C 题:“智慧政务”中的文本挖掘应用

摘要

近年来，随着网络平台的发展，政府通过微信、市长信箱等网络问政平台了解民意。因此，运用文本挖掘技术对民意进行分析，提高留言处理效率对政府工作有重大意义。

对于问题 1，通过数据预处理删除文本中的特殊字符后分词并去除停用词，利用 TF-IDF 算法生成每个留言的文本向量，对支持向量机分类模型进行训练，得到一个较好的分类模型，并对模型做出评价。

对于问题 2，观察到留言的主题已经涵盖大部分信息，因此选用留言主题来对特定地点或特定人群的留言进行分类。首先利用 jieba 分词对样本中所有的主题分词，通过 TF-IDF 算法得到每个留言的权重向量，得到关于所有留言主题的特征矩阵。其次利用层次聚类算法，对所有的留言进行聚类，选出包含样本个数大于 15 的 23 大类，再对 23 大类进行筛选，保留筛选后的 13 类。最后定义三个指标，分别为出现次数，时间分散度和互动数，按照出现次数*80%+时间分散度*10%+互动数*10%计算出热度指数，给出排名前五的热点问题以及对应的热地问题明细。

对于问题 3，观察到回复意见中有许多重复留言中的部分，经处理后保留回复中的核心内容。根据核心内容，选用了完整性、相关性和时效性三大一级指标和八个二级指标来对留言答复的质量进行评价。

关键词：中文分词、支持向量机、分层聚类、指标体系

目录

一 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道。各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。本文通过收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，利用自然语言处理和文本挖掘的方法解决下面的问题。

二 目标分析方法与流程

1、群众留言分类

1.1 问题目标

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。因此要建立关于留言内容的一级标签分类模型，完善电子政务系统。

1.2 问题解决

1.2.1 数据预处理

在题目给出的数据中，留言中出现了大量的空格、字母和数字。为了后续的分词处理能达到理想的效果，在 python 中读入数据后，利用正则表达式删除所有的空格、换行符。由于日期和形如“A3 区”的地名与留言的划分归类无关，而且分词后会产生很多噪音，因此删除留言中的日期、包含字母或数字的地名、其他字母以及阿拉伯数字。

1.2.2 中文分词

在对留言进行数据挖掘之前，要先把非结构化的文本信息转化为计算机能够识别的结构化信息，因此首先对留言进行中文分词。本文采用的分词工具为 python 的中文分词包 jieba 进行分词。Jieba 分词是基于 Trie 数结构实现高效词图扫描，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)，同时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，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使其在使用中展现了不错的效果。并在分词后去除无实际含义停用词。

1.2.3 文本向量化

在对留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。本文采用 TF-IDF 算法，把留言信息转化为权重向量。TF-IDF 的具体算法如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

词频（TF）= 某个词在文本中出现的次数

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

词频（TF）= 某个词在文本中出现的次数/文本的总词数

或词频（TF）= 某个词在文本中出现的次数/该文本出现次数最多的词的出现次数

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

逆文档频率（IDF）= log（语料库的文本总数/包含该词的文本数 + 1）

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）× 逆文档频率（IDF）

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本的关键词。

计算出 TF-IDF 值后，用每个词的 TF-IDF 值形成特征矩阵。

1.2.4 建立分类模型

本文使用线性核的支持向量机建立留言的分类模型，共分类七类，分别为: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。并用 0、1、2、3、4、5、6 依次对七个类进行标记。

线性支持向量机的算法原理如下：

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2)...(x_n,y_n)\}$，其中 $x_i \in R^n$，

$y_i \in \{-1,+1\}, i=1,2...n$；

输出：分离超平面和分类决策函数

（1） 选择惩罚参数 C>0，构造并求解凸二次规划问题

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

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

$$0\le\alpha_i\le C,\ i=1,2,\ldots,N$$

得到最优解：$\alpha^*=(\alpha_1^*,\alpha_2^*,...,\alpha_N^*)^T$

（2） 计算：

$$\boldsymbol{w}^*=\sum_{i=1}^{N}\alpha_i^* y_i\boldsymbol{x}_i$$

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

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i\left(\boldsymbol{x_i}\cdot\boldsymbol{x_j}\right)$$

（3）求分离超平面：

$$\boldsymbol{w}^*\cdot\boldsymbol{x}+b^*=0$$

分离决策函数为：

$$f(\boldsymbol{x})=sign\left(\boldsymbol{w}^*\cdot\boldsymbol{x}+b^*\right)$$

在 python 中使用 sklearn 库中的分类器对进行分类，对支持向量机的惩罚系数 C 进行调参，得到最优值为 C=0.48。支持向量机在十折交叉验证下准确率为 0.90685，且具有较好的稳定性。

1.2.5 模型评价

使用 F-Score 对分类模型进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

本文使用的支持向量机模型对每一类的 F-Score 如下：

	precision	recall	f1-score	support
0	0.88	0.90	0.89	620
1	0.95	0.94	0.94	278
2	0.87	0.89	0.88	189
3	0.95	0.95	0.95	467
4	0.94	0.94	0.94	579
5	0.89	0.86	0.87	360
6	0.94	0.91	0.92	270
avg / total	0.92	0.92	0.92	2763

平均 F-Score 为 0.92。

2、热点问题挖掘

2.1 问题目标

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。对某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给

出评价结果。

2.2 问题解决

2.2.1 数据预描述

由题目的要求可知，首先要对相似的留言进行合并，找出留言反映较多的问题。通过观察所给的数据，发现留言详情文本信息存在大量噪声特征，如果把这些数据也引入进行分词、词频统计等，则必然会对聚类结果的质量造成很大的影响。而留言主题精简且已经概括大部分信息，所以在本部分主要使用留言主题，找出留言主题相似的样本。

2.2.2 留言主题分词

由于留言主题属于非结构化的文本信息，要对其进行后续的建模必须转化为结构化数据。中文文本词与词之间没有明显的界限，因此在提取文本的特征时，要对中文进行分词操作。这里采用的分词工具为 python 的中文分词包 jieba 进行分词。部分分词结果如图 2.1 所示

```
0              [A3, 区, 一米阳光, 婚纱, 艺术摄影, 是否, 合法, 纳税, 了, ？]
1     [咨询, A6, 区, 道路, 命名, 规划, 初步, 成果, 公示, 和, 城乡, 门牌,...
2        [反映, A7, 县, 春华, 镇金鼎村, 水泥路, 、, 自来水, 到户, 的, 问题]
3          [A2, 区, 黄兴路, 步行街, 大, 古道, 巷, 住户, 卫生间, 粪便, 外排]
4     [A, 市, A3, 区, 中海, 国际, 社区, 三期, 与, 四期, 中间, 空地, 夜...
5     [A3, 区麓, 泉, 社区, 单方面, 改变, 麓, 谷, 明珠, 小区, 6, 栋, 架...
6                  [A2, 区富, 绿, 新村, 房产, 的, 性质, 是, 什么, ？]
7                      [对, A, 市, 地铁, 违规, 用工, 问题, 的, 质疑]
8                         [A, 市, 6, 路, 公交车, 随意, 变道, 通行]
9     [A3, 区, 保利, 麓, 谷林语, 桐梓, 坡路, 与麓, 松路, 交汇处, 地铁, 凌...
10    [A7, 县, 特立, 路, 与, 东四, 路口, 晚, 高峰, 太堵, ，, 建议, 调整...
11    [A3, 区, 青青, 家园, 小区, 乐果, 果, 零食, 炒货, 公共, 通道, 摆放,...
12    [关于, 拆除, 聚美龙楚, 在, 西地省, 商学院, 宿舍, 旁, 安装, 变压器, 的,...
13                     [A, 市利保, 壹号, 公馆, 项目, 夜间, 噪声, 扰民]
14    [A, 市, 地铁, 3, 号线, 星沙, 大道, 站, 地铁, 出入口, 设置, 极, 不...
Name: 留言主题, dtype: object
```

图 2.1 部分分词结果

由分词结果可以看出，里面存在标点符号，无意义词等，这会对后续分析增加复杂度，但又对结果不提供有效信息，所以需要对这些停用词进行过滤。

2.2.3 停用词过滤

为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words（停用词）。停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。本部分使用网上下载的停用词表，并且在停用词内加入“区”，“县”等无意义的单词，过滤部分结果如图 2.2 所示：

```
0                                [A3，一米阳光，婚纱，艺术摄影，合法，纳税]
1                    [咨询，A6，道路，命名，规划，初步，成果，公示，城乡，门牌]
2                              [A7，县，春华，镇金鼎村，水泥路，自来水，到户]
3                        [A2，黄兴路，步行街，古道，巷，住户，卫生间，粪便，外排]
4                 [A3，中海，国际，社区，三期，四期，空地，夜间，施工，噪音，扰民]
5            [A3，区麓，泉，社区，单方面，改变，麓，谷，明珠，小区，栋，架空层,...
6                                      [A2，区富，绿，新村，房产，性质]
7                                          [地铁，违规，用工，质疑]
8                                      [路，公交车，随意，变道，通行]
9            [A3，保利，麓，谷林语，桐梓，坡路，与麓，松路，交汇处，地铁，凌晨，...
10           [A7，县，特立，路，东四，路口，晚，高峰，太堵，建议，调整，信号灯,...
11           [A3，青青，家园，小区，乐果，果，零食，炒货，公共，通道，摆放，空调...
12                    [拆除，聚美龙楚，西地省，商学院，宿舍，旁，安装，变压器，请求]
13                          [市利保，壹号，公馆，项目，夜间，噪声，扰民]
14                       [地铁，号线，星沙，大道，站，地铁，出入口，设置，不合理]
Name: 留言主题，dtype: object
```

图 2.2 部分停用词过滤结果

2.2.4 生成 TF-IDF 向量

将所有的分词传入到 sklearn 模块中的 TF-IDF 向量生成函数，最终生成一个 4326x7512 的稀疏矩阵。到此，非结构化的文本已经转化为结构化的数据，数据与处理部分完成。

2.2.5 层次聚类算法

层次法（Hierarchicalmethods）先计算样本之间的距离。每次将距离最近的点合并到同一个类。然后，再计算类与类之间的距离，将距离最近的类合并为一个大类。不停的合并，直到合成了一个类。其中类与类的距离的计算方法有：最短距离法，最长距离法，中间距离法，类平均法等。比如最短距离法，将类与类的距离定义为类与类之间样本的最短距离。

层次聚类算法根据层次分解的顺序分为：自下底向上和自上向下，即凝聚的层次聚类算法和分裂的层次聚类算法（agglomerative 和 divisive），也可以理解为自下而上法（bottom-up）和自上而下法（top-down）。自下而上法就是一开始每个个体（object）都是一个类，然后根据 linkage 寻找同类，最后形成一个“类”。自上而下法就是反过来，一开始所有个体都属于一个“类”，然后根据 linkage 排除异己，最后每个个体都成为一个“类”。这两种路方法没有孰优孰劣之分，只是在实际应用的时候要根据数据特点以及你想要的“类”的个数，来考虑是自上而下更快还是自下而上更快。至于根据 Linkage 判断“类”的方法就是最短距离法、最长距离法、中间距离法、类平均法等等（其中类平均法往往被认为是最常用也最好用的方法，一方面因为其良好的单调性，另一方面因为其空间扩张/浓缩的程度适中）。为弥补分解与合并的不足，层次合并经常要与其它聚类方法相结合，如循环定位。

Hierarchical methods 中比较新的算法有 BIRCH（Balanced Iterative Reducingand Clustering Using Hierarchies 利用层次方法的平衡迭代规约和聚类）主要是在数据量很大的时候使用，而且数据类型是 numerical。首先利用树的结构对对象集进行划分，然后再利用其它聚类方法对这些聚类进行优化；ROCK（A

Hierarchical ClusteringAlgorithm for Categorical Attributes）主要用在 categorical 的数据类型上；Chameleon（A Hierarchical Clustering AlgorithmUsing Dynamic Modeling）里用到的 linkage 是 kNN（k-nearest-neighbor）算法，并以此构建一个 graph，Chameleon 的聚类效果被认为非常强大，比 BIRCH 好用，但运算复杂度很高，为 O(n^2)。

凝聚型层次聚类的策略是先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到所有对象都在一个簇中，或者某个终结条件被满足。绝大多数层次聚类属于凝聚型层次聚类，它们只是在簇间相似度的定义上有所不同。这里给出采用最小距离的凝聚层次聚类算法流程：

(1) 将每个对象看作一类，计算两两之间的最小距离；

(2) 将距离最小的两个类合并成一个新类；

(3) 重新计算新类与所有类之间的距离；

(4) 重复(2)、(3)，直到所有类最后合并成一类。

层次聚类的优点有距离和规则的相似度容易定义，限制少，不需要预先制定聚类数，可以发现类的层次关系，可以聚类成其它形状。缺点是计算复杂度太高，奇异值也能产生很大影响，算法很可能聚类成链状。

2.2.6 层次聚结果分析

层次聚类法最终获得 1242 簇。由题目可知，本题首先要将在某段时间特定地点或者特定人群集中反映的热点问题提取出来，而热点问题有一个非常重要的特点就是反映的人数要多。因此，将样本量大于 15 的簇提取出来，也就是相似样本大于 15 的类别提取出来，就可以找到出现数目较多的热点问题候选类。最

终保留下的类别有 23 个，通过筛选最终确定热点候选类别有 13 中，其中出现数量最多的一类总共有 54 条，部分结果如图 2.3 所示：

	ID	反对数	点赞数	留言主题	留言时间	留言用户	留言编号	
0	1	0	0	关于伊景园滨河苑捆绑销售车位的维权投诉	2019/8/23 12:22	A00090519	190337	投诉伊景园.滨河苑开发商捆绑销售车位！A市武广新城片区下的伊景园.滨河苑是广铁集团铁路职工的...
1	1	1	12	A市伊景园滨河苑协商要求购房同时必须购买车位	2019/8/16 9:21	A909171	191001	商品房伊景园滨河苑项目是由A市政府办牵头为广铁集团铁路职工定向销售的楼盘，作为集团的一名退休...
2	1	0	0	请政府救救广铁集团的职工吧	2019/9/1 20:32	A909188	192739	实在搞不懂买个单位福利房-伊景园滨河苑这么麻烦，认购都都认购了好几次，从最开始的五万，到交开...
3	1	0	0	车位捆绑违规销售	2019/8/16 14:20	A909237	195511	对于伊景园滨河苑商品房，A市广铁集团违规捆绑车位销售至今，买房必须买车位我们反映多次一直没有...
4	1	0	0	关于广铁集团铁路职工定向商品房伊景园滨河苑项目的问题	2019/8/10 18:15	A909199	195995	尊敬的市政府领导，您好！我是广铁集团基层职工，我要反应的问题是关于广铁集团铁路职工定向商品房...
5	1	0	0	投诉A市伊景园滨河苑捆绑车位销售	2019/8/7 19:52	A00095080	196264	A市伊景园·滨河苑现强制要求购房者捆绑购买12万车位一个，不买就取消购房资格，国家三令五申禁...
6	1	0	0	关于A市武广新城违法捆绑销售车位的投诉	2019/8/1 22:32	A00095080	199190	武广新城为铁广集团的定向商品房，在未取得预售资格强行逼迫职工缴纳18.5万购房款且不签正式购...
7	1	0	0	家里本来就困难,还要捆绑买卖车位	2019/8/21 18:12	A909192	204960	我是广铁集团铁路职工,因家人身体欠佳,常年就医家里早已经是捉襟见肘,现在跟广铁集团内部购买伊...
8	1	0	1	伊景园滨河苑捆绑车位销售合法吗？！	2019/8/14 9:28	A909234	205277	广铁集团强制要求职工购买伊景园滨河苑楼盘时捆绑购买12万一个的车位，不买车位，不能购买房子！...
9	1	0	2	坚决反对伊景园滨河苑强制捆绑销售车位	2019/8/3 10:03	A909168	205982	我坚决反对伊景园滨河苑捆绑销售车位！原本广铁集团与市政府和开发商协议，可以给铁路职工优惠定向...

图 2.3 样本量最多的一簇聚类部分结果

13 个类别的数量如图 2.4 所示：

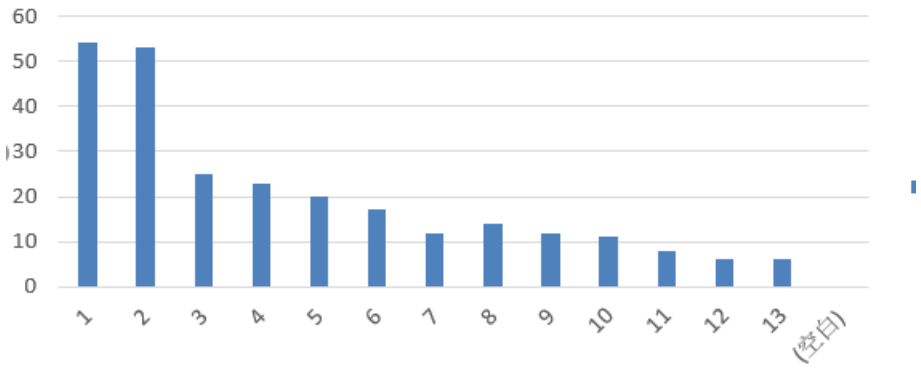

图 2.4 13 个类别的数量结果

2.2.7 热度指标的定义与热度指数的计算

2.2.7.1 热度指标的定义

根据题目的要求，对于计算热度指数首先要定义热度指标，本文定义以下三个热度指标来计算热度指数。

问题出现数：样本中反映同一件事物的数量。反映同一件事情的样本量越多，

说明这件事情关注的人越多，热度越高。反映的人数在热度评价中重要程度较高，因此这一指标的权重为 80%。

时间分散度：时间跨度的长短程度。某一件事情在一个较短的时间内集中反映，说明这件事情的热度较高，这一指标的权重为 10%，计算公式为：

时间分散度=100/同一件事情的时间跨度

互动数：利用反对和点赞数来衡量问题关注度的指标。反对和点赞的人越多，说明问题关注的人越多，这个指标为某一类别所有反对和点赞数之和，权重为 10%。

13 个类别的指标原始数据：

	原始数据表			
类别	问题出现数	时间跨度	时间分散度	互动数
1	54	56	1.79	26
2	53	207	0.48	53
3	25	382	0.26	46
4	23	330	0.30	37
5	20	305	0.33	6
6	17	355	0.28	23
7	14	361	0.28	68
8	12	226	0.44	120
9	12	297	0.34	23
10	11	902	0.11	15
11	8	256	0.39	0
12	6	110	0.91	19
13	6	51	1.96	35

图 2.5 13 个类别的指标原始数据

为避免数据量纲的影响，在计算热度指数时所有数据均进行了归一化处理。归一化之后：

	归一化数据表		
类别	问题出现数	时间分散度	互动数
1	1.00	0.91	0.22
2	0.98	0.20	0.44
3	0.40	0.08	0.38
4	0.35	0.10	0.31
5	0.29	0.12	0.05
6	0.23	0.09	0.19
7	0.17	0.09	0.57
8	0.13	0.18	1.00
9	0.13	0.12	0.19
10	0.10	0.00	0.13
11	0.04	0.15	0.00
12	0.00	0.43	0.16
13	0.00	1.00	0.29

图 2.6 数据归一化处理

2.2.7.1 热度指数的计算

热度指数的计算公式为：

热度指数=问题出现数*80%+时间分散度*10%+互动数*10%

计算结果如图 2.7 所示

热度指数表				
类别	问题出现数	时间分散度	互动数	热度指数
1	1.00	0.91	0.22	0.91
2	0.98	0.20	0.44	0.85
3	0.40	0.08	0.38	0.36
4	0.35	0.10	0.31	0.32
5	0.29	0.12	0.05	0.25
6	0.23	0.09	0.19	0.21
7	0.17	0.09	0.57	0.20
8	0.13	0.18	1.00	0.22
9	0.13	0.12	0.19	0.13
10	0.10	0.00	0.13	0.10
11	0.04	0.15	0.00	0.05
12	0.00	0.43	0.16	0.06
13	0.00	1.00	0.29	0.13

图 2.7

排名前 5 的热点问题为：

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	0.8275	2019/7/7至2019/9/1	伊景园滨河苑	伊景园滨河苑捆绑销售车位
2	2	0.823995	2019/7/3至2020/1/2	万家丽南路丽发新城居民区	A市万家丽南路丽发新城居民区附近搅拌站扰民
3	3	0.374624	2018/11/15至2019/12/2	人才购房补助	A市人才购房补助发放问题
4	4	0.33887	2019/1/8至2019/12/4	A5区魅力之城小区	A5区劳动东路魅力之城小区噪音油烟扰民
5	5	0.266224	2019/2/24至2019/12/26	A市居民	A市大量麻将馆扰民

图 2.8

热点问题详情保存在热点问题详情.xls 文件中。

2.2.8 结果分析

由结果可以看出，对留言主题进行分类并根据热度指标进行排序，可以看出亟需解决的热点问题集中在商贸旅行，劳动和社会保障，城乡建设等分类。在答复意见的回复当中，平均问答时间差长达 20 天，时间跨度从 30 分钟至 3 年，可以看出问政留言平台答复时效性仍待改善。同时，由于答复时间不及时，出现同一用户多次前来平台留言同一问题的情况。

3、答复意见的评价

3.1 问题目标

相关部门对留言给出了答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

3.2 问题解决

3.2.1 数据预处理

在答复意见中，可以发现每条答复都具有先复述一遍留言问题，在具体答复问题。为了使后续分析更准确，不受无用信息的干扰，在做分析前提取了每条答复中具体回复的部分。例如：”现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先…”这类留言，通过正则表达式定位“如下”二字，删除前面的复述内容。而形如“已收悉”的留言，则完整保留。

3.2.2 指标的选取和计算公式

对于数据中相关部门对留言的答复意见，回复内容的详细程度与回复的文本长度有直接关系。简短内容的回复信息量一般不够，较长的文本一般提供较多的有用信息，因此在构建信息完整性的角度上，我们选取了回答长度、回答分词后词语个数，平均句子长度和问答对长度比四个指标来衡量。除了文本内容详实，信息丰富，问题与回复的相似性也是一大衡量角度。一般来说，问题与回复相同或者相近的词语越多，越可能避免答非所问的情况发生。在这一点我们选用词向量相关性、问答重叠词个数来衡量。此外，时效性也是质量回复的一大评判标准，优质的回复应该具有及时的特点。最终，我们选用了完整性、相关性和时效性三个角度和对应的八个二级指标来对留言答复的质量进行评价，具体指标如表 3.1 所示

答复质量评价指标	
一级指标	二级指标
完整性	回答长度
	回答分词后词语个数（去停用词）
	平均句子长度
	问答对长度比
相关性	文本向量相关性
	问答对重叠词个数
	问答对重叠词个数（去停用词）
时效性	问答时效性

指标的描述和计算：

回答长度：即回复的长度，用回复的总字数衡量，无需对回答内容进行深入的分析，而且先前的工作发现它在预测回答质量中表现十分显著。

回答分词后词语个数（去停用词）：这个特征代表了在分词并且移除停用词后

回答中的词语个数。

平均句子长度：这个特征表示回答中每个句子的平均长度，我们期望一个好的回答的句子具有合理的长度，既包含足够的信息，又不冗余。在这个指标中，我们选用句子的长度除以句号的个数，计算公式为：

平均句子长度=回复句子长度/回复句号个数

问答对长度之比：这个特征给出了一个问题和它的回答的原始长度之比，我们期望一个好的回答有一个足够的长度，这样可以包含一些详细的解释来支持它的可信度。计算公式为：

问答对长度之比=回复句子长度/留言详情长度

文本向量相关性：回复和留言生成的向量之间的相关性。通过把回复和留言的文本转化成向量，利用余弦相似性计算出两个向量的相似度，衡量文本间的近似程度。

问答对重叠词个数：回复与留言之间的重叠词个数。该特征反映了回复和留言的相关性，我们期望一个好的回答和问题具有很大的相关性。

问答对重叠词个数(去停用词)：回复与留言之间去除停用词后的重叠词个数。

问答时效性：答复时间与留言时间之差，衡量回复的速度和及时性。由于留言时间差相差巨大，且时间差与效率成反比，所以在计算时使用公式：

问答时效性=1/log（时间差*100）

本部分的数据内容为市民留言及回复，总共有 2816 条数据。计算之后的指标原始部分数据为：

回答长度	词语个数	平均句长度	答对长度	向量相关	问答对重叠词个数	问答对重叠词个数（去停用词）	问答时效性
319	88	59	0.932749	0.240791	75	36	0.31421129
232	74	35.33333	1.487179	0.027872	25	1	0.315607208
315	96	73.5	1.852941	0.278651	33	13	0.315559008
235	70	34.5	4.051724	0.22811	19	9	0.315491846
126	34	27.5	3.230769	0.480743	23	12	0.312900594
171	38	52.66667	1.195804	0.03921	17	2	0.286353332
172	55	32.2	0.519637	0.176171	67	14	0.276848123
591	188	45.91667	2.281853	0.174483	59	20	0.289421112
471	145	55	2.021459	0.114185	47	20	0.311489231
190	61	58.33333	0.497382	0.239341	86	28	0.311470518
454	139	52.5	2.316327	0.070073	33	15	0.259765606
393	124	69.6	7.415094	0.278743	14	10	0.287030908
106	32	48	0.302857	0.100685	57	16	0.311806789
66	16	30.5	0.111486	0.031775	82	5	0.361244976

为了避免量纲对最后结果的影响，考虑到部分指标值存在异常值，所有将数据标准化并求和计算质量评价指数。作为一个探索性研究，暂不考虑各个指标的权重参数影响，质量指数的结果直接由各个指标之和计算而来。部分结果如下：

	回答长度	回答分词	平均句子	问答对长	词向量相	问答对重	问答对重	问答时效	质量指数
0	0.04615	-0.06066	0.670559	-0.16378	0.747075	0.133012	0.389197	-0.32558	1.435972
1	-0.18613	-0.17349	-0.3719	-0.08224	-1.00615	-0.4271	-0.58372	-0.31349	-3.14422
2	0.035471	0.003809	1.309251	-0.02845	1.058823	-0.33749	-0.25015	-0.3139	1.477364
3	-0.17812	-0.20573	-0.40861	0.294909	0.642655	-0.49432	-0.36134	-0.31449	-1.02504
4	-0.46914	-0.49585	-0.71694	0.174177	2.722895	-0.44951	-0.27795	-0.33694	0.150743
5	-0.34899	-0.46362	0.39159	-0.12509	-0.91278	-0.51672	-0.55592	-0.56696	-3.0985
6	-0.34632	-0.32661	-0.50992	-0.22453	0.214982	0.043393	-0.22235	-0.64931	-2.02068
7	0.772358	0.745243	0.094268	0.034627	0.201081	-0.04623	-0.05557	-0.54038	1.20541
8	0.451972	0.398703	0.494368	-0.00367	-0.29542	-0.18065	-0.05557	-0.34917	0.460566
9	-0.29826	-0.27826	0.641194	-0.2278	0.735135	0.256237	0.166816	-0.34933	0.645727
10	0.406584	0.350349	0.384248	0.039697	-0.65865	-0.33749	-0.19455	-0.79733	-0.80714
11	0.243722	0.229463	1.137465	0.789535	1.059578	-0.55033	-0.33354	-0.56109	2.014805
12	-0.52253	-0.51197	0.186034	-0.25641	-0.40659	-0.06863	-0.16676	-0.34642	-2.09327
13	-0.62933	-0.64092	-0.5848	-0.28455	-0.974	0.211428	-0.47253	0.08194	-3.29277
14	-0.33831	-0.35079	0.39159	-0.08098	-0.95976	-0.49432	-0.41694	-0.37176	-2.62127
15	0.841775	0.833893	0.743972	-0.09283	0.94477	0.782746	0.972947	-0.79026	4.237014

3.2.3 结果分析

将计算得到的质量指数放回附件 4 中，得到回复质量评价结果表，再对质量评价指数从高到低排序，评分较高的部分回复：

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	质量评价指数	排名
88359	UU0082390	海J9县的油茶种植政	2019/1/5 15:52:1	产品品牌建设的决	以产业发展为目标，以品牌建设为动力，以“原生态、纯天然、	2019/1/7 17:00:37	71.6282557	1
96757	UU0082390	海J9县的油茶种植政	2019/1/5 15:56:2	产品品牌建设的决	求为导向，以新型组织为主体，以产业发展为目标，以品牌建设	2019/1/16 8:38:52	70.57095094	2
96762	UU0082390	产品品牌建设扶持政	18/12/15 23:28:	有机农产品品牌建	设，以新型组织为主体，以产业发展为目标，以品牌建设为动力，	2019/1/3 10:53:46	66.91819818	3
9744	UU0081323	中医医院在北京园御	16/10/25 12:47:	下的原售楼部（北	卫委《关于医疗机构设置审批前公示》模版，公示的内容为：医疗	2016/11/24 9:34:40	34.04248419	4
138410	UU008758	交易分局局长邓斌文	016/9/29 18:39:4	的领导是否能够反	代表们你一言我一语一直哭诉到中午12点半，廖组长最后提出四点	2017/1/9 9:25:47	31.21321433	5
4331	UU0082338	装修价格以及[政府	18/10/24 10:55:4	文件的计算表在[	费的3.0%）。你反映的该单位工程费用计算表列有1.4其他费用00	2018/11/6 10:21:48	30.88339581	6
18413	UU008835	7县江背镇乌川湖村发	015/12/30 2:26:4	给黄刚明一家，	洞设计较低导致出行不便的一次性补偿协议，协议明确要求黄刚明	2016/1/7 10:59:47	29.75360034	7
11819	UU008167	号线井标、评标过程	14/11/18 14:07:	扩主导地位，商务	进行了澄清。 二、“评标办法中的商务加分太离谱”问题。	2014/12/23 16:44:49	26.9189088	8
9225	UU0081831	“恒大城”小区配套	017/3/12 11:16:2	的人生发展。	月26日，区教育局参加了由洞井街道组织召开的恒大城幼儿园协调	2017/3/31 16:33:12	25.73127173	9
5607	UU008777	造文明畅通新A市的	018/6/21 15:23:4	享的香港，A市要	整治工作为工作重心，坚持交通违法整治的强日常、优结构、勤	2018/7/6 14:57:36	25.02344657	10
9199	UU008144	育局拖延恒大城公立	017/3/14 14:37:5	到，让洞井街道的	区教育局接待了恒大城业主的多轮次来电来访，向广大业主宣传相	2017/3/24 13:46:11	22.4728884	11
9128	UU008424	和馨园租户相关问题	017/3/22 12:55:1	是委会与业主代表	人员，仅缴纳了房屋租金，未额外缴纳物业管理费用，而业主物业管	2017/4/14 15:32:24	22.03354943	12
8764	UU008593	圣国际小区的几点问	017/5/11 21:47:2	根本看不见），西	按照有关规定给予优惠。国家鼓励、支持企业事业组织、社会团	2017/6/12 16:23:36	21.69912382	13
25582	UU008353	道如八字槽门一带增	016/2/19 8:43:04	就如飞地一块	辖围内，无法通过本栏目处理程序进行处理的信件，将被视为无效	2016/2/19 9:13:53	20.42365658	14

图　得分较高的一部分数据

评分较低的部分回复：

12152	UU0081903	运输驾驶员接受远程	014/7/11 19:15:2	老乡都可以通过网	网友：您好！留言已收悉	2014/8/13 17:50:52	-6.559666735	2806
11974	UU008334	汽车西站至金洲大道	2014/9/16 8:41:4	下班都搞得心惊胆	网友：您好！留言已收悉	2014/10/20 9:45:37	-6.568543226	2807
12165	UU008271	造成为烂尾楼，请求	014/7/9 10:02:2	无法收房。	网友：您好！留言已收悉	2014/8/13 17:39:32	-6.570831252	2808
12163	UU0082207	或915路公交调整经过	014/7/9 13:43:4	市大学一来可以	网友：您好！留言已收悉	2014/8/13 17:43:11	-6.574914138	2809
12387	UU0081236	金洲新区商新安置房	014/4/10 13:38:4	经有三年多了，五	网友：您好！留言已收悉	2014/5/16 15:55:37	-6.584033423	2810
12458	UU00854	镇头镇连山村交通不	014/3/7 22:39:3	不断的扩宽提质，	网友：您好！留言已收悉	2014/4/14 12:13:50	-6.600243749	2811
30019	UU008151	备全款购买二手房事	016/11/3 10:00:1	房，房产局资金监	已收悉	2016/11/22 12:25:56	-6.606719186	2812
12415	UU0081509	含浦镇白鹤社区的加	014/3/27 17:16:3	导这些难道都不是	网友：您好！留言已收悉	2014/5/9 17:28:09	-6.634546204	2813
12142	UU0082324	古曲南路交叉路口建	014/7/14 10:09:4	段是一个下坡路	网友：您好！留言已收悉	2014/8/27 16:11:28	-6.644756558	2814
114346	UU0081119	城锦豪庭景园小区商	013/6/3 13:05:4	但购房补充合同	已收悉	2013/7/5 16:47:46	-6.787753208	2815
159285	UU0081173	保、残疾人补助的相	014/6/14 22:59:3	为，致使享有低保	你好，请向当地民政部门询问。2017年8月18日	2017/8/18 9:27:48	-7.098088971	2816

图　得分较低的一部分数据

由结果可以看出，质量较高的回复内容比较详实，与留言的相关性较高，可解释性也较强，较多有用信息，回复有理有据，较好地解决了市民所反映的问题。而质量较低的留言回复简短，提供很少有用信息，对网友只有收到回复或是转交给其他部门，并没有很好地解决相关问题。从结果来看，本题定义的指标体系取得了较好的评价结果。

从回复结果评价中可以大概看出，质量不高的回复大概 270 条左右，所占比例大概 9%。这说明，在回复中还是存在一小部分劣质回复。在实际中，应该对这些得分不高的评价进行再一次复核，厘清回复质量不高的原因，更进一步提高网络服务的质量。

三 结论

对网络问政平台上的民众留言及政府答复进行文本挖掘，能帮助政府能及时、准确了解民意，缓解以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。

从以上结果可以看出，利用支持向量机分类器能较准确的把民众留言划分到城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生七个一级标签中，便于将各类别的留言分放到对应的政府部分进行处理，提高了政府运行的效率。通过分层聚类算法提取热点问题，能方便政府部门优先处理较急较重的问题，更好的为民服务。同时可以看出亟需解决的热点问题集中在商贸旅行，劳动和社会保障，城乡建设等分类，建议政府着重完善这几个方面的建设。而且政府对一些问题长达三年后才回复，建议缩短回复时间。最后为政府的回复建议建立评价指标体系，为政府在回复留言时给出参考。

四 参考文献

[1] 杨开平,李明奇,覃思义.基于网络回复的律师评价方法[J].计算机科学.2018
[2] 谢青.苍南县国税局纳税咨询服务质量评价研究[D].福建农林大学.2016
[3] 王宝勋.面向网络社区问答对的语义挖掘研究[D].哈尔滨工业大学.2013
[4] 胡泽.在线问诊服务回答质量评价方法研究[D].哈尔滨工业大学.2019
[5] 于书鳗.网络问政平台的“回应性陷阱”—基于 C 省“地方领导留言板”的实证研究[D].吉林大学.2019
[6] 刘高军,马砚忠,段建勇.社区问答系统中“问答对”的质量评价[J]. 北方工业大学学报.2012
[7] 贾佳,宋恩梅,苏环.社会化问答平台的答案质量评估——以“知乎”、“百度知道”为例[J].信息资源管理学报.2012
[8] 邹沁含,庞晓阳,黄嘉靖,刘司卓.交互文本质量评价模型的构建与实践[J].开放学习研究.2020
[9] 来社安,蔡中民.基于相似度的问答社区问答质量评价方法[J].计算机应用与软件.2013
[10] 李少温.基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[D].华中科技大学.2019