“智慧政务”中的文本挖掘与应用

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。因此，运用数据挖掘技术和自然语言处理技术提升政府的管理水平和施政效率具有重大的意义。

针对问题 1，通过 class_id 将一级标签转换成 id（0~6）。利用 jieba 中文分此工具对留言主题信息进行分此，并通过 TF-IDF 算法量化文本。通过比较逻辑回归、多项式朴素贝叶斯、线性支持向量机和随机森林 4 种模型的分类效果，确定采用线性支持向量机模型。最后利用 F1 值评估线性支持向量机模型。

针对问题 2，通过对文本进行处理后，有效的按照文本重复率进行分类，可以获得目前留言的主要问题。再对信息进行有效挖掘，使得民众留言得到最大化效率。

针对问题 3，我们选择对答复实效性与答复相关性进行研究，通过对相关文献的参考，我们利用 with open 函数选取了相关列进行研究。通过利用指针进行文本相关率研究。对于答复实效性，我们通过 DataFrame 与 DateTime 函数进行时间差计算，得到相关分析结果，最后利用簇状统计图对该程序答复质量进行评价。

关键词 支持向量机

1、挖掘目标

本次建模目标是利用线性支持向量机对群众留言进行分类、

达到以下三个目标：

利用 TF-IDF 算法和文本分词方法对非结构化的文本数据进行挖掘，根据 F1 值的评估结果评价分类模型的性能。

方便后续工作人员方便快捷的定义，合理的对留言进行分类。提高有关部门的工作效率，同时可以最大程度的了解到民生目前面临的主要问题。

对相关部门对留言的答复意见质量进行评估，并实现智慧政务答复系统评价方案。

2、挖掘方法与过程

2.1 问题 1 的挖掘方法与过程

2.1.1 数据分析

2.1.1.1 标签类目数量和分布

将题目给出附件 2.xlsx 文件转换为附件 2.csv 文件，可以方便后续建模的一系列处理。附件 2.scv 有 9210 条记录，7 个标签。标签内容及对应的文本数量如表 1，分布如图 1。

表 1:标签内容及对应的文本数量

一级标签	count
城乡建设	2009
劳动和社会保障	1969
教育文本	1589
商贸旅游	1215
环境保护	938
卫生计生	877
交通运输	613

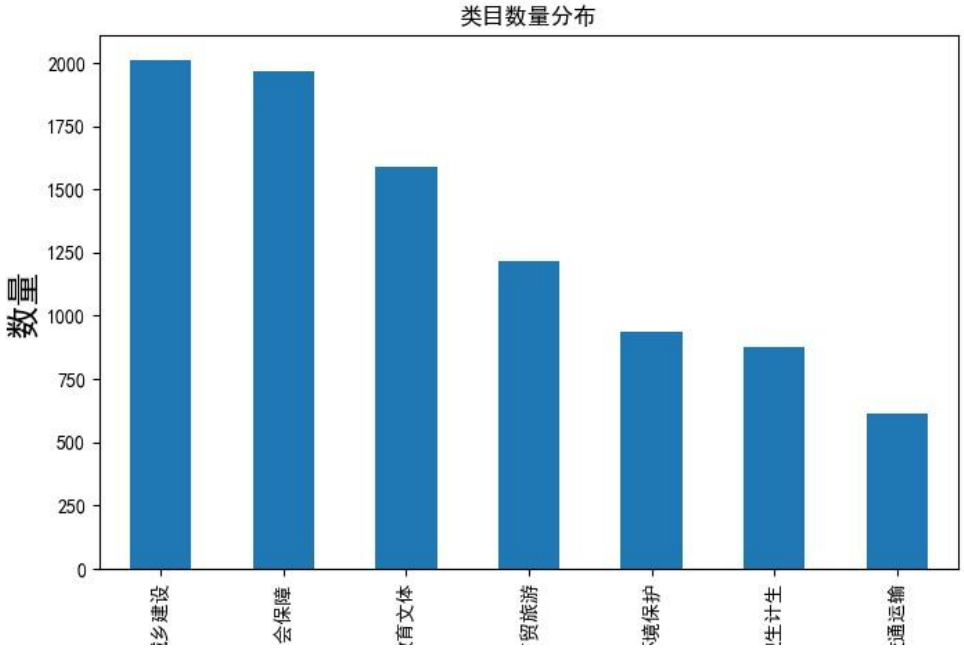


图 1:标签类目数量分布图

由图 1 所示，城乡建设与交通运输类文本数量差异较大，城乡建设类文本数量超过交通运输类的 2 倍，分布不是很均匀。

2.1.2 数据预处理

2.1.2.1 留言文本的去空

建模之前，需要对题目给出的数据进行数据预处理。将附件 2.xlsx 转换为附件 2.csv 文件后，运用 isnull()函数对一级标签和留言主题判空后，去空。虽然这两列没有空值，但是鉴于问题 3 的研究，这里不建议去重。

2.1.2.2 一级标签转换成 class_id

将附件 2 表格中的一级标枪转换成 class_id，便于以后的分类模型的训练。一级标签转换结果如表 2。

表 2：一级标签转换成 class_id

一级标签	class_id
城乡建设	0
劳动和社会保障	4
教育文本	3
商贸旅游	5
环境保护	1
卫生计生	6
交通运输	2

2.1.2.3 对留言主题进行中文分词

在对留言主题文本进行挖掘分类之前，先要把非结构化的文本信息转换为计算能够识别的结构化信息。在附件 2 表中，以中文文本的方式给出了数据。为了便于转换，先要删除文中标点符号、无意义的常用词得到 clean_theme，再对对其中的留言主题进行中文分词得到 cut_theme。

本文采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于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。另外，采用停用词列表对文本进行过滤，例如吧，吗，呢等一些高频但无法反应出文本主要意思的常用词。

在分词的同时，采用了 TF-IDF 算法，将文本向量化。

2.1.2.4TF-IDF 算法[1]

在对留言主题文本分词后，需要把这些词语转换为向量，以供挖掘分析使用。本文采用 TF-IDF 算法，把留言主题文本转换为权重向量，在单词计数的基础上，降低了常用高频词的权重，增加罕见词的权重。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建 立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本 中的分布越集中，说明该分词在区分该文本内容属性能力越强。

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）× 逆文档频率（IDF）

实际分析得出 TF-IDF 值与一个词在留言主题文本中出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。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主题文本的关键词。

2.1.2.5 计算 cut_theme 的 TF-IDF 特征值

文本使用 sklearn.feature_extraction.text.TfidfVectorizer 方法来抽取文本的 TF-IDF 的特征值，用参数 ngram_range=(1,2)表示除了抽取评论中的每个词语外,还要抽取每个词相邻的词并组成一个“词语对”,如: 词 1，词 2，词 3，词 4，(词 1，词 2)，(词 2,词 3)，(词 3，词 4)。如此就扩展了特征集的数量,利于提高分类文本的准确度。另外，用参数 norm='l2'———一种数据标准划处理的方式，将数据限制在一点的范围内，比如说(-1,1)。部分文本的向量如表 3.

表 3:TF-IDF 特征值

(9210,567893)	
(0, 34783)	0.38992145246512155
(0, 55129)	0.38992145246512155
(1, 23692)	0.3107737611568129
(1, 22011)	0.3107737611568129
(2, 52506)	0.34968334638276066

(9208, 3861)	0.23492267442427844
(9209, 13951)	0.20278564955724215

如表 3 所示，features 的维度是(9210,56783)，这里的 9210 表示总共有 9210 条留言主题信息，56783 表示特征数量，包括全部留言主题中的所有词语数+词语对(相邻两个单词的组合)的总数。记录(0,34783) 0.380.38992145246512155，表示某条 0 类（城乡建设类）留言主题提取的词汇表为 34783，特征值为 0.38992145246512155。

2.1.2.6 各分类中关联度最大的两个词语和两个词语对

采用卡方检验和 sklearn 中的 chi2 方法找出每个分类中关联度最大的两个词语和两个词语对，如表 4。

表 4：一级标签的最相关词语与最相关词语对

一级标签	最相关词语	最相关词语对
城乡建设	公积金	拖欠 工程款
	房产证	住房 公积金
劳动和社会保障	职工	社保 问题
	社保	退休 人员
教育文本	补课	教师 招聘
	教师	培训 机构
商贸旅游	电梯	传销 组织
	传销	小区 电梯
环境保护	排放	严重 污染
	污染	污染 严重
卫生计生	独生子女	再婚 家庭
	医院	人民 医院
交通运输	的士	出租车 管理
	出租车	滴滴 出行

由表 4 所示，各类一级标签对应的最相关词语和最相关词语对比较合理。

2.1.3 群众留言分类

分类过程中，本问不考虑标签之间的关联性。

2.1.3.1 文本分类器的选择

本文尝试逻辑回归、多项式朴素贝叶斯、线性支持向量机和随机森林 4 种模型，并使用箱型图评估该 4 中模型的准确率。

首先训练朴素贝叶斯分类器，编写 myPredict()函数，预先观察下分类效果，如表 5。

表 5：朴素贝叶斯分类结果

留言主题文本	分类结果
A 市西湖建筑集团占道施工有安全隐患	城乡建设
E 市崑山培英学校乱收费，一涨就一千	教育文体
D 市一滴滴出行司机无网约车驾驶证	交通运输

L5 县医院的药品以次充好流入县内各大医院	卫生计生
J9 县沙田的化工厂、砖厂的污染破坏性强	环境保护
咨询小孩城乡居民医保政策	劳动和社会保障
E10 县万塘乡瑞康家庭生活馆诱骗老人购买昂贵的仪器	商贸旅游

贝叶斯分类的结果不错。

评估 4 种模型准确率

表 6:4 种模型准确率

模型名称	准确率
线性支持向量机	0.808035
逻辑回归	0.730510
多项式朴素贝叶斯	0.642562
随机森林	0.363518

由表 x6 所示，随机森林准确率最低，因为随机森林属于集成分类器，一般不适合处理高纬度数据（如文本数据），另外 3 个分类器的平均准确率都在 60%以上。其中，线性支持向量机的准确率最高，本文选择线性支持向量机为留言主题分类的模型。

2.2 问题 2 的挖掘方法与过程

2.2.1 热点问题的挖掘

2.2.1.1.过程：

（1）数据预处理：

空白信息：在民生留言中，针对是否存在空的留言信息进行了排查。

无用信息：在留言中会使用很多标点符号，例如空格，标点符号等，这些再我们进行分词统计的时候都属于无用信息

（2）提取数据：

针对附件三中的民众留言问题，单独提取出相应的留言板块，方便后续的留言分析与对应词云生成。

(3)通过清洗器对数据进行清洗：

通过清洗器可以循环读出存储在 csv 表格中的留言内容，注意清晰后再存入文档，方便分此阶段的 jieba 分词。

(4)对数据进行分析：

利用 jieba 进行中文分词，使用,atplotlib 生成词云图片，使用 pandas 对 csv 进行写入，有效避免溢出，同时使用了 PIL 搞笑图片处理库，自定义背景图片模板，以及使用 re 正则表达式，对数据进行清理和提取

如下图可以相对主观的看到哪个问题的热度比较高（相对来说比较直观）

（5）结果：

1 扰民 3311
2 施工 1672
3 噪音 1214
4 投诉 933
5 西地 897
6 公司 755
7 地铁 728
8 咨询 688
9 建议 688
10 社区 641

思路：在数据里面找到反应统一问题的相似的留言，进行总结问题的归并。

通过分析可以得出，前五名最受关注的问题关键词为：

关键词排名	关键词	相关问题出现次数
关键词 1	扰民	3311
关键词 2	施工	1672
关键词 3	噪音	404
关键词 4	投诉	933
关键词 5	西地	897

热度排名	时间范围	地点/人群	问题描述
1	2019/08/07 至	居民区/居民	施工，投诉，扰民，噪音

	2019/10/21		
2	2017/08/09 至 2019/09/01	城市/居民	街道，地铁，咨询，建议
3	2017/06/12 至 2018/04/11	生活区/上班	建议，业主，公司，公交
4	2018/09/07 至 2019/01/03	家属楼/学校居民	西地，建设，夜间，线
5	2017/09/03 至 2018/04/13	生活区/父母	安全，幼儿，老师，费用

我们在此处列举了五个留言较多的问题，同时针对这五类问题各自挑选了具有代表性的例子。

（1）热点问题 1：

以关键词排名第一的扰民为例，通过 excell 软件筛选出与扰民相关的留言内容后，再进行对留言内容归并，得到了如下：

1	施工	926
2	投诉	711
3	西地	672
4	附近	565
5	反映	563

可以看出施工是主要扰民的问题，再通过 excell 表格对施工相关内容进行筛选

1	投诉	731
2	西地	680
3	地铁	595
4	公司	573
5	咨询	516

由此可以看出针对施工，扰民，工地，噪音等问题是民生目前面临的最主要的热度问题 1。通过对投诉问题进行筛选后，筛选部分相关问题。在热点问题留言明细表中，分别给出了 5 个热点问题的相关例子。

（2）热点问题 2：

针对施工进行单独分类排除热点 1 后，我们对热点 2 施工进行讨论分析：以下为用户热度排名第二的一些主要问题：

1	街道	1169
2	公司	946
3	地铁	899
4	咨询	860
5	建议	860

（3）由图我们可以看出很多用户针对热点 2 施工方面有很多困扰，以下我们列举出了几个关于热点问题 2 的留言例子：热点问题三：通过对热点问题 2 进行筛选后，对剩余的留言进行了分类，可以得到剩余的关键词数量：

1	建议	1456
2	业主	1325
3	公交	1311
4	公司	1296
5	苑	1237

（4）同理可以得到热点问题 3，4，5 以及相应的例子。

2.3 问题 3 的方法与过程

2.4.1 数据分析

通过对题目附件 4.xlsx 的分析可得，答复意见的质量主要由答复时效、答复完整性、答复相关性等方面决定，于是，我们选取了答复时效性与相关性进行研究。

主要研究列如下[2]：

表 x:

留言主题	留言详情	留言时间	答复时间	答复意见
A2 区景蓉华苑物业管理有问题	2019 年 4 月以来，位于 A 市 A2 区桂....	2019/4/25 9:32:09	2019/5/10 14:56:53	向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”...
关于 A 市公交站点名称变更的建议	2019 年 4 月以来，位于 A 市 A2 区桂花....	2019/4/23 17:03:19	2019/5/9 9:49:42	关于来信人建议“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”...

2.4.2 答复时效性研究

2.4.2.1 数据预处理

首先利用 pycharm 读取原始文件，去掉与答复时效性无关的其他列，只选取留言时间与答复时间两列。利用 Datetime 函数与 str 函数转化原文件中的时间属性，并存在新的 DataFrame 中。

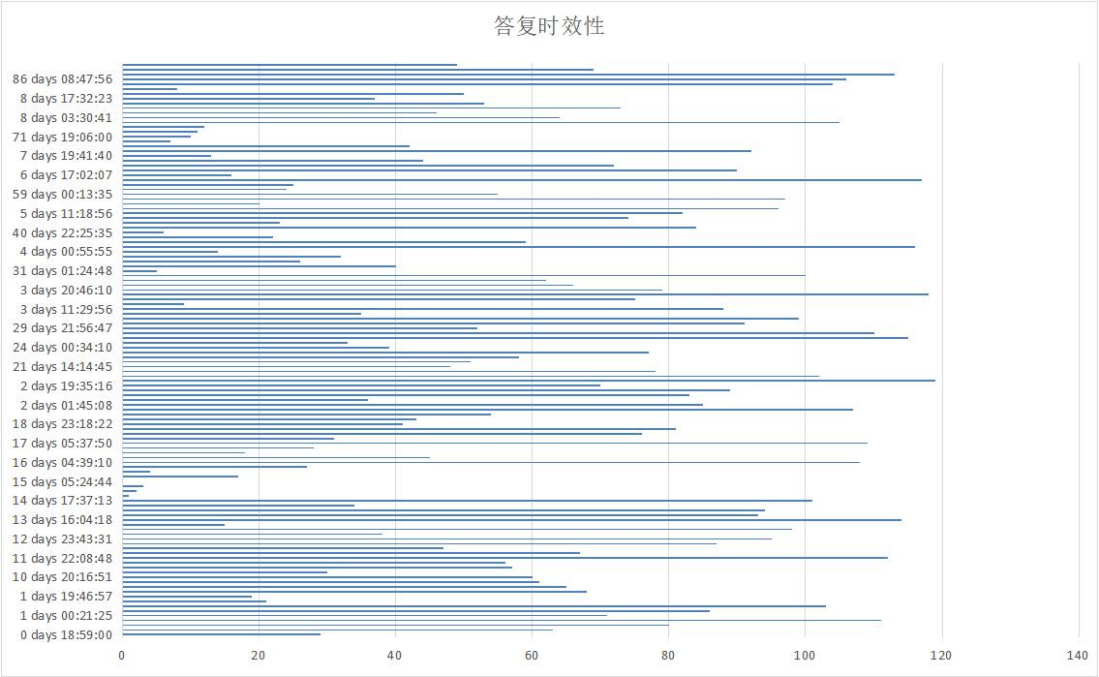
2.4.2.2 答复时效性方法实现

通过分析，我们认为时效性指留言时间与答复时间的差值，故利用 python 代码对这一模块进行了编程。通过 map(lambda x : str_time(x)) 与 data['result'] = data['t2'] - data['t1']语句实现时间之间的差值计算。最后利用 data.to_excle 函数输出计算后的新文件并进行储存。

使用新生成的 excle 文档，对时差进行升序排列，并利用 excle 生成簇状统计图。分析答复时效性。

2.4.2.3 答复时效性结果评估

通过 excle 生成的簇状统计图，我们可以发现，该系统总体答复时差大部分处于 10 天以上，最长时差为 86 天，最短时差为 18:59:00，故而该答复时效性比较差，且不稳定。



2.4.3 答复相关性研究

通过对相关数据的分析，我们认为，对于答复相关性的研究，可以从留言主题与留言详情之间分词的重叠性入手，即当居民提出问题时的关键词句与答复意见关键词重叠比率进行研究。

2.4.3.1 答复相关性评价方案与可能结果评估

对于分词问题，可以利用 python 的 jiaba 库进行中文分词，为了防止干扰结果准确率，故而不进行停用词的过滤。利用指针进行分词的自动匹配，找出最长重叠串并进行结果统计。对每个主题进行循环，取 KNN 均值来判断该系统相关率的大小。

如果相关率<30%，则说明该系统答复相关性较差；当相关率>30%，且<60%时，说明答复相关性一般；当相关率>60%时，说明该系统答复相关性较好。

3、结果分析

3.1 问题 1 的结果分析

3.1.1 线性支持向量机的模型评价

3.1.1.1 混淆矩阵

采用线性支持向量机分类后的混淆矩阵如图 y2

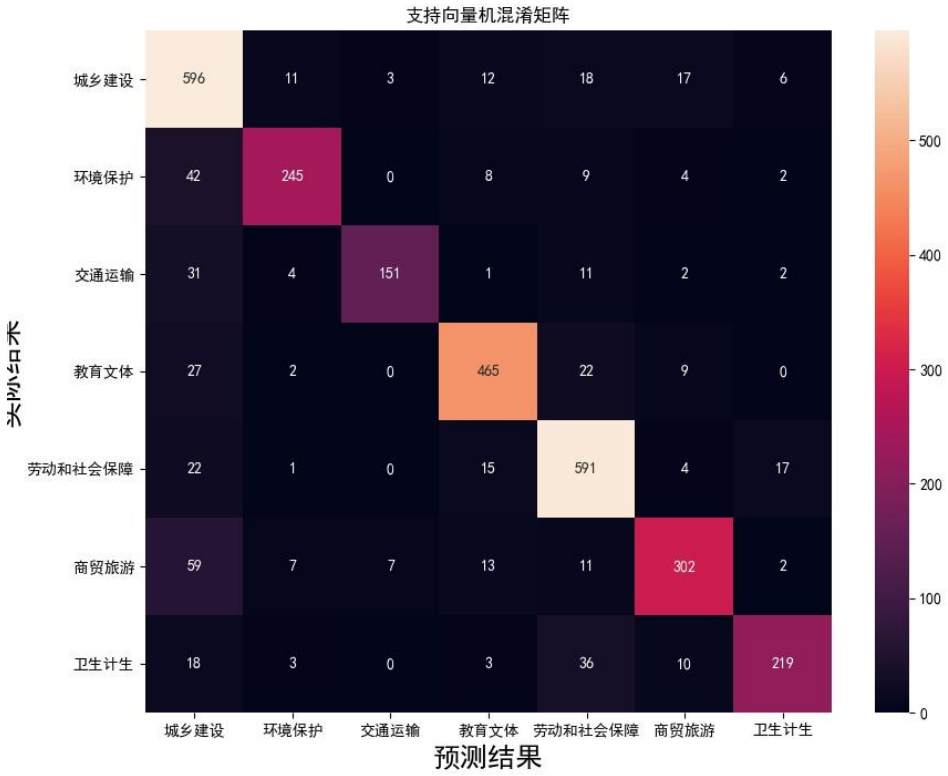


图 y2:线性支持向量机的混淆矩阵

混淆矩阵主对角线表示预测正确的数量，除主对角线外其余都是预测错误的数量。由图 y2 所示，交通运输类预测最准确，只有 10 例错误

3.1.1.2F1 值评估模型预测结果

多分类模型一般不使用准确率(accuracy)来评估模型的质量,因为 accuracy 不能反应出每一个分类的准确性。当训练数据不平衡(有的类数据很多,有的类数据很少)时，accuracy 不能反映出模型的实际预测精度，本文借助于 F1 值指标来评估模型，如表 x7。

表 x7：F1 值评估线性支持向量机模型

一类标签	查准率	查全率	F1	support
城乡建设	0.75	0.90	0.82	663
环境保护	0.90	0.79	0.84	310
交通运输	0.94	0.75	0.83	202
教育文体	0.90	0.89	0.89	525
劳动和社会保障	0.85	0.91	0.88	650
商贸旅游	0.87	0.75	0.81	401
卫生计生	0.88	0.76	0.82	289

由表 x7 所示，教育文体类的 F1 值最大（0.89），商贸旅游类的 F1 值最低，但是有 0.81，模型整体还是不错的。

4、结论

对群众留言信息进行分类挖掘研究，了解各类社情民意，对广大民众有重大意义，同时也是文本分析的一个课题、一个难题。传统的人工留言划分和热点整理已经不能满足数据量庞大的群众留言文本数据。本文采用根据 TF-IDF 算法和线性支持向量机，对群众留言进行合理的分类，以便后续将群众留言分派至相应的职能部门。

5、参考文献

[1] 张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.
[2] 张黄群.　如何撰写稿件答复意见[C].　.学报编辑论丛（第十一集）.:华东地区高校学报研究会,2003:232-233.