基于文本挖掘的网络“智慧政务”问题分析

摘要

随着网络问政的发展，以留言为代表的网络问政平台得到了快速的发展，但是随着反应民意的文本数据量不断攀升，给主要依靠人工来进行留言分类的相关部门的工作带来了极大挑战。

本文旨在通过建立留言文本适合的通用模型，得出留言的分类，构建相应的分类评估体系，找出某一时间段的“热点问题”，分析相关部门给出的答复与留言文本的相似度，提升网络问政的服务质量。

针对问题一，对群众留言信息进行处理，对所有词的词频进行统计，随后对词频进行降序排序，初步浏览分词，将无用的词加入到停用词库中以剔除，减小误差。按照原有标签对数据进行抽样，并建立通用模型。用 python 中机器学习模块对每级标签处理后的数据进行模型训练和测试，利用 F-Score 评价法对模型中的分类方法进行评价，当达到一定置信程度后，即可得到优化模型及优化后的分类方法，处理文本数据，以便将留言分派至相应的职能部门处理。

针对问题二，首先确定包含点赞数，反对数以及留言数量的热度指标体系，并根据实际数据计算出网络上各问题在各指标的取值。利用层次分析法，求出相应权重，利用此权重建立新的热度算法模型，对基于问政平台的留言问题热度值进行计算，求出热度趋势值，得到热度趋势图。用 python 算法对文本相似程度高的文本进行筛选和整合，初步生成包含热点问题表和热点问题明细表的热点表。通过计算热度指数按照高低对热门问题排序绘制出热点问题明细表，并依照关键词以及时间范围对明细表整理得到题中所求热点问题表。

针对问题三，通过对原数据进行处理，从三个角度进行分析。将留言问题和答复置入训练好的模型当中，可以给问题留言和答复分别进行一级标签分类，对问题和文本之间的相似度进行分析；再对答复时间和留言时间进行对比，可以分析出答复的即时性评价；最后对答复的标准化词语进行筛选得到答复完整性得分评估。依据变异系数法建立模型，从而给出一套从相关性、完整性、即时性等角度进行评价的评价模型，对答复意见的质量以与评分。

关键词：F-Score 评价法，高斯朴素贝叶斯算法，层次分析法，文本相似性热度统计

目录

1.网络问政留言平台中的问题及具体分析

1.1 问题叙述

1.针对网络问政平台中大量的群众留言，仅依靠人工根据经验对留言进行分类，存在工作量大、效率低，且差错率高等问题。根据题目本身要求需建立关于留言内容的一级标签分类模型，对留言进行分类处理，以便后续将留言分派至相应的职能部门处理。

2.网络问政留言平台中群众问题冗杂，工作人员需要优先处理被大量反应的热点问题。因此需要根据热点问题的定义，针对群众留言，对问题以及留言文本进行识别分类，建立合理的热度评价指标，为政府提供排名前五的热点问题和具体留言信息，以便优先处理，提高群众满意度。

3.当相关部门针对留言问题给出相应答复后，从相关性，完整性，可解释性等角度对答复意见做出一套评价方案，构建相应模型和指标来计算和评价该方案，以便提高网络问政的服务质量。

1.2 问题具体分析

1.2.1 群众留言分类的分析

针对群众留言分类问题，对群众留言信息进行深度处理，按照原有标签对文本进行抽样，将文本转化为数据信息，并建立通用模型。用 python 中机器学习模块对每一级标签的处理后的数据进行模型训练和测试，利用 F-Score 评价法（综合评价指标）对模型中的分类方法进行评价，当达到一定置信程度后，即可得到优化模型及优化后的分类方法，在一定程度上解决工作量大等问题。针对政务系统影响范围和深度的分析，要开发一套面向政务系统留言问题热度分析系统，系统的主要功能如图 1-1 所示。其中包括管理平台、文本数据采集器、留言详情分析器、分析平台和相关库。管理平台主要是对文本数据采集器、留言详情分析器等进行集中管理和控制，主要功能包括网络留言事件管理、留言状态和留言分类等属性进行配置等；文本数据采集器是根据配置，通过给定文本数据，检

索等待分析留言内容，自动采集检索结果，对留言内容进行智能解析，自动去重，消除冗余，抽取相关留言文本信息内容，转换成结构化数据，存储到相关库中进行分析，供留言详情分析器调用；留言详情分析器是根据留言的热度计算模型，对库中的留言详情数据进行统计分析，并将计算结果提供给分析平台使用；分析平台主要是对留言事件的热度和数据源进行可视化分析，热度以指数形式进行显示。但我们的分析过程中可能存在一些难点：

①文本语义带来的词语交叉，比如:交通局的亲属拖欠我们工资

②多分类问题带来的难度，转化为多个二分类分析

③数据不平衡带来的影响，进行数据增强

④长文本的无意义表达太多，需要转为短文本、关键句

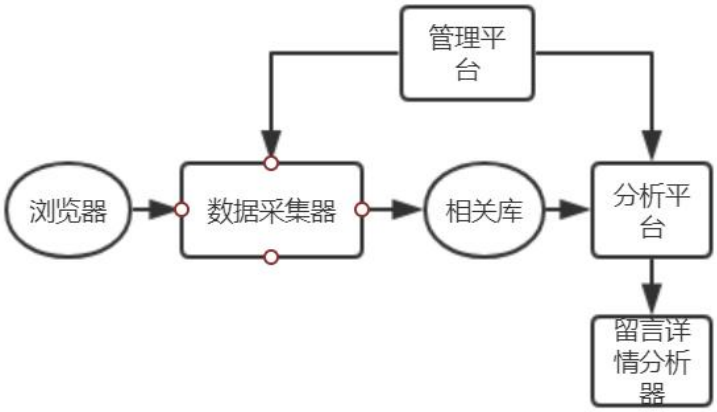

图 1-1 网络平台分析留言流程图

1.2.2 热点问题挖掘的分析

针对热点问题挖掘，给出“热点问题”的定义，进行问题识别，思考如何从众多留言中识别出相似的留言得到群众反映较为集中的几个问题。其次对“热点问题”的留言信息进行分析和分类，把特定地点或人群的数据归并，把相似的留言归为同一问题，进行热度评价。定义热度评价指标和计算方法，对指标排名，为得到一个热度排行，可根据点赞数和反对数以及问题出现频率等因素，建立热度算法模型，考虑其中各因素间的联系，确定热度指数算法，筛选和整合相似程度高的文本并按留言数量排序，通过计算热度指数按照高低对热门问题排序绘制出热点问题明细表，并按照关键词以及时间范围对明细表整理得到热点问题表。但是地点、人群的表达多样化，在识别上增加了一定的难度。相似的计算复杂，特征多、两两之间计算相似计算量大也是一大难题。

1.2.3 答复意见评价的分析

针对答复意见的评价，通过对留言文本和答复内容之间的相似度进行分析，进而在每一条问答消息中建立通用模型，给出一套评价方案，且从相关性（答复意见的内容是否与问题相关）、完整性（是否满足某种规范）、可解释性（答复意见中内容的相关解释）等角度进行评价，如若相似度高则可说明相关性程度高等，从而对答复意见的质量以与评价。那如何将相关性、 完整性、可解释性等描述量化，构建什么指标来计算和评价需要我们深入思考。

2. 数据预处理及原理概要

2.1 数据处理流程

（1）建立通用模型，导入测试数据。

（2）对无用行列进行剔除并删除冗余数据。

（3）对留言详情中 7 个标签进行标注，从每个标签中抽取一定数量文本作为样本，合并为一个样本集。

（4）为提高分词后样本的准确性，删除所有留言中出现次数较多且对分析无用的“停用词”。

（5）对样本集中留言文本进行分析，采用 python 中国内较为有名的结巴库对所有留言详情进行分词处理，将每一句话按库中定义方法分成一个个词语。

2.2 结巴分词原理概述

（1）基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词的情况所构成的有向无环图 (DAG)。

（2）采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。

（3)对于未登录词,采用了基于汉字成词能力的 HMM 模型,使用了 Viterbi 算法。

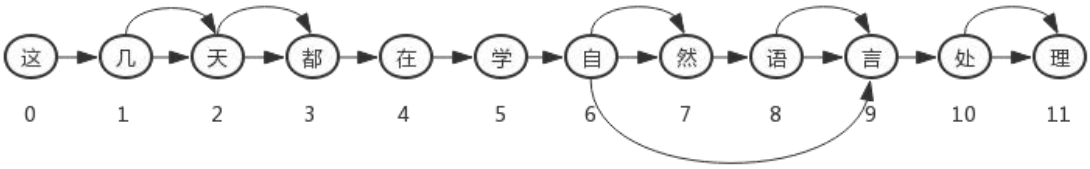

图 2-1 成词有向无环图

3. 问题解决及结果

3.1 基于机器学习的群众留言分类

3.1.1 分词统计

首先，遍历已准备好的样本数据，对所有词的词频进行统计，随后对词频进行降序排序，初步浏览大部分词，将显而易见的无用的词加入到停用词库中以剔除，减小对模型训练造成的累计误差。下面以标签为“城乡建设”的样本数据为例，绘制词云图以及词频表。

由词云图（图 3-1）和词频表（表 3-1），结合主观意识可以看出，这几个频数较高的分词确实从某一种程度上与“城乡建设”标签相联系。

表 3-1（城乡建设）词频表

分词	频数
业主	853
小区	757
开发商	485
政府	483
建设	441
部门	420
房屋	413

图 3-1（城乡建设）词云图

3.1.2 模型训练

为了建立一个能自动识别并对所有测试文本进行标签分类的模型，利用 TF-IDF 权重策略，即词频逆文档权重方案。如果某个词或者短语在一条留言文本中出现频率高，而在其他留言文本中很少出现，则认为此词或者短语具有很好

的类别区分能力，以此作为依据进行分类，具有较高的说服力。此外可以通过增加词频信息以及对向量进行归一化处理，避免句子长度不一致的问题。权重策略文档中的高频词具有表征此文档较高的权重，若该词是高文档频率词，则具有一定的代表性。

首先将分词转换成词频向量，后转换成 TF-IDF 权重矩阵，提取特征，构建模型，利用高斯朴素贝叶斯算法及 sklearn 库对模型进行训练。

3.1.3 朴素贝叶斯原理概述

在高斯朴素贝叶斯中，每个特征都是连续的，并且都呈高斯分布（正态分布）。图像如图 3-2 所示：

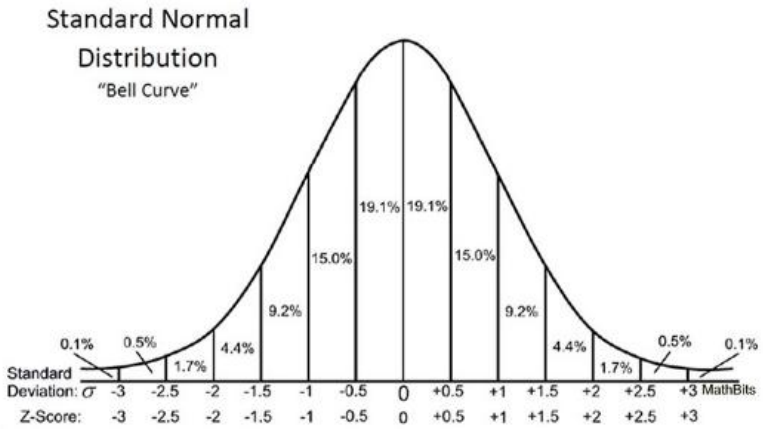

图 3-2 正态分布图

GaussianNB 实现了运用于分类的高斯朴素贝叶斯算法。特征的可能性(即概率)假设为高斯分布:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2})$$

参数σ_y 和μ_y 使用最大似然法估计。

3.1.4 模型评价及通用模型构建

将抽取的样本分为两类，80%为训练样本，20%为测试样本。通过朴素贝叶斯训练和测试后，导出对应的一级标签分类结果如下表所示：

利用 F-Score 评价法（综合评价指标）对模型得出的测试结果的分类方法进

行评价，其中涉及到两个度量值：查准率和查全率，用这两个度量值来评价结果的质量。定义式如下：

1.查准率 = 提取出的正确文本信息条数 / 提取出的总文本信息条数

$$即\quad precision = \frac{tp}{tp + fp}$$

2.查全率 = 提取出的正确文本信息条数 / 样本中的文本信息条数

$$即\quad recall = \frac{tp}{tp + fn}$$

（两者取值在 0 和 1 之间，数值越接近 1，查准率或查全率越高。）

3.f1 = 查全率 * 查准率 * 2 / (查准率 +查全率)

$$即\quad \mathrm{f1} = \frac{2 * p * r}{p + r}$$

（f1 的值即为查准率和查全率的调和平均值）

据 sklearn 库中的 classification_report,可直接得到三者结果如表 3-2 所示：

表 3-2 综合评价得分表

	Precision（查准率）	Recall（查全率）	**F1-score**	Support
交通运输	0.77	0.37	**0.50**	91
劳动和社会保	0.71	0.77	**0.74**	304
卫生计生	0.78	0.52	**0.63**	117
商贸旅游	0.67	0.64	**0.66**	188
城乡建设	0.65	0.78	**0.71**	303
教育文体	0.72	0.80	**0.76**	244
环境保护	0.83	0.75	**0.79**	155
精度			**0.71**	1402
宏平均	0.73	0.66	**0.68**	1402
加权平均	0.72	0.71	**0.71**	1402

通过表 3-2 分析可得知：环境保护的 f1-score 分数达到了 0.79，可以说明对于检索出来的留言文本绝大多数是准确的，但对于交通运输的 f1-score 分数仅只有 0.50，二者之间存在较大误差。造成此误差的重要原因是模型的训练仍

有一定的缺陷。综合查准率和查全率以及 f1-score 的平均值来看，此模型的本身建立并无问题，需要更加精细化才可将模型训练到投入使用。

将留言利用我们建立的模型进行初步的一级标签分类，为验证分类结果的准确性，我们在所有分类好的留言数据中，每个标签所有数据的 80%作为样本，将样本的 80%用于机器训练，剩余的 20%作为测试样本，以上过程进行三次后取平均值，以减小误差，得到相应的结果如下表 3-3：

表 3-3 模型测试结果表

训练次数	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生	总计
1	92	92	89	95	91	90	79	628
2	92	81	84	93	91	80	83	604
3	93	96	88	94	86	85	84	626
平均值	92.333	89.667	87	94	89.3333	85	82	619.333

通过表 3-3 我们可以看出，得出的结果与我们所假设的结果吻合率高达 80%，这证明我们建立的模型的可行性较高。吻合度未达到 90%以上一定程度上因为数据量较小，无法避免更加精细的误差。

表 3-4 测试样表展示表

留言编号	留言用户	留言主题	留言时间	留言详情	正确标签	测试结果
161743	U0001157	转此件，还	12/1/122:30:	生活带来了	环境保护	劳动和社会保障
117649	U0006711	广告公司喷	7/9/2517:02	了，是一到家	环境保护	环境保护
117537	U0005348	新村的机碳厂	8/12/115:20	上上班。发出	环境保护	环境保护
161372	U0007189	保部门打游击	8/11/211:19	污染，排放不	环境保护	环境保护
161761	U0004574	农村小纸厂大	1/9/317:44:	造纸厂搬到我	环境保护	环境保护
60611	U000575	邦化工厂直接	3/7/1312:21	不可能天天守	环境保护	环境保护
161343	U0004599	地省明珠选	9/3/1714:10	当作地方政府	环境保护	环境保护

3.2 基于相似文本热度统计算法的热点问题挖掘

3.2.1 热度指数模型建立

首先通过分析留言在问政平台中的点赞数，反对数以及问题出现的次数用层次分析法得到三者对应的权重，利用此权重构建新的模型，对基于问政平台的留言问题热度值进行计算，求出热度趋势值，得到热度趋势表，接着在此基础上，导出留言问题热度表。

再通过对“热度”一词分析，不妨假设留言数量 Q，点赞数 M，反对数 N 三个因素中存在一定模型关系。在对于群众问政留言体系当中，由于民生问题，点赞数和反对数都可以认定为该问题受到了一定的关注，即存在一定的热度。可认为这三个因素都与热度成正比关系，给出权重系数分别为 a，b，c。基于 Hacker News 社区网站对投票的热点排序算法，经过改进和调整得到适用于我们模型的排名算法，公式如下：

$$score = \frac{a \cdot Q + b \cdot M + c \cdot N}{T^{G}}$$

利用层次分析法通过查阅相关浏览器资料[6]和有关热度排序权值可计算出 a,b,c 权重值以及相关系数分别为：

表 3-5 层次分析法权重及相关系数表

a	b	c	CI	CR
0.7306	0.1884	0.0810	0.0324	0.0624

因为 CR ＜ 0.1,所以可以认为判断矩阵的一致性检验可以接受，故可采用此权重模型对热度指数进行计算。

其中 T 表示第一条留言与最后一条留言相间隔的时间（单位：天）

G 为“重力因子”（gravityth power），指随着时间间隔差的不断增加，让热度不断的衰减的因素。

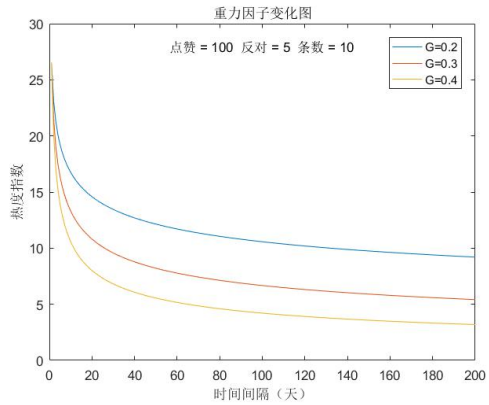

图 3-3 重力因子变化图

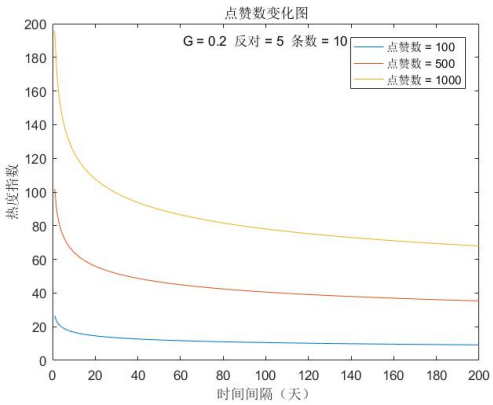

图 3-4 点赞数变化图

如图 3-3,3-4 所示，分析热度指数模型，共有三个因素影响热度指数的大小：

第一个因素是留言数，点赞数和反对数的综合分析值；第二个因素是第一条留言与最后一条留言的时间差 T；第三个因素是重力因子 G，它的数值大小决定了热度随时间下降的速度。

在保证其他条件不变的情况下，通过图示可得点赞数越多，热度越高；随着时间间隔推移热度逐渐衰退；对于重力因子 G 考虑到问政系统对于热度的应该衰退的较为缓慢，我们经过测试选出热度下降最为平缓的结果 G=0.2。

通过建立此模型，对相关文本分类后，可以计算每一热点信息所有的热度，如图 3-5 对候选热度点的展示，最后对热度进行排名可以得到“热点问题表”。

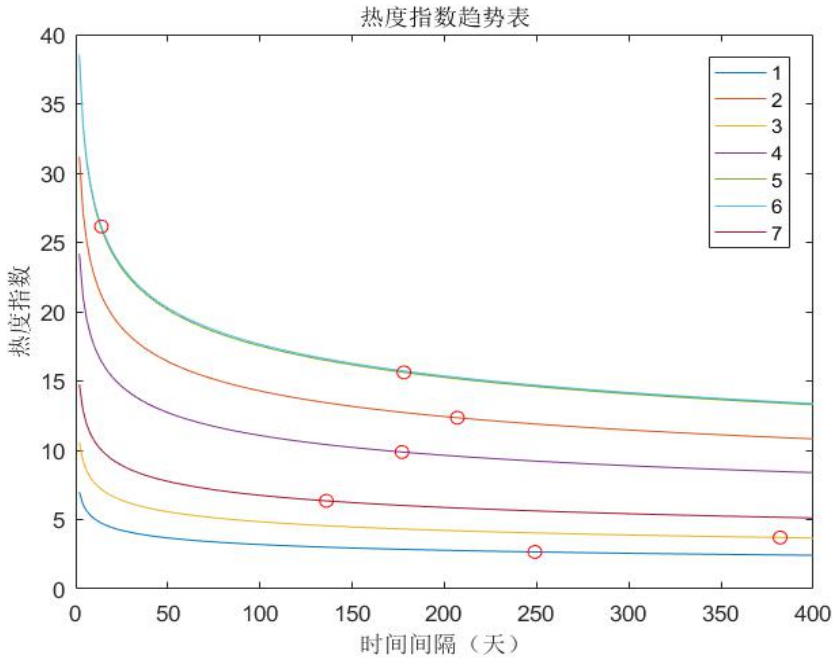

图 3-5 热度指数趋势图

3.2.2 热点留言算法原理

算法实现简要介绍：

（1）分析样本数据，通过 openpyxl 获取数据信息，通过 jieba 分词进行处理，利用 jieba 分词和自定义词典以及排除信息，形成一个二维数组。

（2）使用 gensim 中的 corpora 模块，将分词形成后的二维数组生成词典。

（3）将二维数组通过 doc2bow 稀疏向量，形成语料库。

（4）使用 LsiModel 模型算法，将语料库计算出 Tfidf 值。

（5）获取词典 token2id 的特征数。

（6）计算稀疏矩阵相似度，建立一个索引。

（7）读取 excel 行数据，通过 jieba 进行分词处理。

（8）通过 doc2bow 计算测试数据的稀疏向量。

（9）求得测试数据与样本数据的相似度。

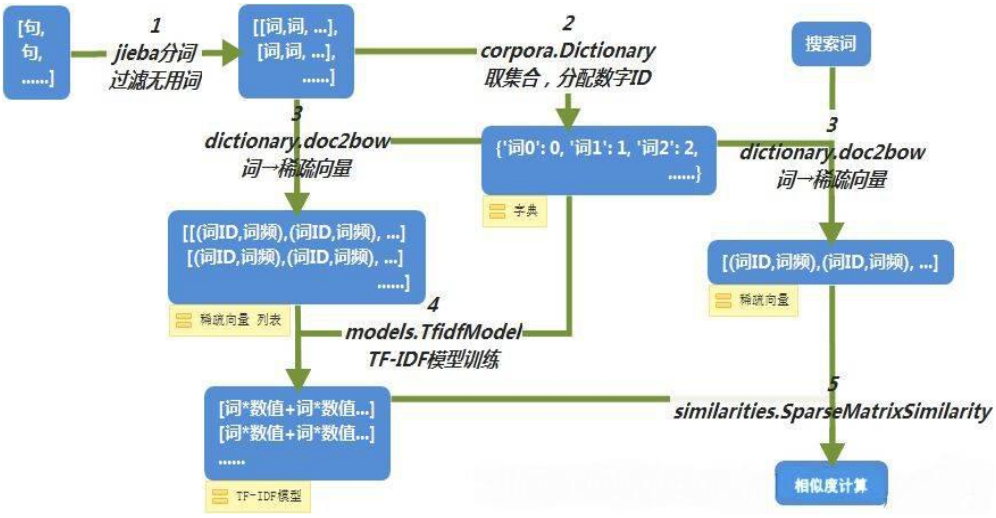

图 3-6 算法流程图

因数据量庞大，仅用上述算法直接对所有数据同时分析时，相似度处理效果较差。在此情况的基础上，决定利用问题一模型适用性强的特点，将问题二的数据首先带入问题一的模型中，使其分类于城乡建设，交通运输等产生共七类子数据。大大减少热点留言算法中的单次相似度处理量，提高算法的精度。

表 3-6 子数据示例表（环境保护）

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	一级标签
188546	A0006817	A2区佳兆业	2019/1/2313	敬爱的领导	0	0	环境保护
188774	A00048792	A2区政府东	2019/6/1823	多年来A2区	0	0	环境保护
188780	A00094754	请依法解决	2019/10/201	尊敬的县领	0	0	环境保护
188809	A909139	A市万家丽商	2019/11/191	A市万家丽商	0	1	环境保护
188876	A00013435	咨询A7县榔	2019/2/2611	尊敬的领导	0	0	环境保护
189587	A00018292	对A8县高新	2019/1/1318	我们是西地	0	0	环境保护
189635	A00029819	A1区桐阴里	2019/7/1521	桐阴里夜间	0	0	环境保护
190108	A909240	丽发新城小	2019-12-21	丽发新城小	0	1	环境保护
190196	A00060165	西地省师大	2019/10/111	自从师大附	0	0	环境保护
190213	A00031618	请A市加快	2019/1/1612	地处时代倾	0	0	环境保护
190523	A00072847	A市丽发新城	2019/12/261	领导您好！	0	0	环境保护

将7类处理好的子数据分别带入热点留言算法中，根据每类数据调整sim（相似度函数）的控制值，通过算法自动创建一个“热点表”。如图所示，在表（留言热点表，以环境保护为例）中，会自动将留言主题汇总，相关留言数累加，分析出主要关键词创建一个带导航链接的表格。

表3-7 导航链接示例表

序号	留言主题	留言数量	导航-链接到留言热点明细sheet表
1	洛丽发新城居民区附	11	A市万家丽南路丽发新城居民区附近搅拌站扰民
2	桐阴里小区一直夜间	9	A1区桐阴里小区一直夜间施工!
3	城附近修建搅拌厂，产	8	A2区丽发新城附近修建搅拌厂严重污染环境!
4	路附近工地昼夜施工	7	A2区刘家冲路附近工地昼夜施工严重扰民!
5	近违规乱建混凝土搅	5	A2区丽发新城附近违规乱建混凝土搅拌站谁来监
6	城小区临街门面油烟直	4	魅力之城小区临街门面油烟直排扰民!
7	鹰小学附近噪音扰民影	4	A4区金鹰小学附近噪音扰民影响上课!
8	的A2区丽发新城附近	4	噪音灰尘污染的A2区丽发新城附近环保部门不作
9	加快商业中心建设刻	4	请A市加快商业中心建设刻不容缓!
10	加快自来水深度净化改	3	请A市加快自来水深度净化改造力度!
11	违建搅拌站，彻夜施工	3	A市丽发新城违建搅拌站彻夜施工扰民污染环境!

点击表中的任意热点明细可以得到表（A市万家丽南路），如表3-7，3-8所示，可以明确的看到热点第一条的热点明细展示，均为万家丽南路丽发新城居民区附近拌站扰民问题。其他消息均可以此法查阅明细内容。因相似性分类得到的结果难以在一条中涵盖所有问题，但是很完善的提供了问题指向，可以通过后续操作快速找到所有问题。

表 3-8 子连接示例表

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
188809	A909139	A市万家丽南路丽发新城居民区附近搅拌站扰民	2019/11/1918	A市万家丽南路	0	1
190108	A909240	丽发新城小区旁边建搅拌站	2019-12-2115	丽发新城小区旁	0	1
213464	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民	2019-12-1012	我是暮云街道丽	0	0
217700	A909239	丽发新城小区旁的搅拌站严重影响生活	2019-12-212:	开发商把特大型	0	1
233158	A909242	丽发新城小区旁建搅拌厂严重扰民！	2019-12-58:4	本人是丽发新城	0	0
243692	A909201	丽发新城小区附近的搅拌站噪音严重扰民	2019-11-1511	领导您好！我小	0	2

3.2.3 热点表的建立和展示

依据题目中的样本表展示，用模型对样本数据处理后，通过对算出来的七类热点表进行人工的统计和修正，给出留言数量最多的若干类问题作为可能热点问题。对点赞数远大于其他值的问题一并提取作为可能热点问题，所有汇总为可能热点问题，将所有可能问题的留言数量，点赞数，留言时间间隔数据放入热点指数模型中，取热点指数排名前五的问题，整理最终可以得到以下题目所求表 3-9，表 3-10：

表 3-9 排名前五的热点问题表

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	26.15	2019/1/11至2019/7/8	西地省A市案件受害人	A市58车贷特大集资诈骗案
2	2	15.62	2019/8/23至2019/9/6	A4区绿地海外滩小区居民	A4区绿地海外滩小区距长赣高铁最近只有不到30米
3	3	12.35	2019/7/3至2020/1/26	A2区丽发新城小区居民	A2区丽发新城小区搅拌厂垃圾噪音污染
4	4	9.88	2019/7/7至2019/12/31	A市伊景园滨河苑居民	A市伊景园滨河苑捆绑销售车位
5	5	6.35	2019/7/21至2019/12/4	A5区劳动东路魅力之城小区	A5区劳动东路魅力之城小区门面油烟噪音扰民

因为每个热点问题的留言都有很多，针对每个问题的留言明细表已放在附件中，仅在此展示每个问题的一条留言详情。

表 3-10 热点问题留言明细表（简表）

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	尊敬的胡书记：您好!A4区p2p公司58车贷，非法经营近四年。在受害人要求下，于去年8.20立案侦察，至今已6个月整。未发一字立案公告和案件进展财产处置通报…	0	821
…	…	…	…	…	…	…	…
2	263672	A00041448	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？	2019/9/5 13:06:55	您好，近日看到了渝长厦高铁最新的红线征地范围以及走向经过，其经过北三环的地方紧挨着绿地海外滩小区二期，我测算了一下距离，最近的位置只有30米不到，这严重…	0	669
…	…	…	…	…	…	…	…
3	208714	A00042015	A2区丽发新城附近修建搅拌站，污染环境，影响生活	2020-01-02 00:00:00	尊敬的领导：您好！作为一名居住在A2区丽发新城的业主，和小区内的每一位业主一样，最近我们正面临一个十分头痛的问题：我们小区附近百米范围内修建了搅拌厂，严重污染…	0	4
…	…	…	…	…	…	…	…
4	276016	A909181	车位属于业主所有，不应该被捆绑销售！	2019-08-06 00:00:00	尊敬的胡书记，您好！我叫陈玉春，身份证号**********，现实名投诉广铁集团铁路职工定向商品房伊景园滨河苑在销售中捆绑房子和12万的车位一对一销售…	0	2
…	…	…	…	…	…	…	…
5	272122	A909113	A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处	2019/08/01 16:20:02	局长：你好，A5区劳动东路魅力之城小区的一楼，开了几家夜宵快餐店，厨房内多个灶台油烟随意排放，有的店灶台烧的柴火，气味难闻。您不难想象，用餐时段没有净化的柴火味，烧烤炭火味…	0	6
…	…	…	…	…	…	…	…

3.2.4 算法模型的优缺点

（1）优点：算法本身具有较高的自动化程度，可以自动生成所需数据表格，超链接功能的加入使得数据更直观的呈现。

（2）缺点：模型训练的程度不够完善，得到的数据结果仍需要人工调整修正，但是相对于传统人工分类而言，大大减少了工作量，一定程度上达到了提高工作效率的目的。

3.3 答复意见评价体系构建

3.3.1 评价模型建立

通过对现有的文献以及问题的分析，将从三个角度：相关性、即时性、完整性建立模型。给三者赋予相应权重，此处利用变异系数法。变异系数法是直接利用各项指标所包含的信息，通过计算得到指标的权重。在评价指标体系中，指标取值差异越大的指标，也就是越难以实现的指标，这样的指标更能反映被评价单位的差距。公式如下：

$$v_i = \frac{\sigma_i}{x_i}$$

（v_i是指标的变异系数（标准差系数）；σ_i是指标的标准差；x_i是指标的平均数。）

各项指标的权重为：

$$w_i = \frac{v_i}{\sum v_i}$$

3.3.2 评价模型求解

（1）相关性：通过对原数据进行处理，将留言问题和答复置入训练好的模型当中，可以给问题留言和答复分别进行一级标签分类，若二者预测的一级标签相同则可以判断为答复的相似度较高，可说明相关性强。在进行一次完整的训练预测后，我们又对依次剔除一个一级标签内容的样本进行共计 8 次训练和预测，以减少可能因训练次数过少存在的误差。将测试样本中的数据进行分类对比后，用 0-8 表示两者预测分类相同的次数，即相关性得分。

表 3-11 相关性得分表

编号	留言主题	留言详情	答复意见	留言分类预测	答案分类预测	相关性	相关性总得分
44203	咨询C市电动车免	听说这个月中旬	您好，根据《	商贸旅游	交通运输	0	4
41479	投诉C2区国际商贸	关于C2区国际商	您好！ 接到您	城乡建设	城乡建设	1	7
72278	反映E9县长铺镇乡	我是居住在绿洲	您好！	城乡建设	城乡建设	1	6
87831	投诉J市邦万装修	我投诉J市邦万装	您的留言已收悉	劳动和社会保障	劳动和社会保障	1	5
87827	J市金伯利大厦小	关于反映J市金伯	您的留言已收悉	城乡建设	劳动和社会保障	0	0
144848	投诉M5市汇源煤气	保护环境人人有	网友您好：您的	环境保护	教育文体	0	0
32599	呼吁B2区千亿大道	B2区千亿大道和	“UU0082382”	城乡建设	劳动和社会保障	0	1
129504	强烈呼吁L4县安北	强烈呼吁老市场	“UU008468”	城乡建设	劳动和社会保障	0	2
47549	建议让C市每个学	为什么现在初中	“UU0082117”	教育文体	教育文体	1	8

（2）即时性：根据答复时间和问题的留言时间算出相应的时间差，通过对回复时间差分布情况的分析，得到如图 3-7：

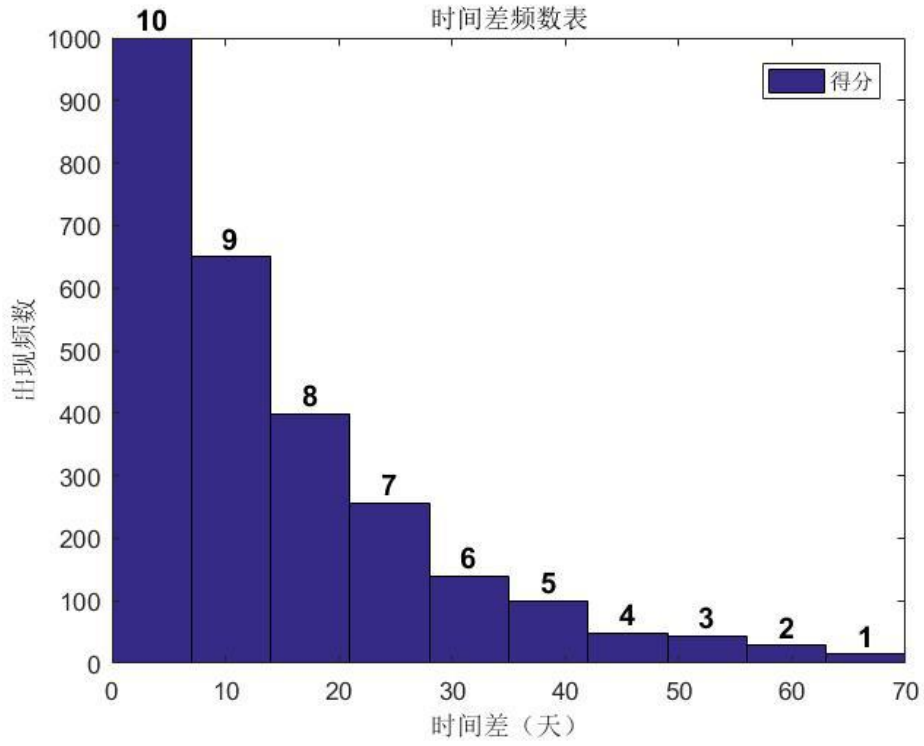

图 3-7 时间差频数直方图

根据回复时间差分布情况，建立时间差评分体系如表 3-12：

表 3-12 时间差评分体系表

时间差（天）	0-7	7-14	14-21	21-28	28-35	35-42	42-49	49-56	56-63	63-70	>70
时间差评分	10	9	8	7	6	5	4	3	2	1	0

根据时间差评分体系，得出所有问题的时间差评分，得到表 3-13：

表 3-13 时间差评分明细表

编号	留言主题	留言详情	答复意见	留言时间	答复时间	答复时间差	时间差评分
2549	华苑物业管	却以交20万保	费，在业主大	/4/25 9:3	5/10 14:5	15.22550926	8
2554	洛洋湖段怎么	意带来很大景	且换填后还存	/4/24 16:0	/5/9 9:49	14.73993056	8
2555	市民营幼儿园	是加大了教师	职工要依法	/4/24 15:4	/5/9 9:49	14.75636574	8
2557	享受人才新政	想买套公寓	（含），首次	/4/24 15:0	/5/9 9:49	14.77930556	8
2574	交站点名称含	小学”，原	的问题。公交	/4/23 17:0	/5/9 9:51	15.70012731	8
2759	浦镇马路卫	冲到右边，就	明卫生较差	8:37:20	/5/9 10:0	31.05888889	6
2849	小区盼望早	区惠民装电梯	府办公室下发	/3/29 11:5	/5/9 10:1	40.93443287	5
3681	湾社区居民的	天寒地冻的路	设施设备采购	/12/31 22:	/1/29 10:5	28.52153935	6
3683	住宅楼无故停	相关准确开立	查后，西地	/12/31 9:5	/1/16 15:2	16.23244213	8
3684	洋湖壹号小区	地方做立体绿	求完成了建	/12/31 9:4	/1/16 15:3	16.23965278	8
3685	托街道大托铺	审批通过《温	地征收补偿款	/12/30 22:	/3/11 16:0	70.73336806	0

（3）完整性：由主观意见定义几个答复评价的关键词，对相关答复进行完整性分析，通过寻找创建出标准格式词表，建立一个完整答复标准模型进行评分，使用 Excel 中的 FIND 函数以及 IFERROR 函数可以得到每一条答复的完整性评分，相应的评分如表 3-14 所示：

表 3-14 完整性得分表

编号	留言主题	留言详情	答复意见	完整性评分
44203	动车免费上	免费上牌办理	时号牌注册	2
41479	国际商贸城	责任”。对买	收即将完成，	3
72278	长铺镇交通通	了大量的车辆		4
87831	邓万装修公司	集搬迁项目工	题转相关部门	2
87827	居业主合法权	和热水，但	题转相关部门	2
144848	公司排放刺	皮，严重违	已通过网络电	4
41481	一中每个周末	，而且寒假	局关于做好	4
87841	山景区交通事	产，房屋也涨。另外，鉴		3
37211	石奄村叶金砰	的事故像这两	路交通安全	4
92936	房屋办证登记	去办理。于是	财产公证，如	5
127914	景秀龙园小区	龙园小区入	已转交至住建	3

表 3-15 完整性评分表

标准格式词	分数
网友/网民	1
祝	1
您好/你好	1
谢谢/感谢	1
尊敬/亲爱	1
回复如下	1
收悉	1

（4）在综合评分时，为将相关性进行更明显的数据化表示，得出相应指标权重分配结果展示：

表 3-16 指标权重表

	个案数	平均值	标准差	变异系数	**权重**
答复时间差	2816	7.10	2.473	0.348	**0.204**
完整性评分	2816	4.29	1.687	0.393	**0.231**
相关性评分	2816	5.18	4.997	0.964	**0.565**

由表 3-16 可以看出相关性的占比最大，而完整性和时间差的权数相对较小，反映出相关部门给出的答复意见，与问题的相关性还是较吻合的，但是在某种规范或者在及时回复的情况上情况不太乐观。总结以上数据可以发现，相关部门的回复还是具有一定的缺点：一是政府回复及时性不强；二是政府回复内容的针对性弱，象征性或形式化回应现象较为明显；三是政府回复效果很差，会有部分答非所问现象。

表 3-17 为最终通过此评分模型计算出的样表展示，总评分结果越高越能够说明答复评价比较令人满意，如果分数较低问政平台的答复可能需要从三个角度进行查漏补缺。

表 3-17 综合评分表

编号	留言主题	留言详情	答复意见	留言分类预测	答案分类预测	相关性	留言时间	答复时间	答复时间差	完整性评分	相关性总得分	时间差评分	综合评分
126597	且鑫源百货对	重扰民，影	板已在长乐	城乡建设	城乡建设	1	/3/12 14:2	/3/18 17:	6.12177	7	8	10	8.189
16799	各中小学的	，但据天气	初三年级期	教育文体	教育文体	1	/1/22 15:3	/1/23 16:2	1.03316	7	8	10	8.189
16176	照中学课程	部改为主课，	要求其必须	教育文体	教育文体	1	/12/4 18:0	/12/12 10:	7.68417	7	8	9	7.945
18596	的房子问题多	尊敬的政府领导：	好！来信收	城乡建设	城乡建设	1	/10/15 9:1	/10/21 9:	6.00269	6	8	10	7.89
36222	师希望今年职	称；同期	标准，中级	教育文体	教育文体	1	12/21 15:	12/27 10:	5.78082	6	8	10	7.89
4134	业医师资格	导致资格证发	与省卫计委	劳动和社会保障	劳动和社会保障	1	11/14 15:	/11/20 9:4	5.76657	6	8	10	7.89
20258	在A市定点医	尊敬的领导：	，基金支付	劳动和社会保障	劳动和社会保障	1	10/23 16:	10/29 11:	5.75955	6	8	10	7.89
19633	长浏大通道间	的盼盼路，	里的方案，因	城乡建设	城乡建设	1	/6/18 20:4	/6/24 14:4	5.75063	6	8	10	7.89
25567	时代中心房屋	小区人车不	供电、供气	城乡建设	城乡建设	1	/3/11 16:3	/3/17 9:1	5.69529	6	8	10	7.89

3.3.3 评价模型的优化

（1）由于一级标签分类比较法，存在误差较大，导致一级标签对答复意见的分类不准。需要建立一个答复体系模型，对答复文本进行训练，再和留言问题进行对比，使用机器学习算法对相似度进行分析统计。

（2）对于完整性和即时性的评分具有一定的主观因素，需要寻找一个合理的量化方式将打分按照一定的方式自动化，生成相应答复建议的评分，给网络问政平台提出缺点和改进的方式。

4.模型评估

4.1 模型的优缺点

（1）模型的优点

（ｉ）该模型针对现有数据，有完整的训练模型，且具有一定的通用性，测试文本无需大幅调整格式，直接将测试文本置入程序当中，可以直接生成所需表

格。对于第一问可以生成分类好标签的的文件；对于第二问可以汇总后生成热点问题表和热点问题明细表。

（ii）热度模型和评价模型具有说服力，可以准确的将文本按照热度顺序进行排列，评价模型通过变异系数算出的权数，可以对答复进行合理的评分，辨析评价是否合规，从而改进答复方式。

（2）模型的缺点

（ i ）由于是文本分析而不是直观的数据分析，中国汉字博大精深将其转换成数字矩阵进行分析只能寻找出文本的相似性，很难对留言中可能包含的深层含义和情感进行理解分析，当前模型只能尽可能解决人工繁杂的文本标签分类工作，文本筛选整合工作。

（ii）鉴于数据的限制，只使用朴素贝叶斯等机器学习模型进行训练，模型可能存在的训练不到位导致误差无法消除，答复与留言内容的相关性分析不够完善，相关系数可能需要重新建立更合理的模型。

（iii）由于资源和人力所限，只是提取了部分数据作为样本，进行模型的研究，数据比较粗糙。因此对于留言问题热度模型的验证显得比较薄弱。并且对于留言问题类别进行了初步的划分，但由于数据量不甚理想，因此划分不够准确。

4.2 模型的推广与改进

（1）模型的推广

本文基于网络问政平台的留言文本分类处理问题，可推广用于其他相似行业的网络反馈信息评估，例如建立一个垃圾短信处理分类系统、电商商品评价体系建立等基于文本的信息分析的案例，调整本模型后可分析解决基于机器学习的大量数据的文本分析类问题。

（2）模型的改进

模型的改进可以通过修正优化算法以提高机器学习的精确度减小误差；在建立模型中考虑更多的相关因素和变量增加模型的通用性和可行性；对代码算法的改进，提高算法计算速度，简化及优化代码使其通用化以便处理多种不同名但同类型的问题。

参考文献

[1] 梁柯,李健,陈颖雪,刘志钢.基于朴素贝叶斯的文本情感分类及实现[J].智能计算机与应用,2019 年 9 月,第 9 卷第 5 期：150-157.

[2]张敏.短文本语义相似度计算研究[J].基金项目,2019 年 4 月，第 35 卷第 10 期：39-43.

[3]秦彩杰,管强.一种基于 F-Score 的特征选择方法[J].宜宾学院学报，2018 年 6 月，第 18 卷第 6 期：4-8.

[4]谢达.基于投票的微博用户影响力量化算法 WeiRank 的设计与实现[D].湖北省:华中科技大学,2013 年:8-12.

[5]胡改丽等.我国网络舆情热度分析文献综述[J].情报科学,2016 年 1 月,第 34 卷第 11 期：160-166.

[6]谢宜瑾.基于层次分析法的网络舆情热点评估方法研究[J].研究与开发,2015 年 4 月,006:27-29.

[7]孙飞显等.政府负面网络舆情热度定量评价方法-以新浪微博为例[J].情报杂志，2015,34(8):137-141.