基于 KNN 算法的“智慧政务”R 语言文本挖掘应用

摘要

随着网络问政平台成为广大人民群众反应问题的主要方式和渠道，如何处理政务大数据，从而提升政府政务水平，便成为了“智慧政务”的重要问题。本文以互联网公开渠道收集而来的政务信息作为主要的研究对象，将群众留言进行文本挖掘和分词处理，利用 KNN 模型群众留言分类问题，并运用 F-score 方法对分类方法预测进行了评价。定义热度表达式，利用数据挖掘方法，并找出了关注度最高的 5 个问题列举在文中表格。评价答复意见主要考虑了答复的相关性，仍然使用 KNN 模型，对群众留言对应的一级标签分类和答复内容对应的一级标签分类进行比较，找出答复内容对应的一级标签分类占原问题分类的比例，发现该比例较高，答复意见效果较好。

关键词：KNN 算法、智慧政务、F-score、文本挖掘、文本分词处理、R 语言

目录

1 引言

1.1 问题重述

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。

附件给出了收集自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。请利用自然语言处理和文本挖掘的方法解决下面的问题。

1、群众留言分类问题

请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型，并使用 F-Score 方法对分类方法的效果进行评价。

2、热点挖掘问题

某一时段内群众集中反映的某一问题，称为热点问题。请根据附件 3，讲某一时段内反映特定地点和特点人群问题的留言进行分类，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，另存为“热点问题表.xls”。给出相应热点问题的留言信息，保存为“热点问题留言明细表.xls”。

3、答复意见的评价

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。

1.2 本文主要工作和创新点

本文基于题目附件所给的数据，从题目的分析角度和已有的学术基础上，结合文本挖掘中的文本分词和自然语言处理方法、数据挖掘中最常用的 KNN 分类算法，完成了群众留言分类问题的预测和评价，以及答复意见内容的分类。通过数据分析和筛选，定义热度公式找出了热度前 5 的事件相关信息，操作简单可行。

（一）通过文本挖掘和自然语言处理的角度。本文对数据预处理后的群众留言信息，并生成文档-词条矩阵，将大量非结构化的文本数据转化为结构化的数字矩阵。这样做便于接下来的统计性文本分析。

（二）使用 KNN 分类模型进行分类预测。KNN 分类模型是数据挖掘模型中经典的核心模型，使用 KNN 分类模型能够快速训练并且预测出位置观测的类别，简单易行。

（三）定义热度函数完成重点事件分类和筛选。定义了热度函数，热度函数值高的事件排在最前面，将事件与热度函数联系，通过热度函数的结果，一目了然地看到热点事件的具体程度。

2 群众留言分类问题

2.1 问题分析

处理网络问政平台的群众留言，需要将这些群众留言的内容，按照内容分类三级标签体系进行分类，这样能够让问题交给对应的职能部门解决，从而提高整个政务系统的工作效率。但是人工分类的方法，应对政务大数据显然苍白无力。考虑文本分词，将非结构化的留言内容结构化，并将结构化的内容，使用 KNN 算法对未知类别进行分类，便是处理问题的一种主要思路。

2.2 数据预处理

本文以互联网公开渠道收集而来的政务信息作为主要的研究对象，研究事件从 2019 年 1 月 1 日，到 2020 年 1 月 8 日。数据具有全面性和较高的权威性，以及较高的质量。本文使用 R 语言（Rstudio 软件）进行研究，是因为 R 语言简单易用，拓展包功能十分丰富，利于进行数据分析。

2.2.1 数据提取和数据转换

本题需要建立关于留言内容的一级标签分类模型。从题目中所给出的“附件 2.xlsx”表格数据以及数据结构看，一级分类的内容只与留言内容有关，且留言主题和留言内容具有高度的相似性，因此在分类的过程中，不将留言主题纳入自变量的范围中。

利用 R 语言 readxl 包的 read_xlsx 函数，将附件 1 的数据中留言详情列和一级标签列的全部内容，导入并保存到变量 data_1st 当中，留作本题的数据源。将一级标签转换为因子型，发现群众留言的主要问题为城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生这 7 大类问题。同时使用 summary 函数，分析这七大类问题的数量分布。结果如下：

```
> summary(data_1st)
   留言详情                    一级标签
 Length:9210        城乡建设        :2009
 Class :character   环境保护        : 938
 Mode  :character   交通运输        : 613
                    教育文体        :1589
                    劳动和社会保障:1969
                    商贸旅游        :1215
                    卫生计生        : 877
>
```

图 1　群众反映问题的数量分布

可以看到，在总共 9210 条留言当中，城乡建设、劳动和社会保障、教育文体这三大类问题，是群众向政府反映的主要问题类型。

2.2.2 划分训练集和测试集

划分训练集和测试集的方法有很多，主要有：留出法、交叉验证法、自助法（Bootstrap 法）等。但是划分出来的训练集和测试集，在结构上都需要大体相同才能够继续进行分类学习，否则会出现结构性问题。

为了解决结构性问题，我们在 R 语言中使用留出法，结合 sample 函数抽样和变量索引，随机选取 data_1st 中的 75%内容为训练集，剩余 25%的内容作为测试集。训练集和测试集的数据概要如下：

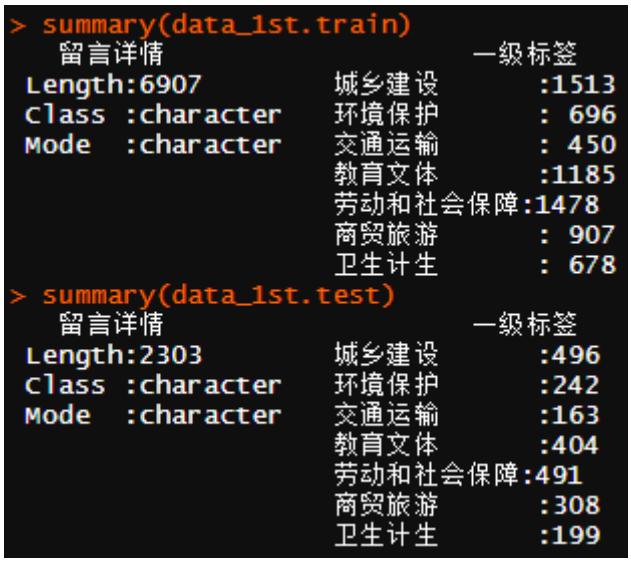

图 2 训练集和测试集概要

从概要中看到，训练集和测试集结构差别不大，因此能够接着进行下一步分析。

2.3 文本挖掘

在分好测试集和训练集之后，将进行文本挖掘。文本挖掘分析的主要步骤，包括文本分词，生成文档-词条矩阵、进行统计分析。

2.3.1 文本分词

对上述测试集和训练集分别进行文本分词操作，具体用到了 Rwordseg 包的 segmentCN 函数。Rwordseg 包是一个常用的进行中文文本分析和挖掘的拓展包，segmentCN 函数用来

将一个词组、一句话甚至是一篇文章分成多个文字和词语，从而将非结构化的文本数据转换为结构化的文本数据。

2.3.2 去除终止词

终止词是在自然语言里最为常见的词汇，在所有文本当中都会相当频繁地出现，比如英文当中常见的“is”，中文当中常见的“的”，这些都是常见的终止词。然而，这些终止词对于类别的分类却没有太大的作用，甚至对分类效果产生负面影响。因此需要从文本分词后再做一遍去除终止词的操作，才能让分类效果更加显著。

2.3.3 生成文档-词条矩阵

R 语言中的 tm 组件，支持对文本进行组织和预处理，还提供了分析组件所需要的基础架构。在本题中，还需要将刚经过分词的训练集和测试集合并起来，并用 tm 包中的函数 VectorSource 将文本向量化（即构建特征向量，每一个句子对应一个分词和频率构成的向量当中），最后添加进语料库当中，这里需要使用 tm 包中的 corpus 函数实现。

2.4 构建 KNN 模型

2.4.1 定义和模型原理

KNN 模型，全称近邻分析模型，是一种用于数据分类预测的模型。

为了预测一个新观测 X_0 输出变量 y_0 的取值，近邻分析法的基本思想主要如下：

首先，在已有数据中找到与 X_0 相似的若干个（如 K 个）观测，如$(X_1, X_2, \cdots, X_k)$。这些观测称为 X_0 的近邻

然后，对近邻$(X_1, X_2, \cdots, X_k)$的输出变量$(y_1, y_2, \cdots, y_k)$，计算诸如算术平均值（或加权均值，或中位数，或众数）之类的统计量，然后将统计量的数值，作为新观测 X_0 输出变量取值 y_0 的预测值。

典型的近邻分析方法是 K-近邻法(KNN)，KNN 方法将样本包含的 n 个观测数据看成为 p 维（p 个输入变量）特征空间中的点，并根据 X0 的 K 个近邻的$(y_1, y_2, \cdots, y_k)$依函数计算如下统计量：

$$\hat{y}_0 = \frac{1}{K} \sum_{X_i \in N_k(X_0)} y_i$$

这个统计量的结果，就作为新观测输出变量取值的预测值。

2.4.2 常用距离和 R 语言实现

KNN 方法中常用的距离有欧氏距离、闵可夫斯基距离、绝对距离、切比雪夫距离、夹

角余弦距离等。在 R 语言中，实现 KNN 算法所使用的距离函数是欧氏距离。

R 语言中 class 包中的 knn 函数用来实现 KNN 模型，其基本格式如下：

knn(train=训练样本集, test=测试样本集,

cl=输出变量, k=近邻个数 K, prob=TRUE/FALSE,

use.all=TRUE/FALSE)

利用 knn 函数，就能够使用 KNN 近邻分析模型，对未知分类的观测进行分类预测。

2.5 分析结果和 F-score 方法评价

测试集分析结果预览如下，详细的预览结果保存在测试文件中。

```
> knn_predict
  [1] 城乡建设       城乡建设       交通运输       城乡建设       城乡建设       交通运输       交通运输
  [8] 城乡建设       城乡建设       城乡建设       交通运输       交通运输       商贸旅游       商贸旅游
 [15] 交通运输       交通运输       交通运输       环境保护       劳动和社会保障 商贸旅游       交通运输
 [22] 交通运输       交通运输       交通运输       城乡建设       交通运输       城乡建设       交通运输
 [29] 卫生计生       交通运输       交通运输       交通运输       交通运输       交通运输       交通运输
 [36] 交通运输       劳动和社会保障 城乡建设       交通运输       交通运输       交通运输       交通运输
 [43] 商贸旅游       交通运输       交通运输       劳动和社会保障 城乡建设       城乡建设       交通运输
 [50] 交通运输       交通运输       交通运输       交通运输       城乡建设       交通运输       交通运输
 [57] 交通运输       教育文体       交通运输       交通运输       交通运输       交通运输       交通运输
 [64] 城乡建设       劳动和社会保障 城乡建设       交通运输       交通运输       交通运输       交通运输
 [71] 交通运输       城乡建设       交通运输       城乡建设       交通运输       交通运输       商贸旅游
 [78] 城乡建设       城乡建设       交通运输       商贸旅游       交通运输       交通运输       劳动和社会保障
 [85] 交通运输       劳动和社会保障 交通运输       劳动和社会保障 交通运输       交通运输       劳动和社会保障
 [92] 交通运输       城乡建设       城乡建设       交通运输       城乡建设       城乡建设       交通运输
 [99] 环境保护       交通运输       交通运输       城乡建设       交通运输       商贸旅游       交通运输
[106] 卫生计生       环境保护       交通运输       城乡建设       劳动和社会保障 交通运输       交通运输
[113] 城乡建设       交通运输       城乡建设       城乡建设       交通运输       交通运输       交通运输
[120] 城乡建设       城乡建设       城乡建设       交通运输       交通运输       交通运输       交通运输
[127] 商贸旅游       交通运输       交通运输       交通运输       交通运输       交通运输       交通运输
[134] 卫生计生       城乡建设       交通运输       交通运输       交通运输       交通运输       交通运输
[141] 交通运输       城乡建设       交通运输       交通运输       交通运输       商贸旅游       交通运输
[148] 交通运输       卫生计生       交通运输       交通运输       城乡建设       城乡建设       交通运输
[155] 城乡建设       教育文体       交通运输       卫生计生       交通运输       交通运输       劳动和社会保障
[162] 交通运输       商贸旅游       商贸旅游       交通运输       城乡建设       交通运输       劳动和社会保障
[169] 教育文体       商贸旅游       商贸旅游       商贸旅游       交通运输       交通运输       交通运输
[176] 城乡建设       交通运输       交通运输       交通运输       交通运输       城乡建设       交通运输
[183] 交通运输       卫生计生       交通运输       交通运输       交通运输       劳动和社会保障 交通运输
[190] 交通运输       交通运输       交通运输       交通运输       商贸旅游       交通运输       交通运输
[197] 交通运输       商贸旅游       劳动和社会保障 交通运输
Levels: 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
```

图 4 测试集的预测分类结果

2.5.1 F-score 方法原理

通常使用 F-score 方法，对分类结果进行评价：

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

其中 P_i 为第 i 类查准率，R_i 为第 i 类查全率。

3 热点问题挖掘

3.1 问题分析

热点问题，指的是在某一个指定的时间段内，群众集中反映的某一问题。这个问题可能问法不一，说法不同，但某一问题在更大的层面上说，是指某一大类的问题，更具体地讲，就是某一段时间之内，集中反映的一级分类问题。

3.2 定义热度指标并整理事件热度

假定看完留言的网民都会选择支持或者反对，则本文定义关注度作为热度指标的衡量标准，具体公式为：

$$关注度 = 支持数 + 反对数$$

按照这个具体公式，我们选出来具体的关注度最高的五个问题，列表如下，并保存在“热点问题表.xls”当中

表 1 热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	208636	2097	2019.8.19	汇金路别墅小区/业主	配套治理混乱，存在安全隐患
2	223297	1767	2018.4.1-2019.4.12	梅溪湖金毛湾/小区业主	配套治理混乱
3	220711	821	2018.8.20-2019.2.21	A4 区 p2p 公司/受害者	全国车贷诈骗案件
4	217032	790	2019.2.21-2019.2.25	A4 区 p2p 公司/受害者	集资诈骗案
5	194303	733	2019.2.21-2019.3.1	A4 区 p2p 公司/受害者	诈骗案件没有跟进

表 2 热点问题留言明细表

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
208636	A0007711	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019/8/19 11:34:04	我是 A 市 A5 区汇金路五矿万境 K9 县 24 栋的一名业主，我们小区一开始的定位是一个高端别墅小区，实行人车分流管理，物业也标榜 37 度五星级服务，到目前为止小区群租房泛滥成灾，有直接十几个工人租住毛坯房的，也有二房东将毛坯房租过来擅自改成好多格子间,加装 n 块电表的，也有毛坯房直接搬煤气罐入户做饭的，有穿个内裤到处晃悠的，有没装烟机打开门做饭导致整个楼道烟味呛鼻的，我们小区高层大部分都是刚需买房，或者是因为结婚，或者是因为给孩子读书，这些群租房的存在给我们的安全带来极大的隐患以及不确定性，就好像埋在小区的不定时炸弹，说不定啥时候就爆了。为什么 A 市处理了那么多群租房，而我们小区的群租房一直是无人问津的状态，多次投诉，社区说管不了，物业说管不了，12345 也说管不了，到底有谁能来管管，有谁能保证我们人民群众的安居？难道非得等群租房起火爆炸了或者是出现了严重的治安事故才有人来管么？为什么这么大的隐患存在投诉无数次都无人问津无人处理的情况，每个部门都在推卸责任，三房两厅被改成五个单间的格子间，请问是谁批准的，这一波二房东有组织有计划的侵占我们业主的正当权益，给我们业主埋下了炸弹请问是谁允许的？我们小区也曾发生过狗咬人，请问有人对养宠物的情况进行过排查吗？我们小区黄谷路路口每天车辆通行不计其数，数量非常庞大，然而孩子们上学这是必经的路口，每天大人过马路都胆战心惊的，更何况孩子呢？为什么这样一个危险的路口没有人行天桥或者地下通道？五矿万境K9县小区作为一个封闭式的高端小区现在物业管理及其混乱，小区门禁几乎没用，保安见人就开门从不询问，跟物业投诉过多次，却始终无果，小区的保洁也是一直跟物业投诉然后也是一直无果，对于一个这样的物业，收取着高额的物业服务费却几乎没尽到物业该尽的责任，这样一个物业谁评的五星？谁能来监管，谁能保证我们业主的基本安全和卫生？关于之前一直强调的人车分流的小区，现在小区儿童的休闲区域各种快递车外卖车以及业主的电动车都往上面开，孩子的安全谁来保障？小区将电动摩托车充电桩安装在地下室中间，地下室阴暗潮湿视线不好制动也不好，极容易发生交通事故，这样安排合理？电摩充电一旦发生	0	2097

				火灾小区的车辆是不是都要遭殃？24 栋 25 栋每个单元消防楼梯都只到一楼，在负一楼和一楼中间唯一的楼梯在小区地下车库中间，然后一旦发生火灾地下室的人该如何逃生？盼请回复，还正政府出面帮我们解决这些实际问题，只有让我们安居了我们才能乐业。		
223297	A0008752 2	反映 A 市金毛湾配套入学的问题	2019/4/11 21:02:44	书记先生：您好！我是梅溪湖金毛湾的一名业主，和其他业主一样因为当初金毛湾的承若学校都是金毛建的，果断买了梅溪湖金毛湾。楼盘当时承若小学配套周南小学或实验三小，中学配套周南中学或西雅中学。2018 年 4 月 16 号学区划分没有金毛湾，诸多业主进行了维权，最终由市教育局缪副局长做出了相关解释：因为金毛湾没有交房，待交房后经过测量后会在 2019 年上半年早一点安排配套入学，学校也就在周南中学和西雅中学两所学校之内，因为楼盘离这两所学校最近。得到缪副局长的解释后所有业主也就停止了维权，期盼着今年出具配套入学的相关文件。2018 年 5 月 10 日单位回复网友业主（hnistch）信中提到鉴于金毛湾项目交付时间为 2018 年 12 月 31 日,A 市教育局暂未将金毛湾楼盘纳入配套入学，A3 区教育局向市教育局汇报，明确于 2018 年 7 月 31 日前将金毛湾纳入周南梅溪湖中学或西雅中学配套入学范畴。但是，时至今日仍然没有相关配套入学文件出台，十几个业主赴市教育局咨询得到的结果是：今年报名系统里有你就报，明年或以后不能保证，配套文件以后也没有了，楼盘的配套是开发商给你们的，不是教育局。周边楼盘同样没有交房的建发央筑同样没有交房却已经划入了学区，金毛在梅溪湖的诸多楼盘中只有金毛湾学位至今还没有着落，现在又到开学季，诸多业主为悬而未决的学位问题很是担心。恳请市局领导为民做主，解决金毛湾 2800 户业主担心的学位问题，不胜感激！市民：周某	5	1762
220711	A0003168 2	请书记关注 A 市 A4 区 58 车贷案	2019/2/21 18:45:14	尊敬的胡书记：您好!A4 区 p2p 公司 58 车贷，非法经营近四年。在受害人要求下，于去年 8.20 立案侦察，至今已 6 个月整。未发一字立案公告和案件进展财产处置通报，全国仅此一家！6000 受害人分布于全国，许多人一生积蓄出借在平台，焦急万分，急切盼望有案情消息总是失望，四处诉求也无效。此种办案明显违背两高一部《指导意见》及时公布案件进展与资产处置，维护受害人权利的规定！在此，我们盼望您在百忙中给予关注，	0	821

				督促 A 市公安尽早发布案件侦办进展通报，追赃挽损成果，尽全力为受害人挽回损失，以安民心！谢谢！		
2170322	A00056543	严惩 A 市 58 车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	胡市长：您好！西地省展星投资有限公司设立 58 车贷 https://baidu.com/亿。2018 年 8 月 6 日，58 车贷爆雷，其法定代表人、大股东苏纳和董事长邢明向（化名邢 ze）（夫妻关系）外逃美国。随即，几十位 58 车贷受害者向 A 市 A4 区公安分局报案，未立案，经侦毛钧回复没有办案经费，对报案人也不给报案回执。8 月 20 日，全国各地 100 多位受害者聚集西地省政府门口，A4 区公安分局才被迫以“非法吸收公众存款罪”立案，但并不给报案人立案通知书。当时，经侦毛钧承诺：办案会通报，重大信息会披露。五个月过去，其承诺无一实现。A4 区公安分局未公布过任何案情，除了冻结部分银行账户，未查封任何涉案人员房子、车子、证券等其他资产，也不公布冻结银行存款情况。9 月份，查封了 58 车贷的办公室，但之后，受害者发现，该办公室已经解封，人员随便进出。对涉案人员，9 月份就说要通缉外逃的邢 ze 苏纳，却现在还在办手续，未通缉，对留在国内涉案人员，未采取任何措施。办案民警毛骏还为他们一一辩解，说总经理刘顺是 17 年才进来的，进来后对 58 做了合规工作，上了银行存管系统，对其发假标的事却不提，意思是刘顺不止无罪，还有功；说大股东苏纳和小股东、苏纳弟弟苏吕是挂名；说担保公司法定代表人陈杰雄是替邢 ze 背锅的，但受害人查证，苏吕是线下资产端负责人，陈杰雄在 58 有高额工资，两人的银行账户用于 58 借贷中的资金往来；羊毛头子孙开，其自称是为邢 ze 维稳，毛骏说他是出借人。毛骏话中，涉及近 5600 人、3 亿多的诈骗案都是出逃美国的美国公民邢 ze 一人所为，所有中国人，不管是外逃美国的苏纳，还是留在国内的苏吕、刘顺、陈杰雄、孙开等人都是无罪的。由于未控制人员，未查封资产和 58 网站，涉案人员可自由修改网站，利用网站进行各种活动，公然打着经侦旗号，多次组织、操控伪代表选举，妄图用代表大事化小、小事化了。涉案人员有组织地打电话给受害者，威胁说不投票视为投票，不投票的不给兑付。受害人向毛骏反映，毛骏不仅不理，还公然替涉案人员说话。受害人发现，毛骏跟羊毛孙开保持着紧密联系，孙开多次去毛骏办公室，出来后，以经侦名义散布各种有利于涉案人员行动的消息。因为 58 车贷大股东苏纳是西地省体育局子弟，对 58 车贷诈骗案反常现象，受害者怀疑西地省官场有苏纳、苏吕关系网干涉办案。特此举报，希望调查 58 车贷案中，地方政府、官员的腐败、违法行为，打掉保护伞。	0	790

194343	A0001061	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	胡书记：您好！58车贷案发，引发受害人举报投诉，也引起市领导的重视，公布了受害人的留言，使受害人深受感动，也看到了希望。但是，A市A4区经侦并没有跟进市领导的留言，案件调查进展报告、司法审计报告还是没有公布。鉴于此情，垦请胡书记关注此案，督促A4区经侦，领会市领导意图，尽快跟进领导留言。	0	733

4 答复意见的评价

4.1 问题分析

本问题需要根据附件4中给出的相关部门对留言的答复意见数据，对答复意见的质量给出一套评价方案。这里主要考虑答复的相关性。

4.2 解决思路

针对本问题，仍然需要像解决问题1的思路，对附件4当中的答复文本数据进行数据处理，包括文本挖掘中文本分词的处理，去除终止词，转换为文本矩阵，然后再用KNN模型对答复的文本数据的一级分类标签进行预测，如果预测值跟需要解决问题的一级标签相同或者相近，则认为这个答复意见是有效且相关的。

参考文献

[1] 吴宏胜.基于可信计算和 UEBA 的智慧政务系统[J].信息网络安全,2020,(1):89-93. DOI:10.3969/j.issn.1671-1122.2020.01.013.

[2]蒲泓宇,马捷,黄山.基于业务流的智慧政务多源信息协同结构分析 ——以长春市为例[J].情报资料工作,2020,41(1):13-23. DOI:10.12154/j.qbzlgz.2020.01.002.

[3] 王勤英.基于云安全技术的智慧政务云解决方案[J].中国新通信,2019,21(15):129. DOI:10.3969/j.issn.1673-4866.2019.15.104.

[4] 郭明良.智慧政务视频信访平台的开发与设计[J].企业科技与发展,2019,(7):57-59. DOI:10.3969/j.issn.1674-0688.2019.07.024.

[5] 云南永兴元科技有限公司.一种综合智慧政务信息服务及管理服务智能终端:CN201921644005.2[P].2020-03-31.

[6]李薇.智慧政务门户网站建设研究[J].科学与信息化,2019,(8):50-51.

[7]汪玉凯.数字政府的到来与智慧政务发展新趋势 ——5G 时代政务信息化前瞻[J].人民论坛,2019,(11):33-35. DOI:10.3969/j.issn.1004-3381.2019.11.011.

[8] 陈海波,李研.中国联通智慧政务大脑生态体系创新与实践[J].通信管理与技术,2018,(6):47-50. DOI:10.3969/j.issn.1672-6200.2018.06.024.

[9] 张利.智慧政务框架下的大数据共享实现与应用研究[J].通讯世界,2019,26(3):220-221. DOI:10.3969/j.issn.1006-4222.2019.03.141.

[10]熊磊.智慧政务大数据统一平台解决方案[J].信息技术与标准化,2018,(1):24-27.

[11]

附录：R 语言运行代码

```
# open with encoding UTF-8.
library(tmcn)
library(HMM)
library(Rwordseg)
library(jiebaRD)
library(jiebaR)
library(readr)
library(readxl)
library(NLP)
library(tm)
library(class)

### 第 1 题：分类
## 导入数据
data_1 <- as.data.frame(read_xlsx('D:/2020_C/C 题全部数据//附件 2.xlsx'))
# head(data_1)

## 数据提取 # 这里提取附件二中的详情内容和分类内容。
data_1st <- data_1[,5:6]
## 添加顺序标记
data_1st$选择序号 <- 1:dim(data_1)[1]
## 数据转换
data_1st$一级标签 <- factor(data_1st$一级标签)

## 划分训练集和测试集：留出法。这里确定随机划分比例：data_1st 中的 75%划分为训练集，其余为测试集
# training set:
data_1st.train <- data_1st[sample(nrow(data_1st), 0.75*dim(data_1st)[1], replace = F),]
summary(data_1st.train)
# testing set:
# 选取训练集中已经抽取的选择序号
d1_train_index <- data_1st.train[,3]
data_1st.test <- data_1st[-d1_train_index,]
data_1st.test <- data_1st.test[,-2]
```

```
dim(data_1st.test)[1]

## 对训练集和测试集进行分词处理
## 第一次分词处理
# training set:
segwords_1.train <- segmentCN(data_1st.train$留言详情[1:dim(data_1st.train)[1]]) # TRUE
# testing set:
segwords_1.test <- segmentCN(data_1st.test$留言详情[1:dim(data_1st.test)[1]]) # TRUE

## 第二次分词处理

# 数据清洗：设置停用词函数，去除停用词
stopword <- stopwordsCN()
RemoveStopword <- function(x, stopword){
  temp <- c()
  x = x
  stopword = stopword
  index = 1
  while(index <= length(x)){
    if(length(stopword[which(stopword == x[index])]) < 1 ){
      temp <- c(temp, x[index])
    }
    index <- index + 1
  }
  return(temp) # 返回经过处理后的分词向量。
}

# 去除停用词
# training set:
segwords_1.train_temp <- lapply(segwords_1.train, RemoveStopword, stopword)
# testing set:
segwords_1.test_temp <- lapply(segwords_1.test, RemoveStopword, stopword)

## 生成文档-词条矩阵
segwords_1_all <- c()
```

```
segwords_1_all[1:length(segwords_1.train)] <- segwords_1.train_temp
segwords_1_all[(length(segwords_1.train)+1):dim(data_1st)[1]] <- segwords_1.test_temp
corpusAll <- Corpus(VectorSource(segwords_1_all))
segwords_1_all.dtm <- DocumentTermMatrix(corpusAll, control = list(wordLengths = c(2,Inf)))
dtmAll_matrix <- as.matrix(segwords_1_all.dtm)

# Classfication：KNN 算法
rownames(dtmAll_matrix)[1:6907] <- as.character(data_1st.train$一级标签)
rownames(dtmAll_matrix)[6908:9210] <- ''
tr <- dtmAll_matrix[1:6907,]
te <- dtmAll_matrix[6908:9210,]
CL <- factor(rownames(tr))

Knn_predict <- knn(tr[1:400,], te[1:400,], CL[1:400])

# 第二题：热点问题挖掘
# 数据导入
data_2 <- as.data.frame(read_xlsx('G:/C 题全部数据//附件 3.xlsx'))
# 数据转换
data_2$留言时间 <- as.Date(data_2$留言时间)
summary(data_2$留言时间)
data_2$关注度 <- data_2[,6] + data_2[,7]  # 假定看过留言的网友都点赞或者反对。
data_2_focus <- data_2[order(data_2$关注度, decreasing = T),]

# 选择关注度最高的 5 个事情
data_2_focus[1:5]

# 第三题：评价指标
# 答复相关性
# 考虑使用（1）中的 Knn 算法，对回复内容进行分类，如果对应一级标签跟原问题一级标签符合，就算是答复有相关性。
data_3 <- as.data.frame(read_xlsx('G:/C 题全部数据/附件 4.xlsx'))
data_3$答复时间 <- as.Date(data_3$答复时间)
data_3_replyset <- data_3[,6]
```

```
## 第一次分词处理
# testing set:
segwords_2.test <- segmentCN(data_3_replyset[1:length(data_3_replyset)]) # TRUE

# 第二次处理：去除停用词
# testing set:
segwords_2.test_temp <- lapply(data_3_replyset, RemoveStopword, stopword)

## 将回复内容转换成生成文档-词条矩阵
segwords_2_all <- c()
segwords_2_all[1:length(segwords_2.test_temp)] <- segwords_2.test_temp
corpusAll_2 <- Corpus(VectorSource(segwords_2_all))
segwords_2_all.dtm <- DocumentTermMatrix(corpusAll_2, control = list(wordLengths = c(2,Inf)))
dtmAll_matrix2 <- as.matrix(segwords_2_all.dtm)

## Knn 模型
rownames(dtmAll_matrix2)[6908:9210] <- ''
tr2 <- dtmAll_matrix
te <- dtmAll_matrix2
CL <- factor(rownames(tr))

Knn_predict2 <- knn(tr2[1:200,], te[1:200,], CL2[1:200])
Knn_predict2
```