“智慧政务”中的文本挖掘应用

摘要

近年来，随着互联网络的快速发展和不断普及，人们的生活习惯发生了深刻的变化，人们愈来愈倾向于借助网络这个平台发表自己的观点及看法。以微信、微博、市长信箱、阳光热线等途径逐步成为政府了解民意、汇聚民智、凝聚民气的重要信息来源方式。与此同时文本挖掘目前已经被广泛应用于各个领域，建立基于自然语言处理技术的智慧政务系统已经成为致力于社会治理创新发展的新天地。群众留言作为一种重要的信息源，数据充足丰富，并且大多数的群众留言内容出于主动，具备有效性、准确性和真实性。基于此，通过解析网络平台上获得的留言信息，运用文本挖掘技术，使政府的管理水平和施政效率获得更大提升具有重大意义。

对于问题 1，对文本数据进行去重，去重即仅保留重复文本中的一条记录，得到不重复的留言文本数据。利用 jieba 中文分词工具对每条留言内容进行分词，将其向量化，然后根据自行收集整理的停用词词库去除部分特征符号及无意义的语气词及词语等，使得文本向量的维数降低，进而依据所提取的关键词生成不同表征意义的词云，再利用 TF-IDF 进行特征值计算，检验数据的拟合度和关联度应用于训练集样本，结合四种模型，使用 F-Score 对四种分类模型进行评价，得出各条留言所属分类类别。

对于问题 2，初始化数据，将附件 3 中的时间列单元格转化为日期时间格式，而后导入数据。利用 TF-IDF 计算文本相似度。通过对所有留言内容两两进行相似度比较，大于某一阈值即归为一类，即首先从众多留言中识别出相似留言。再从留言主题及内容中提取表示地点或人群的特征信息，将相似留言问题归类。然后定义热度内涵构建热度指标模型，结合所建构的热度指标模型对归类后留言问题进行分析，给出排名输出留言文本序号，检查计算结果、不断调优，筛选出题目所需表格。

对于问题 3，初始化数据，通过根据 Levenshtein 距离相似度比较每条留言内容与答复的相似距离，获取留言内容与答复的长度，依据所设定指标，建立评价方案，依据相关性，完整度，解释度对答复意见的质量进行评价。

关键词：jieba 分词 自然语言处理 TF-IDF 算法 Levenshtein 距离 LinearSVM

Abstract

In recent years, with the rapid development and popularization of Internet, people's living habits have undergone profound changes, and people are increasingly inclined to express their views and opinions with the help of the network platform. Wechat, Weibo, mayor's mailbox, sunshine hotline and other channels have gradually become an important source of information for the government to understand public opinions, gather people's wisdom and morale. Simultaneously, text mining has been widely applied in various fields. The establishment of intelligent government system based on Natural Language Processing technology has become a new field of social governance innovation and development. As an important information source, the mass message has abundant data, and most of its contents are active, effective, accurate and authentic. Based on this, it is of great significance to analyze the message information obtained on the network platform and use the text mining technology to improve the management level and governance efficiency of the government.

For question 1, the text data is de duplicated, and only one record in the duplicate text is retained to get the non duplicate message text data. Using the Chinese word segmentation tool of 'Jieba' to segment each message content and quantify it, and then according to the self collected and sorted out stop words thesaurus to remove some feature symbols and meaningless modal words and words, so as to reduce the dimension of the text vector, and then generate word clouds with different representation meanings according to the extracted keywords, and then use TF-IDF to calculate the feature value and test the number According to the fitting degree and correlation degree of the training set samples, combined with four models, four classification models are evaluated with F-score, and the classification category of each message is obtained.

For question 2, initialize the data, convert the time column cells in attachment 3 to the date time format, and then import the data. TF-IDF is used to calculate text similarity. Through the similarity comparison of all messages, if the similarity is greater than a certain threshold, it will be classified into one category, that is, the similar messages will be identified from many messages first. Then extract the characteristic information of the place or crowd from the topic and content of the message, and classify the similar message problems. Then define the connotation of heat degree to build the heat degree index model, and analyze the problem of classified messages based on the heat degree index model, give the rank output message text number, check the calculation results, constantly optimize, and screen out the table required by the topic.

For question 3, we initialize the data, and then we compare the similar distance between each message content and the reply according to the Levenshtein Distance similarity, obtain the length of the message content and the reply, establish the evaluation scheme according to the set indicators, and evaluate the quality of the reply according to the correlation, integrity and interpretation.

Keywords: Jieba word segmentation, natural language processing, TF-IDF algorithm, Levenshtein Distance, linearSVC

1 挖掘目标

本次建模目标是利用题目所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大幅意见的文本数据，通过 jieba 中文分词工具对每条留言内容进行分词，通过随机森林分布、线性支持向量机（SVM）、贝叶斯、逻辑回归等模型达到以下目标：

1 通过对群众留言进行预处理数据清洗、jieba 中文分词、去除停用词、绘制词云、特征提取建立关于留言内容的一级标签分类模型，并结合 F-Score 对所选定模型不断优化和测试。

2 依据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，定义恰当的热度评价指标，并按规定格式给出评价结果。

3 针对相关部门对留言的大幅意见，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。

2 分析方法与过程

问题 1 分析方法与过程

图 1 文本挖掘流程图如下。

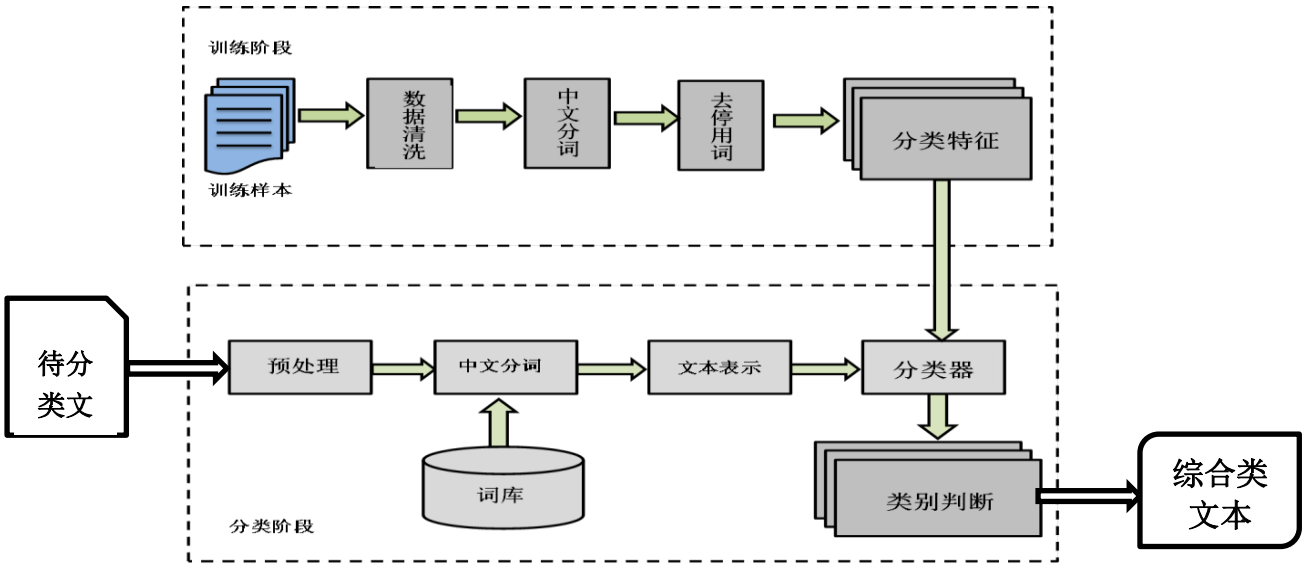


图 1 文本挖掘流程图

本用例主要包括以下步骤：

步骤一：数据预处理

（1）群众留言信息的去重、去空

在数据的的储存和提取过程中，由于技术和某些客观的原因，会造成一些留言文本内容缺失或重复记录等情况，因此首先需要对文本数据进行去重，去重即仅保留重复文本中的一条记录，得到不重复的留言文本数据。由图 2 重复信息可看出本例所分析的留言文本信息数据中存在重复信息。

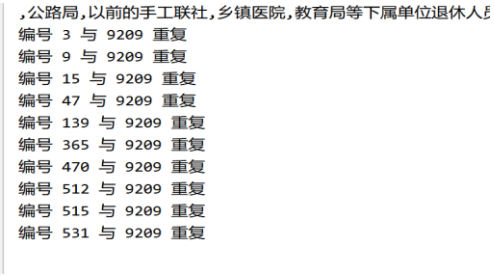


图 2 重复信息

（2）对留言信息进行中文分词

群众留言数据大多是非结构化数据，计算机不能直接进行数据的分析和处理，如果完全依靠人工方法去获取所需的信息要花费大量的时间和精力，想要全面了解留言内容，快速精准提取留言中有价值的信息，将会十分困难。因此，需先要把非结构化的留言文本信息转换为计算机能够识别的结构化信息。本文用的中文分词是 python 的中文分词包 jieba 进行分词。其分词效率高，分词准确，使用方便，是目前常用的中文分词解决方案，支持最大概率法（Maximum Probability），隐式马尔科夫模型（ Hidden Markov Model ），索引模型（ Query Segment）,混合模型（Mix Segment）共四种分词模式，同时有词性标注、中文姓名识别、关键词提取、文本 Simhash 相似度比较等功能。Jieba 采用的是基于词典规则来标注词性的，所以任意一个词在 Jieba 里有且只有一个词性，如果一个词有一个以上的词性，那么它的标签就变成了一个集合。前缀词典实现的高效词图扫描，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，使得能更好的实现中文分词效果。结合文本数据，通过添加自定义词典使得分词结果更贴切实际。表 1 分词结果为在 jieba 包中，通过自定义词典后的分词效果。

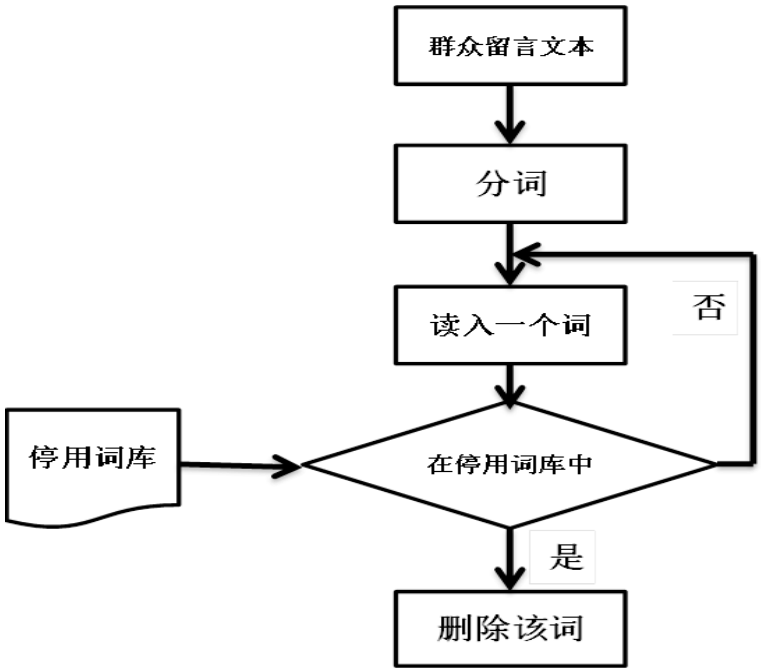
表 1 分词结果

源数据：
A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。

强烈请求文明城市 A 市，尽快整改这个极不文明的路段。
分词后：
“A3” “区” “大道” “西行” “便道” “未管所” “路口” “至” “加油站” “路段” “人行道” “包括” “路灯” “杆” “被” “圈” “西湖” “建筑” “集团” “燕子山” “安置房” “项目” “施工” “围墙内” “每天” “尤其” “上下班” “期间” “这条” “路上” “人流” “车流” “极多” “安全隐患” “非常大” “强烈” “请求” “文明城市” “A 市” “尽快” “整改” “这个” “极不文明” “的” “路段”

（3）去除停用词

完成文本分词后，并非所有的词都会被保留下来作为特征项。在留言文本当中，总是会出现一些出现频率很高但是对表达文本含义没有很大贡献的词，如“哈哈”、“啦啦啦”、“的”、“了”等，称之为停用词。如果停用词包含在文本当中会给后续的分析带来很多噪声，因此有必要将这类词过滤掉。过滤停用词的工作需要建立停用词表，即将分词结果中与停用表中一样的词剔除，并根据实际情况在现有停用词表中做增删添改。对于分词后文本中的每一个词，都需要检查该词是否存在于停用词表中，有则去除无则保留。停用词过滤流程如图 3 停用词过滤流程。



（4）绘制词云

词云作为一种可视化手段。由于其在分析文本数据时具有美观性、高效性，因此更利于我们将从留言文本中所提取的关键信息展示出来。“词云”就是对网络文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或

“关键词渲染”，从而过滤掉大量的文本信息，使浏览网页者只要一眼扫过文本就可以领略文本的主旨。通俗讲，即频率越高的字体越大。通过数据的预处理及中文分词，我们使用 python 中的 wordcloud 完成词云绘制的操作。

步骤二：数据分析

在对留言文本信息分词后，需要把这些词语转换为向量，以便挖掘分析使用。这里采用 IF-IDF 算法，找出每条留言文本信息的关键词，将留言文本转换为权重向量。

（1）TF-IDF 算法

词频逆文本频率算法（term frequency - inverse document frequency 以下简称 TF-IDF）是一种普遍应用在搜索排序和数据挖掘等领域的特征加权算法。TF-IDF 用来对文档中的每一个词进行重要性评价，并给出该词对于该文档的重要性权值。TF-IDF 评估方法是统计待评估词在某一文本中的出现频次以及该词在整个文本库中出现的频率，然后按照与某一文本中的词频正相关与文本库中该词频率负相关的原理进行词权重的估计。TF-IDF 的计算表达式是：TF * IDF，TF 指词频(Term Frequency)，IDF 指逆文本频率(Inverse Document Frequency)。TF 是指词语在文本中的频次统计，IDF 是指包含词语的文本数越小则 IDF 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留言文本信息中的关键词。

其中 TF、IDF 的计算方法分别如下所示。

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

式中 $n_{i,j}$ 指词 t_i 的词频。

$$idf_i = \log \frac{\left|D\right|}{1 + \left| j : t_i \in d_j \right|}$$

式（4-5）中 $\left|D\right|$ 指语料库中总的文本数，$1 + \left| j : t_i \in d_j \right|$ 指包含有词 t_i 的总文本数。

TF-IDF 的计算方法是直接获取 $f_{i,j}$ 和 idf_i 的乘积，表达式如下所示

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

经过 TF-IDF 算法分析，词语分散的分词词表转换为“词-权重”的词表形式。

步骤三：数据分类

分类是文本挖掘中常用的方法。有监督的分类算法分为经典机器学习算法和神经网络算法。经典的文本分类算法有朴素贝叶斯算法（Naïve Bayesian,NB），线性支持向量机（Linear Support Vector Machine,LSVM）,随机森林（Random Forest,RF），逻辑回归(Logistics Regression,LR). 有监督的分类方法由于其有明确的评价标准，并且在实际运用中取得了较好的效果，已经成为了文本分类领域的主流。其中，SVM 分类器通用性较好，且分类精度高、 分类速度快，应用前景较为广泛。因此，在本项目中，我们选择线性 SVM 作为分类器对训练集样本数据进行测试，而后使用贝叶斯算法、线性支持向量机、随机森林、逻辑回归四个模型对数据进行分类对比。

（1）贝叶斯算法

贝叶斯分类的基础就是贝叶斯定理，这个定理解决的是在日常生活中经常遇到的一些具有不确定性的问题。贝叶斯分类算法中最简单也是最常用的就是朴素贝叶斯算法，它的思想基础是：对于已知要分类的目标，只要得知它在所有条件下，每个属性中出现的概率就可以判断出属于哪个类别，哪个概率大，分类目标就属于哪个类别。

根据朴素贝叶斯分类算法原理，可以将分类过程总结为以下步骤：

第一步：为朴素贝叶斯分类做准备工作，主要是根据具体情况确定特征属性，并对每个特征属性进行类别的划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。由人工输入所有待分类数据，输出特征属性和训练样本。

第二步：计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。

第三步：使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。

在实际应用中，把留言看成一个分类问题。首先收集大量的留言文本信息当作样本，然后使用贝叶斯分类器对收集到的样本进行有指导的学习，最后使用训练好的贝叶斯分类器对新写入的留言进行分类。

（2）线性支持向量机

支持向量机是一种二分类模型，基本定义是在特征空间上的间隔最大的线性分类器。它的基本思想是对原始特征映射到高维特征空间，再从新的特征空间中搜索超平面，在保证正确划分的同时使硬间隔或者软间隔最大化，这样的超平面分类误差较小，对于未知数据有很好的分类预测能力。直观的 SVM 超平面分

类效果如图 4。

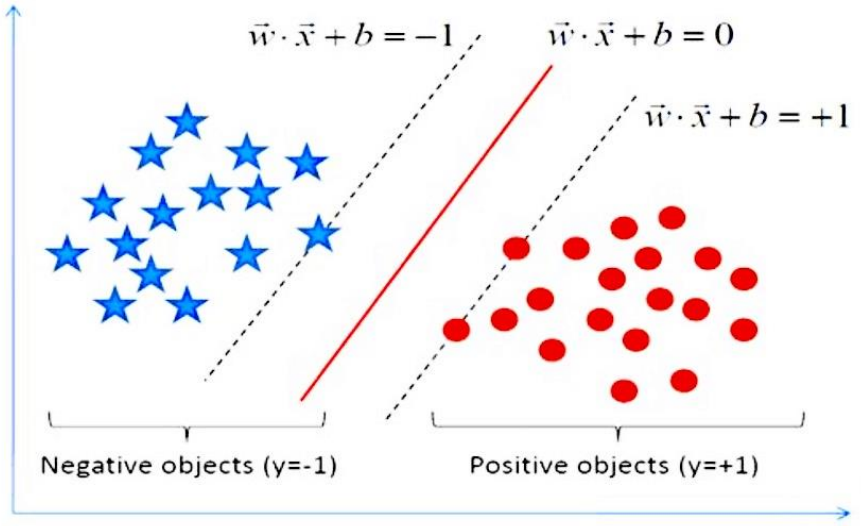


图 4 SVM 超平面分类效果

（3）随机森林

随机森林是由很多决策树模型 $h(\chi,\Theta_k),k=1,2\ldots$ 组合而成的分类模型，基本思想是袋装和随机特征子空间方法，其中 χ 为输入自变量，Θ_k 则是独立同分布的随机变量。通过 K 轮训练后，得到 $h_1(\chi),h_2(\chi),\ldots,h_k(\chi)$ 分类序列，它们一起组合成了一个集成分类系统。在给定的输入变量 χ 下，每个决策树都有一票的投票权利来选择最终的结果。最后的分类结果为分类序列中结果数量最多的那个。

（4）逻辑回归

逻辑回归的原理源于线性回归，但是两者目的不同。逻辑回归利用线性回归的结果来预测正确类的对数概率。其中某类型的事件，其几率（odds）发生概率 p 和不发生概率 $1-p$ 进行比值，对数几率表达式为 $\log(p/(1-p))$。逻辑回归算法的目标是得到某个表达式能够让输入得到映射，从而预测出其所属的分类。

步骤四：数据的拟合度与关联度

关联分析是数据挖掘中最常用的分析方法之一，关联分析的主要作用是发现潜藏在数据中的重要的、有价值的信息。利用关联分析技术挖掘出的知识发现数据中隐藏的规律，进一步对模型的准确度进行检验。

问题 2 分析方法与过程

由于所给留言集合比较大，对于捕捉当前最新最热的热点问题造成了困难。我们的热度评价模型主要是对相似留言归类为一个问题，计算其热度指数。热度评价的流程图如图，图 5 热度评价的流程图所示：

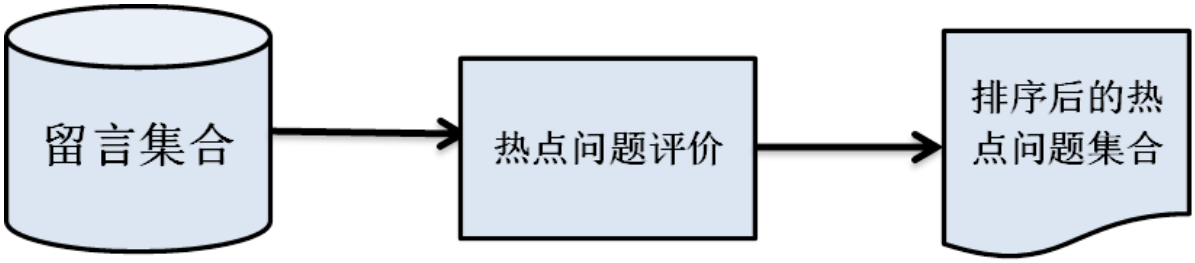


图 5 热度评价的流程图

1 数据预处理

初始化数据，将附件 3 内的时间列单元格转化为自定义时间日期格式(“yyyy-mm-dd hh:m:s”), 导入数据。

2 文本相似度计算

计算出留言中每个词的 TF-IDF 值，然后结合相似度计算方法（一般采用余弦相似度）计算两条留言的相似度，将大于某一阈值的留言归为相似留言（本例中规定相似度阈值为 70%），即将相似留言合并为一类。

2.1 利用 TF-IDF 计算相似留言步骤如下：

1)使用 TF-IDF 算法，找出两条留言的关键词
2)每条留言各取出若干个关键词（比如 20 个），合并成一个集合，计算每条留言对于这个集合中的词的词频（为了避免文章长度的差异，可以使用相对词频）
3)生成两条留言各自的词频向量
4)计算两个向量的余弦相似度，值越大就表示越相似

2.2 余弦相似度

余弦相似度就是通过一个向量空间中两个向量夹角的余弦值作为衡量两个个体之间差异的大小。把 1 设为相同，0 设为不同，那么相似度的值就是在 0~1

之间，所有的事物的相似度范围都应该是 0~1，余弦相似度的特点是余弦值接近 1，夹角趋于 0，表明两个向量越相似。

向量余弦公式：

$$\sin(A,B)=\cos(\theta)=\frac{A\bullet B}{\|A\|\|B\|}=\frac{\sum_{k=1}^{n}A_k\times B_k}{\sqrt{\sum_{k=1}^{n}A_i^2}\times\sqrt{\sum_{k=1}^{n}B_i^2}}$$

3 构建热度指标

某一时段内群众集中反映的某一问题可称为热点问题。本文利用 TF-IDF 算法的余弦相似度来衡量群众留言的相似性，找出相似度超过 70%的相似留言。当在某一时段内群众留言描述的事件频繁出现，则表明该事件引起了热议，需要重点关注。根据所给数据，定义热度指标为：

$$热度=类别内的留言数量+（赞-踩）\times 0.5$$

4 数据筛选

对留言文本数据的留言主题以及内容进行检索，寻找与地点相关的信息，例如省、市、区等，以及人群信息比如学生，居民等。某些特定的场所也可以表面人群信息，如医院->医生，学校->学生等。通过检索地点或人群信息，结合上述热度指标对留言文本数据进行统计，计算热度指数，得出热点问题留言明细表，并从中筛选出排名位于前 5 的热点问题表。

问题 3 分析方法与过程

第一步：导入数据。根据 Levenshtein 距离相似度比较各个留言内容与答复的相似距离。Levenshtein 距离编辑距离又被称作 Levenshtein 距离，是本文中用于衡量每条留言内容与答复的相似距离的工具。编辑距离可以定义为将原字符串 W 转变成目标字符串 X 所必需的最少的基本操作的次数，也就是通常所说的最小代价。它通过动态规划的方式获得最佳版本路径的成本，其中计算最佳版本路径所需的三个基本操作成本分别是：删除操作、替换操作与插入操作。

第二步：获取每条留言内容及相对应的答复的长度。

第三步：根据以下指标，判断相关性，完整度，解释度。其中留言内容长度为$len1$，

答复意见为$len2$，相似距离为s。

表 1 判断指标

	相关性	完整度	解释度
$len1 \geq len2$ 且$s > 100$	低	低	低
$len1 \geq len2$ 且$50 < s \leq 100$	中	低	低
$len1 \geq len2$ 且$s \leq 50$	高	低	低
$len1 <$2$len2$ 且$s > 300$	低	中	低
$len1 <$2$len2$ 且$200 < s \leq 300$	中	中	低
$len1 <$2$len2$ 且$100 < s \leq 200$	中	中	中
2$len2 < len1 \leq$ 2$len2$ 且$s \leq 100$	高	中	高
2$len2 < len1 \leq$ 2$len2$ 且$s > 400$	中	高	中
2$len2 < len1 \leq 2 * len2$ 且$s \leq 400$	高	高	高

结果分析

问题 1 结果分析

本节通过对留言内容类别统计可得下图表。由图该图表可以直观的看出城乡建设、社会保障、教育文体、商贸旅游、环境保护、卫生计生、交通运输七个类别各个类的留言数据数量明显不同。其中城乡建设和社会保障数量相近，环境保护和卫生计生相差不多，交通运输类别的留言内容最少。

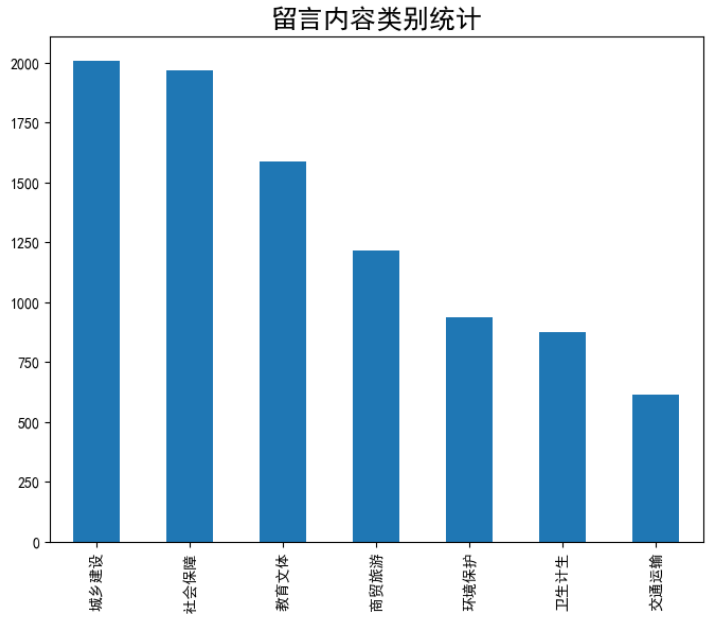


图 6 留言内容类别统计

词云结果

经过数据预处理、中文分词及去除停用词等处理，运用 python 中 wordcloud 得到所提起各类别的关键字如下图所示。

城乡建设高频词汇

环境保护高频词汇

环境保护高频词汇

交通运输高频词汇

教育文体高频词汇

劳动和社会保障高频词汇

商贸旅游高频词汇

卫生计生高频词汇

图 7 各个类别词云

从上图我们可以看出城乡建设的高频词汇为“领导”“业主”“开发商”“小区”等词汇，这与实际是相对吻合的。环境保护的高频词汇为“居民”“污染”“环保局”“生产”等词汇。交通运输的高频词汇为“出租车”“快递”“公司”“司机”等词汇。教育文体的高频词汇为“学校”“老师”“教师”“教育”等词汇。从上图我们可以看出劳动和社会保障的高频词汇为“职工”“工作”“单位”“领导”等词汇。商贸旅游的高频词汇为“公司”“部门”“电梯”等词汇。卫生计生的高频词汇为“领导”“医院”“医生”“患者”等词汇。

经过上述分析，根据题意需对分类模型给予评价并不断调试优化，提高其准确率。

通常使用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

通过将四个分类模型应用于留言文本信息的实验，使用 F-Score 对四个模型进行检验，其准确率如下图所示。从图中可看出，LSVC 的准确率最高，大致接近于 90%。

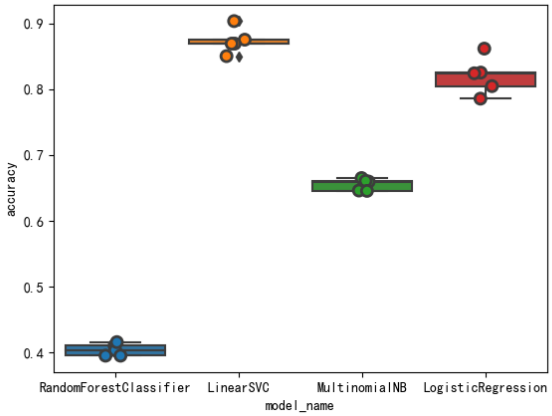


图 8 模型准确度比较图

使用 LSVC 模型对数据进行测试，计算得到的 F-Score 如下图所示。为了更好地说明分类模型的准确性，对数据进行预测，实际结果与预测结果混淆矩阵如图所示。

	precision	recall	f1-score	support
城乡建设	0.82	0.95	0.88	663
环境保护	0.96	0.93	0.95	310
交通运输	0.95	0.70	0.81	202
教育文体	0.94	0.93	0.94	525
劳动和社会保障	0.90	0.94	0.92	650
商贸旅游	0.91	0.83	0.87	401
卫生计生	0.95	0.85	0.89	289
accuracy			0.90	3040
macro avg	0.92	0.88	0.89	3040
weighted avg	0.91	0.90	0.90	3040

图 9 LSVC 模型 F-Score 图

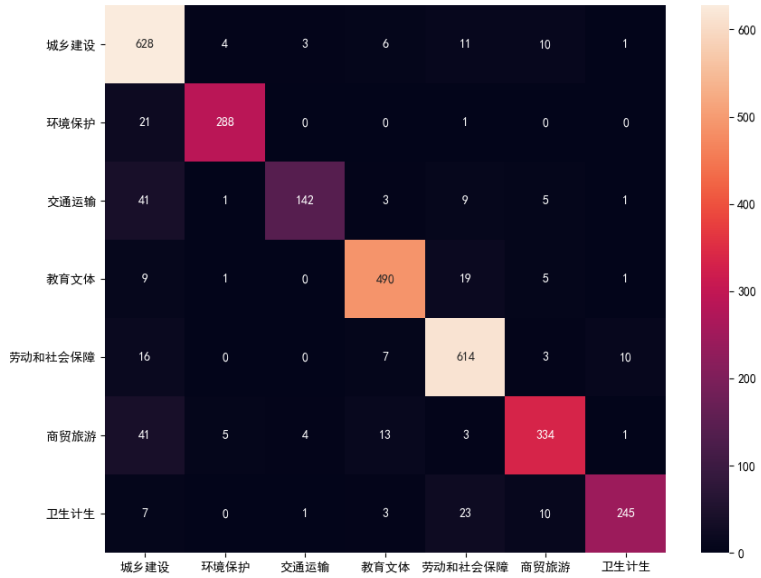


图 10 实际结果与预测结果混淆矩阵

问题 2 结果分析

通过计算留言中每个词的 TF-IDF 值，将大于某一阈值的留言归为相似留言，依据所定义的热度指标，对留言文本数据的留言主题以及内容进行检索，结合上述热度指标对留言文本数据进行统计，计算热度指数，得出热点问题留言明细表。

表 2 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	163.5	2019-01-09 至 2020-01-02	A 市 A2 区居民	A2 区黄兴路步行街大古道巷住户卫生间粪便外排
2	2	163.5	2019-01-01 至 2020-01-08	A 市 A3 区居民	A 市 A3 区中海国际社区三期与四期中间空地夜间施工噪音扰民
3	3	163.5	2019-01-02 至 2020-01-06	A3 区居民	A3 区青青家园小区乐果果零食炒货公共通道摆放空调扰民
4	4	163.5	2018-11-15 至 2020-01-06	A 市	咨询异地办理出国签证的问题
5	5	163.5	2019-01-01 至 2020-01-04	A 市 A3 区	A 市 6 路公交车随意变道通行

表 3 留言明细表

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	188039	A00081379	A2 区黄兴路步行街大古道巷住户卫生间粪便外排	2019-08-19 11:48:23	靠近黄兴路步行街，城南路街道、大古道巷、一步两搭桥小区（停车场东面围墙外），第一单元一住户卫生间粪……	0	1
1	189345	A00077163	A4 区东风街道蚌塘社区自来水管网改造设计存在严重缺陷	2019-07-31 06:52:29	我是 1996 年购的安居房，在东风街道蚌塘社区 9 栋 704 室，根据购房合同 089 第 5 条：室……	0	0

1	189635	A00029819	A1 区桐阴里小区一直夜间施工	2019-07-15 21:27:02	桐阴里夜间施工扰民，桐阴里小区长期以来一直夜间施工，对周边的住户影响较大，希望有关部门出面解决。	0	0
1	190077	A000112913	A3 区欣胜园小区急需水改电改和车位改造	2019-03-14 15:27:28	尊敬的胡书记：您好！我们是 A 市 A3 区欣胜园小区居民。我们小区建于 2000 年代初，因为当初的规划与设计的原因，小区的变……	0	0
…	……	……	……	……	……		

问题 3 结果分析

通过根据 Levenshtein 距离相似度比较每条留言内容与答复的相似距离，获取留言内容与答复的长度，依据所设定指标，建立评价方案，依据相关性，完整度，解释度对答复意见的质量进行评价。评价结果如下图所示：

留言编号	留言内容	答复意见	相关性	完整度	解释度
2549	2019 年 4 月以来，位于 A 市 A2 区桂花坪街道的 A2 区公安分局宿舍区……	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…	低	中	低
2554	潇楚南路从 2018 年开始修，到现在都快一年了，路挖得稀烂……	网友“A00023583”：您好！针对您反映 A3 区潇楚南路洋湖段怎么还没修好的问……	中	中	低
2555	地处省会 A 市民营幼儿园众多，小孩是祖国的未来……	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收……	低	中	低
2557	尊敬的书记：您好！我研究生毕业后根据人才新政落户 A 市，想……	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及……	中	中	低
2574	建议将“白竹坡路口”更名为“马坡岭小学”,原“马坡岭小学”……”	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人	中	中	中

结论：

对群众留言进行分析研究，了解群众的社会需求及问题，对政府提高管理水平及施政效率有重大意义，同时也是文本分析的一个难题。本文通过采用随机森林分布、支持向量机（SVM）、贝叶斯、逻辑回归四个分类模型构建关于留言内容的一级标签分类模型。使用 F-Score 对分类方法进行优化抉择，深入分析群众留言所传达的含义。

由分析结果可以看出，留言文本信息归为城乡建设、社会保障、教育文体、商贸旅游、环境保护、卫生计生、交通运输七大类别，对各类别分别提取关键词构建特征向量，利用 IF-IDF 算法计算文本相似度，定义热度评价模型，可以发现，热点问题多属于与城乡建设和社会保障相关的领域内。

通过利用编辑距离判断每条留言及对应答复的相关性、完整度、及解释度、得出相关部门对留言的答复意见质量相对较低，需进一步提升改进。

参考文献：

[1]王国薇. 基于深度学习的文本分类方法研究[D].新疆大学,2019.

[2]朱英龙. 基于机器学习的文本分类算法[D].西安科技大学,2019.

[3]彭湃. 自然语言处理—中文词和短文本向量化的研究[D].华中师范大学,2019.

[4]崔鹏程. 基于文本挖掘的学术文献内容智能识别方法研究[D].北京交通大学,2019.

[5]潘成龙. 基于深度学习的文本分类问题研究[D].华中科技大学,2019.

[6]宋艳青. 基于词向量的文本分类方法研究[D].燕山大学,2019.

[7]白治龙. 基于 Hadoop 的文本分类方法研究[D].河南科技大学,2019.

[8]杨静. 基于文本挖掘的网络招聘信息分析[D].山东师范大学,2019.

[9]仲远. 自然语言处理在信息检索中的研究和应用[D].江苏科技大学,2019.

[10]吉祥. 数据挖掘中关联规则算法的研究[D].江苏科技大学,2019.

[11]程颖涛. 基于深度学习的自然语言处理中问题分析的研究[D].西安邮电大学,2018.

[12]洪陈建. 序列和文本的熵压缩结构研究[D].西安电子科技大学,2018.

[13]戴舜. 自然场景下文本提取方法的研究[D].北方工业大学,2018.
[14]解波. 基于自然语言处理及机器学习的文本分类研究[D].云南大学,2018.
[15]何梦娇. 基于文本挖掘的苏州交通舆情分析[D].苏州大学,2018.
[16]崔哲. 基于朴素贝叶斯方法的文本分类研究[D].河北科技大学,2018.
[17]孙鹏. 基于文本挖掘上市银行年报情感词频与业绩关系实证研究[D].辽宁大学,2018.
[18]黄钊炜. 面向主题的文本挖掘研究与应用[D].华中科技大学,2018.
[19]张平霞. 基于文本挖掘的MOOC讨论区学习评价研究[D].重庆师范大学,2018.
[20]张丽. 文本挖掘中关键词与文本摘要自动提取研究[D].青岛理工大学,2018.
[21]范佳健. 微博评论信息的聚类分析[D].安徽大学,2017.
[22]于岗. 涉警网络舆情热度评价指标体系构建研究[D].中国人民公安大学,2017.
[23]张培华. 基于自然语言处理的社交网络数据挖掘研究[D].华北电力大学,2017.
[24]任朋启. 文本挖掘在产品评论中的研究与应用[D].江苏科技大学,2017.
[25]刘召明. 基于 Elasticsearch 的新闻实时词云系统设计与实现[D].华中科技大学,2016.
[26]黄情. 基于文本挖掘的网络舆情分析应用研究[D].大连海事大学,2016.
[27]丁诗晴. 基于在线网站评论的中文文本挖掘[D].华中科技大学,2016.
[28]张帆. 贝叶斯算法在校园留言板垃圾过滤中的应用研究[D].郑州大学,2016.
[29]唐靓. 拓扑保持的词云布局算法研究[D].合肥工业大学,2016.
[30]黄旭. 基于F-score模型的财务报告舞弊识别研究[D].北京邮电大学,2015.
[31]王川. 基于自然语言处理的作文自动评分系统研究[D].武汉理工大学,2015.
[32]王培培. 编辑距离快速算法研究[D].东北大学,2011.
[33]赵小龙. 粒计算在数据挖掘中的应用研究[D].西南交通大学,2007.