问政事务的模型研究与综合评价方法

摘要

近年来，随着科技的发展，网络在人们生活中也日益普及，中国公民以网民的身份通过互联网行使知情权、参与权、表达权和监督权进行网络问政也日益成为一大趋势，网络问政平台也逐渐成为了政府为人民工作的重要手段。在大量的问题反馈中，运用更加高效的分类方法显得越来越重要。

因此，分类计数也应随着科技的发展而得到优化，为了解决人工分类效率低下的问题，我们需要问政平台的后台程序能够通过数学模型对大量的文本数据进行处理，从而实现对留言类别的划分。此时，智慧政务系统将是一种新的工作趋势，我们选择了一种完整、合理的分类体系，以大大提高问政平台处理和文本挖掘的方法，对于不同的问题进行分析。

针对问题一，我们首先对于附件 1 进行数据挖掘，在充分利用其数据基础上，对其进行分类处理。接着对于附件 2，我们首先利用汉语词法分析系统 ictclas 的分词功能及 excel 的筛选功能对我们的 csv 数据进行预处理，利用数学统计学公式及空间向量模型得到各个类别的特征词序列，从而完成文本筛分器的组件，并代回到文本数据中进行回测，计算查准率与查全率，从而通过 F-Score 对此分类方法进行评价。

对于第二问，我们主要也是利用分词系统与 excel 对 csv 数据进行预处理，用数学公式得到合理的关键词，对其进行文本处理，得到可计算的序列，利用空间向量模型进行相似度计算，由此经过几类热门话题的相似度分析，我们得出能够筛选出最全、最准的热门问题的标准相似度，将相似度在此标准上的事件个数作为热度评价的项目，将热度评价排名前五的称为热门问题。之后再代回文本数据，得到题目所要求我们做出的热点问题留言明细表。

针对第三问，我们先对留言文本数据进行预处理，用数学公式计算特征性得出合理的关键词，利用空间向量模型得出各个方面的问题类别，再对应类别对该类别中的留言所对应的回复进行问题粒度的特征分析，将该类留言问题的回答基于关键词进行文本特征分析，再结合文本长度与回答的非重复字符数进行比较，对目标回复文本从相关性、完整性与可解释型进行质量评测。

关键字：分类、空间向量模型、F-Score 评价法、关键词

Abstract

In recent years, with the development of science and technology, the Internet has become increasingly popular in people's lives. It has increasingly become a trend for Chinese citizens to exercise their right to know, to participate in, to express and to supervise their political affairs online as Internet users, and the online political platform has gradually become an important means for the government to work for the people. In a large number of problem feedback, the use of more efficient classification methods is becoming more and more important.

Therefore, classification counting should be optimized with the development of science and technology. In order to solve the problem of low efficiency of manual classification, We need to ask the background program of the political platform to process a large amount of text data through mathematical model, so as to realize the classification of message categories. At this time, the intelligent government system will be a new working trend, we choose a complete and reasonable classification system, in order to greatly improve the political platform processing and text mining methods, for the analysis of different issues.

For the first problem, we first conduct data mining for annex 1, and classify it on the basis of making full use of its data. Then for attachment 2, we first use the word segmentation of Chinese lexical analysis system ictclas function and filter function of excel for our CSV data preprocessing, using mathematical statistics formulas and vector space model to get the key sequence of each category, so as to complete the text sifter components, and to back to back to the text data measuring, computing precision and recall rate, thus by F - Score for evaluation of this classification method.

For the second question, we mainly use word segmentation system and excel to CSV data preprocessing, use mathematical formula to get reasonable keywords and text processing, get sequence, to calculate similarity calculation using space vector model, which after several types of the hot topics on the similarity analysis, we concluded that can filter out most complete, most accurate hot issues the standard of similarity, the similarity criteria on this event number as heat evaluation project, called heat evaluation in the top five of the hot issues. Then substitute back to the text data to get the list of hot topic comments that the topic asked us to make.

Against the third question, we first left a message text data preprocessing, characteristic was calculated by the mathematical formula reasonably come to the key words, use of the space vector model that all aspects of the category, leave a message in the corresponding category in the category of the response analysis of the characteristics of the particle size, will the class message answer based on keyword text characteristic analysis, combined with the text and answer not repeat number of characters, the target response text from relevance, integrity and quality evaluation can be interpreted.

Keywords: classification, spatial vector model, F-Score evaluation method, keywords

目　录

图　录

表　录

第一章 问题描述

1.1 问题描述

随着网络的日益普及，互联网在中国民众的政治、经济和社会生活中扮演着日益重要的角色，中国公民以网民的身份通过互联网行使知情权、参与权、表达权和监督权，这就是网络问政。而微信、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，这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。

在各大网络问政平台，其后台会每日接受来自民众的留言、热线等，而管理人员则会对其进行汇总，而我们的任务，就是对采集到的数据进行更为准确、全面地所属类别的划分，并建立定义合理的热度评价指标旨在挖掘出某个特定时段的热点问题，并基于此基础上得出题目要求的热点问题留言明细表。除此之外，还需对相关部门做出答复意见的质量给予一套完整、合理的评价体系。

首先，我们需要从附件二所给的数据中，应用自然语言处理技术对各类别的信息内容进行高频词的筛查，经过分析比较后的结果作为留言内容的一级标签分类模型，并根据模型计算得出参考量采用 F-Score 的方法对我们模型进行评价。然后对附件三中的问题进行统计筛查出热点问题并对其所对应留言进行分析，归纳总结后建立定理合理的热点评价指标。最后根据附件四中相关部门对留言的答复意见，对答复进行特征性分析，得出可从的相关性、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。

1.2 论文的结构安排

本文共分为四章，各章节的内容安排如下所示：

第一章，对本论文所需要解决的具体问题进行归纳描述，并完整介绍整篇论文的结构。

第二章，合理地对留言内容的文本数据进行分析，先对文件进行预处理，排除干扰词汇，提取具有价值的关键字词，加工处理形成向量，引用空间向量模型（VSM）完成文本分类器，再进行数据回测，得到此方法的准确率与查全率，最后通过 F-Score 方法对模型进行评价。

第三章，对 csv 数据进行预处理，得到合理的关键词，对其进行文本处理，得到可计算的序列，利用空间向量模型进行相似度计算，由此经过几类出现次数多的问题进行的相似度分析，得出能够筛选出最全、最准的热门问题的标准相似度，之后再代回文本数据，筛选出相似度超过此标准的问题，对留言问题进行归类，得到题目所要求我们做出的热点问题留言明细表。

第四章，对留言文本数据进行预处理，得出合理的关键词，利用空间向量模型得

出各个的问题类别，再对各类别中的留言所对应的回复进行问题粒度的特征分析，将该类留言问题的回答基于关键词进行文本特征分析，再结合文本长度与回答的非重复字符数进行比较，对目标回复文本从相关性、完整性与可解释型进行质量评测。

第二章　数据的探索分析

2.1 标签数据预处理

2.1.1　标签数据挖掘

关于群众留言的问题，工作人员处理网络问政平台的群众留言中规定了一定的划分体系，并对其进行分类，这三类都有不同的分支，针对所提供的分类三级标签体系，我们进行了分析，如下表。

表 1　处理网络问政平台群众留言问题的三级分类表

	处理网络问政平台群众留言问题的三级分类
一级分类	城乡建设、党务政务、国土资源、环境保护等
二级分类	安全生产、城市建设和市政管理、城乡规划、村镇建设等
三级分类	事故处理、安全生产管理、安全隐患、园林绿化环卫等

2.1.2　一级分类表的分析

对于处理网络平台群众留言问题表的分析，不同的分类所对应的不同问题数量也有所不同，人们对于社会的需求形成了表格统计的不同问题的分类。不同的分类的格局也有缩不同，对于表格进行分析，我们可以看出三级分类的分支比一级分类和二级分类的多，我们针对一级分类表进行详细分析，如表：

表 2　一级分类表

名称	城乡建设	农村农业	政法	教育文体	劳动和社会保障
计数	65	56	55	45	42
名称	民政	党务政务	经济管理	卫生计生	商贸旅游
计数	38	32	31	27	25
名称	环境保护	交通运输	纪检监察	国土资源	科技与信息产业
计数	24	22	21	19	15

对于一级分类表进行分析，我们可以看出，在一级分类表中，城乡建问题占比最大，统计数为 65，占比 12.5%，科技与信息产业问题占比最小，统计数占比 2.9%。通过对于一级分类的分析，我们可以看出一级分类分支问题的归类数量不多，进行饼状图统计，我们可以看出其不同问题的分布情况，如图。

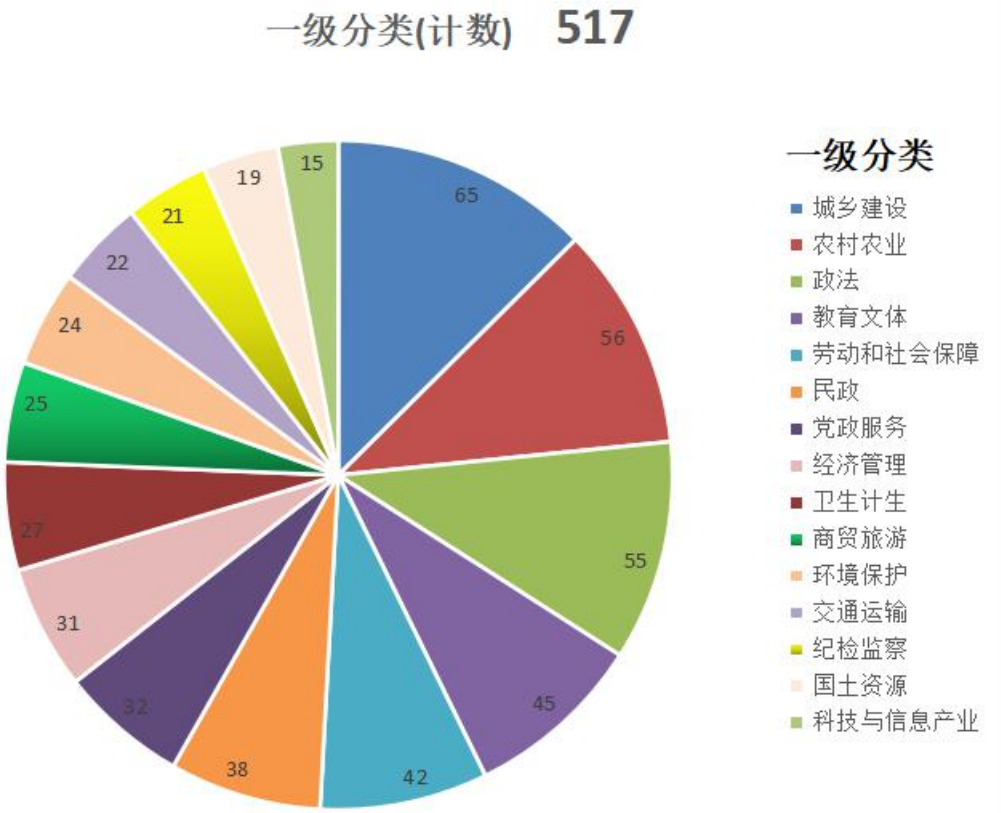

图 1　一级分类饼状图

2.1.3　二级分类表的分析

对于二级分类数据的分析，得到了下表的数据。

表 3　二级分类表（出现次数在 6 次及以上）

名称	城市建设和市政管理	住房保障与房地产	环境污染	社会治安	教育行政管理
计数	16	16	11	10	9
名称	市场监管	环保管理	医政监管	安全生产	国有土地上房屋征收与补偿
计数	9	8	8	7	7
名称	劳动关系	林业管理	人口计生	水利水电	城乡规划
计数	7	7	7	7	6
名称	城镇职工社会保险	工资福利	建设管理	金融财税	劳动保护
计数	6	6	6	6	6
名称	商业贸易	贪污贿赂	退休政策及待遇	文化	宣传舆论
计数	6	6	6	6	6
名称	优抚	质检检验检疫	仲裁与调解		
计数	6	6	6		

对于二级分类表进行分析，二级分类表的问题归类数量远大于一级分类表的问题数量，通过对于数量的统计，我们做出了问题出现次数在 6 次以上的表格并加以分析。出现次数最多的是城市建设和市政管理问题以及住房保障与房地产问题，出现次数均为 16，占比 3%，出现次数最少的是交通运输的问题，出现次数仅为 1，占比为 0.1%，进行饼状图的统计，我们同样可以看出其不同问题的分布情况，如图。

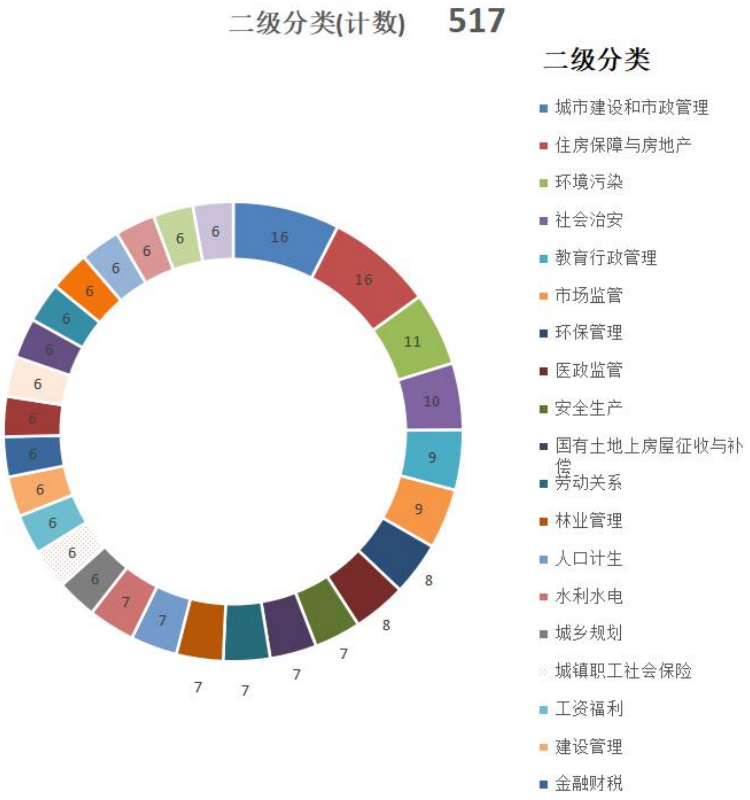

图 2　二级分类饼状图

2.1.4 三级分类表的分析

三级分类数据的所涉及的类别数量较为庞大，通过对其分析，可以看出每个问题出现的频率不相上下，除了其它问题，出现次数最多为 3 次，出现次数最少为 1 次，通过对出现次数为 2 及 2 次以上的数据进行分析，同样得到了下表的数据：

表 4　三级分类表（出现次数在 2 次及以上）

名称	安置补偿	安全生产管理	安全隐患	房屋拆迁
计数	3	2	2	2
名称	规划设计	回迁房	棚户区和城中村改造	审批手续
计数	2	2	2	2
名称	事故处理	异地升学	住房公积金	
计数	2	2	2	

通过对于三级分类表的分析，出现次数最多的问题是安置补偿，次数为 3 次，占比 0.5%，大部分问题出现次数为 1 次，进行饼状图统计，我们得到了下图：

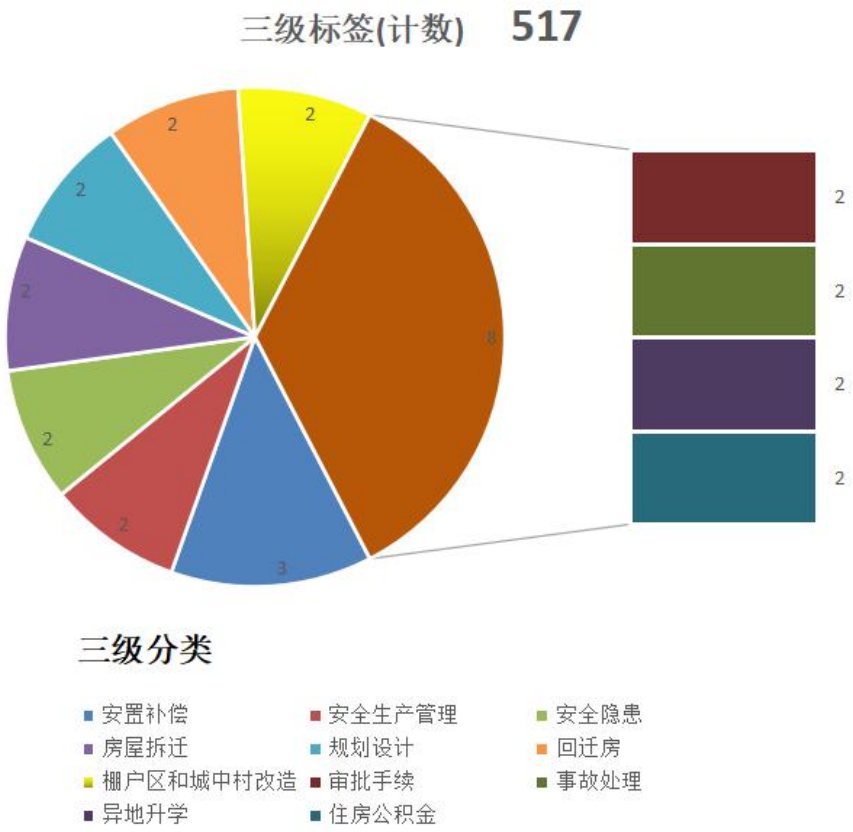

图 3　三级分类饼状图

2.2 留言一级分类模型的建立

2.2.1 挖掘目标

由于各大网络问政平台的后台每日都会接受大量来自民众的留言、热线等可以反映社情民意相关的文本内容，且数据量的不断攀升使得以往主要依靠人工来进行留言划分的工作带来了极大挑战。而本节旨在挖掘能够实现对各大网络问政平台所接收到民众的留言信息内容按一级类别标签进行精确、快速分类的模型，使其将初始的 csv 文档进行精简化、特征化处理，再根据数学算法及数学模型，构建一个完整的文本筛分体系，实现文本筛分器的组建，最终达到高精度、高效率处理留言信息内容一级分类的效果。

2.2.2 模型构思

对于显得让人力不从心的传统文本分类，我们构建了一个挖掘模型该模型可根据特征词在类别中出现的文档频率对特征词初步排序。我们希望利用逐渐成熟的搜索引擎，能在给定的分类体系下，自动地根据文本的内容确定文本关联的类别。对于模型的研究，我们引入了向量空间模型（VSM）来表示文档，根据特征词在类别当中所出现的文档频率，我们将特征词进行了排序，并对其进行加工处理。

我们之所以选择基于文档频率的方法来完成特征选择，是因为它不同于传统的文档频率方法，传统的文档频率方法通常容易使得多种误差产生的概率加大。我们通过使用向量空间模型，标注表格中文本所出现的特征词，对于特征词的处理，我们也运用了巧妙的方法。中文文档和英文文档不同，词语和词语之间没有建个符号，而且其中包含了大量伴随文本产生的标间符号或者其它字符。

因此，我们对表格中的文本进行了预处理，我们对每篇文章做了分词的处理，我们将文本的句子按照词性的不同进行了划分，并且将多余的符号进行删除，因为在数据挖掘的过程中，出现了大量的与分类无关的特征向量，所以我们需要将其删除，筛选挖掘出对类别划分贡献较大的关键词，从而而能够在这些特征词中进一步挑选出对类别贡献较大的特征词所构成的特征向量。

通过统计贡献较大的特征词在一个类别中出现的数量，我们可以得到这个特征词与该类别的相关度，从而保留出现频率较大的特征词。根据特征词，我们会在表格文本中对其完成分类，通过补充高频词，也可以使得文本的分类性能被大大提高。

最后，我们会进行系统的评估，通过计算准确率，判断文本中与人工分类结果吻合的文本所占的比率，通过计算查全率，计算出人工分类结果应有的文本中分类系统系统温和的文本所占的文本所占的比率，从而评价本模型筛分的精准度。

2.2.3 留言数据的预处理

对于留言数据的预处理，我们运用了我们的模型，建立了关于留言内容的一级标签分类模型。

我们首先准备了表格的数据，并将它整合成一篇文档。我们使用向量空间模型（VSM）来表示文档。我们将文档中出现的特征词用$T\text{erm}_1$，$T\text{erm}_2$……$T\text{erm}_n$进行表示，并将其表示出来。

$$Doc \to (Term_{1,} Term_2, Term_3 \cdots\cdots, Term_n)$$

我们对于文档进行处理，将中文文档与英文文档的分词方法区分开来，建立一个处理中文文档的方法。基于中文文档的词语与词语之间并没有间隔符号，我们使用了 ICTCLAS 软件将文档进行了初步分词，使得每一段连续的中文文档被分割成为特征词列表，形成一个特征词体系。

对于文档的细分中，我们发现我们得到的特征词列表中存在不同的标点以及各式各样的非中文文本的符号，于是我们将这些符号予以删除，只保留文档中原有的中文文本，使其组成一个由特征向量组成的文档。

我们为每一个特征向量标明了它的词性，并且从我们得到得特征词中，我们使用软件选择性地删减了对分类几乎无贡献地词性，因为这些特征词并不能很好地反映类别信息，甚至可能会误导我们得出的分类结果，使得我们分类结果的精确性得不到保障。例如量词、数词、感叹词等单词。留下的特征词，我们用他们作为表示文档特性的贡献性指标，用其来表示文档的特征向量。

选择出我们所需要的特征向量后，我们将特征向量进行了统计，我们发现特征向量在文本中可能出现多次。出现次数多的特征词比出现次数少的特征词更能够体现文档所表达的主题，于是我们选择了出现匹女极高的特征向量，并计算这些特征向量所出现的频率。

$$TF_i^{'} = \frac{\lg(TF_i + 1)}{\max\limits_{J \le j \le n_t} \lg(TF_j + 1)}$$

在我们选择的这个公式中，$1 \le i \le n_t, n_t$作为特征词出现的数量，TF_i为第 i 个特征词在这篇文章出现的次数。我们因此借鉴了模糊特征的思想，将每个特征值的TF^3值进行二次化。举个例子，比如说我们选择的一个特征词在一篇文档中出现，我们便将它的TF^3值设为 0，如果一个特征词在文档中出现频率为高频，我们便将它的TF^3值设为 1。

$$TF_i^{'} = \begin{cases} 0 & 0 < TF_i^{'} \le 1 \\ 1 & T_{\max} < TF_i^{'} \le 1 \end{cases}$$

在该式子中，我们设定了一个阈值$T_{\max}$，如果一个特征词的TF^3值大于该阈值，我们就认为该特征词以高频出现。[i]

通过对留言数据的预处理，我们将文档中不同分类的高频词进行了计数，并将它确定下来，加以分析。

2.2.4 高频词的提取

通过对表格中的一级分类表前进行分析，我们可以看出表格提供了其中一级分类表的类别，其中包含城市建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生这几个方面，通过分析不同类别，我们将使用分类模型对这几个类别进行特征词的提取，以下是这几个类别问题反馈出现的次数及频率。

表 5　一级标签分类表

一级标签	问题计数	频率
城乡建设	2009	0.218
环境保护	1969	0.213
交通运输	1589	0.172
教育文体	1215	0.131
劳动和社会保障	968	0.101
商贸旅游	877	0.95
卫生计生	613	0.66

通过对一级标签的类别初步划分，我们计算出了不同类别所占的频率，并用饼状图将其表示出来。

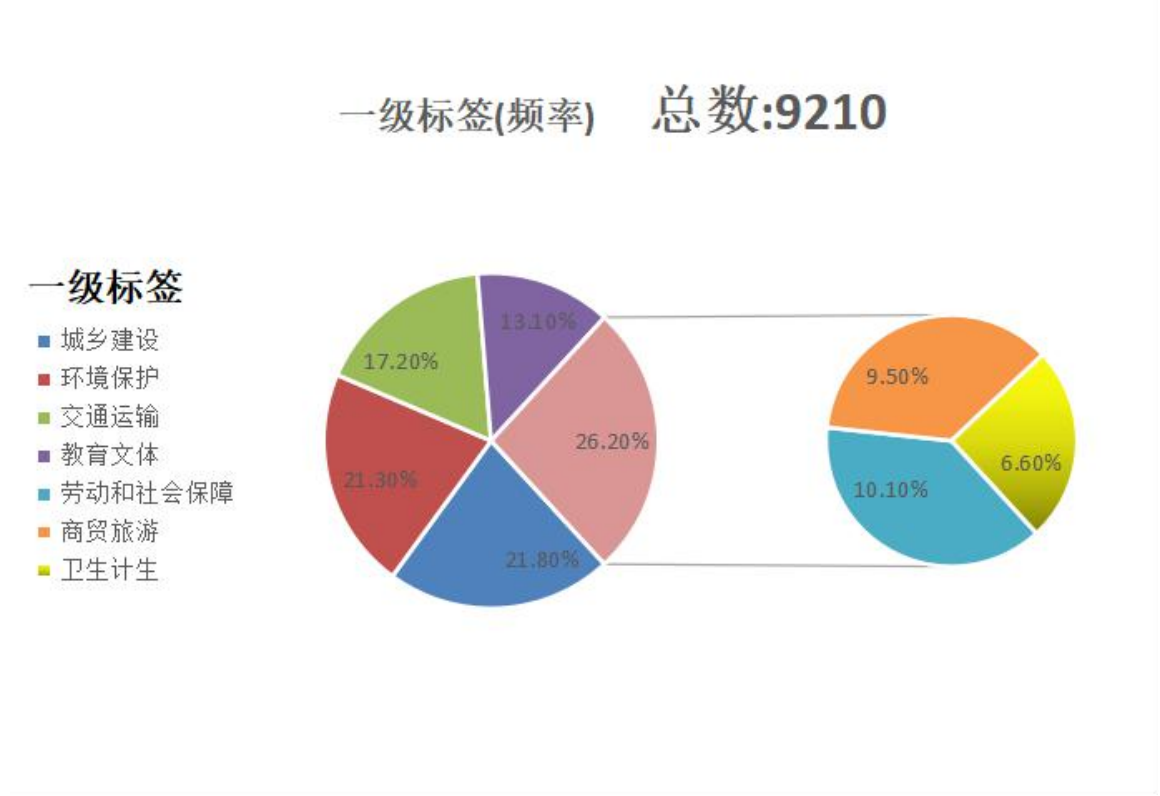

图 4　一级标签饼状图

2.2.5 异常数据的剔除

在数据预处理中，我们筛除了对分类基本上是没有贡献的词例如数词、量词、感叹词等，最后筛查出一系列出现频率较高且有可能对文本分类产生价值的词汇。但在最后依照类别进行具有针对性的筛查，在所得到的结果中我们发现，各大类别所对应留言内容中出现频率排名较前的词汇有一部分是高度重叠的。

从下图我们可以观察到，一级标签每个类别的留言内容出现频率排名前三的词汇，除了环境保护出现频率排名第三的词汇是“污染”，其余的都是“市”、“县”、“的”按不同顺序的排列，虽然它们出现的次数都远超后面的词汇，但显而易见，这样没有特征性、指向性的词汇，无法加入我们每个类别留言内容的特征词序列中，以免对筛分结果造成影响。同理，“区”这个词汇在城乡建设的留言内容中出现频率位居该标签词汇中的第四，次数但其在商贸旅游中同样位居第四，出现次数也高达 168 次，而在环境保护、交通运输中也位居第五，分别高达 188 次与 174 次。

因此，对于这些出现频率虽然高，占某类别词汇比例相对较大但不能很好反映类别信息，甚至会误导分类结果的词汇我们不予以对其进行作为特征词序列的选择。但出于对样本完整性角度的考量，在计算有效高频词的出现频率时，我们仍然将这部分词汇出现的次数，纳入词汇总次数的计算。

表 6　各一级标签类别出现次数排名前十的词汇及它们出现次数表

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	
词汇	计数	词汇	计数	词汇	计数	词汇	计数	词汇	计数	词汇	计数	词汇	计数
市	780	县	397	县	273	市	643	的	786	市	502	的	299
县	749	市	300	市	232	的	576	市	655	县	418	市	225
的	608	污染	254	的	179	县	559	县	464	的	416	县	225
区	371	的	244	出租车	111	教师	275	问题	280	区	168	医院	133
问题	228	区	188	收费	72	区	174	职工	276	电梯	160	省	96
不	192	严重	168	不	67	问题	164	省	263	传销	158	西地	87
小区	172	环境	109	交通	64	不	138	西地	240	省	100	二	85

2.3 一级分类模型的数据分析

2.3.1 一级标签——城乡建设的分析

从下面的城乡建设高频关键词计数表中我们可以清晰直观地看到，“小区”在城乡建设这一类别的留言主题中出现的次数高达 172 次，频率为 0.0085。而“改造”次之，在本类别的留言主题中出现 103 次，频率为 0.051，即使是图表中出现次数相对是最低的“建筑”，其出现次数也达到了 71 次，频率超过 1/500。除此之外，从语义角度而言，以下词汇都具有很强的指向性，例如“工程”、“建设”等从表达上的无法避免会用到更甚至于光看语意上即可秒选的词汇，这有助于文本分类更精确、更高效率地完成。

表 7 一级标签——城乡建设高频关键词计数表

关键词	计数	频率
小区	172	0.0085
改造	103	0.0051
工程	95	0.0047
路	84	0.0041
公积金	83	0.0041
房产证	82	0.0040
规划	80	0.0039
建设	80	0.0039
质量	72	0.0035
建筑	71	0.0035

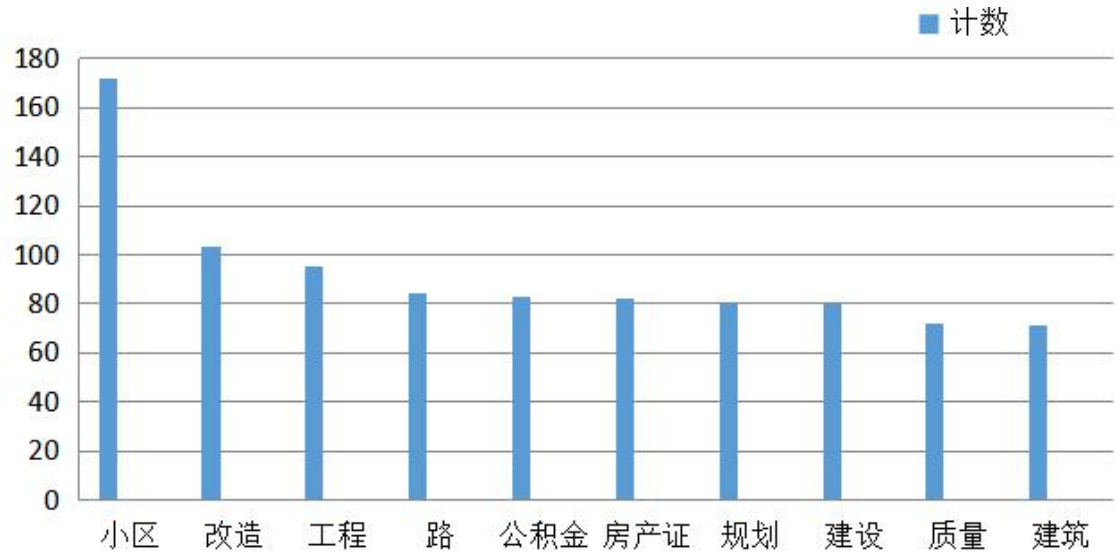

图 5 城乡建设高频关键词统计图

2.3.2 一级标签——环境保护的分析

从下面的环境保护高频关键词计数表中我们可以直接明了地看到，“污染”在环境保护这一类别的留言主题中出现的次数高达 254 次，频率为 0.026，超过 1/50。而“环境”，在该本类别的留言主题中也出现 109 次，频率为 0.011。而“养猪场”出现的次数相对来说最少，达到了 39 次，频率也超过了 1/5。从语义角度而言，污染等词汇使用的广泛性与会比养猪场这样的特有名词来得大，这也与我们的统计结果相符合，说明在一定指向性范围内，日常使用率高的词汇更有利于我们对文本的类别筛分。

表 8 一级标签——环境保护高频关键词计数表

关键词	计数	频率
污染	254	0.026
环境	109	0.011
厂	85	0.0086
排放	75	0.0076
噪音	74	0.0075
扰民	52	0.0052
污水	49	0.0049
养猪场	39	0.0039
非法	36	0.0036
垃圾	29	0.0029

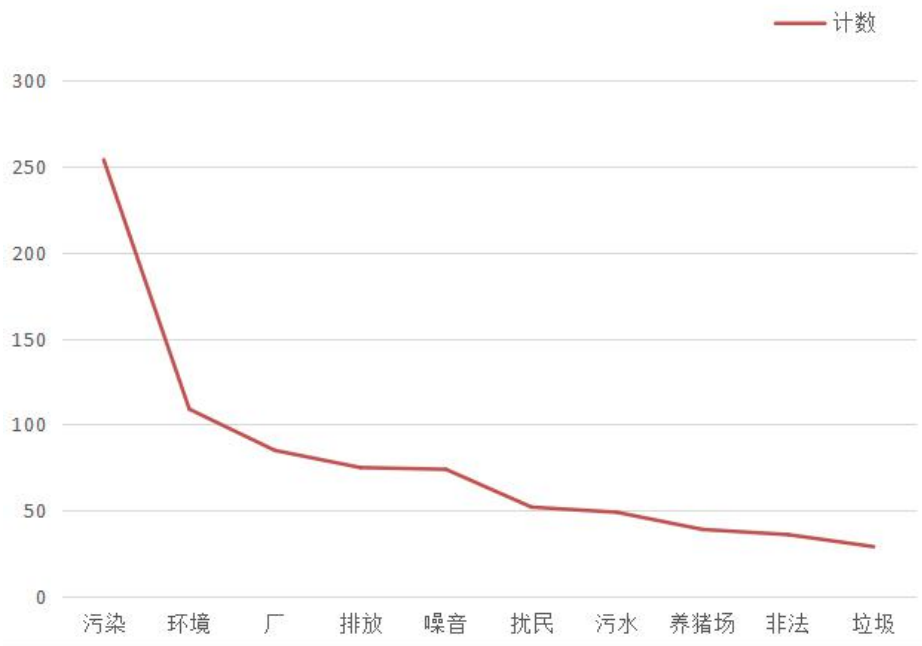

图 6 环境保护高频关键词计数图

2.3.3 一级标签——交通运输的分析

从下面的交通运输高频关键词计数表中我们可以直观地看到，这样人工专有制造物造成的指向极为明确的词汇，在交通运输这一类别的留言主题中出现的次数高达 166 次，频率为 0.029，接近 1/30。而“路面”、“道路”、“公路”这样与“出租车”、“的士”有异曲同工的词汇，在该本类别的留言主题中也出现 108 次，频率为 0.019，接近 1/50。而“物流”这样与交通运输没有第一联系的词汇其出现的次数最少为 22 次，频率为 0.0039。说明在交通分类运输这一标签的关键词选择上要更倾向于与交通、运输达成第一联系的词汇。

表 9　一级标签——交通运输高频关键词计数表

关键词	计数	频率
出租车/的士	166	0.029
路面/道路/公路/	108	0.019
路收费	72	0.013
交通	64	0.011
快递	60	0.011
客运	44	0.0078
邮政	43	0.0076
司机	27	0.0048
物流	22	0.0039

图 7　交通运输高频关键词计数图

2.3.4 一级标签——教育文体的分析

从以下表格可以清晰直观地看出，关键词“学校/小学/中学/一中”在该类别留言内容中出现次数高达 386 次，频率为 0.025，而关键词“教师/老师”以及“教育”次之，在该类别留言内容中出现 370 次，频率为 0.024。通过对于表格的分析，我们可以清晰地看出与教育文体有关的几个高频词，这些词语都是教育文体类别的代表性指标。

表 10 一级标签——教育文体高频关键词计数表

关键词	次数	频率
学校/小学/中学/一中	386	0.025
教师/老师	370	0.024
教育	135	0.0087
学生	127	0.0082
补课	121	0.0078
教育局	108	0.0070
违规	66	0.0042
招生	64	0.0041
民办	43	0.0027
文化	41	0.0026

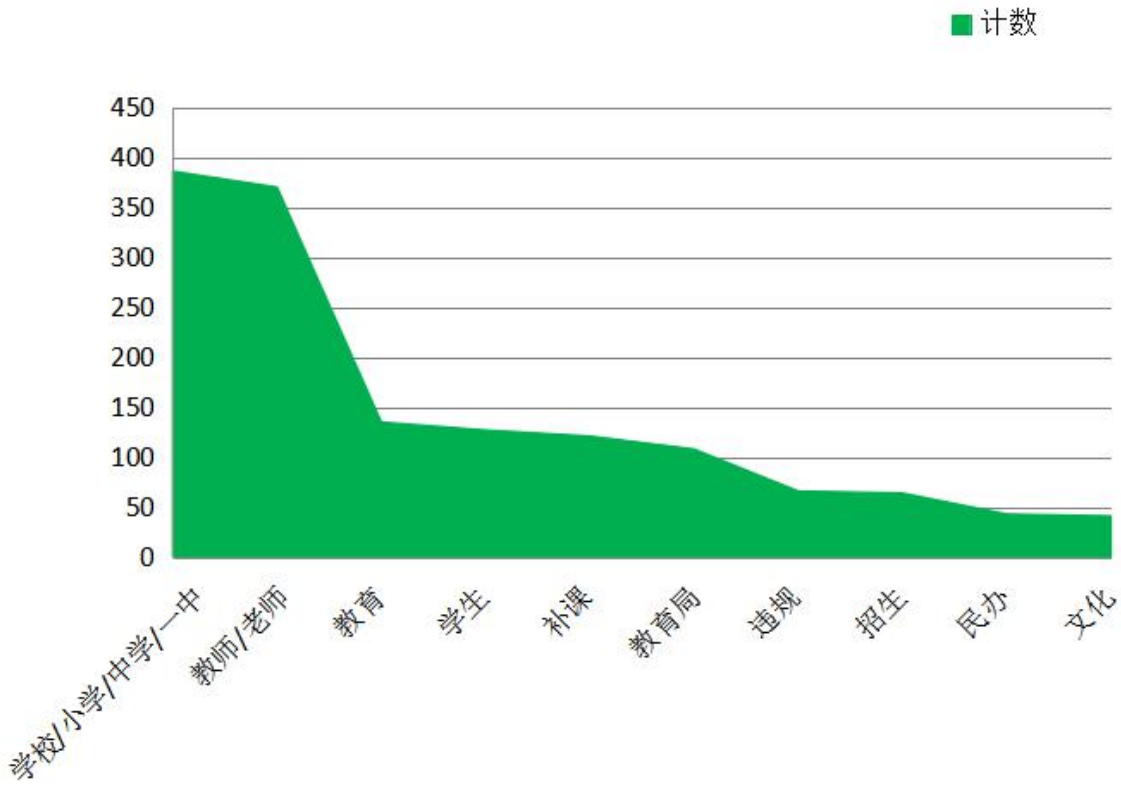

图 8 教育文体高频关键词计数图

2.3.5 一级标签——劳动和社会保障的分析

从以下表格可以清晰直观地看出，关键词“职工/员工”在该类别留言内容中出现次数高达 419 次，频率为 0.020，而关键词“社保/医保”次之，在该类别留言内容中出现 320 次，频率为 0.016。通过对于表格的观察，劳动和社会保障无疑是公司职工的重点关注对象，无论是在职员工，亦或是退休员工，对于工资保障以及社会的福利待遇问题都是能够体现劳动和社会保障运输的关键性问题。

表 11 一级标签——劳动和社会保障高频关键词计数表

关键词	次数	频率
职工/员工	419	0.020
社保/医保	320	0.016
工资	216	0.010
退休	209	0.010
人员	187	0.0091
保险	152	0.0074
工作	110	0.0053
养老	108	0.0052
单位	105	0.0051
政策	84	0.0040
待遇	80	0.0039
医疗	70	0.0034

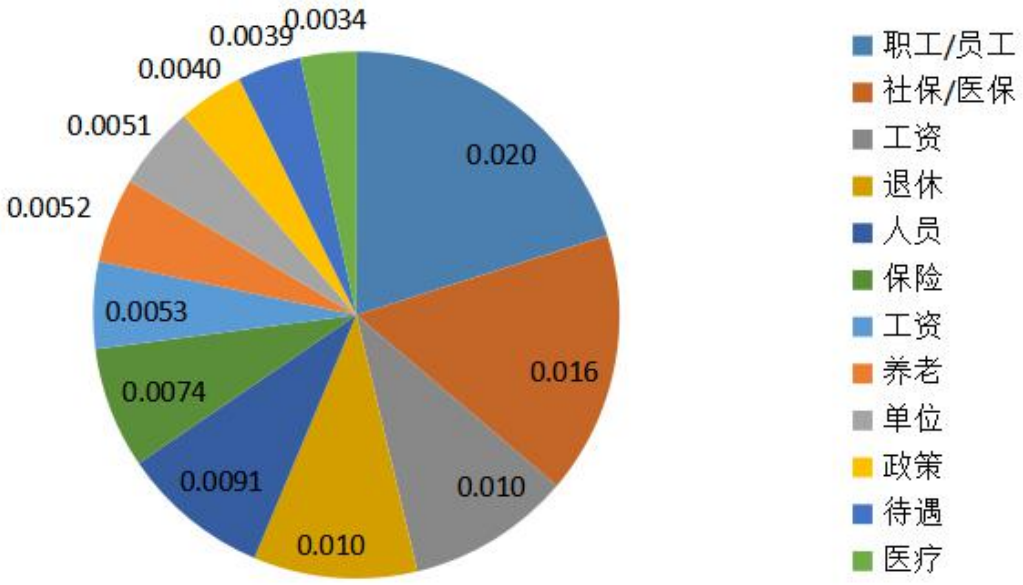

图 9 劳动和社会保障高频关键词饼状图

2.3.6 一级标签——商贸旅游的分析

从以下表格可以清晰直观地看出，关键词“电梯”在该类别留言内容中出现次数高达 168 次，频率为 0.011，而关键词“电梯”次之，在该类别留言内容中出现 158 次，其频率与出现次数最多的关键词不相上下，为 0.011.通过分析该表格，在商贸旅游的反馈问题中，电梯的建设是我们所需要关注的，传销的力度需要加大力量进行把控，小区问题以及收费问题也屡屡存在，解决好商贸旅游的问题，需要特别注意表中不同现象的发生。

表 12 一级标签——商贸旅游高频关键词计数表

名称	计数	频率
电梯	160	0.011
传销	158	0.011
收费	100	0.0069
垄断	94	0.0065
小区	72	0.0050
乱	69	0.0048
市场	53	0.0037
投诉	51	0.0035
价格	49	0.0034
旅游	49	0.0034
质量	44	0.0031
景区	43	0.0030

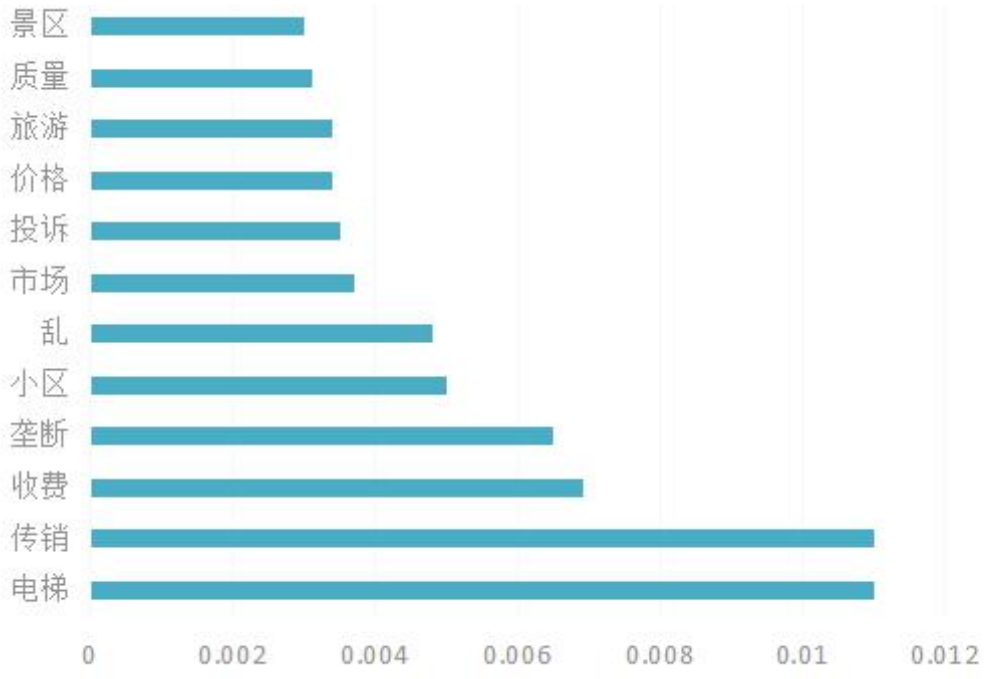

图 10 商贸旅游高频关键词频率图

2.3.7 一级标签——卫生计生的分析

从以下表格可以清晰直观地看出，关键词“医院/胎/人民医院”在该类别留言内容中出现次数高达 200 次，频率为 0.024，而关键词“政策”次之，在该类别留言内容中出现 58 次，其频率为 0.0063。在表格中，与医院有关的词都与卫生计生的高频词有很大的关系，可见得，医院出现的问题问题在卫生计生中占了很大的比例。

表 13 一级标签——卫生计生高频关键词计数表

名称	计数	频率
医院/胎/人民医院	220	0.024
政策	58	0.0063
医疗	53	0.0058
生育	50	0.0055
医生	45	0.0050
独生子女	39	0.0043
卫生	35	0.0039
人员	32	0.0035
超生	27	0.0030
计划生育	27	0.0030

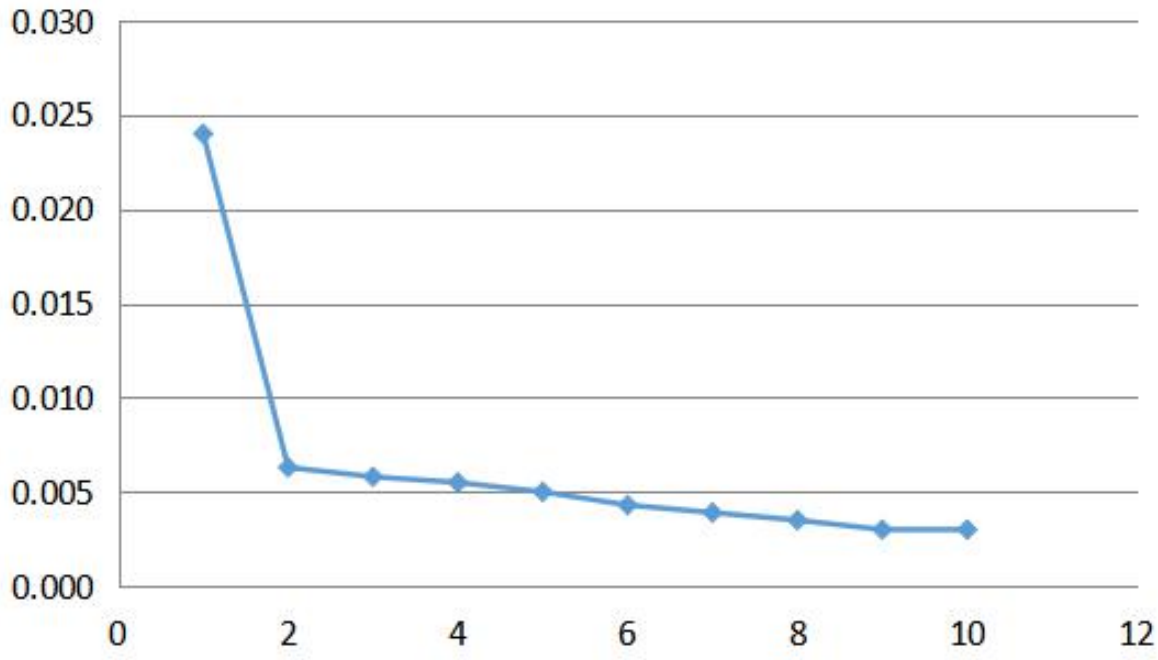

图 11 卫生计生高频关键词频率图

2.4 分类模型的评价

对于文本分类的评估是十分重要的，评估文本分类系统的标志性指标是分类的精准度，假设标志性分类完全正确，并且我们予以排除个人思维差异的因素，我们提供的分类方法与标志性分类越接近，分类的精准度就会越高。

查准率和查全率是两个很重要的指标，这两个指标表明了分类质量的两个不同的方面，这两者必须一同纳入考虑，查全率和查准率之间具有紧密的不可拆分的关系，在极端情况下，将文档集合中的所有文档返回为结果集合的系统有 100 % 的查全率，但是查准率可能会变得很低。而另一方面，一个系统如果只能返回唯一的文档，将会有很低的查全率，但是它的准确率却可能达到 100%却可能达到 100 %。通常，总是假定查全率为一个适当的值，然后按查准率的高低来衡量系统的有效性，所以我们将这两个指标作为我们对文本评估的必经之路。同时，我们也介入了 F-Score 方法对我们的分类方法进行评估，以确定我们分类方法的精准度。

假设人工分类结果为：

1. 城乡建设类：小区、改造、工程、路、公积金、房产证、规划、建设、文化

2. 环境保护类：污染、环境、厂、排放、噪音、扰民、污水、养猪场、乱

3. 交通运输类：出租车/的士、路面/道路/公路、路收费、交通、快递、客运、违规

4. 教育文体类：学校/小学/中学/一中、教师/老师、教育、学生、补课、教育局、违规、招生、待遇

5. 劳动和社会保障类：职工/员工、社保/医保、工资、退休、人员、保险、工作、养老、单位、公积金

6. 商贸旅游类：电梯、传销、收费、垄断、小区、乱、市场、投诉、价格、客运

7. 卫生计生类：医院/胎/人民医院、政策、医疗、生育、医生、独生子女、卫生、投诉

2.4.1 模型查准率的测定：

计算模型的查准率可以评价我们的分类方法的正确程度，准确率是所有判断的文本中与人工分类结果吻合的文本所占的比率。

准确率在我们的分类方法中指的是分类正确的文本数除以我们实际分类的文本数，在对于分类过程中，异常数据或者无意义数据的剔除都会导致我查的产生，使用准确率表示指在一定条件下筛选出的特征值数量，与实际文档所提供的特征值数量所占的比，准确率的数学公式如下。

$$P_i = \frac{\text{分类正确的文本数}}{\text{实际分类的文本数}}$$

P_i（第i类的查准率）

则:

评估城乡建设的文本分类系统的准确率

$$P_1=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}=\frac{172+103+95+84+83+82+80+80}{172+103+95+84+83+82+80+80+41}\approx 95.00\%$$

评估环境保护的文本分类系统的准确率

$$P_2=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}=\frac{254+109+85+75+74+52+49+39}{254+109+85+75+74+52+49+39+69}\approx 91.44\%$$

评估交通运输的文本分类系统的准确率

$$P_3=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}=\frac{166+108+72+64+60+44}{166+108+72+64+60+44+66}\approx 88.62\%$$

评估教育文体的文本分类系统的准确率

$$P_4=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}=\frac{386+370+135+127+121+108+66+64}{386+370+135+127+121+108+66+64+80}\approx 94.51\%$$

评估劳动和社会保障的文本分类系统的准确率

$$P_5=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}=\frac{419+320+216+209+187+152+110+108+105}{419+320+216+209+187+152+110+108+105+83}\approx 95.65\%$$

评估卫生计生的文本分类系统的准确率

$$P_7=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}=\frac{220+58+53+50+45+39+35}{220+58+53+50+45+39+35+51}\approx 90.74\%$$

2.4.2 模型查全率的测定

同理，计算模型的查全率是表示人工分类结果应有的文本中分类系统吻合文本所占的比率。

查全率是衡量某一项系统从所研究的文献的集合中所检索出相关文献成功度的一项指标，即检测出的相关文献量和检索系统中相关文献总量的百分比，查全率的公式如下。

$$R_i=\frac{\text{分类的正确文本数}}{\text{应有文本数}}$$

R_i（第i类的查全率）

则：

评估城乡建设的文本分类系统的查全率：

$$R_1=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{172+103+95+84+83+82+80+80}{172+103+95+84+83+82+80+80+72+71}\approx 84.49\%$$

评估环境保护的文本分类系统的查全率：

$$R_2=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{254+109+85+75+74+52+49+39}{254+109+85+75+74+52+49+39+36+29}\approx 91.90\%$$

评估交通运输的文本分类系统的查全率：

$$R_3=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{166+108+72+64+60+44}{166+108+72+64+60+44+43+27+22}\approx 84.82\%$$

评估教育文体的文本分类系统的查全率：

$$R_4=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{386+370+135+127+121+108+66+64}{386+370+135+127+121+108+66+64+43+41}\approx 94.25\%$$

评估劳动和社会保障的文本分类系统的查全率：

$$R_5=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{419+320+216+209+187+152+110+108+105}{419+320+216+209+187+152+110+108+105+84+80+70}\approx 88.64\%$$

评估商贸旅游的文本分类系统的查全率：

$$R_6=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{160+158+100+94+72+69+53+51+49}{160+158+100+94+72+69+53+51+49+49+44+43}\approx 85.56\%$$

评估卫生计生的文本分类系统的查全率：

$$R_7=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{220+58+53+50+45+39+35}{220+58+53+50+45+39+35+32+27+27}\approx 85.32\%$$

2.4.3 F-Score 综合测评

通过计算出查全率和查准率，我们可以使用 F-Score 对分类方法进行评价。从以上计算来看，查全率和查准率较高，均符合我们的的要求，在保证查全率的

情况下，我们希望提高查准率，反之，在保证查准率的情况下，我们希望提高的是查全率。在两者都要求很高的情况下，我们可以使用 F-Score 进行综合测评，来更加具有说服力地评价我们的分类方法。

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

P_i（第i类的查准率）R_i（第i类的查全率）

则：

故总体文本分类系统的准确率

$$\overline{P} = \frac{\sum_{i=1}^{7} P_i}{7} = \frac{95.00\% + 91.44\% + 88.62\% + 94.51\% + 95.65\% + 94.82\% + 90.74\%}{7} \approx 92.97\%$$

总体文本分类系统的查全率

$$\overline{R} = \frac{\sum_{i=1}^{7} R_i}{7} = \frac{84.49\% + 91.90\% + 84.82\% + 94.25\% + 88.64\% + 85.56\% + 85.32\%}{7} \approx 87.85\%$$

$$F_1\text{测试值} = \frac{\text{准确率} \bullet \text{查全率} \bullet 2}{\text{准确率} + \text{查全率}} = \frac{92.97\% \bullet 87.85\% \bullet 2}{92.97\% + 87.85\%} \approx 90.34\%$$

第三章　热度问题的挖掘

3.1 留言数据的预处理

3.1.1　留言数据的分析

基于第一章所引用的空间向量模型进行对数据的筛分，在热点问题的挖掘中，我们仍然采用该模型对其进行分类。

附件 3 所表示出的问题为 2019 年全年以及 2020 年年初的群众反应的问题，我们对这些问题进行了预处理，以便为我们在热点问题的筛查起到推动作用。

3.1.2 问题数量的统计

我们首先对留言数据进行了预处理，通过对某一时段内群众集中反应的某一问题，我们可以对其进行筛查，并选出热点问题。通过对不同月份的研究，我们发现了不同月份中群众所反应的问题量有所波动。于是，我们将不同月份作为横轴，问题的数量作为 y 轴，建立了平面直角坐标系，并将问题的数量以折线图的方式表现出来，如下图。

图 12　不同月份的问题数量变化折线图

从我们绘制的折线图，我们可以看到不同月份变化的规律。在 2019 月份，我们可以看出，一月份到二月份的问题数量变化急剧下降，二月份到三月份又有所回升，三月份与一月份的问题数量值不相上下。之后的两个月无大波动，到了六月和七月，问题的变化先从下降到上升。九月到十月份，问题的数量发生了显著的下降，之后又在十二月份回升，由于 2020 年的一月份统计的天数不够多，所以不纳入比较的范围内。

通过图表可以看出，除去十二月份，七月份的问题数量最多，高达 417 个群众反映问题，二月份的数量最少，为 256 个。

通过计算不同月份问题的平均数，我们可以看出每个月平均可以收到 354 条问题，如下：

$$\overline{X}=\frac{X_1+X_2++X_3+X_4+X_5+X_6+X_7+X_8+X_9+X_{10}+X_{11}+X_{12}}{12}$$

解得：

$$\overline{X}=\frac{359+256+368+397+388+339+417+391+387+297+281+366}{12}=354$$

将这些问题的数量统计出来，有助于我们之后热点问题的筛分。

3.2 文本特征的提取

3.2.1 留言问题的处理

与之前的方法相同，我们将留言问题进行了处理，运用了我们的模型，建立了关于留言问题的分类模型。

我们将附件 3 表格的数据进行了处理，并将它整合成一篇文档。对于文档的处理，我们仍然存在许多的干扰因素，我们利用 ICTCLAS 软件对文档进行分词，仍然使用模型的方法，将多余的标点以及非文本的符号剔除，处理直至剩下的词基本为文本特征词。

这篇文档中存在许多的我们使用向量空间模型（VSM）来表示文档。我们将文档中出现的特征词用 $Term_1$，$Term_2$ ……$Term_n$ 进行表示，并将其表示出来。

$$Doc \rightarrow (Term_{1,}Term_2,Term_3 \cdots\cdots,Term_n)$$

3.2.2 关键词的处理

对于热点问题的筛分，我们利用有指向性的特征向量进行问题的选择，首先，我们把具有指向性的关键词标出，并提取出来，如下。而此处的数学计算，我们主要采用 TF-IDF 函数进行目标词汇对该话题的指向性大小的衡量从而进行特征词的选择。

表 14　指向性关键词计数表

关键词	计数
扰民	278
噪音	147
幼儿园	87
拆迁	67
公交车	65
油烟	53
伊景园	40
经济学院	14
温斯顿	9

图 13　指向性关键词计数折线图

3.3　热点问题的发现

3.3.1　热点问题的选择

在构建一级标签分类模型中同样使到该方法,而它在该模型中是将每条留言文本的内容刻画为一维向量的形式,然后再凭借各个向量之间的重合度度量，将目标文本划分到不同的集合中，用向量来表示文本，从而简化了文本中的关键词之间的复杂关系，而文档用十分简单的向量表示，使得模型具备了可计算性。。而对于筛选热点问题，由于热点问题其信息量比单纯的分类别来的大，而空间向量构建方法应用到这种信息量大的话题发现中时算法中往往会引起“高维灾难”问题,即一维向量中引入非文本特有的特征,如噪声,从而影响相似度的准确性。因此我们抓住问题结构中的时间、地点和关键词三个因素，构建三维空间向量模型，使得每个问题的都能在该构建空间的一个向量得到体现，通过向量之间一定的距离，我们可以得出不同的留言问题之间的相关性，从而找到具有到达一定相关性留言问题的集合，即我们的热点问题。

3.3.2　基于向量空间模型聚集的方法

向量空间模型将一条留言问题映射为一个特征向量 V(d)=(t1, ω1(d)，;…; tn, ωn(d))，其中 ti(i=1,2, …,n)为一列互不雷同的留言问题，ωi(d)为 ti 在 d 中的权值，一般被定义为 ti 在 d 中出现频率 tfi(d)的函数，即　。

$$\omega_i(d)=\psi(tf_i(d))$$

在信息检索中常用的词条权值计算方法为 TF-IDF 函数，其中 N 为所有留言问题的数目，ni 为含有词条 ti 的文档数目。TF-IDF 公式有很多变种，下面是一个常　用的 TF-IDF 公式：

$$\psi = tf_i(d) \times \log\left(\frac{N}{n_i}\right)$$

$$\omega_t(d) = \frac{tf_i(d)\log\left(\frac{N}{n_i}+0.1\right)}{\sqrt{\sum_{i=1}^{n}(tf_i(d))^2 \times \log^2\left(\frac{N}{n_i}+0.1\right)}}$$

根据 TF-IDF 公式，各类留言问题集中包含某一关键词的留言问题越多，说明它区分留言问题主题的能力越低，其权值越小；另一方面，某一类留言问题主题中某一关键词出现的频率越高，说明它区分留言问题属性的能力越强，其权值越大。

两留言问题之间的相似度可以用其对应的向量之间的夹角余弦来表示，即文档 di，dj 的相似度可以表示为

$$Sim(d_i, d_j) = \cos\theta = \frac{\sum_{k=1}^{n}\omega_k(d_i)\times\omega_k(d_j)}{\sqrt{\left(\sum_{k=1}^{n}\omega_k^2(d_i)\right)\left(\sum_{k=1}^{n}\omega_k^2(d_j)\right)}}$$

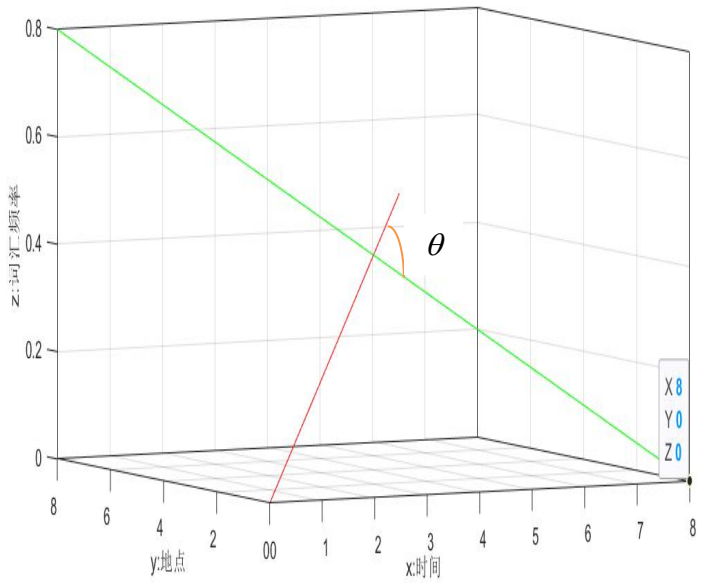

图 14　向量关系图

进行查询的过程中，先将查询条件 Q 进行向量化，主要依据布尔模型：当 ti 在查询条件 Q 中时，将对应的第 i 坐标置为 1，否则置为 0，即

$$q_i = \begin{cases} 1 & t_i \in Q \\ 0 & t_i \notin Q \end{cases}$$

从而文档 d 与查询 Q 的相似度为

$$Sim(Q,d) = \frac{\sum_{i=1}^{n} \omega_i(d) \times q_i}{\sqrt{\left(\sum_{i=1}^{n} \omega_i^2(d)\right)\left(\sum_{i=1}^{n} q_i^2\right)}} \text{ ii}$$

3.3.2 基于热词抽取并聚类的方法

目前提出了许多通过热词抽取与聚类的方式来发现热点话题的算法，这类算法需要我们提前提取出当前热词并对其进行聚类处理。其中热词的提取可以通过提取出所有的词的增量 TFIDF 值来找到热度较高的词。聚类方法可以通过基于密度的聚类策略发现热度比较高的词集合，但是该方法不同的探测次序会严重影响到集合之间边缘点的归属，而对热度比较高的词选择实施层次聚类，消耗时间大，不利于这种大数量级数据的挖掘。[iii]

3.4 文本表示

3.4.1 热点问题的选择

通过对热点问题挖掘方法的探索，我们将我们所选择的具有指向性的关键词进行研究，热点问题指的是在某一段时间内群众集中反映的某一问题可以称为热点问题我们使用基于向量空间模型聚集的方法对热点问题进行挖掘，我们将对热点问题中所具有指向性的关键词进行搜索，并作出热点问题表。

首先，我们将具有指向性的关键词进行分开探索。首先，我们通过将关键词介入附件 3 的表格中，得到了不同关键词所对应的不同问题，这些问题出现在了不同的时间以及不同的范围内。

对于不同的关键词对应的不同问题，我们需要指定约束的因素，寻找出符合题目提供的某一段时间内群众集中反映的某一问题。所以，我们将地点进行约束，对于不同地点，我们选择反应相似的问题构成一个整体，并将它们约束在一个时间范围内。

首先，我们确定了不同关键词所对应的不同频率出现就较高的地点。

表 15　关键词所对应的地点表

关键词	地点
扰民	魅力之都
噪音	魅力之都
幼儿园	A7 县幼儿园
拆迁	西湖街道
公交车	X205 路公交车
油烟	魅力之都
伊景园	伊景园河滨苑
经济学院	A 市经济学院
温斯顿	温斯顿英语机构

通过对于地点的确定，我们得出了不同地点在一定时间范围内所发生的相似性问题。我们建立了空间坐标系，将 X 轴设定为时间轴，将 Y 轴设定为时间发生的次数，将 Z 轴设定为该类时间发生的频率，将文本内容转化为可进行数学计算的数据，则两留言问题之间的相似度依然可以用其对应的向量之间的夹角余弦来表示，即文档 di，dj 的相似度可以表示为

$$Sim(d_i,d_j)=\cos\theta=\frac{\sum_{k=1}^{n}\omega_k(d_i)\times\omega_k(d_j)}{\sqrt{\left(\sum_{k=1}^{n}\omega_k^2(d_i)\right)\left(\sum_{k=1}^{n}\omega_k^2(d_j)\right)}}$$

夹角越小，相似度越大。[iv]因此我们在处理若干留言问题数据后，得到最为合理的相关度，并将它作为我们留言问题的热度评价指标。

3.4.2　热点问题的评价指标

通过空间向量模型中向量之间的余弦计算，我们可以间接性地确定出两个问题的相似度，并指定出了热点问题评价指标。我们将其应用到我们的文件中，我们将之前所标记的标志性问题作为我们的评价指标，并将其它问题与其进行相似度计算，排除相似度较低的数据，如下。

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	
196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019/8/7 19:52:14	申禁止捆	0	0	相关度
279070	A00095080	投诉A市伊景园滨河苑开发商违法捆绑销售无产权车位	2019/8/31 6:33:25	立即制止开	0	0	68.97%
218709	A00010669	A市伊景园滨河苑捆绑销售车位	2019/8/1 22:42:21	房合同，强	0	1	79.07%
289950	A00044759	投诉A市伊景园滨河院捆绑销售车位	2019/7/7 7:28:06	一对一购买	0	0	80.43%
223247	A00044759	投诉A市伊景园滨河苑捆绑销售车位	2019/7/23 17:06:03	。2.希望	0	0	86.96%
283879	A00044759	A市伊景园滨河苑项目捆绑销售车位	2019/7/18 20:27:40	员工，经	0	0	73.91%
246407	A00099597	举报广铁集团在伊景园滨河苑项目非法绑定车位出售	2019-09-01 14:20:22	非法绑定高	0	0	59.13%
239032	A909169	请维护铁路职工权益取消伊景园滨河苑捆绑销售车位的要求	2019-09-01 10:03:10	要求一户一	0	1	56.45%
260254	A909173	投诉A市伊景园滨河苑开发商违法捆绑销售无产权车位	2019-08-30 18:10:23	产市场乱象	0	0	68.97%
236301	A909197	和谐社会背景下的A市伊景园滨河苑车位捆绑销售	2019-08-30 16:32:12	和谐社会么	0	0	65.45%
209571	A909200	伊景园滨河苑项目绑定车位出售是否合法合规	2019-08-28 19:32:11	而单个车	0	0	56.60%
222209	A00017171	A市伊景园滨河苑定向限价商品房项目违规捆绑销售车位	2019-08-28 10:06:03	相要挟，	0	0	57.14%
244243	A909198	关于伊景园滨河苑捆绑销售车位的投诉	2019-08-24 18:23:12	但属于城	0	0	63.92%
276460	A909170	A市伊景园滨河苑捆绑销售车位是否合理？	2019-08-24 17:23:11	没有购买车	0	0	69.31
218739	A909184	A市伊景园·滨河苑欺诈消费者	2019-08-24 00:00:00	车位定金，	0	0	58.82%
190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019-08-23 12:22:00	件，强制	0	0	60.19%
207243	A909175	伊景园滨河苑强行捆绑车位销售给业主	2019-08-23 12:16:03	不买车位就	0	0	74.23%
258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	2019-08-23 11:46:03	广铁集团	0	0	80.85%
286304	A909196	无视职工意愿、职工权益的A市伊景园滨河苑车位捆绑销售行为	2019-08-23 10:23:23	？有考虑	0	0	56.25%
214975	A909182	关于房伊景园滨河苑销售若干问题的投诉	2019-08-22 00:00:00	的成本价要	0	3	50.00%
289473	A00010343	反对滨河苑房子和车位捆绑销售	2019-08-22 00:00:00	很难兼顾的	0	0	52.27%
289588	A909183	投诉A市伊景园.滨河苑开发商	2019-08-21 21:00:21	价，行为极	0	0	71.43%
244528	A909235	伊景园滨河苑开发商强买强卖！	2019-08-21 19:05:34	，但现在基	0	2	45.45%
268299	A909193	惊！！A市伊景园滨河苑商品房竟然捆绑销售车位	2019-08-21 15:32:33	买房子的同	0	0	65.45%
251844	A909167	投诉伊景园滨河苑项目违法捆绑车位销售	2019-08-20 13:34:12	车位销售	0	1	84.00%
255507	A909195	违反自由买卖的A市伊景园滨河苑车位捆绑销售行为	2019-08-20 12:34:21	按成本价	0	0	61.95%
234633	A909194	无视消费者权益的A市伊景园滨河苑车位捆绑销售行为	2019-08-20 12:34:20	子的同时一	0	0	62.07%
230554	A909174	投诉A市伊景园滨河苑捆绑车位销售	2019-08-19 10:22:44	的，购房	0	0	100.00%
191001	A909171	A市伊景园滨河苑协商要求购房同时必须购买车位	2019-08-16 09:21:33	一名退休职	1	12	54.55%
205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	2019-08-14 09:28:31	12万一个	0	1	74.23%
268626	A909186	A市伊景园滨河苑坑害购房者	2019-08-10 12:31:01	捆绑价值1	0	0	60.24%
205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	2019-08-03 10:03:10	违法捆绑销	0	2	62.00%
251601	A909187	A市伊景园滨河苑诈骗钱财	2019-08-01 22:42:21	购金后不与	0	0	60.00%
285897	A909191	武广新城伊景园滨河苑违法捆绑销售车位，求解决	2019-08-01 20:06:52	要求捆绑购	0	0	60.00%
188801	A909180	投诉滨河苑针对广铁职工购房的霸王规定	2019-08-01 00:00:00	一个，首	0	0	40.00%
224767	A909176	伊景园滨河苑车位捆绑销售！广铁集团做个人吧！	2019-07-30 14:20:08	什么预购不	0	0	53.57%
213584	A909172	投诉A市伊景园滨河苑定向限价商品房违规涨价	2019-07-28 13:09:08	无视法律法	0	0	56.07%

图 15　A 市伊景园河滨苑捆绑车位销售问题图

通过欧几里德距离的算法，我们可以得出不同的问题与我们所标明的标志性问题的相似度。于是，我们得到了我们剔除相似度较低的数据后得到的数据。

3.4.3　特殊情况的分析

当然，我们也对特殊情况进行了分析。由于表达的方式不同，可能同一个问题是同一个意思，但是算出的相似度却不符合我们的标准，于是，我们便将它作为特殊情况的问题，如下：

260766	A00042274	A市温斯顿英语（梅溪新天地校区）店大服务差，退款难	2019/7/11 10:24:11	了，退款仍	0	0	
233066	A00072941	A市温斯顿楚府英语恶意拖欠退款	2019/6/11 15:52:48	人到温斯	0	0	55.00%
223497	A00072941	A市民办培训机构乱象丛生，温斯顿英语培训强设霸王条款	2019/6/11 0:43:37	人到温斯	0	1	35.29%

图 16　特殊情况分析图

通过对特殊情况进行分析，我们可以看出 A 市温斯顿英语（梅溪新天地校区）店大服务差，退款难与 A 市民办培训机构乱想丛生，温斯顿英语培训强设霸王条款均为表达同一个方向，只是后者表达不够清晰明了，所以我们依然将其纳入我们的问题考虑范围，不给予剔除，以保证我们模型的查全率。

3.4.4　热度指数的排序

我们计算了不同问题平均数的相关度，如下

魅力之都油烟扰民问题平均相关度计算：

$$\overline{X_1}=\frac{\begin{array}{l}71.24\%+76.38\%+84.12\%+92.57\%+89.90\%+84.65\%\\+72.17\%+76.38\%+100\%+74.12\%+82.57\%+89.09\%\\+74.65\%+71.24\%\end{array}}{14}=80.64\%$$

魅力之都房屋噪音扰民问题平均相关度计算：

$$\overline{X_2}=\frac{80.61\%+85.21\%+70.38\%}{3}=78.73\%$$

魅力之都商铺营业噪音扰民问题平均相关度计算：

$$\overline{X_3}=\frac{85.56\%+100\%+72.70\%+72.70\%+75.56\%+100\%}{6}=84.42\%$$

A7 县幼儿园收费问题平均相关度计算：

$$\overline{X_4}=\frac{81.51\%+80.74\%+85.10\%+81.44\%+93.04\%}{5}=84.36\%$$

西湖街道拆迁问题平均相关度计算：

$$\overline{X_5}=\frac{\begin{array}{l}79.13\%+76.60\%+83.55\%+89.41\%+87.33\%+76.00\%\\+87.33\%+79.60\%+79.31\%+82.20\%\end{array}}{10}=82.04\%$$

X205 路公交车问题平均相关度计算：

$$\overline{X_6}=\frac{100\%+76.10\%+76.67}{3}=84.25\%$$

A3 区油烟扰民问题平均相关度计算：

$$\overline{X_7}=\frac{\begin{array}{l}95.32\%+85.16\%+83.30\%+91.84\%+86.43\%+83.83\%\\+77_{\circ}14\%+74.24\%\end{array}}{8}=84.65\%$$

伊景园滨河苑捆绑销售问题平均相关度计算：

$$\overline{X_8}=\frac{\begin{array}{l}78.97\%+89.07\%+80.43\%+86.96\%+73.91\%+79.13\%\\+76.45\%+88.97\%+85.45\%+86.60\%+77.14\%+83.92\%\\+89.31\%+78.82\%+80.19\%+74.23\%+80.85\%+76.25\%\\+80.00\%+82.27\%+81.43\%+75.45\%+85.45\%+84.00\%\\+81.95\%+72.07\%+100.00\%+74.55\%+84.23\%+80.24\%\\+72.00\%+80.00\%+80.00\%+70.00\%+83.57\%+86.07\%\end{array}}{3}=80.82\%$$

经济学院强制学生实习问题平均相关度计算：

$$\overline{X_9}=\frac{81.26\%+80.78\%+73.33\%+72.08\%+92.31\%+85.71+86.67}{7}=71.73\%$$

温斯顿机构拖延退费问题平均相关度计算：

$$\overline{X_{10}}=\frac{83.03\%+74.44\%+85.00\%+75.29\%+76.51\%+90.23\%+76.97\%+71.03\%}{8}=79.06\%$$

最后，我们将不同关键词所对应的相似度较高的问题进行导出，并统计它们的热度指数，平均相关度进行统计，得到了以下的表格。

表 16　不同问题的平均相关度及热点指数表

问题	平均相关度	热点指数
魅力之都油烟扰民问题	80.64%	15
魅力之都房屋噪音扰民问题	78.73%	4
魅力之都商铺营业噪音扰民问题	84.42%	15
A7 幼儿园收费问题	84.36%	6
西湖街道拆迁问题	82.04%	11
X205 路公交车问题	84.25%	4
A3 区油烟扰民问题	84.65%	9
伊景园滨河苑捆绑销售问题	80.82%	37
经济学院强制学生实习问题	81.73%	8
温斯顿机构拖延退费问题	79.06%	9

根据热点指数的不同，我们对其进行了排序，并选出了热点指数排名前五的热点问题，导入并做成热点问题表和热点问题留言明细表于附件中，如下。

表 17　热点指数排行表

问题	热点指数	热点排名
伊景园滨河苑捆绑销售问题	37	1
魅力之都油烟扰民问题	15	2
西湖街道拆迁问题	11	3
A3 区油烟扰民问题	9	4
温斯顿机构拖延退费问题	9	4
经济学院强制学生实习问题	8	5
魅力之都商铺营业噪音扰民问题	6	6
A7 幼儿园收费问题	6	6
魅力之都房屋噪音扰民问题	4	7
X205 路公交车问题	4	7

3.5 问题二的总结

对于问题二，我们在对文本数据进行预处理后，根据 TF-IDF 公式选择对热点问题贡献率大、指向性强的特征词进行研究，得到可以代表某一留言问题的特征词向量，再将它们导入降维后的空间向量模型进行它们之间夹角余弦的计算，将夹角都处于一定合理范围内的同一问题反映的留言之间进行相关度的计算，对比分析后将能够合理筛选出热点问题的相关度大小定为热度评价指标。接着我们对所给文本针对关键字指标探索出相应的问题，在这些问题里，我们制定了约束条件，运用空间向量模型，对问题的相关度进行了求解，并且根据前面指定的标准选择出了某一时段内群众集中反映的某一问题可称为热点问题的问题集并且制成了表格。

第四章：留言答复的分析

4.1 数据的处理和提取

4.1.1 关键词的提取

基于第一、二章所引用的空间向量模型进行对文本数据的筛选与分析，在优质评论的挖掘中，我们仍然采用该模型对此问题进行文本处理进而分类。

我们首先对留言数据进行了预处理，利用汉语词法分析系统 ICTCLAS 对留言文本按照词性进行分词，再通过 excel 筛查词频，得到一系列高频词。筛除掉出现次数高但所反映的问题分布散乱的词汇，筛选出指向性强的词汇作为我们的特征词，如下图所示。再按照问题将他们做成特征向量以此代表问题进行分析。此处的数学计算，我们主要采用 TF-IDF 函数进行目标词汇对该话题的指向性大小的衡量从而进行特征词的选择。得出对留言主题贡献大的且具有针对性的词汇通过对提供的文本数据中群众反应的各出现次数较多的问题，我们就可以对其进行筛查与排序，筛选出出现次数较多的问题，进而对其评论文本进行分析，找出体现多角度回答的优质评论。

在找出特征词后，由于我们后面会运用到空间向量进行数据相关度的计算，因此此处我们进行文本处理，我们使用向量空间模型（VSM）来表示文档。我们将文档中出现的特征词用 $Term_1$，$Term_2 \cdots\cdots Term_n$ 进行表示，并将其表示出来。

$$Doc \to (Term_1, Term_2, Term_3 \cdots\cdots, Term_n)$$

表 18 关键词的次数及其频率统计表

关键词	次数	频率（%）
搬迁	13	0.050
地铁	27	0.10
地铁	27	0.10
买房	9	0.034
公交车	56	0.21
补课	43	0.16
养老	21	0.080
扰民	72	0.28
自来水	23	0.088
小学	54	0.21
高速	15	0.057
驾校	13	0.050
医保	28	0.11
报销	23	0.088
松雅	11	0.042

4.1.2 关键词的次数及其频率统计

根据下面留言问题关键词出现频率条形图，我们可以简单明了地看到出现的词汇大部分的频率都超过了 0.05，且基本都具有很强的话题指向性，对问题相关度的计算都具有很大的贡献。

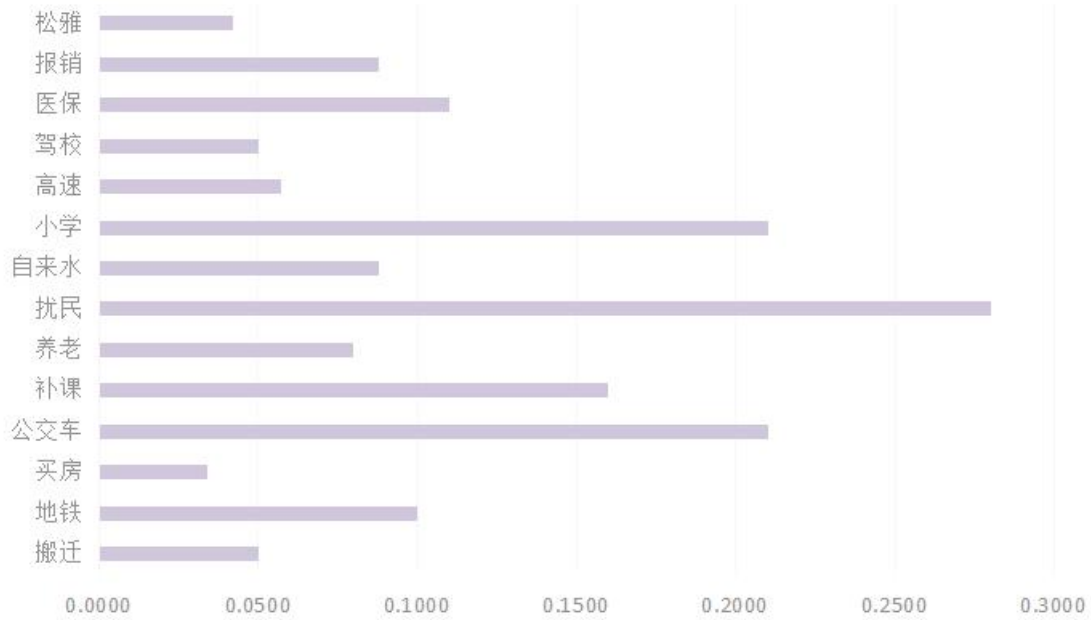

图 17 关键词出现频率条形图

4.1.3 留言问题的统计

留言数据进行了预处理后，通过对某一时段内群众集中反应的某一问题，我们可以对其进行筛查，更容易选出聚集性问题。而通过对不同月份的研究，我们发现了不同月份中群众所反应的问题量有所波动。除此之外，由于部分留言中的问题涉及到政策等具有时效性的部分，于是为了确保我们参考评论的准确性与完整性，我们将不同月份作为横轴，问题的数量作为 y 轴，建立了平面直角坐标系，并将问题的数量以折线图的方式表现出来，如下图。

表 19　月份所对应的问题数统计表

月份	问题数
2020	4
2019	886
2018	847
2017	453
2016	227
2015	130
2014	103
2013	109
2012	52
2011	5

图 18　2011 年至 2020 年逐年问题数折线统计图

4.2 相似问题的筛选

4.2.1 问题的归类

通过对关键字的筛选，我们选择除了具有指向性的几个特征向量，也就是特殊关键词。

我们同样通过构建空间向量直角坐标系，将我们筛选的特殊关键词接介入表格，选出有一定关联的问题。空间向量模型中向量之间的余弦计算。接下来，我们将对不同类别的问题进行筛选，筛选出同一类别且相似度较高的问题。

我们通过对不同问题类别中，分别标注不同的具有标志性的问题，以它为模板，进行相似度的计算。通过引用公式，我们计算了两个问题在三维坐标系里所形成的夹角，夹角越小，相似度越大。因此我们在处理若干留言问题数据后，得到最为合理的相关度。

我们对不同类别的问题进行了归类，例如 A 市的公交问题：

表 20　A 市公交问题汇总表

留言编号	留言用户	留言主题
2574	A0009233	关于 A 市公交站点名称变更的建议
3713	UU0081227	建议增开 A 市 261 路公交车
4157	UU008816	对 A 市公交线路的建议
4180	UU0081030	投诉 A 市 908 区线公交车对老人不友好
4192	UU008524	关于 A 市公交管理的建议
4197	UU0081611	关于 A 市 brt 公交规划的咨询
4931	UU0082182	建议 A 市公交开通乘车码，可以移动支付
5667	UU0081085	建议 A 市有关部门把公交站台等进行全面检查排查
5873	UU0081434	建议 A 市公交管理部门完善公交相关事项
6174	UU0081797	请求 A 市相关部门将 128 路公交车末班车时间延长至 21:45
6322	UU0081207	希望增加 A7 县筑梦园小区到 A 市内的公交车
6519	UU0081406	反映下雨天 A 市公交车站等车，车子水花问题
6610	UU0082032	希望尽快解决 A 市公交卡网上充值业务
6760	UU0081338	希望咱们 A 市公交也能引入移动支付

4.2.2 异常数据的处理

面对问题的选择，我们同样面临着异常数据的剔除，相似问题可能因为表达的方式差别太大而被剔除，而我们也对剔除的数据进行了检索，并将表达意思相同，却因表达方式差异太大而被剔除的问题进行了还原：

例如：

A市金科世界城业主关于小孩“就近入学”的请求
举报A7县教育局在小一学区划分中没有遵循“就近入学”原则
咨询A7县泉星社区东城名苑小孩就近入学问题
咨询A7县星沙东城名苑小区就近入学问题
咨询B市小学生入学事项
B市华晨格林水岸业主咨询小孩入学问题
B2区翡翠公园业主小孩不能入学
建议将C4市原一职改为小学缓解小学入学难问题
对C市小学入学政策的咨询
关于在C市买房的业主小孩小学入学问题的咨询
咨询C市小学入学年龄限制问题
咨询C市幼升小入学问题
咨询F5县小孩入学问题
关于J市小孩入学难问题的反映
对L6县小学入学政策的质疑
咨询L6县小学的入学政策
咨询外地户口小孩在L2县城的入学条件
关于L5县小学入学申请资料的咨询
咨询M2县永丰镇绍塘村拆迁户小孩的入学问题

图 19　异常数据的还原及标注图

4.2.3 相关度的处理

通过对问题进行处理，我们分析了答复意见的相关度，我们同样使用了相关度的计算公式，在建立空间坐标系后，将 X 轴设定为时间轴，将 Y 轴设定为时间发生的次数，将 Z 轴设定为该类时间发生的频率，将文本内容转化为可进行数学计算的数据，则政府的答复意见和问题的关系可以用 di 和 dj 表示，即文档 di，dj 的相似度可以表示为

$$Sim(d_i,d_j)=\cos\theta=\frac{\sum_{k=1}^{n}\omega_k(d_i)\times\omega_k(d_j)}{\sqrt{\left(\sum_{k=1}^{n}\omega_k^2(d_i)\right)\left(\sum_{k=1}^{n}\omega_k^2(d_j)\right)}}$$

夹角越小，相似度越大。我们将政府的答复相关度进行算出，并分析答复的相关性，我们可以得到了不同问题的政府留言相关性回馈。

针对公交问题，我们进行了分析，公交问题是许多市民所会反应的问题，公交车问题的类别与反馈的方向不同，使得政府的留言方向与完整度不同，通过统计政府的答复意见，我们得到了以下不同问题答复意见的相关度：

答复意见	答复时间	相关度
坡岭”的问题。公交站点的	2019/5/9 9:51:30	86.14%
于驾驶员工作时间长，劳动	2019/1/14 14:33:17	85.14%
输局将认真考虑，最大限度	2018/11/27 14:48:01	80.38%
公交站点，驾驶员正常停靠	2018/11/29 14:58:43	83.71%
直要求企业加强对驾驶员	2018/11/27 14:45:43	95.68%
的意见和建议，市城乡规划	2018/11/27 14:43:34	87.14%
支持“楚行一卡通”app，	2018/9/18 9:23:28	77.89%
护的户外公交站设施在汛	2018/6/28 14:07:08	73.64%
信息，市交通运输局运管	2018/6/12 14:33:59	80.00%
表示目前A市公交驾驶员缺	2018/5/7 14:45:57	90.36%
公交到开元路和万家丽路	2018/5/7 14:35:30	70.00%
业，要求公交车驾驶员遇	2018/4/8 16:26:35	89.60%
口。公交ic卡管理中心已	2018/3/23 15:40:08	86.96%
。公交ic卡管理中心已委	2018/2/22 15:02:17	90.45%

图 20　相似问题答复意见及其相关度

通过对不同问题的政府留言相关度进行计算，我们得到了如下的表格：

表 21　不同问题所对应的平均相关度表

反馈问题	平均相关度
L3 县搬迁问题	82.07%
A 市地铁站点问题	83.21%
A 市买房问题	82.71%
A 市公交问题	84.08%
C 市补课问题	80.97%
A 市小区扰民问题	85.06%
A 市自来水问题	77.57%
小学入学问题	49.12%
A 市高速问题	90.98%
A 市驾校退费问题	96.88%
医保报销问题	80.85%
松雅湖交通安全问题	80.82%
养老问题	82.53%

4.3 问题三的总结

4.3.1 相关性与完整性的计算

假设人工筛选的关键词为：公交车、补课、小学、扰民、医保、地铁、自来水、报销、环境

则该文本分类系统的准确率

$$P=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}=\frac{56+43+54+72+28+27+23+23}{56+43+54+72+28+27+23+23+40}\approx 89.07\%$$

该文本分类系统的查全率

$$R=\frac{\text{分类的正确文本数}}{\text{应有文本数}}=\frac{56+43+54+72+28+27+23+23}{56+43+54+72+28+13+27+9+21+23+15+13+23+11}\approx 79.90\%$$

$$F1\text{测试值}=\frac{\text{准确率}\bullet\text{查全率}\bullet 2}{\text{准确率}+\text{查全率}}=\frac{89.07\%\bullet 79.90\%\bullet 2}{89.07\%+79.90\%}\approx 84.24\%$$

4.3.2 留言答复的分析

根据前面文本预处理以及通过建立空间向量模型进行留言问题相关度的计算，我们可以筛选出各主题所包含的问题，从而可以进行类似问题的评论文本比较。

根据我们通过空间向量模型所找的筛选出的出现次数多的留言问题，我们对应他们的回复，分别考察了其相关的特征，包括回答长度、回答的非重复字符数和关键词相关度等。其中优质回答和其他答案的回答长度与回答的非重复字符数这两个特征分布差异较大，对优质回答与一般回答的区分力较强。而由于我们需要处理的留言问题文本数据量庞大，其中包含着众多或简单或复杂的问题，这就导致了有些简单问题的优质回答与一般回答的回答长度是差不多的，而一些复杂问题则是优质回答的详细而明了，一般回答就粗糙且模糊。因此我们只通过这两个指标对留言问题回答进行初步评价。

因此，我们可以通过基于问题粒度的特征，通过比较一个问题的不同回答的某种特征的相对大小，做出相对性的衡量。令问题 q 的第 x 个回答的 f 特征的值为 fx，q 问题的回答总数为 n，则第 x 个回答的特征的基于问题粒度特征的定义为：

$$QG(f_x)=\frac{f_x}{\max(f_1,f_2,\cdots,f_n)}$$

因而通过问题粒度的特征，我们可以将我们前面分类好的留言问题的回答基于关键词进行文本特征分析，结合文本长度与回答的非重复字符数进行比较，对目标回复文本从相关性、完整性与可解释型进行质量评测。

参考文献

i 杨凯峰, 张毅坤, 李燕. 基于文档频率的特征选择方法[J]. 计算机工程, 2010, 036(017):33-35,38.

ii GUAN, Lei, 关磊, MO, Sha. BOOKMARK INTELLIGENT CLASSIFICATION METHOD AND SERVER: WO 2012.

iii 魏德华. 微博热点话题发现问题的研究及实现[D].福州大学,2017.

iv 孔维泽, 刘奕群, 张敏, et al. 问答社区中回答质量的评价方法研究[J]. 中文信息学报, 2011(01):5-10.