“智慧政务”中的文本挖掘

摘要

近几年来，随着微博、微信、市长信箱、阳光热线等一系列的网络问政平台的发展，居民们可以通过这些网络平台对政府进行了解民意、汇聚民意以及去凝聚民意，有了这些居民的反映，政府也要及时的去了解群众所反映的内容，然后接纳和吸取好的群众意见，改正坏的旧习。由于各类社情民意相关的文本数据量不断地提高，在以往，主要是通过人工来进行读取、筛选、汇总、总结等工作的，但是随着数量的递增，要处理这些数据对人工读取来说已经是一项巨大的考验，人工读取需要消耗大量的人力、物力以及财力，还有很多人为的误差，如今随着云计算、大数据以及人工智能的发展，这些问题得到了解决。针对群众的问政留言数据，我们将会用自然语言和文本挖掘的的方式解决。

针对任务一，在处理网络问政的群众留言时，我们以附件一的参考体系对附件二的信息先进行分类，以便于后续读取数据的方便，基于附件一的划分体系，我们了解到三级标签是二级标签的真子集，而二级标签是一级标签的真子集，所以，在建立关于留言内容一级标签的分类模型时，我们可以采取从三级标签入手，然后到二级标签，最后到一级标签，逐一突破，在分类时，由于在短期内大量发送信息,容易导致短时间内大量数据的堆积,并且这些数据信息参差不齐,为了精准的提取有效信息,本文引入了 LDA 主题模型,主题模型的使用可以在大批量的数据中快速找到相关信息。

针对任务二，我们知道了某一时段内群众集中反映的某一问题称为热点问题，在这里，我们也可以利用 LDA 主题模型参与热点问题的查找，首先，我们将附件三按照时间和点赞数进行数据排序，利用热点问题的相关数据算出热点，及时发现热点问题，定义合理的热度评价指标，并给出评价结果，然后进行总结，有助于相关部门进行有针对性地处理，提升部门服务效率。

针对任务三，主要是针对附件四，用户的留言在相关部门得到之后，要进行一定的答复，答复必须做到不重不漏，这样做的目的有助于民众发挥积极性，所以我们要从回复的内容中，从相关性、完整性、可解释性以及历史性等方面进行分析。这里可以应用 SPSS 软件对留言详情和答复意见进行相关性分析。

关键词：LDA 主题模型、 热点、SPSS 软件

目录

一、 问题分析

本题给出了四个附件数据，附件一是问题的三个级别的分类共有三列，每一列是一个级别的分类。附件二、三、四都是一些留言问题的记录，其中的记录信息包括：留言、编号、用户、主题、时间、详情、答复、点赞数等等。

问题一是让我们根据附件一中的标签体系对附件二中的具体的留言内容建立模型进行分类。同时在应用我们建立的模型分类之后，使用指定的方法来对我们建立的模型进行检验。

在这个问题中我们利用关键词来进行匹配，在我们说的话中每一句话都有一个最关键的词语很多时候这个词语往往决定着我们这句话想说的是什么，在这个问题中我们需要考虑，由於中国文化的博大精深，所以我们的词语需要放在句意中来进行理解，也就可能有些信件利用关键词没法分类。

问题二中，由于信件很多同时通过网络来反馈民生问题的方式方便快捷不需成本，而由于信件数量的庞大，所以有可能导致一些急需解决的紧要问题没法得到及时的解决，针对这个问题提出了热点问题，在这里我们需要定义一个指标，当某一个问题提出的人数达到一定的标准或者说点赞这个问题的人数达到标准，那么我们就称为这个问题是热点问题是需要我们赶快解决的。在这个问题中我们需要注意怎样制定合适的指标来判定一个问题是否为热点问题。

问题三是根据附件四来解决的，当人们提出问题提出反馈就需要得到解决需要得到回答，在附件四中给出了部分信件留言问题、时间、内容等的基础上增加了回复时间及回复内容。

要判断这个回答的质量高不高我们需要从以下几个方面来进行判断，第一个判断的指标是答复的相关性，也就是这个答案与我们的问题有没有关系；第二个判断的指标是完整性，即是答案回答的完不完整；第三个指标是可解释性，就是答案能不能解决问题；最后一个指标是及时性，这个指标可以用来判断该政府解决问题的效率。

通过这些指标我们就可以得出，相关部门对留言的答复的质量如何，从而进行改进。

二、 数据准备

2.1 剔除异常样本

根据本文的问题观察所给数据我们会发现，在附件所给的数据中会出现，同意留言用户针对同一问题在不同的时间内提交了同一封信件。在这里为了不对我们的信息处理造成误导我们选择将重复的信件保留且只保留一份。

2.2 删除用不到的数据

通过分析观察在解决问题之前由于本次数据庞大，一些用不到的指标会增加我们解决问题的难度，所以我们仅保留分析使用到的指标。经过将重复的问题删减之后，我们保留的数据有：

附件二中保留数据留言主题、标签体系。附件三中保留有言主题、点赞数。附件四中保留留言主题、留言时间、留言详情、答复意见、答复时间。

2.3 构造分析需要的指标

基于分析需求构造如下指标：

1）关键词

关键词一词来源于英文中的 keywords,它特指与某网页主题有关，并且可通过搜索引擎搜索到某类网站的词语，也是某个媒体在制作使用索引的过程中，所用到的词。现在很多网络搜索索引最主要的搜索方式之一就是关键词搜索，简单来说它就是一个访问者或者调研者们了解的产品、服务或者公司等的具体名称的一个用语。它有核心关键词（网站的核心产品和内容的关键词)、目标关键词(从核心关键词以及它的衍生的关键词中提取出来的“主打”关键词。

2）查准率

查准率是指被检出的相关文献占被检出文献的百分比，是用来描述系统拒绝不相关文献的能力或者检索精确度的一种尺度。它应用于很多领域，比如在

本题中，它就是在观点挖掘领域的应用，用它来衡量分类器识别出正面观点的性能，它的计算公式：查准率=识别出的真正的正面观点数/所有的识别为正面的观点的条数。

3）查全率

查全率是指被检出的相关文献占总文献内的相关文献的百分比。是用来描述系统检出文献能力的一种尺度。在很多方面都会应用到查全率这一指标，比如本文中正是它在观点挖掘方面的应用，可以用来衡量文本分类器的的性能。计算方式为：查全率=识别出的真正的正面观点数/样本中所有的真正正面观点的条数。

4）吻合度分析

吻合度分析指用来评价用数学模型去分析其某一现象及其规律的这种模型优劣的一种标准。在本篇文章中将它用来评价题中所用来评价解决分类问题、热点问题等建立的数学模型。

5）热度

热度也就是指数，它反应的时每天搜索量、浏览量或者点赞量等等的多少。

6）相关性分析

相关性分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。相关性的元素之间需要存在一定的联系或者概率才可以进行相关性分析。根据附件四，这里研究的是留言详情和答复意见的相关性分析。

7）完整性

完整性是指确保系统中所包含的信息均完整真实和可靠。在本题是指相关部门对留言的所有答复，应保持一致、完整、可靠和准确。

8）可解释性

可解释性通常情况下是指在我们需要了解或者是解决一件事情的时候，可以获得所需要的足够的可以理解的信息。在本题中意指在了解和评价答复意见时，可以获得相关部门给出的所有可以理解的答复信息。

9）及时性（timeliness）

在日常或工作中，我们提出的问题对于我们本身来说，希望得到更高效率的解决的。回答是否及时是检验工作的效率高低的一个重要因素。在本篇文章中，及时性也就是用来判定相关部门的回答是否及时，是检验回答质量的一个因素指标，工作效率的体现。

指标名称	含义	符号表示
关键词	具有概括性，目标性和重要性的简短的语句	G
查准率	识别出的真正的正面观点数/所有被识别为正面的观点的条数	Z
查全率	识别出的真正的正面观点数/样本中所有的真正正面观点的条数	Q
吻合度分析	用来评价用数学模型去分析其某一现象及其规律的这种模型优劣的一种标准	W
热度	也就是指数，它反应的时每天搜索量、浏览量或者点赞量等等的多少。	R
相关性分析	两个或多个具备相关性的变量元素进行分析，	X

	从而衡量两个变量因素的相关密切程度	
完整性	确保系统中所包含的信息均完整真实和可靠。在本题是指相关部门对留言的所有答复，应保持一致、完整、可靠和准确	A
可解释性	我们需要了解或者是解决一件事情的时候，可以获得所需要的足够的可以理解的信息	K
及时性	检验工作的效率高低的一个重要因素	J

2.4 保留数据处理

由于本次数据文字比较多为了方便，我们需要对保留的数据进行处理首先需要提取关键词，同时将文字数据化否则没法利用工具来进行计算。

三、 模型假设

为了便于研究问题，将对题目中某些条件进行合理假设。

1）假设网络爬虫提取的数据是正确可靠的。

2）由于不了解群众留言的意图，假设每个群众的留言信息都是内心真实意见。

3）为了结果的准确性，假设题中和附件中所给出的所有信息，包括留言时间、留言主题、点赞数等等均真实可信。

4）为了能对答复意见的质量给出一套合理准确的评价方案，假设相关部门对每一条留言都给出了相应的答复意见。

四、 问题一

4.1 信息的初步处理

在附件二中我们可以看到部分信息是重复的，例如编号 303 和 319 用户是同一个人，并且反映的内容是同一个内容。为排除重复信息带来的影响，首先我们需要对信息进行初步处理，通过电子表格的操作可以将重复信息用其他颜色标注出来，进而将重复的内容去掉。

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
24	A00074011	A市西湖建筑集团占道施工有安全隐患	20/1/6 12:09:	围墙内。每天尤其上	城乡建设
37	U0008473	市在水一方大厦人为烂尾多年，安全隐患严	20/1/4 11:17:	着，不但占用人行道	城乡建设
83	A00063999	投诉A市A1区苑物业违规收停车费	9/12/30 17:06	决。不知程明物业如	城乡建设
303	U0007137	A1区蔡锷南路A2区华庭楼顶水箱长年不洗	19/12/6 14:40:	品，霉是一种强致癌物	城乡建设
319	U0007137	A1区A2区华庭自来水好大一股霉味	19/12/5 11:17:	品，霉是一种强致癌物	城乡建设
379	A00016773	投诉A市盛世耀凯小区物业无故停水	19/11/28 9:08:	物业不是为业主服务的	城乡建设
382	U0005806	咨询A市楼盘集中供暖一事	9/11/27 17:14	月亮岛片区近年规划	城乡建设
445	A00019209	3区桐梓坡西路可可小城长期停水得不到解决	9/11/19 22:39	求帮助至今没有找到	城乡建设
476	U0003167	反映C4市收取城市垃圾处理费不平等的问题	9/11/15 11:44	在的物业公司也未给	城乡建设
530	U0008488	A3区魏家坡小区脏乱差	9/11/10 18:59	人让人好好休息一下，	城乡建设
532	U0008488	A市魏家坡小区脏乱差	9/11/10 12:30	人让人好好休息一下，	城乡建设
673	A00080647	十村小区第四届非法业委会涉嫌侵占小区业主	9/10/24 11:29	法业委会主任不敢出	城乡建设
994	U0005196	A3区梅溪湖壹号御湾业主用水难	19/9/18 22:43:	别的城市都已经一门	城乡建设
1005	U0006509	A4区鸿涛翡翠湾强行对入住的业主关水限电	19/9/18 13:36:	房地产公司和金晖物业	城乡建设
1110	A00099772	施工导致A市锦楚国际星城小区三期一个月价	19/9/9 11:07:	是无通知，突然断电	城乡建设
1309	U0005083	A6区润和紫郡用电的问题能不能解决	19/8/21 15:12:	起之后，我们的用电	城乡建设
1440	A0003288	A市锦楚国际新城从6月份开始停电好多次了	19/8/6 10:28:	的生活，而且我们小	城乡建设
1775	U0002150	给A9市城区南西片区城铁站设立的建议	19/7/4 18:52:	达A市，并且规划有地	城乡建设
1783	U0004763	请A6区政府加大对滨水新城的绿化建设	19/7/4 14:25:	的或者几个半大小树	城乡建设
1827	U000613	A5区楚府线几个小区经常停电	19/7/1 20:14:	已停电三次。说是	城乡建设
2603	A00099650	省建望集团及西地省辉东安建工程有限公司	19/4/20 16:50:	已（2015年～2016年	城乡建设
3607	A00046529	A2区山水嘉园1栋三单元群租房扰民	19/1/8 10:08:	隐患，投诉给物业公	城乡建设
3742	A00013884	区杜鹃文苑小区外的非法汽车检测站要开业	18/12/26 10:13	（省级）；水务局：	城乡建设
3800	U0001518	C市联合修建中速磁悬浮（最高时速150km	18/12/20 1:23:	与京广高铁B市西站，	城乡建设
3874	U0007328	5区嘉华路旁露天垃圾池子臭味熏天污水横流	18/12/11 21:40	无专人值守和管理	城乡建设
3980	U0001518	市地铁2号线西延二期暂缓修建，改建中速磁	18/12/1 1:05:	至I市的中速磁悬浮	城乡建设
3981	U0001518	A市城铁站至A市火车南站、A市黄花国际机	18/12/1 0:14:	A市南站、A市东站(	城乡建设
4042	U0007328	路两旁人行道脏乱差，露天垃圾池子臭味重	18/11/25 12:23	区（大桥二区安置点	城乡建设

如上图处理后，我们需要将无关的数据去除，形成新的电子表格，而我们本次分析所需要的信息保留。新的表格命名为表格一，文件形式如下：

A	B	C
留言主题	留言详情	一级标签
A市西湖建筑集团占道施工有安全隐患	围墙内。每天尤其上	城乡建设
市在水一方大厦人为烂尾多年，安全隐患严	着，不但占用人行道	城乡建设
投诉A市A1区苑物业违规收停车费	决。不知程明物业如	城乡建设
A1区蔡锷南路A2区华庭楼顶水箱长年不洗	品，霉是一种强致癌物	城乡建设
A1区A2区华庭自来水好大一股霉味	品，霉是一种强致癌物	城乡建设
投诉A市盛世耀凯小区物业无故停水	物业不是为业主服务的	城乡建设
咨询A市楼盘集中供暖一事	月亮岛片区近年规划	城乡建设
3区桐梓坡西路可可小城长期停水得不到解决	求帮助至今没有找到	城乡建设
反映C4市收取城市垃圾处理费不平等的问题	在的物业公司也未给	城乡建设
A3区魏家坡小区脏乱差	认让人好好休息一下，	城乡建设
A市魏家坡小区脏乱差	认让人好好休息一下，	城乡建设
一村小区第四届非法业委会涉嫌侵占小区业主	法业委会主任不敢出	城乡建设
A3区梅溪湖壹号御湾业主用水难	别的城市都已经一门	城乡建设
A4区鸿涛翡翠湾强行对入住的业主关水限电	房地产公司和金晖物业	城乡建设
施工导致A市锦楚国际星城小区三期一个月停	是无通知，突然断电	城乡建设
A6区润和紫郡用电的问题能不能解决	起之后，我们的用电量	城乡建设
A市锦楚国际新城从6月份开始停电好多次了	的生活，而且我们小	城乡建设
给A9市城区南西片区城铁站设立的建议	达A市，并且规划有七	城乡建设
请A6区政府加大对滨水新城的绿化建设	的或者几个半大小树	城乡建设
A5区楚府线几个小区经常停电	，已停电三次。说是	城乡建设
省建望集团及西地省辉东安建工程有限公司	吧（2015年～2016年	城乡建设
A2区山水嘉园1栋三单元群租房扰民	隐患，投诉给物业公	城乡建设
区杜鹃文苑小区外的非法汽车检测站要开业	（省级）；水务局：	城乡建设
C市联合修建中速磁悬浮（最高时速150km/	与京广高铁B市西站，	城乡建设
5区嘉华路旁露天垃圾池子臭味熏天污水横流	无专人值守和管理	城乡建设
市地铁2号线西延二期暂缓修建，改建中速磁	至I市的中速磁悬浮	城乡建设
A市城铁站至A市火车南站、A市黄花国际机	A市南站、A市东站(	城乡建设
冬两旁人行道脏乱差，露天垃圾池子臭味熏天	（大桥二区安置房	城乡建设

4.2 基于网络爬虫的关键词提取

网络爬虫别名网络机器人网络蜘蛛更经常被称为网页追逐者，是一种按照一定规则自动的抓取信息的程序或脚本。

在生活中我们想要了解事物除了询问别人更主观的方法则是上网查询，而百度、搜狐等浏览器是我们常常用到的，当我们用这些浏览器进行查询时，细心的小伙伴们会发现。这些浏览器就像是我们肚子里的蛔虫一样有时候我们只打了一两个字相关的信息就会出现。甚至我们查询一些信息时只需要打出一两个字就可以查询，这就是利用关键词进行查询的方法。

首先，根据附件一中的三级标签体系，通过网络给出每个三级标签中标签对应的相连词，所有这些词对应着成为一级标签中的关键词。关键词的提取是文本挖掘领域的一个分支，利用算法来提取关键词的方法大致分为两类：分别是有无监督关键词提取

无监督顾名思义不需要人工进行标注，而是利用其他的某种方法来进行发现关键词，进而进行关键词的提取。这种方法主要是先选出几个候选词，然后通过候选池的比较评分，从中选出分数较高的几个候选词作为关键词。具体的算法有 LDA、TF-IDF、TextRank 等。

有监督的关键词提取方法，把关键词提取看作为二分类问题，同样的是先提取出几个候选词，每个候选词对应着标签，是否为关键词。在这个基础上我们就将一些词语作为储存，当一篇文档进入数据库时通过利用已经分好标签的词语对比，就可以选出作为关键词的候选词。

1、将给定的文本按照整句分割，

即

$$T=[A_1，A_2,\cdots，A_m]$$

2、将划分后的句子进行词性标注处理并将句子分词。保留指定词性的单词去掉介词等无关紧要的词语，主要保留名词、动词、形容词等

在这里是a_{ij}经过筛选后的候选关键词

$$A_i=[a_{i1},a_{i2},\cdots,a_{im}]$$

3、构造关键词 G＝（P，R），P 根据第二步中得到的候选词组成的节点集合。将两个词语看做两个点，根据两个点之间的边来判断。

4、根据公式迭代节点的权重直到该权重收敛。

5、对节点的权重进行排序从而得到最重要的几个词作为候选关键词。

6、根据上一步得到的后选关键词在原来的文本中进行记录标记，若有相连的关键词组成词组，作为多词关键词。

一级标签对应的关键词有：

一级标签分类	关键词
城乡建设	交通 医疗 安全生产管理
党务政务	公平公正 贪污 团结 和谐
国土资源	天气预报 阴晴圆缺 震源深度 房屋倒塌
环境保护	臭氧层 温室效应 二氧化碳
纪检监察	奢靡之风 攀比 打击报复 治安
交通运输	拐卖 收费站 高速路
经济管理	安全隐患 网路贷款 保险
科技与信息产业	通信保障 联通 移动 欠费 超额
民政	慈善 安置 救灾募捐 优抚 救助 社会服务
农村农业	草原 扶贫 惠农 防疫 林业 农产品 农耕
商贸旅游	旅游管理 商业贸易 市场监管 质检
卫生计生	卫生 计生 食药品监管 医患 医政
政法	法律服务 行政复议 户籍证件 交通警务
教育文体	教师待遇 教育行政 教育体制 考试招生
劳动与社会保障	劳动 基金 工资 保险

4.3 吻合度对比

通过 4.2 的方法可以分别得到关于标签和留言内容的关键词，通过对比关键词之间的相关度，就可以得到标签与留言内容间的对应。

利用相关度：

当 X，Y 关联大时，MI（x，y）大于 0；当关系弱时，等于 0，小于 0 时，X 和 Y 称为互补关系。

$$MI（X,Y）=\log 2p(x,y)/p（x）p(y)$$

假设所有留言内容的集合为{C}，共有 N 篇文章，含有某个关联词 x 的文章总数是 nx，有关键词 y 的文章总数是 ny， 相关性的计算公式为：

$$Corr(X,Y)=Math.\log 10(N/Nx)*Math.\log 10(N/Ny)*Nxy/(Nx+Ny-Nxy)$$

五、 任务二

5.1 热点问题挖掘

LDA 主题模型

LDA(Latent Dirichlet Allocation)中文翻译为：潜在狄利克雷分布。LDA 主题模型是一种文档生成模型，是一种非监督机器学习技术。它认为一篇文档是有多个主题的，而每个主题又对应着不同的词。一篇文档的构造过程，首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词，这样就生成了这篇文档的第一个词。不断重复这个过程，就生成了整篇文章（当然这里假定词与词之间是没有顺序的，即所有词无序的堆放在一个大袋子中，称之为词袋，这种方式可以使算法相对简化一些）。

LDA 模型（LDATopic Model）能够识别在附件三里面的留言主题，并且挖掘语料里隐藏信息，并且在留言主题聚合、从非结构化文本中提取出重要信息、特征选择内容，这对我们进行附件三热点挖掘问题有很大的作用。

a. LDA 模型优点：

b. ①算法快速，简单、是解决聚类问题的一种重要的算法

c. ②对处理大数据集，该算法保持可伸缩性和高效性③有利于处理大量的数据，是一种很好的主体建模算法。

d. 缺点：

e. ①A 模型没有考虑所要提取的关键词在文档中的位置②对于较长的文档，不匹配的主题也比较困难。

5.2 热点挖掘

我们先是利用 excel 的排序功能，将附件三按照时间顺序给排列出来，然后，我们可以自定义一个坐标轴，以 时间作为坐标轴的横坐标，以点赞数和反对数作为纵坐标。利用 MATLAB 的画图功能，我们就可以得到一个关于时间-点赞、反对的一个坐标轴，然后，我们自定义一个关于热点与总的数量和点赞数之间的一个函数关系 Y=D/S （D 表示点赞数，S 表示点赞数与反对数的总和），我们对 Y 的值进行一个分类，当 $50\% < Y < 100\%$时，我们就认为这些范围内的话题是热度的 $30\% < Y < 50\%$和 $0\% < Y < 30\%$不具有热度的。

导入 - C:\Users\lqdn\Desktop\附件三.xlsx

导入 视图

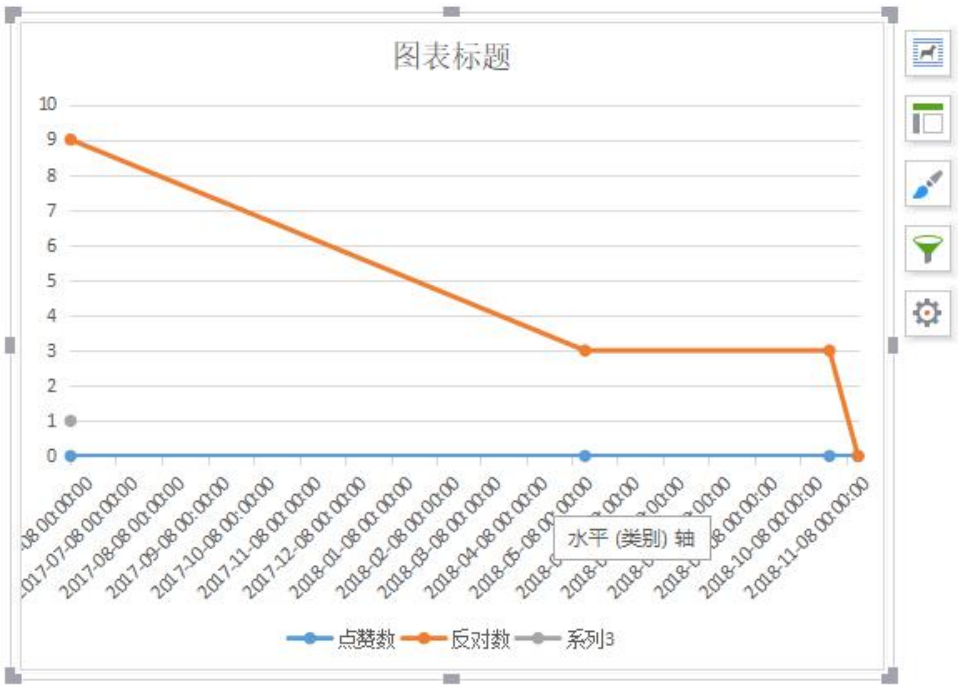
单个 | 左侧/右侧 | 顶端/底端 | 自定义 | 选项卡位置 | 按窗口大小收缩选项卡 | 按字母顺序排序 | 平铺 | 文档选项卡

Untitled

	A	B	C	D	E	F	G
	VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7
	数值	元胞	元胞	日期时间	元胞	数值	数值
1	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
2	188006	A0001029...	A3区一米...	2019/2/28...	座落在A市...	0	0
3	188007	A00074795	咨询A6区...	2019/2/14...	A市A6区道...	0	1
4	188031	A00040066	反映A7县...	2019/7/19...	本人系春华...	0	1
5	188039	A00081379	A2区黄兴...	2019/8/19...	靠近黄兴路...	0	1
6	188059	A00028571	A市A3区中...	2019/11/2...	A市A3区中...	0	0
7	188073	A909164	A3区麓泉...	2019/3/11...	作为麓泉社...	0	0
8	188074	A909092	A2区富绿...	2019/1/31...	`二高一部...	0	0
9	188119	A00035029	对A市	A市6路公交车随意变道通行 已转换为[类型: 元胞, 值: A市6路公交车随意变道通行]			
10	188170	A88011323	A市6路公...	2019/12/2...	12月21日...	0	0
11	188249	A00084085	A3区保利...	2019/9/17...	保利麓谷林...	0	0
12	188251	A00013092	A7县特立...	2019/10/1...	近来，下午...	0	0
13	188260	A00053484	A3区青青...	2019/5/31...	还我宁静我...	0	0
14	188396	A00047580	关于拆除聚...	2019/4/15...	桐梓坡589...	2	1
15	188399	A00097934	A市利保壹...	2019/7/3 ...	您好，我想...	0	0
16	188409	A0003274	A市地铁3...	2019/6/19...	尊敬的领导...	0	4
17	188414	A00096844	A4区北辰...	2019/8/1 ...	您好！我是...	0	0
18	188416	A00029753	请给K3县...	2019/06/2...	K3县的乡...	0	0
19	188451	A00013004	A7县春华...	2019/4/11...	我是春华镇...	0	2
20	188455	A00035902	咨询异地办...	2019/5/16...	书记您好！ ...	0	0

Sheet1

附件三.xlsx



5.3 评价结果

六、 任务三

6.1 相关性分析

相关性分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。相关性的元素之间需要存在一定的联系或者概率才可以进行相关性分析。在这里研究的是留言详情和答复意见的相关性分析。

对相关性进行分析我们考虑转化为：判断问题与答复的关联程度，当问题答复的关联程度高则代表答复意见的相关性比较高。在这里我们利用关键词来进行判断，而两个词语之间的关联程度也就是相似程度，根据互信息的观点有：

由于在信息论与概率论中，判断两个变量之间依赖性是由两个随机变量的互信息或者转移信息（transinformation）来判断的。

与相关系数不同，互信息不仅仅用于实值变量，互信息更加的一般，并且他还决定着联合分布与被分解的边缘分布的乘积的相关联系程度。可以表示两个事件集合之间相关性。

由于互信息的一般性，所以它可以用于度量一些与语言相关的问题，在信息论当中还可以用户信息来衡量两个词语之间的关联程度也可以用来计算词语与类别之间的关联程度。

计算公式：

$$
\begin{aligned}
&I（X，\ Y）=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\\
&=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)}-\sum_{x,y} p(x,y)\log\ p(x)\\
&=\sum_{x,y} p(x)p(y\mid x)\log\ p(y\mid x)-\sum_{x,y} p(x,y)\log\ p(x)\\
&=\sum_{x} p(x)[\,p(y\mid x)\log\ p(y\mid x)]-\sum_{y}\log\ p(y)[\sum_{x} p(x,y)]\\
&=-\sum_{x} p(x)H(Y\mid X=x)-\sum_{y}\log\ p(x)p(y)\\
&=-H(Y\mid X)+H(Y)
\end{aligned}
$$

6.2 完整性分析

完整性是指确保系统中所包含的信息均完整真实和可靠。在本题是指相关部门对留言的所有答复，应保持一致、完整、可靠和准确。

根据附件四的留言主题，我们需要对答复意见进行检验，是否符合留言主题，这里需要一一检验.

6.3 可解释性

可解释性通常情况下是指在我们需要了解或者是解决一件事情的时候，可以获得所需要的足够的可以理解的信息。在本题中意指在了解和评价答复意见时，可以获得相关部门给出的所有可以理解的答复信息。

根据附件四的信息，我们将用卡方检验对答复意见进行检验。从卡方检验中得到它的相关性，通过相关性来判断它的可解释性，从而判断问题是否得到解决。

1、卡方检验

卡方检验是数理统计中用来检验两个变量独立性的方法，它是一种用来确定两个分量之间是否存在相关性的统计方法，一般经典的卡方检验是用来检验定性自变量对定性应变量的相关性的。

2、基本思路

（1）原假设：两个变量是独立的。

（2）计算出我们实际观察值与理论值之间的偏离程度。

（3）如果计算得到的偏差足够小，小于之前设定的阈值，则就接受原假设；否则将否定原假设，则可认为两个变量是相关的。

3、计算公式

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

A 为实际值，T 为理论值

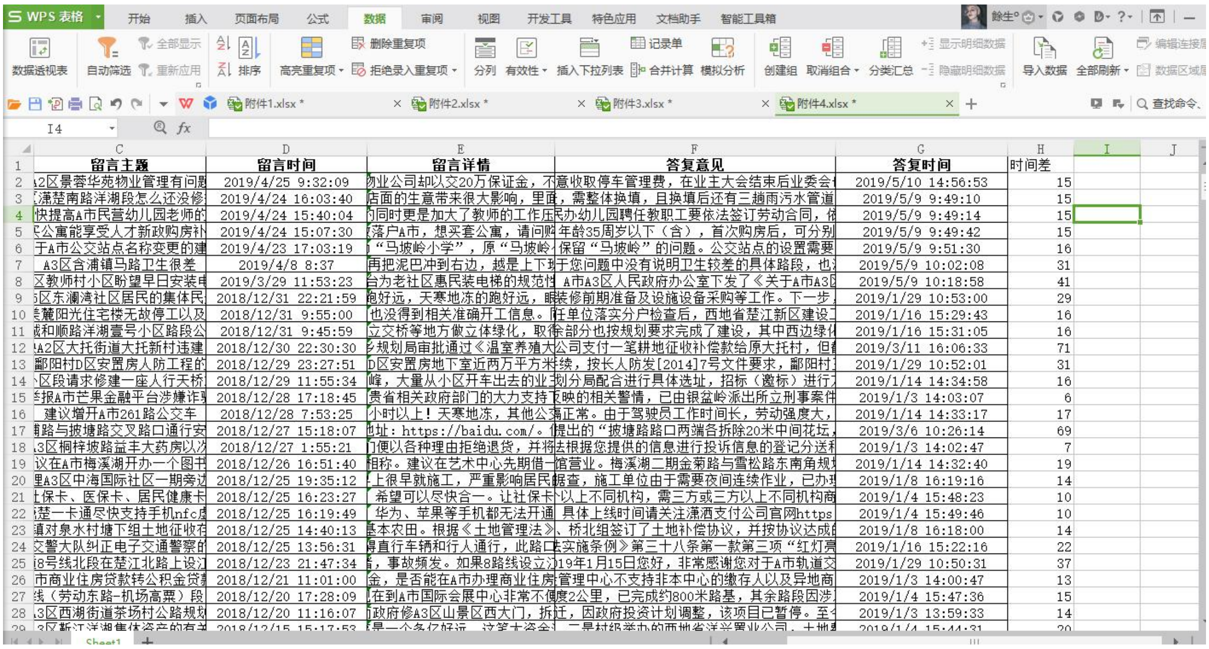
6.4 及时性

在答复的质量指标中，及时性是指有关部门对问题给出答复的及时程度，检验及时性的方法也很简单。在附件四中利用计算公式计算答复时间与问题时间相差的天数，所得结果越小这证明越及时。

计算公式：

D＝d2-d1。

通过计算，所得表格形式如：



	留言主题	留言时间	留言详情	答复意见	答复时间	时间差
2	A2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	15
3	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	，需整体换填，且换填后还有三趟雨污水管道	2019/5/9 9:49:10	15
4	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	15
5	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	15
6	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	16
7	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	31
8	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	41
9	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	29
10	麓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。	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	16
11	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	16
12	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，但着	2019/3/11 16:06:33	71
13	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	31
14	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	16
15	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下	反映的相关警情，已由银盆岭派出所立刑事案件	2019/1/3 14:03:07	6
16	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	正常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	17
17	路与披塘路交叉路口通行安	2018/12/27 15:18:07	地址：https://baidu.com/。	提出的“披塘路路口两端各拆除20米中间花坛，	2019/3/6 10:26:14	69
18	3区桐梓坡路益丰大药房以次	2018/12/27 1:55:21	门便以各种理由拒绝退货，并将	去根据您提供的信息进行投诉信息的登记分送私	2019/1/3 14:02:47	7
19	议在A市梅溪湖开办一个图书	2018/12/26 16:51:40	相称。建议在艺术中心先期借一	馆营业。梅溪湖二期金菊路与雪松路东南角规划	2019/1/14 14:32:40	19
20	理A3区中海国际社区一期旁边	2018/12/25 19:35:12	上很早就施工，严重影响居民	据查，施工单位由于需要夜间连续作业，已办理	2019/1/8 16:19:16	14
21	社保卡、医保卡、居民健康卡	2018/12/25 16:23:27	希望可以尽快合一。让社保卡	以上不同机构，需三方或三方以上不同机构商	2019/1/4 15:48:23	10
22	楚一卡通尽快支持手机nfc虚	2018/12/25 16:19:49	华为、苹果等手机都无法开通	具体上线时间请关注潇洒支付公司官网https	2019/1/4 15:49:46	10
23	镇对泉水村塘下组土地征收存	2018/12/25 14:40:13	基本农田。根据《土地管理法》	、桥北组签订了土地补偿协议，并按协议达成的	2019/1/8 16:18:00	14
24	交警大队纠正电子交通警察的	2018/12/25 13:56:31	得直行车辆和行人通行，此路口	法实施条例》第三十八条第一款第三项“红灯亮	2019/1/16 15:22:16	22
25	8号线北段在楚江北路上设立	2018/12/23 21:47:34	路，事故频发。如果8路线设立	2019年1月15日您好，非常感谢您对于A市轨道交	2019/1/29 10:50:31	37
26	市商业住房贷款转公积金贷款	2018/12/21 11:01:00	金，是否能在A市办理商业住房	管理中心不支持非本中心的缴存人以及异地商	2019/1/3 14:00:47	13
27	线（劳动东路-机场高票）段	2018/12/20 17:28:09	在到A市国际会展中心非常不便	度2公里，已完成约800米路基，其余路段因涉	2019/1/4 15:47:36	15
28	3区西湖街道茶场村公路规划	2018/12/20 11:16:07	政府修A3区山景区西大门，拆	迁，因政府投资计划调整，该项目已暂停。至今	2019/1/3 13:59:33	14

6.5 答复评价

通过对以上四个评价指标的分析，有关部门的答复的质量与相关性、完整性、可解释性，及时性都呈正相关。

七、 参考文献

[1] 张厚粲，徐建平. 现代心理与教育统计学.北京师范大学.1988.

[2] 盛骤，谢式千，潘承毅.概率论与数理统计.高等教育出版社

[3] NLP 之关键词提取,CSDN

[4] 于成龙，于洪波.网络爬虫技术的研究.牡丹江大学.2011.

[5] 王光宏，蒋平，数据挖掘综述[J]同济大学学报(自然科学版),2004,32(2):246-252

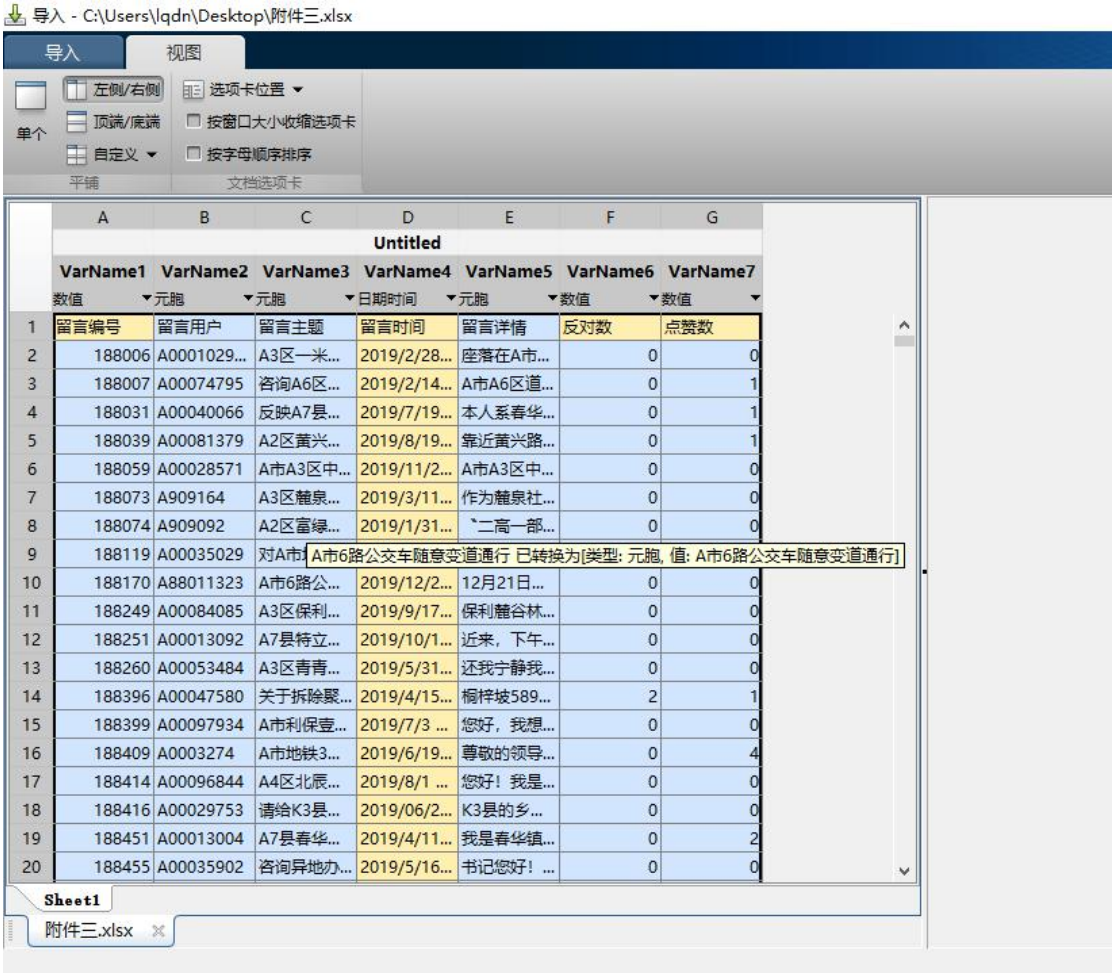
[6] 陈文伟等.数据挖掘技术[M].北京:北京工业大学出版社

[7] 毛国君等.数据挖掘原理与算法[M].北京:清华大学出版社,2005.

[8] 字等编著.数据仓库原理与实践.北京：人民邮电出版社，2003

[9] textRank 杂谈 - Together_CZ 的博客 - CSDN 博客，网页

八、 附录




```
from jieba import analyse

tfidf = analyse.extract_tags

analyse.set_stop_words('stopword.txt')

text =

keywords = tfidf(text, topK=10, withWeight=False, allowPOS=())

print('结果为：')

print([keyword for keyword in keywords])
```