

基于朴素贝叶斯分类的智慧政务文本挖掘

目录

摘要.....	3
一. 挖掘目标.....	5
1.1 挖掘背景.....	5
1.2 挖掘目标.....	5
二、 符号说明.....	6
三. 问题分析.....	7
3.1 问题一的分析.....	7
3.2 问题二的分析.....	7
3.3 问题三的分析.....	8
四. 模型建立与求解.....	9
4.1 问题一模型的建立（群众留言分类）.....	9
4.1.1 朴素贝叶斯分类模型：.....	9
4.1.2 F-score 评价方法的建立：.....	11
4.2 问题二模型的建立（热点问题的挖掘）：.....	12
4.2.1 热点问题描述.....	12
4.2.2 输出相应的表.....	13
4.3 问题三模型的建立（答复意见的评价）：.....	14
4.3.1 答复意见的相似性：.....	14
4.3.2 答复意见的完整性：.....	16
4.3.3 答复意见的可解释性：.....	17
4.3.4 综合评价方案：.....	17
五. 总结.....	18
六. 参考文献.....	18

摘要

近年来,各种各样的网络问政平台慢慢的成为了政府了解、处理、解决民众问题的重要渠道,而各类不同数据量不断攀升,给以往主要靠人工来进行划分和整理的相关部门的工作带来了极大挑战。于此同时,随着相关技术的发展,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,能极大的推动政府的管理水平和施政效率。

本文根据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答复意见。打算利用自然语言处理和文本挖掘的方法解决下面主要的三个问题。

群众留言分类问题,本文可以从附件 1 中看出有有城乡建设、党务政务、国土资源等 15 个一级分类,对此根据附件 2 给出的 7 个一级分类,本文首先截取留言详情和一级分类两列数据,对留言详情进行分词,并自定义分词词典,去除停用词,就可得到新的留言详情数据,进而建立一级标签朴素贝叶斯分类模型,得到分类模型的准确率大约为 86.5%。

热点问题挖掘问题,首先,关于热点问题描述可以先将问题中出现超过一定次数的高频词看作为热点问题,对留言主题字段进行分词并去除停用词,然后利用字典类型的方式对留言主题进行一一匹配,统计出地点或人群的高频词,接着对前五名的留言时间进行排序,在同一地点或人群的条件下结合点赞数与反对数得出热点问题描述。其次,输出相似地点对应的留言主题以及对应的留言编号、用户、时间等数据。

答复意见的评价问题,本文需要从三个方面进行综合计算评价,首先可以通过余弦相似性来计算留言详情和答复意见的相似度,其次,可以通过对高频词语出现的频数计算留言详情和答复意见的完整性和可解释性,进行对其标准化处理,最后通过加权平均得到答复意见的质量指数。

关键词: 特征向量化、朴素贝叶斯分类、余弦相似性

Abstract

In recent years, a variety of online political platforms have gradually become an important channel for the government to understand, deal with and solve public problems, and the amount of different data continues to rise. In the past, it has brought great challenges to the work of the relevant departments which mainly rely on manual division and arrangement. At the same time, with the development of related technologies, the establishment of a smart government system based on natural language processing technology has become a new trend of innovative development of social governance, which can greatly promote the management level and efficiency of the government.

This article is based on the records of people's political messages from open sources on the Internet, as well as the responses of relevant departments to some of the messages left by the masses. We intend to use natural language processing and text mining to solve the following three main problems.

With regard to the classification of mass messages, this paper can see from Annex 1 that there are 15 first-level classifications such as urban and rural construction, party affairs and government affairs, land and resources, and so on. According to the seven first-level classifications given in Annex 2, this paper first intercepts the two columns of data of message details and first-level classification, divides the message details, defines a word segmentation dictionary, and removes stop words, and then you can get new message details data. Furthermore, a naive Bayesian classification model with first-level label is established, and the accuracy of the classification model is about 86.5%.

Mining hot issues, first of all, with regard to the description of hot issues, we can first regard the high-frequency words that appear more than a certain number of times as hot issues, segment the topic field of the message and remove the stopped words. Then use the dictionary type to match the topic of the message one by one, count the high-frequency words of the place or crowd, and then sort the message time of the top five. Under the condition of the same place or crowd, the hot issues are described by combining the positive and negative numbers. Secondly, output the message subject corresponding to the similar location, as well as the corresponding message number, user, time and other data.

For the evaluation of reply comments, this paper needs to make a comprehensive calculation and evaluation from three aspects. Firstly, the similarity of message details and response comments can be calculated by cosine similarity, and secondly, by calculating the frequency of high-frequency words, we can calculate the details of messages and the integrity and explainability of responses, and standardize them. Finally, the quality index of responses can be obtained by weighted average.

Keywords: feature vectorization, naive Bayesian classification, cosine similarity

一. 挖掘目标

1.1 挖掘背景

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，能够提高政府办公、监管、服务、决策的智能化水平，形成高效、敏捷、便民的新型政府，是电子政务发展的高级阶段，是提高党的执政能力的重要手段。

此题需要根据自互联网收集的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答复意见，分析群众留言，热点问题，以及答复评价，利用自然语言处理和文本挖掘的方法进行解决。

随着智慧政务的普及，可以让用户提交诉求的时间缩短，让领导能更快的处理解决问题，在推动决策更精准、服务更高效、办事更便捷等方面取得较大进展。

1.2 挖掘目标

根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。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，给出排名前 5 的热点问题，给出相应热点问题对应的留言信息。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、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

二、 符号说明

C_i	留言详情样本所属的类别
P_i	一级标签第 i 类的查准率
R_i	一级标签第 i 类的查全率
n	一级分类的类别数
D	热度指数
d_1	留言高频词的出现次数
d_2	留言的点赞数
d_3	留言的反对数
y	相似度
k	比例系数
θ	夹角
Z	质量指数
m	相似度
v	完整性指数
w	可解释性指数
a	留言详情文本向量
b	答复意见文本向量
F_1	预测的准确率

三. 问题分析

3.1 问题一的分析

对附件 2 的数据进行分析，截取留言详情和一级分类两列数据，对留言详情进行分词、去除停用词，得到新的留言详情数据；将文本特征向量化，利用 CountVectorizer 向量化工具，定义数值；利用朴素贝叶斯分类，把数据拆分为训练集和测试集，建立一个朴素贝叶斯分类器，利用分好词的留言详情数据进行训练，进而根据测试集的留言详情数据预测出测试集的一级分类值。

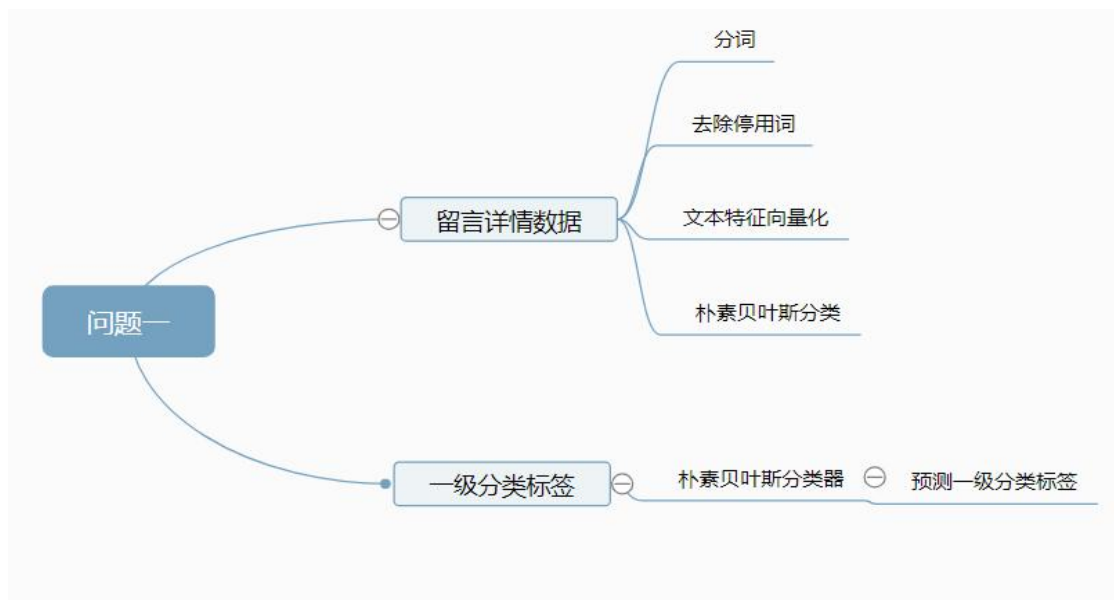


图 1 问题一分析路线图

3.2 问题二的分析

热点问题挖掘的解决，我们首先把留言的高频词出现超过四次的看作为热点问题，因为当高频词出现三次或以下的情况可能是没有完全去除停用词时，这时的高频词就不能计算留言问题出现的普遍次数。定义热度评价指标，即为热度指数=留言主题的高频词出现的次数+留言的点赞数+留言的反对数。然后对留言主题进行提取地点或人群，利用 jieba 库分词然后删除停用词获取到地点或人群，

然后用字典类型的形式把留言问题描述与地点或人群一一对应起来。

通过地点或人群出现前五的高频词找出在附件三对应留言的点赞数和反对数，进而提取，然后用 collections 库里边的 counter 词频计数，统计出地点或人群的词频，最后把地点或人群的词频数加上点赞数再加上反对数即为对应留言的热度指数。

首先对附件三的留言时间数据进行预处理，改成相同的数据类型长度，通过输出到 Excel 表格中，对其前五名的留言时间进行排序，从而得出留言时间范围在最小值与最大值之间。输出前五的热点或人群高频词，在同一地点或人群中，点赞数与反对数之和最大的留言主题即为热点问题描述。

通过对留言主题字段进行分词，去除停用词，然后利用字典索引的方式对留言主题进行一一匹配。统计高频词，根据高频词输出相似地点的留言主题，进而再输出对应的留言编号，留言用户，留言时间等数据。

3.3 问题三的分析

通过余弦相似性来计算留言详情和答复意见的相似度 y ，当余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，夹角 θ 与相似度 y 成反比的关系， k 为比例系数。

$$y = \frac{k}{\theta}$$

通过观察答复意见文本的特点，结合文本的完整性规律发现答复文本开头结尾都会出现的词语，以及具体答复的细节，根据答复意见完整性的规律统计出现词语的频数，例如“你好”，“答复如下”，“感谢”，“谢谢”等。词语的频数对其进行标准化处理，数值越大，其答复意见的完整性普遍就越高。

通过观察答复意见文本的特点，结合文本的可解释性规律发现答复文本绝大多数会出现的词语，根据答复意见可解释性的规律统计出现词语的频数，例如“法律”，“法规”，“调查”，“文件精神”等。词语的频数对其进行标准化处理，数值越大，其答复意见的可解释性普遍就越高。通过相似度、完整性、可解释性加权平均可得到质量指数，综合评价每一条答复意见的质量。

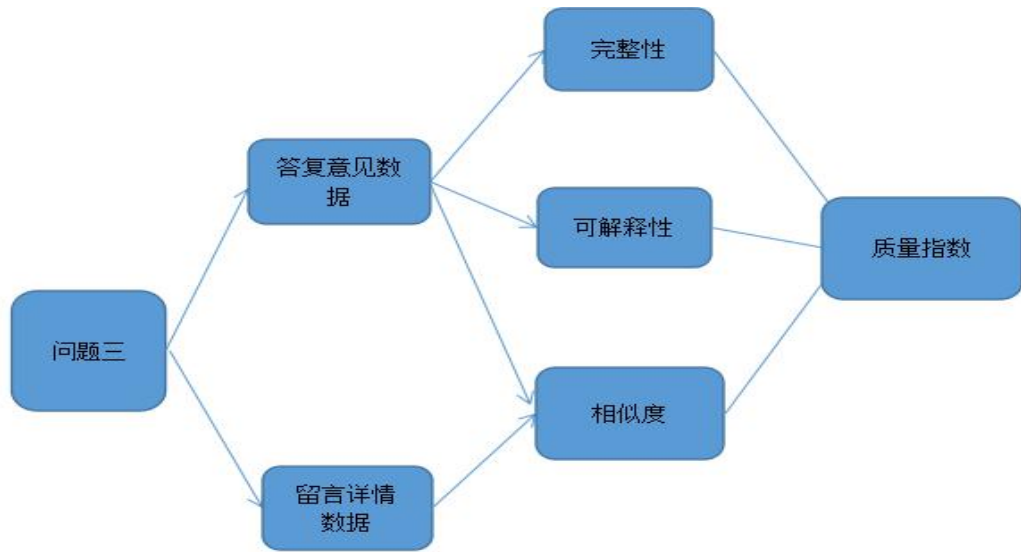


图 2 问题三分析路线图

四. 模型建立与求解

4.1 问题一模型的建立（群众留言分类）

4.1.1 朴素贝叶斯分类模型：

假设数据样本有 n 个留言详情，分别为 F_1 、 F_2 、...、 F_n ，有 m 个一级分类标签，分别为 C_1 、 C_2 、...、 C_m ，留言详情样本所属的类别为 C_i 。

在已知 F_n 的情况下，计算留言详情样本属于哪个一级分类的概率，可以表示为：

$$P(C_i|F_n) = \frac{P(C_i F_n)}{P(F_n)}$$

根据全概率公式，将计算所得的最大概率值 $P(C_m|F_n)$ 对应的一级分类标签作为留言详情样本的最终一级分类，可以表示为：

$$P(C_i|F_n) = \arg \max \frac{P(c_i) * P(Fn|C_i)}{\sum_{i=1}^m P(C_i)P(Fn|C_i)}$$

假设所有留言详情样本都彼此独立，由于 $P(Fn)$ 对于所有的类别都是相同的，所以可表示为：

$$P(C_i|F_n) = p(F_1|C_i)P(F_2|C_i)\cdots P(F_n|C_i)$$

首先先通过分词去除停用词得到新的留言详情数据。

```

0      A3区 大道 西行 道 未管 路口 加油站 路段 人行道路 路灯 杆 圈 西湖 建筑 湖
1      书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电等 后 烂尾 护栏
2      市政府 市 交警支队 市 安监局 市 环保局 A3区 市 A3区 杜鹃 文苑 小区 严
3      胡书记 您好 感谢您 百忙之中 这份 留言 父亲 5.1 A6区 金星 北路 明发 工
4      K8 县 丁字街 乱 摆摊 前段时间 丁字街 交通 好 几天 最近 丁字街 做生意
5      南门 街 前段时间 整改 劝阻 摆摊 占道 情况 改善 情况 好 几天 慢慢 以前
6      现 K8 县冷 江东 路 蓝波 旺 酒店 前面 外墙 装修 搭 架子 无人 施工 路政
7      九亿 城区 景观 点 很漂亮 每到 晚上 人到 玩耍 . 唯有 两个 公厕 却 灯
8      石期 市镇 农贸市场 旁边 公厕 旱厕 里面 脏 乱 差 臭气熏天 老百姓 厕所
9      李 书记 您好 感谢您 阅读 十二五 期间 非 省会 地级市 轨道交通 建设 席卷
10     易 市长 您好 感谢您 阅读 十二五 期间 非 省会 地级市 轨道交通 建设 席卷
11     媒体报道 市 公交 地铁 爱心卡 一卡通 残疾人 爱心卡 乘坐 地铁 刷卡 时 反
12     地铁 号线 施工 导致 万家 丽路 锦楚 星城 小区 三期 一个月 停电 10 来次
13     尊敬 你好 A6区 润 紫 郡 今年年初 小区 竖起 一道道 高压线塔 筑起 高压
14     市 A5 区 朝晖路 锦楚 新城 三区 月份 一共 停电 次 每次 说 原因 停电 线
15     西地省 地区 阴冷 潮湿 气候 近年 气候 逐渐 恶劣 月亮岛 片区 近年 楚江
16     胡书记 冬天 市 湿冷 冬天 真是 受不了 诶 太冷 被子 感觉 潮潮 洗衣服 难
17     尊敬 市委 市政府 市是 一座 名城 一座 幸福感 幸福感 市委 市政府 想民 想
18     K9 县城 公交线路 新 公交车 试运行 中 市民 出行 一项 重大 民生 省市 城
19     K6 县 路 路 公交车 延迟 晚上 21 点 以后 晚上 19 点 以后 路 路 公交车
20     K6 县 公交车 破旧不堪 这是 明显 最让人 愤怒 车人 无 看似 插卡 见过 乘
21     你们好 上周 提交 请求 迎丰 公园 人性 关怀 角度 考虑 延后 清晨 路灯 熄
22     L市 城北 中坡山 国家森林公园 一座 自然 风光秀丽 市民 好去处 遗憾 公园

```

图 3 分词并去除停用词后的留言详情数据

通过建立的朴素贝叶斯分类模型，一级分类有 7 个不同的指标，把它定义为数值 1 到 7, 分别为“城乡建设”：1, “环境保护”：2, “交通运输”：3, “教育文体”：4, “劳动和社会保障”：5, “商贸旅游”：6, “卫生计生”：7, 以示例数据附件

三的数据作为测试得出每条留言的一级分类标签预测值。

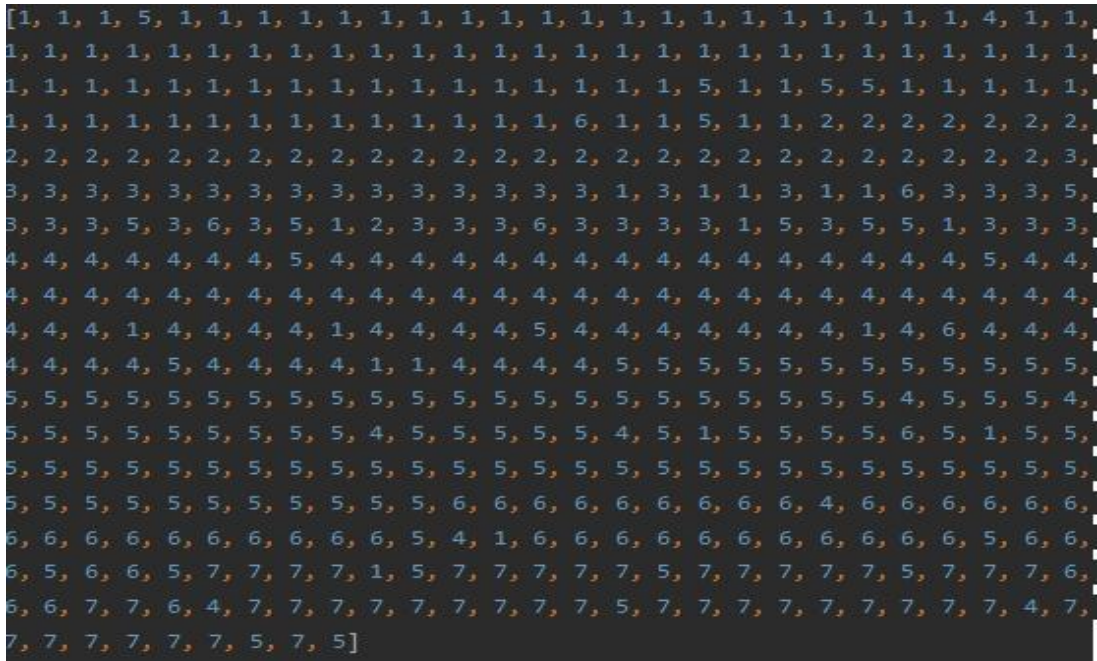


图 4 一级分类标签的预测值

4.1.2 F-score 评价方法的建立:

假设一级分类标签有 i 类，分别是城乡建设，环境保护，交通运输，教育文体，劳动和社会保障，商贸旅游，卫生计生七类一级分类标签， P_i 为第 i 类的查准率， R_i 为第 i 类的查全率， n 为一级分类的类别数，因此

$$F1 = \frac{1}{n} \sum_{i=1}^n \left(2 \frac{P_i R_i}{P_i + R_i} \right)$$

利用 F-Score 方法对朴素贝叶斯分类模型进行评价，进而得出模型的准确率，通过计算得出在**没有去除停用词**时一级分类预测准确率大约为 76.1%。

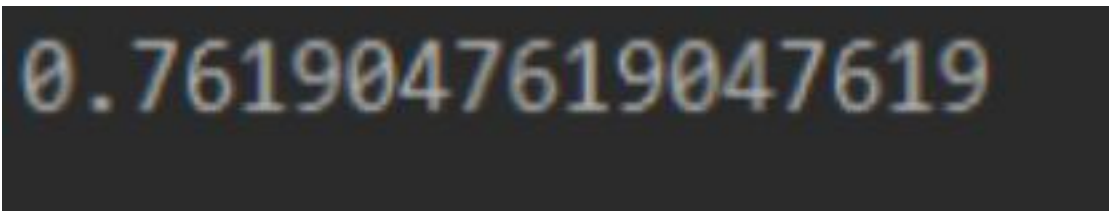


图 5 F-score 模型评价在**没有去除停用词**的准确率

通过计算得出去除停用词后一级分类预测准确率大约为 86.5%。

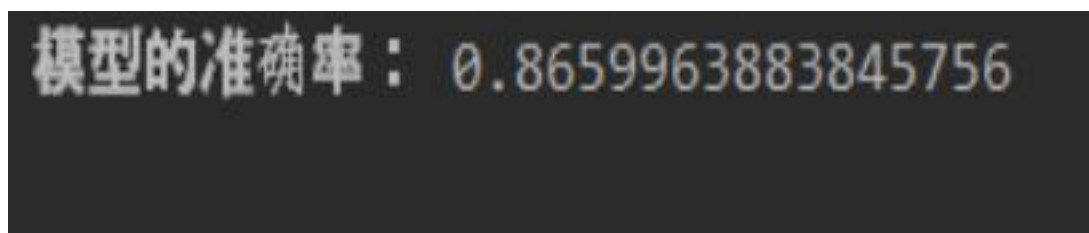


图6 F-score 模型评价在去除停用词后的准确率

4.2 问题二模型的建立（热点问题的挖掘）：

4.2.1 热点问题描述

首先把问题的高频词出现超过四次的看作为热点问题，定义热度评价指标，即热度指数 D ，假设留言主题的高频词的出现次数为 d_1 ，留言的点赞数为 d_2 ，留言的反对数为 d_3 ：

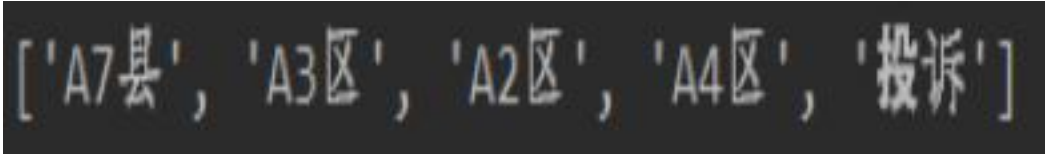
$$D = d_1 + d_2 + d_3$$

然后对留言主题进行提取地点或人群，利用 jieba 库分词然后删除停用词获取到地点或人群，然后用字典类型的形式把留言问题描述与地点或人群一一对应起来。

0	A3区 一米阳光 婚纱摄影 纳税
1	A6区 道路 命名 初步 公示 城乡 门牌
2	A7县 春华 镇金鼎村 水泥路 自来水 到户
3	A2区 黄兴路 步行街 古道 巷 住户 卫生间 粪便 外排
4	A3区 中海 三期 四期 中间 空地 夜间 施工 噪音扰民
5	A3区 泉 单方面 改变 麓谷明珠 架空层 性质
6	A2区 富绿 新村 房产 性质
7	地铁 违规 用工 质疑
8	公交车 随意 变道 通行
9	A3区 保利 谷林语 桐梓 坡路 与麓 松路 交汇处 地铁 凌晨 点 施工 扰民
10	A7县 东四 路口 晚 高峰 太堵 调整 信号灯 配时
11	A3区 青青 家园 乐果 果 零食 炒货 通道 摆放 空调 扰民
12	拆除 聚美龙楚 西地省 商学院 宿舍 旁 安装 变压器
13	市利保 壹号 公馆 夜间 噪声 扰民
14	地铁 号线 星沙 地铁 出入口 不合理
15	A4区 北辰 非法 住 改商 何时能
16	K3县乡村医生 卫生室 执业
17	A7县 春华 石塘 铺村 党员 家开 麻将馆
18	异地 办理 出国 签证
19	投诉 温斯顿 英语 拖延 退费
20	A6区 乾源 停车场 违章 乱建 现象
21	A7县 星城 幢 非法经营 家庭旅馆
22	A2区 佳兆业 水新 垃圾 无人
23	沙坪 街上 有无 证 理疗 馆 骗取 老人 钱财
24	市德鸿 餐饮店 拖欠工资 维权 难

图7 分词并去除停用词后的留言主题数据

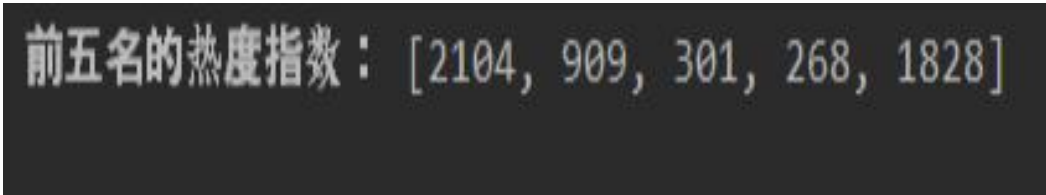
通过地点或人群出现前五的高频词找出在附件三对应留言的点赞数和反对数，进而提取，然后用 collections 库里边的 counter 词频计数，统计出地点或人群的词频。



['A7县', 'A3区', 'A2区', 'A4区', '投诉']

图 8 留言前五的地点或特定人群的高频词

把地点或人群的词频数加上点赞数再加上反对数即为对应留言的热度指数，进而得出前五的热度指数。



前五名的热度指数：[2104, 909, 301, 268, 1828]

图 9 热点问题留言高频词前五名的热度指数

首先对前五名的热度指数进行排序，对附件三的留言时间数据进行预处理，改成相同的数据类型长度，通过输出到 Excel 表格中，对其前五名的留言时间进行排序，从而得出留言时间范围在最小值与最大值之间。

时间范围
2019/1/22至2020/1/8
2019/3/28至2020/1/7
2019/2/28至2020/1/8
2019/8/27至2020/1/6
2019/1/9至2020/1/7

图 10 热点问题留言前五名的时间范围

4.2.2 输出相应的表

输出前五的热点或人群高频词，在同一地点或人群中，点赞数与反对数之和最大的留言主题即为热点问题描述；根据高频词输出相似地点的留言主题，进而再输出对应的留言编号，留言用户，留言时间等数据，整理出热点问题表。

热度排名	问题ID	热度指数	时间范围	地点或人群	问题描述
1	1	2104	2019/1/22至2020/1/8	A7县	A7县留地安置失地农民
2	2	1828	2019/3/28至2020/1/7	A3区	A3区西湖街道茶场村五组启动征地拆迁
3	3	909	2019/2/28至2020/1/8	A2区	A2区丽发新城违规乱建混凝土搅拌站监管
4	4	301	2019/8/27至2020/1/6	A4区	A4区凤舞路新河大厦建筑工地噪音扰民居民
5	5	268	2019/1/9至2020/1/7	投诉	投诉A6区万润滨江天著阻止准正常组合贷

图 11 热点问题表

通过输出高频词出现四次的留言，把同一点或人群的留言归为一个问题，进而根据高频词的次数对其进行问题编号，输出热点问题明细表。

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	188031	A0004006	A7县星沙乐园C1门面油烟排	2019/7/19 18:19:54	本人系春	0	1
1	188251	A0001309	A7县金科代时物业公司违	2019/10/19 11:02:40	近来，下	0	0
1	188451	A0001300	A7县财政评审投资强制算	2019/4/11 17:54:25	我是春华	0	2
1	188535	A0006177	A7县中南汽车乱象丛生	2019/6/13 15:28:44	尊敬的各	0	0
1	188780	A0009475	A7县供销社原泰阳商城丢	2019/10/20 14:59:13	尊敬的县	0	0
1	188799	A0001073	A7县开元东四路交叉口未	2019/5/22 12:24:16	我是橄榄	0	1
1	188876	A0001343	A7县恒基凯旋门万婴格林	2019/2/26 11:29:49	尊敬的领	0	0
1	188887	A0008566	A7县黄兴蓝田新村农村用	2019/7/17 10:19:08	尊敬的领	1	4
1	189864	A0001003	A7县泉星公园开工多长	2019/8/1 12:45:07	路口镇区	0	0
1	190019	A0001042	A7县泉塘E楚水湾有人晚	2019/6/4 18:08:56	领导好，	0	0
1	190033	A0001069	A7县泉塘再一所公办中学	2019/11/3 20:19:18	尊敬的领	0	5
1	190156	A0004057	A7县星沙三区噪音扰民	2019/3/25 11:53:00	本人于20	0	0
1	190346	A0005873	A7县星沙中贸城欺诈骗退还	2019/7/10 22:46:03	我们是A7	0	1
1	190754	A0008998	A7县社塘路白改黑及铁建	2019/3/13 18:03:53	本人家住	0	0
1	190978	A0002565	A7县楚龙街道龙塘安置区	2019/4/8 9:24:39	请问东六	0	1
1	191153	A909097	A7县星沙四区京塘路旧城	2019/12/15 18:43:57	我们是西	0	0
1	191794	A0005420	A7县黄兴大塘组村民家耕	2019/3/11 17:32:46	A7县泉塘	0	2
1	191885	A0001040	A7县江背农村集体经济界	2019/10/27 13:33:33	你好，沈	0	0
1	191993	A0008352	A7县色情行业泛滥酒店色	2019/5/7 9:57:30	A7县城区	0	1
1	192029	A0002484	A7县楚龙路餐馆音乐声超	2019/8/18 18:24:25	孩子马上	0	0
1	192043	A0005190	A7县中石化加油站爱义行	2019/5/13 10:45:14	请问A7县	0	1
1	192366	A0001035	A7县右拐远大路黑漆漆无	2019/4/30 17:17:00	社塘路东	0	2
1	192493	A0001039	A7县高桥镇范琳村路灯不	2019/3/12 10:07:21	昨天路过	0	0
1	192636	A0001799	A7县黄集镇合心村灌溉山	2019/7/7 11:55:14	张县长	0	1
1	192685	A909099	A7县泉塘小广播扰民	2019/11/10 12:56:27	领导好，	0	1
1	192898	A0002033	A7县人民法院公正处理楚	2019/12/28 20:00:56	星沙吾悦	0	1
1	193056	A0003267	A7县撤县改区进度市辖区	2019/12/9 21:33:26	建议雅韵	0	3
1	193286	A0001031	A7县留地安置失地农民	2019/4/17 11:13:12	沈书记：	0	32
1	193337	A0008034	A7县春天物业放任垃圾堆	2019/10/16 13:03:13	泉星公园	0	12
1	193680	A0009244	A7县福临冯家坡村菜水塘	2019/7/8 15:17:54	尊敬的A7	0	0
1	193723	A0008032	A7县星沙灰埠16垃圾站恶	2019/6/20 23:44:24	809路方	0	1
1	193788	A0005335	A7县星沙徐记海鲜强势霸	2019/9/9 11:16:56	近年来，	0	1
1	193972	A0008434	A7县华广大厦购房合同天	2019/2/15 10:21:40	我是A7县	0	0
1	194022	A0004210	A7县妇幼保健院急诊门前	2019/7/8 10:39:54	我们是诺	0	1
1	194168	A0003083	A7县智慧桥幼儿园拒绝普	2019/8/4 22:12:46	举报：A7	0	3
1	194204	A0009523	A7县楚绣城c1羊肉串太吵	2019/12/12 16:11:00	黄花收费	0	0
1	194373	A0003100	A7县星沙一桥早堵晚不见	2019/12/16 13:24:36	问问A7县	0	15
1	194554	A0007545	A7县国泰九龙湾2000住户	2019/11/4 21:07:28	三一大道	0	0

图 12 部分热点问题明细表

4.3 问题三模型的建立（答复意见的评价）：

4.3.1 答复意见的相似性：

首先进行分词，并列出所有的词，计算词频之后写出词频向量。假设 a, b 是留言详情和答复意见的文本向量，利用余弦定理可知

$$\cos \theta = \frac{a^2 + b^2 - c^2}{2ab}$$

假定 a 向量是 [x1, y1]，b 向量是 [x2, y2]，则可以写成以下形式：

$$\cos \theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} + \sqrt{x_2^2 + y_2^2}}$$

假定 A 和 B 是两个 n 维向量，A 是 [A1, A2, ..., An]，B 是 [B1, B2, ..., Bn]，则 A 与 B 的夹角 θ 的余弦就等于：

$$\cos \theta = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} + \sqrt{\sum_{i=1}^n (B_i)^2}} = \frac{A \cdot B}{|A| \times |B|}$$

当余弦值越接近 1，则夹角越接近 0 度，也就是两个向量越相似，夹角 θ 与相似度 y 成反比的关系，k 为比例系数。

$$y = \frac{k}{\theta}$$

通过附件 4 给出的数据根据余弦相似性得出每条答复意见与留言详情的相似度如下：

答复意见与留言详情的相似度: [0.618, 0.712, 0.703, 0.822, 0.787, 0.799, 0.622, 0.782, 0.803, 0.783, 0.476, 0.663, 0.736, 0.852, 0.569, 0.731, 0.66, 0.863, 0.78, 0.783, 0.826, 0.9, 0.815, 0.848, 0.9, 0.785, 0.783, 0.645, 0.725, 0.84, 0.847, 0.849, 0.643, 0.853, 0.844, 0.717, 0.763, 0.884, 0.856, 0.84, 0.816, 0.809, 0.732, 0.859, 0.708, 0.841, 0.59, 0.48, 0.711, 0.699, 0.867, 0.662, 0.698, 0.779, 0.857, 0.699, 0.728, 0.857, 0.802, 0.838, 0.761, 0.772, 0.509, 0.702, 0.846, 0.655, 0.858, 0.412, 0.811, 0.774, 0.649, 0.888, 0.758, 0.762, 0.653, 0.818, 0.875, 0.812, 0.845, 0.573, 0.85, 0.892, 0.719, 0.735, 0.796, 0.508, 0.679, 0.707, 0.82, 0.798, 0.797, 0.781, 0.792, 0.792, 0.845, 0.811, 0.805, 0.703, 0.529, 0.655, 0.87, 0.845, 0.423, 0.873, 0.865, 0.681, 0.79, 0.879, 0.844, 0.829, 0.86, 0.856, 0.801, 0.853, 0.79, 0.889, 0.724, 0.776, 0.887, 0.553]

图 13 每条答复意见与留言详情的相似度

4.3.2 答复意见的完整性:

通过观察答复意见文本,结合文本的完整性规律,筛选出答复文本的开头以及结尾都会出现的高频词语,例如“你好”,“答复如下”,“感谢”等,对其进行统计频数,并标准化处理,数值越大,则其答复意见的完整性指数普遍就越高,进而得出每条答复意见的完整性指数。

答复意见的完整性: [1.0, 0.286, 0.286, 0.143, 0.571, 0.286, 0.286, 0.143, 0.143, 0.286, 0.143, 0.286, 0.286, 0.286, 0.429, 0.286, 0.286, 0.0, 0.0, 0.286, 0.286, 0.0, 0.143, 0.143, 0.0, 0.143, 0.143, 0.429, 0.571, 0.286, 0.143, 0.429, 0.286, 0.286, 0.286, 0.286, 0.286, 0.571, 0.286, 0.143, 0.143, 0.143, 0.286, 0.143, 0.143, 0.143, 0.571, 0.143, 0.429, 0.0, 0.143, 0.143, 0.143, 0.286, 0.143, 0.143, 0.143, 0.286, 0.286, 0.143, 0.286, 0.143, 0.143, 0.143, 0.0, 0.286, 0.286, 0.143, 0.429, 0.286, 0.143, 0.286, 0.143, 0.143, 0.143, 0.429, 0.143, 0.143, 0.143, 0.143, 0.286, 0.143, 0.143, 0.143, 0.143, 0.143, 0.143, 0.429, 0.429, 0.143, 0.0, 0.143, 0.143, 0.143, 0.143, 0.143, 0.143, 0.143, 0.143, 0.143, 0.0, 0.143, 0.0, 0.0, 0.143, 0.0, 0.286, 0.286, 0.286, 0.286, 0.286, 0.143, 0.143, 0.571, 0.143, 0.286, 0.143, 0.0, 0.143]

图 14 每条答复意见的完整性指数

4.3.3 答复意见的可解释性:

通过观察答复意见文本,结合文本的可解释性规律,可以看出,答复文本均为根据…调查,得出…结果,并进一步解释,通过…法律法规,…条例等的格式,对这些高频词语进行统计频数,并进行标准化处理,数值越大,则其答复意见的可解释性普遍就越高,进而得出每条答复意见的可解释性指数。

答复意见的可解释性: [0.714, 0.143, 0.571, 0.0, 0.0, 0.0, 0.143, 0.0, 0.0, 0.0, 0.143, 0.0, 0.143, 0.143, 0.0, 0.143, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.286, 0.0, 0.143, 0.286, 0.0, 0.0, 0.143, 0.143, 0.143, 0.0, 0.286, 0.0, 0.0, 0.0, 0.0, 0.143, 0.0, 0.0, 0.286, 0.0, 0.0, 0.0, 0.286, 0.143, 0.857, 1.429, 0.143, 0.143, 0.0, 0.429, 0.0, 0.0, 0.857, 0.0, 0.143, 0.0, 0.0, 0.0, 0.0, 0.143, 0.143, 0.143, 0.0, 0.286, 0.143, 0.571, 0.571, 0.0, 0.143, 0.143, 0.143, 0.143, 0.286, 0.143, 0.143, 0.0, 0.286, 0.0, 0.0, 0.0, 0.286, 0.286, 0.143, 1.143, 1.429, 0.857, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.143, 0.0, 0.143, 0.0, 1.143, 0.429, 0.0, 0.0, 0.0, 0.0, 0.0, 0.143, 0.143, 0.143, 0.0, 0.571, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

图 15 每条答复意见的完整性指数

4.3.4 综合评价方案:

通过计算出的答复意见相似度 m、完整性指数 v、可解释性指数 w 进行加权平均得出每一个答复意见的质量指数 Z,质量指数在[0,1]之间,加权平均得到的数值越大,其答复意见的质量越高。

$$Z = \frac{m + v + w}{3}$$

根据附件 4 给出的数据用综合评价方案计算出每条答复意见的质量指数。

答复意见的质量指数: [0.777, 0.38, 0.52, 0.322, 0.453, 0.362, 0.35, 0.308, 0.315, 0.356, 0.254, 0.316, 0.388, 0.427, 0.333, 0.387, 0.315, 0.288, 0.26, 0.356, 0.371, 0.3, 0.319, 0.426, 0.3, 0.357, 0.404, 0.358, 0.432, 0.423, 0.378, 0.474, 0.31, 0.475, 0.377, 0.334, 0.35, 0.533, 0.381, 0.328, 0.415, 0.317, 0.339, 0.334, 0.379, 0.376, 0.673, 0.684, 0.428, 0.281, 0.337, 0.411, 0.28, 0.355, 0.619, 0.281, 0.338, 0.381, 0.363, 0.327, 0.349, 0.353, 0.265, 0.282, 0.377, 0.409, 0.381, 0.471, 0.556, 0.306, 0.359, 0.391, 0.348, 0.349, 0.456, 0.368, 0.387, 0.318, 0.425, 0.286, 0.331, 0.345, 0.383, 0.388, 0.361, 0.598, 0.75, 0.569, 0.416, 0.409, 0.313, 0.26, 0.312, 0.312, 0.377, 0.318, 0.364, 0.282, 0.605, 0.409, 0.338, 0.282, 0.189, 0.291, 0.288, 0.322, 0.311, 0.436, 0.377, 0.562, 0.382, 0.381, 0.315, 0.332, 0.454, 0.344, 0.337, 0.306, 0.296, 0.232]

图 16 每条答复意见的综合评价质量指数

五. 总结

本文的主要目的利用数据挖掘与数学建模技术建立有关“智慧政务”的一系列相关模型。

首先,通过对群众的留言建立一级标签分类模型进行分类,为热点问题的挖掘做准备。先对留言详情进行分词,得到新的留言数据,然后利用文本特征向量化之后对其进行朴素贝叶斯分类,进而预测出对应的一级分类标签值;

其次,通过对地点或人群的词频计数以及点赞数反对数定义热度指标,输出前 5 高频地点或人群,根据高频词输出相似地点的留言主题,进而得以输出其他对应的数据;

最后,通过分析计算得到答复意见的综合评价方案。先通过余弦相似性计算答复意见的相关性,其次通过答复意见文本的特点,利用词频统计对其进行标准化处理,计算答复意见的完整性指数以及可解释性指数,最后对三个因素进行加权平均得出答复意见的质量指数。

本文所建立的模型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减轻了负担,对提升政府的管理水平和施政效率也具有极大的推动作用。

六. 参考文献

- [1] 刘顺祥. 从零开始学 Python 数据分析与挖掘[M]. 北京:清华大学出版社, 2018. 76-271
- [2] 程婧, 刘娜娜. 一种低频词词向量优化方法及在短文本分类中的应用[J]. 计算机科学, 2020
- [3] 梁志剑, 谢红宇, 安卫刚. 基于 BiGRU 和贝叶斯分类器的文本分类[J]. 计算机工程与设计, 2020
- [4] 许甜华, 吴明礼. 一种基于 TF-IDF 的朴素贝叶斯算法改进[J]. 计算机技术与发展, 2019
- [5] 刘彧. 基于贝叶斯理论的文本分类技术的研究与实现[C]. 吉林大学, 2009

- [6] 简杨君. Python 中文分词及词频统计[EB/OL]. <https://www.jianshu.com/p/7ad0cd33005e>
- [7] baidu-liuming. 带你彻彻底底搞懂朴素贝叶斯公式[EB/OL]. <https://blog.csdn.net/fishermin/article/details/79509025>
- [8] 王树义. 用 Python 提取中文关键词[EB/OL]. <https://www.jianshu.com/p/cf383fd471bb>