“智慧政务”中的文本挖掘应用

摘要：近年来，随着网络的发展，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。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。本文针对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。

针对问题一：本文首先根据附件一的分类三级标签体系，在利用 excel 将附件 2 的数据进行数据预处理，运用 python 进行分词，并按照附件一级标签进行分类。再利用 F-Score 对分类方法进行评价：

针对问题二：在附件 3 的数据中，发现了很多重复的数据。根据数据预处理，在原始的数据上进行去空格、去重处理，在此基础上利用 python 进行中文分词；利用 TF-IDF 算法,找出每个问题描述的关键词，把问题描述信息转换为权重向量。采用 K-means 算法对问题进行分类,利用 Knn 算法找出与各中心相似的元素，根据个数多的判定所属类别；统计相关数据，分类筛选汇总，预测热点问题。

针对问题三： 运用 TF-IDF 算法,找出每个问题描述的关键词，把问题描述信息转换为权重向量。再利用 excel 绘制线性相关图表，完整性图表，以及可解释性图表，即可直观的看出答复意见的质量。

关键词：数据预处理；python；TF-IDF；excel；

Abstract: in recent years, with the development of the network, WeChat, weibo, mayor’s mailbox, sunshine hotline and other network politicalplatform gradually become an important channel for the government to understand public opinion, gather people’s wisdom, and condense

people's spirit. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the

establishment of intelligent government affairs system based on natural language processing technology has become a new trend of socialgovernance innovation and development, which plays a great role in promoting the management level and efficiency of the government.

According to the attachment, this paper gives the record of people's political message collected from the open source of the Internet, and thecomments of relevant departments on the comments of some people. Use natural language processing and text mining to solve the following problem

问题重述

问题一：群众留言分类问题，工作人员在处理问政平台群众留言时，按照一定的划分体系进行分类，便于相应部门的处理。请根据附件一将群众留言进行分类。

问题二：热点问题挖掘，某一时间段内群众集中反映的问题称为“热点问题”。热点问题挖掘，有助于相关部门进行针对性处理，提高服务效率。。请根据附件 3 将某一时段内反映特定地点或特定 人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出 排名前 5 的热点问题，并保存为文件“热点问题表.xls”。

问题三：针对附 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。

问题分析

问题一：本文首先根据附件一的分类三级标签体系，在利用 excel 将附件 2 的数据进行数据预处理：去重、去空格。运用 python 进行分词，提取关键词，并

按照附件一级标签进行分类。再利用 F-Score 对分类方法进行评价：

问题二：热点问题挖掘，某一时间段内群众集中反映的问题称为“热点问题”。热点问题挖掘，有助于相关部门进行针对性处理，提高服务效率。。请根据附件 3 将某一时段内反映特定地点或特定 人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。

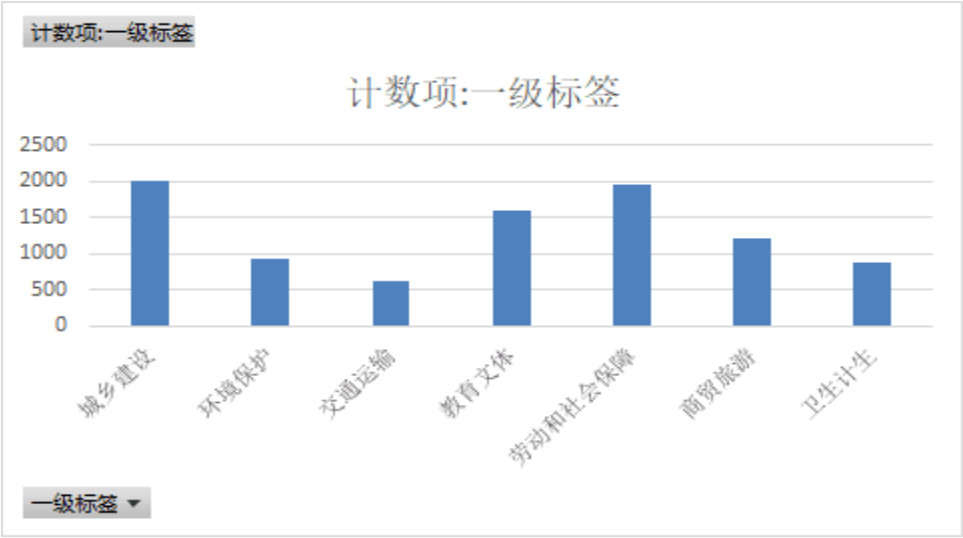
问题三：根据本题所给出的附件 4，首先将数据进行预处理。针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

总体流程与步骤

问题一：根据附件 2 的数据，本文首先根据附件一的分类三级标签体系利用 excel 建立模型，能够直观的知道各相关部门的数据，利用 excel 进行数据预处理：数据清洗、分词，添词典去停用词、绘制词云。

三级标签分类模型：

一级标签	计数项：一级标签
城乡建设	2009
环境保护	938
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



第二题：本题主要包括如下步骤:

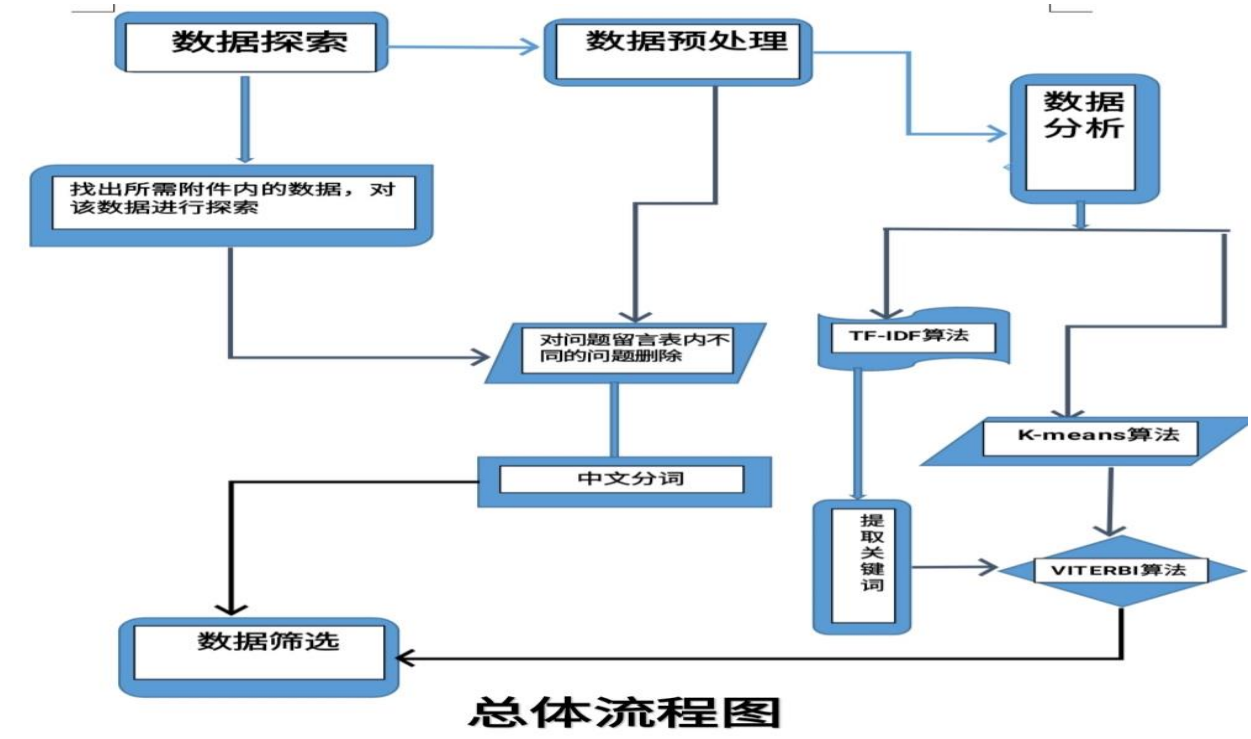
步骤一：数据探索，在题目给出的数据中，发现了很多重复的数据。

步骤二：数据预处理，在原始的数据上进行去重处理，在此基础上利用 python 进行中文分词。

步骤二：数据分析，在对留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用.这里采用 TF-IDF 算法,找出每个问题描述的关键词，把问题描述信息转换为权重向量。采用 K-means 算法对问题进行分类,利用 Knn 算法找出与各中心相似的元素，根据个数多的判定所属类别。

步骤三:数据筛选，统计相关数据，分类筛选汇总，预测热点问题。

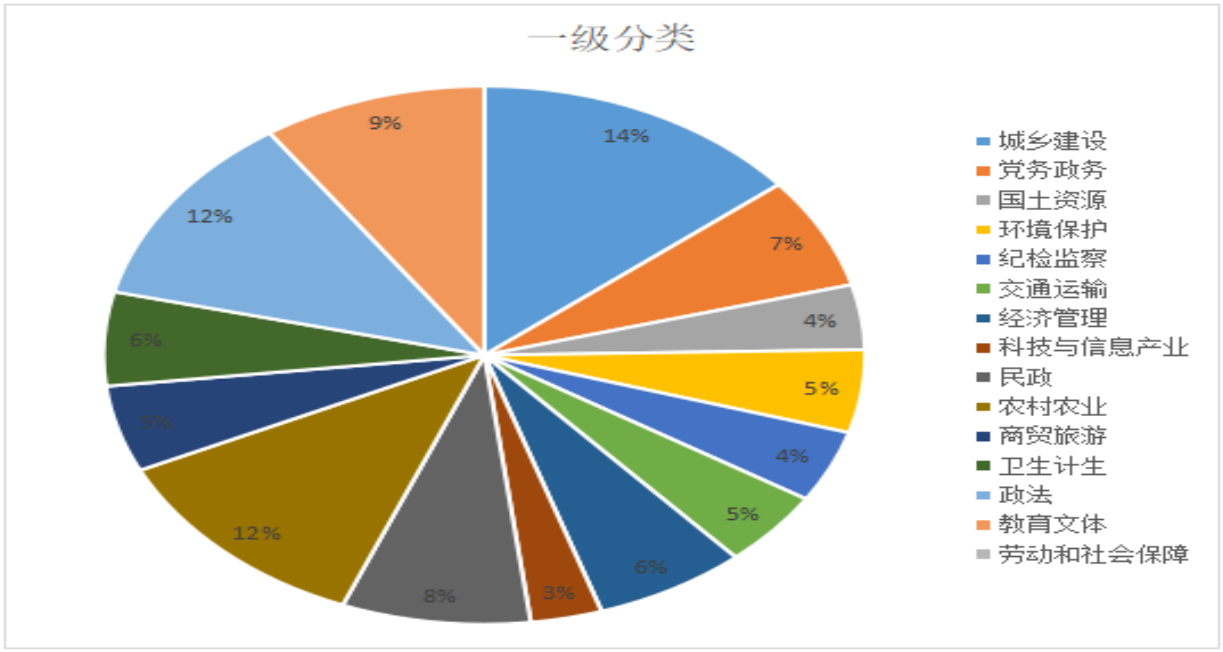
总体的流程图如下：



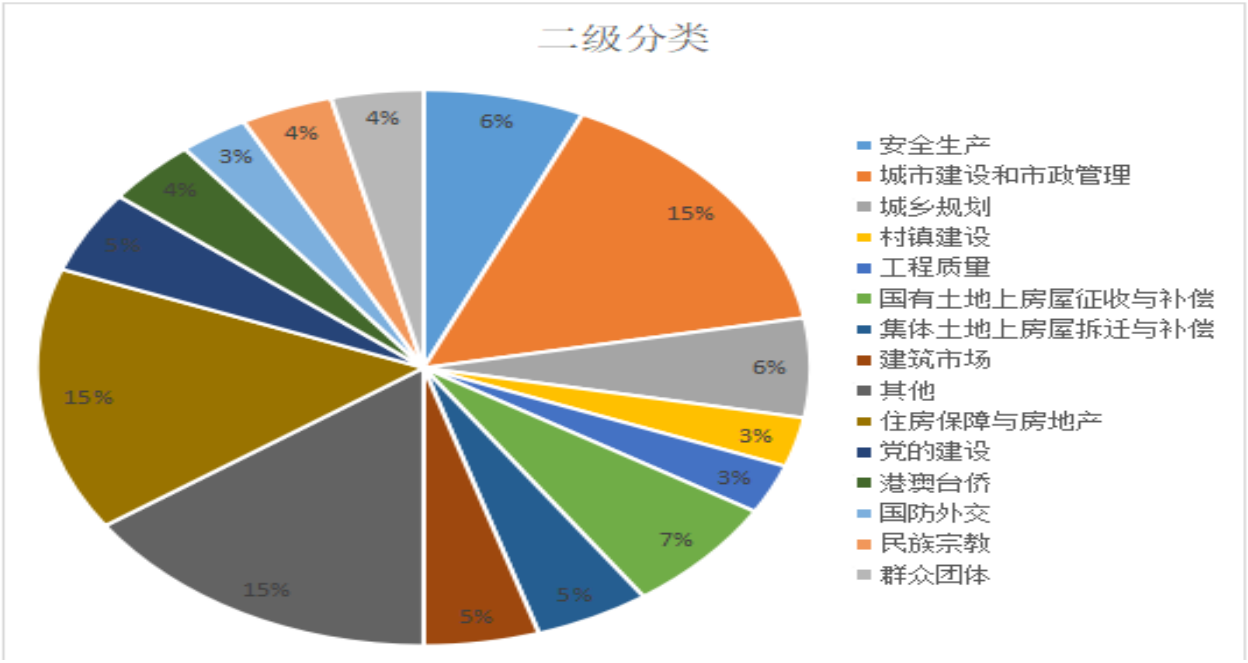
总体流程图

根据附件一的分类三级标签体系，利用附件 3 的数据建立模型，各级相关部门所占的比例如下图：

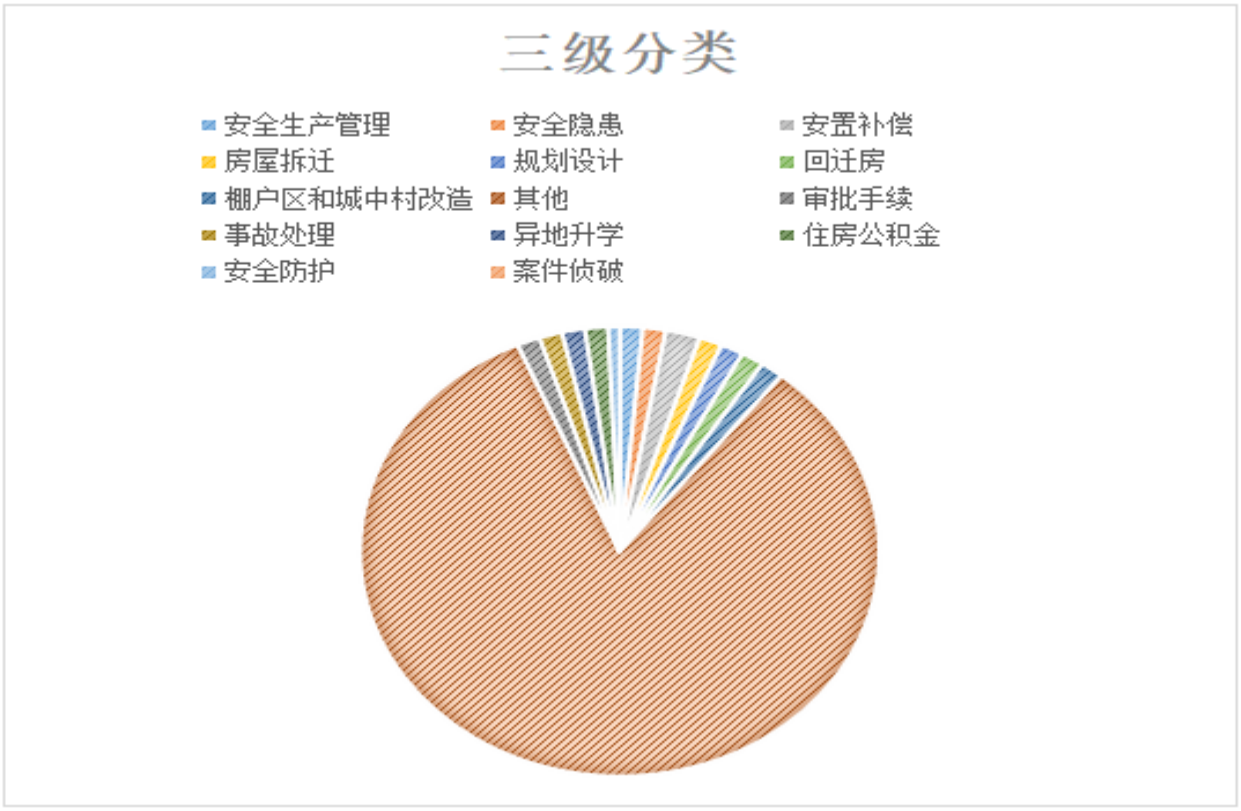
一级分类：


一级分类
14%
7%
4%
5%
4%
5%
6%
3%
8%
12%
5%
6%
12%
9%
城乡建设
党务政务
国土资源
环境保护
纪检监察
交通运输
经济管理
科技与信息产业
民政
农村农业
商贸旅游
卫生计生
政法
教育文体
劳动和社会保障

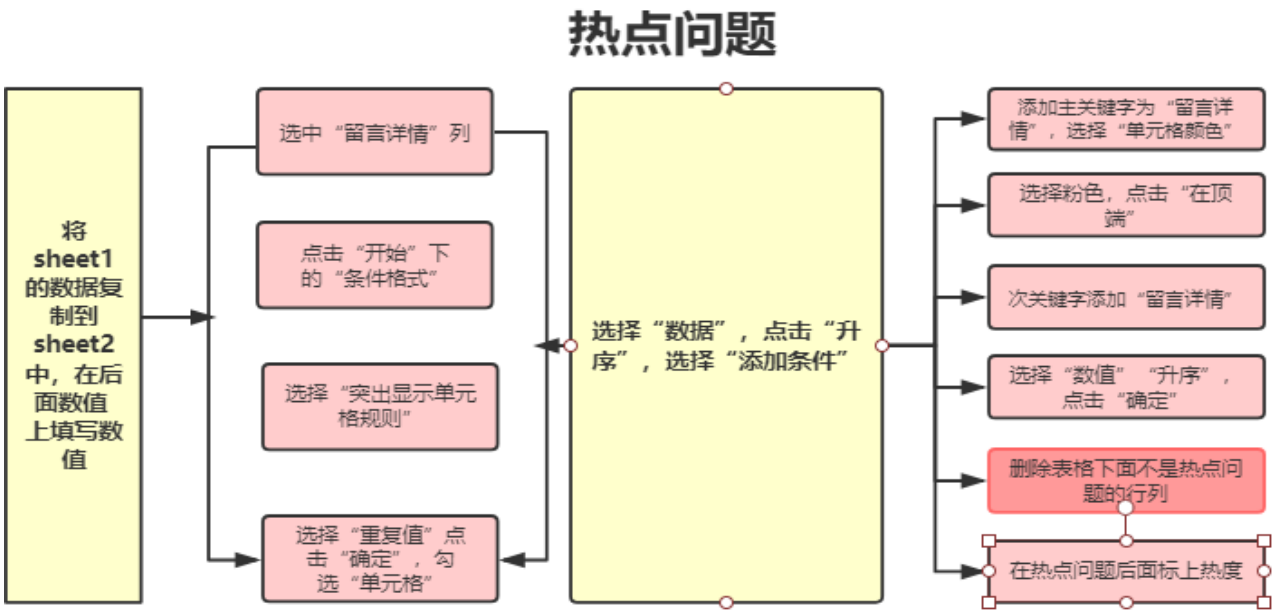
二级分类：


二级分类
6%
15%
6%
3%
3%
7%
5%
5%
15%
15%
5%
4%
3%
4%
4%
安全生产
城市建设和市政管理
城乡规划
村镇建设
工程质量
国有土地上房屋征收与补偿
集体土地上房屋拆迁与补偿
建筑市场
其他
住房保障与房地产
党的建设
港澳台侨
国防外交
民族宗教
群众团体

三级分类：



利用 wps 进行热点问题分析，其流程图如下：



主要步骤如下:

步骤一：先将 sheet2 的数据复制到 sheet1 中，在后面数值上填写数字。

步骤二：选中“留言详情”，点击“开始”下的“条件格式”。

步骤三：选择“突出显示单元格规则”，点击“重复值”，点击“确定”，勾选单元格。

步骤四：选择“数据”，点击“升序”，点击“添加条件”。

步骤五：添加主要关键字为“留言详情”，选择“单元格颜色”，选择“粉色”，点击“在顶端”。

步骤六：次关键字添加“留言详情”，“数值”，“升序”，点击“确定”。

步骤七：删除表格下面没有相同问题的几行。

	A	B	C	D	E	F	G	H
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	热度
2	239737	A000107463	A市能否设立南塘城轨公交站？	2019/10/31 21:19:59	南塘小学，A市一中城	0	0	3
3	264088	A000107463	A市能否设立南塘城轨公交站？	2019/10/31 21:17:22	南塘小学，A市一中城	0	1	
4	343985	A108051	A市能否设立南塘城轨公交站？	2019-10-31 21:19:59	南塘小学，A市一中城	0	0	
5	219572	A00045546	问A市什么时候能普及5G网络	2019/5/14 11:19:22	的城区建成。看消息	1	4	3
6	248712	A00045546	问A市什么时候能普及5G网络	2019/05/14 11:22:13	的城区建成。看消息	0	0	
7	316619	A235259	问A市什么时候能普及5G网络	2019-05-14 11:22:13	的城区建成。看消息	0	0	
8	188396	A00047580	龙楚在西地省商学院宿舍旁安装	2019/4/15 16:23:09	《中小学校校园环境	2	1	3
9	198874	A00047580	地省商学院宿舍、A3区一小旁安	2019/4/15 16:13:32	《中小学校校园环境	0	0	
10	273805	A00047580	地省商学院宿舍、A3区一小旁安	2019/4/15 16:11:19	《中小学校校园环境	0	1	
11	201447	A00020543	距我家仅3米，相关政府部门为	2019/7/22 17:04:08	全统计，整栋房屋开	0	2	5
12	226753	A00020543	7距我家仅3米，相关部门为何	2019/7/23 11:03:10	全统计，整栋房屋开	0	0	
13	229903	A00020543	距我家仅3米，相关政府部门为	2019/7/22 17:05:04	全统计，整栋房屋开	0	0	
14	273925	A00020543	距我家仅3米，相关政府部门为	2019/7/18 10:47:31	全统计，整栋房屋开	0	3	
15	278907	A00020543	距我家仅3米，相关政府部门为	2019/7/18 10:48:28	全统计，整栋房屋开	0	0	
16	190812	A00095451	市江山帝景新房有严重安全隐	2019/5/30 17:34:02	暴天气后过道全部是水	0	0	3
17	289893	A00095451	市江山帝景新房有严重安全隐	2019/5/30 17:20:53	暴天气后过道全部是水	0	0	
18	319659	A023956	市江山帝景新房有严重安全隐	2019-05-30 17:34:02	暴天气后过道全部是水	0	0	
19	227100	A00018469	市能不能提高医疗门诊报销范	2019/6/12 8:07:49	，小孩体弱多病各种	0	0	3
20	282746	A00018469	市能不能提高医疗门诊报销范	2019/06/12 08:23:01	，小孩体弱多病各种	0	1	
21	321736	A9992521	市能不能提高医疗门诊报销范	2019-06-12 08:23:01	，小孩体弱多病各种	1	0	
22	284147	A909113	力之城小区一楼的夜宵摊严重	2019/07/21 10:29:36	维护社会和谐稳定，各	0	3	4
23	360107	A0283523	力之城小区一楼的夜宵摊严重	2019-07-21 10:29:36	维护社会和谐稳定，各	3	0	
24	272122	A909113	小区一楼的夜宵摊严重污染附	2019/08/01 16:20:02	维护社会和谐稳定，各	0	6	
25	360108	A0283523	小区一楼的夜宵摊严重污染附	2019-08-01 16:20:02	维护社会和谐稳定，各	6	0	

表 1：选出 7 个群众集中反映的问题，其中两个因为是同一人反映，所以视为不是热点问题

	A	B	C	D	E	F	G	H
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	热度
2	284147	A909113	路魅力之城小区一楼的夜宵摊严重污染	2019/07/21 10:29:36	得要维护社会和谐稳定，合法维	0	3	4
3	360107	A0283523	路魅力之城小区一楼的夜宵摊严重污染	2019-07-21 10:29:36	得要维护社会和谐稳定，合法维	3	0	
4	272122	A909113	之城小区一楼的夜宵摊严重污染附近的	2019/08/01 16:20:02	得要维护社会和谐稳定，合法维	0	6	
5	360108	A0283523	之城小区一楼的夜宵摊严重污染附近的	2019-08-01 16:20:02	得要维护社会和谐稳定，合法维	6	0	
6	239737	A000107463	A市能否设立南塘城轨公交站？	2019/10/31 21:19:59	区，南塘小学，A市一中城南中	0	0	3
7	264088	A000107463	A市能否设立南塘城轨公交站？	2019/10/31 21:17:22	区，南塘小学，A市一中城南中	0	1	
8	343985	A108051	A市能否设立南塘城轨公交站？	2019-10-31 21:19:59	区，南塘小学，A市一中城南中	0	0	
9	219572	A00045546	请问A市什么时候能普及5G网络？	2019/5/14 11:19:22	覆盖的城区建成。看消息介绍，	1	4	3
10	248712	A00045546	请问A市什么时候能普及5G网络？	2019/05/14 11:22:13	覆盖的城区建成。看消息介绍，	0	0	
11	316619	A235259	请问A市什么时候能普及5G网络？	2019-05-14 11:22:13	覆盖的城区建成。看消息介绍，	0	0	
12	190812	A00095451	A市江山帝景新房有严重安全隐患	2019/5/30 17:34:02	雨雪天气后过道全部是水和雪	0	0	3
13	289893	A00095451	A市江山帝景新房有严重安全隐患	2019/5/30 17:20:53	雨雪天气后过道全部是水和雪	0	0	
14	319659	A023956	A市江山帝景新房有严重安全隐患	2019-05-30 17:34:02	雨雪天气后过道全部是水和雪	0	0	
15	227100	A00018469	A市能不能提高医疗门诊报销范畴	2019/6/12 8:07:49	不让，小孩体弱多病各种开支，	0	0	3
16	282746	A00018469	A市能不能提高医疗门诊报销范畴	2019/06/12 08:23:01	不让，小孩体弱多病各种开支，	0	1	
17	321736	A9992521	A市能不能提高医疗门诊报销范畴	2019-06-12 08:23:01	不让，小孩体弱多病各种开支，	1	0	

表 2：去掉同一个人反映的问题，剩下 5 个热点问题

3	343985	A108051	A市能否设立南塘城轨公交站？	2019-10-31 21:19:59	A2区南托街道东三线上南塘小区公交首末站就在城轨桥下，距暮云站和A1区南站(野生动物园)各约2.5公里，五百米范围内有A1区悦府，南台一号，南舍2号，幸福庄园，南方嘉城，格林香山，水电八局，暮云2业园等，还有新西地省科技学院新校区，南塘小学，A市一中城南中子等等，。如果在此加建一个城轨公交站，，第一可节省新建地铁巨额投资；二，可缓解只有单一公交的尴尬，公交车的准时性是不及轨道交通的。三，随着另外几五六百米范围内小区几年内入住，和学校学院使用，高峰期是公交车无法解决的。	0	0
4	316619	A235259	请问A市什么时候能普及5G网络？	2019-05-14 11:22:13	A市A2区之前宣布，5G站址布点1432个，全区5G网络基础建设基本完成。这意味着西地省首个5G基础网络全覆盖的城区建成。看消息介绍，5G下载速度比4G快40倍到60倍，希望5G网络能尽快覆盖A市全市，请问在这方面A市有什么计划和举措吗？	0	0
5	319659	A023956	A市江山帝景新房有严重安全隐患	2019-05-30 17:34:02	我是江山帝景a14栋的住户。看到新房，却怎么也高兴不起来，房子有严重安全隐患。每层入户过道东西两侧为敞开式栏杆防护，雨雪天气后过道全部是水和雪，而且水已经进入室内，进入电梯井，电梯将还存在严重的安全隐患。强烈要求开发商安装推拉式护窗，保证住户的人身安全。	0	0

表 3：部分热点问题留言明细

问题三： 本题针对留言问题，根据附件 4，从相关性、完整性、可解释性等角度检测相关部门给出的答复意见质量是否合格。

将附件 4 的数据进行数据预处理：去重，去空格。

相关性：将市民留言内容与相关部门的答复意见利用 python 进行关键词分词，在运用 TF-IDF 将文本转换为向量，再利用 excel 进行线性相关分析，构造相关模型，根据模型可以直观的知道答复意见与留言内容是否具有相关性。

完整性：将市民留言详情进行关键词，提出答复意见中的关键词并分词，利用 TF-IDF 将文本转换为向量，再利用 excel 进行拟合，建立模型。根据模型判断答复意见是否具有完整性。

可解释性：首先将数据预处理，去重，去除空格，将数据进行分词，提取关键词，构建词云图，根据词云图，直观的了解答复意见的有用性。

参考文献

[1] 赵琳瑛.基于隐马尔科夫模型的中文命名实体识别研究.西安电子科技大学.2007

[2]朱志远基于数据挖掘的网络招聘系统是设计与实现.电子科技大学硕士学位论文.2013

[3]曹卫峰中文分词关键技术研究.南京理工大学硕士学位论文.2009

[4]杨虎.面向海量短文文本去重技术的研究与实现国防科学技术大学.2007