“智慧政务”中的文本挖掘应用

摘要

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。因此，建立基于自然语言处理技术的智慧政务系统对提升政府的管理水平和施政效率具有重大意义。

对于问题 1 群众留言分类，首先提取附件 2 中留言详情与一级标签内容，并对数据进行乱序和去重。利用正则表达式对数据进行无用序列去除，并通过 jieba 中文分词和停用词表对数据进行切分，将分词用空格连接成字符串作为模型的自变量；将去重数据的对应标签保存，作为模型的因变量。再利用 sklearn 中的数据集按 8:2 的比例随机切分。对训练集和测试集的数据计算 TF-IDF 权值，通过互补朴素贝叶斯模型训练，最后通过 f1_score 计算模型精确率和召回率的调和平均值。

对于问题 2 热点问题挖掘，通过附件 3 数据显示，在留言主题和留言内容中，留言用户会表述其想反映的现象的地点及主要情况。通过对每条留言主题和留言内容的文本相似度分析，将附件 3 中的留言作问题归类，类内问题相似度较高，类外相似度较低。在同一个问题中以合理的热度评价指标对问题进行热度值计算并排序。

对于问题 3 答复意见的评价，分析附件 4 可知，较完整的答复包含留言用户昵称，且回复内容与用户的留言详情相关性高。首先提取出用户昵称、留言内容和答复意见，合并用户昵称、留言主题与详情作为关键词，答复意见作为比对文本。将关键词和比对文本分别进行 jieba 分词和去停用词处理，在用封装好的函数计算关键词与比对文本相似度作为答复意见质量。

关键词：中文分词　去重　TF-IDF 算法　文本相似度　SinglePass 单遍聚类

目录

1、挖掘目标

党的十九大报告提出，需加快前沿技术创新，为数字中国、智慧社会建设提供有力支撑。在现代数字社会的环境下，许多网络问政平台沉淀海量与社会经济、公民生活密切相关的数据，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，依靠人工来进行留言划分和热点整理无疑会降低效率。基于这个问题，发展基于自然语言处理技术的智慧政务服务刻不容缓。利用文本挖掘与分析技术，直观、快速、精准地获取留言中的热点问题，并进行及时有效地反馈，这将有助于相关部门进行有针对性地处理，提升服务效率。

本次建模目标是利用四个附件中的数据，通过 jieba 中文分词工具对数据进行分词、互补朴素贝叶斯模型、（LDA 模型）、文本相似度计算，达到以下三个目标：

1）利用 jieba 分词对附件 1 中的数据进行分词，通过多项式贝叶斯模型训练，建立关于留言内容的一级分类模型，并用 F-score 对分类方法进行评价。

2）

3）针对附件 4 相关部门对留言的答复意见，提取用户昵称和留言内容，通过 gensim 库中的文本相似度分析对答复意见的质量做出评价。

2、分析方法与过程

2.1 问题 1 分析方法与过程

2.1.1 流程图

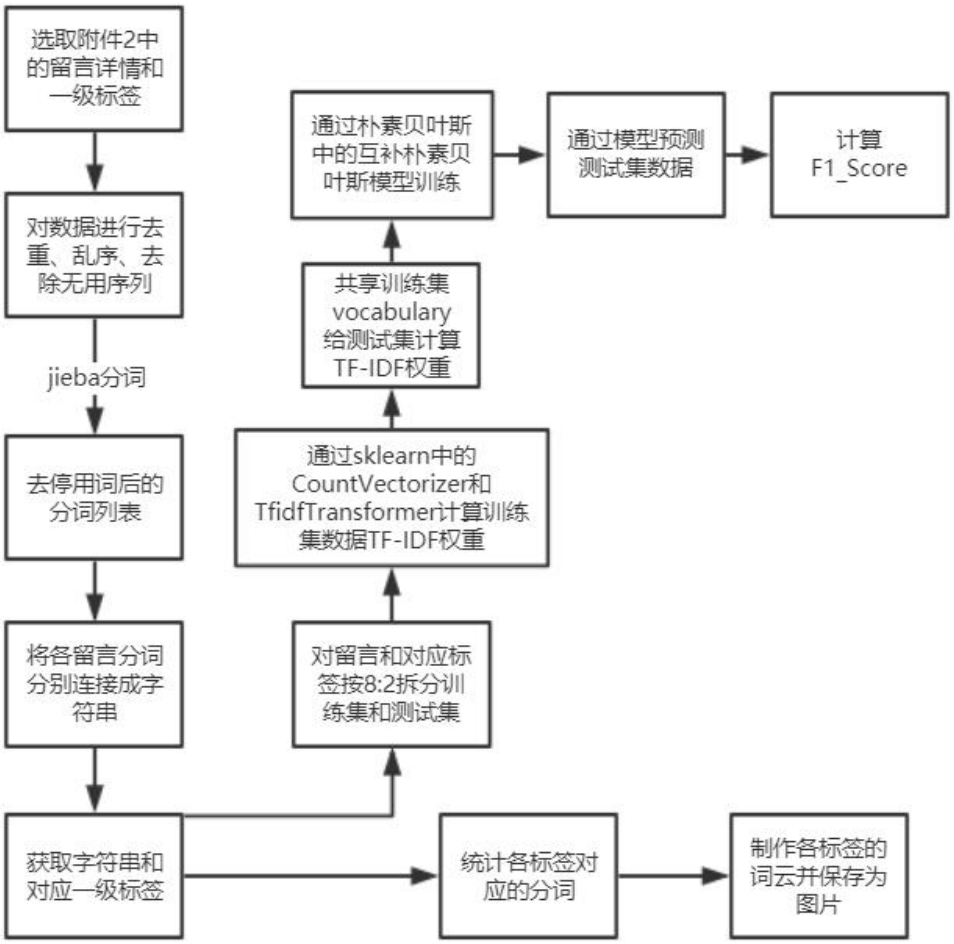

图 1.问题 1 流程图

2.1.2 数据预处理

2.1.2.1 留言详情的提取、去重

首先通过 pandas 读取附件 2 中的第 1、5、6 列，即留言编号、留言详情、一级标签。留言编号做数据的索引。虽然在分割数据集时使用了随机切分训练集和测试集函数，但是仍然为提取出的数据进行乱序排列，以便增大数据的无序程度，在分割数据集时抽取到训练集和测试集的数据不同标签数量更随机，更好的训练模型。接着对留言详情进行去重，减少重复数据对标签预测的影响，缩短数据处理和模型训练的时间。

2.1.2.2 留言详情去除无用序列和去停用词分词操作

观察留言详情数据，可以发现其中含有许多包含“\t”“\n”的无用字符串序列，利用正则表达式和匿名函数分别对每条留言进行去除无用序列操作。

通过 jieba 分词对去除无用序列后的数据进行分词，得到包含标点符号等停用词的分词列表。读入停用词表“stopword.txt”，再增加词表中没有的停用词，再次对切分后的数据进行过滤，去除停用词，得到 data_after_stop 分词列表。

2.1.2.3 得到训练模型需要的数据

通过去停用词后的数据索引，提取对应的一级标签 labels。将每条留言去停用词后的分词列表用空格连起来构成字符串 data_str。这样，就可以对 data_str 和 labels 进行监督式机器学习。

2.1.3 各标签词云制作

通过数据处理得到的分词 data_after_stop，统计各个标签中不同分词出现的次数。读入词云背景图，利用 wordcloud 库进行词云绘制，并将每个标签的词云图保存成图片，命名为”标签名”.jpg。

2.1.4 模型构建

2.1.4.1 分割数据集

通过 sklearn 中的 train_test_split 函数对数据处理得到的 data_str 和 labels 进行训练集和测试集的随机切分，其中训练集和测试集的比例为 8:2。

2.1.4.2 TF-IDF 算法

要对文本数据进行模型训练，首先要将文本转换为词频向量。采用 TF-IDF 权重策略，将数据集转换为权重向量。权重策略文档中的高频词应具有表征此文档较高的权重，除非该词也是高文档频率词。TF-IDF 算法的具体原理如下：
第一步，计算词频。

$$词频(TF)=某个词在文章中的出现次数$$

考虑到文章有长短之分，为了便于不同文章的比较，进行“词频”标准化。

$$词频(TF)=\frac{某个词在文章中的出现次数}{文章的总词数} \qquad (1)$$

或者

$$词频(TF)=\frac{某个词在文章中的出现次数}{该文出现次数最多的词的出现次数} \qquad (2)$$

第二步，计算逆文档频率。
这时，需要一个语料库（corpus），用来模拟语言的使用环境。

$$逆文档频率(IDF)=\log(\frac{语料库的文档总数}{包含该词的文档数+1}) \qquad (3)$$

如果一个词越常见，分母就越大，逆文档频率就越接近 0。分母加 1 时为了避免分母为 0（即所有文档都不包含该词），然后对得到的值取对数。

第三步，计算 TF-IDF。

$$TF-IDF=词频(TF)\times 逆文档频率(IDF) \qquad (4)$$

可见，TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语言中的出

现次数成反比。

2.1.4.3 模型训练

由于训练集样本和测试集样本上提取的 feature 维度不同，首先让两个 CountVectorizer 共享 vocabulary 使训练集和测试集样本向量长度保持一致。再通过 sklearn 计算出训练集和测试集的 TF-IDF 权重。通过朴素贝叶斯中的互补朴素贝叶斯训练模型，用模型预测测试集样本，并将预测结果和 f1_score 输出到文件 tag_classify.txt。

2.2 问题 2 分析方法与过程

2.2.1 基本流程

热点问题挖掘是先进行话题发现，并对问题按照合理的热度评价指标进行热度评价后，以热度指数对已归类的问题进行排列，排名靠前的问题即为被更多群众关注的问题，即热点话题。热点问题是在某一时间段群众集中反映的问题，在社会舆论中关注度较大，因此热点问题的发掘能让政府更加关注当前的民生热点，及时处理，提高政府办事效率。

问题发现是通过发现问题并将相关内容的问题归为一类，通过一定的聚类模型和方法，留言信息被聚集到有限的话题类簇中，类内相似度高，类间相似度低，即将不同的留言按照相似度算法进行融合。

基本步骤：

（1）利用 jieba 分词对留言进行分词处理并去除停用词（stopword），即诸如“的”、“是”等语句中最常用的字词，其在语句中出现的频率很高，但对找到结果没有帮助，因此需要把这类词过滤。

（2）提取文本关键词，对留言中出现的词进行特征权重（TF-IDF）计算

（3）文本相似度计算（如余弦相似度）。

（4）文本聚合，将留言内容进行聚类。

（5）得出“问题”，通过聚类得出的类簇，每个类簇代表了相类似的问题，结合时间及群众对留言的反应（赞成和反对）进行热度计算。

（6）按照热度降序排列得出“热点话题”。

2.2.2 模型说明

2.2.2.1 SinglePass 单遍聚类方法

SinglePass 算法在处理流数据时有较大的优势，虽然本问题处理的数据非流数据，但在不确定留言问题主题数的情况下，利用 SinglePass 聚类可以将相似的文本内容先归类作为参考。

SinglePass 算法的主要处理过程如（图 2-2-3-1）所示。按一定顺序读取数据，在初始时类的数量为 0，读取第一条文本内容时，新建一个类，并将该文本内容归入；接着逐条文本数据与已有数据进行比较，如果按照一定相似度计算的方法得出的相似度小于阈值，则归入该已有类，否则新建一个类，将其归入。流程如下：

（1）以第一个留言为种子，建立一个主题；

（2）将留言 A 文本向量化；

（3）将该留言 A 与已有的所有话题均做相似度计算，如余弦相似度。

（4）找出该留言 A 具有最大相似度的已有主题；

（5）若相似度值大于阈值，则把留言 A 加入到有最大相似度的主题中，跳转至 7；

（6）若相似度值小于阈值，则留言 A 不属于任一已有主题，创建一个新的主题类别，同时将当前文本归属到新创建的主题类别中；

（7）聚类结束，读取下一篇留言，重复操作。

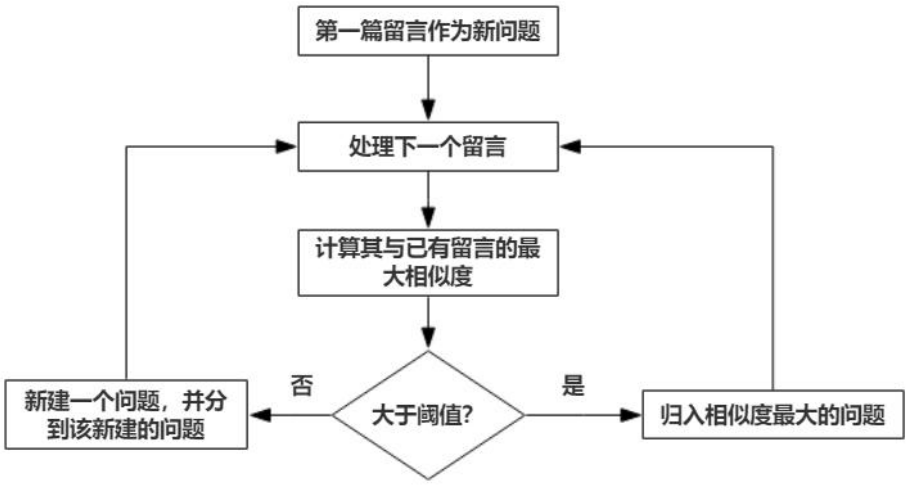

图 2

2.2.2.2 LDA 模型

Latent Dirichlet Allocation：主题模型，可以将文档集中每篇文档的主题按照概率分布的形式给出。同时它是一种无监督学习算法，在训练时不需要手工标注的训练集，需要的仅仅是文档集以及指定主题的数量 k 即可。

主题模型是一种典型的词袋模型，即它认为一篇文档是由一组词构成的集合，词与词之间没有顺序以及先后关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

通过 LDA 模型+gensim 对留言内容进行主题词划分。

2.2.3 热度评价指标及热点问题

为了挖掘热点问题，在聚类得到一个个类簇也就是一个个问题的基础上，我们还需要考虑附件 3 中群众留言的点赞数和反对数。群众对某一留言表示赞成说明其认同该留言反映的现象或者遇到类似的问题，群众对某一留言表示反对说明其反对该留言反映的问题或者认为该现象不实，因此我们定义一个问题反映量 ref(i),i 表示第 i 个问题，其计算公式为：

$$ref(i) = \text{问题内留言的总数} + \text{问题内留言的赞成总数} - \text{问题内留言的反对总数}$$

问题的热度主要通过留言数、点赞数、反对数及留言时间跨度进行评估，评估过程中，问题反映量越多，表示该问题受群众关注度越大，那么它的热度就越高。此外，时间因子也是影响问题热度的重要因素，一般来说，如果在问题反映量相近的情况下，如果某一问题的时间跨度越短，越能表示该问题是集中反映的，

表示它在一定时间内热度越高。因此我们定义一个问题的热度值 hot(i)，i 表示第 i 个问题，其计算公式为：

$$hot(i)=\frac{e^{ref(i)\bullet\frac{1}{\Delta t}\bullet\alpha}-e^{-ref(i)\bullet\frac{1}{\Delta t}\bullet\alpha}}{e^{ref(i)\bullet\frac{1}{\Delta t}\bullet\alpha}+e^{-ref(i)\bullet\frac{1}{\Delta t}\bullet\alpha}}$$

其中，hot(i)表示问题 i 的热度值，其值越大，该问题就越热，ref(i)表示问题 i 的问题反映量，△t 表示问题发表的时间跨度，本模型将同一问题中留言发表的最后时间与最初时间之差作为时间因子（△t），α为时间权值修正因子。

对于每个聚类得到的问题，计算其 hot 值，并按降序排列，排名越前的问题热度越高，越迫切需要政府解决。

2.3 问题 3 分析方法与过程

2.3.1 分析答复意见

从附件 4 中的答复意见列可以看出，大多数完整答复内容中包含留言网友的昵称，内容也和网友留言的内容相关度较高，所以考虑用文本相似度来评价答复意见的质量。将用户昵称、留言主题和详情结合作为搜索词，答复内容作为比对文本，比较关键词和文本之间的相似度。

2.3.2 gensim 使用流程

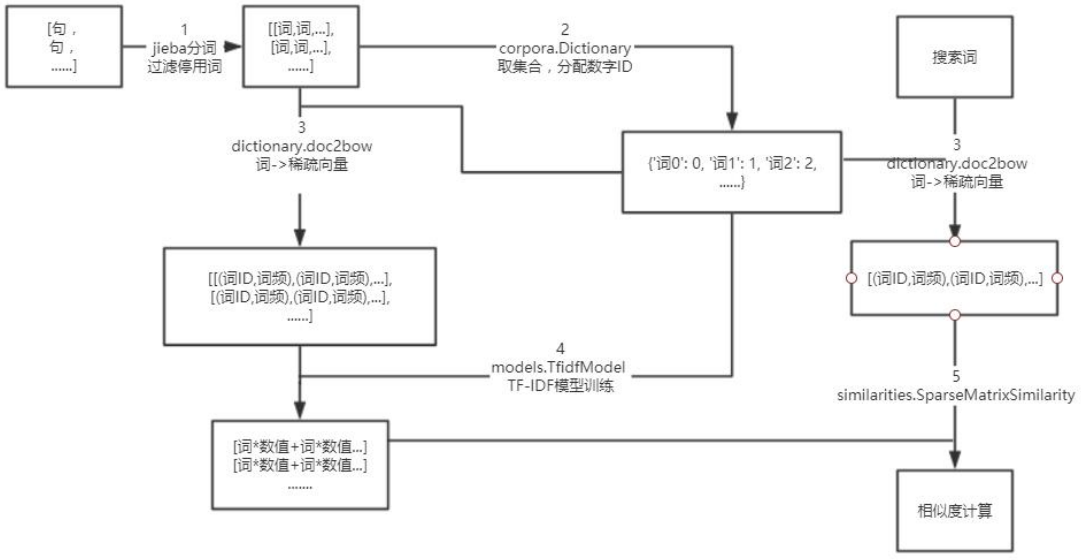

图 3

Gensim 提供了一个发现文档语义结构的工具，通过检查词出现的频率。Gensim 读取一段语料，输出一个向量，表示文档中的一个词。词向量可以用来训练各种分类器模型。

2.3.3 定义相似度函数

函数 calc_similarity 的两个形参分别是用户留言提取出的搜索词和答复内容的文本分词列表。通过 gensim 库比较两个文本的相似度并返回相似度值。

2.3.4 评价答复意见

首先提取附件 4 中的留言用户、留言主题、留言详情、答复意见，合并留言主题和留言详情，去除无用序列。遍历每组留言，将搜索词和答复文本进行如问题 1 中的分词、去停用词操作，调用 calc_similarity 函数，比较每组留言和答复之间的相似度并输出到文件 similarity.txt。

3、结果分析

3.1 问题 1 结果分析

3.1.1 词云图结果

根据各个标签生成的词云图，可以看出对于标签“交通运输”，关键词有“出租车”、“司机”、“快递”、“车”、“路”等；对于标签“劳动和社会保障”，关键词有“劳动者”、“社保”、“职工”、“工资”等；对于标签“卫生计生”，关键词有“治疗”、“医院”、“医生”、“生育”等；对于标签“商贸旅游”，关键词有“景区”、“旅游”、“经营”等；对于标签“城乡建设”，关键词有“小区”、“业主”、“生活”、“规划”、“开发商”等；对于标签“教育文体”，关键词有“学校”、“教师”、“学生”、“教育”、“招生”等；对于标签“环境保护”，关键词有“环评”、“污染”、“排放”、“环保局”等。

3.1.2 模型分析

通过训练互补朴素贝叶斯模型，并预测测试集样本。计算 F_Score 为 0.86。

```
卫生计生  卫生计生
劳动和社会保障    劳动和社会保障
教育文体  教育文体
劳动和社会保障    劳动和社会保障
商贸旅游  商贸旅游
f1_score: 0.8595982928347681
```

图 4

3.2 问题 2 结果分析

3.2.1 关键词提取及文本聚类分析

通过 LDA 模型统计语料库的问题关键词的共现频率和出现比例，如（图 5、图 6）所示，在问题反映关键词中，“经营”、“销售”、“车位”、“噪音”、“西地省”等问题地点、现象等关键信息词的出现比例较高。

```
(0, '0.017*"A7" + 0.012*"噪音" + 0.010*"城" + 0.010*"二期" + 0.009*"A6"')
(1, '0.019*"A3" + 0.012*"建设" + 0.008*"学院" + 0.007*"公园" + 0.007*"公司"')
(2, '0.030*"A7" + 0.012*"A3" + 0.010*"咨询" + 0.007*"县星沙" + 0.007*"街道"')
(3, '0.026*"A7" + 0.013*"A3" + 0.011*"滨河" + 0.010*"车位" + 0.010*"苑"')
(4, '0.020*"A7" + 0.016*"A3" + 0.013*"路" + 0.012*"A2" + 0.010*"投诉"')
(5, '0.028*"扰民" + 0.019*"栋" + 0.016*"A3" + 0.013*"噪音" + 0.012*"A7"')
(6, '0.022*"A7" + 0.012*"A9" + 0.011*"房屋" + 0.010*"有限公司" + 0.010*"解决"')
```

图 5

```
(0, '0.028*"A7" + 0.021*"建设" + 0.012*"加快" + 0.011*"社区" + 0.010*"A4" + 0.007*"居民" + 0.006*"经营"')
(1, '0.029*"A7" + 0.022*"A3" + 0.019*"扰民" + 0.012*"施工" + 0.010*"A2" + 0.010*"单元" + 0.009*"幼儿园"')
(2, '0.015*"A1" + 0.013*"销售" + 0.011*"滨河" + 0.011*"苑" + 0.011*"车位" + 0.010*"学生" + 0.010*"景园"')
(3, '0.020*"A3" + 0.014*"噪音" + 0.014*"扰民" + 0.013*"A7" + 0.012*"城" + 0.011*"栋" + 0.011*"A9"')
(4, '0.032*"A7" + 0.016*"A3" + 0.013*"路" + 0.012*"A2" + 0.009*"扰民" + 0.009*"A5" + 0.008*"大道"')
(5, '0.030*"A7" + 0.011*"A2" + 0.010*"A6" + 0.010*"安置" + 0.008*"A5" + 0.008*"拆迁" + 0.007*"影响"')
(6, '0.019*"西地省" + 0.011*"A3" + 0.010*"公司" + 0.010*"有限公司" + 0.009*"房屋" + 0.008*"拖欠" + 0.008*"A7"')
```

图 6

通过 SinglePass 聚类方法，依赖文本相似度对问题进行聚类，结果如（图 7）所示。在不同类簇中，如“车位捆绑销售”、“投资公司诈骗”、“丽发新城搅拌站扰民”、“油烟扰民”等问题类簇中数量相对较多。

由此可见在附件 3 的留言问题中，与上述关键词及关键问题类相近的留言出现比例较大，结合热度评价指标将留言数结合留言点赞数、反对数及书检查进行热度计算。

```
--------------------------------------------------------------------------------
【主题索引】:11
【主题声量】: 127
【主题关键词】:  内幕,雅苑,御景,活禽,霸王,员工,家园,世景,华庭,使用权
【主题中心句】 :
['投诉', '市伊', '景园', '滨河', '苑', '捆绑', '车位', '销售']
['投诉', '市伊', '景园', '滨河', '苑', '捆绑', '销售', '车位']
['投诉', '市伊', '景园', '滨河', '苑', '开发商', '违法', '捆绑', '销售', '产权', '车位']
--------------------------------------------------------------------------------
【主题索引】:24
【主题声量】: 122
【主题关键词】:  危房,事业,cnccd,货币,百姓,银盆岭,地址,风险,钱财,业主
【主题中心句】 :
['西地省', '展星', '投资', '有限公司', '涉嫌', '诈骗']
['西地省', '聚利人', '普惠', '投资', '有限公司', '涉嫌', '诈骗', '巨额', '资金']
['西地省', '惠普', '利聚人', '投资', '有限公司', '涉嫌', '诈骗', '巨额', '资金']
--------------------------------------------------------------------------------
【主题索引】:1
【主题声量】: 151
【主题关键词】:  合法,中海,国际,泥土,新村,公馆,荣盛,沥青,市场,云塘路
【主题中心句】 :
['丽发', '新城', '小区', '搅拌站', '噪音', '扰民']
['投诉', '丽发', '新城', '小区', '违建', '搅拌站', '噪音', '扰民']
['市丽发', '新城', '小区', '搅拌站', '噪音', '扰民', '污染环境']
--------------------------------------------------------------------------------
【主题索引】:13
【主题声量】: 121
【主题关键词】:  贷案,油烟,高源,饺子,故事,废气,餐饮店,兴旺,搅拌罐,沙岭
【主题中心句】 :
['A4', '四方', '坪', '双拥', '路口', '先福', '安置', '小区', '门面', '经营', '汽车', '钣', '喷', '业务', '油漆
', '气味', '重']
['A4', '四方', '坪', '怡然', '翠园', '退房', '钱', '不到']
['A4', '区广福园', '小区', '蓝鸟', 'KTV', '噪音', '扰民']
--------------------------------------------------------------------------------
【主题索引】:7
【主题声量】: 74
【主题关键词】:  司法,a8,西地省,民工,梅溪正,园林,卫生,实务,农华,答案
【主题中心句】 :
['西地省', '科技', '职业', '学院', '未经', '沟通', '强制', '学生', '搬离', '宿舍']
['西地省', '科技', '职业', '技术', '学院', '女生宿舍', '条件', '极差']
['商贸', '旅游', '职业', '技术', '学院', '强制', '学生', '实习']
```

图 7

3.2.2 热点问题分析

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	0.88	9/07/07至2019/09	A市伊景园滨河苑	A市伊景园滨河苑捆绑销售车位
2	2	0.67	9/11/13至2020/01	丽发新城小区	丽发新城小区附近违建搅拌站噪音扰民
3	3	0.51	9/01/02至2019/12	A市魅力之城小区	A市魅力之城油烟噪音扰民
4	4	0.49	9/01/17至2019/09	A7县	A7县星沙违建之风盛行无人管
5	5	0.33	9/01/08至2019/12	西地省聚利人普惠投资有限公司	地省聚利人普惠投资有限公司涉嫌诈骗巨额资

表 1

如表 1 所示，通过 $\mathrm{hot}(i)$ 计算不同类别的问题的热度值， 进行排序后得出附件 3 中群众在特定地点较集中反映的问题有：

A 市伊景园滨河苑捆绑销售车位；

A 市丽发新城附近建搅拌站噪音扰民；

3、A 市魅力之城油烟噪音扰民；

4、A7 县星沙违建之风盛行无人管；

5、西地省聚利人普惠投资有限公司涉嫌诈骗巨额资金。

热点问题的及时发现，政府部门可针对群众集中反映的现象进行及时处理，保障群众的民生问题得到快速解决。

3.3 问题 3 结果分析

根据输出的相似度文件如图 8，找到编号为 179893 的留言如图 9，可见答复意见并没有对留言内容做出相关解释。

179893	0.0
180537	0.29483384
181267	0.49999997
181603	0.37796447
184423	0.33667007
185799	0.40509573
185986	0.36313653

图 8

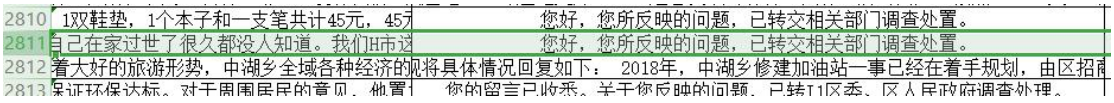

图 9

4、 参考文献

[1]Yan Leng,Weiwei Zhao,Chan Lin,Chengli Sun,Rongyan Wang,Qi Yuan,Dengwang Li. LDA-based data augmentation algorithm for acoustic scene classification[J]. Knowledge-Based Systems,2020,195.

[2]Changxuan Wan,Yun Peng,Keli Xiao,Xiping Liu,Tengjiao Jiang,Dexi Liu. An association-constrained LDA model for joint extraction of product aspects and opinions[J]. Information Sciences,2020,519.

[3]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.

[4]刘惠,赵海清.基于 TF-IDF 和 LDA 主题模型的电影短评文本情感分析——以《少年的你》为例[J].现代电影技术,2020(03):42-46.

[5]黄晨晨,索朗拉姆,拉姆卓嘎,群诺.基于 SVM 的藏文微博文本情感分析研究与实现[J].高原科学研究,2020,4(01):92-96.

[6]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.

[7]王洋.基于手机商品评论文本的情感分析与挖掘[J].企业科技与发

展,2019(05):130-132.
[8]孙昕. 基于文本挖掘对商品评论的分析[D].华中师范大学,2018.
[9]黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[J].软件工程,2020,23(03):1-3.
[10]郝淼,谭红,张成梅,于杰,黄伟.基于 TF-IDF 方法融合生物医学同义词的相似度计算方法[J].贵州科学,2019,37(06):91-96.
[11]姚佳奇,徐正国,燕继坤,熊钢,李智翔.基于标签语义相似的动态多标签文本分类算法[J/OL].计算机工程与应用:1-8[2020-05-08].http://kns.cnki.net/kcms/detail/11.2127.TP.20200324.2241.029.html.