“智慧政务”中文本数据挖掘与分析

摘要

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。而自然语言处理（NLP）作为人工智能的一个重要领域得到了飞速发展，因此本文通过数据分析，构建基于自然语言处理技术的分类模型，以解决此类问题，最后对答复意见的质量给出一套评价方案并加以实现。

针对第一题，主要对留言主题和留言详情进行去重、中文分词和去停用词等的数据预处理，以 7:3 的比例把处理后数据划分为训练集(data_text)和训练集(data_train)，然后通过 Word2vec 模型进行合成词向量，最后构建基于自然语言处理技术的支持向量机(Support Vector Machine,SVM)分类模型，从而使得留言内容达到自动分类的效果。

针对第二题，基于 word2vec 模型，对每一个时间间隔的留言内容计算句子语义相似度，再筛选其中区间相似度最高的留言文本，进行量化处理，并对每个时间间隔留言内容获取长尾关键词，最终将两种处理结果综合得出热度问题的评价指标计量。

针对第三题，我们以相关性、完整性、可解释性、及时性为主要指标构建指标体系，并对其进行相关性分析，采用层次分析法对主要指标进行计算权重，基于权重构建评价的得分标准，最后基于评价方案对附件 4 的留言内容进行评价并得出评价结果。

在研究方面，我们首先对数据进行去重，利用 python 的 jieba 工具进行中文分词，然后利用 Word2vec 的模型进行特征处理，最后通过支持向量机（Support Vector Machine SVM）建立模型，通过 F1 值调整模型，最后使得留言内容通过 SVM 模型使得留言内容进行自动分类。通过文本得相似度的量化等方法定义热度指标，从而得出热度问题表和热度问题留言明细表，最后建立评价方案对附件 4 的答复意见的质量进行评价。

关键词：文本挖掘　Word2vec 模型　SVM 模型　评价方案

目录

一、引言

1.1 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，政务热线话务量每年上升，各类社情民意相关的文本数据量也不断攀升，期间产生的数据量、信息量非常巨大。以前政府处理这些文本数据需要依靠人工进行留言划分、热点整理，他们的工作是有极大挑战力的。如话务员记录的相关内容，可能会不准确不精准，这需要耗费大量的人力、时间去重新听取话务员录音，记录评价话务员说话语气和工作态度，再进一步人工提取信息、分类整合。以往政府的操作流程，会耗费大量的人力、物力，花费大量时间去整理。

随着大数据、 云计算、人工智能等技术的发展，国家陆续发布的关于云计算、大数据、物联网、互联网+和信息惠民等一系列推进信息化的文件，制定了“互联网+”行动计划，把互联网的创新成果与经济社会各领域深度融合，推动技术进步、效率提升和组织变革，并确定了提升公共服务水平“互联网＋”益民服务的具体任务。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地推动作用。以最便利的公共服务推进政府治理的科学化和现代化。如何更好地利用这些文本数据，让政府更好的结合经济、社会发展的热点问题，有效推动市场发展、为政府决策提供有力支撑，从文本数据中获取有价值的信息和知识，再通过算法文本挖掘中最重要最基本的应用是实现文本的分类和聚类，前者是有监督的挖掘算法，后者是无监督的挖掘算法。

二、分类模型的构建

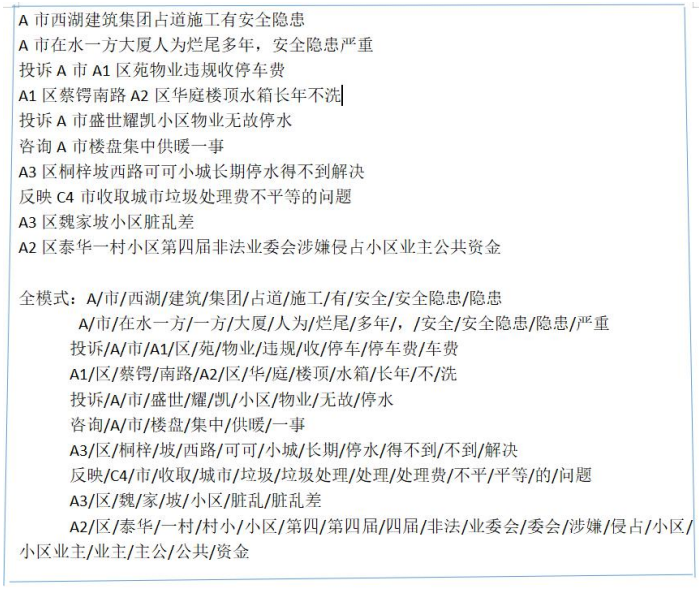
2.1 数据预处理

2.1.1 初步处理

通过对附件 2 的数据进行初步分析，对留言内容进行以下处理：（一）数据去重。留言数据存在重复，由于数据保存在 Excel 表，直接用 excel 进行去重处理，分别选中留言标题和留言内容两列进行去重，一共去除重复数据 344 条。（二）数据去噪。对于无法判断所属类别的数据，统一作为噪声数据去除处理。具体按以下三种情形进行处理:一是剔除只有街道名、小区名等的标题，如“K8 县”、“B9 市南桥镇潼塘村”；二是去除无分类特征词的留言标题，如“您好”、“深度探讨”、“ Potato）”。降噪过程共计删除留言 30 条，经过去重降噪处理共除去 374 条数据，剩余 8836 条留言内容数据，作为后续词向量模型训练输入的语料文本来源。（三）转为短文本。对于附件 2 的有些长文本的无意义表达太多，所以对长文本提取关键字，转为短文本。（四）语义带来的词语交叉处理。在留言内容中存在一词多义的现象，不同留言使用同一词汇，这需结合具体语境分到不同类别，因此采用考虑语境与目标词汇映射关系的 Word2vec 模型进行处理[1]。

2.1.2 分词

将初步处理完的 excel 表中的留言内容保存至 csv 文件。在中文中，只有字、句和段落能够通过明显分节符进行简单的划界，而对于词、词组来说，它们的边界模糊，没有一个形式上的分界符，所以，进行文本挖掘时，首先对文本分词，即将连续的字序列组按照一定的规范重新合成词序列的过程。本文采用 pycharm 的结巴分词工具对其进行分词处理，导入用户自定义词典，经过相关测试，此分词的精确度高达 98.56%，分词的部分结果示例如图 1 所示。


A 市西湖建筑集团占道施工有安全隐患
A 市在水一方大厦人为烂尾多年，安全隐患严重
投诉 A 市 A1 区苑物业违规收停车费
A1 区蔡锷南路 A2 区华庭楼顶水箱长年不洗
投诉 A 市盛世耀凯小区物业无故停水
咨询 A 市楼盘集中供暖一事
A3 区桐梓坡西路可可小城长期停水得不到解决
反映 C4 市收取城市垃圾处理费不平等的问题
A3 区魏家坡小区脏乱差
A2 区泰华一村小区第四届非法业委会涉嫌侵占小区业主公共资金

全模式：A/市/西湖/建筑/集团/占道/施工/有/安全/安全隐患/隐患
A/市/在水一方/一方/大厦/人为/烂尾/多年/，/安全/安全隐患/隐患/严重
投诉/A/市/A1/区/苑/物业/违规/收/停车/停车费/车费
A1/区/蔡锷/南路/A2/区/华/庭/楼顶/水箱/长年/不/洗
投诉/A/市/盛世/耀/凯/小区/物业/无故/停水
咨询/A/市/楼盘/集中/供暖/一事
A3/区/桐梓/坡/西路/可可/小城/长期/停水/得不到/不到/解决
反映/C4/市/收取/城市/垃圾/垃圾处理/处理/处理费/不平/平等/的/问题
A3/区/魏/家/坡/小区/脏乱/脏乱差
A2/区/泰华/一村/村小/小区/第四/第四届/四届/非法/业委会/委会/涉嫌/侵占/小区/小区业主/业主/主公/公共/资金

图 1 部分留言主题及分词效果

2.1.3 停用词过滤

经过中文词这一步骤后，将原来的文本处理成词的集合，即 $\mathrm{d}=\left(\partial_1;\partial_2;...;\partial_n\right)$，

其中 n 为文本 d 中出现词语的个数。由于留言内容中含有对文本含义表达无意义的词语，应进行删除，以消除其对文本挖掘工作的不良影响，此类词成为停用词。停用词的两个特征为：一是极其普遍、出现的频率高；二是包含的信息量低，对文本标识无意义。例如“在、的、啊、了”等等，因此通过去停用词操作，把文本中大量的语气助词、虚词、助词、非法字符等去除，得到分词后的每条留言文本数据。结果部分示例如图 2 下：

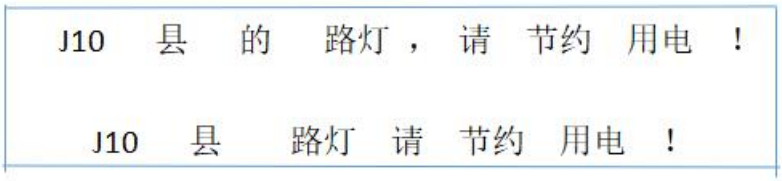


图 2 停用词效果前后对比

2.1.4 特征处理

留言内容预处理之后会产生许多特征词，如果直接使用预处理后的特征词进行挖掘，不但会造成特征表达上维度灾难，而且也会得不到高质量的聚类结果，因此，我们进一步开展特征提取，从而为后续的挖掘以及最终的分类带来更好的效果。

对于预处理后的留言数据，本文使用 Word2vec 模型训练留言数据获取文本向量。Word2vec 是快速处理词向量的方法，其中包含两种模型，一种是 CBOW（Continuous Bag-of-Words Model），一种是 Skip-gram 模型。其中 CBOW 的模式是已知上下文，估算当前语言的模型，而 Skip-gram 是已知当前词语估算上下文的语言模型，这两种语言模型如图 3 所示[2]。

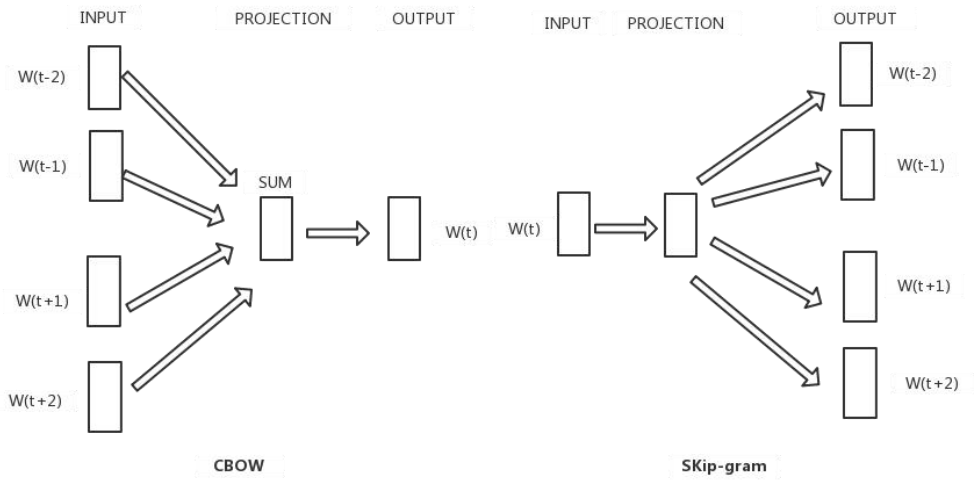


图 3 Word2vec 模型图

CBOW 学习目标是最大化对数似然函数：

$$L=\sum_{\mathrm{w}\in\mathrm{c}}\log\mathrm{p}(\mathrm{w}\,|\,\mathrm{Context}(\mathrm{w}))$$

SKip-gram 的概率模型可以写成：

$$\mathrm{p}(Context(w)\,|\,w)=\prod_{u\in Context(w)}p(u\,|\,w)$$

其中 c 代表语料库，w 代表语料库中任意一词，u 代表 w 上下文中的一个词语。

本文采用预处理后的留言内容文本作为训练语料库，采用 SKip-Gram 模型将分词处理后的词采用 Word2vec 训练成词向量。通过 python 调用 gensim 库中 gensim.models.word2vec 包，模型训练部参数设置如下：词向量维度 size=110，词向量上下文最大距离 window=5，需计算词向量的最小词频 min_count=5，训练完毕后，每个词都训练成一个 110 维的向量，保存模型并进行测试。

Word2vec 模型训练完毕后，进行词向量合成从而得到特征矩阵。其中合成的模式有几种模式：①直接进行词向量累加；②去重去噪后进行词向量的累加；③对词向量进行累加后平均。由于留言内容长短不一，为了避免长度差距带来的对词向量的合成的影响，我们采取累加后平均方式进行合成词向量，即将每一条留言内容包含的每个词的词向量相加再除以词的向量，从而得到词向量。

2.2 分类模型的构建

2.2.1 SVM 算法的原理

经典的支持向量机（Support Vector Machine SVM）方法是解决二类分类问题，支持向量机的分类原理如下：

假如给定一个特征空间的训练数据集：

$$T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$$

其中，$x_i\in x\subseteq R^n, y_i\in C=\{-1,1\}, i=1,2,3,...,n$，$(x_i)$是第 i 个实例，$y_i$是$x_i$的类别标签，当$y_i=\pm 1$时，$x_i$被成为正例，当$y_i=-1$时，$x_i$被成为负例，$(x_i,y_i)$称为样本点。当训练数据集是线性可分时，那么 SVM 的训练目标是在特征空间中找到一个能够将不同类别的样本实例分到两侧的超平面。然而，一般情况下这样的超平面不只有一个，而是存在无穷多个。因此，SVM 规定将距离两类数据的间隔最大的超平面视为最优越平面，如图 4 所示。虽然 B 和 D 也能够正确分类正例和负例，但是 B 和 D 不是最优越平面，最优越超平面是 C 唯一确定，该超平面能够正确区分正例和负例样本点，同样距离这些样本点的间隔最大[3]。 当在低维无法找到合适的界限时，就把数据投射到多维中，通过建立合适的核函数，完成投射，从而完成留言内容的分类，如图 4 所示

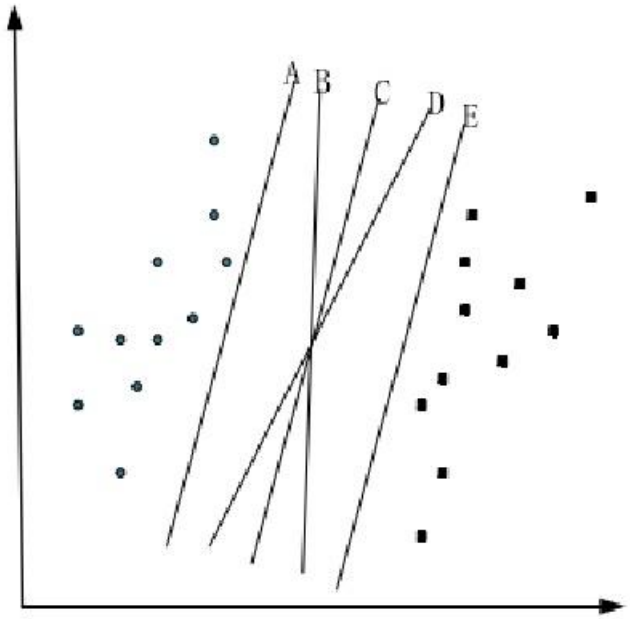


图 4 支持向量机原理

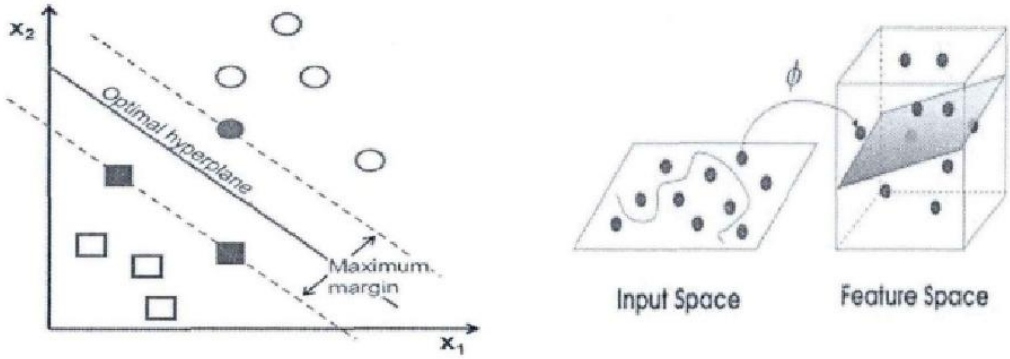


图 5 支持向量机图解

2.2.2 留言内容分类

留言内容的分类属于典型的多类分类问题，因此，我们采取将多类分类问题分解成多个两类问题求解，通过决策函数确定分类结果。我们采用 python3.7 调用 sklean 库中的 svm 包实现多类分类算法，其基本思想是找出能使不同类别样本数据的分类间隔最大的超平面。其具体过程如下：

（1）对上述得到的 110 维词向量进行归一化处理，通过 0-1 标准化对数据进行线性变换，使所有结果落在[0,1]区间，从而得到归一化处理的 110 维词向量$X_1\ X_2 \cdots X_{100}$。其次对留言内容的一级标签进行处理，SVM 模型作为有监督的分类预测模型，需将对数据标签Y_i追加到数据特征。将城市建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生七个类别分别标号 1、2、3、4、5、6、7、8、9，与归一处理后的 110 维词向量形成每一条留言内容样本的词向量表示$X_1\ X_2 \cdots X_{100}\ Y_i$，$其中 Y_i \in 1,2,\cdots,9$。

（2）将处理后的数据按照 7:3 的比例划分为训练集(data_train)和测试集(data_text)，采用 python 软件进行 SVM 模型的训练与测试。所谓模型的训练，

即寻找支持向量，确定最大超平面的过程。通过模型的建立，使得留言内容根据留言主题的特征词自动分类。

2.3 模型的评价

本模型采用的评价指标主要包括准确率以及 F1-Sore。准确率（Accuracy）即对于给定的测试数据集，模型正确分类的样本数与总样本之比。准确率在某些场合下确实可以有效地评价一个模型的好坏，然而在一些极端的情况，却显得不那么重要了。例如，某些地区的总人数 110 万人，而人群中患有某种病的人数只有 110 人，一个负责检测行人是否患有该种病的模型只需要持续将行人归为“不患病”一类，即可超过 99%的概率。

$$\text{Accuracy}=\frac{\text{“预测正确”的样本数}}{\text{总样本数}}$$

由此可知，单纯使用准确率是不足够的，我们引入 F1-Sore 来解决上述现象，作为我们评价模型好坏的另一个指标。F1-Sore 是精确率（precision）和召回率（recall）的调和函数，其公式为：

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

其中P_i为第 i 类的查准率，R_i是第 i 类的查全率。

通过改变参数的取值，探究其与准确率之间的关系，如图 6 所示，随着参数的增大，平均准确率呈现先上升后下降的趋势。通过模型的训练与测试，最后得到 F1 值为 86。

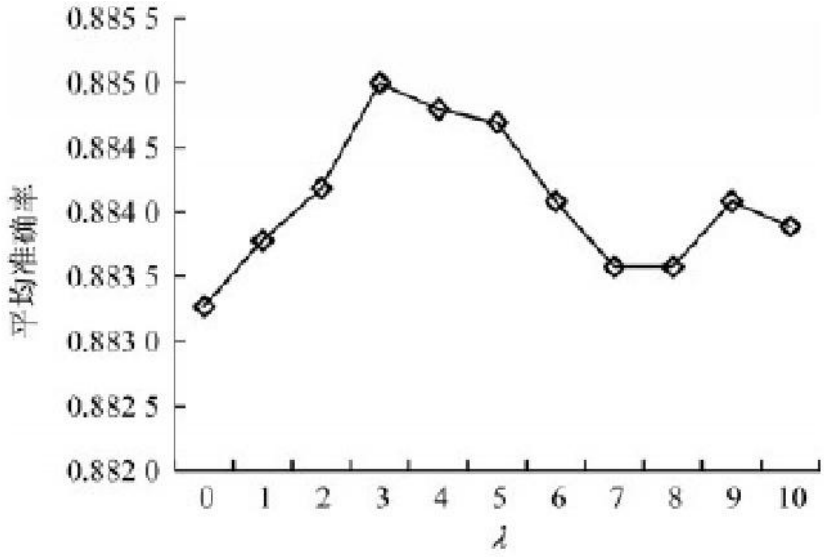


图 6 参数取值与准确率关系图

三、热度问题的处理

3.1 数据预处理

对附件 3 进行规范化文本处理，这些都需要用文本预处理来净化数据。我们

首先进行用户留言时间的划分归类，题目中说明某一时段内群众集中反映的某一问题可称为热点问题，设某段时间以月为单位，分别分为 15 个时间间隔，再对每个间隔的留言主题运用 Word2Vec 进行句段级别的相似度计算，最后利用留言文本相似度量化。

（1）附件 3 的数据是用户的留言评论文本，留言内容格式长短不一，存在口头表达方式和语言标点符号错用乱用情况，且留言时间的划分也需要剔除权重低无实际意义的数据。所以留言文本预处理的主要功能是对含有干扰项的留言样本按规则进行归一化处理，如口语化语气词、繁体字、错乱符号、数字序列等，再将原始留言信息样本根据匹配转换规则转换为统一的文本序列。以下是例子：留言样本为：###我是一名在 A 市某地铁站上班的安檢員%，我是由中介公司介绍来上班的，安检员岗位分两个班次，一个班从 630 上到 3，30，&一个班从 3：30 上到 11.，一个班九个小时 80-----90 元，各种保险福利没有就算了，还经常強制讓我們加班？！不加班还扣！扣身份证！Wtdnjjm 一天二拾肆个小时上班就上了拾捌个小时，还不算上班路途的时间，加班完班第二天继续上早班的话，就只能睡三四个小时，最基本的睡眠都保证不了，同时还有许多未成年和我一样在地铁安检岗位上班，希望有关劳动監察部門能夠解救我們。

处理方法如下：

•处理错乱字符：将待测留言文本中的#、&、%、--、，、Wtdnjjm、！、I•等特殊符号剔除

•处理数字序列：将待测留言文本中的捌—8、拾—10 按照数字对照表进行转换，另再根据上下文相应的数字序列句意将缺少字符的数字赋予相应的字符，如 630—6：30。

• 处理繁体字：直接选中有繁体字句段，点击繁转简，将繁体转换成简体，希望有关劳动監察部門能夠解救我—希望有关劳动监察部门能够解救我。

（2）因为附件 3 给予的数据中有 2017、2018 年份的个别月份的极少数组数据，而问题提供的热度问题的定义是某一时间段，明显这些数据是无实际探究意义的，所以我们选择剔除这些部分的数据。问题如成分词和词性标注结果；然后按照停用词表删除停用词、无用词（根据问题一的方法），一个句子中能够真正表达句子含义的词为名词、名词性短语、动词和动词性短语等，电话号码、银行卡号等在策略提取中的意义较小，可以通过正则等方式识别并使用统一的词汇进行替换，最后只保留上述 4 种词性的词汇[4]。

3.2 热度问题评价指标的建立

3.2.1 数据相似度分析

本题用 word2vec 模型的计算句子语义相似度的程序，在计算句子语义相似度的时候，都是以句子对的形式输入到网络中，定义为两个网络结构分别来表征句子对中的句子，然后通过曼哈顿距离，欧式距离，余弦相似度等来度量两个句子之间的空间相似度[5]。如图 7。并在每一个间隔选取句子，每组两个留言文本进行相似度计算，随机选取其中间隔运行结果如图 8 所示。

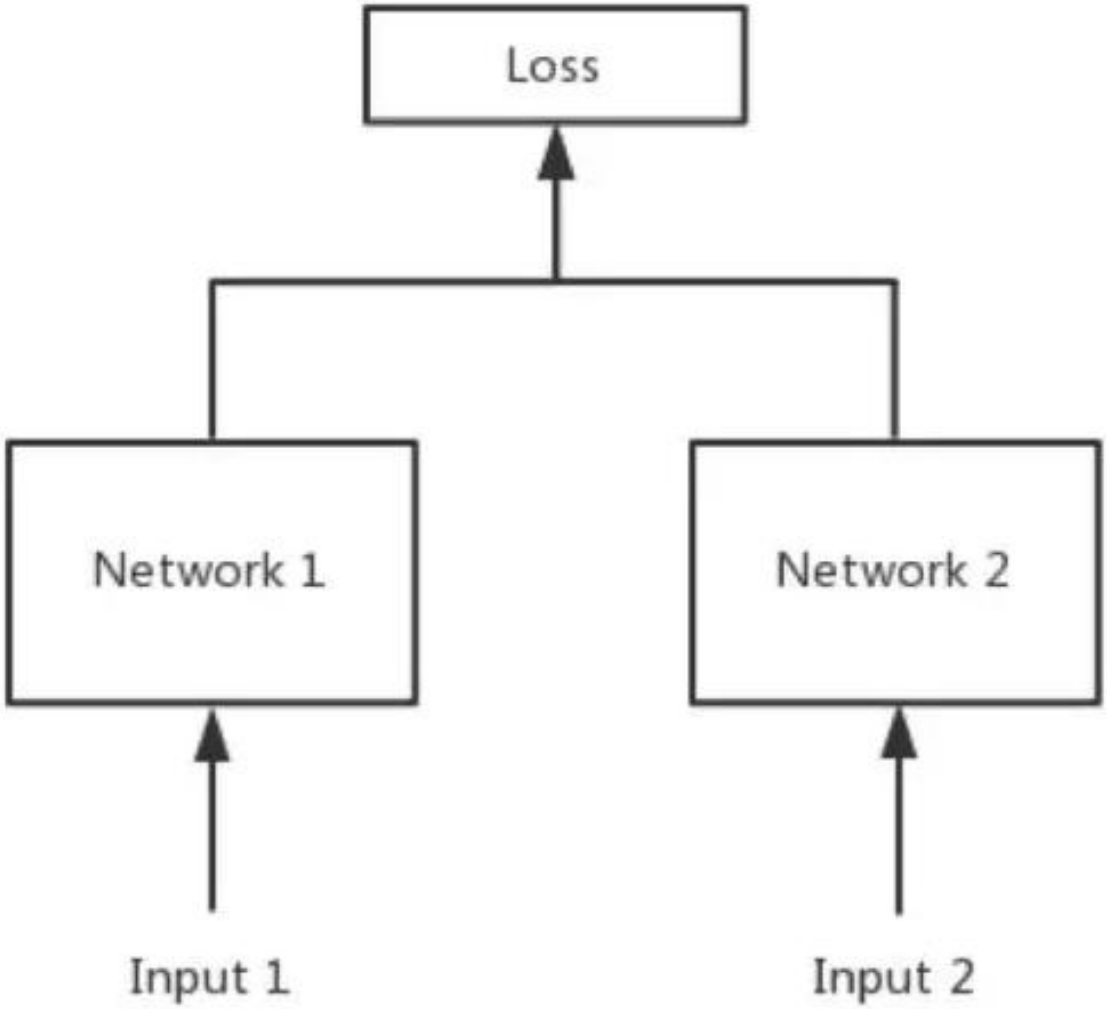


图 7　句义相似度挖掘结构图

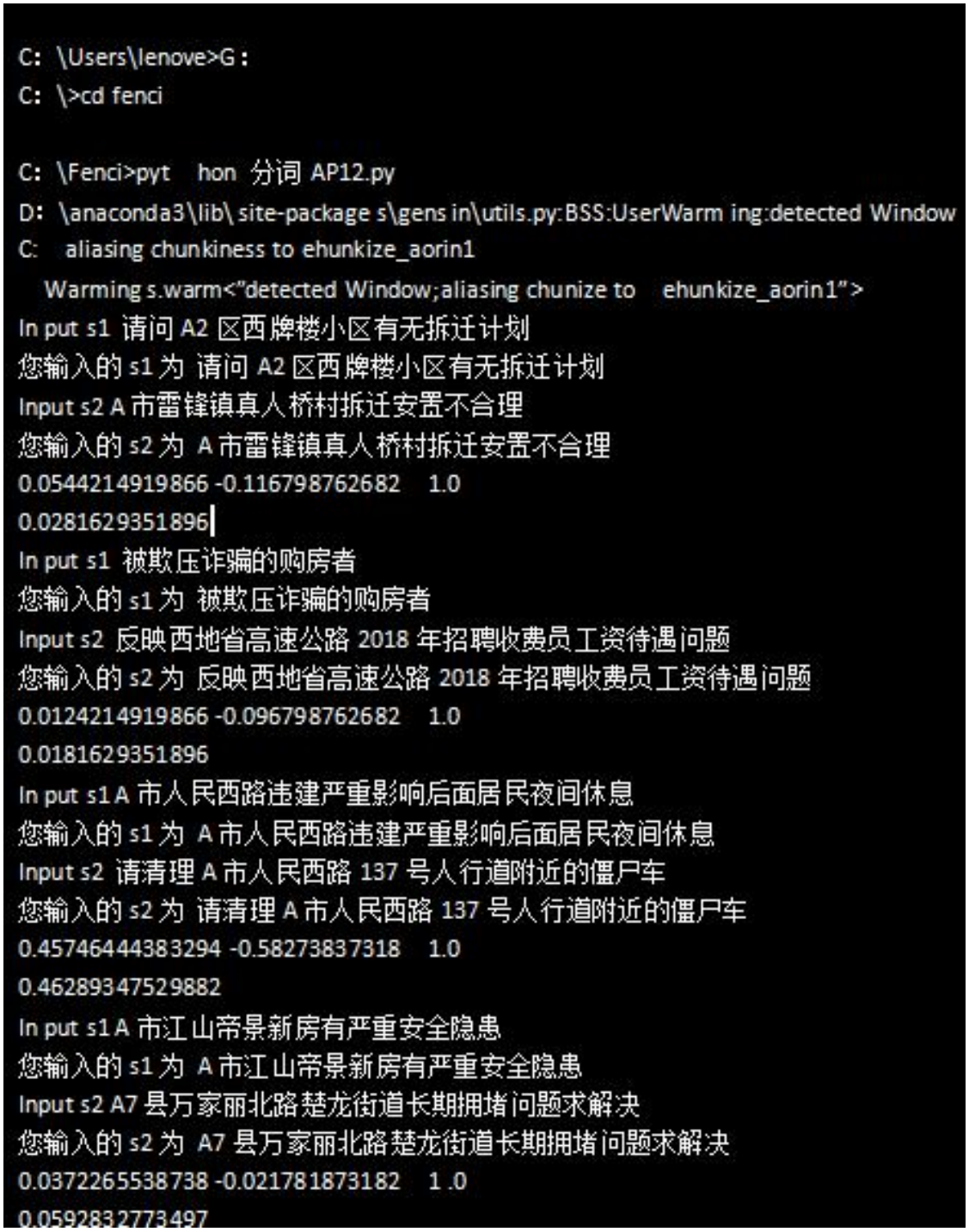


图 8 运行结果

3.2.2 长尾关键词

长尾关键词是指网站上非目标关键词但也可以带来搜索流量的关键词。如 1：目标词-厦门旅游，长尾词-厦门旅游哪里好？厦门旅游你有什么推荐？厦门旅游胜地有哪些？；这些以长词或短语的词就是长尾词。2：目标词-雨果培训，长尾词-雨果培训去哪家？参加雨果培训有什么意见吗？雨果培训课程大纲；这些以长词或短语的词就是长尾词。对于小型站点或小型企业站点来说，优化重心要放在目标关键词上面，长尾关键词数量少，带来的流量也没有几个，大多数流量都

来自于目标关键词（主词、核心词）。根据附件 3 的留言内容通过问题一的数据预处理利用词云提取关键词，就可以得到每一个时间间隔的长尾关键词，站点多集中目标关键词权重，拥有良好的排名和流量，也可以作为热度问题的指标参考的一部分，一般来说，网站优化人员在做小型站点或企业网站优化时。如图 9 就是 2019 年 6 月的留言主题的长尾关键词挖掘结果。

长尾关键词挖掘结果

序号	种子词	拓展词	整体搜索量	PC搜索量	移动搜索量
1	咨询A市人才购房补贴政策	成都人才补贴政策	7	0	0
2	A市高配置的电动老年代步车为何不能上牌?	可以上牌的老年代步车	126	7	112
3	A市交警推出的12123app不实用	12123app怎么交罚款	21	0	21
4	A市交警推出的12123app不实用	12123手机违章app	14	0	14
5	A市高配置的电动老年代步车为何不能上牌?	老年代步车需要上牌吗	21	0	21
6	咨询人防车位产权的问题	人防车位产权属于谁	378	70	308
7	咨询人防车位产权的问题	人防车位有产权吗	21	0	21
8	咨询人防车位产权的问题	非人防车位有产权吗	28	0	21
9	A市交警推出的12123app不实用	深圳交警app	273	21	245
10	咨询A7县星沙旧城改造项目问题	旧城改造案例	70	42	28
11	咨询A市人才购房补贴政策	高层次人才购房补贴	7	0	0
12	A市高配置的电动老年代步车为何不能上牌?	老人四轮电动代步车价格	7	0	7
13	A市高配置的电动老年代步车为何不能上牌?	老年四轮代步车哪个品牌好	21	0	21
14	咨询人防车位产权的问题	人防车位使用权	35	7	21
15	A市交警推出的12123app不实用	广东交警app	14	0	7
16	A市高配置的电动老年代步车为何不能上牌?	电动老人车四轮车价格	14	0	7
17	A市交警推出的12123app不实用	北京交警12123手机app	21	0	21
18	A市交警推出的12123app不实用	广州交警app	105	0	98
19	咨询人防车位产权的问题	人防车位产权	35	14	14
20	A市交警推出的12123app不实用	北京交警app	1148	112	1036
21	A市交警推出的12123app不实用	12123交管下载	70	7	56

图 9 长尾关键词挖掘结果

3.2.3 热度问题评价指标

留言内容数据与软件模型的分析，人对群问题的留言进行不同时间段归类，根据每个间隔的相似度计算结果，查找出句义相似度最高的一组留言文本，再筛选出此留言文本所表达同样意思的留言主题，对留言主题实行量化，并对相应时间段留言主题进行长尾关键词处理，综合两者的文本挖掘结果，由此确定每一个间隔表示某一段时间的某个特定群体集中反映的合理的问题热度评价指标，得出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”

四、评价方案

4.1 主要指标的确定

评价指标体系反映了各个因素对政府答复意见质量的影响，只有合理地构建

指标体系，才能有效地衡量答复意见的质量，为了更好地评价政府答复意见的质量，在构建答复意见的质量评价指标体系时，其主要指标如表 1 所示：

主要指标	含义
相关性	相关性是指两个变量的关联程度，而这里的相关性是指对于留言内容的答复是否与留言问题之间的关联程度，如果答复意见与留言内容呈正相关，则答复意见具有好的质量
完整性	完整性是指针对留言内容，政府给出的答复是完整的，能够完整地回答公众的问题。答复的完整性会影响到政府的服务质量以及人们的满意度
可解释性	由于留言的群众有农民、学生等，他们对一些专业名词并不熟悉，如果答复中含有大量的专业名词，会导致一些留言群众看不懂，这将会影响人们对智慧政务的满意度
及时性	当评价对象本身或人们对评价对象的认知程度随时间发生变化时，人们就会面临评价体系的失效问题。时效性是指通过答复的时间与留言的时间之间的比较，从而可知答复是否具有时效性，及时回答留言的问题

表 1 主要指标及其含义

4.2 不同评价指标的相关性分析

通过相关性分析，判断主要指标之间是否存在过强的关联性，如果指标之间有较强的关联性，则说明指标之间存在冗余。因此当建立一个特征矩阵时，应检验其指标之间的相关系数，保留最直接能体现评价结果的有效指标，从而减少工作量。利用附件 4 的数据对不同指标的相关性进行分析，相关性的计算结果如表 2 所示

序号	相关性	完整性	可解释性	时效性
相关性	1			
完整性	0.0403	1		
可解释性	0.2960	0.0540	1	
及时性	0.0201	0.0101	0.0055	1

表 2 指标相关分析结果

由相关性分析结果显示，每个指标之间均不存在强相关性，相关性最大的相关性和可解释性两个指标之间的相关性也仅为 0.296。不存在冗余特征，所以根据这所有四个指标对答复意见进行评价。

4.3 权重确定方法与结果

对于影响答复意见质量的因素，我们采用层次分析法对指标进行计算其权重。首先我们确定目标层为答复意见质量好，将四个指标作为方案层，从而构建

出层次分析结构，如图 10 所示。

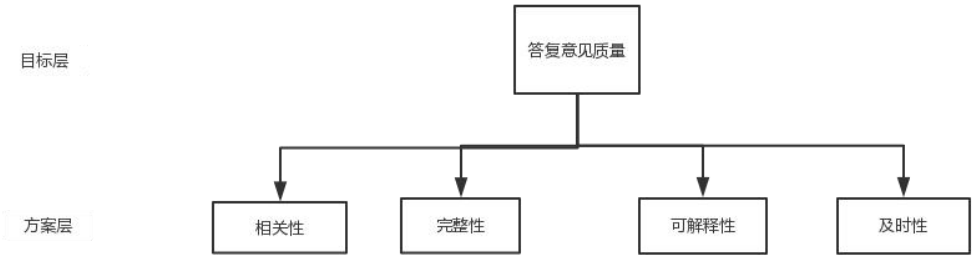


图 10 层次结构

对于方案层各指标元素，给出其相对重要性，构建判断矩阵。

方案层元素对目标层的相对重要性判断矩阵：

$$A=\begin{bmatrix} 1 & 7 & 6 & \frac{7}{5} \\ \frac{3}{7} & 1 & 3 & 3 \\ \frac{1}{7} & \frac{1}{3} & 1 & 1 \\ \frac{1}{7} & 5 & 5 & 1 \end{bmatrix}$$

最后通过计算得出其权重为：

	答复意见质量
相关性	0.196
完整性	0.030
可解释性	0.035
及时性	0.023

4.4 答复意见质量的评估标准

根据权重的分析，我们通过答复意见与留言内容的之间关系以及留言群众对答复的满意程度进行最终的评估标准的确定，如表 3 所示

分数	指标
0	答复意见与提问没有相关性（答非所问）
	答复意见只有一部分，没有完整性
	答复意见的可解释性差
	答复的时间与提问时间相差甚远
0-50	相关性、完整性、可解释性和及时性只具有相关性一个

50-70	相关性、完整性、可解释性和及时性只具有相关性与其他三个之一
70-90	相关性、完整性、可解释性和及时性只具有相关性与其他三个之二
90-100	相关性、完整性、可解释性和及时性只具有其中四个

表 3 答复意见质量的评估标准与相应的得分

4.5 评价结果

根据评价方案，以相关性、完整性、可解释性和及时性对附件 4 答复意见进行评价，大部分的答复意见均具有相关性、完整性、可解释性和及时性，其分数大多在 80 分-100 分之间，答复意见质量好。其部分结果如表 4 所示。

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	分数
25918	UU0081182	1区南路路	/4/21 1:4	区南路A市	UU0081182	/5/28 15:1	0
25431	UU0081037	能否根据房	5/24 15:1	方面称无	16年6月12	6/12 10:5	
131948	UU0081287	业管理混	/7/8 14:2	老总及其亲	情况的，》	/8/5 10:0	80
140568	UU0081853	处M2县套	9/12 12:1	货车在M2	月14日转交	10/23 10:1	
45807	UU0082021	枣镇莲云公	3/9/4 9:0	现在基础处理，2018		/12/14 8:4	
46512	UU0081952	区路灯能	1/19 13:4	灯就关了	区需要加强	/1/25 17:0	95
47572	UU0082226	关于编制	9/25 15:1	。听闻在龙	的相关政	9/25 17:0	

表 4 部分评价结果

五、结语

信息服务是政府公务活动的重要内容，政府通过多种方式来获取信息以了解民意，并通过对信息的分析来采取相关措施，虽然互联网的普及为公众反映诉求提供了更多的途径，但是也导致网络问政平台的信息量不断增加，庞大的信息量无疑给信息的分类整理加大了难度，也给相关工作人员带来了压力，增加了出差错的可能性，纯人工的处理方式已经不再适用，但随着大数据等技术的发展，对留言信息的分类和分析已经有了更便捷有效的方式。

本文正是通过构建 Word2vec 模型以及 SVM（Support Vector Machine）模型，将大量数据应用 python 等软件处理来实现对留言信息的自动分类，以减少人工，再基于 word2vec 模型对具体留言内容进行分析，进而找出一段时间内公众关注的热点问题，以利于有关部门能够依据公众关注度对留言中所述问题进行分批处理，来及时了解所面临的问题的发生地等基本特征，具有一定的科学性和合理性，一定程度上提高了对群众问政留言的处理效率。而针对答复意见，本文中运用层次分析法对建立的评价指标体系进行权重评估，以此作为答复意见的评价标准，较为客观的反映了政府对群众留言的关注度和反馈的及时准确性。

但是本论文依然存在以下几点不足：（1）在数据方面，本文所用数据仅有 9210 条，经过去重、去噪等处理后仅剩 8836 条，与其它研究中的数万数据相比，数据处理量显然还不够大。(2)在方法方面，虽然 SVM(Support Vector Machine)模型实现了留言信息的自动分类，且与其它文本分类模型相比具有较大优势，但

其模型训练复杂度高，容易出现不确定因素来影响测试结果的准确性。

六、参考文献

[1]段尧清;姚兰.政媒融合问政平台非正式文本自动分类匹配研究[J].情报理论与实践,1-9.
[2]丁世涛;卢军;洪鸿辉;黄傲;郭致远.基于 SVM 的文本多选择分类系统的设计与实现[J].计算机与数字工程,2020,v.48;No.363,152-157.
[3]宿绍勋.基于 SVM 文本分类的传销识别研究[C].北京工业大学,2019.
[4]艾楚涵;姜迪;吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技
术,2020,v.44;No.341,73-78.
[5]钱芸芸;杨文忠;姚苗;李海磊;柴亚闯.融合主题相似度权重的主题社区发现模型[J].计算机工程与应用,1-11.