“智慧政务”中的文本挖掘应用

摘要

随着互联网的进一步发展，人们都能通过网络渠道，用手机电脑等在论坛、贴吧、微博等形式自由发表言论，来表达自己的意愿和对政府的建议等，对于政府来说，处理政务的方式也逐渐从单一的线下传递变成了网络问政，如微信、微博、市长信箱等一系列的网络问政平台成为了反映民意、了解民意的重要途径。大数据和网络问政推动了网络社会的崛起，网络虚拟空间已经退隐为一种现代生活无处不在的背景。但是随着这类方式的普及和越来越受欢迎，各类民意相关的数据也日趋增加，对于这样大量文本的分类和处理，仅仅依靠人工划分是要消耗掉大量的人力物力。因此，在大数据、人工智能发展的时代，采用智慧政务是必不可少的，就此问题，我们将对文本进行分类，划分提取等。在大量的文本信息中心，可能很难从某段句子中找到我们想要的结果，所以在本文采用了设置特征词（以这些词的意思为核心，再去查询相关的信息），利用 TF-IDF 算法提取各词并计算特征词权重，再以这些特征词词为抽取的研究对象，然后在大量的留言中用 LVM 算法筛选出使用敏感词的相关留言，并根据不同特征词词对其进行选取和分类，通过对信息的筛选就相对比较容易得到民众的意愿，对于政府来说也能提升管理水平和效率，在此过程中采用了 DFA 敏感词提取法。

关键词： 设置特征词　特征词权重　LVM 算法文本分类

目录

一、问题的背景与挖掘目标

1.1 问题的背景

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。近年来，虽然网络问政风起云涌、不断发展，大数据给学术界也同样带来了巨大的挑战和机遇，但国内学术界对其深入研究并不多，尤其是大数据和网络问政相结合的文字是少之又少。因此“智慧政务”的关键还是在于对大数据的应用，如何才能有效率的提取网络中大量的民众的有效意愿。

1.2 挖掘目标和意义

随着媒体融合和大数据时代的发展，传统问政与网络问政方式的结合，实现了内容与技术形式的顺畅对接，也使大数据的应用和网络问政发展之间的关系更加紧密。从云计算技术延伸到大数据技术、云搜索;从网络问政建设发展到微博问政、电视问政的兴起，以及相关的网络发言人平台的形成，都体现出网络问政在网上办公的新形式和新需求。

目前，还需进一步完善网络问政的各种条件，其中包括网络问政内容的搜集、抓取、分析、研判等工作，并将网上的社情民意的信息转化为有意义、有价值的政治决策参考。

网络问政拓宽了公民参与民主决策利民土监督的渠道，有利于维护公民的合法权益，提高了公民的政治参与意识、社会责任感和主人翁意识。保障了人民当家作主的权利的落实，推动了社会主义政治文明。政府通过网络虚心听取群众意见，自觉接受群众监督，有利于决策的科学化、民主化，有利于提高政府依法行政水平，树立政府权威，真正做到权为民所用，利为民所谋。

二、问题探索分析

2.1 问题 1 分析方法与过程

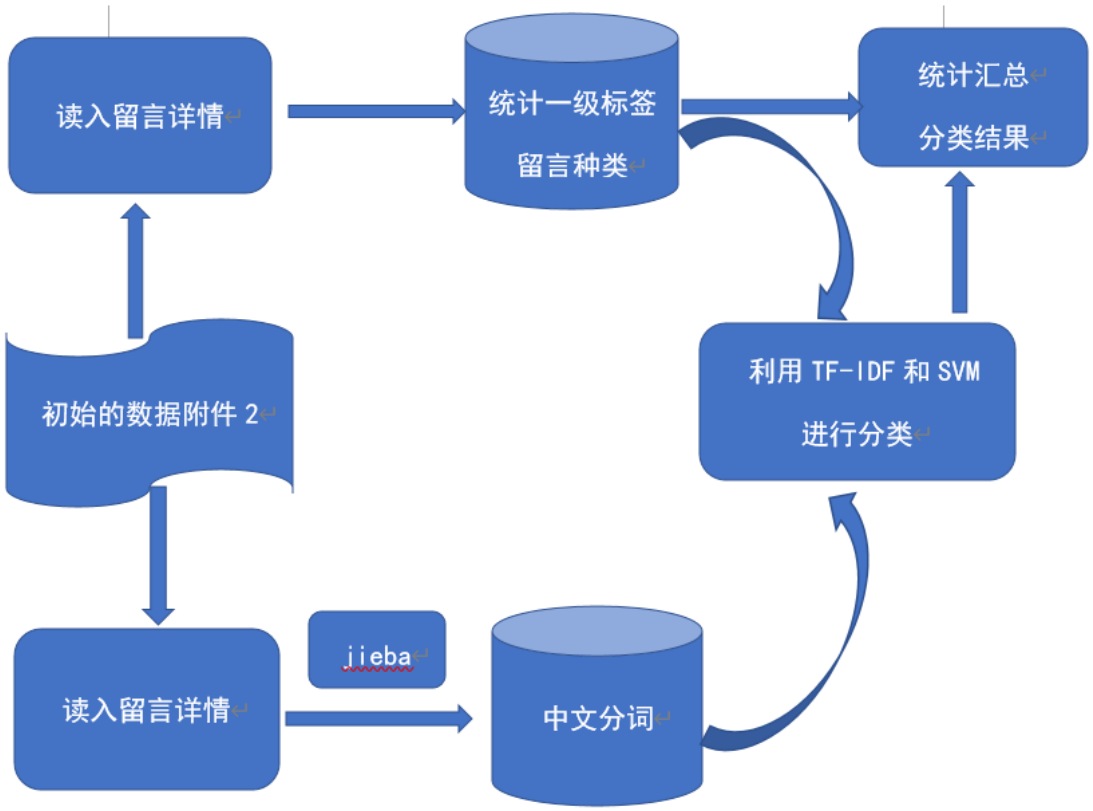
2.1.1 问题解析

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，为进一步提高效率和减少不必要的工作量，可以运用大数据人工智能等先对群众的留言进行分类处理，根据留言相关的内容，建立关于留言内容的一级标签分类模型，主要运用利用 TF-IDF 提取关键词

2.1.2 流程图

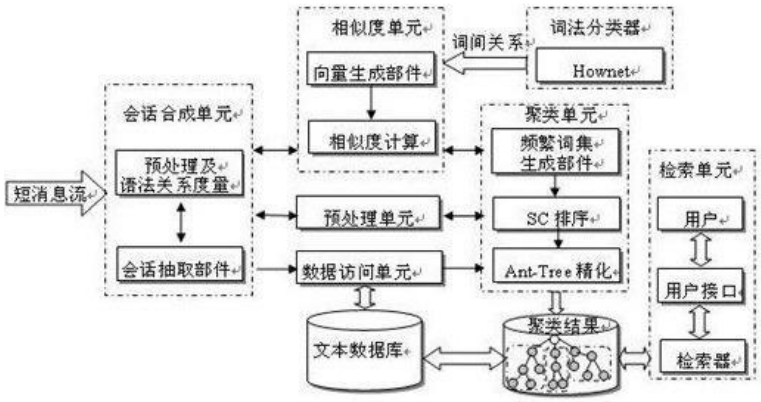
图 1：问题 1 流程图



2.1.3 TF-IDF 简介：

TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。

图 2：tf-idf 原理图



我们使用 tf-idf 算法抽取不同不同标签的关键词。统计每个词汇在文章中出现

的频率 TF（term frequency），频率高的就是具有代表性的词汇。

$$词频(TF) = \frac{某个词在文中出现的次数}{文章的总词数}$$

但是类似于“的”，“是”，“你我他”这样的停用词需要排除，这些词排除过后，还会发现有些词出现的频率是一样的，但这并不意味着它们的重要性是相同的。如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，正是我们所需要的关键词。即在词频的基础上，我们需要对该词分配一个权重也就是逆文档频率。所以除了考虑词汇的频率之外，还需考虑词汇在其他文档当中出现的概率，词汇的重要性应该和该概率是反相关的，我们用公式来衡量：

$$逆文档词频(IDF) = \log\left(\frac{语料库的文档总数}{包含该词的文档数 + 1}\right)$$

故为衡量词汇的重要性，利用 TF-IDF 来计算

$$TF - IDF = 词频(TF) \times 逆文档频率（IDF）$$

2.1.4 TF-IDF 模型实现和结果分析

2.1.4.1 文档处理

读取文档后需对文档进行预处理，处理步骤如下

读取文档 → 删除不需要的字符（如回车符\n、制表符\t、空格等）→ 转换成 unicode 格式 → 对文档分词 →转换成 utf-8 格式写入 txt 文档

2.1.4.2 tf-idf 算法

①对文档进行分词：由于中文需要分词，jieba 分词是 python 里面比较好用的分词工具，所以选用 jieba 分词。

②词频统计，计算 tf 值

③计算 idf 值

④tf×idf，即 tf-idf 的实现

2.1.4.3 参数设置

表 1：参数设置表

参数	含义
copus	需处理的字符串
tf	分词后的词频统计
corpus	读入处理的文档

2.1.4.4 结果

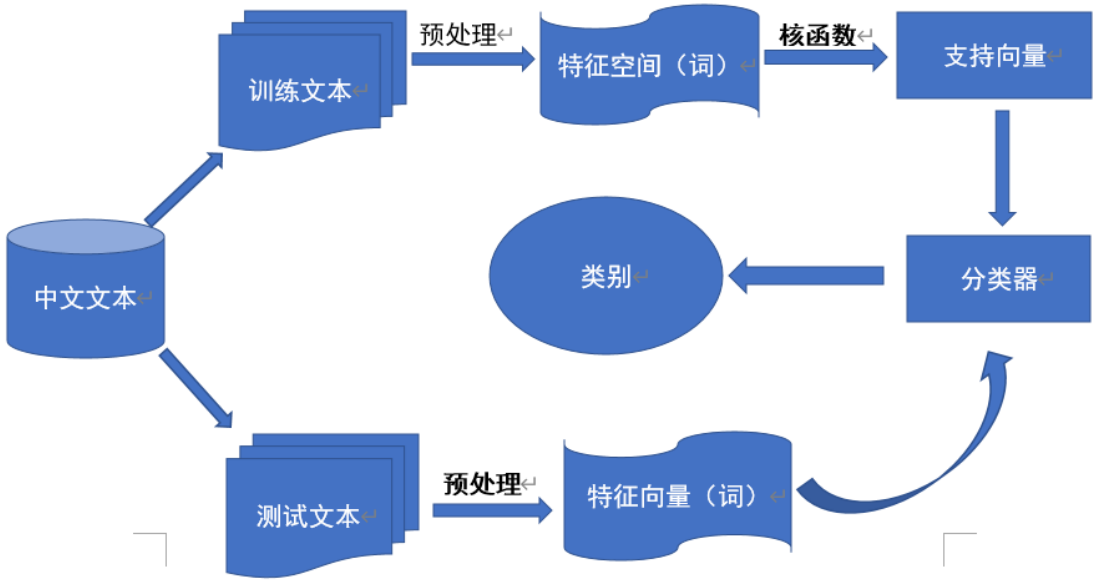
表 2：城市建设类关键词的特征选择结果

特征词	城市道路	维修	建设	安全隐患	非常	整改
权重	3.11	3.11	3.11	3.11	3.11	3.11
特征词	摆摊	民众	干净	卫生	公厕	市容市貌
权重	3.11	3.11	3.11	3.11	3.11	3.11
特征词	路段	人行道	危机	过往行人	车辆	安全
权重	3.11	3.11	3.12	3.11	3.11	3.11
特征词	城管	绿化	公共资源	搬迁	重建	强揽
权重	3.11	3.11	3.11	3.11	3.11	3.11
特征词	商户	乱	招标	秩序	营商	占用
权重	3.11	3.12	3.11	3.11	3.11	2.7
特征词	工程	严重	正常	干扰	环境	路段
权重	3.11	3.11	3.11	3.11	3.11	3.11
注：因为运行结果保留16位数，而次表保留2位小数所以结果看起来一						

2.1.5 群众留言的分类

生成留言的 TF-IDF 权重向量后，根据每类留言的 TF-IDF 权重向量对留言就行分类。采用 LVM 算法将留言分为 7 类

图 3：SVM 中文文本分类模型



2.1.5.1 LVM 的原理如下：

SVM 理论的主要思想是在样本空间中建立一个最优超平面，将两类样本的隔离边缘最大化，是结构风险最小化（ SRM ）的近似实现。 虽然 SVM 不利用问题所属的领域知识，但仍能够提供良好的泛化性能。

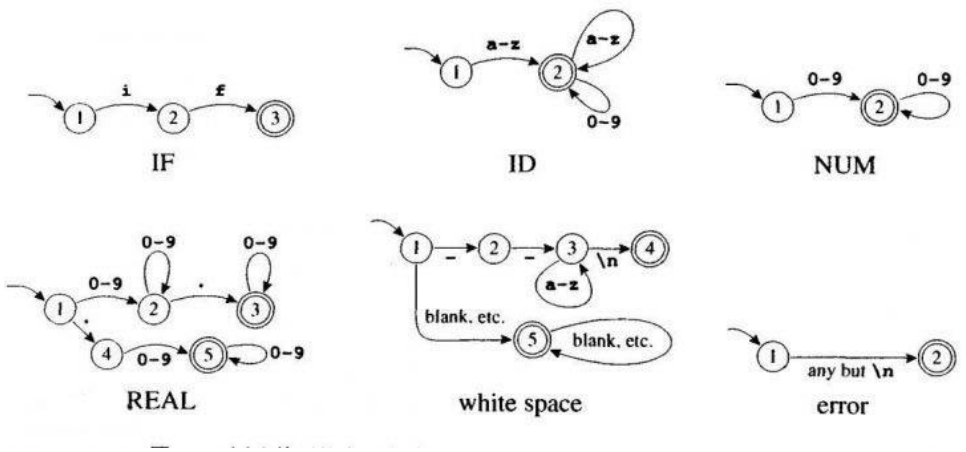
SVM 的目标是在样本点所在的向量空间中， 找到一个满足分类要求的最优分类超平面，这个超平面能把不同类的样本分开， 使其在满足分类精度的同时，两侧的空白区域（分类间隔） 最大化。对于分类间隔最大化的原因，可以简单理解为最大化的分类间隔下，对于未知点的分类更准确，也即泛化能力更强。获得了这个最优超平面后，就可以使用它对于训练样本所属类别进行决策。

在分类问题中， 存在着线性不可分的样本， 是不能够直接求得决策曲面的。SVM 对低维空间中线性不可分的样本集进行映射，让样本获得在高维特征空间中线性可分的特性， 并在这个高维特征空间中求得最优超平面， 最终实现样本的分类。

2.1.5.2 LVM 的算法步骤如下：

文本预处理模块

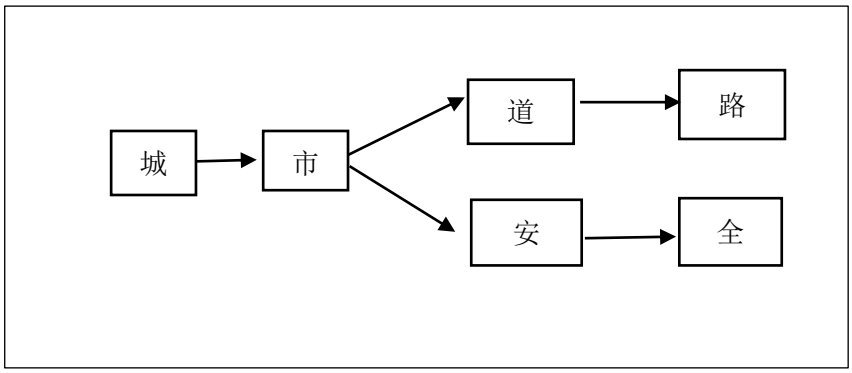
将中文文本转换为特征向量的形式，以便分类函数训练模块和文本分类模块使用。
分类函数训练模块
构造分类函数，根据附件 2 的留言详情，求解二次规划，确定分类函数。将文本预处理得到的特征向量带入求解参数，明确分类函数的参数 ω 和 b 值，确定分类函数。
文本分类模块
将预处理模块的测试文本特征向量，带入分类函数计算结果，当结果大于 0 为正文本，反之为负文本。
2.2 问题 2 分析方法与过程
2.2.1 问题解析
由于民众反映的问题很散很杂，这会导致在接收和处理问题时没有针对性，不能高效率的解决和改善相关问题，所以为了精准找到这些留言背后的关键性问题所在，并及时做出相关的回应，需要进行热点问题挖掘，在此过程中更，可以采取文本聚类的方法将相关的信息在第一时间整合在一起，然后再反馈给有关部门、热点问题挖掘中，热点关键词[1]的提取是尤为重要的，直接决定了挖掘热点问题的准确性，热点关键词具有以下特点，有代表性、简洁性、时效性、集中性等，能够在最短的时间内反应最大的信息量。热点关键词的提取可以采用自动提取和手动设置，但在本文中，我们采用自动提取的方法，在留言中筛选出文本出现频率高的词作为热点关键词，找出热点关键词后，再将相关的留言整合，反馈相关的问题。
2.2.2 热点问题挖掘步骤
将文档向量化，构建初始的相似度矩阵，利用 TF-IDF 对留言进行分词、特征项权值计算、留言详情向量表示以及构建相似度矩阵。
按某一时间段内特征词出现的权重对留言详情进行筛选，权重越高的特征词的留言进行整合。用相似度最接近的留言的 ID 来表示事件，作为最终列表
2.2.3 使用 DFA 算法对文档中特征词进行标记
2.2.3.1.DFA 概念
DFA 全称为：Deterministic Finite Automaton,即确定有穷自动机。其特征为：有一个有限状态集合和一些从一个状态通向另一个状态的边，每条边上标记有一个符号，其中一个状态是初态，某些状态是终态。但不同于不确定的有限自动机，DFA 中不会有从同一状态出发的两条边标志有相同的符号。
图 4：DFA 原理图



2.2.3.2.关键词词构造

以“城市道路“，“城市安全”为例

图 5：关键词构造示例图

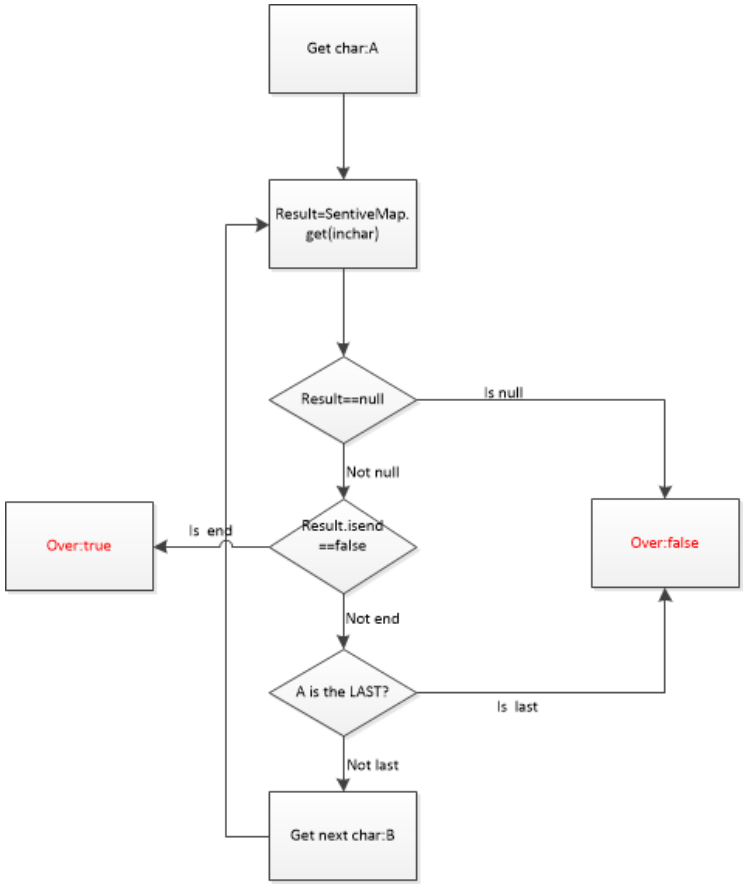


2.2.3.3.基于关键词库搜索

将含有标记的元素筛选出来即为一种类别

对分类方法进行评价

图 6：DFA 算法图



2.3 问题 3 分析方法与过程

2.3.1 问题解析

相关部门对留言的答复意见，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，对意见答复采用聚类的方法进行分类，然后根据不同分类的点赞数归结不同的答复的质量。

三、结果分析

3.1 问题 1 结果分析

对附件 2 给出的数据，建立关于留言内容的一级标签分类模型，将留言分为了以下 7 类分别为：劳动和社会保障、教育文体、卫生计生、商贸旅游、交通运输、环境保护和城乡建设等；通过描述统计表可以看出，劳动和社会保障、城乡建设和教育文体等方面的留言比较多，分别占比为 21%、20%和 19%，说明市民对这三方面的问题比较关注，但环境保护方面的留言相对较少，说明市民对环境方面的意见相对较少。

图 7：一级分类各类标签所占比例

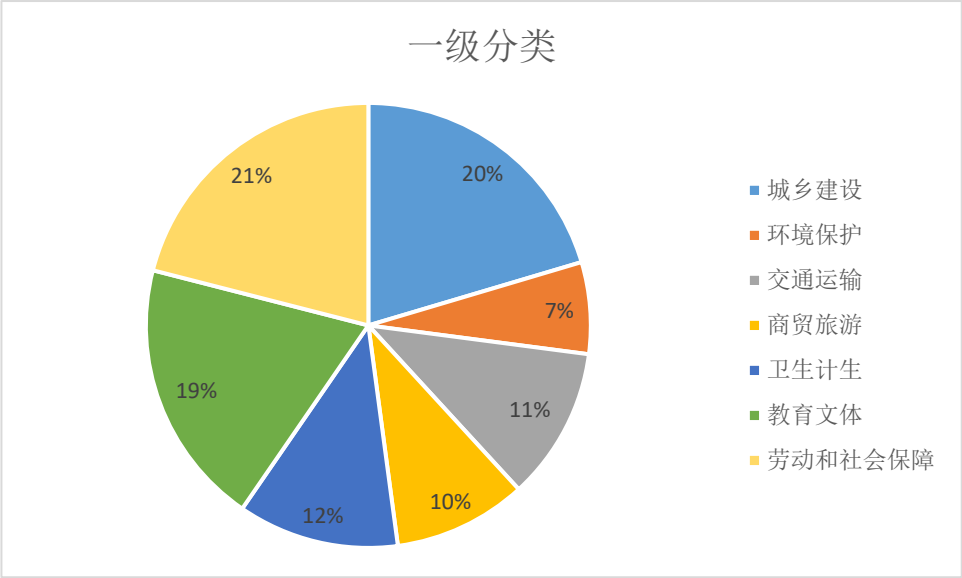
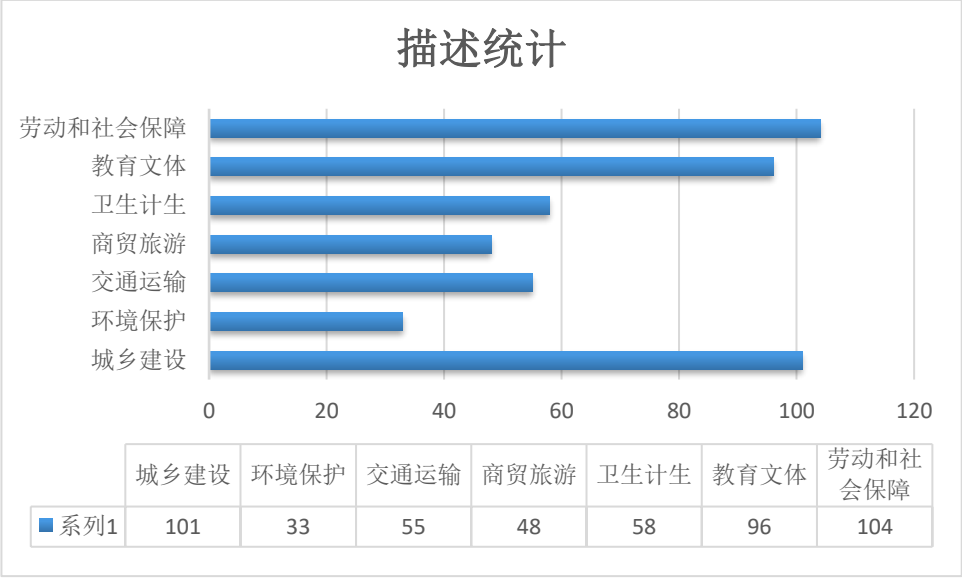


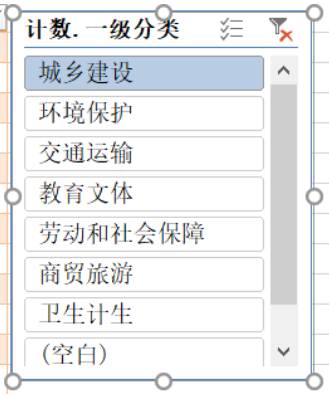
图 8：各类表标签描述统计



点击对应的切片选项就会出现对应的留言分类。

表 3 结果展示表

计数. 留言用户	计数. 留言主题	计数. 留言详情	计数. 一级分类
A00074011	A市西湖建筑集团占	A3区大道西行便道	城乡建设
A000107866	A市在水一方大厦人	位于书院路主干道	城乡建设
A00013884	A3区杜鹃文苑小区/	A市政府、市交警引	城乡建设
A0009647	民工在A6区明发国	胡书记，您好，感	城乡建设
A00047862	K8县丁字街的商户	K8县丁字街的	城乡建设
A00093415	K8县南门街干净整	南门街前段时	城乡建设
A00096769	K8县冷江东路蓝波	现K8县冷江东	城乡建设
A00055421	K8县九亿广场的公	九亿广场是城	城乡建设
A00091339	K4县石期市镇老农	石期市镇老农贸市	城乡建设
A0006699	K市域轨道交通规划	李书记您好，	城乡建设
A0006699	关于K市域轨道交通	易市长您好，	城乡建设
A00072155	请问A市乘坐地跌是	看某媒体报道	城乡建设



3.2 问题 2 结果分析排名前五的热点问题

表 4 前 5 热点问题表

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
360114	A0182491	A市经济学院(	42894.73009	书记您好	9	0
360108	A0283523	A5区劳动东路	43678.68058	局长： 你	6	0
360101	A324156	A5区劳动东路	43674.53424	尊敬的政府：	4	0
360100	A324156	魅力之城小区	43713.52015	魅力之城小区	3	0
360113	A3352352	A市经济学院引	43237.3556	A市经济学院引	3	0

3.3 问题 3 结果展示

留言答复总体可以分为三类：

1. 留言所描述的问题得到解决，予以回复
2. 留言所描述问题未解决单交由相关部门及解决
3. 留言表述不清楚或该问题早已解决重复留言

四、总结与展望

通过对此次问题的挖掘，DFA 模型通过设置关键词，来对网络留言进行分类，极大地减少了人工劳动力，节约了时间成本，同时对于相关的热点话题的讨论也能及时作出相应的回复，极大地提高了政府的办事效率，对人们生活水平的进一步提升有很大的帮助，可以让用户提交诉求的时间缩短，并在用户提交内容的第一时间直接专递给领导，让领导能更快的处理解决问题。诉求马上办就是一个智慧政务的工具，可以缩短用户和领导直接的沟通时间，让领导能第 一时间就能知晓并着手帮助用户解决问题。线上收集解决问题，线上无法处理版的线下沟通解决，有效的提高了事情的处理进度及人员的利用率，并且数据清晰明了易整理，会是未来智慧政务发展的趋势。在此过程中也有一些问题尚待解决，如数据的不完整性，还有对有些小众的问题的还没有好的解决办法，但相对来说，已经对智慧政务的发展有推动的作用。

五、参考文献

[1] 张寿华、刘振鹏.网络舆情热点话题聚类方法研究[D].河北大学.数学与计算机学院.2013-03
[2] 阮一峰. TF-IDF 与余弦相似性的应用（一）：自动提取关键词[EB/OL].http://www.ruanyifeng.com/blog/2013/03/tf-idf.html,2013-3-15
[3]休耕.基于 DFA 算法、RegExp 对象和 vee-validate 实现前端敏感词过滤[EB/OL]. https://www.cnblogs.com/xiugeng/p/11169244.html,2019-7-12
[4] 李博.网络热点事件挖掘及特征描述研究[D].国防科学技术大学.2010-11
[5] 向亮.基于的中文文本算法[EB/OL].https://www.happyxiangyang.com/p/ ，2017-6-15