2020 年“泰迪杯” 数据分析职业大赛

参

赛

作

品

作品名称：C 题：“智慧政务”中的文本挖掘应用

文本挖掘在“智慧政务”中的应用

摘要

近年来，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，基于互联网公开来源的群众问政留言数据。而这些数据的统计与分析，有助于政府更高效、更快速地开展针对工作，对于提高市政管理的效率有着重要的意义。

针对问题 1，首先利用 python 将附件 2 中的数据进行一定的数据清洗，留言去重，去换行符，以及将留言中无意义的常用词（stopword）进行删除，方便进一步的数据处理。然后通过统计各一级标签的数据量，得到各类别的占比，再采用随机抽样的方式，对七个一级标签里的数据进行等量数据提取。再通过平滑聚集、数据概化、规范式等方式将数据转换成适用于数据挖掘的形式。再通过 PyEchart 和 jieba 分词处理，建立停用词表，过滤标点以及特殊符号，提取高频词来绘制词云图，通过词云图分析，预估各大一级标签的留言内容的特点。然后，将数据重新排序，寻找重叠词并将其转换为具体的概率值，再进行权重词处理，进行归类并可视化，并进行简单的结果分析。最后，选用 KNN 分类回归模型、线性支持向量机 SVM 模型、多项式朴素贝叶斯分类模型三个模型进行模型构建，并利用 F-Score 算法分别对三个模型评价，选用评价结果最优的模型。

针对问题 2，先从时间维度上分析，通过 python 将时间转化为时间戳并将其设置为索引，进行归类聚类分析，最后进行数据合并，筛选出排名前 5 的热点问题。然后，对热点问题进行有针对性地处理，进行留言统计，从多维角度进行分析，得到“热点问题留言明细表”。

针对问题 3，通过相关性、完整性、可解释性等多维度的构思，设计一套具体的针对答复意见质量的评价方案。

关键词：分类标签　模型构建与评价　热点问题　评价方案

目录

1 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。

1.2 挖掘目标

根据附件 2 里的数据进行高频词统计、绘制词云图并进行模型构建，并采用 F-Score 算法进行模型评估，选出最优模型。并将附件 3 里的数据清洗，选出排名前 5 的热点问题及其留言明细。最后构建一套合理的答复意见质量评价方案。

2 问题分析并求解

2.1 群众留言分类

2.1.1 数据预处理。

对于原始数据的完整性（检查群众留言数据个体是否有遗漏，所有调查项目或指标是否齐全,有无矛盾现象，是否反映客观实际情况）和准确性(检查数据是否有错误，计算是否正确)两个方面进行审核。

首先对原始数据进行清洗，留言去重，去换行符。将一级分类换成 id（0 到 6），删除文本的标点符号与特殊符号，无意义的常用词（stopword），因为这些词与符号对系统分析预测文本的内容没有帮助。通过定义删除字母、数字、汉字意外的所有字符的函数方式，记载停用词，过滤留言文本内容。

	留言主题	一级分类
3804	咨询2013年幼师资格证是参加全国统考还是当地自考的相关问题	教育文体
4579	咨询C3县农村小学教师岗位设置的问题	教育文体
4965	咨询下我小孩是否还能完成9年义务教育	教育文体
6066	D9县乡镇卫生院的职工又有几个月没有发工资了	劳动和社会保障
7758	L12市安江镇的液化气站被严重垄断	商贸旅游
8993	K1区嘉雯牙科擅自行医	卫生计生
669	E3区公园建设指挥部的拆迁补偿协议上说的补偿款我至今没拿到！	城乡建设

在一级分类列中总共有0个空值。

留言主题列中总共有0个空值。

图 1 部分数据示例效果图

2.1.2 统计数据量

统计各类别的数据量，通过绘制分裂饼图，直观看出各类别的比例。

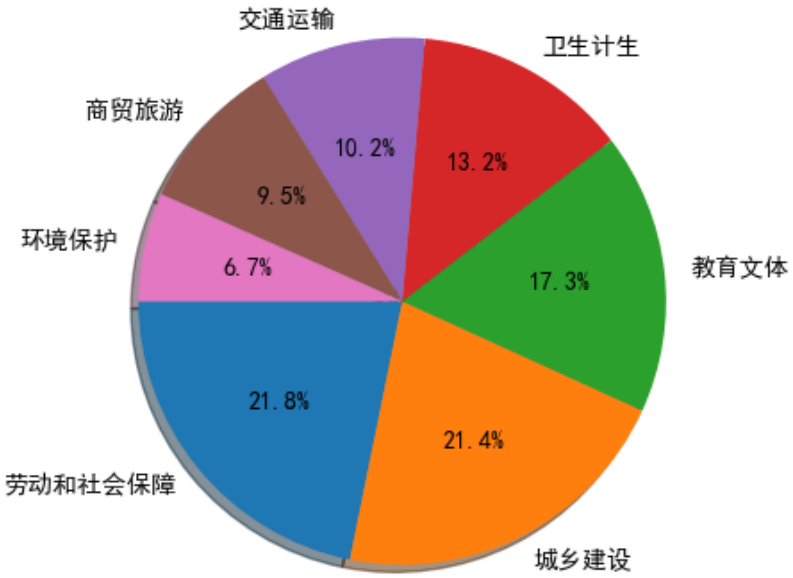

图 2 类别占比饼图

结果分析:

通过观察类别占比饼图发现，劳动和社会保障、城乡建设、教育文体类占比较多，而商贸旅游、环境保护类占比较少。

由于各类占比不一样，分布不均匀，为了实现数据平衡，我们采取随机抽样的方式，同时对城乡建设、环境保护、交通运算、教育文体、劳动保障、商贸旅

游、卫生计生，这 7 个类别进行提取等量数据，再通过平滑聚集，数据概化，规范式等方式将数据转换成适用于数据挖掘的形式。

2.1.3 绘制词云图

利用 PyEchart，将数据可视化处理，并利用 jieba，进行分词处理，过滤标点符号与一些特殊符号，建立停用词表，提取高频词，绘制词云图，通过词云图分析，预估留言内容的特点，以便后续将群众留言分派至相应的职能部门处理。

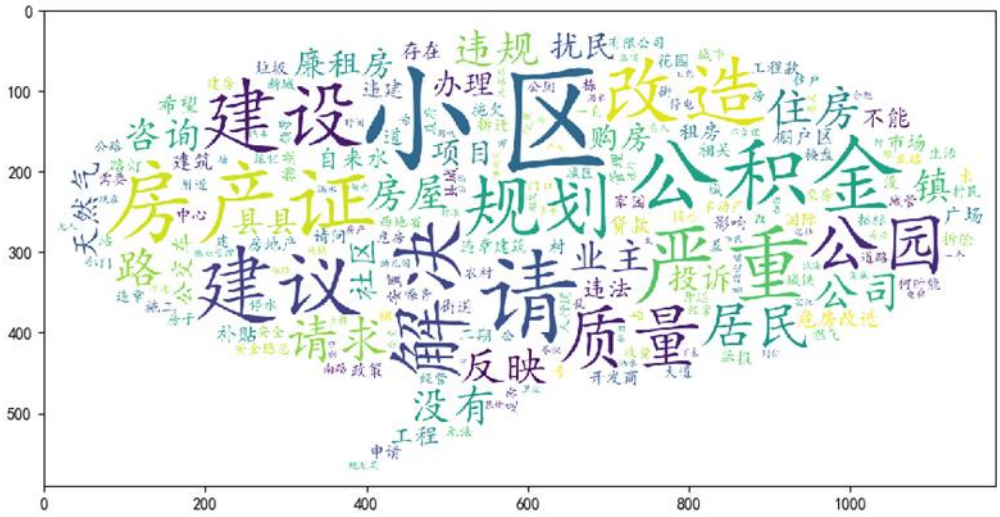

图 3 留言词云图

结果分析：

通过留言文本词云图分析，可以看出留言文本的关键词有“ 建设 ”、“小区”、“房地产“等，当前大多数热门问题都围绕着城乡建设这一大类来讨论的。基本都是房产的问题，由此可见，民众对于房产的关心还是很迫切的。民众在对规定提出问题的情况下，反应出当前城乡建设这一块的问题还是比较严重。但是也有一定民众对城乡建设较为满意，证明了城乡建设工作还是有落实。同时，也可以看到词云图中存在极多的环境污染问题，反应着我们的环境建设还是不到位，影响民众的生活。至于交通运输和文体教育这两大类的问题比较少，但是这两类的问题更重要，需要我们优先处理。

2.1.4 数据重新排序

通过 python 寻找重叠词，将其转换为具体的概率值，再进行权重词处理，最后进行归类.

2.1.5 建立分类标签

建立关于留言内容的一级标签分类模型，将各级分类进行可视化处理 。

分类结果:

序号	一级标签	数量
0	城乡建设	2009
1	劳动和社会保障	1969
2	教育文体	1589
3	商贸旅游	1215
4	环境保护	938
5	卫生计生	877
6	交通运输	613

表 1 一级标签统计

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

表 2 文本转化为数据标签

```
城乡建设        65
农村农业        56
政法          55
教育文体        45
劳动和社会保障     42
Name: 一级分类, dtype: int64
[65 56 55 45 42]
Index(['城乡建设', '农村农业', '政法', '教育文体', '劳动和社会保障'], dtype='object')
城市建设和市政管理     16
住房保障与房地产      16
其他            16
环境污染          11
社会治安          10
Name: 二级分类, dtype: int64
其他       116
安置补偿       3
事故处理       2
房屋拆迁       2
回迁房        2
Name: 三级分类, dtype: int64
Index(['留言编号', '留言用户', '留言主题', '留言时间', '留言详情', '一级分类'], dtype='object')
```

图 4 一级标签展示

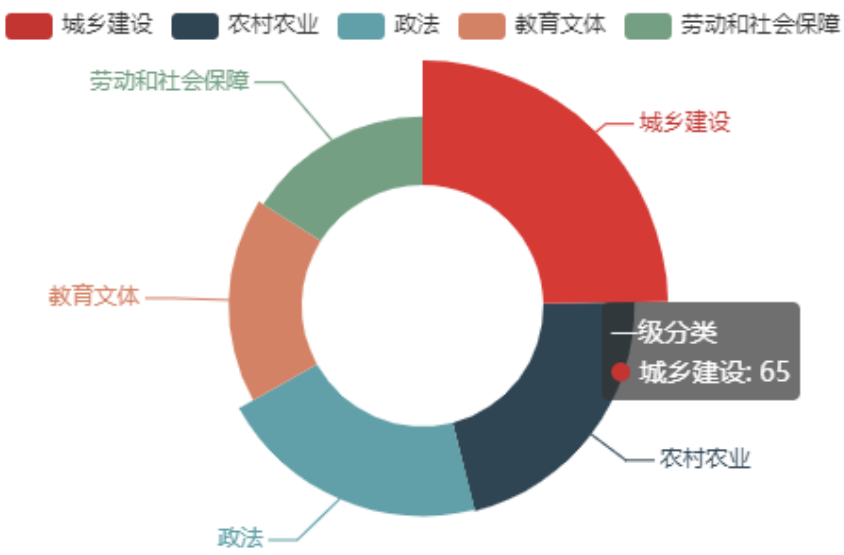

图 5 光标定位一级分类图

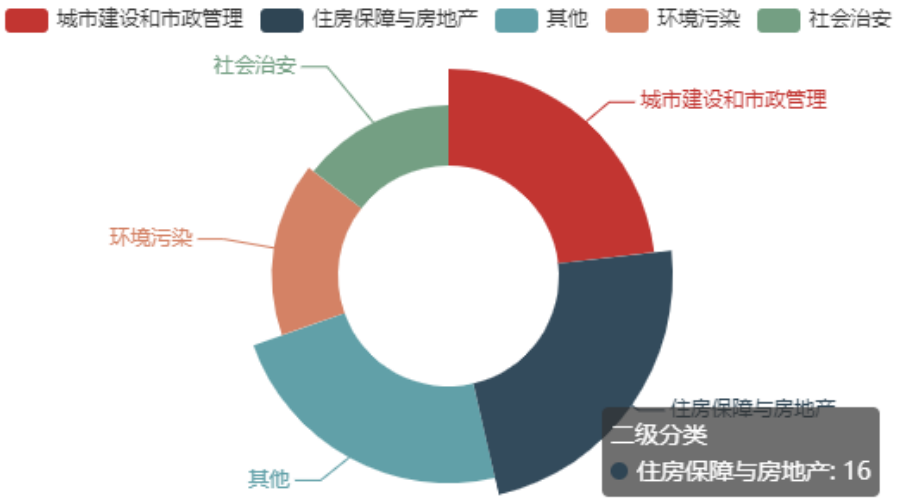

图 6 光标定位二级分类图

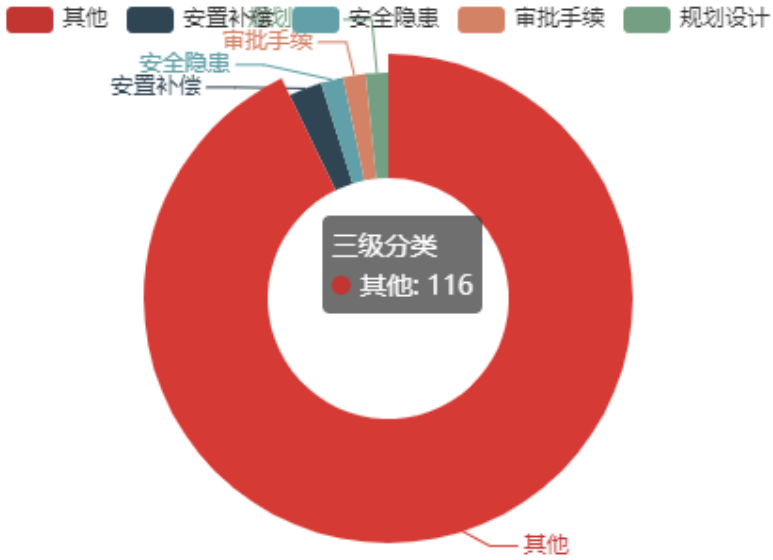

图 7 光标定位三级分类图

特征 分类 1

变量	值	概率
一级标签	劳动和社会保障	
一级标签	城乡建设	
一级标签	环境保护	
一级标签	卫生计生	
一级标签	商贸旅游	
一级标签	交通运输	

表 3 一级标签特征分类 1 表

特征 分类 2

变量	值	概率
一级标签	教育文体	
一级标签	商贸旅游	
一级标签	环境保护	
一级标签	卫生计生	
一级标签	交通运输	

表 4 一级标签特征分类 2 表

分类 1: 分类 1　　分类 2: 分类 2

分类 1 和 分类 2 的对比分数

变量	值	倾向于 分类 1	倾向于 分类 2
一级标签	劳动和社会保障		
一级标签	城乡建设		
一级标签	教育文体		
一级标签	商贸旅游		
一级标签	交通运输		
一级标签	卫生计生		
一级标签	环境保护		

表 5　分类 1 与分类 2 的对比分数表

分类 1: 分类 1　　分类 2: 非 分类 1

分类 1 和 非 分类 1 的对比分数

变量	值	倾向于 分类 1	倾向于 非 分类 1
一级标签	教育文体		
一级标签	劳动和社会保障		
一级标签	城乡建设		
一级标签	商贸旅游		
一级标签	交通运输		
一级标签	卫生计生		
一级标签	环境保护		

表 6　分类 1 与非分类 1 的对比分数表

分类 1: 分类 2　　分类 2: 非 分类 2

分类 2 和 非 分类 2 的对比分数

变量	值	倾向于 分类 2	倾向于 非 分类 2
一级标签	劳动和社会保障		
一级标签	城乡建设		
一级标签	教育文体		
一级标签	商贸旅游		
一级标签	交通运输		
一级标签	卫生计生		
一级标签	环境保护		

表 7　分类 1 与非分类 2 的对比分数表

结果分析：

由分类标签统计可知，一级分类中城乡建设依然占据了大部分，它是民众较为关心的点。政府应该着重民众所提出的热点问题作出措施。（1）例如小区的物业管理费、房价的涨幅以及楼盘的定价是否合理等。通过市场调研对相应的指标分析，做出为百姓谋福利的事情。（2）教育与法规问题也较受关心，民众对于制度的改革提出了许多的意见。大量民众反应目前的学校存在乱收费的现象，政府应该规范学校的管理，严禁教师贪污。（3）城市是在农村的基础上建成的，要想把城市建设的更好，就要同步推进农村的发展。民众们对于农村的建设工作存在较大的意见，主要集中于乡村道路的建设以及便民设施是否齐全，少部分民众关心于补助金的制度问题。

二级分类中可见，城市建设以及房产占据了两大热门点。（1）民众们最关心的就是自己住哪，住的好不好，舒服不舒服。所以政府应该着重关注城乡建设和房产市场的问题。一定要从法制法规入手，合情合理地做事，争取为百姓着想。（2）污染问题也存在一定的关注性。噪音污染问题是大家最常反应的热点问题，噪音扰民直接影响着民众的休息体验和生活质量，政府要狠抓工地以及房屋装修的工作时间，减少此类问题。（3）治安是城市正常工作的基本要求，民众需要一个安全的环境才能迸发出更大的工作激情。政府需要针对一些热点问题，例如偷抢诈骗等犯罪行为，对前科人员实时追踪，加强日常巡逻。

而在三级标签中，其他类占绝大多数，安置补偿、安全隐患、审批手续等类占比较小。

通过聚类分析，得到各级特征分类对比表格。通过观察表 4 可以发现，劳动和社会保障与城乡建设的概率占比一样多，占有 42%，环境保护、卫生计生、商贸旅游、交通运输四个类别占比较少。通过观察表 5 可以发现，一级标签的分类 2 中各类占比差距较小，比较均匀，其中教育文体占有 34%，而交通运输占有 12%。通过分类 1 与分类 2 的对比分数表中，发现劳动和社会保障、城乡建设都倾向于分类 1，而其他类别倾向于分类 2。通过表 7 可以发现，教育文体基本上倾向于非分类 1，劳动和社会保障、城乡建设都倾向于分类 1，其他类别较倾向于非分类 1，其中卫生计生倾向比例为 7.531，环境保护倾向比例为 5.425，但在非分类 2 的比较分数表中，占比发生了一定的变化，劳动和社会保障、城乡建设都倾向于非分类 2，其他类别倾向于分类 2。

2.1.6 模型评估

对提取到的特征值进行特征选择，去除冗余特征，即特征降维，计算第 i 个正类、负类、所有特征值的平均值，使用 F-Score 算法，通过比较下列各模型的

精确率（Precision）和召回率（Recall），将其作为评估指标进行评价，为减少工作量、提高效率，降低差错率等方面，达到更加优化的效果。

评价构建模型（用不同的机器学习模型，并评估他们的准确率）

使用 F-Score 算法对模型进行评价：

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

其中 Pi 为第 i 类的查准率， Ri 为第 i 类的查全率。

一、KNN 分类回归（KNeighborsClassifier）模型。

处理步骤：

利用 Scikit-Learn 对数据进行逻辑回归分析。

1、特征选择：

（1）给出各个特征的 F 值和 p 值，选出 F 值大的或者 p 值小的；

（2）递归特征消除 Scikit-Learn 提供了 RFE 包，还有 RFECV，利用交叉验证对特征进行排序；

（3）稳定性选择，表现在随机逻辑回归模型上，对训练数据进行多次采样拟合回归模型，即在不同的数据子集和特征子集上运行特征算法，不断重复，最终选择得分高的重要特征。这是稳定性选择方法。得分高的重要特征可能是由于被认为是重要特征的频率高（被选为重要特征的次数除以它所在的子集被测试的次数）

2、筛选出的特征建立逻辑回归模型，训练模型，随机逻辑回归模型筛选结果，将数据集拆分成两部分，分别为用户模型的构建与模型的测试，返回了模型在测试集上的混淆矩阵，计算出模型的测试数据集上的预测准确率。

模型结果：

accuracy	**0.7365131578947368**				
	precision	recall	f1score	support	
	城乡建设	0.67	0.83	0.74	663
	环境保护	0.81	0.69	0.74	310
	交通运输	0.74	0.65	0.69	202
	教育文体	0.75	0.74	0.74	525
	劳动和社会保障	0.76	0.81	0.78	650
	商贸旅游	0.75	0.66	0.70	401
	卫生计生	0.77	0.59	0.67	289
	accuracy	0.74	3040		
	Macro	avg	0.75	0.71	0.72

weighted	Avg	0.74	0.74	0.73	3040

表 8　KNN 分类回归（KNeighborsClassifier）模型结果

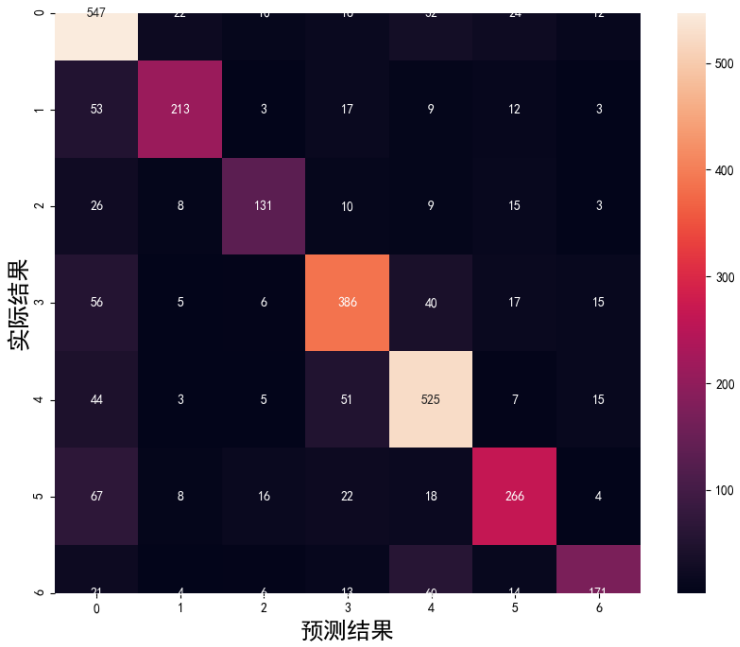

图 8 KNN 分类回归（KNeighborsClassifier）模型效果图

F-Score 评价：

Knn 分类回归模型的正确率是 0.7365131578947368，从混淆矩阵的预测结果来看，查全率与查准率相当。

模型二：线性支持向量机 SVM 模型。

处理步骤：

1、数据标准化后，将数据集分割为训练数据集和测试数据集；

2、用线性核函数初始化一个 SVM 对象，并训练线性 SVM 分类器，使用模型进行训练和测试，然后输出测试后的精度 。

模型结果：

accuracy	**0.8450657894736842**				
	precision	recall	f1score	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
	accuracy	0.85	3040		
	Macro	avg	0.87	0.82	0.84
weighted	Avg	0.85	0.85	0.84	3040

表 9　线性支持向量机 SVM 模型结果

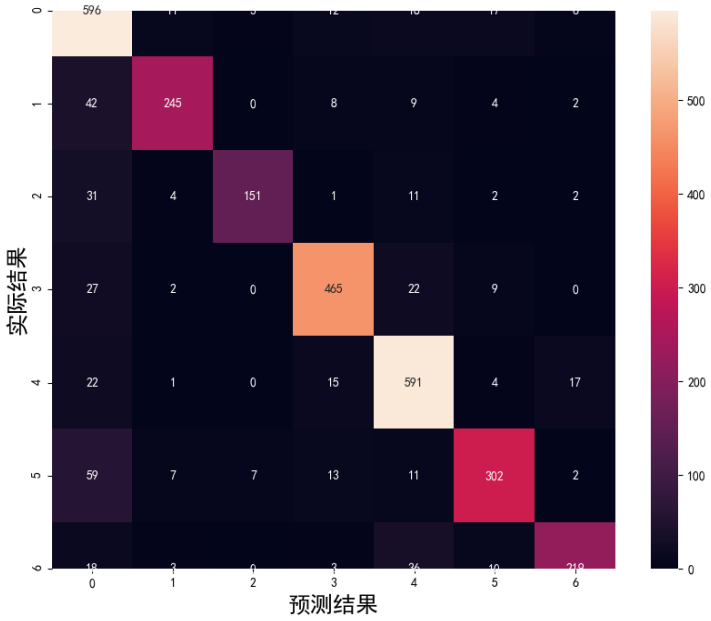

图 9　线性支持向量机 SVM 模型效果图

F-Score 评价：

线性支持向量机 SVM 模型的正确率为 0.8450657894736842，从混淆矩阵的预测结果来看，查全率和查准率相当。

模型三：多项式朴素贝叶斯模型。

处理步骤：

在多项分布朴素贝叶斯模型中，特征向量 X 的特征通常为离散型变量，并且假定所有特征的取值是符合多项分布的，可用于文本分类。在多项式模型中， 设某文档 d=(t1,t2,...,tk)，tk 是该文档中出现过的单词，允许重复，

则先验概率：P(c)= 类 c 下单词总数/整个训练样本的单词总数；

类条件概率：P(tk|c)=(类 c 下单词 tk 在各个文档中出现过的次数之和+1)/(类 c 下单词总数+|V|)。V 是训练样本的单词表（即抽取单词，单词出现多次，只算一个），|V|则表示训练样本包含多少种单词。

P(tk|c)可以看作是单词 tk 在证明 d 属于类 c 上提供了多大的证据，而 P(c)则可以认为是类别 c 在整体上占多大比例(有多大可能性)。

模型结果：

accuracy	**0.6703947368421053**				
	precision	recall	f1score	support	
	城乡建设	0.53	0.93	0.67	663
	环境保护	1.00	0.37	0.54	310
	交通运输	1.00	0.12	0.22	202
	教育文体	0.91	0.77	0.83	525
	劳动和社会保障	0.61	0.95	0.74	650
	商贸旅游	0.93	0.46	0.62	401
	卫生计生	0.99	0.25	0.39	289
	Accuracy	0.67	3040		
	Macro	avg	0.85	0.55	0.58
weighted	Avg	0.79	0.67	0.64	3040

表 10　多项式朴素贝叶斯模型结果

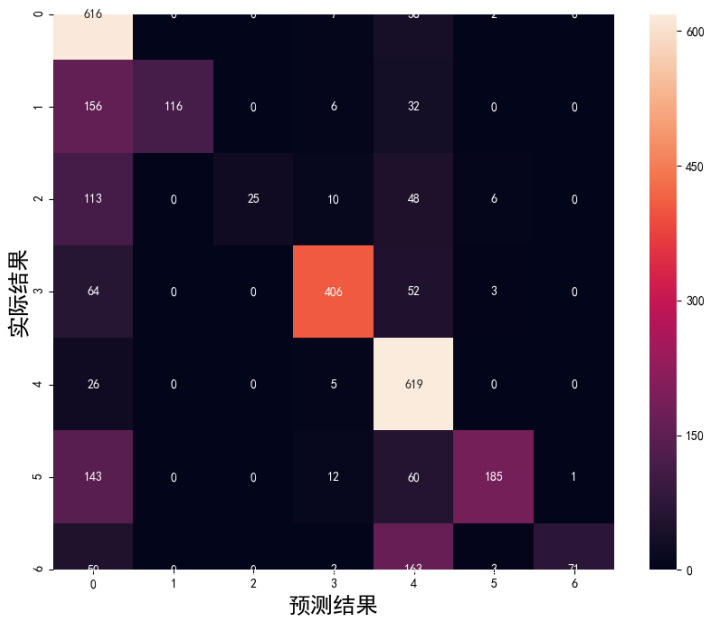

图 10 多项式朴素贝叶斯模型效果图

F-Score 评价：

多项式朴素贝叶斯模型的正确率是 0.6703947368421053，从混淆矩阵的预测结构来看，查全率比查准率高。

模型比较：

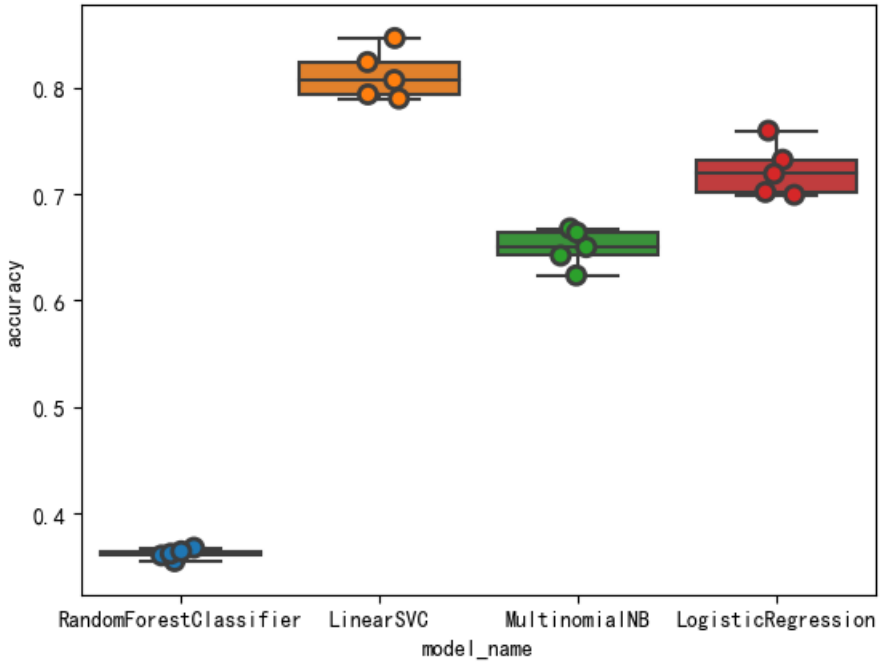

图 11 各类模型的 accuracy 比较图

结果分析：

经过以上模型评估的比较，LinearSVC 模型的 accuracy 对应值最高，即支持向量机模型模拟的效果较好，其精确度=0.8450657894736842。RandomForestClassifier 模型的 accuracy 对应值最低，即精确度较低。

各类模型的精确度不同的原因，取决与各模型的评价标准，以及优缺点。

对于 KNN 分类回归（KNeighborsClassifier）模型，对数据没有假设，准确度较高，对 outlier 不敏感，可以用于非线性分类，新数据可以直接加入数据集，但对于样本容量大的数据集计算量比较大，样本不平衡时，预测偏差比较大，且每次分类都需要重新进行一次全局运算。

对于线性支持向量机 SVM 模型，能够处理非线性特征的相互作用，无需依赖整个数据，但观测数据较多。效率不是很高，对于缺失数据非常敏感。

对于多项式朴素贝叶斯模型，对大量数据训练具有较高的速度，支持增量式运算，即可实现实时的对新增的样本进行训练，但其分类决策存在错误率，且使用了样本属性的独立性的假设，导致样本属性的有关联性处理的效果不是很好，

2.2 热点问题挖掘。

处理步骤：

1、对某一时段内的特定地点或特定人群问题的留言问题，在时间维度上进行分析，将时间转化为时间戳，并将其设置为索引，进行归类聚类分析，将索引按照时期划分，提取某一时段内问题出现频率高的问题，最后进行数据合并，筛选出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。

2、对热点问题进行有针对性地处理，以提升服务效率。合理筛选部分热点问题，进行留言统计，从多维角度进行分析，并按表 2 的格式给出相应热点问题 对应的留言信息，并保存为 ‘热点问题留言明细表.xls”。

处理结果：

热度排名	问题ID	热度指数	时间范围	地点	问题描述
1	1	10.07750079	2019-01-21/2019-01-27	A市	A市爱心卡可以刷地铁吗
2	2	9.222372609	2019-06-10/2019-06-16	A市	请问A市地铁3号线什么时候开通?
3	3	8.975899408	2019-03-25/2019-03-31	A市	A市地铁四号线何时会试运营?
4	4	8.965137828	2019-09-09/2019-09-15	A市	A市地铁1号线北延线还建吗
5	5	8.710724003	2019-06-03/2019-06-09	A4市	A4区的植基路为什么一直是断头路?

表 11 热点问题表

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	229932	A00085827	A市爱心卡可以刷地铁吗	2019/1/23 13:18	爱心卡可以刷地铁吗	0	0
2	222275	A00014325	请问A市地铁3号线什么时候开通?	2019/6/16 9:03	地铁号线什么时候开通	0	0
3	201307	A00012819	A市地铁四号线何时会试运营?	2019/3/28 20:02	请问市地铁四号线时间试运营?	0	0
4	260435	A00074973	A市地铁1号线北延线还建吗	2019/9/10 7:55	地铁号北延线不是月开建吗，都官宣几次，玩人吗?	1	0
5	220961	A00015167	A4区的植基路为什么一直是断头路?	2019/6/4 17:18	请问区植基路为什么一直是断头路？请问何时会拉通?	0	0
6	247370	A00095128	A市地铁7号线经C5市路会不会修高架桥?	2019/1/27 12:40	想问一下，地铁号线，经市路，会不会和地铁号线一样，地铁上面	0	0
7	224355	A000106222	A7县小塘路还修吗?	2019/3/23 7:32	想问一下泉塘这边的小塘路从漓楚路到南山路这一截还能拉通吗?	4	0
8	265913	A000106937	A7县北横线还修吗，为什么一直未动?	2019/10/25 18:49	请问北横线还修吗，为什么一直未动?	2	0
9	225630	A00025416	请问A市汽车南站何时会完工投入使用?	2019/7/15 11:49	市汽车南站到底什么时候可以完工并投入使用?	0	0
10	225511	A00093151	A2区碧云复绿工程为什么一年多一直在填渣土	2019/7/23 17:17	市区碧云复绿工程为什么一年多一直在填渣土?	1	1
11	244529	A000100163	A3区枫林三路的华润万家烂尾快7年了，无人问津	2019/7/25 15:50	区枫林三路的华润万家烂尾快年了，没人管。	6	0
12	204713	A00085667	请问A7县新安路泉塘公交首末站规划在哪个位置	2019/1/22 21:23	请问县新安路泉塘公交首末站规划在哪个位置?	4	0
13	214326	A00094280	在A市摇号买房，摇到了不买的话会有什么影响吗?	2019/4/25 10:45	符合刚需条件，在市摇号买房，摇到了不买的话会有什么影响吗?	0	0
14	202747	A00033382	请问西地省这边有哪些市县可以接种HPV九价?	2019/2/1 10:50	很多地方都开始可接种九价了，请问西地省这边有哪些市县可以接	0	0
15	196230	A00018062	A2区肉联厂18栋这边的平房什么时候拆迁?	2019/8/14 14:40	咨询区黑石铺冷库对面肉联厂栋的平房什么时候拆迁?	3	0
16	191249	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03	建议将“白竹坡路口”更名为“马坡岭小学”。原“马坡岭小学”取消，保	0	0
17	237798	A000107775	2020西地省城乡居民医疗保险的报销比例是多少，门	2019/11/20 15:11	西地省城乡居民医疗保险的报销比例是多少，门诊有没有报销?	0	0
18	258471	A000104357	建议A市引进智能红绿灯	2019/5/9 9:14	智能红绿灯能随时根据车流量调整红绿灯时长，建议市引进西安产	0	0
19	191800	A00050382	A4区秀峰街道1094围墙外渣土车日夜运渣土填农田，	2019/4/19 20:35	区秀峰街道围墙外渣土车日夜运渣土填农田、严重扰民！	0	0
20	233044	A00084627	A2区丽江路下穿京广铁路涵洞怎么停工了?	2019/11/22 16:03	市丽江路下穿京广铁路涵洞不是说年度完工么？怎么半年了都没动	0	0
21	282149	A00041304	请在A3区政府附近规划地铁口	2019/10/30 13:39	整个市区域区政府怎么只有区政府没有规划地铁口。	1	0

表 12 热点问题留言明细表

结果分析：

通过观察热点问题表可以看出，热点指数最高达到 10.0775007907449，热点问题的时间范围集中在 2019 年 1 月 21 日到 2019 年 1 月 27 日，其地点主要集中在 A 市，大多问题关于地铁交通方面。

通过观察热点问题留言明细表可以看出，留言时间从 2017 年 6 月 8 日到 2020 年 1 月 26 日，且在 2017 年只有一条留言，大多数时间集中在 2019 年到 2020 年，具有一定的实时性，留言点赞数最高为 2097 次，其次为 1762 次，有少部分的留言点赞数较高，但留言反对数最高为 53，绝对多数在 15 次以下。

2.3 答复意见的评价 。

方案思路：

1、 研究答复意见数据的相关性，将答复意见进行分类，分成肯定性、否定性、不定性三类，进行逻辑归纳，变量之间是否具有正（负）相关性，因果关系等问题，其变量取值之间是否存在规律，并探寻数据集里隐藏的相关关系网。将其数据可视化。

2、 研究答复意见数据的完整性，将“空值”、“不知道”、“无意义”、“不相关问答复意见“等值，进行筛选排除。

3、研究答复意见数据的可解释性，通过大量的数据来构造和验证提出的一些假设，构建规则，再将得到的模型的行为进行实际验证。

4、进行留言内容与答复意见情感分析，提取积极与消极样本，再进行分词处理，通过以上对留言内容与留言答复意见的总体倾向进行预估判断，拟定一套合适的评价方案。

具体方案：

1、完整性角度。

答复内容具有一定的完整性，留言答复的用语符合礼节，首先，向用户问好，并提到用户的相关问题，并以恰当的方式，例如具体时间的发生的事件举例说明。并提出采取的措施，最后言道感谢的话语，如感谢您对我区工作的理解和关心。这不仅是对留言用户的尊敬，更是让人民的生活更加和谐美好。

通过代码提取答复内容中的关键词，并列出所用到的敬语以及格式词的文本txt，在与提取的关键词做对比，先筛选一定的优秀答复作为训练集，在训练及统计后得出优秀答复的所占敬语的词频比例，以及检查必用词是否存在。最后得出本答复是否优秀，并根据比例打分。（此条满分 2 分）

2、及时性角度。

提取问题的提出时间与相应的答复时间，再根据任务一划分的一级标签，得出各大类、小类的平均答复时间，再以平均答复时间作为基准，对每一条答复作出回复时间的评价。最后得出本答复是否及时，并根据比例打分。（此条满分 1 分）

3、行动性角度。

结合每类问题的答复及合乎情理的一些问题，部门采取了适当的措施，予以答复，例如“在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。”利用结巴分词，提取并验证是否存在行动词，例如已要求、已开展、调查等，从而对答复内容中的实际行动进行评价。根据每小类的平均行动词占比，对每一条

答复作出行动能力的评价。最后得出本答复是否行动，并根据比例打分。（此条满分 2 分）

4、可解释性角度。

对于不符合现实的答复，选择澄清事实，并分点说明理由，让其答复内容更加具有可信性，例如对“G 市残疾人创业求支持”、“B 市泰民米粉厂存在严重安全隐患”等主题的部分答复。通过文本提取，圈定所有答复中未能成功解决问题的答复，并检索是否有理由说明。最后得出本答复是否存在理论，并根据比例打分，成功解决问题的记满分。（此条满分 2 分）

5、解决性角度。

部门向部分带有疑惑的留言，给予了相关问题的联系方式、解决方案等帮助，如“请咨询厅基层工作处，电话：0731-0000-00000000”。通过相应解决关键词，对每一条答复进行文本检索，并统计每一小类的平均解决率。通过答复与相应小类的答复解决率相比较，最后得出本答复是否解决问题，并根据比例打分。（此条满分 2 分）

6、合法性角度。

结合当地法律规定，进行合理且合法地回复用户，例如“B 市男职工生育险其配偶能报吗？”等主题的部分答复。通过文本检索，检测答复中是否提及法律法规，以及此答复的是否合法。最后得出本答复是否合理合法，并根据比例打分。（此条满分 1 分）

总而言之，部门可以通过一系列指标进行各小项的量化评分，以满分十分对每条答复相应机械化打分，并向民众提供答复评价评分表，再通过与机械打分的情况不断对比学习，进行优化，最后得出完整的评价体系和评价方案。

3 总结

本文的主要目的是利用数据挖掘与数据建模技术，将文本数据合理清洗及挖掘，建立分类标签模型并提取热点问题，最后对答复意见质量对出评价。

通过一系列的模型建立，方便了政府的工作，能够对热点问题针对性地作出答复，也能高效地进行分类，将指定的问题分类到正确的应对部门。最后能智能评价答复意见质量，将答复质量量化，一目了然。