“智慧政务”中的文本挖掘应用

摘要

近年来，微信、微博、市长信箱等网络问政平台已然成为了政府了解民意、汇聚民气、凝聚民气的重要渠道。而随着各类社情民意相关的文本数据量不断攀升，给依靠人工来进行留言划分和热点整理的相关部门工作带来了极大挑战。因此，运用自然语言处理和文本挖掘的方法对解决此类问题具有重大意义。

针对问题 1：通过 spyder 软件读取附件 2 中的数据，将新的“主题”这一列作为决策属性，“一级标签”作为类别属性。然后通过 spyder 中的 jieba 分词函数和 CountVectorizer 函数预处理“留言主题”这列数据。将处理后的数据分为训练集和测试集，利用 fit_transform 对训练数据进行拟合和标准化，transform 对测试数据进行标准化，然后使用 sklearns 中的 sklearn.naive_bayes 方法对我们的数据集做出预测，实现朴素贝叶斯。最后得到查准率 P 和查全率 R 的值分别为 0.86 和 0.78，因此 $F-Score=\frac{2\times P\times R}{(P+R)}=0.82$，然后利用 F-Score 对该朴素贝叶斯分类方法进行评价。

针对问题 2：根据附件 3 中“留言主题”这一列的信息，利用问题一中的朴素贝叶斯预测方法对特定人群问题的留言进行归类。并将其新存为表格“附件 3 预测结果.xlsx”。对于热度评价指标，我们通过锁定问题的关键字，然后找出它在留言中出现频数来作为该类问题的热度指数，并将热度指数作为一列添加进附件 3 并存为新的表格“附件 3 热度指数.xlsx”。通过得到的结果将热度指数进行排序，找出热度指数最高的前五个问题，并将需要的信息提取出来得到“热点问题表.xlsx”和“热点问题留言明细表.xlsx”。

针对问题 3：提取附件 4 中“留言主题”，“留言详情”和“答复意见”并将其存为.txt 形式，画出“留言详情”和“答复意见”词云图，找出它们各自的关键词；作“留言主题”与“答复意见”，“留言详情”与“答复意见”间的相似性。以此分析相关部门对留言的答复意见答复的相关性、完整性、可解释性等。

关键词：spyder 软件；jieba 分词函数；朴素贝叶斯；相似性

Abstract

In recent years, online questioning platforms such as WeChat, Weibo, and Mayor ' s Mailbox have become important channels for the government to understand public opinion, gather popular opinion, and condense popular opinion. As the amount of text data related to various social conditions and public opinion continues to rise, it has brought great challenges to the work of relevant departments that rely on manual division of messages and hotspot sorting. Therefore, the use of natural language processing and text mining methods is of great significance to solve such problems.

Aiming at Problem 1: Read the data in Attachment 2 through spyder software, and use the new "Subject" column as the decision attribute, and the "First Class Label" as the category attribute. Then use the jieba word segmentation function and CountVectorizer function in spyder to pre-process the "message subject" column of data. The processed data is divided into training and test sets. Fit_transform is used to fit and normalize the training data. Transform normalizes the test data, and then uses the sklearn.naive_bayes method in sklearns to make predictions on our data set to achieve Naive Bayes. Finally, the values of the precision and recall are 0.86 and 0.78 respectively, so the values are, and then the naive Bayes classification method is evaluated using F-Score.

Regarding Question 2: According to the information in the "Message Subject" column in Annex 3, use the naive Bayesian prediction method in Question 1 to classify the message of a specific group of people. And save it as a table "Annex 3 prediction results. Xlsx". For the heat evaluation index, we lock the keyword of the question, and then find out how often it appears in the message as the heat index of this type of problem, and add the heat index as a column to Annex 3 and save it as a new table "Annex

3 heat Index.xlsx ". According to the obtained results, the heat index is sorted to find the top five questions with the highest heat index, and the required information is extracted to obtain the "hotspot question list.xlsx" and "hotspot question message list.xlsx".

For question 3: extract the "message subject", "message details" and "answer comments" in Annex 4 and save them as .txt format, draw the "message details" and "answer comments" word cloud maps and find out their respective Keywords; make "message subject" and "reply opinion", similarity between "message details" and "reply opinion". This is to analyze the relevance, completeness, and interpretability of the relevant department's response to the message.

Keywords: Spyder Software ; jieba Wrd ; Segmentation Naive Bayes Classifier ; Similarity.

目录

1、挖掘意义及目标

1.1 挖掘意义

智慧政务办公云平台是利用现代信息技提高政府智能化水平的一种形态。智慧政务建设能有效挖掘、分析处理各种政务信息，节省大量的政务成本，提高政府行政效率，增加政务的透明性和公正性，为社会经济运行态势和社会管理创新提供有力的方法手段，这是转变政府职能的创新性手段，也是推进国家治理智能化、精细化、高效化的重要方法途径和选择。

智慧政务政务指的是可以实现政务服务高效化，数据实时化，响应及时化的政务工具。可以简单、高效的解决百姓的各类问题，数据统计清晰明了。

1.2 挖掘目标

本次建模目标是根据互联网公开渠道发布的数据，利用 spyder 中的 jieba 分词函数对“留言主题”进行分词，通过 CountVectorize 函数进行预处理、scikit-learn 实现朴素贝叶斯以及相似度分析等主要方法来完成以下三个目标：

1）为了方便将群众留言分派至相应的职能部门进行处理，解决工作量大、效率低、差错率高等问题。建立关于留言内容的一级标签分类模型。

2)根据附件 3 将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。

3）针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性和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

总体流程图:

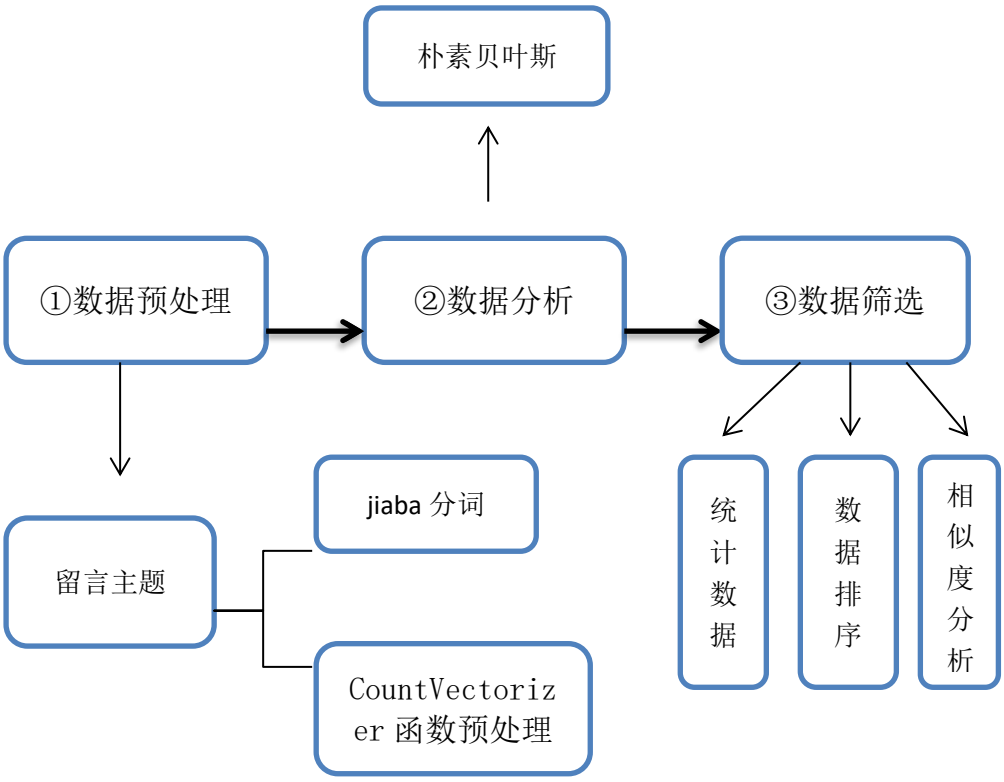

图 1　总体流程图

本文主要包括如下步骤：

步骤一：数据预处理，在给出的数据中，对附件中的“留言主题”，“留言详情”及“答复意见”等文本数据利用 jieba 进行分词。

步骤二：数据分析，根据附件 3 中“留言主题”这一列的信息，通过朴素贝叶斯分类方法对特定人群问题的留言进行归类。

步骤三：数据的筛选，通过相关数据，来将热度指数进行排序，找出热度指数最高的前五个问题。最后通过“留言主题”与“答复意见”，“留言详情”与“答复意见”作相似性分析。

2.1 问题 1

2.1.1 数据预处理

利用 spyder 中的 jieba 分词函数对附件 2 中的“留言主题”进行分词，然后将分词后的文本以“主题”作为列名添加到附件 2 的数据中，在 scikit-learn 中实现 Bag of Words，因此使用 CountVectorizer 函数预处理该列数据，然后把“主题”这一列作为决策属性，“一级标签”作为类别属性，通过 train_test_split 函数将这两列数据拆分成训练集和测试集各占$\frac{2}{3}$和$\frac{1}{3}$，并利用 fit_transform 对训练数据进行拟合和标准化，transform 对测试数据进行标准化。

2.1.2 朴素贝叶斯

朴素贝叶斯法发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC 模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为 NBC 模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给 NBC 模型的正确分类带来了一定影响。

（1）贝叶斯公式：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

其中B是观察样本个体，A为被预测个体所属类别

（2）朴素贝叶斯：

$$P(D|C) = P(D1|C) \times P(D2|C) \times \cdots \times P(Dn|C)$$

其中C表示某种类别，用D代表数据集中的一篇文档，$D1, D2 \cdots, Dn$为D中的特征

2.1.3 F-Score

1）二元分类模型个案预测的四种结局

真阳性：预测为真，实际也为真

伪阳性：预测为真，实际为假

真阴性：预测为假，实际也为假

伪阴性：预测为假，实际为真

		真实值		总数
		p	n	
预测输出	p'	真阳性 (TP)	伪阳性 (FP)	P'
	n'	伪阴性 (FN)	真阴性 (TN)	N'
总数		P	N	

表 1

样本的数量$N = P + N = TP + FP + FN + TN$

2）准确率，正确率，召回率

准确率（accuracy）：所有的预测正确的占总的比重

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

精确率（也叫查准率，precision)：真正正确的占所有预测为正确的比例

$$precision = \frac{TP}{TP + FP}$$

召回率（也叫查全率，recall）：真正正确的占所有实际为正确的比例

$$recall = \frac{TP}{TP + FN}$$

F-score 值:F-score 值为算数平均数除以几何平均数，且越大越好

$$\frac{2}{\mathrm{F-score}}=\frac{1}{precision}+\frac{1}{recall}$$

即：

$$\mathrm{F-score}=\frac{2PR}{P+R}$$

其中P表示$precision$，R表示$recall$

准确率和召回率是互相影响的，理想情况下肯定是做到两者都高，但是一般情况下准确率高、召回率就低，召回率低、准确率高。

2.2 问题 2

2.2.1 数据预处理

将附件 3 中“留言主题”利用 spyder 中 jieba 函数进行分词，并将其存为数据框中新的一列“主题”，通过 transform 对该列文本数据进行标准化。

要挖掘群众集中反映的问题，也就是热点问题，首先提取附件 3 中“留言主题”这一列，但是这一列中不仅包含了留言的内容，还有反映的群众的具体位置，所以我们通过分词对其进行拆分，将群众位置和留言的内容区分开，然后分别统计出每个地区在不同时间段群众反映的“热点问题”。

2.2.2 数据预测与整理

对预处理的数据，使用问题一中的朴素贝叶斯分类模型，通过 sklearns 中的 sklearn.naive_bayes 方法对附件 3 中反映的特定地点或特定人群问题的留言进行归类，将分类结果以列名“一级标签预测结果”存储，并将其保存到新的表格“附件 3 预测结果.xlsx”。

通过锁定附件 3 中留言问题的关键字，然后找出它在留言中出现频数来作为该类问题的热度指数，并将热度指数以列名“热度指数”添加进附件 3 并存为新的表格“附件 3 热度指数.xlsx”，最后再借助 R 软件画“热度指数”变化趋势图，以便于相关部门确定对这几个热度问题的重视程度。

将表格“附件 3 热度指数.xlsx”按照热度指数进行排序，从而得到热度指

数排序最高的前五个问题，即排名前 5 的热点问题。然后将数据分析整理后分别得到“热点问题表.xls”和“热点问题留言明细表.xls”两个文件。

2.3 问题 3

2.3.1 数据预处理

为了分析相关部分的答复意见的相关性、完整性、可解释性等。将附件 4 中的“留言详情”，“留言详情”，“答复意见”提取出来分别保存为三个.txt 文件。利用 re.findall 函数将文本中的无用符合删掉，然后通过 jieba 函数对三个对它们进行分词。

2.3.2 词云图及相似度

在作词云图时，首先将不需要的修饰词从文本中移除，然后利用 collections.Counter 函数统计词频。选择背景图后，通过 wordcloud 画出“留言详情”和“答复意见”两张词云图，从词云图中找出它们各自的关键词。关键词的匹配度可以粗略的判断相关部门答复意见的相关性、完整性，可理解性。

对分词后的文档整理成指定格式后计算词频，对频率低的词语进行过滤，通过语料库建立词典，将文档通过 doc2bow 转化为稀疏向量，将新语料库通过 tf-idf 模型处理，计算稀疏矩阵的相似度，得到相似度结果。个方法在一定程度上可以看出相关部门答复意见的相关性。

3、结果分析

3.1 问题 1

通过对附件 2 数据预处理后，使用 scikit-learn 实现朴素贝叶斯得到关于留言内容的一级标签分类模型。

利用 accuracy_score,precision_score,recall_score,f1_score 三个函数

得到准确率，查准率和查全率分别为：

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} = 0.82$$

$$precision = \frac{TP}{TP + FP} = 0.86$$

$$recall = \frac{TP}{TP + FN} = 0.78$$

因此：

$$\mathrm{F-score} = \frac{2PR}{P + R} = \frac{2 \times 0.86 \times 0.78}{0.86 + 0.78} = 0.82$$

从得到的$\mathrm{F-score} = 0.82$，我们认为利用朴素贝叶斯得到关于留言内容的一级标签分类模型效果较好。

3.2 问题 2

1、根据问题一中的朴素贝叶斯分类模型得到附件 3 关于留言内容的一级标签结果：

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	主题	一级标签预测结果
188006	A00010294	A3区一米	2019/2/2	座落在A市	0	0	A3 区 一	教育文体
188007	A00074795	咨询A6区	2019/2/1	A市A6区建	0	1	咨询 A6	城乡建设
188031	A00040066	反映A7县	2019/7/1	本人系春	0	1	反映 A7	城乡建设
188039	A00081379	A2区黄兴	2019/8/1	靠近黄兴	0	1	A2 区 黄	城乡建设
188059	A00028571	A市A3区中	2019/11/	A市A3区中	0	0	A 市 A3	城乡建设
188073	A909164	A3区麓泉	2019/3/1	作为麓泉	0	0	A3 区麓	城乡建设
188074	A909092	A2区富绿	2019/1/3	〝二高一	0	0	A2 区富	城乡建设
188119	A00035029	对A市地铁	2019/5/2	我是一名	0	0	对 A 市	城乡建设
188170	A88011323	A市6路公	2019/12/	12月21日	0	0	A 市 6 路	城乡建设
188249	A00084085	A3区保利	2019/9/1	保利麓谷	0	0	A3 区 保	城乡建设
188251	A00013092	A7县特立	2019/10/	近来，下	0	0	A7 县 特	城乡建设
188260	A00053484	A3区青青	2019/5/3	还我宁静	0	0	A3 区 青	城乡建设
188396	A00047580	关于拆除	2019/4/1	桐梓坡58	2	1	关于 拆除	城乡建设
188399	A00097934	A市利保壹	2019/7/3	您好，我	0	0	A 市利保	城乡建设
188409	A0003274	A市地铁3	2019/6/1	尊敬的领	0	4	A 市 地铁	城乡建设
188414	A00096844	A4区北辰	2019/8/1	您好！我	0	0	A4 区 北	城乡建设
188416	A00029753	请给K3县	2019/06/	K3县的乡	0	0	请 给 K3	卫生计生
188451	A00013004	A7县春华	2019/4/1	我是春华	0	2	A7 县 春	城乡建设
188455	A00035902	咨询异地	2019/5/1	书记您好	0	0	咨询 异地	劳动和社会保障
188467	A00050188	投诉A市滥	2019/3/2	退费之日	0	1	投诉 A 市	教育文体

表 2　附件 3 预测结果（部分数据，详见“附件 3 预测结果.xlsx）

2、将热度指数添加进附件 3 后得到结果：

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	热度指数
188006	A00010294	A3区一米	2019/2/2	座落在A市	0	0	5
188007	A00074795	咨询A6区	2019/2/1	A市A6区道	0	1	2
188031	A00040066	反映A7县	2019/7/1	本人系春	0	1	6
188039	A00081379	A2区黄兴	2019/8/1	靠近黄兴	0	1	3
188059	A00028571	A市A3区中	2019/11/	A市A3区中	0	0	5
188073	A909164	A3区麓泉	2019/3/1	作为麓泉	0	0	1
188074	A909092	A2区富绿	2019/1/3	“二高一	0	0	1
188119	A00035029	对A市地铁	2019/5/2	我是一名	0	0	1
188170	A88011323	A市6路公	2019/12/	12月21日	0	0	5
188249	A00084085	A3区保利	2019/9/1	保利麓谷	0	0	1
188251	A00013092	A7县特立	2019/10/	近来，下	0	0	1
188260	A00053484	A3区青青	2019/5/3	还我宁静	0	0	7
188396	A00047580	关于拆除	2019/4/1	桐梓坡589	2	1	3
188399	A00097934	A市利保壹	2019/7/3	您好，我	0	0	1
188409	A0003274	A市地铁3	2019/6/1	尊敬的领	0	4	6
188414	A00096844	A4区北辰	2019/8/1	您好！我	0	0	3
188416	A00029753	请给K3县	2019/06/	K3县的乡	0	0	2
188451	A00013004	A7县春华	2019/4/1	我是春华	0	2	5
188455	A00035902	咨询异地	2019/5/1	书记您好	0	0	7
188467	A00050188	投诉A市温	2019/3/2	退费之日	0	1	2

表 3　附件 3 热度指数（部分数据,详见“附件 3 热度指数.xlsx”）

3、将“附件 3 热度指数”文件进行整理后得到

（1）

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	29	2019/1/8 至 2019/12/4	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	14	2017/6/8 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
3	3	10	2018/5/17 至 2019/11/27	A 市西湖街道茶场村	拆迁事宜及相关规划
4	4	10	2019/1/6 至 2019/9/12	A 市长房云时代	房子建立与维修，垃圾场的处理，幼儿园相关事宜
5	5	8	2019/1/2 至 2019/7/21	A 市北辰三角洲	住房及车库问题

表 4　热点问题表

从热点问题表知，排名前 5 的热点问题地点是，“魅力之城”，“经济学院”，“西湖街道茶场村”，“长房云时代”，“北辰三角洲”，热度指数分别为 29，14，10，10，8。主要反应的问题分别是“小区临街餐饮店油烟噪音扰民”，“学校强制学生去定点企业实习”，“拆迁事宜及相关规划”，“房子建立与维修，垃圾场的处理，幼儿园相关事宜”，“住房及车库问题”。

（2）将热点问题表中的“热度指数”通过 R 软件画折线图

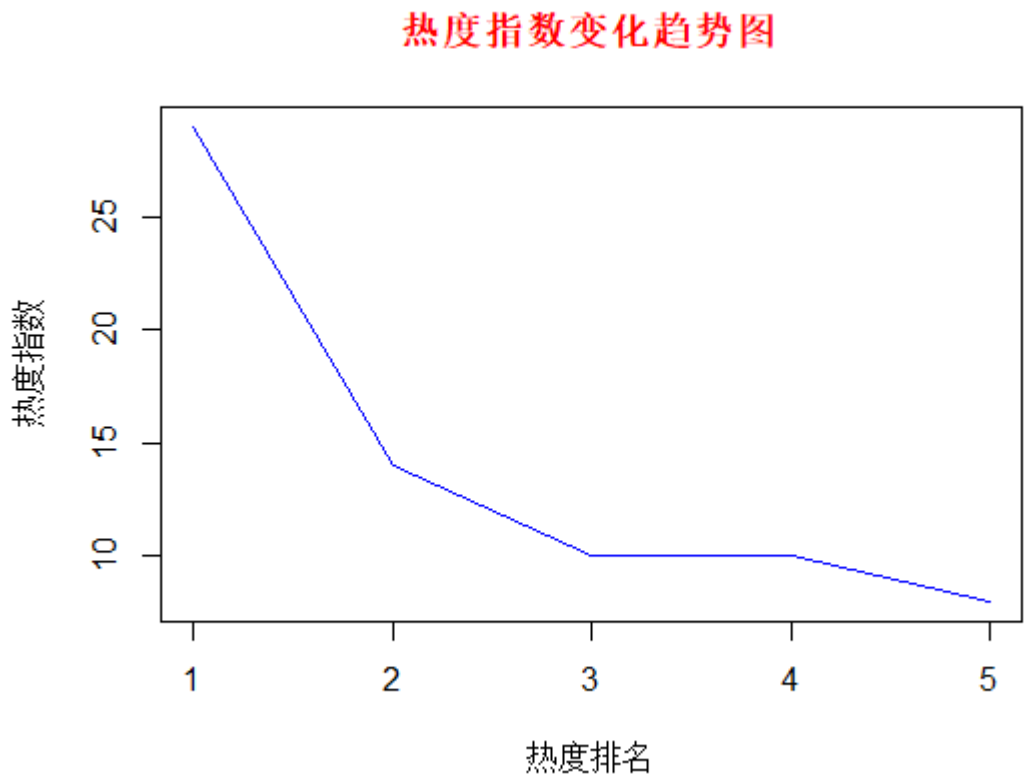

图 2　热度指数变化趋势

上图可以看出，热度指数第一的比第二高很多，第二比第三高较多，后面的热度指数都相差不大了。因此相关部门在解决问题或是作答复意见时一定要着重前两个热度指数较高的问题。

（3）

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	189381	A000109815	A市万科魅	2019/12/4	A市万科魅	0	0
1	195095	A00039089	魅力之城小	2019/09/05	魅力之城小	3	0
1	198084	A00022429	A市万科魅	2019/10/23	本人花近1	0	0
1	205168	A00022429	A市万科魅	2019/10/24	A市万科魅	1	0
1	232892	A00015335	A市万科魅	2019/1/8 9	您好！我是	1	0
1	233338	A00022429	A市万科魅	2019/11/13	我是万科魅	0	0
1	236303	A00022429	A市万科魅	2019/10/29	我是万科魅	0	0
1	236798	A00039089	A5区劳动东	2019/07/28	尊敬的政府	4	0
1	240330	A00087099	A市万科魅	2019/3/21	本人2018年	0	0
1	242792	A909115	A5区魅力之	2019/08/26	我们是魅力	1	0
1	245136	A909117	万科魅力之	2019/09/04	您好：我是	0	0
1	246362	A909114	A市魅力之	2019/08/26	2019年5月	0	0
1	246598	A00054842	A5区劳动东	2019/09/25	A5区劳动东	1	0
1	253314	A00089954	反映A5区万	2019/4/17	尊敬的市委	0	0
1	263498	A00015643	A市万科魅	2019/4/17	市长：您好	0	0
1	268914	A0006238	A5区劳动东	2019/09/10	A5区劳动东	0	0
1	272122	A909113	A5区劳动东	2019/08/01	局长：	6	0
1	284147	A909113	A5区劳动东	2019/07/21	局长：	3	0

表 5　热点问题留言明细表（部分数据，详见“热点问题留言明细表.xlsx”）

表中可以知道每个地区的热点问题具体是哪些，例如：A 市万科魅力之城小区热点问题主要是“商铺无排烟管道，小区内到处油烟味，近百户楼板开裂墙面开裂，开发商未通知业主就进行车位开盘销售活动...”，这些问题为热点问题。想知道某个地区热点问题的详细情况，只要通过这个表格就可以看出来。

3.3 问题 3

1、词云图

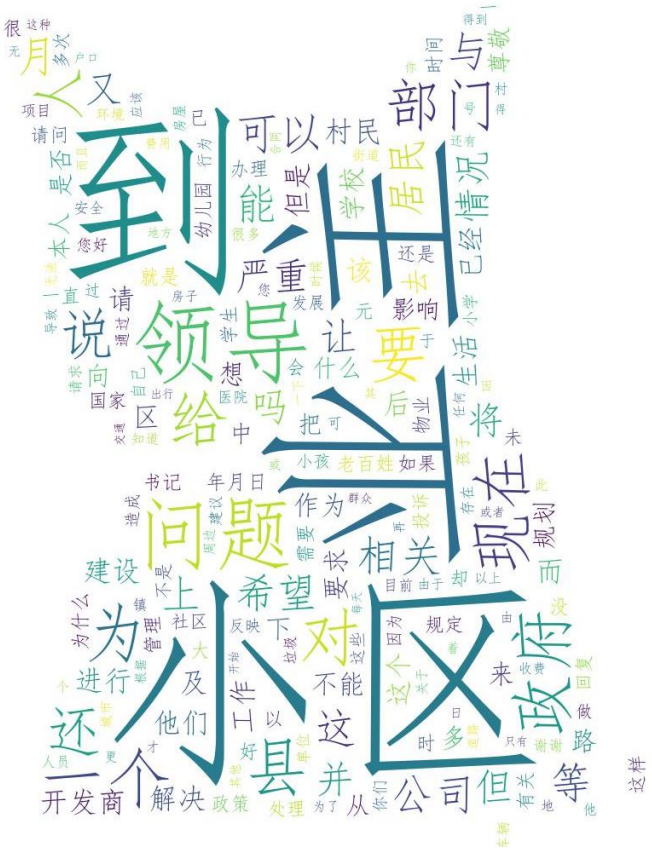

图 3　留言详情词云图

图 4　答复意见词云图

两个词云图我们可以明显看到，留言详情中出现最多的字眼是“小区，业主，领导，问题”等，答复意见中出现最多的字眼是“问题，工作，进行，情况”等。

2、相似度结果

留言详情与答复意见相似度	留言主题与答复意见相似度
0.3937	0.0005

表 6　相似度结果

表 6 我们可以发现，留言详情与答复意见相似度有 0.3937，而留言主题与答复意见相似度更低，只有 0.0005，因此我们初步判断，相关部门对部分群众留言的答复意见相关性不是很强，可能对于很多问题还没有给出合理的建议或是回答过于官方。

4、结论

对于留言的分类问题进行了一系列的分析研究之后，了解民意问题以及对热点问题的发掘对于提升政府的管理水平和施政效率具有极大的推动作用。由于存在留言类型多、分类不明确的问题给政府了解民意情况带来了一定的挑战，因此本文通过朴素贝叶斯将留言进行分类，以便了解群众的留言内容。最后得到查准率P和查全率R的值分别为 0.86 和 0.78，因此$F-Score=\frac{2\times P\times R}{(P+R)}=0.82$，说明此方法是比较可靠有用的。对政府的管理机构以及了解民意起到一定的作用。

对于在留言内容上的热点问题的发掘，我们可以知道群众反应的最多的问题有：油烟扰民、噪声严重、空气的过度污染。学生反映的最多的问题有垃圾桶的位置摆放、心中所想的实习与学校背道而驰。对于住宅小区，反映的是拆迁、房子裂缝问题。

相关部门依据群众留言内容进行了一定的答复，针对不同的问题有不同的解释。通过“留言主题”与“答复意见”，“留言详情”与“答复意见”做相似性分析可以知道，该答复意见的质量是有一定保障的。

5、给政府及相关部门的建议

通过对表格中的群众留言发现，对于居民而言，热点问题大多是跟生活环境息息相关的，例如群众反映的油烟扰民问题、噪声问题、空气污染等问题；而对于学生来说，大多数问题都是学生学习上的问题，例如对于实习就产生了很多的问题，而从有关部门的回复来看，回复太过于官方，概括性太强。

所以，从这个热点问题的挖掘结果来看，建议有关部门在收到到留言的时候根据各问题的热度指数给予不同的重视程度，及时给群众合理答复并有效解决好每一个问题，对不同的社会群体，不同的问题，找出更优的解决方案。

6、参考文献

[1]邢江豪,覃楚岳,刘力赫.基于 SOA 的物联网数据应用架构[J].电子世界. 2017(09)

[2]迪莉娅.基于云计算的电子政务大数据管理研究[J].图书馆理论与实践. 2013(12)

[3]王柯.智慧城市建设过程中政府管理机制与创新研究[J].中国高新区,2018(11):200.

[4]张亚丽.基于云计算架构下智慧政务平台的设计与实现[J].中国新通信,2018,20(14):118-119.

[5]刘文富.智慧政务:智慧城市建设的政府治理新范式[J].中共南京市委党校学报,2017(01):62-68.

[6] 王克照,智慧政府之路[M]. 清华大学出版社 , 2014.