactivation words and pycorrector misspellings processing.

For the classification of the mass message in question one, the text is first divided into characters using char granularity, then use Tensorflow to create character embedding, generate word vectors for further in-depth supervision training. Next, the convolutional features (message text vectors) are extracted by the CNN model, and the probability of the message object to each category label is output using the Softmax classifier after connecting all the features at the pooling level, thus determining the message category. Finally, we evaluate the model using the F-Score coefficient indicator.

For the mining of the second heat problem, after completing the wordwise quantification of the text using the Word2Vec model, the combination of the Naming Entity Recognition (NER) and TF-IDF algorithm to extract the keywords of each mass message, then through the cosine distance to the text message similarity calculation, and then use the DBSCAN algorithm to clustering the text, and select the best clustering cluster number to obtain clustering results. Based on the relevant features of all the hot issues obtained by clustering, the entropy quantile is used to calculate the hotness of the issue and sort the top five hot issues.

For the evaluation of the responses in question 3, the evaluation system for the responses was first constructed, and four evaluation indicators were set, based on the data set itself and the topic requirements, namely, time frame, content relevance, formal completeness and explainability of the details. Analytic Hierarchy Process (AHP) was used to assign weights to the indicators and to set the scoring criteria for each indicator in the form of a Likert scale. All indicator values for each response were then determined by keyword matching, and a composite score was finally calculated to yield an evaluation of the quality of the message responses.

Keywords: jieba word-participle; char-CNN; Word2Vec; DBSCAN clustering; AHP

目录

1 挖掘目标

在网络问政兴起的大背景下，传统网络媒体、新兴的社会化媒体、政府网站纷纷推出不同形式的网络问政平台。

本次建模目标是利用某政务网站上发布的实际群众留言及政务人员回复信息数据，其中包括结构化和非结构化文本数据，在对数据进行基本的预处理包括长文本分析、中文分词以及停用词过滤后，首先根据题目给出的政务分类标签训练卷积神经网络，对留言内容进行分类和预测；其次用题目所给留言数据训练 word2vec 词向量，从而得到每条文本的向量表示以及文本之间的相似度，再采用 DBSCAN 聚类方法得到热点问题簇，通过熵值法筛选出前 5 个群众热点反映问题；接着，对政务人员的答复给出四个评价指标，通过李克特量表和层次分析法为每条回复的质量进行评价。最后结合以上三个问题的解答，对群众填写留言申请和政务人员的答复提出参考意见。

2 分析方法和步骤

总体流程图：

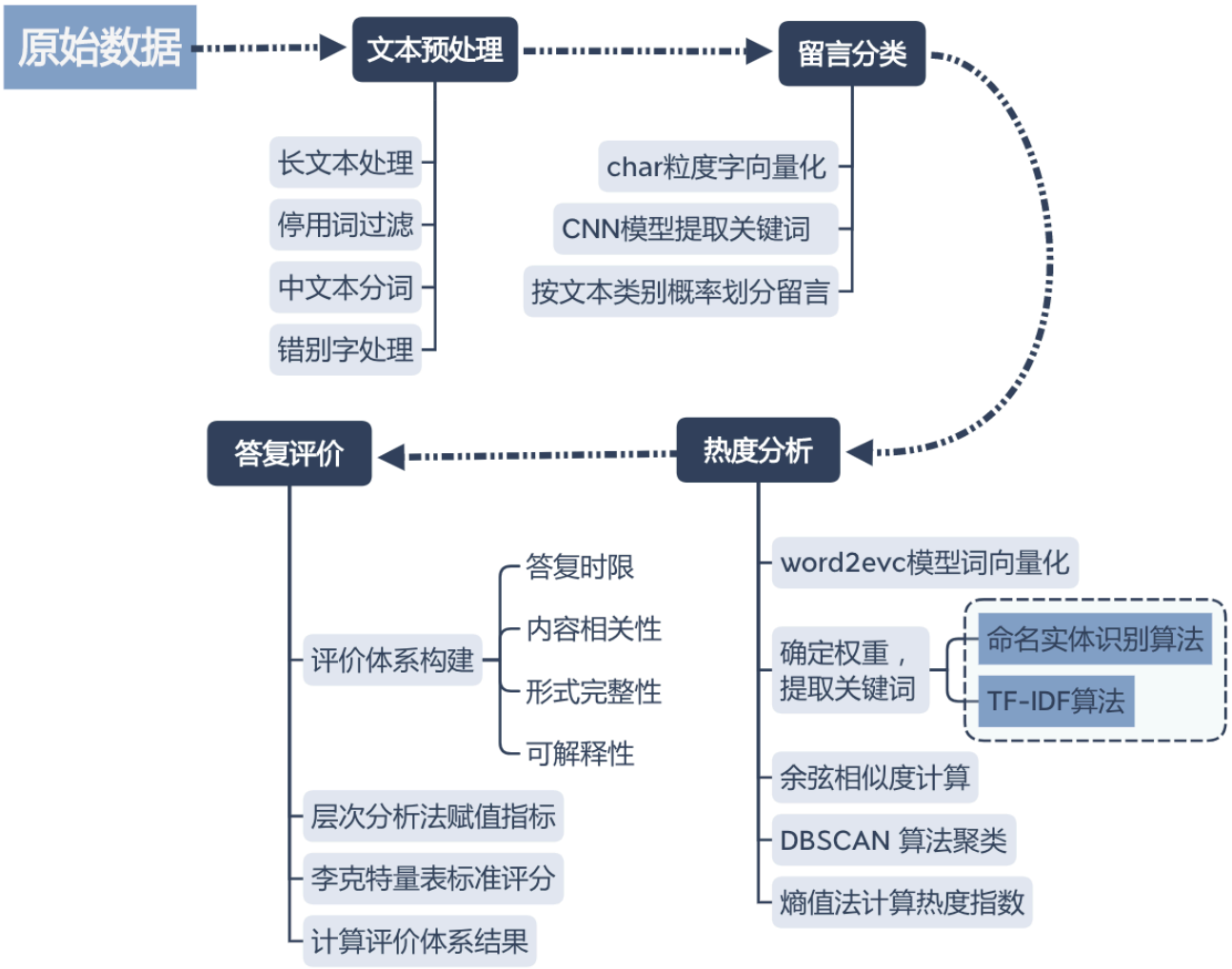

图 1: 总流程图

步骤一：文本预处理。对附件 2 结构化文本数据数值化处理，将留言主题和留言详情相结合，并进行长文本处理、文法纠正和错别字与不规范用语等噪音处理，而后进行中文文本分词，停用词过滤，以便后续分析。

步骤二：文本分类。利用 char 粒度将文本分字后，再通过 Tensorflow 转化为字向量，由此每段留言能构造出一个二维矩阵。接着，基于 CNN 模型，通过具有多个过滤器的卷积层，提取输入层（即字矩阵）的多层次特征，结合附件 1 给出的三级标签划分体系，根据输出结果对每段留言进行归类。

步骤三：热度分析。首先将留言文本转化为词向量，通过 TF-IDF 算法提取文本关键词，并结合使用 TF-IDF 和命名实体识别确定每个关键词的权重，得到文本向量后进行两两余弦相似度计算以及 DBSCAN 完成留言的聚类。根据聚类成的热点问题所有的特征，构建热度评价指标，利用熵值法计算每个问题的综合得分，排名得出热点问题。

步骤四：答复评价。从答复的时限、内容相关性、形式完整性以及可解释性四个方面，建立答复意见的评价体系，再利用层次分析法设置每个指标的权重，按照李克特量表的形式设定指标评分的标准。

3 文本预处理

3.1 数据描述

近年来，互联网、微博等新兴媒体形式的普及给予了公共讨论更大的空间，越来越多的普通民众开始通过网络表达各种意见和诉求。一方面，公民权利意识在更开放的信息环境中逐渐觉醒，公众参与社会热点、基层治理的热情和需求增加；另一方面，政府、官员也愈来愈重视民意在网络中的表达，并通过各种途径对其进行回应，网络问政的时代已经到来。

党的十七大、十八大报告强调，要保障公民的知情权、参与权、表达权和监督权，从各个层次、各个领域扩大公民有序的政治参与；健全权力运行制约和监督体系，加强党内监督、民主监督、法律监督、媒体监督，让权力在阳光下运行。网络问政的兴起，为公民参政议政提供了一个公平、透明、快捷的舆论交流平台，成为公民行使知情权、参与权、表达权和监督权的重要渠道，网络问政的制度化、规范化、常态化也成为推进我国民主建设的必然选择。

在网络问政兴起的大背景下，传统网络媒体、新兴的社会化媒体、政府网站纷纷推出不同形式的网络问政平台。如微信、微博、市长信箱、阳光热线等使政府能够更深入、更准确、更便捷地了解民意汇聚民智、凝聚民气。题目给出的数据是收集自互联网公开来源的各类社情民意相关的文本数据。由于此类文本的提供主体是人民群众，它与其他社交平台上的文本形式不同：

1. 文本长度不一。由于各人的表达习惯和问题反馈内容不同，留言文本长度有长有短。
2. 群众留言具有不规范性，其文法通常是非正式的，语言偏向口语化和生活化。
3. 留言文本常带有缩写、拼写错误、不规范用语、及表情符号等噪音，增加了政务人员对信息理解和事件发现的困难程度。同时因为文本是从互联网公开来源爬取的，存在许多网页信息如网址等无关信息。
4. 本内容中通常包含着显著的个人意图和明显的个人主义感情色彩，容易在当前主题中产生新主题，形成主题漂移或交叉，并呈现出“长尾现象”，最终导致数据分布出现严重的不平衡。

同时，通过观察所给数据，可以发现数据量比较大，且附件 2 中的字段大多为文本格式，故需要将其量化成数值形式才能对其进行分析。并且，针对于附件中留言内容存在大量噪声特征，如果把这些数据也

引入分词、词频统计乃至文本分类等，则必然会对分类结果造成很大的影响。于是，本文首先对数据进行预处理。

3.2 文本预处理

本文将文本数据的预处理分为四个部分：

3.2.1 长文本分析

观察附件中每条留言记录可知，一部分留言详情存在留言长度过长、不相关内容 (即噪音) 过多等问题，因此需要对长文本进行关键句提取来减少噪音，同时降低文本向量计算维度。

为了分别得到每条留言记录中留言详情的关键句，我们使用 Python 的 TextRank4ZH 模块包进行关键句提取。TextRank 算法是由网页重要性排序算法 PageRank 算法迁移而来：PageRank 算法根据万维网上页面之间的链接关系计算每个页面的重要性；TextRank 算法将词视为“万维网上的节点”，根据词之间的共现关系计算每个词的重要性，并将 PageRank 中的有向边变为无向边。而 TextRank4ZH 模块是针对中文文本的 TextRank 算法的 Python 算法实现，它能够从一个给定的文本中提取出该文本的关键词和关键词组，并使用抽取式的自动文摘方法提取出该文本的关键句。

而后，我们对文本长度超过 300 个中文字符的输入数据进行长文本分析处理，从每段数据中提取出 5 条关键句代替原来的长文。通过上述文本处理后，可以在保留文本主要特征和中心思想的前提下，以牺牲微量的文本信息为代价过滤掉大部分文本噪音，使得留言内容更加纯净。长文本分析示例如图 2 所示。

3.2.2 去停用词

停用词过滤是一个常见的预处理过程。

停用词指语言中一类没有实际含义的词或字，比如“的”，“甚至”，“不仅”，“吧”等。一个句子去掉了停用词并不影响理解。不仅提高计算速度，对留言分类的准确性也大有好处。因而大多数情况下不能成为搜索的关键词，因而创建索引时，这种词会被去掉而减少索引的大小。本文采用哈工大拓展停用词表，生成附件中文件，并基于此进行去停用词处理。

3.2.3 中文文本分词

在英文语句中，最基本的单位是单词 (word) , 单词与单词之间以空格作为天然的分隔符，而中文的最基本单位是汉字，但是单个汉字往往不能表达特定的意思，真正能表达语义的是若干个汉字组合成的词。因此，在使用计算机进一步处理之前，需要对中文语句进行分词使其分割成一个个基本的语义单位。

故本文在对政务留言信息进行挖掘分析之前, 先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件 2 中，以中文文本的方式给出了数据。为了便于转换，先要对这些政务留言信息进行中文分词。

本文采用 python 的中文分词包 jieba 进行分词。jieba 分词主要是基于统计词典，构造一个前缀词典；然后利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)；通过动态规划算法，计算得到最大概率路径，也就得到了最终的切分形式。而对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果 [2]。

对附件 2 中的前 5 条留言详情文本采用 jieba 分词后得到的结果如图 3 所示。

原留言详情	长文本分析后留言关键句
我住在梅溪湖壹号御湾 4 楼，自 2019 年 8 月份住进来每天晚上都会停水，白天水很小，到了用水高峰期就会水越来越小直至停水，要停到 12 点左右才开始有水。夏天很热的天一家人没洗澡就着汗液睡觉，现在天气逐渐转凉家里 70 多岁的老人想洗个热水澡要么停水要么水压小到打不亮燃气。这一个多月打了物业公司电话、自来水客服电话、社区投诉电话都没有用，据自来水工作人员说 4 楼以下水没有加压所以这样。我想咨询一下 A 市自来水公司的负责人，别的城市都已经一门心思用"智慧水务"、"一次都不用跑"等方式提升百姓生活体验，为什么 A 市的自来水公司还停留在像梅溪湖这样的新建小区连正常的生活用水都保障不了？而且多次反应没有效果，没有任何解决方案和计划。这样的工作方式是否与"让老百姓在宜居的环境中享受生活"理念背道而驰？这样的作风是否有违为人民服务的宗旨？恳请 A 市自来水公司的领导能够听到百姓的呼声，恳请各位领导能够体谅梅溪湖低楼层住户的心情，彻底解决用水难的问题。	我住在梅溪湖壹号御湾 4 楼，自 2019 年 8 月份住进来每天晚上都会停水，白天水很小，到了用水高峰期就会水越来越小直至停水，要停到 12 点左右才开始有水。 这一个多月打了物业公司电话、自来水客服电话、社区投诉电话都没有用，据自来水工作人员说 4 楼以下水没有加压所以这样。 而且多次反应没有效果，没有任何解决方案和计划。 这样的工作方式是否与"让老百姓在宜居的环境中享受生活"理念背道而驰？ 这样的作风是否有违为人民服务的宗旨？

图 2: 长文本分析示例

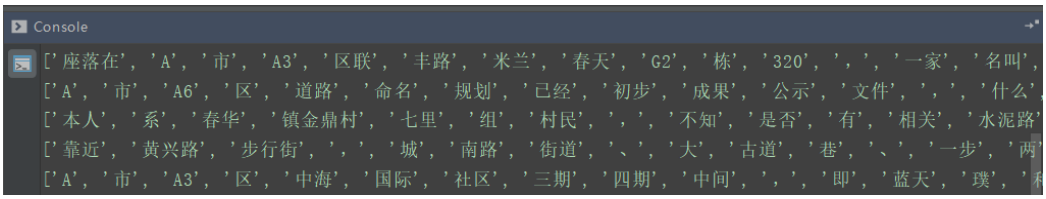

图 3: jieba 分词

将哈工大停用词表应用在该分词结果上后得到新的词语列表：

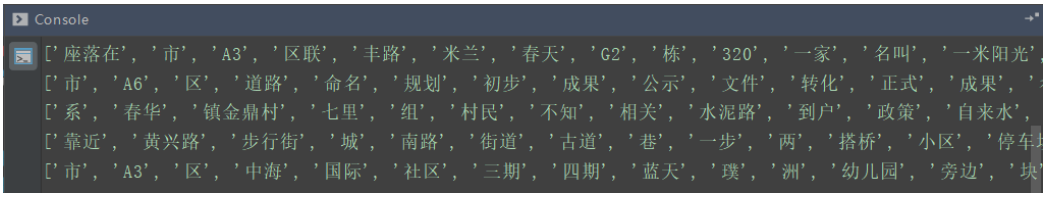

图 4: 去停用词

, |连 |摩托车 |都 |不能 |正常 |的 |通过 |, |城管 |局 |的 |领导 |就 |不 |能 |制定 |一个 |措施 |来 |制止 |这种 |形 |为 |吗 |? |

城乡 |建设 |K8 |县 |南门 |街 |干净 |整洁 |了 |几 |天 |, |又 |是 |老 |样子 |了 | | |南门 |街 |前段 |时间 |经过 |整改 |劝阻 |摆摊 |占 |道 |的 |情况 |改善 |了 |很多 |, |但是 |情况 |好 |了 |几 |天 |又 |慢慢 |的 |和 |以前 |一样 |了 |, |只要 |有 |人 |带头 | |后面 |慢慢 |又 |摆 |出 |来 |, | 很多 |商 |户 |现在 |干脆 |用 |钩子 |把 |一些 |货物 |挂 |门口 |屋檐 |下 |的 |电线 |上 |, |上 |有 |政策 |下 |就 |有 |对策 |, |城管 |来 |检查 |就 |稍微 |好 |点 |, |城管 |一 |走 |又 |摆 |出 |来 |又 |是 |老 |样子 |, |希望 |有关 |部门 |采取 |强硬 |点 |的 |措施 |, |每次 |都 |不痛不痒 |的 |整治 |一下 |根本 |起 |不 |到 |什么 |效果 |。 |现在 |二 |小 |门口 |那 |条 |路 |也 |成 |了 |马路 |市场 |了 |, |卖 |小菜 |. |卖 |鱼 |的 |. |卖 |水果 |的 |成堆 |了 |。 |
城乡 |建设 |K8 |县 |冷 |江东 |路 |蓝 |波 |旺 |酒店 |外 |墙 |装修 |无 |人 |施工 | | |现 |K8 |县 |冷 |江东 |路 |蓝 |波 |旺 |酒店 |前面 |的 |外 |墙 |装修 |搭 |着 |架子 |无 |人 |施工 |路政 |已 |在 |酒店 |门口 |搞 |了 |三 |个 |多 |月 |了 |严重 |影响 |了 |酒店 |的 |正常 |营业 |, |酒店 |找 |了 |施工 |队 |的 |人员 |了解 |情况 |什么 |时间 |可以 |搞 |好 |给 |我们 |一个 |正常 |的 |营业 |没有 |一个 |施工 |人员 |可以 |给 |我们 |答复 |, |酒店 |每月 |的 |房租 |就 |是 |三万 |多 |加上 |工资 |水电 |等等 |费用 |一个 |月 |就 |十 |几万 |酒店 |已经 |三 |个 |月 |没 |有 |正常 |营业 |了 |这些 |损失 |酒店 |已经 |无法 |? |承受 |了 |, |望 |有关 |部门 |弥补 |我们 |的 |的 |损失 |给 |我们 |一个 |正常 |的 |营业 |? | |? |
城乡 |建设 |K8 |县 |九亿 |广场 |的 |公厕 |要 |安装 |照明灯 | | |九亿 |广场 |是 |城区 |人民 |休闲 |娱乐 |的 |主要 |场所 |, |景观 |点 |也 |很 |漂亮 |, |每 |到 |晚上 |很多 |人 |到 |那里 |去 |玩耍 |. |但是 |唯有 |两 |个 |公厕 |却 |没有 |灯 |. |黑 |黑 |的 |, |有些 |就 |在 |外面 |附近 |大小便 |, |这样 |影响 |非常 |不好 |, |如果 |说 |灯 |不 |好 |管理 |, |完全 |可以 |和 |景观 |灯 |并网 |, |同时 |开关 |. |希望 |能 |及时 |解决 |好 |, |谢谢 |! |? |
城乡 |建设 |K4 |县 |石 |期 |市镇 |老 |农贸市场 |旁边 |的 |公厕 | (|早 |厕 |) |里面 |脏 |、 |乱 |、 |差 |石 |期 |市镇 |老 |农贸市场 |旁边 |的 |公厕 | (|早 |厕 |) |里面 |脏 |、 |乱 |、 |差 |, |臭气 |熏 |天 |, |老百姓 |上 |个 |厕所 |无从 |下脚 |。 |公厕 |长 |年 |无 |人 |管理 |, |漏 |雨 |, |已 |成 |危房 |。 |这 |座 |早 |厕 |, |不仅仅 |是 |气味 |难闻 |, |到 |了 |夏天 |蚊蝇 |乱 |飞 |, |更 |重要 |的 |是 |安全 |卫生 |隐患 |大 |。 |石 |期 |市 |要 |发展 |, |公厕 |要 |革命 |, |希望 |领导 |真抓实干 |, |给 |个 |确切 |的 |回复 |, |让 |广大 |民众 |用 |上 |干净 |卫生 |的 |公厕 |。 | (|希望 |网友 |们 |多 |多 |关注 |此 |问题 |。 |) |
城乡 |建设 |K |市 |域 |轨道 |交通 |规划 |建议 | | |李 |书记 |您好 |, |感谢 |您 |的 |阅读 |。 |十二五 |期间 |, |非 |省会 |地级 |市 |的 |轨

图 5: 部分分词结果示例

3.2.4 错别字纠正

网络平台上的政务留言内容均为群众自主留言获得。民众的文化水平程度不一，且在输入过程中可能受到输入法或习惯用语等因素影响，书面表达会出现别字、人名地名错误、拼音错误、知识性错误、用户发音或方言纠错、重复性错误以及口语化问题。其中几类别字错误类型展示如下：

错误类型	举例
别字	感帽，随然，传然，呕土
人名/地名错误	哈蜜（正：哈密）
拼音错误	咳数（ke shu）→ ke sou
知识性错误	广州黄浦（埔）
重复性错误	在 上 上面 上面 那 什么 啊
口语化问题	呃 ，啊，那用户名称是叫什么呢？

图 6: 别字错误类型示例

中文纠错一般分为两步走，第一步是错误检测，第二步是错误纠正。

- 错误检测部分: 先通过结巴中文分词器切词，由于句子中含有错别字，所以切词结果往往会有切分错误的情况，这样从字粒度和词粒度两方面检测错误，整合这两种粒度的疑似错误结果，形成疑似错误位置候选集；
- 错误纠正部分：遍历所有的疑似错误位置，并使用音似、形似词典替换错误位置的词，然后通过语言模型计算句子困惑度，对所有候选集结果比较并排序，得到最优纠正词。

本文使用 python 中的中文文本纠错工具 pycorrector 进行错别字处理。此模块包可进行音似、形似错字（或变体字）纠正，可用于中文拼音、笔画输入法的错误纠正。

pycorrector 依据语言模型检测错别字位置，通过拼音音似特征、笔画五笔编辑距离特征及语言模型困惑度特征纠正错别字。

3.3 文本表示

样本数据集以非结构化文本数据为主，需要通过建立合适的模型转化为适合计算机计算的结构化数据。在实际的操作当中，都对文本做了一种叫做“词袋”(bag of words) 的假设，也就是认为文本可由文本中的词的种类及其出现的频率来表示，而词与词之间的相对位置则可以忽略。这虽然对文本做了很大程度的简化，但是实验表明这种简化在绝大多数情况下都能取得良好的效果，因此是合理的。

文本表示 (Test Expression) 也称为文本特征表达。它不仅要求能够真实准确地反映文档的内容，而且要对不同的文档具有区分能力。目前常用的文本表示模型有向量空间模型、布尔模型和概率模型等。

3.3.1 文本向量空间模型

向量空间模型（Vector Space Model，VSM）最早是由 Salton 和 McGill 于 20 世纪 60 年代末提出的，是目前在文本挖掘技术中最常用的表示模型。

- 文档 (document)：通常是文章中具有一定规模的片段，如句子、句群、段落、段落组直至整篇文章。
- 项/特征项 (term/feature term)：特征项是 VSM 中最小的不可分的语言单元，可以是字、词、词组或短语等。每一个特征项将被嵌入成一个维度为 n 的向量，表示为 $t_i = (k_{i1}, k_{i2}, ..., k_{in})$；则整篇文本 D 被划分为

$$\boldsymbol{A} = \begin{bmatrix} lk_{11} & k_{21} & \cdots & k_{n1} \\ k_{12} & k_{22} & \cdots & k_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1n} & k_{2n} & \cdots & k_{nn} \end{bmatrix}_{n \times n}$$

- 项的权重 (term weight)：对于含有个特征项的文档，每一特征项都依据一定的原则被赋予一个权重，记为 $W = (w_1, w_2, ..., w_n)$，其中 w_i 表示第 i 个特征项在整个文档中的重要程度。
- 文档 D 可用它含有的特征项在每个对应维度上进行加权相加所表示，简记为 $D = AW^T$，即得到维度为 n 的向量 D。

向量空间模型由于其对文档出色的描述能力以及简单易操作等特点，成为文本表示领域最重要的模型之一，广泛应用于自动摘要、文本分类、文本聚类等领域中。

4 群众留言分类

网民留言数量较多，且涉及地区、内容各有不同，因此在处理之前需要对留言信息进行搜集整理。部分留言中存在所述事实不明、内容空泛、逻辑不通等现象，政府需要通过筛选整理出拟处理回复的留言，并根据不同的领域进行分类，方便下一步不同主管部门进行认领或主管部门分配任务。题目给出了统一分类体系，以便后续将群众留言分派给对应的政府职能部分进行处理。

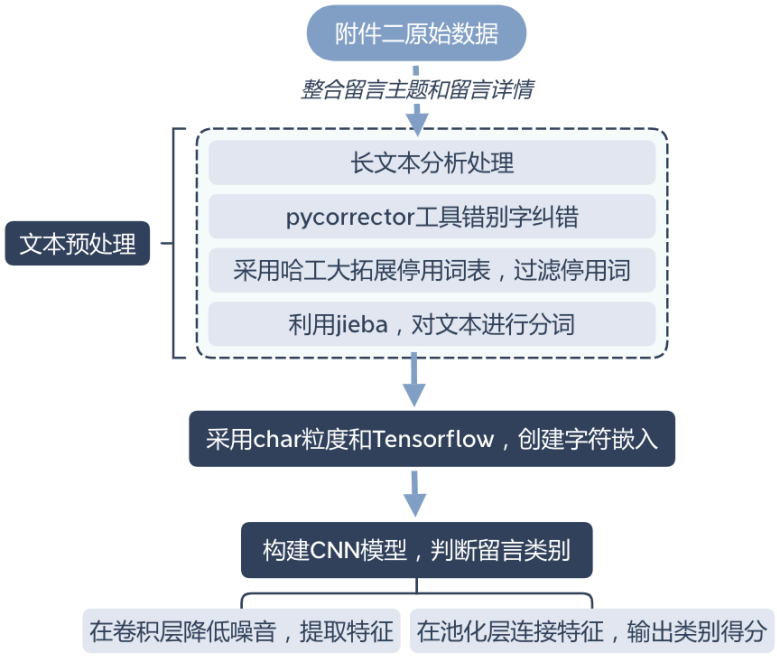

图 7: 问题一解决方案流程图

一级分类	二级分类	三级分类
城乡建设	安全生产	事故处理
城乡建设	安全生产	安全生产管理
城乡建设	安全生产	安全隐患
城乡建设	城市建设和市政管理	园林绿化环卫
城乡建设	城市建设和市政管理	城管执法
城乡建设	城市建设和市政管理	居民服务设施
城乡建设	城市建设和市政管理	城市公共设施
城乡建设	城市建设和市政管理	其他
城乡建设	城市建设和市政管理	公共汽车
城乡建设	城市建设和市政管理	公园管理
城乡建设	城市建设和市政管理	冬季采暖

图 8: 部分划分体系图片

4.1 char 粒度文本向量化

文本挖掘算法不能直接在原始文本形式上处理，所以我们需要对文本内进行处理，即将文本转化为更易计算机识别的信息，即对文本进行形式化处理。由于本题数据集中的样本主要由最广大互联网用户产生，文风千奇百怪，充斥各种语法错误和错别字，故本题采用 char 粒度（即字符粒度）对文本进行向量化处理。在自然语言处理 NLP 中，嵌入（Embedding）可以将离散输入对象转化为有用的连续向量。词嵌入把词转换成实数向量，字符嵌入把字符转换成实数向量。

本文在 TensorFlow 中创建字符嵌入。首先为汉字表中每个字分配一个整数，那么预处理分词过的文本就会形成相应的整数 ID 组合。接着为了将这些字 ID 映射到向量，我们创建嵌入变量并使用 tf.gather 函数进行嵌入，使得我们文本的每个汉字都能对应一个向量。同时经过训练后，在汉字表中所有的字的嵌入都能被学到，即包含在文本中的所有字符的嵌入。由此，我们得到了 char 粒度的向量化结果。

4.2 CNN 模型

卷积神经网络模型 (Convolutional Neural Networks,CNN) 在 1987 年由 Alexander Waibel 为首的科学家小组提出，是一类包含卷积计算且具有深度结构的前馈神经网络，近年来被广泛应用于自然语言处理领域，是深度学习的代表算法之一。卷积神经网络仿造生物的视知觉机制构建，可以进行监督学习和非监督学习，其隐含层内的卷积核参数共享和层间连接的稀疏性使得卷积神经网络能够以较小的计算量对格点化特征，如文字向量空间模型和像素进行学习。除了传统神经网络模型的输入层 (Input) 与输出层 (Output) 外，CNN 模型的网络拓扑结构还包含卷积层 (Convolutional layer)、池化层 (Pooling layer) 和全连接层 (Fully-connected layer)。

4.2.1 训练样本

在使用卷积神经网络进行建模分析时，我们采用有监督学习的方式对训练样本进行训练学习。在附件二中，对于每条留言记录，我们将留言主题与留言详情用字符串连接的方式拼接在一起形成一条新的数据。在对得到的数据进行上述错别字处理后，我们对文本长度超过 300 个中文字符的输入数据进行长文本分析处理，每段数据提取出 5 条关键句代替原来的长文，同时对长度不足 300 的文本在前面补 0，使得所有文本长度相同。通过上述文本处理后，可以在保留文本主要特征和中心思想的前提下，以牺牲微量的文本信息为代价过滤掉大部分文本噪音，使得留言内容更加纯净。

在本题的文本向量化模型中，我们将预处理后的每段留言记录作为卷积神经网络的输入，其对应的一级标签 (label) 作为输出来训练卷积神经网络。在本题中，我们为每个一级标签赋予一个互不重复的 ID，具体内容为 {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}。

4.2.2 CNN 模型结构

卷积神经网络的低层是由卷积层和子采样层交替组成，在保持特征不变的情况下减少维度空间和计算时间，更高层次是全连接层，其输入是由卷积层和子采样层提取得到的特征，最后一层是输出层，可以是一个分类器，采用逻辑回归、Softmax 回归，支持向量机等进行分类，也可以直接输出一个结果。一个完整的卷积神经网络除了输入层和输出层外应包括三个阶段：第一阶段为卷积层，第二阶段为探索层（即激活层），第三阶段为池化层。模型结构如下图 3-1 所示。

卷积层 卷积层和子采样层是特征提取的核心模块，与其他前馈神经网络类似，卷积神经网络采用梯度下降的方法，应用最小损失函数对网络中的节点的权重参数逐层递减，通过反向递推，不断调整参数使得损失函数结果逐渐变小，从而提升整个网络的特征描绘能力，即对每段留言文本类别判断的精准度和准确度不断提高。通过卷积层的运算，可以将输入信号在某一特征上增强，从而实现特征的提取，也可以排除干扰因素，从而降低特征的噪声。卷积层同时对其输入使用多个过滤器，使之能够检测到输入的多个特征。在本题留言文本分类的 CNN 模型中，我们使用的过滤器的宽度为整个文本向量的维度，过滤器的尺寸就是过滤器的高度，即过滤器一次性卷积的字符数。如下图所示：

激活层 卷积结束后，给与相应的激活函数。引入激活函数的主要目的是解决线性函数能力表达不够的问题。线性整流层作为神经网络的激活函数可以在不改变卷积层的情况下增强整个神经网络的非线性特征，不改变模型的泛化能力的同时提升训练速度。常见的卷积神经网络的激活函数有 Sigmoid、tanh、ReLU 函数。

池化层 池化层是一种向下采样的形式，在神经网络中也称之为子采样层。池化的结果是特征减少，参数减少，但其目的并不仅在于此。常用的池化方法有平均池化、最大池化、随机池化三种。池化后，参数量和运算量也会减少，同时也减少全连接的数量和复杂度，一定程度上可以避免过拟合。

全连接层 在 CNN 结构中，经多个卷积层和池化层后，连接着 1 个或 1 个以上的全连接层。全连接层中的每个神经元与其前一层的所有神经元进行全连接．全连接层可以整合卷积层或者池化层中具有类别区分性的局部信息．为了提升 CNN 网络性能，全连接层每个神经元的激励函数一般采用 ReLU 函数。

一个完整的文本分类的 CNN 模型如下图所示：

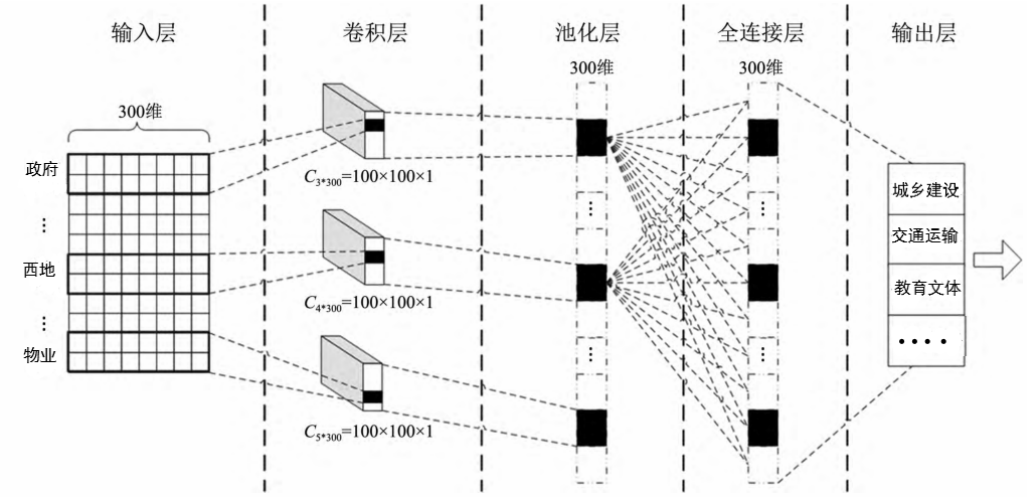

图 9: CNN 模型结构图

4.2.3 CNN 模型的构建

1) 本题中，我们令 CNN 模型直接从字符学习，无需任何预先训练的词向量嵌入。我们利用 Tensorflow 框架，通过 tf.getvariable 函数生成留言文本组成的汉字表中的所有汉字的向量，得到字向量表 vocab，其中每个字的向量维度为 128 维。这种向量生成方式对于特定分类任务而言具有较好的针对性。接着，我们将预处理后的附件二中的 9210 条留言数据作为输入层输入到字符级卷积神经网络模型 (Char-CNN) 中，每一条留言数据根据 vocab 中的字向量生成对应的文本向量，得到 9210×128 的矩阵。

2) 输入后第一层为卷积层，在输入的 9210×128 矩阵上，定义的滑动窗口卷积操作为：

$$c_i = f(\sum_{1}^{n-h} X_i W_i + b) \tag{1}$$

其中，n 代表 9210 条留言数据数量，X_i, W_i 的宽度相同，都是 28，前者代表输入矩阵的第 i 行到第 $i+h-1$ 行，后者表示整个 filter 各行的权重值。此处我们选择 filter 的尺寸为 3，即 X_i, W_i 的高度都为 3。在这里，b(bias) 是一个偏置,c_i 为标量。将一个 filter 卷积操作得到的结果拼接起来，可以组成一个 (n-1) 维的特征向量

$$c = (c_1, c_2, c_3, ..., c_{n-1}) \tag{2}$$

此处卷积层的激活函数 f 采用非线性激活函数

$$Relu(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } x >= 0. \end{cases} \tag{3}$$

其函数图如图 10 所示：

选择 ReLU 函数的原因是 ReLU 函数是非线性饱和函数，输出范围无限的，梯度下降速度快，训练时间更短。而 Sigmoid 和 tanh 函数是饱和线性函数，结果达到一定范围后不再变化。ReLU 则只需要一个阈

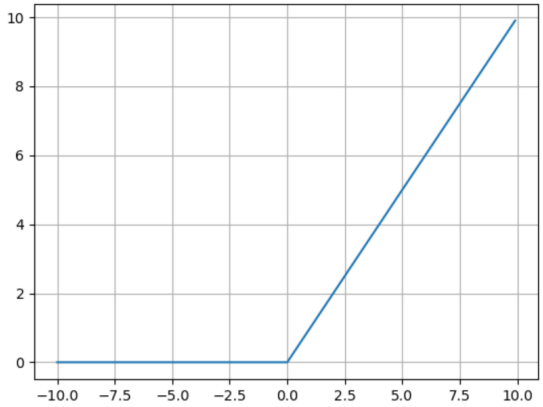

图 10: ReLu

值就可以得到激活函数，不需要对输入归一化来防止达到饱和。

3) 第三层为池化层，本题中，我们采用了 Max-pooling 池化方法，把上层中的每个 filter 产生的特征向量中选出最大的一个作为特征值来减少数据空间大小，即 $c^* = max\{c\}$。将每个 filter 生成的所有 c^* 输入到最后的全连接层，并使用 Softmax 分类器输出每个文本类别的概率。

4.2.4 参数设置

1、我们将附件二中每个类别的 80% 作为训练集，20% 作为测试集。

2、CNN 模型超参设置：

- 卷积核 (filter) 数目：256
- 卷积核尺寸 (size)：3
- 最大迭代次数 (epoch)：10
- 每批训练大小:32
- 显示中间结果的周期 (show)：20
- 学习率 (lr):1e-3

3. 将训练集样本数据输入到卷积神经网络中进行训练，每 20 步记录一次损失函数损失情况，统计结果如图 11。

4. 通过损失函数记录情况可以得到，损失率成梯度下降趋势，最终趋近于 0，说明模型的拟合性较强。

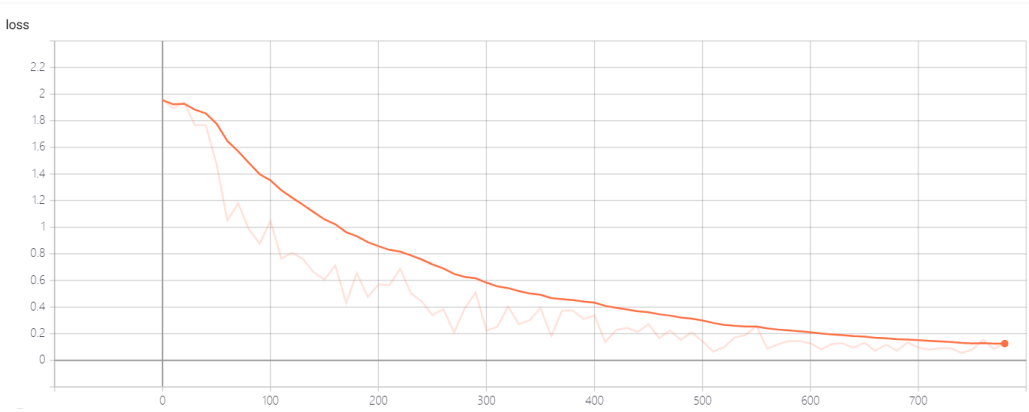

图 11: 损失函数

4.3 分类方法评价

F1 指标是综合平衡查准率和查全率的评价指标，能够全面的表示某一类别的分类或者聚类效果，而宏平均是通过对每一个类别的性能指标的进行算术平均得到的，代表的是某一种分类或聚类方法最终总体的一个的分类或聚类效果。

查准率是所有文本中与被正确分类的文本占分到该类的文本总量的比例，公式如下:

$$\text{查准率}\,(Precision) = \frac{\text{正确分类的文本数}}{\text{分到该类的总文本数}}$$

查全率是正确分到该类的文本占所有应该属于该类的文本数量的比例，公式如下:

$$\text{查全率}\,(Recall) = \frac{\text{正确分类的文本数}}{\text{属于该类的总文本数}}$$

查准率和查全率反映了聚类效果的两个侧面，但是二者往往存在此消彼长的关系，例如过高的准确率很可能造成较低召回率，而过高的召回率又很可能造成较低的准确率，因此二者必须综合考虑，这就产生了评价标准 F-score。

F-score 是根据查准率 (Precision) 和查全率 (Recall) 综合之后得出的，定义如下:

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

其中 P_i 是第 i 类的查准率，R_i 是第 i 类的查全率。

4.4 结果分析

4.4.1 F-score 评价指标

针对第一题所用模型的评价指标主要包括测试结果准确率和 F-Score。其中测试结果准确率（Test Accuracy）是最常见的衡量模型好坏的评价指标，即模型“预测”分类正确的样本数与总样本数之比。在本次测试中，模型测试结果准确率为 88.6%，结果值较高。但仅仅靠准确率来评价模型在一些极端场合下是无效的，例如共有 100 条留言，只有环境保护和教育文体两种类别，假设前者所有留言为 1 条，其余都为教育文体。训练得到的模型准确率可超过 99%，但显然该分类模型没有特征判别能力。

通过测试结果计算得到 F-Score 为 0.87。结合两项评价指标表现较好，说明从整体测试结果来看，分类模型的特征辨别能力较强，即模型能够较好满足分类需求。

4.4.2 实验结果分析

接下来具体分析测试结果中每个样本留言所涉及到的分类标签，其对应的查准率、查全率以及 F1-score 三个评价指标的数据如下表所示：

一级分类	Precision	Recall	F1-Score
城乡建设	0.85	0.83	0.84
环境保护	0.92	0.94	0.93
交通运输	**0.8**	**0.68**	**0.73**
商贸旅游	0.84	0.8	0.82
卫生计生	0.91	0.92	0.91
教育文体	**0.96**	**0.95**	**0.95**
劳动和社会保障	0.88	**0.95**	0.91

图 12: 结果示例表

查准率能反映模型分类的准确性，即是模型正确归类给某标签的留言数占模型归类给该标签的留言总数的百分比；而查全率则是反映模型分类的全面性，是指模型正确归类给某标签的留言数占系统中该标签实际所有的留言的百分比。理论上可以利用这两个指标对分类模型的性能水平进行评价。但是通过观察上表，发现除了极个别情况以外，几乎所有的标签对应查准率与其查全率都很高，且相差无几。因此需要通过两者的调和平均即 F1-Score 分值来衡量模型分类某个标签的有效性。

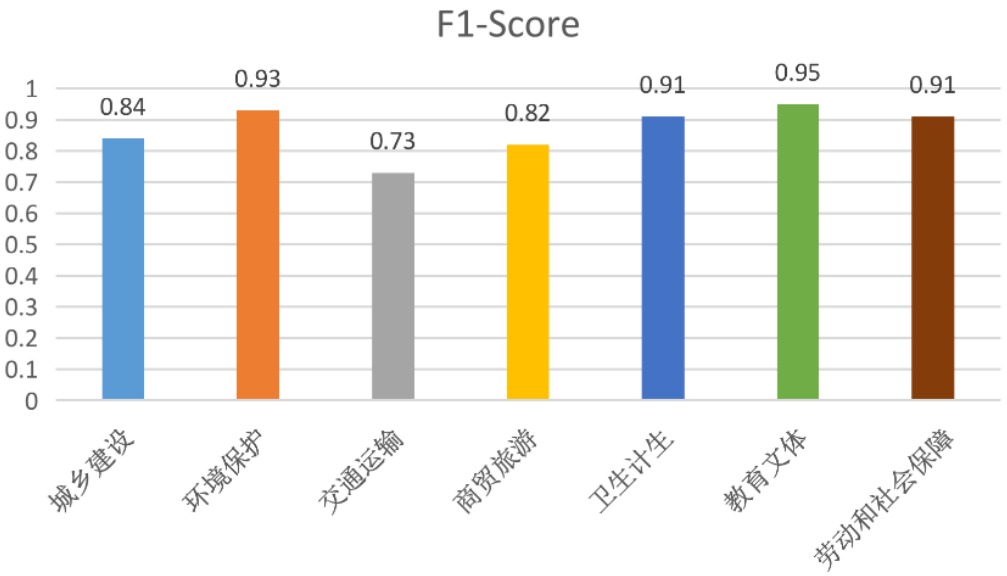

图 13: 实验数据统计图

由上图可知，交通运输标签所对应的 F1-Score 分值最低，为 0.73，其中查全率很低（0.68），其与查准率（0.8）相差较大。造成这种现象的原因可能是划分体系本身存在分类相交的问题，例如在附件一给出的“交通运输标签”三级分类中，“建设管理”类的“规划建设”、“设施维护”三级指标都分别与另一个标签——“城乡建设”里的“城市建设和市政管理”、“城乡规划”类的三级指标有容易混淆的地方，这就导致分类特征不清晰，分类模型误将涉及以上内容的留言归到了城市建设标签中，即使得交通运输标签的查全率

很低。同时这也可能是城市建设的查准率在所有标签中也是较低的原因。

特别地，教育文体标签较其他分类标签获得了较高的 F1-Score 分值，说明该类别标签匹配准确率最高，且当该类别的留言较少时也能被准确判别出来。分值同样较高的还有环境保护、卫生计生以及劳动和社会保障，均在 0.9 分以上，说明至少在定义这些类别时都划分得比较清晰，因此使得分类模型容易区分开来这些类别，保证了其查准率和查全率都在较高的水平。

5 热点问题挖掘

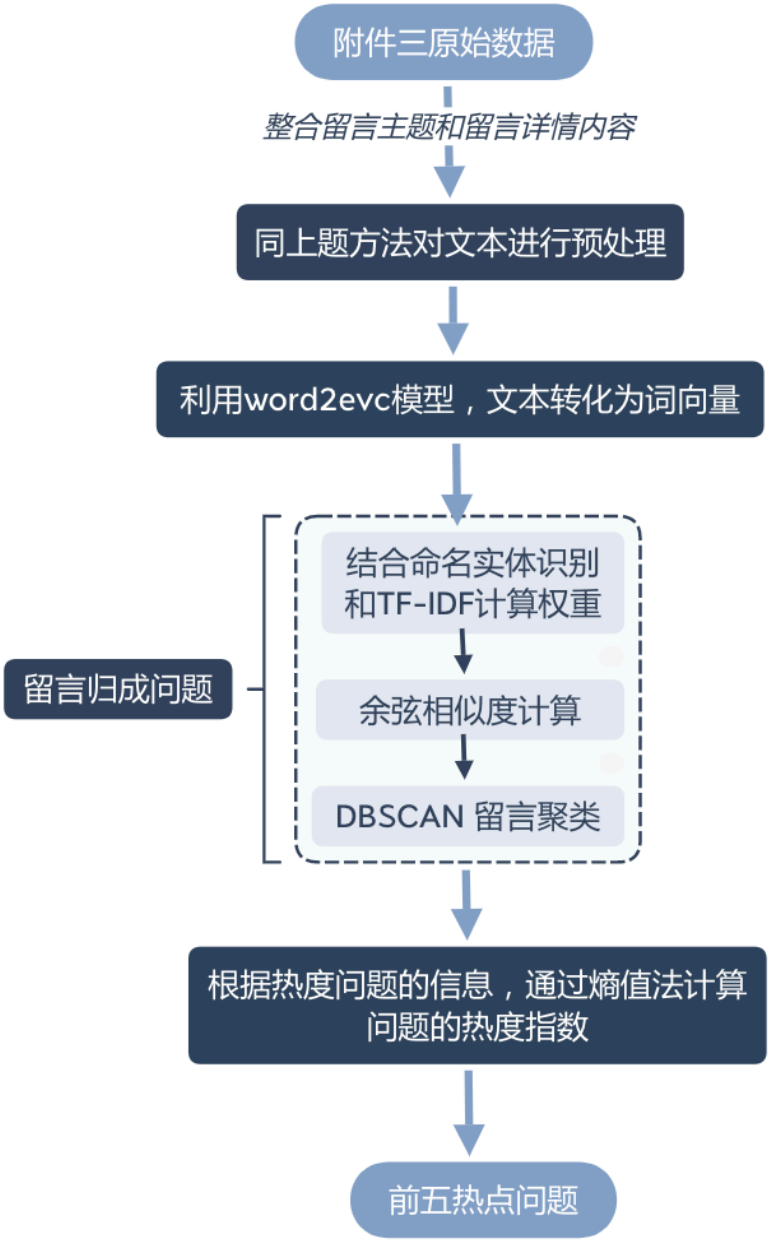

图 14: 问题二解决方案流程图

5.1 热点问题

某一时段内群众集中反映的某一问题可称为热点问题。对于群众问政进行热点挖掘，能够为政府相关部门及时了解社情民意并进行反馈、预警和引导提供必要的信息；同时利于对症下药，为群众排忧解难，提升政务服务效率。

热点话题挖掘与分析也吸引了国内外众多学者和科研人员的关注和研究。在国外，与之紧密相关的是话题检测与追踪 (Topic Detection and Tracking，简写为 TDT)”，它的目的是研究从大量新闻数据流中发现重要信息。1996 年 DARPA 首次提出了 TDT 的概念。TDT 的目标是研究如何在计算机自动地对新闻报道进行处理和组织。TDT 主要有以下几个方面的研究 [7]:

1. 报道切分: 将新闻报道流按照其描述的内容切分为不同事件或话题的报道。
2. 关联检测: 对两篇新闻报道进行比较以评价是否为阐述同一话题的相关内容。
3. 话题跟踪: 计算后续输入的报道与系统已存在话题的相似性，如果该报道属于某一话题则将此报道加入到该话题中。

本文从以上三个方面入手，对群众问政留言内容进行热点话题挖掘。

5.2 原始数据集预处理

将附件三中原始数据中留言主题和留言详情内容相结合，按照上题文本预处理使用的方法，采用哈工大拓展停用词表过滤停用词，再利用 jieba 对文本进行分词，最后进行错别字处理。

5.3 word 粒度文本向量化

关于文本向量化，考虑到本题需要对文本进行特征提取，而在问题一中使用的 char-CNN 模型不利于关键词抽取，并且会带来维度过高的问题。故在此题本文使用 word2vec 模型将文本转化为词向量。

5.3.1 词向量概念

将 word 映射到一个新的空间中，并以多维的连续实数向量进行表示叫做“Word Represention”或“Word Embedding”。自从 21 世纪以来，人们逐渐从原始的词向量稀疏表示法过渡到现在的低维空间中的密集表示。用稀疏表示法在解决实际问题时经常会遇到维数灾难，并且语义信息无法表示，无法揭示 word 之间的潜在联系。而采用低维空间表示法，不但解决了维数灾难问题，并且挖掘了 word 之间的关联属性，从而提高了向量语义上的准确度。

5.3.2 Word2Vec 词向量模型

Word2Vec 是”Word to Vector” 的简称，即通过神经网络机器学习算法来训练词向量。word2vec 工具是由 Google 团队在 2013 年发表的。word2vec 工具主要包含两个模型：跳字模型（skip-gram）和连续词袋模型（continuous bag of words，简称 CBOW），以在训练过程中求出 word 所对应的 vector（向量）。

由于 CBOW 模型和 skip-gram 模型互为镜像。故本文仅对 CBOW 模型进行建模说明。

输入层是由 one-hot 编码的输入上下文 $x1, ..., x_C$ 组成，其中窗口大小为 C，词汇表大小为 V。隐藏层是 N 维的向量。最后输出层是也被 one-hot 编码的输出单词 y。被 one-hot 编码的输入向量通过一个 $V \times N$ 维的权重矩阵 WW 连接到隐藏层；隐藏层通过一个 $N \times V$ 的权重矩阵 W 连接到输出层。接下来，我们假设我们已知输入与输出权重矩阵的大小。

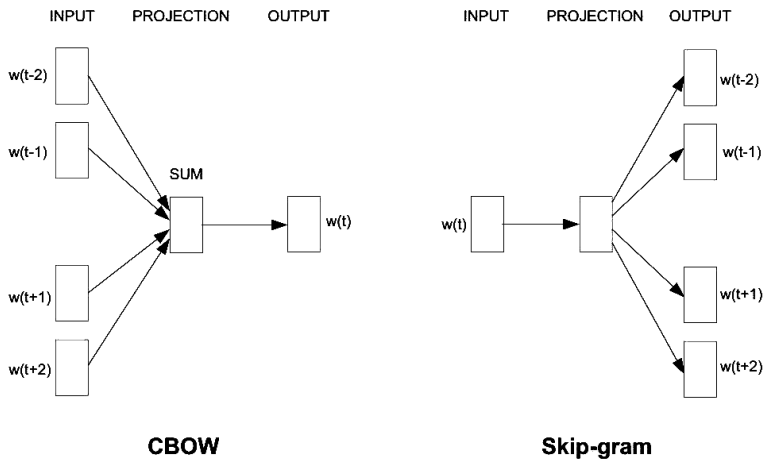

图 15: CBOW 模型和 skip-gram 模型示意图

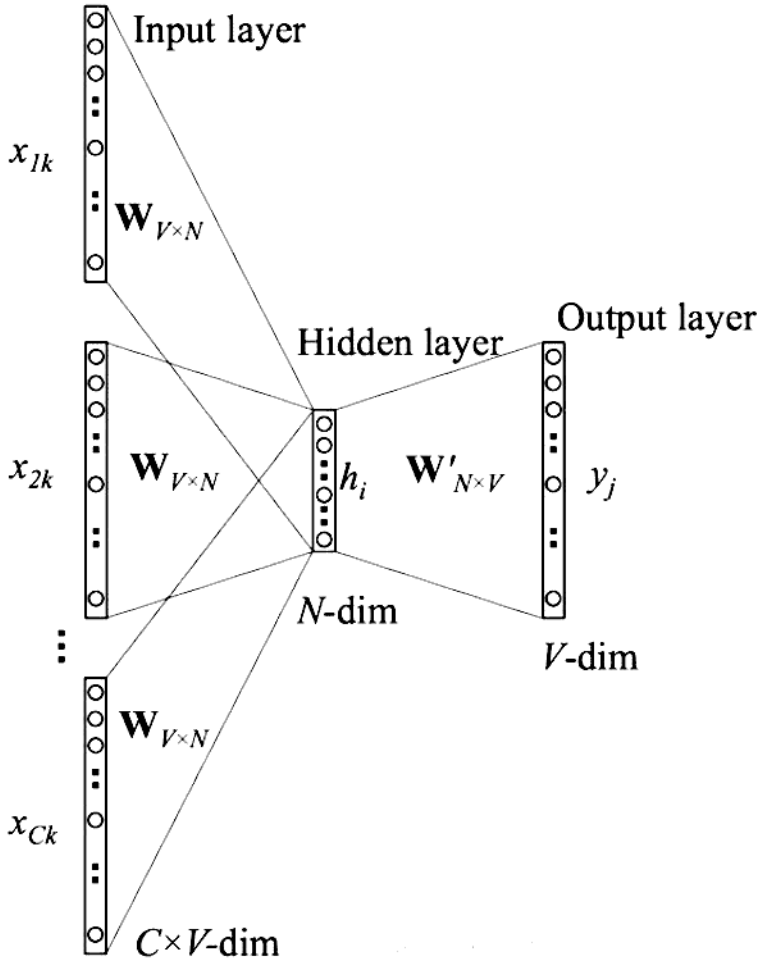

图 16: CBOW 神经网络模型

- 第一步：去隐藏层 h 的输出：

$$h = \frac{1}{C} W \left(\sum_{i=1}^{C} x_i \right)$$

此处输出就是输入向量的加权平均。

- 第二步：计算在输出层每个结点的输入：

$$u_j = v_{wj}^{'T} \cdot h$$

其中 $v_{wj}^{'T}$ 是输出矩阵 W' 的第 j 列。

- 第三步：计算输出层的输出，输出 y_j 如下：

$$y_{c,j} = p\left(w_{y,j}|w_1, ..., w_c\right) = \frac{\exp\left(u_j\right)}{\Sigma_{j'=1}^{V} \exp\left(u'j\right)}$$

而在学习权重矩阵 W 与 W 过程中，我们可以给这些权重赋一个随机值来初始化。然后按序训练样本，逐个观察输出与真实值之间的误差，并计算这些误差的梯度。并在梯度方向纠正权重矩阵。这种方法被称为随机梯度下降。但这个衍生出来的方法叫做反向传播误差算法。此处我们直接采用 Word2vec 包进行训练，不做赘述。由此利用 Word2vec 进行词语训练得到的词向量中包含丰富的词语语义语法信息及词语间相互关系信息。

在本题中，我们将附件三所有的数据作为 Word2Vec 语料库进行词向量训练。因此可对特定的文本聚类任务，训练一个词向量，以此保证训练的针对性和有效性。文档中部分词向量如下所示：

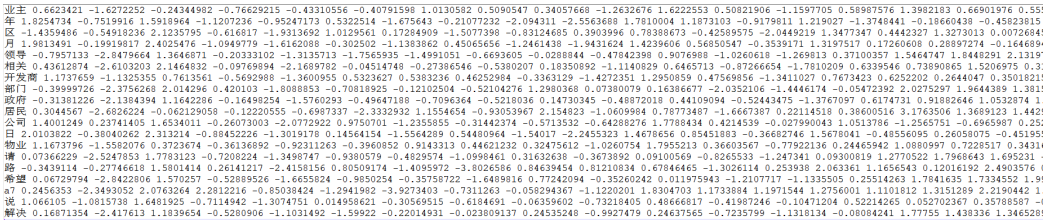

图 17: 词向量表

5.3.3 关键词提取

在对词向量训练完毕后，我们首先计算出所有词语的词频—逆向文档频率（简称 TF-IDF），并根据该值进行降序排列，取前 TF-IDF 值最大的前 10 个词语作为该段文本的关键词，即对文本进行降维处理。

词频 (Term Frequency, TF) 是词语在文本中出现的频率，如果某一个词在一个文本中出现的越多，它的权重就越高，基本公式:

$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{k,j}}$$

以上式子 $n_{i,j}$ 中是该词在文件 d_j 中的出现次数，而分母则是在文件 d_j; 中所有字词的出现次数之和。

逆向文档频率 (Inverse Documentation Frequency，IDF) 是指在少数文本中出现的词的权重比在多数文本中出现的词的权重高，因为在聚类中这些词更具有区分能力。它的基本公式如下:

$$idf_i = \log \frac{N}{|\{j : t_i \in d_j\}|}$$

其中，N 为语料库中的文件总数, $|\{j : t_i \in d_j\}|$ 为包含词语 t_i 的文件数目 (即 $n_{i,j} \neq 0$ 的文件数目)。如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用 $1 + |\{j : t_i \in d_j\}|$

在 Shannon 的信息论的解释中，如果特征项在所有文本中出现的频率越高，它所包含的信息熵越小；如果特征项集中在少数文本中，即在少数文本中出现频率较高，则它所具有的信息熵也较高。最后可以得出：

$$w_{ij} = tf_{ij} \times idf_i$$

5.4 权重计算

5.4.1 命名实体识别算法

命名实体识别 (Named EntitiesRecognition, NER) 是自然语言处理 (Natural LanguageProcessing, NLP) 的一个基础任务。其目的是识别语料中人名、地名、组织机构名等命名实体。由于这些命名实体数量不断增加，通常不可能在词典中穷尽列出，且其构成方法具有各自的一些规律性，因而, 通常把对这些词的识别从词汇形态处理 (如汉语切分) 任务中独立处理，称为命名实体识别。

由此我们可以得到一段文本向量的计算公式：

$$d(i) = \sum_{j=11}^{10} A_j W_j$$

在对词向量进行训练中，本文以关键词的 TF-IDF 权值与命名实体识别算法为依据来进行权重的确定，从而将 Word2Vec 模型进行加权求和，计算出整段文本的向量表示。即在一段文本里所有特征词进行筛选，如果正在进行向量化的词 (A_j) 是实体（如人名、地名等），我们将它的权重直接赋为最高值 1，即 $W_j = 1$；若不是实体，则取其 TF-IDF 值作为它的权重，即 $W_j = w_j$。

5.5 文本相似度

相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要研究文本个体将的差异大小，也就是需要对文本信息进行相似度计算，将根据相似特性的信息进行归类。当将文本用向量表示的时候，通常的方法是利用一些距离公式来得到两个文本向量之间的距离，据此得到两个文本之间的相似度。当将文本表示为两个 n 维向量，即文本 $A = A_1, A_2, ..., A_n$，$B = B_1, B_2, ..., B_n$ 时，为求文本 A、B 之间的相似度值比较常用的距离公式有如下几种:

(1) 余弦相似度通过向量空间中两向量之间的夹角余弦值大小来度量文本相似度，它侧重于两个向量的方向差异，取值范围为 [-1,1], 计算公式如下:

$$S(A,B) = \cos\theta = \frac{\sum_{k=1}^{n} A_k B_k}{\sqrt{\sum_{k=1}^{n} A_k^2}\sqrt{\sum_{k=1}^{n} B_k^2}}$$

而余弦距离就是用 1 减去余弦相似度获得的，取值范围为 [1,2]，即

$$dis(A,B) = 1 - S(A,B)$$

(2) 欧式距离也称欧几里得距离，衡量两个向量的绝对距离，公式如下:

$$dis(A,B) = \sqrt{\sum_{k=1}^{n} (A_k - B_k)^2}$$

考虑到余弦距离衡量的是维度间取值方向的一致性，注重维度之间的差异，不注重数值上的差异，而欧氏度量的只是数值上的差异性。故本文选择余弦距离进行两个文本向量之间的相似度比较。

5.6 热点簇选择

网络热点话题的发现主要是基于对语料进行聚类，也就是将待处理的语料按照其所描述的事件或话题分配到相应的类中，因此高效的聚类算法是网络热点话题发现的重中之重。与单纯的文本聚类不同，网络热点话题的发现需要考虑到语料的规模特征、动态增加的特征、时间的特征等。

5.6.1 DBSCAN 算法选择

通过前期观察数据可知，附件三中的留言的主题比较分散，离群值较多；且每一热点簇的相关留言数量较少，只有 15-45 条左右。基于附件三中所有 4326 条数据，将会产生一百多个热点簇。若使用常用的聚类方法 k-means 进行文本聚类，则需要指定聚类簇数 k，并且初始聚类中心对聚类影响很大。这是因为 k-means 把任何点都归到了某一个类，对异常点比较敏感。而 DBSCAN 能剔除噪声，需要指定邻域距离阈值 eps 和样本个数阈值 MinPts，可以自动确定簇个数。这样就能有效地对噪音较多的留言文本进行聚类，并且避免了因聚类簇数较多产生的聚类效果的问题。

5.6.2 DBSCAN 算法基本思路

DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）是一种基于密度的空间聚类算法。该算法将具有足够密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇，它将簇定义为密度相连的点的最大集合。

5.6.3 DBSCAN 算法具体实现

DBSCAN 算法步骤具体如下所示：

1. 寻找核心点形成临时聚类簇。

   扫描全部样本点，如果某个样本点 R 半径范围内点数目 $\geq$ 最少点数 MinPts，则将其纳入核心点列表，并将其密度直达的点形成对应的临时聚类簇。

2. 合并临时聚类簇得到聚类簇。

   - 对于每一个临时聚类簇，检查其中的点是否为核心点，如果是，将该点对应的临时聚类簇和当前临时聚类簇合并，得到新的临时聚类簇。
   - 重复此操作，直到当前临时聚类簇中的每一个点要么不在核心点列表，要么其密度直达的点都已经在该临时聚类簇，该临时聚类簇升级成为聚类簇。
   - 继续对剩余的临时聚类簇进行相同的合并操作，直到全部临时聚类簇被处理。

在本题中，我们将向量化后的留言详情文本数据作为 DBSCAN 聚类算法的样本点，考虑到相似文本在同一向量空间中的方向相近，我们使用余弦距离作为距离度量方式。为了确定参数 R 和 MinPoints 的最佳取值，我们首先计算出所有文本两两之间的余弦相似度，得到的余弦相似度矩阵如下所示：

在观察上述余弦相似度矩阵之后，通过非经验方法，我们针对本问题进行逐步尝试多组 R 和 MinPoints 的取值，通过轮廓系数 (Silhouette Coefficient) 来评价聚类结果的好坏。轮廓系数是衡量一个样本点与它所属簇相较于其他聚类的相似程度，其计算公式如下：

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

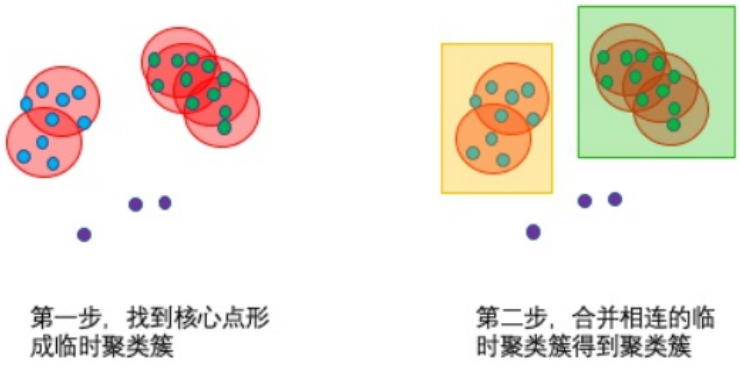

图 18: DBSCAN 算法实现步骤

余弦相似度矩阵：

```
[[1.         0.         0.         ... 0.         0.         0.        ]
 [0.         1.         0.03811413 ... 0.         0.         0.        ]
 [0.         0.03811413 1.         ... 0.         0.         0.        ]
 ...
 [0.         0.         0.         ... 1.         0.87979952 0.72633155]
 [0.         0.         0.         ... 0.87979952 1.         0.63902615]
 [0.         0.         0.         ... 0.72633155 0.63902615 1.        ]]
```

图 19: 余弦相似度矩阵

对于第 i 个样本点而言，a_i 表示 i 到它所属簇中所有其他样本点对象的平均距离，体现了 i 与它所属簇的凝聚度；同时，b_i 表示 i 和任意一个不包含 i 样本点的簇中所有对象的平均距离，体现了 i 和其他簇之间的分离度。从上面公式可以看出，轮廓系数的取值为 [-1,1]，其值越大越好，且当值接近于 0 时，则表示聚类结果中有簇重叠的情况。

确定好聚类评价方法后，我们对所有文本向量进行 DBSCAN 聚类，聚类结果如下表格所示：

组号	R	MinPoints	簇数	噪声比	轮廓系数
1	0.3	3	65	93.85%	0.2132
2	0.4	3	108	87.75%	0.11194
3	0.5	3	158	77.55%	0.0931
4	0.5	4	65	85.14%	**0.32132**
5	0.4	4	45	92.58%	0.2364
6	0.3	4	23	96.90%	0.1923

从表格中可以分析得出，所有组的轮廓系数都偏低，不超过 0.3，因为每一热点簇的相关留言问题数量普遍偏少，只有 15-45 条左右，可能会导致簇与簇之间的距离过小。当我们取 R=0.5，MinPts=5 时，聚类结果的轮廓系数最高，因此我们采用组 4 得到的聚类结果进行分析，共得到 65 个热点问题簇，利用 PCA 降维后使所有文本向量聚类后的结果在二维空间中可视化。从该图中可以看出，颜色相同的样本点其在二维空间内的位置相近，且在同一方向上的聚类结果较好。

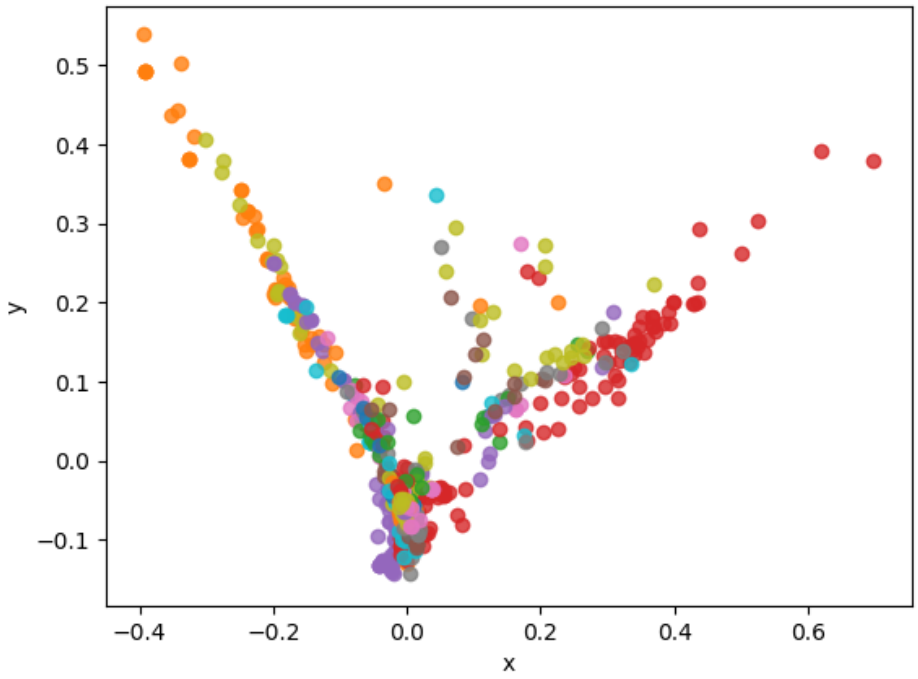

图 20: 聚类结果可视化

以下为任意选取的一部分热点簇中的留言详情对应的留言主题、点赞数以及反对数。

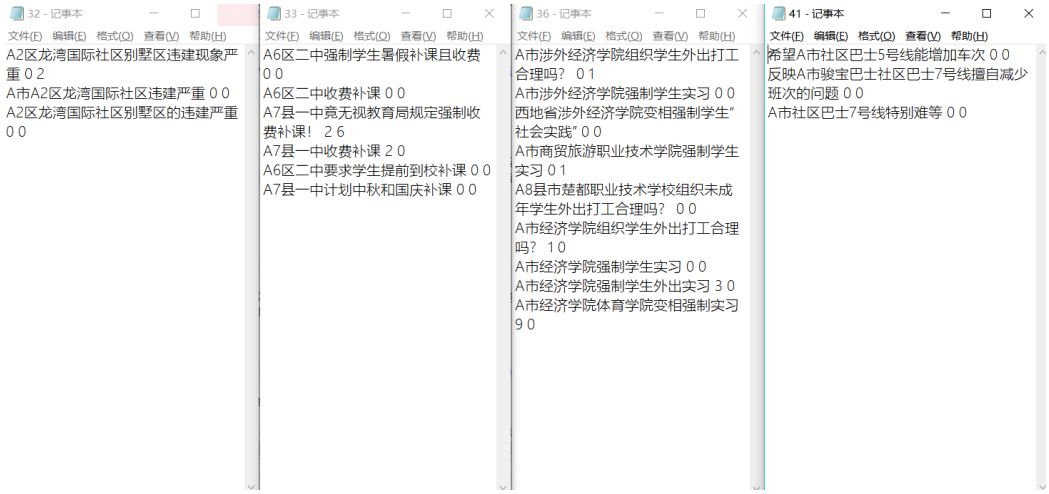

图 21: 聚类部分结果

5.7 熵值法进行热度计算

对于各个特征对问题热度的权重，本文采用熵值法进行定量计算。熵值法是一种基于信息熵的评估方法，它从信息论角度出发，以各特征取值情况来划分数据样本的空间。设样本数据集容量为 m，具有 n 个特征作为指标，则用 X_{ij} 表示为第 i 个样本点第 j 项指标值，其中 i=1，2，...,m ;j=1,2,...,n。各项指标熵值的计算公式为：

首先计算指标比重

$$Y_{ij} = \frac{X_{ij}}{\sum_{i=1}^{m}(X_{ij})} \qquad 1 \le j \le n \tag{4}$$

再利用 Y_{ij} 和样本容量 m，计算第 j 项指标的熵值

$$e_j = -\frac{1}{\ln m}\sum_{i=1}^{m}(Y_{ij}\ln Y_{ij}) \qquad 1 \le i \le m \tag{5}$$

根据各项指标的熵值，可以计算该指标对于热度的贡献和权值，则第 j 项指标的权重 W_j 的计算公式为：

$$W_j = \frac{1-e_j}{\sum_{j=1}^{n}(1-e_j)} \qquad 0 \le e_j \le 1 \tag{6}$$

本文聚类后得到的数据集 Q 为热点问题的集合，其热度的特征主要包括相关留言数量，支持总数和反对总数。为平衡时间对问题热度的影响，通过计算其时间平均值，得到平均留言数量，平均支持数和平均反对数三项指标, 分别为 X_1, X_2, X_3。对数据集 Q 的任一热点问题 Q_i，可描述为 X_{i1}, X_{i2}, X_{i3}，分别代表第 i 个热点问题的平均留言数，平均支持数和平均反对数三项指标。设热点问题 Q_i 的热度为 $F(Q_i)$，利用公式（1）、（2）和（3）可以计算得到 X_{i1} X_{i2} X_{i3} 对于问题热度的权值 W_1 W_2 W_3。则某热度问题 Q_i 的热度计算公式为：

$$F(Q_i) = W_1 \times X_{i1} + W_2 \times X_{i2} + W_3 \times X_{i3} \tag{7}$$

由公式（4）可知，热点问题的平均留言数量，平均支持数越大，平均反对数越低，则该问题的热度就越高，热度排名越靠前。

在本题中，我们通过 DBSCAN 聚类算法得到了 65 个热点问题簇，通过熵值法，我们对每个热点问题簇的留言数量，平均点赞数和平均反对数分别确定了以下权重:

热度指标	留言数量	平均点赞数	平均反对数
权重	0.194118	0.276883	0.528999

接着，通过公式 (7)，我们可以对 65 个热点问题簇进行热度计算，并对其按照热度降序排序后得到如图 22 所示。

从上述表格中可以得出，排在前 5 位的热点问题序号分别为 {1,4,10,12,46}，如图 23 所示。为了更直观地了解各个簇所描述的热点问题，我们将每个热点问题簇中的所有留言详情对应的留言主题，通过提取高频词的方式制作了 5 张“词云图”。

	热点问题序号	留言数量	平均赞成数	平均反对数	总热度
2	1	49	0.04081633	0.979591837	9.004880245
3	4	44	0	0.25	8.40894225
4	10	47	0.06382979	1.872340426	8.15075317
5	12	16	1.0625	1.25	2.738827438
6	46	9	1.444444444	0.222222222	2.029448778
7	2	10	0	0	1.94118
8	33	10	0	0.1	1.8882801
9	13	9	0.11111111	0.555555556	1.483938444
10	15	7	0	0	1.358826
11	24	7	0	0	1.358826
12	34	17	0.05882353	3.764705882	1.324767588
13	14	6	0	0.166666667	1.0765415
14	32	6	0	0.166666667	1.0765415
15	52	6	0	0.166666667	1.0765415
16	26	7	0.42857143	0.857142857	1.024062429
17	5	5	0	0	0.97059
18	54	5	0	0	0.97059
19	6	8	0	1.25	0.89169525
20	60	5	0	0.2	0.8647902
21	55	4	0.25	0	0.84569275
22	58	4	0.25	0	0.84569275
23	35	6	0.66666667	1	0.820297667
24	8	4	0	0	0.776472
25	25	4	0	0	0.776472
26	27	4	0	0	0.776472
27	30	4	0	0	0.776472
28	47	4	0	0	0.776472
29	48	4	0	0	0.776472
30	0	5	0	0.4	0.7589904
31	42	5	0	0.4	0.7589904
32	62	4	0.25	0.25	0.713443
33	7	4	0	0.25	0.64422225

(a) 表 1

	热点问题序号	留言数量	平均赞成数	平均反对数	总热度
34	18	4	0	0.25	0.64422225
35	21	4	0	0.25	0.64422225
36	43	4	0	0.25	0.64422225
37	44	4	0	0.25	0.64422225
38	20	5	0.2	0.8	0.6027674
39	16	4	0	0.5	0.5119725
40	23	5	0	1	0.441591
41	22	4	0	0.75	0.37972275
42	61	4	0	0.75	0.37972275
43	28	8	0.25	2.375	0.365792125
44	3	4	0	1	0.247473
45	36	4	0	1	0.247473
46	57	4	0	1	0.247473
47	63	4	0.25	1.25	0.184444
48	53	7	0	2.285714286	0.149685429
49	56	3	0.33333333	1	0.145649333
50	19	5	0	1.6	0.1241916
51	38	4	0	1.25	0.11522325
52	29	8	0.125	2.875	0.06668225
53	31	4	0	1.5	-0.0170265
54	39	4	0	1.5	-0.0170265
55	11	6	0	4	-0.951288
56	17	5	0	3.8	-1.0396062
57	41	4	0	3.5	-1.0750245
58	40	7	0	5	-1.286169
59	50	4	0	4	-1.339524
60	9	7	0.28571429	5.571428571	-1.509344714
61	59	4	0	7.5	-3.1910205
62	37	5	1	12	-5.100515
63	45	5	0	26.2	-12.8891838
64	51	4	0.5	37	-18.6580495
65	49	5	0	421.6	-222.0553884

(b) 表 2

图 22: 热度计算结果表

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	9.005	2019/11/02至2020/01/25	A2区丽发新城小区	小区附近搅拌站噪音扰民
2	4	8.409	2019/07/07至2019/08/28	A市伊景园滨河苑	定向限价商品房项目违规捆绑销售车位
3	10	8.151	2019/01/04至2019/12/30	A市/A市市民	加快A市全面城市建设
4	12	2.739	2019/07/21至2019/09/25	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民
5	46	2.029	2017/06/08至2019/11/27	A市经济学院学生	学校强制学生去定点企业实习

图 23: Top5 热点问题表

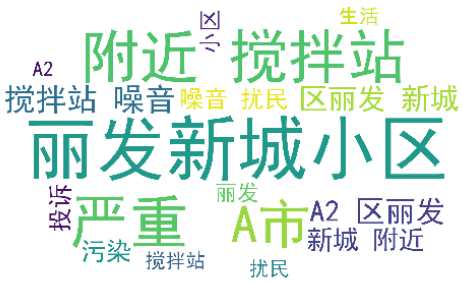

图 24: 热点 1 词云图

5.8 结果分析

与 Top5 热点问题对应的部分留言（前 3 条）展示如下：

通过对网络问政平台的群众留言数据进行深入分析，挖掘出热度指数位列前五的热点问题：居民区附近的搅拌站噪音扰民、开发商捆绑销售车位、城市建设工作的进展、临街餐饮店噪音扰民以及学校强制学生不合理实习。可以看出这些热门问题基本围绕城乡建设类与民政类主题，或许是因为随着城市的快速发展，生活水平的提高，人们越来越注重个人生活质量的追求以及在社会现实生活中个人权益的保护。例如

图 25: 热点 4 词云图

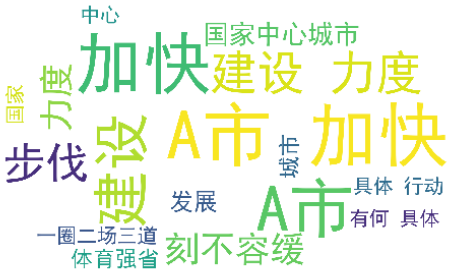

图 26: 热点 10 词云图

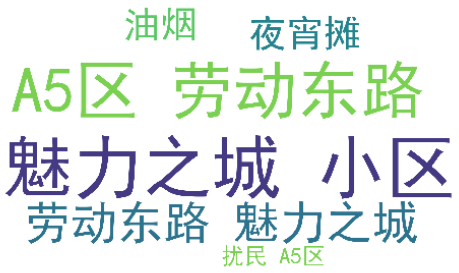

图 27: 热点 12 词云图

在此次热点挖掘中“噪声扰民”俨然成为一大热词，其相关问题在五大热点问题里不仅有居于首位且出现了两次，群众反映强烈能从侧面说明由于城市噪音污染的监督管理工作未得到充分落实，人民生活质量没

图 28: 热点 46 词云图

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	188809	A909139	A市万家丽南路丽发新城居民区附近搅拌站扰民	2019/11/19 18:07:54	0米处建搅拌站，运渣	0	1
1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-11-13 11:20:21	方建搅拌站。可想而	0	0
1	190108	A909240	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康	0	1
问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
4	190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019-08-23 12:22:00	下发文件，强制要求职	0	0
4	195511	A909237	车位捆绑违规销售	2019-08-16 14:20:26	多次一直没有人彻查处	0	0
4	195995	A909199	广铁集团铁路职工定向商品房伊景园滨河苑项目的	2019-08-10 18:15:16	已未给出首付款详细的	0	0
问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
10	191872	A00031618	请A市加快轨道交通建设力度	2019/3/1 15:19:28	高铁网络和数十条地铁	2	9
10	193514	A00031618	请加快A市月亮岛片区公共服务力度	2019/3/20 16:39:22	公办老年大学，公办	0	4
10	195905	A00031618	A市加快招商引资方面有何具体行动?	2019/7/24 13:48:16	口流失严重，近年周边	0	0
问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
12	195095	A00039089	魅力之城小区临街门面油烟直排扰民	2019/09/05 12:29:01	天24小时都是烟。请政	0	3
12	228107	A00010420	A3区梅溪湖看云路润芳园小区油烟扰民	2019/8/28 12:49:18	力强关。目前油烟机清	0	0
12	236798	A00039089	A5区劳动东路魅力之城小区油烟扰民	2019/07/28 12:49:18	没有。每天油烟直排。	0	4
问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
46	242062	A00028889	西地省涉外经济学院变相强制学生“社会实践”	2019/11/27 23:14:33	会实践，起码是从学生	0	0
46	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	知道！学校很小但是这	3	0
46	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	了合同，并且公司也要	9	0

图 29: Top5 热点问题对应部分留言内容

能得到保障，群众呼声体现了他们希望能够敦促政府做好更严格的噪音管理规范，积极规划如何从根本上解决噪声污染的问题。同时，热点问题还反映了霸王条款的猖獗，例如开发商强制捆绑销售车位、学校强制学生外出实习，这些行径已经严重侵害群众利益。而通过留言我们看出买房者和学生持强烈拒绝的态度并向政府平台投诉，这是法治社会公民所具有的正当维权表现。政府有指导公民正确维护合法权利的职责同时也应当加大监督力度以保障群众的权利不受侵犯，如消费者维权部门应当对各类物业管理企业实施更加严格的监管手段；教育局对学校进行定期走访并加强监督教育违法行为。除了维权问题，城市建设也是广大群众关心的热点，随着城区规模的扩张和人流、车流的大幅增加，这给城区道路交通、公共设施服务以及城市招商引资等发展工作增加了不同程度的负担，如果没有得到有效的缓解将会影响城市形象和群众居住环境。城市在不断发展，建设管理工作任重而道远，这既需要政府强有力的领导，也需要广大市民群众的支持和配合。

6 答复意见的评价

网络问政不仅是公民通过网络表达诉求，实现参政议政的新兴民主形式，还是政府和民众之间的交流互动过程。网络问政是我国公共管理实践中产生的创新形式，指通过网络这种新的技术形式，进行党政机关

和民间的平等对话，在网络中显现公民民主需求的政治手段和行为。虽然网络问政最突出的的特色是“问”，但“问”同时隐含了答，展现了一种动态的对话和互动过程。故为了构建充满活力、和谐有序、建设性的网络民主平台，对“答”部分进行评价体系构建并以此进行绩效评估具有非常重要的现实意义。

6.1 评价体系构建

为了进一步合理评价相关部门对留言的答复意见，根据题目要求以及对数据的观察，我们将答复时限、内容相关性、形式完整性和详情可解释性设为一类指标，通过科学的设计，构建了综合评价指标体系。这利于网络平台根据各地区在一段时间内的留言回复情况计算出绩效指数，用统一的标准衡量各地的网络问政情况，将当地政府部门问政指数公开揭露，并定期更新，让监督制约的作用全面覆盖到每个人民群众。通过这种科学设计计算出的指数，网民可以更直观地了解当地区的网络问政发展情况。

6.1.1 评分说明

本文采用李克特量表对评分进行赋值。李克特量表 (Likert scale) 是属评分加总式量表最常用的一种，属同一构念的这些项目是用加总方式来计分，单独或个别项目是无意义的。它是由美国社会心理学家李克特于 1932 年在原有的总加量表基础上改进而成的。该量表由一组陈述组成，每一陈述有” 非常同意”、” 同意”、” 不一定”、” 不同意”、” 非常不同意” 五种回答，分别记为 5、4、3、2、1，每个被调查者的态度总分就是他对各道题的回答所得分数的加总，这一总分可说明他的态度强弱或他在这一量表上的不同状态。

本文将对答复时限、内容相关性、形式完整性和详情可解释性这四个二级指标的具体数值进行一定标准的划分，借鉴李克特量表的形式进行赋分。

6.1.2 指标选择与说明

1. 答复时限

问政讲求给群众落实问题的及时性。虽然群众反映的问题涉及方方面面, 有的问题能立即作出答复，但还有些问题需要线下了解核实才能做出答复或者落实。但出于对民心的安抚，即使在还没有来得及进行问题的实质性处理或者仍在解决过程中时，政府社情民意处应该给群众一个“已收到”的反馈，意在告知群众其反映的问题解决进度，并且给出相应的承诺会对问题进行整改，以起到宽慰群众心理的作用，做到“事事有回应”。各省市的规定普遍对处理留言并给予反馈的时间给出了限定，一般不超过 1 个月，情况复杂未能及时处理的需向上级报告。

本文对附件四中的“答复时间”一栏和“留言时间”一栏中的日期进行差分处理，得到以日为单位的答复时限：

$$\text{答复时限} = \text{答复时间} - \text{留言时间}$$

2. 相关性

相关性指的是答复意见的内容是否与群众留言内容相关。本文从以下三个方面对其进行拆解：

- 措施的有效性：基本事实情况、采取了何种措施、将要进行的工作等
- 回应的权威性：部门主要负责人是否作出回应
- 回应的覆盖面：是否对问题都做出解答

考虑到上文的文本分析工作，可以看出留言内容的相关关键词能够通过特征选择进行获得，并且能够成为留言内容的摘要提取，即覆盖所有关键信息。故在对相关性进行评价时，本文选用关键词匹配算法，对留言

答复时限评分标准	评分
3 日内答复	5
7 日内答复	4
15 日内答复	3
30 日内答复	2
30 日以上内答复	1

图 30: 答复时限量表图

内容与答复内容分别进行文本分析，提取 20 个关键词然后进行匹配，对匹配数量进行打分。

$$\text{分数} = \text{匹配成功的关键词数 (匹配大于等于 5 个关键词按 5 个计算)}$$

3. 完整性

完整性指的是政府答复意见内容是否满足某种规范。根据《政府信息公开条例》，行政机关作出的政府信息公开答复应符合规范形式。我们从政务网上查询规范的答复模板并在附件四中的内容进行比对，总结出较为规范的模板，如下图：

网友“UU008706”您好！您的留言已收悉。现将有关情况回复如下：
（问候 确认收到 开启回复正文）
（回复正文）
梅溪湖一期引进 A 市图书馆分馆，位于梅溪湖创新中心，已开馆营业。梅溪湖二期金菊路与雪松路东南角规划有西地省图书馆新馆，目前正在进行前期筹备工作，具体开馆时间待定。

感谢您对我们工作的支持、理解与监督！（表达感谢）
2019 年 1 月 9 日（日期标注）

图 31: 政府答复意见模板规范说明

对于不属该部门处理的政务信息，我们可以看到在附件四里政府会做已转交的处理（如图）。我们同样对这样的答复意见的完整性进行打分。

基于以上内容，政府答复意见完整性打分标准如下：

4. 可解释性：

可解释性指的是答复内容中的相关解释。在政府对于群众留言内容进行回复过程中，如有相应政府文件和法律法规作为理论支撑，会使得答复内容更具有说服力和可信度。我们将可解释性指标作为虚拟变量。在本题，我们认为只要在答复意见内容出现下图所示索引词，即可认为本段答复意见为可解释性为好，得分为 5；反之得分为 1。

“UU0082365”您好！（问候）您所反映的问题已收悉。（确认收到）

就您所反映的问题我办已转交给相关部门研究处理。（告知留言已转交处理）

在此，感谢您对我们工作的关心与支持。（表达感谢）

2019 年 12 月 17 日（日期标注）

图 32: 已转交的答复意见模板

答复内容类型	答复模式完整程度	评分
直接给出解决内容	出现 5 层模板词语	5
	出现 4 层模板词语	4
	出现 3 层模板词语	3
	出现 2 层模板词语	2
	出现 1 层模板词语或更低	1
转交其他部门或机关	出现 5 层模板词语	5
	出现 4 层模板词语	4
	出现 3 层模板词语	3
	出现 2 层模板词语	2
	出现 1 层模板词语或更低	1

图 33: 完整度量表

类别	索引词
文件引出内容	“根据《”“政府发文”“规定：”
政府文件标题	“条例》”“文件》”“规定》”“方案》”“意见》”“通知》”“核准证》”“法》”“（试行）》”“细则》”“规划》”“协议》”“图》”“批复》”“公告》”“许可证》”“纪律》”“合同》”“标准》”“决定》”

图 34: 理论解释索引词

根据以上评判标准以下是一段得分为 1 的答复意见内容，即具有好的可解释性。由红字高亮内容可见，此条答复内容引用《关于进一步做好义务教育招生入学工作的实施办法》政府文件作为解答支撑，回答有理有据，是好的可解释性的体现，此条答复内容得分为 5。

6.1.3 数据描述

在对附件四的数据进行观察时，我们发现“答复意见”那一栏的内容主要可以分为两类：

网友“UU0082211”您好！您的留言已收悉。现将有关情况回复如下：

2016 年我市在考察外地经验、广泛征求社会各界意见的基础上，由**市委市政府制定下发《关于进一步做好义务教育招生入学工作的实施办法》[政府发文]5 号，文件明确规定**进城务工人员随迁子女入学条件：要求在 A 市城区持续合法居住一年以上（含一年），同时必须参加本市职工基本养老保险至少一年。由于您在来信中提供的信息不够清楚，具体情况请直接电话咨询 A 市教育局基础教育处：0000-00000000。

感谢您对我们工作的支持、理解与监督！
2018 年 2 月 13 日

图 35: 可解释性好的答复意见内容展示

- 对于群众留言具有完整内容的回复
- 对于群众留言内容不完整的回复

对于群众留言不完整的，政府相关部门会告知申请人作出更改、补充。通常是以如下的格式：即对内容

网友“UU0081194”您好！您的留言已收悉。现将有关情况回复如下：因您未留下联系方式及投诉的相关证据材料，市工商局无法根据您提供的信息进行投诉信息的登记分送和处理。您可直接拨打我局消费者投诉举报热线0731-12315进行反映。感谢您对我们工作的支持、理解与监督！2018年12月28日

图 36: 针对不完整留言的答复意见模板

不完整进行说明，并提供下一步解决方案。考虑到若将此类针对不完整留言的答复意见与针对完整留言的答复意见一同进行评分，那么前者出于客观原因，将不具有相关性（即不具有关键词）。故在此处我们将两种答复意见进行筛选并分类，并以分类后的结果对两种情况下的政府答复意见分别进行评分。

6.1.4 评价体系模型

针对完整和非完整的答复意见内容，我们构建了两个 评价模型，即完整内容答复评价体系和非完整内容答复评价体系，具体如下表所示。本文采用层次分析法（AHP）确定评价模型的综合权重，计算步骤如下：

1. 确定评价体系的层次结构模型。本文目标层为答复评价，要素层是影响答复评价的因素即评价体系中操作层指标，方案层则是答复内容对象。
2. 利用 1-9 标度法构建系统层、要素层和指标层的两两判断矩阵。本文主要确定要素层各指标对目标的相对重要性，因此只需构造要素层的判断矩阵。
3. 层次单排序及其一致性检验。W 的元素为同一层次因素对于上一层次某因素相对重要性的排序权值。

- 完整内容答复评价体系

目标层	指标层	操作层指标
完整内容答复评价体系	答复时限	答复时隔长度
	相关性	答复关键词匹配分数
	完整性	答复内容出现模版词语层数
	可解释性	答复内容是否有理论支撑

图 37: 完整内容答复评价体系

在确定层次构建模型之后，构建判断矩阵 M_1。设答复时隔长度、答复关键词匹配分数、答复内容出现模版词语层数、答复内容是否有理论支撑四个指标分别为 A_1、A_2、A_3 和 A_4。构造其对目标层的影响两两比较结果如下：

$$\mathbf{M_1}=\begin{matrix}A_1\\A_2\\A_3\\A_4\end{matrix}\begin{bmatrix}1 & \frac{1}{5} & \frac{1}{3} & \frac{1}{4}\\5 & 1 & 4 & 3\\3 & \frac{1}{4} & 1 & \frac{1}{2}\\4 & \frac{1}{3} & 2 & 1\end{bmatrix}$$, 将矩阵列向量归一化 $$\begin{bmatrix}0.077 & 0.112 & 0.045 & 0.053\\0.385 & 0.561 & 0.545 & 0.632\\0.231 & 0.14 & 0.136 & 0.105\\0.308 & 0.187 & 0.273 & 0.211\end{bmatrix}$$

求行和归一化，得特征向量 $\mathbf{W_1}=\begin{bmatrix}0.072\\0.531\\0.153\\0.245\end{bmatrix}$

最后检验一致性以确认判断矩阵是否在不一致的允许范围内

$\mathbf{M_1W_1}=\begin{bmatrix}0.072\\0.531\\0.153\\0.245\end{bmatrix}=\lambda W_1$ 则 $\lambda=4.117$

$CI=\frac{\lambda-n}{n-1}=0.039$, 已知 $n=4$ 时，$RI=0.9$, 则 $CR=CI/RI=0.043<0.1$, 说明通过一致性检验，即可用其归一化特征向量 w1 作为权向量。因此完整内容答复评价体系各项指标的权重如下表所示：

操作指标层	AHP 权重
答复时隔长度	0.072
答复关键词匹配分数	0.531
答复内容出现模版词语层数	0.153
答复内容是否有理论支撑	0.245

图 38: 评价体系指标及其 AHP 权重

- 完整内容答复评价体系

在确定层次构建模型之后，构建判断矩阵 M_2。设答复时隔长度、答复内容出现模版词语层数、答复内容是否有理论支撑三个指标分别为 B_1、B_2 和 B_3。构造其对目标层的影响两两比较结果如下：

目标层	指标层	操作层指标
非完整内容答复评价体系	答复时限	答复时隔长度
	完整性	答复内容出现模版词语层数
	可解释性	答复内容是否有理论支撑

图 39: 非完整内容答复评价体系

$$\mathbf{M_2} = \begin{matrix} B_1 \\ B_2 \\ B_3 \end{matrix}\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ 2 & 1 & \frac{1}{2} \\ 4 & 2 & 1 \end{bmatrix}$$, 将矩阵列向量归一化 $$\begin{bmatrix} 0.143 & 0.143 & 0.143 \\ 0.286 & 0.286 & 0.286 \\ 0.571 & 0.571 & 0.571 \end{bmatrix}$$

求行和归一化，得特征向量 $\mathbf{W_2} = \begin{bmatrix} 0.143 \\ 0.286 \\ 0.571 \end{bmatrix}$

由于 M_2 显然具有一致阵的性质，因此可用其归一化特征向量 W_2 作为权向量。因此非完整内容答复评价体系各项指标的权重如下表所示：

操作指标层	AHP 权重
答复时隔长度	0.143
答复内容出现模版词语层数	0.286
答复内容是否有理论支撑	0.571

图 40: 评价体系指标及其 AHP 权重

6.2 评价体系结果呈现

6.2.1 答复时限

由于附件 4 中已经给出了各个留言和答复的具体时间，我们在 Excel 中对每条记录的两个日期相减后得到每条留言的答复间隔时间。通过答复时限量表图的赋分方式对每条记录的答复时限进行赋分后，结果如图 42 所示。

观察所得数据我们可以发现，约 80% 以上的答复时限都在 30 日之内，因此也可以看出我们在答复时限量表图中的赋权方式是合理的。

6.2.2 相关性

对附件 4 中的所有留言记录，我们使用 Python 的 TextRank4ZH 模块分别对每个记录的留言详情和回复内容提取 20 个关键字，然后根据成功匹配的关键字数量对每段回复进行打分。图 43 为部分相关性赋分结果。

可以看出，大部分留言记录的关键词匹配次数都小于等于 5 次，因此相关性的赋权方式是合理的。

6.2.3 完整性

一段规范的政府回复模板应包含“问候”、“确认收到”，“回复正文”或“已转交”、“表达感谢”和“回复日期标注”部分，在每个部分中，若相关模板词语出现一次，则该部分的分数就为 1，否则为 0。而完整性得分就是这 5 个部分所有得分的累加之和。对于所有留言回复，我们首先匹配回复内容的“问候”、“确认收到”以及“表达感谢”部分，通过在 Excel 中同时查找多个关键字的操作可以得到每段回复的这三部分的得分。对于属于留言申请部门处理的政务信息，我们用 Excel 匹配“回复正文”等关键词，而对于不属于该部门处理的信息而需要转交给其他部门的留言申请，我们匹配“已转交”等关键词。对于最后的日期标注，由于每条留言的回复日期不尽相同，我们使用 Python 中的正则表达式在每条留言回复的末尾匹配“r(\d{4}-\d{1,2}-\d{1,2})”模式，即年-月-日表达格式。若匹配成功，则“回复日期标注”部分得分为 1，否则为 0。图 44 为部分完整性赋分结果。

6.2.4 可解释性

为了确定可解释性部分得分，我们将“理论解释索引词表”中的词同时作为关键词，使用 Excel 对每段留言回复进行关键词匹配，若某一留言回复中包含表中的任一关键词，则我们令其得分为 5，否则为 1。部分可解释性得分结果如图 45 所示。

6.2.5 答复质量评价

通过以上 4 个指标的得分结果以及 AHP 得到的权重，我们可以得到每条留言回复的总得分。图 41 分别列出了完整留言申请的和不完整留言申请的前 5 条得分最高的留言记录，图 46 和 47 分别列出了两类得分 TOP3 的答复意见详细内容。所有留言记录的得分在附件中给出。

	留言详情	留言时间	答复意见	答复时间	是否转交	答复总得分
2	今年2月份在楚龙辖D	18/3/27 9:46:2	其共同居住生活的	2018/3/29 14:43:22	否	4.025
3	望，原来还有个农贸	17/11/14 12:32	关注。 根据《开	2017/11/15 14:11:02	否	4.023
4	民自愿进行流转的吧	17/5/5 13:26:0	点建设任务中关于	2017/5/8 10:02:18	否	4.019
5	无调整，是仍由渌口	19/6/5 14:43:4	仍由渌口区管理，	2019/6/5 20:51:01	否	4.003
6	是碧波荡漾，波光粼	19/3/6 13:21:0	，提升供水水质，	2019/3/8 17:22:52	否	3.995
7	反都在震动，声音又	18/5/14 22:20:3	检测（四楼1、2	2018/5/29 10:31:48	是	2.43
8	能会不走电，就这	19/10/29 15:21	行现场拍照、登	2019/10/31 9:14:18	是	2.43
9	政策规定父母户口在	18/10/17 10:08	，成年子女不能扌	2018/10/30 15:36:18	是	2.43
10	非常大，现在下河	18/5/23 12:20:3	包含城北体育场改	2018/5/24 11:48:47	是	2.43
11	续投诉。D10县洪杯	19/5/18 16:11:3	的教师介绍学生对	2019/5/22 11:34:17	是	2.287

图 41: 答复质量得分表

从上表可以看出，留言答复时限越短，相关性越强，完整性越好，可解释性越充分，留言回复得分就越高，答复质量就越好。同时，我们可以看到留言申请内容完整的留言回复内容普遍比留言申请内容不完整的得分要高，因此市民在填写留言申请时，完整的留言信息不仅可以让政务人员可以清晰地掌握留言申请中的问题，而且能提高留言回复的质量，使得政务人员可以更好地给出解答方案。

6.3 评价体系反思

答复有效性：指政府答复意见中对群众提出的问题的解决程度。

在对数据的观察中，我们可以发现留言中关于拆迁、城市改造的咨询反复出现这类留言中，网民往往迫切关注问题的进展以及何时能够被解决。这种对信息的需求源于政府内部决策过程和管理过程的相对封闭，而民众对关乎个人利益的事项又比较急切。面对民众对于信息的急切需求，被问政的官员在回复这类留言

时也存在难度，此类问题往往需要一个解决过程或者长期规划，因而政府很难对具体的解决时间给出明确回复，只能使用“尽快处理”、“努力中”等词语来回复。而此类对于留言问题的解答程度我们难以量化。情感分析虽然是自然语言处理中对于主观言论进行感情色彩的评价，但是不适用于对政府答复意见的分析中。原因是，政府相关部门的答复由于表达主体是政府部门，本身具有权威性，在工作人员进行回复的时候，表达一般具有官方色彩，这对于情感的评估是一个难以逾越的障碍。所以对于答复内容中有效性的评估比较困难。

针对以上问题，本文在此提出可能的解决方案：

对于有效性的评估，可以转移对文本长度评估和可解释性评估的指标上。通常来说，越长的政府部门答复意见对于问题的解答会更清楚，同时含有更多理论支撑的政府部门答复意见会更具有参考性与说服力。这对于有效性也是一种保障。

7 结论

十八届三中全会《关于全面深化改革若干重大问题的决定》中提出要创新社会治理体制，发挥政府主导作用，鼓励和支持社会各方面参与，实现政府治理和社会自我调节、居民自治的良性互动。

本文通过对网络问政平台上的留言内容进行分类、运用 DBSCAN 算法对热点问题进行聚类和建立政务答复意见评价体系对网络问政平台的运作进行了深入的了解与研究分析。

对于群众问政留言的分类以及热点问题的聚类工作中我们可以看出，将自然语言处理融入政务工作刻不容缓。群众问政内容信息量大且多、文本噪音嘈杂且种类繁多，自然语言处理技术无疑将便利政务工作的开展。政府可以通过筛选整理出拟处理回复的留言，并根据不同的领域进行分类，方便下一步不同主管部门进行认领或主管部门分配任务；而后可以根据留言内容的不同领域和地区，由不同的部门进行处理，做到分类转办，极大地保障了“事事有回应”的政府对人民的承诺实现。

通过对网络问政数据的分析，可以看出网络问政平台更像是作为政府的留言板，它的主要功能可以归纳为三点：

- 网上答疑，即为网民提供关于公共行政的信息
- 网络谏言，即倾听网民对于政府管理的意见，接纳或给出评价
- 网上行政，即通过网络获取网民投诉、求助信息，并责成有关部门进行处理

同时，从分析结果可以看出，多数群众问政问题集中在民政与城乡建设中。作为一种创新治理体制、鼓励多方参与的尝试的网络问政平台，它既是对回应型政府建设的需求的满足，也是一种协商式的问题解决方式，对于培育社会理性具有正面效应。好如城乡建设与民政方面的问题，若通过政府实地走访获取信息虽是必要，但不能做到事事关心。但网络问政平台的出现，使得社会、民众能够通过“末端”参与公共治理的尝试将成为全方位公众参与的前奏，政府在回应民众“上书”的过程中也能更好的反思相关决策制定的过程，从而促成一种由“末端式”的善后处理向源头式的社会治理的转变。

参考文献

[1] 哈工大拓展停用词表.https://github.com/goto456/stopwords

[2] 赵琳瑛. 基于隐马尔科夫模型的中文命名实体识别研究. 西安电子科技大学.2007

[3] 周昭涛. 文本聚类分析效果评价及文本表示研究 [D]. 中国科学院研究生院 (计算技术研究所)，2005.

[4] 胡学钢，董学春，谢飞，基于词向量空间模型的中文文本分类方法 [J]. 合肥工业大学学报: 自然科学版，2007，30(10) :1261-1264.

[5] 基于多特征的热门微博预测算法研究

[6] 洪宇，张宇，刘挺, 等. 话题检测与跟踪的评测及研究综述. 中文信息学报, 2007(06).

[7] 王宏勇. 网络舆情热点发现与分析研究 [D]. 四川: 西南交通大学，2011.

[8] 张胜, 张鑫, 程佳军等 Chinese Sentiment Classification Using Extended Word2V ec[J].Journal of Donghua University(English Edition),2016,33(05):823-826.

[9] 罗伯特·达尔：《论民主》[M]. 李柏光，林猛译. 北京：商务印书馆,1999 年

[10] 曹学娜，孙祥. 网络问政的条件及推进路径 [J]. 公共行政，2009（10）

[11] 亓莱滨. 李克特量表的统计学分析与模糊综合评判 [J]. 山东科学, 2006(02):21-26+31.

1	留言详情	留言时间	答复意见	答复时间	答复时限	答复时限得分
2	司却以交20万保证金	019/4/25 9:32:	理费，在业主大会	2019/5/10 14:56:53	15.22550926	2
3		19/4/24 16:03:	，且换填后还有三	2019/5/9 9:49:10	14.73993056	3
4	更是加大了教师的工	19/4/24 15:40:	教职工要依法签订	2019/5/9 9:49:14	14.75636574	3
5	A市，想买套公寓，	19/4/24 15:07:	下（含），首次购	2019/5/9 9:49:42	14.77930556	3
6	坡岭小学”，原“马	19/4/23 17:03:	”的问题。公交站	2019/5/9 9:51:30	15.70012731	2
7	泥巴冲到右边，越是	2019/4/8 8:37	有说明卫生较差的	2019/5/9 10:02:08	31.05888889	1
8	社区惠民装电梯的	19/3/29 11:53:	政府办公室下发了	2019/5/9 10:18:58	40.93443287	1
9	，天寒地冻的跑好	18/12/31 22:21	设施设备采购等	2019/1/29 10:53:00	28.52153935	2
10	得到相关准确开工信	18/12/31 9:55:	检查后，西地省楚	2019/1/16 15:29:43	16.23244213	2
11	桥等地方做立体绿化	18/12/31 9:45:	要求完成了建设	2019/1/16 15:31:05	16.23965278	2
12	局审批通过《温室	18/12/30 22:30	地征收补偿款给原	2019/3/11 16:06:33	70.73336806	1
13	置房地下室近两万	18/12/29 23:27	发[2014]7号文件	2019/1/29 10:52:01	30.47511574	1
14	大量从小区开车出去	18/12/29 11:55	具体选址，招标	2019/1/14 14:34:58	16.11069444	2
15	相关政府部门的大力	18/12/28 17:18	，已由银盆岭派出	2019/1/3 14:03:07	5.864143519	4
16	以上！天寒地冻，其	18/12/28 7:53:	驶员工作时间长，	2019/1/14 14:33:17	17.27768519	2
17	https://baidu.com	18/12/27 15:18	路口两端各拆除20	2019/3/6 10:26:14	68.79730324	1
18	各种理由拒绝退货，	18/12/27 1:55:	的信息进行投诉信	2019/1/3 14:02:47	7.505162037	3
19	建议在艺术中心先	18/12/26 16:51	湖二期金菊路与雪	2019/1/14 14:32:40	18.90347222	2
20	早就施工，严重影响	18/12/25 19:35	由于需要夜间连续	2019/1/8 16:19:16	13.86393519	3
21	可以尽快合一。让社	18/12/25 16:23	，需三方或三方以	2019/1/4 15:48:23	9.975648148	3
22	、苹果等手机都无法	18/12/25 16:19	请关注潇洒支付	2019/1/4 15:49:46	9.979131944	3
23	田。根据《土地管理	18/12/25 14:40	了土地补偿协议，	2019/1/8 16:18:00	14.06790509	3
24	车辆和行人通行，此	18/12/25 13:56	三十八条第一款第	2019/1/16 15:22:16	22.05954861	2
25	故频发。如果8路线	18/12/23 21:47	好，非常感谢您	2019/1/29 10:50:31	36.54371528	1
26	是否能在A市办理商业	18/12/21 11:01	持非本中心的缴存	2019/1/3 14:00:47	13.12484954	3
27	A市国际会展中心非	18/12/20 17:28	成约800米路基，	2019/1/4 15:47:36	14.93017361	3
28	修A3区山景区西大门	18/12/20 11:16	资计划调整，该项	2019/1/3 13:59:33	14.11349537	3
29	个多亿好远，这笔大	18/12/15 15:17	办的西地省洋兴置	2019/1/4 15:44:31	20.01849537	2
30	是这样操作的。梅溪	18/12/14 14:29	为A市A3区那么好	2018/12/29 15:05:11	15.02483796	2
31	道路一直没有修好，	18/12/12 8:57:	，中学属于青雅丽	2019/3/15 15:40:09	93.27981481	1
32	，后面才想到在路	18/12/11 15:35	与该车实际停车时	2019/1/4 15:45:01	24.00649306	2
33	立，富吉又说在与住	18/12/11 15:23	3•18”“5•20”	2019/3/15 15:39:44	94.01157407	1
34	不说实话，如果是城	18/12/8 12:16:	细信息及证据，请	2018/12/27 9:23:01	18.87959491	2

图 42: 答复时限分数图

	A	B	C	D	E	F
1	留言详情	留言时间	答复意见	答复时间	匹配关键词	相关性得分
2	司却以交20万保证金	19/4/25 9:32:0	理费，在业主大会	2019/5/10 14:56:53	小区 物业公司 业主 业委	5
3		19/4/24 16:03:1	，且换填后还有三	2019/5/9 9:49:10		1
4	更是加大了教师的工	19/4/24 15:40:3	教职工要依法签订	2019/5/9 9:49:14	教师 幼儿园 市委	3
5	A市，想买套公寓，	19/4/24 15:07:1	下（含），首次购	2019/5/9 9:49:42	市 购房	2
6	坡岭小学”，原“马	19/4/23 17:03:3	”的问题。公交站	2019/5/9 9:51:30	马坡岭 小学 建议 保留 取	5
7	泥巴冲到右边，越是	2019/4/8 8:37	有说明卫生较差的	2019/5/9 10:02:08	没有 含	2
8	社区惠民装电梯的	19/3/29 11:53:3	政府办公室下发了	2019/5/9 10:18:58	电梯 关心 区	3
9	。天寒地冻的跑好	18/12/31 22:21	及设施设备采购等	2019/1/29 10:53:00	澜湾 社区 市	3
10	得到相关准确开工信	18/12/31 9:55:1	检查后，西地省楚	2019/1/16 15:29:43	相关 质量 监督	3
11	桥等地方做立体绿化	18/12/31 9:45:1	要求完成了建设。	2019/1/16 15:31:05	建设 洋湖	2
12	局审批通过《温室养	18/12/30 22:30	地征收补偿款给原	2019/3/11 16:06:33	用地	1
13	置房地下室近两万平	18/12/29 23:27	发[2014]7号文件	2019/1/29 10:52:01	区 房 安置 人防 工程 都	5
14	大量从小区开车出去	18/12/29 11:55	具体选址，招标	2019/1/14 14:34:58	小区 万国	2
15	相关政府部门的大力	18/12/28 17:18	，已由银盆岭派出	2019/1/3 14:03:07	支持	1
16	以上！天寒地冻，其	18/12/28 7:53:3	驶员工作时间长，	2019/1/14 14:33:17	公交车 高峰期	2
17	https://baidu.com	18/12/27 15:18	路口两端各拆除20	2019/3/6 10:26:14	通行 路 新开铺	3
18	各种理由拒绝退货，	18/12/27 1:55:0	的信息进行投诉信	2019/1/3 14:02:47		1
19	建议在艺术中心先	18/12/26 16:51	期二期金菊路与雪	2019/1/14 14:32:40	梅 溪湖	2
20	早就施工，严重影响	18/12/25 19:35	由于需要夜间连续	2019/1/8 16:19:16	施工	1
21	可以尽快合一。让社	18/12/25 16:23	，需三方或三方以	2019/1/4 15:48:23	卡	1
22	、苹果等手机都无法	18/12/25 16:19	可请关注潇洒支付	2019/1/4 15:49:46	nfc 卡 支持 公司	4
23	田。根据《土地管理	18/12/25 14:40	了土地补偿协议，	2019/1/8 16:18:00	村民 人民政府 建设 进行	4
24	车辆和行人通行，此	18/12/25 13:56	三十八条第一款第	2019/1/16 15:22:16	红灯 禁止 道路交通	3
25	放频发。如果8路线	18/12/23 21:47	好，非常感谢您	2019/1/29 10:50:31	市 规划 目前	3
26	是否能在A市办理商业	18/12/21 11:01	持非本中心的缴存	2019/1/3 14:00:47	公积金 缴存 住房 办理 市	5
27	A市国际会展中心非	18/12/20 17:28	成约800米路基，	2019/1/4 15:47:36	东路 机场	2
28	修A3区山景区西大门	18/12/20 11:16	资计划调整，该项	2019/1/3 13:59:33	白云路	1
29	个多亿好远，这笔大	18/12/15 15:17	办的西地省洋兴置	2019/1/4 15:44:31	集体	1
30	是这样操作的。梅溪	18/12/14 14:29	为A市A3区那么好	2018/12/29 15:05:11	干洗店 业	2
31	道路一直没有修好，	18/12/12 8:57:1	，中学属于青雅丽	2019/3/15 15:40:09	小学 城 学校	3
32	，后面才想到在路	18/12/11 15:35	与该车实际停车时	2019/1/4 15:45:01	收费 时间 记录	3
33	立，富吉又说在与住	18/12/11 15:23	3•18”“5•20”	2019/3/15 15:39:44	购房 车位 退房 区	4
34	不说实话，如果县城	18/12/8 12:16:	细信息及证据，计	2018/12/27 9:23:01		1

图 43: 相关性分数图

	A	B	C	D	E	F	G	H	I	J
1	留言详情	留言时间	答复意见	答复时间	日期得分	问候得分	确认收到得分	回复正文或转交得分	感谢得分	完整性得分
2	司却以交20万保证金	19/4/25 9:32:0	理费，在业主大会	2019/5/10 14:56:53	1	1	1	0	1	4
3		19/4/24 16:03:1	，且换填后还有三	2019/5/9 9:49:10	1	1	0	1	1	4
4	更是加大了教师的工	19/4/24 15:40:3	教职工要依法签订	2019/5/9 9:49:14	0	1	1	1	1	4
5	A市，想买套公寓，	19/4/24 15:07:1	下（含），首次购	2019/5/9 9:49:42	1	1	1	1	0	4
6	坡岭小学”，原“马	19/4/23 17:03:3	”的问题。公交站	2019/5/9 9:51:30	1	1	1	0	1	4
7	泥巴冲到右边，越是	2019/4/8 8:37	有说明卫生较差的	2019/5/9 10:02:08	1	1	0	1	1	4
8	社区惠民装电梯的	19/3/29 11:53:3	政府办公室下发了	2019/5/9 10:18:58	1	1	0	1	1	4
9	。天寒地冻的跑好	18/12/31 22:21	及设施设备采购等	2019/1/29 10:53:00	1	1	1	1	1	5
10	得到相关准确开工信	18/12/31 9:55:1	检查后，西地省楚	2019/1/16 15:29:43	1	1	1	1	1	5
11	桥等地方做立体绿化	18/12/31 9:45:1	要求完成了建设。	2019/1/16 15:31:05	0	1	1	1	1	4
12	局审批通过《温室养	18/12/30 22:30	地征收补偿款给原	2019/3/11 16:06:33	1	1	1	1	1	5
13	置房地下室近两万平	18/12/29 23:27	发[2014]7号文件	2019/1/29 10:52:01	1	1	1	1	1	5
14	大量从小区开车出去	18/12/29 11:55	具体选址，招标	2019/1/14 14:34:58	1	1	1	1	1	5
15	相关政府部门的大力	18/12/28 17:18	，已由银盆岭派出	2019/1/3 14:03:07	1	1	1	1	1	5
16	以上！天寒地冻，其	18/12/28 7:53:3	驶员工作时间长，	2019/1/14 14:33:17	1	1	1	1	1	5
17	https://baidu.com	18/12/27 15:18	路口两端各拆除20	2019/3/6 10:26:14	1	1	1	1	1	5
18	各种理由拒绝退货，	18/12/27 1:55:0	的信息进行投诉信	2019/1/3 14:02:47	1	1	1	1	1	5
19	建议在艺术中心先	18/12/26 16:51	期二期金菊路与雪	2019/1/14 14:32:40	1	1	1	1	1	5
20	早就施工，严重影响	18/12/25 19:35	由于需要夜间连续	2019/1/8 16:19:16	1	1	1	1	1	5
21	可以尽快合一。让社	18/12/25 16:23	，需三方或三方以	2019/1/4 15:48:23	1	1	1	1	1	5
22	、苹果等手机都无法	18/12/25 16:19	可请关注潇洒支付	2019/1/4 15:49:46	1	1	1	1	1	5
23	田。根据《土地管理	18/12/25 14:40	了土地补偿协议，	2019/1/8 16:18:00	1	1	1	1	1	5
24	车辆和行人通行，此	18/12/25 13:56	三十八条第一款第	2019/1/16 15:22:16	1	1	1	1	1	5
25	放频发。如果8路线	18/12/23 21:47	好，非常感谢您	2019/1/29 10:50:31	0	1	1	1	1	4
26	是否能在A市办理商业	18/12/21 11:01	持非本中心的缴存	2019/1/3 14:00:47	1	1	1	1	1	5
27	A市国际会展中心非	18/12/20 17:28	成约800米路基，	2019/1/4 15:47:36	1	1	1	1	1	5
28	修A3区山景区西大门	18/12/20 11:16	资计划调整，该项	2019/1/3 13:59:33	1	1	1	1	1	5
29	个多亿好远，这笔大	18/12/15 15:17	办的西地省洋兴置	2019/1/4 15:44:31	1	1	1	1	1	5
30	是这样操作的。梅溪	18/12/14 14:29	为A市A3区那么好	2018/12/29 15:05:11	1	1	1	1	1	5
31	道路一直没有修好，	18/12/12 8:57:1	，中学属于青雅丽	2019/3/15 15:40:09	1	1	1	1	1	5
32	，后面才想到在路	18/12/11 15:35	与该车实际停车时	2019/1/4 15:45:01	1	1	1	1	1	5
33	立，富吉又说在与住	18/12/11 15:23	3•18”“5•20”	2019/3/15 15:39:44	1	1	1	1	1	5
34	不说实话，如果县城	18/12/8 12:16:	细信息及证据，计	2018/12/27 9:23:01	1	1	1	1	1	5

图 44: 完整性分数图

	A	B	C	D	E
1	留言详情	留言时间	答复意见	答复时间	解释性得分
2	司却以交20万保证金	019/4/25 9:32:	理费，在业主大会	2019/5/10 14:56:53	0
3		19/4/24 16:03:	，且换填后还有三	2019/5/9 9:49:10	0
4	更是加大了教师的工	19/4/24 15:40:	教职工要依法签订	2019/5/9 9:49:14	1
5	A市，想买套公寓，	19/4/24 15:07:	下（含），首次购	2019/5/9 9:49:42	1
6	坡岭小学”，原“马	19/4/23 17:03:	”的问题。公交站	2019/5/9 9:51:30	0
7	屁巴冲到右边，越是	2019/4/8 8:37	有说明卫生较差的	2019/5/9 10:02:08	0
8	社区惠民装电梯的	19/3/29 11:53:	政府办公室下发了	2019/5/9 10:18:58	1
9	，天寒地冻的跑好	18/12/31 22:21	及设施设备采购等	2019/1/29 10:53:00	1
10	得到相关准确开工信	18/12/31 9:55:	检查后，西地省楚	2019/1/16 15:29:43	0
11	桥等地方做立体绿化	18/12/31 9:45:	要求完成了建设	2019/1/16 15:31:05	0
12	局审批通过《温室养	18/12/30 22:30	地征收补偿款给原	2019/3/11 16:06:33	0
13	置房地下室近两万	18/12/29 23:27	发[2014]7号文件	2019/1/29 10:52:01	1
14	大量从小区开车出去	18/12/29 11:55	具体选址，招标	2019/1/14 14:34:58	0
15	相关政府部门的大力	18/12/28 17:18	，已由银盆岭派出	2019/1/3 14:03:07	0
16	以上！天寒地冻，其	18/12/28 7:53:	驶员工作时间长，	2019/1/14 14:33:17	0
17	https://baidu.com	18/12/27 15:18	路口两端各拆除20	2019/3/6 10:26:14	1
18	各种理由拒绝退货，	18/12/27 1:55:	的信息进行投诉信	2019/1/3 14:02:47	0
19	建议在艺术中心先	18/12/26 16:51	湖二期金菊路与雪	2019/1/14 14:32:40	0
20	早就施工，严重影响	18/12/25 19:35	由于需要夜间连续	2019/1/8 16:19:16	0
21	可以尽快合一。让社	18/12/25 16:23	，需三方或三方以	2019/1/4 15:48:23	0
22	、苹果等手机都无法	18/12/25 16:19	请关注潇洒支付	2019/1/4 15:49:46	0
23	田。根据《土地管理	18/12/25 14:40	土地补偿协议，	2019/1/8 16:18:00	0
24	车辆和行人通行，此	18/12/25 13:56	三十八条第一款第	2019/1/16 15:22:16	1
25	故频发。如果8路线	18/12/23 21:47	好，非常感谢您	2019/1/29 10:50:31	1
26	是否能在A市办理商业	18/12/21 11:01	持非本中心的缴存	2019/1/3 14:00:47	0
27	A市国际会展中心非	18/12/20 17:28	成约800米路基，	2019/1/4 15:47:36	0
28	修A3区山景区西大门	18/12/20 11:16	资计划调整，该项	2019/1/3 13:59:33	0
29	个多亿好远，这笔大	18/12/15 15:17	办的西地省洋兴置	2019/1/4 15:44:31	0
30	是这样操作的。梅溪	18/12/14 14:29	为A市A3区那么好	2018/12/29 15:05:11	0
31	道路一直没有修好，	18/12/12 8:57:	，中学属于青雅丽	2019/3/15 15:40:09	1
32	，后面才想到在路	18/12/11 15:35	与该车实际停车时	2019/1/4 15:45:01	0
33	立，富吉又说在与住	18/12/11 15:23	3•18”“5•20”	2019/3/15 15:39:44	0
34	不说实话，如果是城	18/12/8 12:16	细信息及证据，请	2018/12/27 9:23:01	0

图 45: 可解释性分数图

答复内容（未转交）	答复得分
网友“UU008545” 您好！您的留言已收悉。现将有关情况回复如下： 涉及到租赁合同作为条件落户的是务工人员落户其中的一项。根据《A市常住户口登记管理规定》务工人员落户登记条件第二项：在A7县、A6区、A9市、A8县有合法稳定住所（含租赁）的人员，可以申请将本人及其共同居住生活的配偶、子女、父母的户口迁入居住地城镇地区。 此外，不是任何派出所都能办理，需前往您租赁房屋所在辖区的公安派出所办理。如您还有疑问，请来电咨询A市公安局人口与出入境管理支队0731-0000-0000000。 感谢您对我们工作的支持、理解与监督！ 2018年3月28日	4.025
网友“UU008374” 您好！来信收悉。现回复如下： 感谢您对我县规划建设工作的关注。 根据《开元中片控制性详细规划》，邮电大楼西南方向规划有部分商业用地。 2017年11月15日	4.023
网友“UU0082272” 您好！来信收悉。现回复如下： 根据《中共A市委办公厅A市人民政府关于印发<A市环城绿带生态圈建设工作实施方案>的通知》[政府发文]25号）文件精神，为全面深入推进三年造绿大行动，A7县政府办下发关于印发《A市环城绿带生态圈A7县段建设工作实施方案》（长县政办函（2016）1号），该项目实施范围涉及到我镇。按照县里文件规定，其中第二点建设任务中关于用地方式规定如下：各有关镇街可灵活采用“只征不转”、“土地流转”等多种模式，我镇采取的是“土地流转”模式。我镇计划流转土地约1500亩，现已有半数农户签订土地流转协议，环城绿带生态圈建设土地流转采取农户自愿原则，对签订土地流转的农户，政府采取按年发放土地流转费用，对有作物的土地补偿移栽费用。 感谢您对我镇工作的大力支持！ 2017年5月8日	4.019

图 46: 第一类得分 TOP3 的答复意见详细内容

答复内容（已转交）	答复得分
网友：您好！您反映的问题，我们已转交新河镇党委。2019年6月27日网友： 您好！关于“D12市新河镇兴旺村房屋征收补贴问题”已收悉，新河镇党委政府高度重视，迅速成立了调查小组进行了调查核实，现将调查结果回复如下： 一、关于反映“咨询D12市新河镇兴旺村房屋征收补贴”等相关问题 通过调查核实，修建**********县道没有经过新河镇兴旺村林角塘组，而是祁常高速修路，贺亮宏的房子在征地红线范围内，截止日前，由于贺亮宏家不肯与祁常高速服务组签订拆迁协议，所以房屋拆迁款无法发放。感谢该网友对我镇工作的支持和监督。 2019年10月25日	2.43
网友：您好！您反映的问题，我们已转交水口山镇党委。2019年5月8日网友：你好！ 关于你反映的“D12市水口山工人运动纪念园建设相关事宜”一事，已收悉。你所反映的情况，领导高度重视，第一时间安排工作人员调查了解。现将有关问题回复如下： 1.水口山工人运动纪念园征地工作已于2018年10月底全部完成。 2.房屋拆迁实物详查工作已全面完成，目前已签订房屋拆迁协议的49户，已签安置协议的31户（大渔村41户，松阳村28户，共69户）。 3.房屋拆迁款2019年5月6日已到12户，（已打款12户），房屋拆除工作准备启动。待中央审批后，同时征地款、拆迁款和住房补贴款全部到位，可启动文化园的建设！ 特此回复。 2019年5月8日	2.43
网友W0081447：反映意外伤的报销比例问题，我局领导特别重视，立即安排专人进行调查核实。现回复如下：3.4月28日下午职工医保中心领导、大地保险公司负责人立即上门与当事人见面，并作了政策宣传、讲解。当事人已经表示理解。感谢网友对医疗保险事业提出的宝贵意见。2019年5月5号网友：您好！你反映的问题，我们已转交医疗保障局。由于您未留下详细资料，故医疗保障局工作人员无法联系您，建议您直接拨打联系电话：0734－7237198。2019年4月28日网友：您好！您反映的问题，我们已转交相关单位。2019年4月28日	2.43

图 47: 第二类得分 TOP3 的答复意见详细内容