“智慧政务”中的文本挖掘应用

摘要

网络问政平台中各类社情民意相关的文本数据量不断攀升，给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困难。

针对群众留言分类，进行三阶段的处理。一是对附件 2 的留言进行分词、去停用词处理，通过*TFIDF* 算法得到分词权重；二是建立基于朴素贝叶斯算法的分类器，对留言进行自动分类，得到关于留言内容的一级标签分类模型；三是用*F - Score*对分类方法进行评价。

针对热点问题的挖掘，先是对附件 3 的留言进行分词处理，生成分词列表，再建立词袋模型，基于文本集建立词典，提取词典特征数，将分词列表集转为稀疏向量集（语料库），进一步的用*TFIDF* 算法处理语料库，再对留言进行稀疏矩阵相似度算法计算，最后给予恰当的热度评价指标，提取排名前五的热点问题。

针对答复意见的评价，在进行完整性评价和可解释性评价时，把答复内容按照格式和解决方案分成两个文本，其中对保存格式的文本进行完整性评价，对保存解决方案的文本进行可解释性评价，同样对他们进行分词、去重复词处理、计算词频，通过进行词频统计，取词频最高的前 50 个词作为完整性和可解释性的关键词，计算关键词的权重，并对其进行匹配，最终对权重进行求和，从而得到完整性评价和可解释性评价。对于相关性评价，采取稀疏矩阵相似度的算法，对留言主题和答复内容进行相似度计算，从而得到对应的相关性评价。

关键词：分词；*TFIDF* 算法；朴素贝叶斯算法

1 背景与挖掘目标

1.1 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

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。我们需要利用自然语言处理和文本挖掘的方法解决下面的问题。

问题 1. 群众留言分类

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。通过 $F\text{-}Score$ 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。

问题 2. 热点问题挖掘

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”。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”。

问题 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。

2 数据获取

2.1 附件 1

附件 1 含有各类一级标签及所属的二三级标签体系。

2.2 附件 2

附件 2 含有若干留言信息，每条信息包括留言编号、留言用户、留言主题、留言时间、留言详情和一级分类等属性。对附件 2 进行排序分类处理，保留每个分类中的留言内容分词、去停用词后的词汇，对词汇进行词频统计，最后根据 *TFIDF* 得出词汇在各个类别中的权重。

2.3 附件 3

附件 3 抽取了附件 2 的部分内容，并在该内容的基础上增添了两个属性，分别为点赞数和反对数。对点赞数及反对数做简单的分析，点赞数越多说明赞同这条留言的人数更多，反对数越多说明不认同这条留言的人数多，做简单的加减运算，留言多的更有可能是热点问题。

2.4 附件 4

附件 4 含有若干留言信息，每条信息包括留言编号、留言用户、留言主题、

留言时间、答复意见和答复时间等属性。一共 2816 条数据，随机抽取 281 条，对数据进行人工分段。把答复内容的格式和解决方案分成两个文本，进行词频统计，得到完整性关键词和可解释性关键词。

3 问题分析

数据挖掘总流程如图 1 所示：

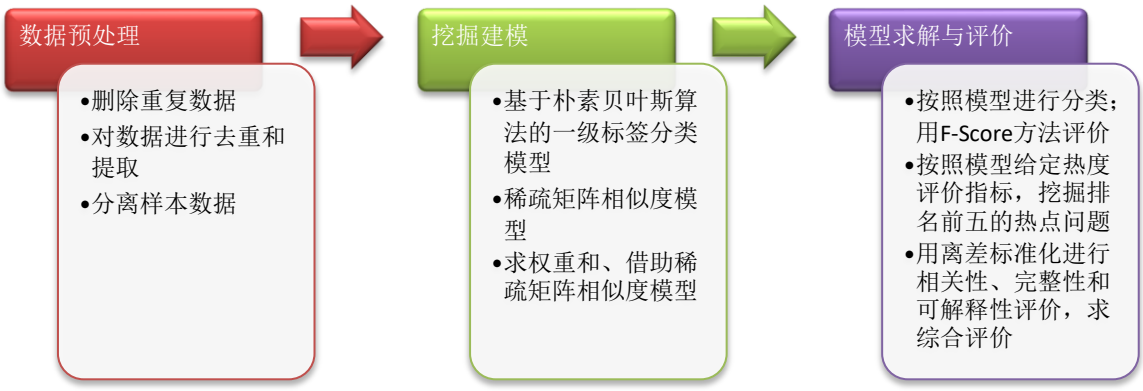


图 1 数据挖掘流程图

数据预处理：对附件 2 和附件 3 的数据进行删除重复项处理，将留言主题和留言详情合并，按分类进行保存。对附件 4 的答复意见进行分离样本数据。

挖掘建模：对附件 2 的数据进行分词和去停用词处理，再对分词进行词频计算，借助*TFIDF* 算法求分词在对应类别的权重，最终选出权重和最大的一类，以此来标记留言对应的类别标签。对附件 3 进行一级标签分类处理，借助稀疏矩阵相似度算法求得每一类留言内容之间的相似度，以便挖掘热点问题。对附件 4 的样本数据进行分词、去重，词频计算得到关键词，计算关键词的权重，对其进行求和，得到完整性评价和可解释性评价；对留言主题和答复内容进行分词、去停用词，借助稀疏矩阵相似度算法求它们之间的相似性。

模型求解：通过一级标签分类模型为每一条留言匹配标签类别。借助稀疏矩阵相似度模型挖掘出排名前五的热点问题。按照对关键词求权重和得到完整性和可解释性，借助稀疏矩阵相似度模型求得留言主题和答复内容的相似性。

模型评价：借助*F - Score*对一级标签分类模型进行评价。反复观察，给定合

理的热度评价指标对文本相似度模型进行评价。利用离差标准化进行相关性、完整性和可解释性评价，最终求三者的均值得到综合评价。

4 问题一的求解

4.1 问题解决思路

对数据进行去重和提取处理，再对数据进行分词，由于在众多数据中，有很多与建模无关的无用词，例如地名，日期还有一些常用符号，这些统称为无用词，故还需要去停用词，通过*TFIDF* 计算各类分词的权重，从而建立一级标签分类模型。求解模型之后，利用*F-Score*对模型进行评价。

流程如图 2 所示：

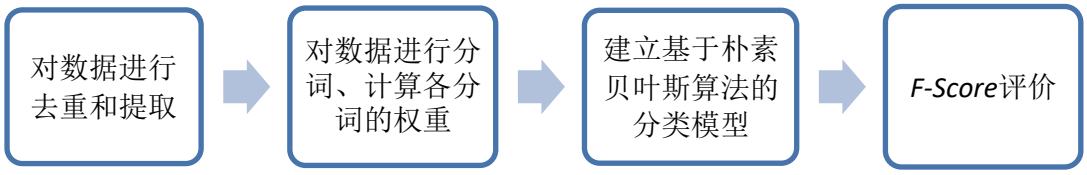


图 2 一级标签分类流程图

4.2 数据预处理

删除重复数据：分析附件 2 的数据，发现存在重复数据，对附件 2 的留言主题、一级标签相同和留言内容 、一级标签相同的数据进行去重，删除留言时间较早的数据；对留言主题或留言内容相同，一级标签不同的留言，进行人工分类，删除一级标签错误的数据。

提取留言详情：对附件 2 中一级标签属性排序，得到各类标签下的留言内容。对附件 2 的留言详情和留言主题进行合并，保存到第 F 列，处理后的附件 2 请浏览附件中的“训练集.xlsx”。提取“训练集.xlsx”中每一类标签下的留言详情和留言主题，即 F 列的内容，保存到对应类的文本文档中,总计有 7 个文本文档，将所有的文本文档放进留言内容文件夹中。

4.3 挖掘建模

对留言进行中文分词，计算各类留言内容的*TFIDF* 的值，作为权重。建立基于朴素贝叶斯分类器的一级标签分类模型，匹配留言分词在各类别中的权重，输出最大的权重和，以此来匹配对应的类别。具体步骤如下：

- **第一步，对数据中的无用词做停用处理**：通过网络资源下载通用“停用词”文档，对这些无用词做停用处理。被停用词详情请浏览附件中的“stopwords.txt”。
- **第二步，对留言内容进行分词**：通过 Python 对留言内容进行分词和去停用词处理，即读取文本文档，将 7 类文本做分词和去停用词处理，对 7 类文本的分词各进行词频统计，观察发现，出现了较多的数字、字母等词汇，这些词汇对结果帮助不大，故将它们删去，只保留有用的词汇。
- **第三步，计算各分类分词的权重**：将词频统计结果放入“result.xlsx”中，运用 excel 的公式计算分词所对应的*TF* 值，根据 *IDF* 的公式计算出 *IDF* 的值，计算公式如下：

$$TF=\frac{\text{词频}}{\text{总词数}},\quad IDF=log(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}}),$$

将每个词的*TF* 与 *IDF* 相乘，得到词的权重。部分数据如图 3 所示，全部数据见附件中“result.xlsx”，相关程序代码为“TFIDF.py”。

	A	B	C	D	E	F	G
1	分词	词频	TF	分词	词频	IDF	TFIDF
2	医生	992	0.017597	医生	312	0.437303	0.007695
3	患者	484	0.008586	患者	141	0.782239	0.006716
4	医院	1959	0.034751	医院	548	0.192677	0.006696
5	手术	333	0.005907	手术	94	0.95833	0.005661
6	生育	462	0.008195	生育	184	0.66664	0.005463
7	病人	355	0.006297	病人	137	0.794737	0.005005

图 3 部分分词结果示意图

- **第四步，借助基于朴素贝叶斯的分类器算法**：对留言进行分词，匹配分词在不同类别的权重，利用 Python 中的 *sum* 函数得到留言在 7 个标签的权重和，再利用 Python 中的 *max* 函数求出最大的权重和，输出权重和最大的值所对

应的标签，并对该留言进行标记。例如：对已知留言进行分词等一系列操作，如若权重和最大值的一类为城乡建设，则城乡建设就是这一留言的一级标签。具体实现过程如下：

Step 1：取“result.xlsx”中各一级标签*TFIDF* 值最大的前 1000 个分词，将其保存在“TFIDF.xlsx”中，作为各一级标签的关键词。

Step 2：读取“附件 2.xlsx”中的留言列，对留言进行分词和去停用词处理，运行后的分词写入“wordCount.xlsx”。

Step 3：通过代码中的 VLOOKUP 操作，将“TFIDF.xlsx”中分词的*TFIDF* 的值匹配给每个分词，并写入 result 文件夹中的 7 个 excel 中，得到该留言的分词在 7 个分类的权重。

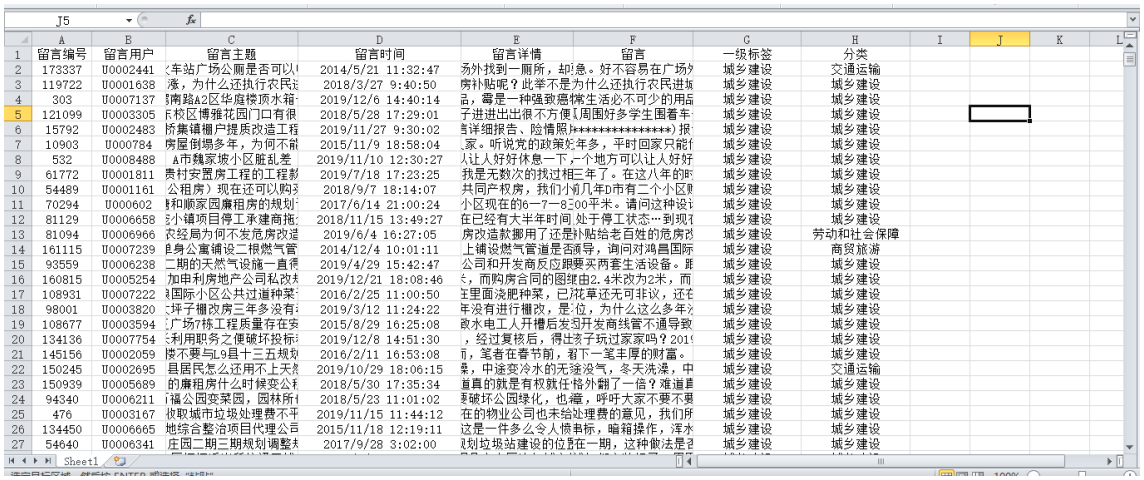
Step 4：计算所有分词在不同分类的权重和，使用 max 函数得到权重最大的分类，对留言做标签标记。

Step2-Step4 皆在“分类器.py”中实现。详情可看附件“分类器.py”。

4.4 模型求解和评价

4.4.1 模型求解

分类器算法部分运行结果如图 4 所示，全部数据见附件“分类器运行结果.xlsx”。



	留言编号	留言用户	留言主题	留言时间	留言详情	留言	一级标签	分类
2	173337	U0002441	车站广场公厕是否可以	2014/5/21 11:32:47	场外找到一厕所，却急。好不容易在广场外		城乡建设	交通运输
3	119722	U0001638	涨，为什么还执行农民进	2018/3/27 9:40:50	房补贴呢？此举不是为什么还执行农民进城		城乡建设	城乡建设
4	303	U0007137	南路A2区华庭楼顶水箱	2019/12/6 14:40:14	品，霉是一种强致癌物”生活必不可少的用品		城乡建设	城乡建设
5	121099	U0003305	校区博雅花园门口有很	2018/5/28 17:29:01	子进进出出很不方便！周围好多学生围着车		城乡建设	城乡建设
6	15792	U0002483	桥集镇棚户提质改造工程	2019/11/27 9:30:02	详细报告、险情照片**********）报		城乡建设	城乡建设
7	10903	U000784	房屋倒塌多年，为何不能	2015/11/9 18:58:04	家。听说党的政策好年多，平时回家只能		城乡建设	城乡建设
8	532	U0008488	A市魏家坡小区脏乱差	2019/11/10 12:30:27	让人好好休息一下，一个地方可以让人好好		城乡建设	城乡建设
9	61772	U0001811	贵村安置房工程的工程都	2019/7/18 17:23:25	我是无数次的找过相三年了。在这八年的时		城乡建设	城乡建设
10	54489	U0001161	公租房）现在还可以购买	2018/9/7 18:14:07	共同产权房，我们小前几年D市有二个小区更		城乡建设	城乡建设
11	70294	U000602	和顺家园廉租房的规划	2017/6/14 21:00:24	小区现在的6—7—8—00平米。请问这种设计		城乡建设	城乡建设
12	81129	U0006658	小镇项目得工承建商拖	2018/11/15 13:49:27	在已经有大半年时间,处于停工状态…到现在		城乡建设	城乡建设
13	81094	U0006966	农经局为何不发危房改造	2019/6/4 16:27:05	房改造款挪用了还是补贴给老百姓的危房改		城乡建设	劳动和社会保障
14	161115	U0007239	身公寓铺设二根燃气管	2014/12/4 10:01:11	上铺设燃气管道是否源导，询问对鸿昌国际		城乡建设	商贸旅游
15	93559	U0006238	二期的天然气设施一直很	2019/4/29 15:42:47	公司和开发商反应跟要买两套生活设备。跟		城乡建设	城乡建设
16	160815	U0005254	加申利房地产公司私改	2019/12/21 18:08:46	七，而购房合同的图纸由2.4米改为2米，而		城乡建设	城乡建设
17	108931	U0007222	国际小区公共过道种菜	2016/2/25 11:00:50	里面浇肥种菜，已沤草还无可非议，还在		城乡建设	城乡建设
18	98001	U0003820	坪子棚改房三年多没有	2019/3/12 11:24:22	年没有进行棚改，是:位，为什么这么多年		城乡建设	城乡建设
19	108677	U0003594	广场7栋工程质量存在安	2015/8/29 16:25:08	改水电工人开槽后发现开发商线管不通导致		城乡建设	城乡建设
20	134136	U0007754	利用职务之便破坏投标	2019/12/8 14:51:30	，经过复核后，得出孩子玩过家家吗？2019		城乡建设	城乡建设
21	145156	U0002059	楼不要与L9县十三五规划	2016/2/11 16:53:08	前，笔者在春节前，看下一笔丰厚的财富。		城乡建设	城乡建设
22	150245	U0002695	县居民怎么还用不上天然	2019/10/29 18:06:15	桑，中途变冷水的无缘没气，冬天洗澡，中		城乡建设	交通运输
23	150939	U0005689	的廉租房什么时候变公租	2018/5/30 17:35:34	道真的就是有权就任‘格外翻了一倍？难道是		城乡建设	城乡建设
24	94340	U0006211	福公园变菜园，园林所	2018/5/23 11:01:02	要破坏公园绿化，也罐，呼吁大家不要不要		城乡建设	城乡建设
25	476	U0003167	收取城市垃圾处理费不平	2019/11/15 11:44:12	在的物业公司也未给处理费的意见，我们所		城乡建设	城乡建设
26	134450	U0006665	地综合整治项目代理公司	2015/11/18 12:19:11	这是一件多么令人愤卑标，暗箱操作，浑水		城乡建设	城乡建设
27	54640	U0006341	庄园二期三期规划调整	2017/9/28 3:02:00	划垃圾站建设的位置在一期，这种做法是否		城乡建设	城乡建设

图 4 分类器运行结果部分数据

4.4.2 评价

利用 $F\text{-}Score$ 对一级标签分类模型进行评价。

$F\text{-}Score$ **算法**（以卫生计生为例）：

Step 1：读取附件 2 第 G 列的一级标签分类、第 H 列分类模型运行结果，依次对各分类进行计算。

Step 2：计算实际是“卫生计生”类的留言总数、分类后的卫生计生类留言数以及分类正确的总数，依次标记为 x, y, z 。

Step 3：利用查准率公式和查全率公式得出“卫生计生”的查准率 P_i 和查全率 R_i，公式如下：

$$P_i = \frac{z}{y}, \quad R_i = \frac{z}{x}.$$

Step 4：根据上述说明，计算其他 6 类的查准率和查全率，最后利用 $F\text{-}Score$ 的公式得出 $F\text{-}Score$ 的值。

运行结果如图 5 所示，将其保存到“F-Score.xlsx”中，详情见附件“F-Score.xlsx”。

```
Python 3.8.1 (tags/v3.8.1:1b293b6, Dec 18 2019, 22:39:24) [MSC v.1916 32 bit (In
tel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
======= RESTART: C:\Users\Administrator\Desktop\泰迪杯\C题\①\验证准确率\查准率
查全率.py =======
城乡建设 查准率： 0.8422344996930632 查全率： 0.6971544715447154
环境保护 查准率： 0.878099173553719 查全率： 0.9361233480176211
交通运输 查准率： 0.4973021582733813 查全率： 0.9388794567062818
教育文体 查准率： 0.9329268292682927 查全率： 0.8
劳动和社会保障 查准率： 0.8680229525299947 查全率： 0.8776371308016878
商贸旅游 查准率： 0.8169014084507042 查全率： 0.7618213660245184
卫生计生 查准率： 0.8201357466063348 查全率： 0.8489461358313818
F-score: 0.810872671937785
>>>
```

图 5 *F-Score* 运行结果

5 问题二的求解

5.1 问题解决思路

流程如图 6 所示：

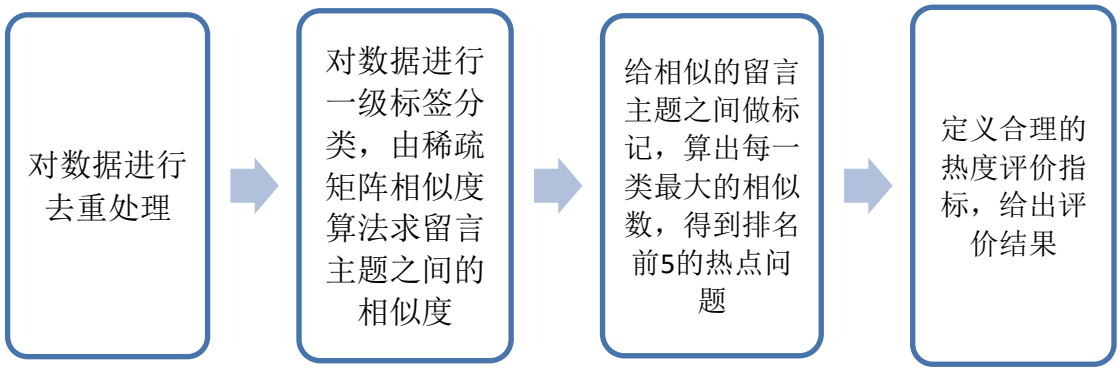


图 6 挖掘热点问题流程图

5.2 数据预处理

数据去重：分析附件 3 的数据，发现存在一些留言的留言编号、留言用户、留言主题、留言时间、留言详情、反对数、点赞数皆重复的重复数据。对附件 3 全部数据进行以全部列为准的“删除重复项”处理。

5.3 挖掘建模

针对该问题，需做以下准备：对数据进行一级标签分类、利用稀疏矩阵相似度的算法求出留言之间的相似度，并给相似留言主题之间贴上标签，求出相似留言主题的相似数，挖掘热点问题。具体步骤如下：

- **第一步，进行一级标签分类**：对附件 3 的留言详情进行一级标签分类，按照一级标签分类建立 7 个工作表，数据请查看“一级标签排名.xlsx”。
- **第二步，计算稀疏矩阵相似度**：（以城乡建设工作表为例）

Step 1：读取附件 3 城乡建设工作表的留言主题，对其进行分词、去停用词处理，由这些词建立词典。

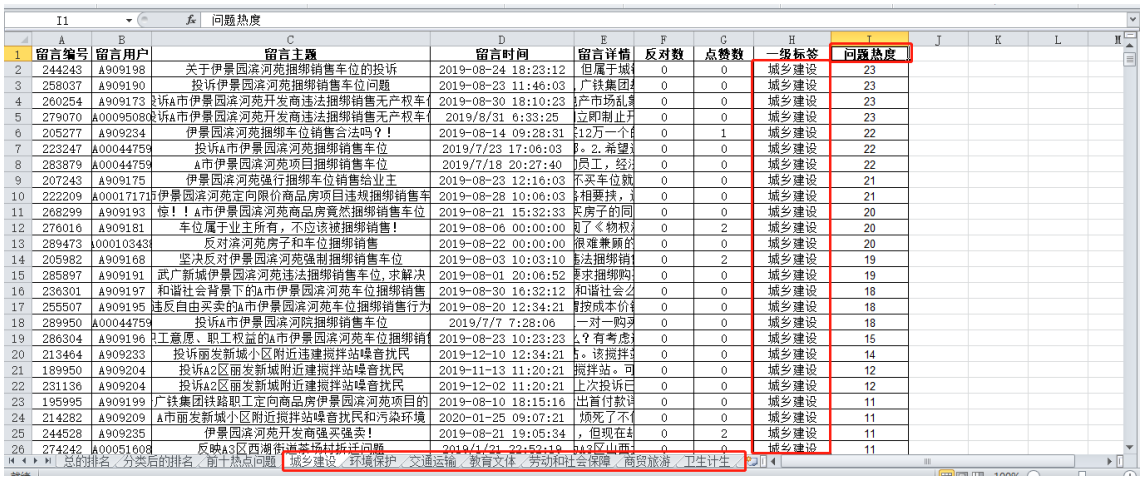
Step 2：对于第一条留言主题，分词和去停用词之后，对语料库进一步处理，得到新语料库，再对其进行*TFIDF* 处理得到对应的*TFIDF* 值，通过 Python 中的 token2id 操作，得到特征数，求出稀疏矩阵相似度，从而建立索引，最终得到第一条留言与其他留言的相似结果。对于第二条到最后一条的留言主题则依次进行上述操作。

Step 3：给留言主题标号，记第 i 条留言的 t 值为t_i。当相似结果r大于 0.4，则t_i为 1；当相似结果r小于 0.4，则t_i为 0。对 t 求和，并将求和的结果S作为第一条留言主题的相似数，即每一条留言主题的热度，填入城乡建设工作表的第 I 列。

$$t_i=\begin{cases}1, r\geq 0.4\\ 0, r<0.4\end{cases},\quad S=\sum t_i .$$

对于其余 6 个工作表，都做上述操作，得到的工作表如图 7 所示，详情见附件“一级标签排名.xlsx”。

以上步骤皆在“问题 2.py”中实现，详情见附件“问题 2.py”。

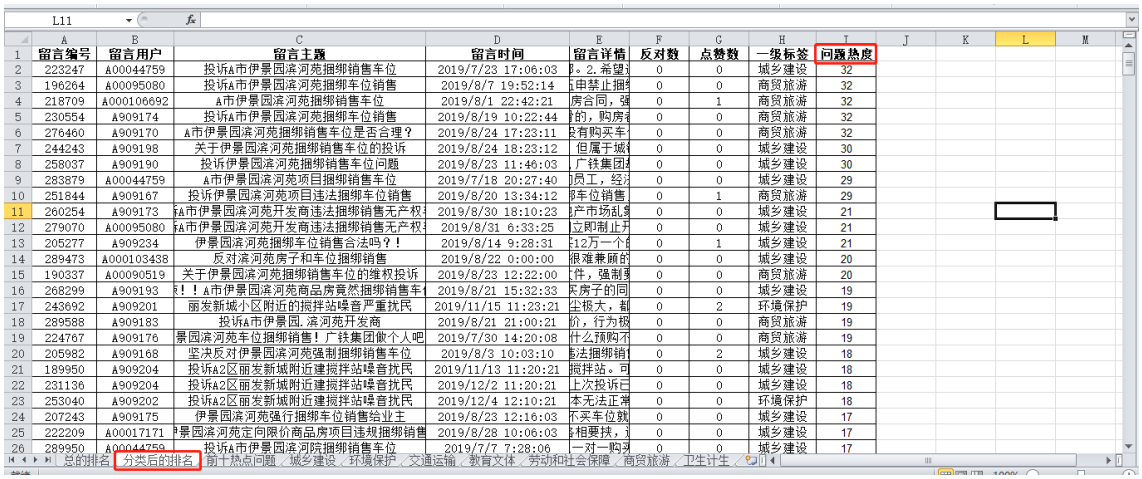


留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	一级标签	问题热度
244243	A909198	关于伊景园滨河苑捆绑销售车位的投诉	2019-08-24 18:23:12	但属于城	0	0	城乡建设	23
258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	2019-08-23 11:46:03	广铁集团	0	0	城乡建设	23
260254	A909173	投诉A市伊景园滨河苑开发商违法捆绑销售无产权车(	2019-08-30 18:10:23	产市场乱	0	0	城乡建设	23
279070	A00095080	投诉A市伊景园滨河苑开发商违法捆绑销售无产权车(	2019/8/31 6:33:25	立即制止开	0	0	城乡建设	23
205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	2019-08-14 09:28:31	12万一个	0	1	城乡建设	22
223247	A00044759	投诉A市伊景园滨河苑捆绑销售车位	2019/7/23 17:06:03	。2.希望	0	0	城乡建设	22
283879	A00044759	A市伊景园滨河苑项目捆绑销售车位	2019/7/18 20:27:40	员工，经	0	0	城乡建设	22
207243	A909175	伊景园滨河苑强行捆绑车位销售给业主	2019-08-23 12:16:03	不买车位就	0	0	城乡建设	21
222209	A00017171	伊景园滨河苑定向限价商品房项目违规捆绑销售车	2019-08-28 10:06:03	相要挟，	0	0	城乡建设	21
268299	A909193	惊！！A市伊景园滨河苑商品房竟然捆绑销售车位	2019-08-21 15:32:33	买房子的同	0	0	城乡建设	20
276016	A909181	车位属于业主所有，不应该被捆绑销售！	2019-08-06 00:00:00	句了《物权	0	2	城乡建设	20
289473	A00010343	反对滨河苑房子和车位捆绑销售	2019-08-22 00:00:00	很难兼顾的	0	0	城乡建设	20
205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	2019-08-03 10:03:10	违法捆绑销	0	2	城乡建设	19
285897	A909191	武广新城伊景园滨河苑违法捆绑销售车位，求解决	2019-08-01 20:06:52	要求捆绑购	0	0	城乡建设	19
236301	A909197	和谐社会背景下的A市伊景园滨河苑车位捆绑销售	2019-08-30 16:32:12	和谐社会么	0	0	城乡建设	18
255507	A909195	违反自由买卖的A市伊景园滨河苑车位捆绑销售行为	2019-08-20 12:34:21	按成本价	0	0	城乡建设	18
289950	A00044759	投诉A市伊景园滨河院捆绑销售车位	2019/7/7 7:28:06	一对一购	0	0	城乡建设	18
286304	A909196	工意愿、职工权益的A市伊景园滨河苑车位捆绑销	2019-08-23 10:23:23	么？有考虑	0	0	城乡建设	15
213464	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民	2019-12-10 12:34:21	5。该搅拌	0	0	城乡建设	14
189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-11-13 11:20:21	搅拌站。可	0	0	城乡建设	12
231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-12-02 11:20:21	上次投诉已	0	0	城乡建设	12
195995	A909199	广铁集团铁路职工定向商品房伊景园滨河苑项目的	2019-08-10 18:15:16	出首付款详	0	0	城乡建设	11
214282	A909209	A市丽发新城小区附近搅拌站噪音扰民和污染环境	2020-01-25 09:07:21	烦死了不	0	0	城乡建设	11
244528	A909235	伊景园滨河苑开发商强买强卖！	2019-08-21 19:05:34	，但现在	0	2	城乡建设	11
274242	A00051608	反映A3区西湖街道英坑村拆迁问题	2019/1/21 22:52:19	A2区山西	0	0	城乡建设	11

图 7 留言分类及其热度部分数据

- **第三步，对热点问题进行挖掘：**

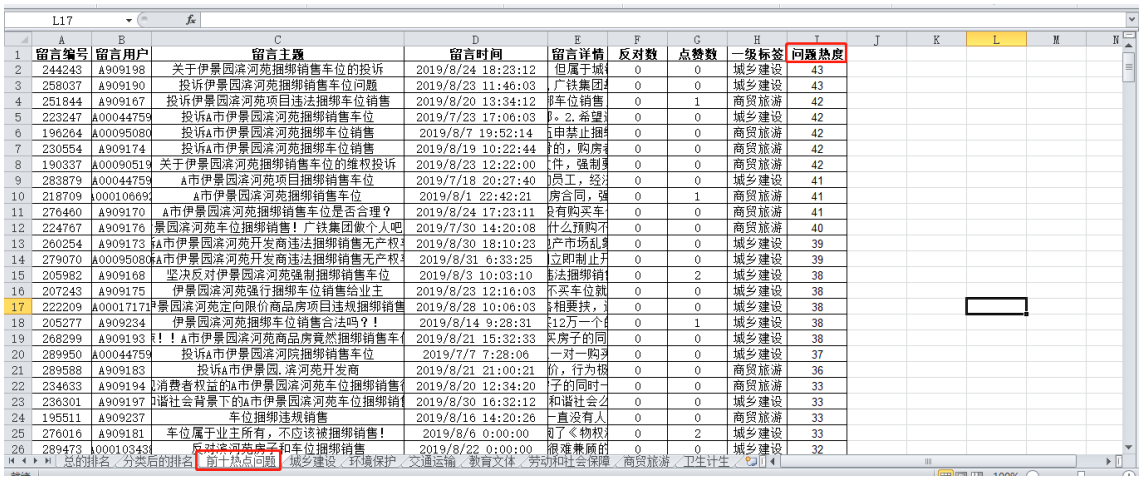
Step 1：对每一类留言主题的热度进行从大到小的排序，由于难以直接取出每一类热度排名前五的留言主题进行保存，故对留言主题热度最大值放大 5 倍，则热度排名前五的留言主题一定包含在所取的数据中。将它们放在“分类后的排名”工作表中。部分结果如图 8 所示，全部数据见附件“一级标签排名.xlsx”。



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	一级标签	问题热度
223247	A00044759	投诉A市伊景园滨河苑捆绑销售车位	2019/7/23 17:06:03	0。2.希望	0	0	城乡建设	32
196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019/8/7 19:52:14	申禁止捆	0	0	商贸旅游	32
218709	A000106692	A市伊景园滨河苑捆绑销售车位	2019/8/1 22:42:21	房合同，强	0	1	商贸旅游	32
230554	A909174	投诉A市伊景园滨河苑捆绑车位销售	2019/8/19 10:22:44	的，购房	0	0	商贸旅游	32
276460	A909170	A市伊景园滨河苑捆绑销售车位是否合理？	2019/8/24 17:23:11	有购买车	0	0	商贸旅游	32
244243	A909198	关于伊景园滨河苑捆绑销售车位的投诉	2019/8/24 18:23:12	但属于城	0	0	城乡建设	30
258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	2019/8/23 11:46:03	广铁集团	0	0	城乡建设	30
283879	A000044759	A市伊景园滨河苑项目捆绑销售车位	2019/7/18 20:27:40	员工，经	0	0	城乡建设	29
251844	A909167	投诉伊景园滨河苑项目违法捆绑车位销售	2019/8/20 13:34:12	车位销售	0	1	商贸旅游	29
260254	A909173	A市伊景园滨河苑开发商违法捆绑销售无产权	2019/8/30 18:10:23	产市场乱	0	0	城乡建设	21
279070	A00095080	A市伊景园滨河苑开发商违法捆绑销售无产权	2019/8/31 6:33:25	立即制止开	0	0	城乡建设	21
205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	2019/8/14 9:28:31	12万一个	0	1	城乡建设	21
289473	A000103438	反对滨河苑房子和车位捆绑销售	2019/8/22 0:00:00	很难兼顾的	0	0	城乡建设	20
190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019/8/23 12:22:00	件，强制	0	0	商贸旅游	20
268299	A909193	！！A市伊景园滨河苑商品房竟然捆绑销售车	2019/8/21 15:32:33	买房子的同	0	0	城乡建设	19
243692	A909201	丽发新城小区附近的搅拌站噪音严重扰民	2019/11/15 11:23:21	尘极大，都	0	2	环境保护	19
289588	A909183	投诉A市伊景园，滨河苑开发商	2019/8/21 21:00:21	价，行为极	0	0	商贸旅游	19
224767	A909176	景园滨河苑车位捆绑销售！广铁集团傲个人吧	2019/7/30 14:20:08	什么预购不	0	0	商贸旅游	19
205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	2019/8/3 10:03:10	法捆绑销	0	2	城乡建设	18
189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/11/13 11:20:21	搅拌站。可	0	0	城乡建设	18
231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/12/2 11:20:21	上次投诉已	0	0	城乡建设	18
253040	A909202	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/12/4 12:10:21	本无法正常	0	0	环境保护	18
207243	A909175	伊景园滨河苑强行捆绑车位销售给业主	2019/8/23 12:16:03	不买车位就	0	0	城乡建设	17
222209	A00017171	景园滨河苑定向限价商品房项目违规捆绑销售	2019/8/28 10:06:03	相要挟，	0	0	城乡建设	17
289950	A00044759	投诉A市伊景园滨河院捆绑销售车位	2019/7/7 7:28:06	一对一购买	0	0	城乡建设	17

图 8 分类后的排名部分数据

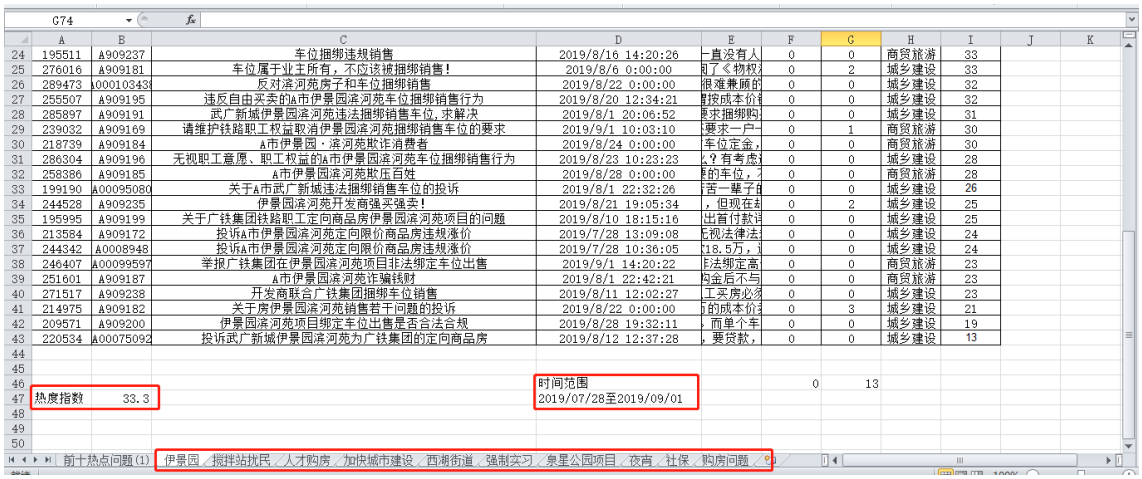
Step 2：对“分类后的排名”工作表进行上述类似的处理，考虑到要将“点赞数与反对数”作为参考，这次取热度排名前十的留言主题。对留言主题热度最大值放大 10 倍，确保排名前十的留言主题在所取数据之中，并将他们放入“前十热点问题”工作表中。部分结果如图 9 所示，全部数据见附件“一级标签排名.xlsx”。



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	一级标签	问题热度
244243	A909198	关于伊景园滨河苑捆绑销售车位的投诉	2019/8/24 18:23:12	但属于城	0	0	城乡建设	43
258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	2019/8/23 11:46:03	广铁集团	0	0	城乡建设	43
251844	A909167	投诉伊景园滨河苑项目违法捆绑车位销售	2019/8/20 13:34:12	绑车位销售	0	1	商贸旅游	42
223247	A00044759	投诉A市伊景园滨河苑捆绑销售车位	2019/7/23 17:06:03	0。2.希望	0	0	城乡建设	42
196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019/8/7 19:52:14	申禁止捆	0	0	商贸旅游	42
230554	A909174	投诉A市伊景园滨河苑捆绑车位销售	2019/8/19 10:22:44	的，购房	0	0	商贸旅游	42
190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019/8/23 12:22:00	件，强制	0	0	商贸旅游	42
283879	A00044759	A市伊景园滨河苑项目捆绑销售车位	2019/7/18 20:27:40	员工，经	0	0	城乡建设	41
218709	A000106692	A市伊景园滨河苑捆绑销售车位	2019/8/1 22:42:21	房合同，强	0	1	商贸旅游	41
276460	A909170	A市伊景园滨河苑捆绑销售车位是否合理？	2019/8/24 17:23:11	有购买车	0	0	商贸旅游	41
224767	A909176	景园滨河苑车位捆绑销售！广铁集团傲个人吧	2019/7/30 14:20:08	什么预购不	0	0	商贸旅游	40
260254	A909173	A市伊景园滨河苑开发商违法捆绑销售无产权	2019/8/30 18:10:23	产市场乱	0	0	城乡建设	39
279070	A00095080	A市伊景园滨河苑开发商违法捆绑销售无产权	2019/8/31 6:33:25	立即制止开	0	0	城乡建设	39
205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	2019/8/3 10:03:10	法捆绑销	0	2	城乡建设	38
207243	A909175	伊景园滨河苑强行捆绑车位销售给业主	2019/8/23 12:16:03	不买车位就	0	0	城乡建设	38
222209	A00017171	景园滨河苑定向限价商品房项目违规捆绑销售	2019/8/28 10:06:03	相要挟，	0	0	城乡建设	38
205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	2019/8/14 9:28:31	12万一个	0	1	城乡建设	38
268299	A909193	！！A市伊景园滨河苑商品房竟然捆绑销售车	2019/8/21 15:32:33	买房子的同	0	0	城乡建设	38
289950	A00044759	投诉A市伊景园滨河院捆绑销售车位	2019/7/7 7:28:06	一对一购买	0	0	城乡建设	37
289588	A909183	投诉A市伊景园，滨河苑开发商	2019/8/21 21:00:21	价，行为极	0	0	商贸旅游	36
234633	A909194	消费者权益的A市伊景园滨河苑车位捆绑销售	2019/8/20 12:34:20	子的同时	0	0	商贸旅游	33
236301	A909197	和谐社会背景下的A市伊景园滨河苑车位捆绑销	2019/8/30 16:32:12	和谐社会之	0	0	城乡建设	33
195511	A909237	车位捆绑违规销售	2019/8/16 14:20:26	一直没有人	0	0	商贸旅游	33
276016	A909181	车位属于业主所有，不应该被捆绑销售！	2019/8/6 0:00:00	了个物权	0	2	城乡建设	33
289473	A000103438	反对滨河苑房子和车位捆绑销售	2019/8/22 0:00:00	很难兼顾的	0	0	城乡建设	32

图 9 前十热点问题部分数据

Step 3：将“前十热点问题”工作表放入“问题 2.py”中运行，反复对比，发现相似度参数取为 0.2 更为合适，从而挑选出排名前十的热点问题，最终得到伊景园等十个问题的十个工作表。部分结果如图 10 所示，全部数据见附件“热点前五汇总.xlsx”。



	A	B	C	D	E	F	G	H	I
24	195511	A909237	车位捆绑违规销售	2019/8/16 14:20:26	一直没有人	0	0	商贸旅游	33
25	276016	A909181	车位属于业主所有，不应该被捆绑销售！	2019/8/6 0:00:00	则了《物权	0	2	城乡建设	33
26	289473	A00010343	反对滨河苑房子和车位捆绑销售	2019/8/22 0:00:00	很难兼顾的	0	0	城乡建设	32
27	255507	A909195	违反自由买卖的A市伊景园滨河苑车位捆绑销售行为	2019/8/20 12:34:21	暂按成本价	0	0	城乡建设	32
28	285897	A909191	武广新城伊景园滨河苑违法捆绑销售车位，求解决	2019/8/1 20:06:52	要求捆绑购	0	0	城乡建设	31
29	239032	A909169	请维护铁路职工权益取消伊景园滨河苑捆绑销售车位的要求	2019/9/1 10:03:10	要求一户一	0	1	商贸旅游	30
30	218739	A909184	A市伊景园·滨河苑欺诈消费者	2019/8/24 0:00:00	车位定金，	0	0	商贸旅游	30
31	286304	A909196	无视职工意愿、职工权益的A市伊景园滨河苑车位捆绑销售行为	2019/8/23 10:23:23	么？有考虑	0	0	城乡建设	28
32	258386	A909185	A市伊景园滨河苑欺压百姓	2019/8/28 0:00:00	的车位，	0	0	商贸旅游	28
33	199190	A00095080	关于A市武广新城违法捆绑销售车位的投诉	2019/8/1 22:32:26	否一辈子	0	0	城乡建设	26
34	244528	A909235	伊景园滨河苑开发商强买强卖！	2019/8/21 19:05:34	，但现在	0	2	城乡建设	25
35	195995	A909199	关于广铁集团铁路职工定向商品房伊景园滨河苑项目的问题	2019/8/10 18:15:16	出首付款	0	0	城乡建设	25
36	213584	A909172	投诉A市伊景园滨河苑定向限价商品房违规涨价	2019/7/28 13:09:08	无视法律法	0	0	城乡建设	24
37	244342	A0008948	投诉A市伊景园滨河苑定向限价商品房违规涨价	2019/7/28 10:36:05	18.5万，	0	0	城乡建设	24
38	246407	A00099597	举报广铁集团在伊景园滨河苑项目非法绑定车位出售	2019/9/1 14:20:22	非法绑定高	0	0	商贸旅游	23
39	251601	A909187	A市伊景园滨河苑诈骗钱财	2019/8/1 22:42:21	购金后不与	0	0	商贸旅游	23
40	271517	A909238	开发商联合广铁集团捆绑车位销售	2019/8/11 12:02:27	工买房必须	0	0	城乡建设	23
41	214975	A909182	关于房伊景园滨河苑销售若干问题的投诉	2019/8/22 0:00:00	的成本价	0	3	城乡建设	21
42	209571	A909200	伊景园滨河苑项目绑定车位出售是否合法合规	2019/8/28 19:32:11	而单个车	0	0	城乡建设	19
43	220534	A00075092	投诉武广新城伊景园滨河苑为广铁集团的定向商品房	2019/8/12 12:37:28	要贷款，	0	0	城乡建设	13
44									
45									
46				时间范围		0	13		
47	热度指数	33.3		2019/07/28至2019/09/01					
48									
49									
50									

前十热点问题(1) 伊景园 搅拌站扰民 人才购房 加快城市建设 西湖街道 强制实习 泉星公园项目 夜宵 社保 购房问题

图 10 十个热点问题的工作表部分数据

Step 4：定义合理的热度指数Q：由于存在对留言主题的点赞数和反对数，对每一热点问题留言主题的热度q乘以 0.7，加上点赞数m与反对数n的差值乘以 0.3，以此作为留言的热度指数。即：

$$Q=q\times 0.7+(m-n)\times 0.3$$

按热度指数对十个热点问题进行从大到小的排序，选取热度指数最高的五个热点问题。

5.4 模型求解

按照挖掘热点问题的步骤，运行“问题 2.py”，从而找出排名前五的热点问题，将其汇总并保存为“热点问题表.xlsx”，数据如图 11 所示，详情见附件“热点问题表.xlsx”。

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	36.8	2019/11/2至2020/1/25	A2区丽发新城附近	投诉A2区丽发新城附近建搅拌站噪音扰民
2	2	33.3	2019/7/28至2019/9/1	A市伊景园滨河苑	伊景园滨河苑强行捆绑车位销售给业主
3	3	24.1	2019/01/04至2019/11/01	A市国家中心城市	请A市加快国家中心城市建设力度
4	4	17.1	2019/01/16至2019/09/27	A市人才	咨询A市人才购房补贴政策
5	5	14.1	2019/08/09至2019/08/26	A市经开区泉星公园	建议A市经开区泉星公园项目规划进一步优化

图 11 前 5 热点问题

对于每个热点问题的留言明细，部分数据如图 12 所示，详情见附件“热点问题留言明细表.xlsx”。

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	272447	A909206	投诉小区附近建设搅拌站	2019/11/20 19:12:22	投诉A市暮云街道丽发新城边上在建大型搅拌站，作为一个逾3万人的大型社区，边上竟然在建大型搅拌站，	0	0
1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/11/13 11:20:21	我是A2区丽发新城小区的一名业主，我要投诉同发投资有限公司在未经小区业主同意的情况下，在离小区不到	0	0
1	231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/12/02 11:20:21	尊敬的领导，我是A2区丽发新城小区的一名业主，再次投诉同发投资有限公司在未经小区业主同意的情况下，	0	0
1	253040	A909202	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/12/04 12:10:21	投诉A2区丽发新城小区附近违建搅拌站！该站每天早6点一直到晚8点设备都在运行，每天耳边都是各种嗡鸣	0	0
1	239336	A909213	A市A2区丽发新城小区遭搅拌站严重污染	2019/12/11 11:44:11	敬爱的领导，您好！最近A2区丽发新城小区不断有粉尘遮天，杂音刺耳，经了解是附近新建搅拌厂的“佳作	0	0

图 12 热点问题留言明细部分数据

6 问题三的求解

6.1 问题解决思路

由于在一个答复意见中，完整性与答复意见的回复格式有关，可解释性与答复意见中解释网民问题的内容有关，故将答复意见分成两个文本，用于提取完整性关键词和可解释性关键词，将关键词的出现频率作为关键词的权重指标。对每个答复意见所拥有的关键词权重进行求和，最终得到答复意见的完整性评价和可解释性评价。

对于相关性评价，可借助稀疏矩阵相似度算法求得留言主题与答复意见的相关性。

流程如图 13 所示：

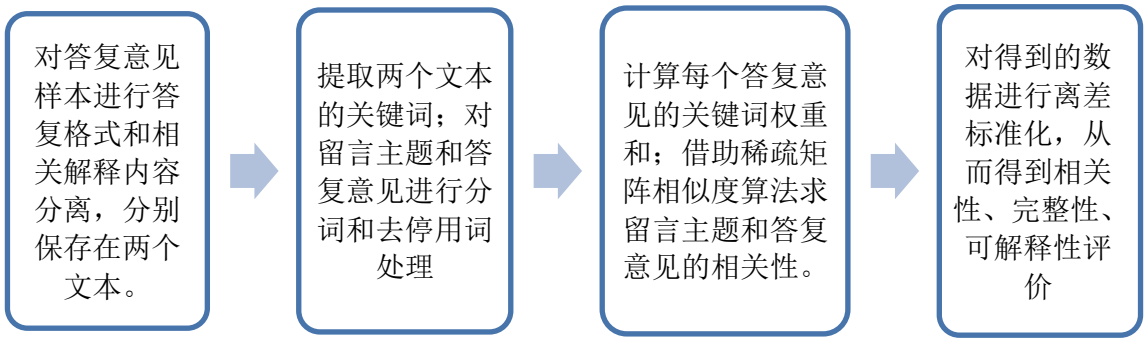


图 13 所示答复意见评价表

6.2 数据预处理

随机抽取样本：分析附件 4 的数据，从附件 4 的 2816 条答复意见中随机抽取 281 条数据作为样本，详情见附件“答复意见样本.txt”。

分离样本数据：人工分离答复意见样本的答复格式和相关解释内容，将分离完的样本保存为“答复意见完整性.txt”与“答复意见可解释性.txt”，详见预处理文件夹。

6.3 挖掘建模

针对完整性评价和可解释性评价：通过词频统计“答复意见完整性.txt” 与“答复意见可解释性.txt”，得到完整性评价与可解释评价的关键词。计算关键词出现的频率，以此作为关键词的权重。对答复意见进行分词、去重复词、保留完整性关键词与可解释性关键词，匹配关键词对应的权重，将所有关键词的权重相加，得到答复意见的完整性评价和可解释性评价。具体步骤如下：

- **第一步，提取关键词：**对“答复意见完整性.txt ”与“答复意见可解释性.txt ”进行分词、去重复词、去停用词、词频统计，取词频最高的前 50 个词作为完整性和可解释性的关键词，保存为“完整性关键词.txt”和“可解释性关键词.txt”，详情见附件“完整性关键词.txt”、“可解释性关键词.txt”。
- **第二步，计算关键词权重：**计算完整性关键词和可解释性关键词对应的词频，将运行结果放入“完整性关键词权重.xlsx”和“可解释性关键词权重.xlsx”，通过 excel 公式计算关键词的出现频率作为对应的权重，公式如下：

$$\text{出现频率}=\frac{\text{词频}}{\text{总文档数}}.$$

- **第三步，对答复意见进行完整性评价和可解释性评价：**运行代码“答复意见完整性.py”和“答复意见可解释性.py”，对答复意见进行分词、去重复词、保留关键词，匹配关键词对应的权重，将所有关键词的权重相加，得到答复意见的完整性和可解释性。将运行结果放入“答复意见评价.xlsx”中的第 I 列和第 J 列。

针对相关性评价：分别对附件 4 中的留言主题和留言详情进行分词、去停用词处理，借助问题 2 稀疏矩阵相似度的算法求得留言主题和答复意见的相关性，运行“留言相关性.py”并把运行结果保存在“答复意见评价.xlsx”中的第 H 列。详情见附件“答复意见评价.xlsx”。

离差标准化：对答复意见的相关性、完整性和可解释性分数进行离差标准化，作为相关性、完整性和可解释性评价。公式如下:

$$x^* = \frac{x - min}{max - min}. \qquad (1)$$

详情见“答复意见评价.xlsx”中的第 K 列、第 L 列和第 M 列。

6.4 模型求解

通过运行代码“答复意见完整性评价.py”、“答复意见可解释性评价.py”、“问题 3 相似性.py”得到相关性评价、完整性评价和可解释性评价，再利用离差标准化，通过 excel 操作运用公式（1），得到离差标准化后的相关性评价、完整性评价和可解释性评价。最终对相关性评价、完整性评价和可解释性评价进行求均值，得到综合评价。部分数据如图 14 所示，详情见附件“答复意见评价.xlsx”。

答复意见	答复时间	相关性	完整性	可解释性	相关性评价	完整性评价	可解释性评价	综合评价
收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	0.362000853	5.326704545	2.289772727	0.408158079	0.723658819	0.489263222	0.54036004
需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	0.184870183	2.084517045	1.171875	0.208442213	0.283191818	0.250398361	0.247344131
办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	0.309548229	4.3984375	1.000355114	0.349017439	0.597549209	0.213749146	0.386771931
年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	0.172543198	6.171164773	1.616832386	0.194543465	0.838382864	0.34547386	0.459466729
保留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	0.112236276	2.021661932	0.19140625	0.126547058	0.274652644	0.040898399	0.147366034
于您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	0.481988698	0.818536932	0.716974432	0.543445076	0.111202239	0.15319827	0.269281862
A市A3区人民政府办公室下发了《关于A市A3	2019/5/9 10:18:58	0.429861248	2.026278409	1.604403409	0.484671071	0.275279815	0.34281812	0.367589668
修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	0.358689547	6.600497159	1.97265625	0.404424562	0.896709765	0.421503908	0.574212745
单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	0.722557664	2.548650568	1.248934659	0.814687992	0.346246623	0.26686395	0.475932855
部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	0.281607002	5.6015625	1.463778409	0.317513542	0.760999614	0.312770316	0.463761157
司支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	0.500792444	3.982954545	1.407315341	0.564646409	0.541103821	0.300705668	0.468818633
续，按长人防发[2014]7号文件要求，鄱阳	2019/1/29 10:52:01	0.67094028	1.338068182	1.760298295	0.75648909	0.181783095	0.37612869	0.438133625
分局配合进行具体选址，招标（邀标）进行	2019/1/14 14:34:58	0.420230985	1.075639205	1.485440341	0.473812893	0.146130838	0.317398892	0.312447541
映的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	0.385643125	3.489701705	1.873934659	0.434814879	0.474093014	0.400409743	0.436439212
正常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	0.146195427	4.974076705	0.586292614	0.164836199	0.675752605	0.125275059	0.321954621
出的“披塘路路口两端各拆除20米中间花坛，	2019/3/6 10:26:14	0.314539164	4.533380682	3.183948864	0.354644747	0.615881899	0.680324759	0.550283802
根据您提供的信息进行投诉信息的登记分送	2019/1/3 14:02:47	0.085757248	6.08984375	2.741832386	0.096691799	0.827335006	0.585856287	0.503294364
营业。梅溪湖二期金菊路与雪松路东南角规	2019/1/14 14:32:40	0.011298059	2.954900568	2.443181818	0.012738628	0.401437669	0.522042644	0.31207298
查，施工单位由于需要夜间连续作业，已办	2019/1/8 16:19:16	0.431438744	3.15625	0.924715909	0.486449707	0.428791972	0.19758707	0.370942917
以上不同机构，需三方或三方以上不同机构	2019/1/4 15:48:23	0.144939944	4.817826705	0.124289773	0.163420635	0.65452528	0.026557402	0.281501106
具体上线时间请关注潇洒支付公司官网htt	2019/1/4 15:49:46	0.166025236	4.793678977	0.158025568	0.187194425	0.651244693	0.03376584	0.290734986
桥北组签订了土地补偿协议，并按协议达成	2019/1/8 16:18:00	0.177381262	5.323153409	2.508877841	0.199998411	0.72317638	0.536080127	0.486418306

图 14 答复意见评价部分数据

7 小结

针对问题一构建的一级标签分类模型，在实现过程中，能够较为方便的对留言进行分类，不足之处在于准确度并不高。

针对问题二构建的文本相似度模型，应用比较广泛，利用稀疏矩阵相似度的算法计算文本的相似性，算法利用率高。不足之处是：在前十问题的提取过程中，过于依赖人工挑选，造成不小的误差。

针对问题三，是基于问题一与问题二的模型求解的。借助文本相似度模型求解出留言主题与答复意见之间的相似性，从而得出相关性评价；对于完整性评价和可解释性评价，则是对答复意见按照格式和答复内容进行分离数据，分别为其匹配对应的权重，根据权重和得出完整性评价和可解释性评价。最终对相关性评价、完整性评价和可解释性评价进行求均值，得到综合评价。

随着这三个问题的解决，对提升政府的留言管理水平和施展针对政策效率具有极大的推动作用。

附录

附件	附件名称	备注
1	stopwords.txt	停用词文档
2	训练集.xlsx	预处理后的附件 2
3	留言内容文件夹	7 类一级标签的留言内容
4	result.xlsx	7 类留言内容分词的 TFIDF 值
5	TFIDF.xlsx	7 类留言内容各 1000 个关键词及对应的权重
6	**分类器运行结果.xlsx**	**将分类器运行结果放入第 H 列**
7	**F-score.xlsx**	**分类模型的 F-score 评价结果**
8	TFIDF 词频.py	对训练集的留言内容进行分词、词频计算
9	分类器.py	对训练集的留言内容进行分类
10	F-score.py	对分类模型进行评价
11	dict.txt	通过语料库将文档的词语进行建立的词典
12	res.xls	问题 2.py 的运行结果
13	一级标签排名.xlsx	各类一级标签的排名、总的排名以及分类后的排名
14	热点前 5 汇总.xlsx	从前十热点问题中挑出 10 个热点问题并放在后面的工作表
15	**热点问题表.xlsx**	**前五个热点问题的最终汇总**
16	**热点问题留言明细表.xlsx**	**前五个热点问题明细的最终汇总**
17	问题 2.py	问题 2 中稀疏矩阵相似度算法实现的代码，对具体问题要更改参数
18	答复意见样本.txt	对附件 4 的 2816 条答复意见随机抽取 281 条作为样本
19	答复意见完整性.txt	对答复意见样本进行分离保留答复格式
20	答复意见可解释性.txt	对答复意见样本进行分离保留相关解释内容

21	答复意见.txt	附件 4 的全部答复意见
22	完整性关键词.txt	完整性评价的关键词
23	可解释性关键词.txt	可解释性评价的关键词
24	完整性关键词权重.xlsx	完整性评价的关键词及对应的权重
25	可解释性关键词权重.xlsx	可解释性评价的关键词及对应的权重
26	**答复意见评价.xlsx**	**答复意见的相关性、完整性、可解释性评价结果**
27	提取完整性关键词.py	对答复意见完整性.txt 进行分词、词频计算
28	提取可解释性关键词.py	对答复意见可解释性.txt 进行分词、词频计算
29	完整性词频计算.py	对答复意见.txt 进行完整性关键词权重计算
30	可解释性词频计算.py	对答复意见.txt 进行可解释性关键词权重计算
31	答复意见完整性.py	对附件 4 的答复意见进行完整性评价
32	答复意见可解释性.py	对附件 4 的答复意见进行可解释性评价
33	答复意见相关性.py	对附件 4 的答复意见进行相关性评价

参考文献

[1] 郭肇毅.文本主题提取及相似度计算系统研究与开发[J].现代信息科技,2017,1(04):20-22.

[2] 王阳,周云才.朴素贝叶斯分类算法的设计与分析[J].电脑知识与技术,2019,15(11):206-208.

[3] 于游,付钰,吴晓平.中文文本分类方法综述[J].网络与信息安全学报,2019,5(05):1-8.

[4] 艾楚涵,姜迪,吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术,2020,44(04):65-70.