“智慧政务”中的文本挖掘应用

摘要

文本数据挖掘的关键在于对数据预处理结果的好坏，中文文本从形式上来看是由汉字组成的字符串，不同的汉字组成不同的词语，句子段落。中文文本和英文最大的区别是英文有空格进行分割，所以对中文文本进行预处理是首要。

我们对文本数据进行数据清洗，删除缺失的数据，去掉重复的数据。再使用基于字符串匹配的分词方法，对文本数据进行分词处理，去停用词，去敬辞等操作。得到干净的中文词组。再利用 TF-IDF 编码将文本数据转化成计算机可以识别的语言。

对于问题一，我们用基于 KNN 算法的分类模型，对附件 2 的用户留言进行分类，利用附件 2 已有的分类标签，我们可以利用所给的 F-score 对建立得分类模型进行评价。得到的准确度是 95.407%。则我们认为分类模型是有效的，分类结果是可以接受的。

对于问题二，我们利用基于关联规则挖掘的分类模型，通过考虑留言的频率、时间和点赞数来建立评价指标，得到关于留言的热度指标。再对留言进行分类处理，得到热点问题明细表热点问题排名前五的留言。

对于问题三，我们通过建立相关性评价指标模型，对相关留言关键问题字眼对答复意见中的关键词进行比对匹配，对量化指标属性赋予权重，对留言和答复意见进行距离的相似度计算，从而建立关于答复意见的评价模型，评价答复意见的完整性，准确性以及答复质量。

关键词：文本数据挖掘 数据预处理 KNN 算法 关联规则挖掘

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。

文本挖掘技术释义。文本挖掘（Text Mining）是数据挖掘的一个分支， 是指从大量文本数据中抽取事先未知 的、可理解的、可操作的、最终可运 用的有价值的知识的过程，其主要用 途是从未经处理的原始文本中提取未 知的知识，同时运用这些知识更好地 组织信息以便将来参考。文本挖掘是 一个多学科混杂的领域，涵盖了多种 技术，包括统计数据分析、数据挖掘、 信息抽取、机器学习等。在现实世界中， 可获取的大部分信息都是通过文本的 形式储存在文本数据库中，由于电子 形式的文本信息飞速增长，文本挖掘 已经成为信息领域的研究热点。

问题分析

问题 1 根据附件 2 中的数据，建立关于留言内容得一级标签分类模型并根据所给的 F-score 对建立得分类模型进行评价。在附件 2 中所给的数据中已经给出一级标签分类，这是作为对所建立分类器进行评价得指标，也即评价指标训练集。需对附件 2 中得评价进行文本数据相应的处理，根据附件 1 中的三级标签、二级标签及一级标签建立分类器模型。也即将附件 1 中的数据作为分类器模型的比对参数，建立比对参数数据表。三级标签、二级标签对应的一级标签作为一级标签分类，同时一级标签本身也是一个分类比对参数。以附件 1 中的分类标签作为分类器的分类比对参数，建立分类比对标签表，对数据处理后的附件 2 中的留言进行关键字比对建立分类器，再根据所给的评价模型 F-score 和附件二中对应的一级标签对建立得分类器进行评价。

问题 2 这是一个热点问题挖掘，建立一个分类器对附件 3 中留言的关键“问题”词进行相似对比并统计词频，但是要保证时间段的统一。建立一个热点问题评价指标模型，根据数据处理的结果和评价指标模型求出评价结果，并给出相应的表格(去前五)，并按相应格式给出相应热点问题对应的留言信息表。

判断是否是热点的关键因素是：频率和时间。某一段时间内群众集中反映的某一问题可称为热点问题。其中“某一段时间”反映的是时间段；而“群众集中反映的某一问题”反映的是在时间段的前提下出现该问题的频率。在热点问题的挖掘中，最为重要的是数据的预处理，尤其是分词和去词， 影响最终的结果。

问题 3 建立相关性评价指标模型，对相关留言关键问题字眼对答复意见中的关键词进行比对匹配，根据建立得模型指标化，给出相关性量化指标；完整性可以根据问题处理的一般步骤对答复意见进行量化处理，给出完整性量化指标；可解释性将其进行量化处理。从不同角度建立不同量化指标，将多元化指标归一化处理，可以很明确的判断其答复意见的质量。

模型假设

1. 假设所有的留言都是真实有效的，所有用户反应的情况真实存在
2. 假设所有用户在留言时都实事求是，不带有个人情感色彩以及夸大瞒报等行为
3. 假设所有的留言没有恶意举报的行为

符号说明

表 1：符号说明

符号	说明
T	训练集
P_i	第i 类查准率
R_i	第i 类查全率
$x_i, \left(i=1,2,\cdots n\right)$	特征向量
$y_i, \left(i=1,2,\cdots n\right)$	实例的类别

模型建立

数据预处理

中文文本从形式上来看是由汉字组成的字符串，不同的汉字组成不同的词语，句子段落。中文文本和英文最大的区别是英文有空格进行分割，所以对中文文本进行分词处理是首要任务，也是重中之重。

中文分词是指利用一定的规律和方法将中文文本分割成词语，常用的分词方法有基于字符串 匹配的、基于理解的和基于统计的分词方法。我们这里使用基于字符串匹配的分词方法。

基于字符串匹配的分词方法是指在已有字典的基础上，按照指定的规则进行匹配，直到完成规则中的“最大”匹配，则识别出一个词。按照匹配的方向不同，基于字符串匹配方式的不同又可以分为：正向最大匹配、逆向最大匹配、双向最大匹配。

首先我们收集字典数据，如图 1，

大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯
杆 圈 西湖 建筑 集团 燕子 山 安置 房 项目
施工 围墙 上下班 期间 条 路上 人流 车流 安全隐患 请求
文明城市 整改 文明 位于 书院 路 主干道 在水一方 大厦 一楼 四楼
人为 拆除 水电 设施 烂尾 多年 护栏 围着 占用 锈迹斑斑 倒塌
危机 过往行人 车辆 牵头 尊敬 苑 火炬 物业 程明 物业管理
有限公司 未经 小区业主 同意 利用 公摊 公共 面积 滥收 停车费
用且 收费 区内 损坏 拒 承担责任 社区 置之不理 妄自 违规
不闻不问 今特请 明查 事实 老百姓 做主 督办 依法 依规 应享

图 1

根据上述分词方法我们可以将用户留言进行分词处理，如图 2

Index	留言详情
0	['大道', '西行', '便道', '未管', '所', '路口', '至', '加油站', '路段', '人行道', '包括', '路灯', '杆', '被', '圈', …
1	['位于', '书院', '路', '主干道', '的', '在水一方', '大厦', '一楼', '至', '四楼', '人为', '拆除', '水电', '等', '设施…
2	['尊敬', '的', '领导', '苑', '小区', '位于', '火炬', '路', '小区', '物业', '程明', '物业管理', '有限公司', '未经', '…
3	['华庭', '小区', '高层', '为', '二次', '供水', '楼顶', '水箱', '长年', '不洗', '现在', '自来水', '龙头', '的', '水'…
5	['我', '在', '年', '购买', '了', '盛世', '耀凯', '小区', '栋', '楼楼', '两层', '共计', '千', '平方', '一直', '以来…
6	['由于', '西地省', '地区', '常年', '阴冷', '潮湿', '的', '气候', '加之', '近年', '气候', '逐渐', '更加', '恶劣', '…
7	['尊敬', '的', '胡书记', '您好', '家住', '桐梓', '坡', '西路', '可可', '小城', '的', '居民', '长期以来', '经常', '…
8	['我们', '是', '梅家田', '社区', '辖区', '内', '的', '小区', '居民', '我们', '每年', '都', '依法', '依规', '向', …
9	['尊敬', '的', '政府', '领导', '您们好', '我', '是', '魏家坡巷', '的', '业主', '多年', '以来', '我们', '小区', '的'…
11	['请求', '依法', '监督', '泰华', '一村', '小区', '第四届', '非法', '业主', '委员会', '涉嫌', '侵占', '小区业主', '公…
12	['我', '住', '在', '梅', '溪湖', '壹号', '御湾楼', '自年', '月份', '住', '进来', '每天晚上', '都', '会', '停水', '白…
13	['尊敬', '的', '领导', '你们好', '我', '是', '捞刀河', '镇', '彭家巷', '社区', '鸦涛', '翡翠', '湾', '的', '一名…
14	['地铁', '号线', '施工', '导致', '万家', '丽路', '锦楚', '国际', '星城', '小区', '三期', '一个月', '停电', '来次',…
15	['尊敬', '的', '领导', '你好', '我', '是', '润', '和', '紫', '郡', '的', '业主', '今年年初', '我们', '小区', '周…

图 2

第二步是去停用词，在中文表达中，我们常常为了使句子更加通顺连贯，能够充分表达意思和情感，通常会加入一些多功能词汇，例如中文中的“的”、“了”、“着”等等，这些词并无实际意义，去掉后不影响文本的意思表达，同时能够提高机器学的速度，为此我们要对中文文本去停用词。

第三步是去问候语句，在中文表达中，人们常常会加入一些敬辞来体现礼貌，或者在某些场合会使用一些模板，例如在本题中某用户留言如下图 3 所示，

尊敬的胡书记：您好！家住A市A3区桐梓坡西路可可小城的居民长期以来经常停水。小区业主、业委会多次找物业、开发商、居委会、自来水厂等单位寻求帮助至今没有找到具体停水原因，也没有相关单位承诺给与解决。停水问题严重影响小区居民日常生活，实在苦不堪言！恳请书记出面解决民生基本问题，在此感激不尽！

图 3

其中“尊敬的胡书记：你好”以及“在此感激不尽”等语句对文本的意思并

无任何作用，所以在文本处理时应该删除这些语句。

第四步，为了让计算机理解文本，我们需要将文本词汇转化成计算机能够识别的“文字”，据此产生了文本表示方法，常用的文本表示方法有 one-hot 编码、词袋模型（Bag of Words）、N-gram 模型和 TF-IDF 技术。这里我们利用 python 中 jieba 库的 TF-IDF 算法对文本进行处理。之后对同一类（如，城乡建设）的所有留言的关键词进行汇总并且去重，后保存成 TXT 文件，构建为这类的词库。

第四步生成词云，对文本数据进行处理后，所得到的数据表现形式为词组，对这些数据进行统计生成词云如下图 4 所示

图 4

问题一模型建立

根据附件一所给的分类标签，我们可以构建分类算法对留言进行分类，常见的分类算法有决策树（ID3 和 C4.5）、朴素贝叶斯、基于关联规则的分类和 KNN 算法，这里我们使用 KNN 分类算法来进行分类。

kNN 算法的核心思想是如果一个样本在特征空间中的 k 个最相邻的样本中的

大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。基本步骤如下：

（1） 输入要训练的数据集，

$$T=\{(x_1,y_1),(x_2,y_2)\cdots,(x_N,y_N)\}$$

其中

$$x_i\in X\subseteq R^n$$

为 n 维的实例特征向量。

$$y_i\in Y=\{c_1,c_2,\cdots,c_K\}$$

为实例的类别，其中，i=1,2,…,N，预测实例 x

（2） 根据给定的距离量度方法（一般情况下使用欧氏距离）在训练集 T 中找出与 x 最相近的 k 个样本点，并将这 k 个样本点所表示的集合记为 N_k(x)；

（3） 根据如下所示的多数投票的原则确定实例 x 所属类别 y：

$$y=\arg\max\sum_{x_i\in N_{k(x)}}I(y_i,c_j),\ i=1,2,\cdots,N;\ j=1,2,\cdots,K$$

上式中 I 为指示函数：

$$I(x,y)=\begin{cases}1,\ if\ x=y\\0,\ if\ x\neq y\end{cases}$$

（4） 输出：预测实例 x 所属类别 y。

这样我们可以得到基于 KNN 算法的分类模型

问题一模型求解

根据上述的分类模型，我们可以对数据进行处理，得到如下结果如图 5

	留言编号	留言用户	留言主题	留言时间	留言详情	一级标签	再分类	结果（对：1，错
4364	172242	U0005269	C4市第一中	2016/8/3 1	第一中学无视省教育厅下发的暑假	教育文体	教育文体	1
6016	131769	U0007530	B市临聘人	2015/9/22	首先我也是初次来这个网站我也不	劳动和社会保障	劳动和社会保障	1
2497	117709	U0007280	K市珊瑚东	2016/8/3 1	我住在珊瑚东路的一栋楼房上楼下有个广	环境保护	环境保护	1
7821	149258	U0006797	M11县锦绣	2019/12/24	锦绣名城各栋楼电梯年检过期很久有的电	商贸旅游	商贸旅游	1
5101	347523	U000470	咨询E7县F	2019/11/26	尊敬的上级领导我叫阳艳美今年岁家住司	教育文体	教育文体	1
4524	228656	U0004572	K3县白水	2017/7/24	作为一名教育工作者今特借平台这	教育文体	教育文体	1
1759	161227	U000684	请查处K8	2012/10/1(	高厅长您好我在这里要向您反映一下房地	城乡建设	城乡建设	1
6927	318319	U0001534	深圳社保如	2019/5/23	我于年起在社保局购买了医疗仪	劳动和社会保障	教育文体	0
1912	174572	U0001696	L市西地省	2012/4/12	我购买的河西西地省西部小商品市场二期	城乡建设	城乡建设	1
6599	255351	U0004264	退休时档案	2018/3/16	胡厅长你好我今年已经到了退休年	劳动和社会保障	劳动和社会保障	1
1244	119776	U0007523	K1区城南	2016/4/26	高局长你好我于年购买了城南路号南苑小	城乡建设	城乡建设	1
6211	159494	A0002014	关于劳务派	2012/6/29	主席你好我是一名企业建筑员工合同形式	劳动和社会保障	劳动和社会保障	1
8018	154048	U0004782	J11市保健	2018/8/13	区几十家保健品理疗店以免费发礼品免费	商贸旅游	商贸旅游	1
5303	21601	U0006312	望好心的领	2013/2/26	对年以后年以前参加工作年以前退	劳动和社会保障	劳动和社会保障	1
1088	108683	U0001667	请问K2区	2013/11/1(	请问马坪农业开发区是什么时候纳入总体	城乡建设	城乡建设	1
2950	3882	U0006002	A市楚府路	2018/12/11	楚府路快速化改造在年启动说是年要全线	交通运输	交通运输	1
8967	219026	U0007839	J市民咨询	2017/5/12	我老婆是再婚之前有两个小孩判给	卫生计生	卫生计生	1
7577	117370	U0007137	K3县检测	2019/6/23	年月政府为规范检测市场价格市发改委根	商贸旅游	商贸旅游	1
2217	67487	U0006842	E8县又兰	2019/7/25	金桥建材砖厂自投资建厂以来一直存在着	环境保护	环境保护	1
3549	4367590	U000749	320国道大	2014/8/8 9	尊敬的陈书记您好您幸苦了从大修国道以	交通运输	交通运输	1
9019	240074	U000585	投诉K7县	2017/10/21	陈主任你好我要投诉江水县妇幼保	卫生计生	卫生计生	1
8081	154133	U0007012	A市西子湖	2017/7/9 1	七月六号晚十点二十几分 西子湖畔沃府	商贸旅游	商贸旅游	1
5831	106161	U0007604	关于省直单	2017/3/30	今年以来我省省直单位公开选拔工	劳动和社会保障	劳动和社会保障	1
1447	137731	U0002484	M市公交车	2018/1/30	公交车自大前年开始渐步使用空调新车票	城乡建设	商贸旅游	0
6739	286415	U0007888	I6市社保局	2018/10/24	社保局未落实上级部门关于职工生	劳动和社会保障	劳动和社会保障	1
								95.41%

图 5

再根据题目所给的评价公式对处理的数据进行查准查全，通常使用 F-Sore 对分类方法进行评价

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率

根据上述公式对附件 2 进行分类的准确率是 95.407%。

部分“异常”数据如下图 6 所示

	留言编号	留言用户	留言主题	留言时间	留言详情	一级标签	再分类	结果（对：1，错	0）
8250	160181	U0004279	H市旅游网	2016/2/24	去旅游后有一肚子话想说于是去了旅游网	商贸旅游	城乡建设	0	0.312591
8768	129127	U0008278	长株潭大市	2015/5/28	我星期天上午在百花农贸市场一个	卫生计生	城乡建设	0	0.795513
5453	44078	A0003634	C市C2区教	2011/12/15	陈书记你好教师的社保基金好几年就没有	劳动和社会保障	城乡建设	0	0.701089
7990	154009	U0002094	B4区黄山路	2019/2/21	黄山路建设家园栋单元电梯掉层我从楼掉	商贸旅游	城乡建设	0	0.594255
6502	236044	U0006967	反映工人等	2014/11/21	年的工人等级考试什么时候发放到	劳动和社会保障	城乡建设	0	0.145866
7983	154000	U0002094	B4区黄山路	2019/3/7 1	黄山路建设家园栋单元电梯掉层我从楼掉	商贸旅游	城乡建设	0	0.672208
5945	120817	U0006847	通过考试的	2015/5/30	胡厅长我是的事业单位的员工岁工	劳动和社会保障	教育文体	0	0.225684
1591	153670	U0005156	M11县建设	2013/3/29	年月下旬中学在围墙施工时垮塌造成人死	城乡建设	劳动和社会保障	0	0.311828
8341	4083	A0001058	西地省独生	2018/11/2(	请问独生子女证有哪些好处我们这一群体	卫生计生	城乡建设	0	0.87391
7890	153860	U0003515	A市天普教	2019/8/29	各位领导年月我从上看到天普教育的注册	商贸旅游	教育文体	0	0.352324
689	75376	U0004590	保护湿地公	2017/4/16	春暖花开万物复苏时值周末江湿地公园人	城乡建设	教育文体	0	0.33043
3201	114303	U0005428	K4县春运期	2014/1/20	尊敬的县长您好春节应该是欢乐团圆的好	交通运输	城乡建设	0	0.475832

图 6

问题二模型建立

利用基于关联规则挖掘的分类器，我们构建热点问题挖掘的模型。

关联规则挖掘发现大量数据中项集之间有趣的关联或者相互联系。在本题中我们提取到的文本词汇即为项集，通过挖掘一些词汇之间的关联来得到所谓的“热点”问题。

我们首先需要对热点进行定性，即何为热点，我们通过考虑频率，点赞数，以及热度指数对热点问题进行定性分析，据此建立关于热点问题的分类标签，再通过基于关联规则的分类器对留言进行分类，筛选出最受关注的热点问题。

问题二模型求解

通过分类模型，我们对所有的留言进行处理，得到热点问题明细表如下图 7 所示

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	188399	A00097934	A市利保壹	2019/7/3	您好，我想	0	0
1	189587	A00018292	对A8县高新	2019/1/13	我们是西地	0	0
1	193286	A00010319	居住在地铁	2019/4/17	沈书记：您	32	0
1	194358	A0005583	A市物业公	2019/2/22	物业公司进	0	0
1	195686	A00091996	A市润万滨	2019/10/1	润万滨江天	16	2
1	196282	A00039390	反映A市地	2019/9/6	张县长，您	16	0
1	197240	A00010713	西地省人民	2019/10/2	西地省人民	0	0
1	197951	A00035453	投诉A市大	2019/7/23	大汉汉园开	5	0
1	198111	A00053914	A3区金晖仪	2019/2/25	金晖优步花	0	0
1	198765	A00010908	A2区鑫远湾	2019/6/2	尊敬的领导	0	0
1	199047	A00078733	举报A8县翠	2019/3/11	西地省A市	0	0
1	199492	A00027830	反映A市美	2019/1/1	我们是美麓	1	0
1	201518	A00010370	投诉A3区中	2019/9/19	您好！我们	0	0
1	201707	A00010319	反映A市地	2019/4/17	胡书记：您	9	0
1	202867	A909147	祥源控股效	2019/9/20	近日，西地	0	0
1	203345	A00095628	17年A3区商	2019/6/26	17年村民天	0	0
1	206818	A00028611	投诉A市东	2019/5/29	投诉西地省	2	1
1	208618	A00033673	居然在A9市	2019/6/30	尊敬的领导	0	0
1	210513	A00010389	A市晟福兴	2019/3/11	尊敬的胡书	9	0
1	215273	A00046349	投诉A3区住	2019/11/1	佳境小区2	0	0
1	216920	A00077723	请解决A7县	2019/5/6	尊敬的政府	2	0
1	219274	A00010888	A市梅溪湖	2019/8/9	本人封明亮	0	0
1	220052	A00063081	A市新东方	2019/1/19	尊敬的领导	1	0

图 7

然后根据属性标签从所有热点问题中筛选出前五的留言如下图 8 所示

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	3206	2019/1/1至2019/9/29	西地省	西地省各处施工带来给居民带来干扰
2	2	2870	2019/11/2至2020/1/26	丽发新城小区	丽发新车小区附近搅拌站噪声大粉尘多给居民生活与健康带来影响
3	3	1586	2019/1/12至2020/1/7	西地省	西地省各处居民区大量开设麻将馆并经营至深夜，影响居民休息
4	4	1273	2019/3/16至2019/9/1	铁路职工	政府优待铁路职工政策在集团执行过程中存在捆绑销售等不良行为
5	5	706	2019/7/21至2019/12/4	A市A5区魅力之城	小区附近餐饮店油烟噪音扰民

图 8

问题三模型建立

建立留言的答复意见评价指标模型，其核心是所给的答复能否解决留言所提到的问题，需要从答复的相关性、完整性、可解释性等角度对答复意见的质量进行评价。

首先对评价模型的属性标签指标化，权衡每个属性的重要程度给予权重，给出量化指标。

其次，评价答复的完整性，我们从答复意见的处理步骤以及解决方案来进行量化处理。然后结合各个角度，建立统一完整的量化指标，判断答复意见的质量。

最后，我们将留言和答复意见进行矢量化处理，通过对比答复意见与留言的距离来判断答复意见的合理性以及完整性。

问题三模型求解

通过上述评价体系，我们以百分制来对答复意见进行打分，对附件 4 中的答复意见打分后的结果如下图 9 所示

	B	C	D	E	F	G	H
	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	打分
2	A00045581	2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	物业公司却以交20万保证金，不	意收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	78
3	A00023583	潇楚南路洋湖段怎么还没修	2019/4/24 16:03:40	店面的生意带来很大影响，里面	，需整体换墙，且换墙后还有三趟雨污水管道	2019/5/9 9:49:10	89
4	A00031618	快提高A市民营幼儿园老师的	2019/4/24 15:40:04	的同时更是加大了教师的工作压	民办幼儿园聘任教职工要依法签订劳动合同，依	2019/5/9 9:49:14	93
5	A000110735	公寓能享受人才新政购房补	2019/4/24 15:07:30	落户A市，想买套公寓，请问购	年龄35周岁以下（含），首次购房后，可分别	2019/5/9 9:49:42	75
6	A0009233	于A市公交站点名称变更的建	2019/4/23 17:03:19	“马坡岭小学”，原“马坡岭	保留“马坡岭”的问题。公交站点的设置需要	2019/5/9 9:51:30	85
7	A00077538	A3区含浦镇马路卫生很差	2019/4/8 8:37	再把泥巴冲到右边，越是上下班	于您问题中没有说明卫生较差的具体路段，也	2019/5/9 10:02:08	69
8	A000100804	区教师村小区盼望早日安装电	2019/3/29 11:53:23	台为老社区惠民装电梯的规范性	A市A3区人民政府办公室下发了《关于A市A3区	2019/5/9 10:18:58	33
9	UU00812	区东澜湾社区居民的集体民	2018/12/31 22:21:59	跑好远，天寒地冻的跑好远，眼	装修前期准备及设施设备采购等工作。下一步，	2019/1/29 10:53:00	64
10	UU008792	襄阳光住宅楼无故停工以及	2018/12/31 9:55:00	也没得到相关准确开工信息。附	任单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	71
11	UU008687	城和顺路洋湖壹号小区路段公	2018/12/31 9:45:59	立交桥等地方做立体绿化，取得	余部分也按规划要求完成了建设，其中西边绿化	2019/1/16 15:31:05	48
12	UU0082204	A2区大托街道大托新村违建	2018/12/30 22:30:30	乡规划局审批通过《温室养殖大	公司支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	87
13	UU008829	鄱阳村D区安置房人防工程的	2018/12/29 23:27:51	D区安置房地下室近两万平方米	续，按长人防发[2014]7号文件要求，鄱阳村	2019/1/29 10:52:01	68
14	UU00877	区段请求修建一座人行天桥	2018/12/29 11:55:34	峰，大量从小区开车出去的业主	划分局配合进行具体选址，招标（邀标）进行	2019/1/14 14:34:58	79
15	UU0081480	举报A市芒果金融平台涉嫌诈骗	2018/12/28 17:18:45	贵省相关政府部门的大力支持！	映的相关警情，已由银盆岭派出所立刑事案件	2019/1/3 14:03:07	97
16	UU0081227	建议增开A市261路公交车	2018/12/28 7:53:25	小时以上！天寒地冻，其他公交	运正常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	66

图 9

模型评价与总结

我们首先解决了用户留言分类问题，通过文本数据挖掘，利用 KNN 算法建立分类模型对用户留言进行一级标签分类。模型有很多优点与不足，分类模型使用 KNN 算法是最为简单和实用的一种方式，尤其是对于文本数据挖掘，且准确率高，存在的不足和缺点也很明显，对数据要求较为苛刻，如果存在一两条错误数据，

且刚好处在分类边缘，则会导致结果大相径庭。另外 KNN 算法的本身的缺陷之一就是维数灾难，当数据量过于庞大会导致处理时间过长和运行缓慢等问题，有待进一步优化。

对于热点问题的建立，我们利用基于关联规则的分类模型，通过频率、时间以及点赞数等属性标签对所谓的“热点”问题进行挖掘，找出频繁项集，就可以得到所需的热点问题。这是关联规则的变形和扩展，并没有直接去寻找热点问题，而是先通过寻找频繁项集，再来确定热点问题。

第三问的评价模型从答复意见的完整性，准确性等多方面各个角度考虑，能够合理有效的评价答复意见的质量。

引用文献

[1]李湘东,徐朋,黄莉,沈祥兴.基于 KNN 算法的文本自动分类方法研究——以学术期刊栏目自动归类为例[J].图书情报知识,2010(04):71-76.
[2]张影.基于数据关联与文本挖掘技术的图书馆文献资源开发利用研究[J].中国中医药图书情报杂志,2019,43(04):48-51.
[3]徐蕾,张科伟.基于文本挖掘的京东商品评论分析[J].内蒙古科技与经济,2020(03):41+43.
[4]贾璇.基于文本挖掘的求职软件顾客评论情感分析[J].科技与创新,2019(17):1-4.
[5]陈曦.文本挖掘技术在社情民意调查中的应用[J].中国统计,2019(06):27-29.

附录

数据预处理

```
# -*- coding: utf-8 -*-
"""
Created on Mon May  4 16:46:04 2020

@author: 16277
"""

import pandas as pd
import os
import re
import jieba
def compress(sentence):
    #一字词压缩：重复 2 次以上
    for i in [1]:
```

```
        for j in range(len(sentence)):#重复开始的位置
            if sentence[j:j+i]==sentence[j+i:j+2*i] and sentence[j+i:j+2*i]==sentence[j+2*i:j+3*i]:
                k=j+2*i
                while sentence[k:k+i]==sentence[k+i:k+2*i] and k<len(sentence):
                    k+=i    #k=k+i
                sentence=sentence[:j]+sentence[k:]
    #二字词或以上
    for i in range(2,int(len(sentence)/2)+1):
        for j in range(len(sentence)):
            if sentence[j:j+1]==sentence[j+i:j+2*i]:
                k=j+i
                while sentence[k:k+i]==sentence[k+i:k+2*i] and k<len(sentence):
                    k+=i
                sentence = sentence[:j]+sentence[k:]
    return sentence
#%%
os.chdir('D:\\360MoveData\\Users\\16277\\Desktop\\泰迪杯\\C 题全部数据')
data=pd.read_excel('附件 2.xlsx')
data00=data
os.sep
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('\t','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('\n','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('\r','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('[0-9]','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('[A-Z]+[市县区]','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('[A-Z]+[0-9]+[市县区]','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('[a-z]','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('[A-Z]','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('&[a-z]+;','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('\u3000','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('\u2022','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('\xa0','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('\xa9','',x))
data00['留言详情']=data00['留言详情'].astype(str).apply(lambda x: re.sub('[~`!#$%^&*()_+-=|\';":/.,?><~ •！ @#￥%……&*（）——+-=“：’；、。，？》《{}]','',x))
ind03=data00['留言详情']==''
sum(ind03)
ind04=data00['留言详情']=='nan'
sum(ind04)
ind05=data00['留言详情'].notnull()
data00=data00.ix[-ind03&-ind04&ind05,:]
```

```
data00['留言详情']=data00['留言详情'].apply(lambda x :compress(x))

#%%
#分词
#jie.cut(data2['评论'])
comment_cut0=data00['留言详情'].apply(lambda x: jieba.lcut(x))

#%%

###去除停用词
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/stoplist.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop #Pandas 自动过滤了空格符，这里手动添加
comment_cut0=comment_cut0.apply(lambda x: [i for i in x if i not in stop])
test=comment_cut0

#%%城乡建设
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/cxjs.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop
cxjs=test.apply(lambda x: [i for i in x if i not in stop])

data00['cxjs']=0
for i in range(len(cxjs)):
    data00['cxjs'][i]=len(test[i])-len(cxjs[i])

#%%环境保护
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/hjbh.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop
hjbh=test.apply(lambda x: [i for i in x if i not in stop])
data00['hjbh']=0
for i in range(len(hjbh)):
    data00['hjbh'][i]=len(test[i])-len(hjbh[i])

#%%交通运输
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/jtys.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop
jtys=test.apply(lambda x: [i for i in x if i not in stop])
```

```
data00['jtys']=0
for i in range(len(jtys)):
    data00['jtys'][i]=len(test[i])-len(jtys[i])

#%%教育文体
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/jywt.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop
jywt=test.apply(lambda x: [i for i in x if i not in stop])
data00['jywt']=0
for i in range(len(jywt)):
    data00['jywt'][i]=len(test[i])-len(jywt[i])

#%%劳动和社会保障
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/ldhshbz.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop
ldhshbz=test.apply(lambda x: [i for i in x if i not in stop])
data00['ldhshbz']=0
for i in range(len(ldhshbz)):
    data00['ldhshbz'][i]=len(test[i])-len(ldhshbz[i])

#%%商贸旅游
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/smly.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop
smly=test.apply(lambda x: [i for i in x if i not in stop])
data00['smly']=0
for i in range(len(smly)):
    data00['smly'][i]=len(test[i])-len(smly[i])

#%%卫生计生
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/wsjs.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop
wsjs=test.apply(lambda x: [i for i in x if i not in stop])
data00['wsjs']=0
for i in range(len(wsjs)):
    data00['wsjs'][i]=len(test[i])-len(wsjs[i])
```

```
#%%
data11=data00[['cxjs','hjbh','jtys','jywt','ldhshbz','smly','wsjs']]
data00['一级标签']='其他'
for i in range(len(data00['一级标签'])):
    if max(data11.iloc[i,:])==data11.iloc[i,0]:
        data00['一级标签'][i]='城乡建设'
    elif max(data11.iloc[i,:])==data11.iloc[i,1]:
        data00['一级标签'][i]='环境保护'
    elif max(data11.iloc[i,:])==data11.iloc[i,2]:
        data00['一级标签'][i]='交通运输'
    elif max(data11.iloc[i,:])==data11.iloc[i,3]:
        data00['一级标签'][i]='教育文体'
    elif max(data11.iloc[i,:])==data11.iloc[i,4]:
        data00['一级标签'][i]='劳动和社会保障'
    elif max(data11.iloc[i,:])==data11.iloc[i,5]:
        data00['一级标签'][i]='商贸旅游'
    elif max(data11.iloc[i,:])==data11.iloc[i,6]:
        data00['一级标签'][i]='卫生计生'

#%%\
data['再分类']=data00['一级标签']
data.to_excel('D:\\360MoveData\\Users\\16277\\Desktop\\泰迪杯\\附件 3.xls')
```

热点问题代码实现

```
# -*- coding: utf-8 -*-
"""
Created on Wed May  6 15:12:40 2020

@author: 16277
"""
from jieba.analyse import *
import numpy as np
import pandas as pd
import jieba
import jieba.analyse
import os
from textrank4zh import TextRank4Keyword, TextRank4Sentence
from snownlp import SnowNLP
import copy
os.chdir('D:\\360MoveData\\Users\\16277\\Desktop\\泰迪杯\\C 题全部数据')
data=pd.read_excel('附件 3.xlsx')
data1=copy.deepcopy(data)
```

```
#%%
def keywords_textrank(text):
    keywords = jieba.analyse.textrank(text, topK=10)
    return keywords


 #%%
keyword=data1['反对数']

for i in range(len(data1)):
    if __name__ == "__main__":
        text = data1['留言详情'][i]
        keyword[i]=keywords_textrank(text)
#%%
with open('D:/360MoveData/Users/16277/Desktop/泰迪杯/stoplist.txt',encoding='utf-8')as f:
    stop = f.read()
stop = stop.split()
stop=[' ']+stop #Pandas 自动过滤了空格符，这里手动添加
keyword=keyword.apply(lambda x: [i for i in x if i not in stop])


#%%
com=keyword
for i in range(len(com)):
    com[i]=list(set(com[i]))


#%%
###

tmp=com.apply(lambda x: ' '.join(x))#去除逗号等符号
tmp2=' '.join(tmp)#去除列表符号
num = pd.Series(tmp2.split()).value_counts()

#%%

for i in range(len(com)):
    for j in range(len(com[i])):
        com[i][j]=num[com[i][j]]
#%%
com1=com.to_frame()
com1['热度']=0
for i in range(len(com)):
    if com[i]!=[]:
```

```
            com1['热度'][i]=sum(com[i])/len(com[i])
        elif com[i]!=[]:
            com1['热度'][i]=0

#%%
data['热度']=com1['热度']
data.to_excel('D:\\360MoveData\\Users\\16277\\Desktop\\泰迪杯\\aaa3.xlsx')
```