

“智慧政务”中的文本挖掘应用

摘要：随着科学技术的发展，网络平台的应用也越来越广泛。人们开始通过各种网络问政平台来反映生活中的各种问题以来引起相关部门的关注与解决，那么运用建立基于自然语言处理技术的智慧政务系统对群众留言进行分类和对热点问题挖掘显得尤为重要。本文将基于数据挖掘技术对三个附件（附件 2 附件 3 附件 4）的留言信息数据进行内在的信息挖掘，提取我们需要进行分析的部分进行深度挖掘与分析。

针对问题一：本文首先将附件 2 中的非结构化数据进行去重去空、中文分词及停用词过滤等数据预处理，然后将一级分类标签转为与之对应的 ID 号，为建立关于留言内容的一级标签分类模型做准备，最后利用 lstm 建模，并绘制该模型的损失函数与准确度趋势图，利用 F 值评价该模型。

针对问题二：第一步对附件 3 中文本数据进行去空去重，jieba 分词及去除停用词等数据预处理，转化成结构化数据。第二步利用 matplotlib, WordCloud 做数据可视化，绘制关于留言主题的词云图。第三步用 K-means 进行层次聚类，制定热度指标，进行热点问题分类与计算热度指数，实现热度排名。第四步利用实体命名方法提取留言主题中的人群和地点，再抽取时间范围，得到关于热点问题的详细信息，存放结果于附件中的热点问题表及热点问题留言明细表。

针对问题三：对附件三中 2816 条相关部门留言答复意见从相关性，完整性的角度作出评价。

关键词：Lstm 建模 jieba K-means 文本聚类 实体命名法 余弦相似度算法

Abstract:With the development of science and technology, the application of network platform is becoming more and more extensive. Since people began to reflect various problems in life through various network political platforms, it has attracted the attention and solution of relevant departments, so it is very important to use the intelligent government system based on natural language processing technology to classify the mass messages and excavate the hot issues. In this paper, based on the data mining technology, the message information data of three attachments (annex 2, annex 3, annex 4) are internally mined, and the parts we need to analyze are extracted for deep mining and analysis.

To solve the problem one, this paper preprocesses the unstructured data in annex 2, such as de-empty, Chinese word segmentation and disuse word filtering, then converts the first-level classification label into the corresponding ID number, prepares for the establishment of the first-level label classification model about the message content, finally uses the lstm modeling, and draws the loss function and accuracy trend map of the model, and evaluates the model by using the F-score.

Aiming at the problem two, the first step is to preprocess the Chinese text data in Annex 3, jieba the word segmentation and deactivating words, and them into structured data. A second step is to use the matplotlib, WordCloud to do data visualization and draw a word cloud map about the topic of the message. And the third step is to use the K-means to cluster the heat level, make the heat index, classify and calculate the heat index to achieve the heat ranking. The fourth step uses the entity naming method to extract the crowd and place in the message topic, then extracts the time range, obtains the detailed information about the hot spot question, stores the result in the attachment hot spot question list and the hot spot question message list.

In response to question 3, the comments of 2816 relevant departments in Annex III are evaluated in terms of relevance, completeness.

Key words: Lstm modeling; jieba; K-means text clustering; entity naming; cosine similarity algorithm

目录

1. 挖掘背景与目标.....	5
1.1 挖掘背景.....	5
1.2 挖掘目标.....	5
2. 问题一分析与解决过程.....	5
2.1 问题一分析.....	5
2.2 问题一的解决过程.....	6
2.2.1 流程图.....	6
2.2.2 一级标签处理.....	7
2.2.3 Lstm 建模.....	9
2.2.4 F 值评价分类模型.....	11
3. 问题二分析与解决过程.....	12
3.1 问题二分析.....	12
3.2 问题二的解决过程.....	12
3.2.1 流程图.....	12
3.2.2 具体解决过程.....	12
3.3.2.1 数据预处理.....	12
3.3.2.2 聚类分析.....	14
3.2.2.3 提取地点，人群，留言时间.....	17
3.2.2.4 热度评价.....	18
3.2.2.5 留言问题明细.....	18
4. 问题三分析与解决过程.....	19
4.1 问题三分析.....	19
4.2 问题三解决过程.....	19
4.2.1 流程图.....	19
4.2.2 解决过程.....	19
4.2.2.1 numpy 应用.....	19
4.2.2.2 相关性.....	20

4.2.2.3 完整性.....	21
4.2.2.4 评价汇总.....	22
5. 参考文献.....	23

1. 挖掘背景与目标

1.1 挖掘背景

随着科学技术的发展，各种网络平台的应用也越来越广泛。近年来，微信、微博、市长信箱、阳光热线等网络问政平台就逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，与以往依靠拨打热线，上交匿名信等方式才能反映群众生活中的问题与想法相比，人们通过留言的方式来反映会更加方便，快捷。但随之而来的问题也出现了：由于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，工作量大，差错率高，效率低等问题亟需解决。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

（1）根据网络问政平台上的群众留言（附件 2）的一级标签分类信息构建分类模型，并对该分类方法进行 F 值评价。以来初步解决大部分电子政务系统还是依靠人工根据经验处理留言分类，存在工作量大、效率低，且差错率高等问题。

（2）对一段时间内集中爆发的，多人反映的热点问题归类与热度评价，以便相关部门更快速地发现问题，有针对性地解决问题，提高服务民众的效率。

（3）针对附件 4 相关部门对群众留言的答复意见，从答复的相关性、完整性、可解释性等多角度对答复意见的质量给出一套评价方案，为相关部门对群众留言答复意见需在哪些方面进行修改提供理论依据。

2. 问题一分析与解决过程

2.1 问题一分析

充分利用附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

问题一旨在根据附件二给出的 9210 条已经按照一级标签分类的群众留言，分析这些数据的留言详情，找到每条留言被分到对应一级标签的原因，并以此构建关于留言内容的一级标签分类模型，并对该模型进行 F 值评价。

2.2 问题一的解决过程

2.2.1 流程图

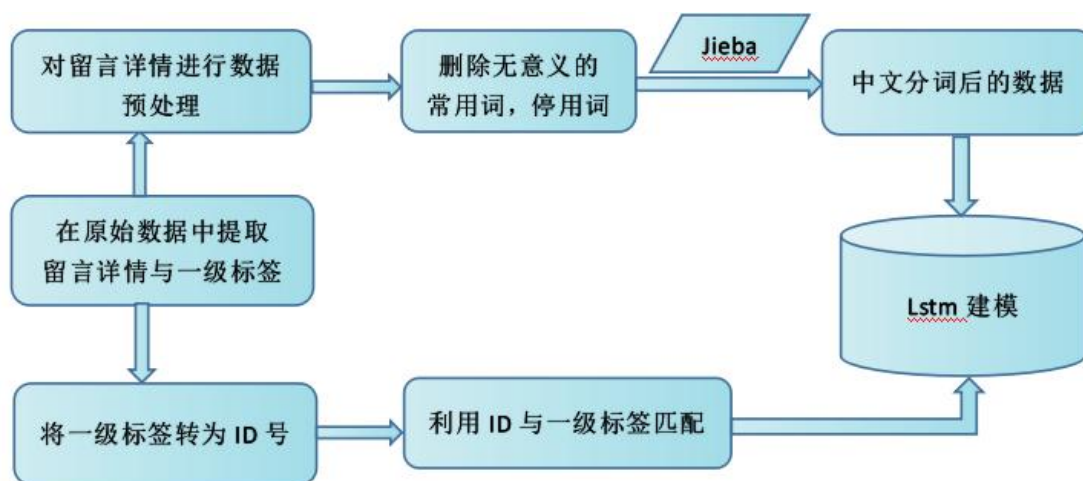


图 1 问题一流程图

2.2.2 具体解决过程

2.2.2.1 数据预处理

(1) 提取关键列

问题一主要目的在于根据附件 2 已经按一级标签分好的群众留言建立分类模型，附件 2 的数据中要用到的就是留言详情与一级标签这两列，则将附件二中的这两列关键列提取出来，同时对附件 2 中总的的数据数进行统计，一共 9210 条留言数据需要分析。

(2) 留言详情去空，去重

留言详情中的空数据与重复数据会影响数据的分析与模型建立，所以将留言详情中的空数据与重复数据去除，得到有 0 个空值，0 个重复值。

(3) 过滤停用词

为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，如：的，了，啊等，这些字或词即被称为 Stop Words（停用词）。停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。在此我们需要对留言详情中对分类过程中无意义，加上还可能影响分类结果的停用词去掉，提高分类结果的准确性。

为防止一些无用的符号没有在停用词表中，我们还自定义了删除除字母，数字，汉字以外的符号的函数，进行二次过滤。得到的新的留言详情为 clean_message。

(4) 对留言详情进行中文分词

使用 jieba 词库进行分词，分词后各词语间由空格隔开

在对留言详情进行挖掘分析之前，先要把非结构化的文本数据转换为计算机能够识别的结构化信息。也为后续的向量表达做准备。在附件 2 中，留言详情以中文文本的方式给出了数据。为了便于转换，先要对这些数据进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)，同时采用了动态规划 查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于 汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。进行分词后的留言详情定义为 cut_message。此时对留言详情的数据预处理完成。

2.2.2.2 一级标签处理

(1) 将留言详情按照一级分类标签分类

将留言详情按照所对应的一级标签分类，得到附件 2 中的群众留言涉及 7 个附件 1 中的一级标签，并统计每个一级标签中包含的留言数。

	一级标签	留言总数
0	城乡建设	2009
1	劳动和社会保障	1969
2	教育文体	1589
3	商贸旅游	1215
4	环境保护	938
5	卫生计生	877
6	交通运输	613

表 1 各一级标签留言数

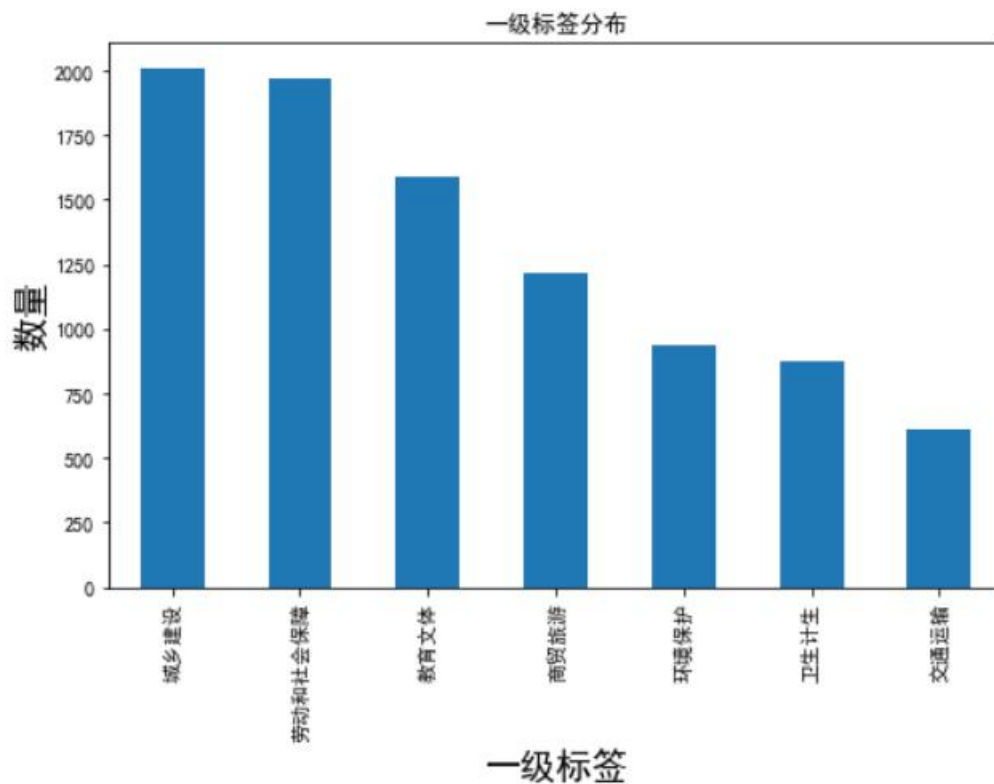


图 2 各一级标签留言数

(2) 将一级标签转为 ID 号

一级标签都是文本内容为非结构化数据，不利于计算机识别，则将一级标签转化为结构化数据 ID 号，为以后的分类模型的训练做准备。

	一级标签	cat_id
0	城乡建设	1
1	劳动和社会保障	2
2	教育文体	3
3	商贸旅游	4
4	环境保护	5
5	卫生计生	6
6	交通运输	7

表 2 各一级标签对应的 ID

2.2.3 lstm 建模

该模型原理及步骤图如下图所示

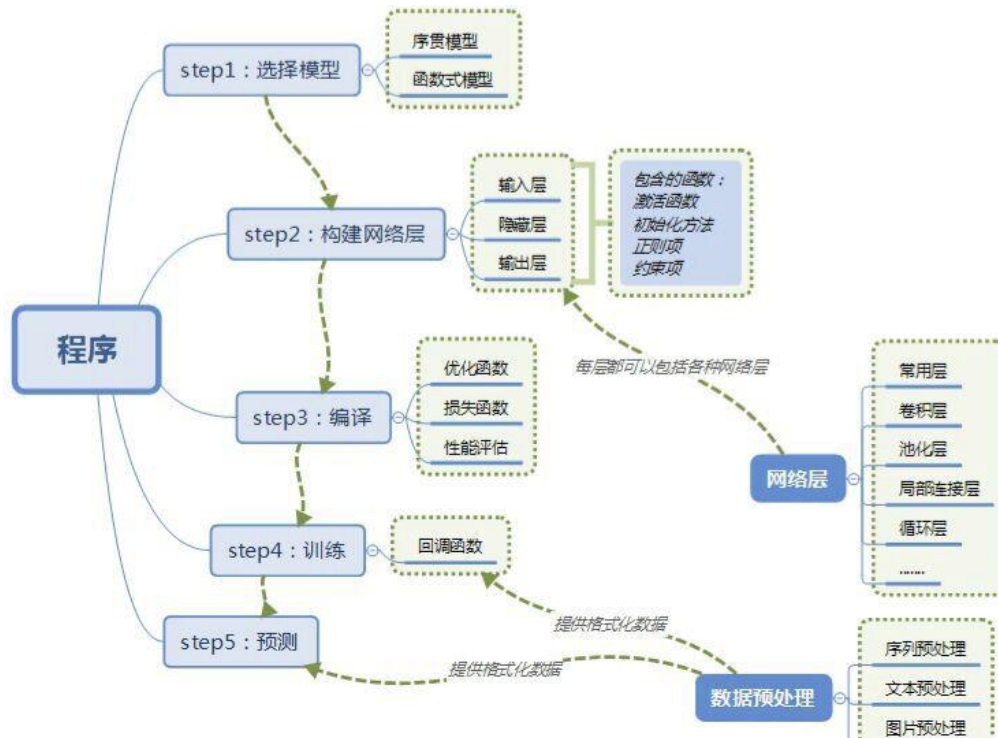


图 3 lstm 建模

具体实现步骤:

- (1) 将 cut_message 进行向量化处理, 将 cut_message 中每条数据转化成整数序列的向量
- (2) 设置最频繁使用的 50000 个词
- (3) 设置 cut_message 每条数据最大的词语数为 250 个, 对于超过的会截去, 不足的将会补 0
- (4) 将附件 2 拆分成训练集和测试集
- (5) 定义 lstm 的序列模型
 - ① Embedding 嵌入层 用长度为 70 的向量来表示每一个词语
 - ② SpatialDropout1D 层 在训练中每次更新时将输入的单元按比例随机设置成 0.2, 这样是为了防止过拟合。
 - ③ Lstm 层, 包含 70 个记忆单元
 - ④ 输出层 包含 7 个分类的全链接层
- (6) 由于是多分类, 将激活函数设置为 softmax ()
- (7) 又由于是多分类, 损失函数为分类交叉熵 categorical_crossentropy
- (8) Lstm 模型定义成功, 开始训练数据, 设置 5 个训练周期, 且 batch_size 为 64

(9) 得到损失函数趋势图与准确度趋势图。

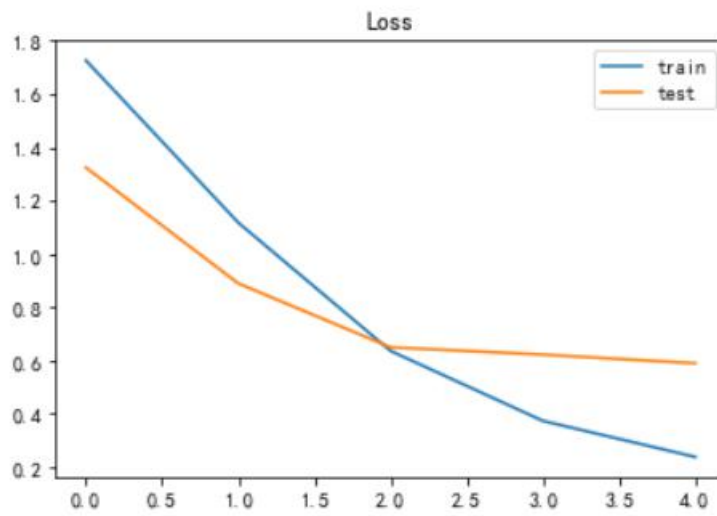


图 4 损失函数趋势图

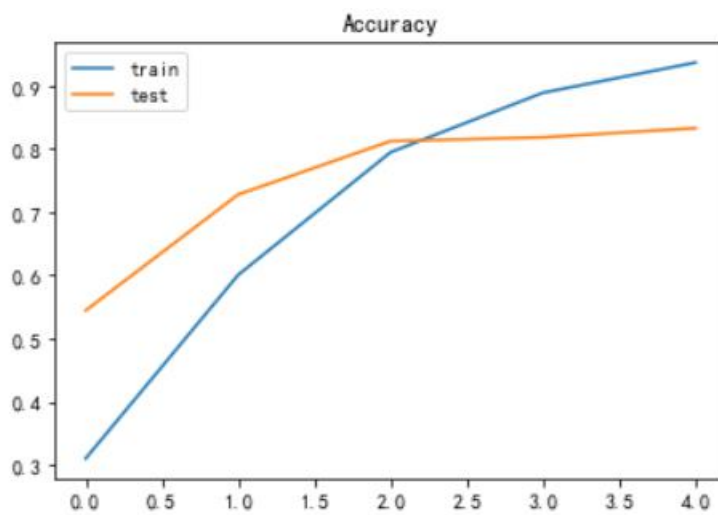


图 5 准确度趋势图

(10) 对测试集进行预测，评价预测效果，计算混淆矩阵，混淆矩阵可视化

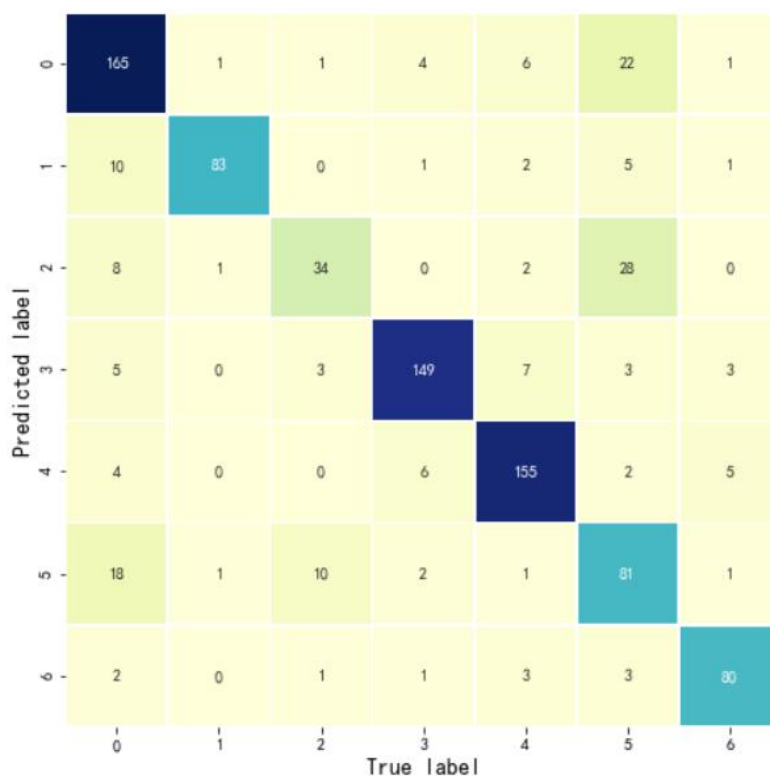


图 6 混淆矩阵

2.2.4 F 值评价分类模型

关于留言内容的一级分类模型已经建立完成，需要对该留言模型做出评价

通常使用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率

编写程序实现 F 值的评价，评价结果如下

	precision	recall	f1-score	support
0	0.82	0.78	0.80	212
1	0.81	0.97	0.88	86
2	0.47	0.69	0.56	49
3	0.88	0.91	0.89	163
4	0.90	0.88	0.89	176
5	0.71	0.56	0.63	144
6	0.89	0.88	0.88	91
accuracy			0.81	921
macro avg	0.78	0.81	0.79	921
weighted avg	0.82	0.81	0.81	921

图 7 F 值评价结果

3. 问题二分析与解决过程

3.1 问题二分析

充分利用附件三中的留言信息，挖掘热点问题，进行热度排名。

问题二旨在根据附件三中的 4323 条留言数据，挖掘出其中一段时间内集中爆发的，多人反映的热点问题并进行归类，制定合理的热度评价指标对热点问题计算热度指数，并以此为依据进行热度排名，并将群众反映的排名前五的热点问题的相关信息挖掘出来。

3.2 问题二的解决过程

3.2.1 流程图

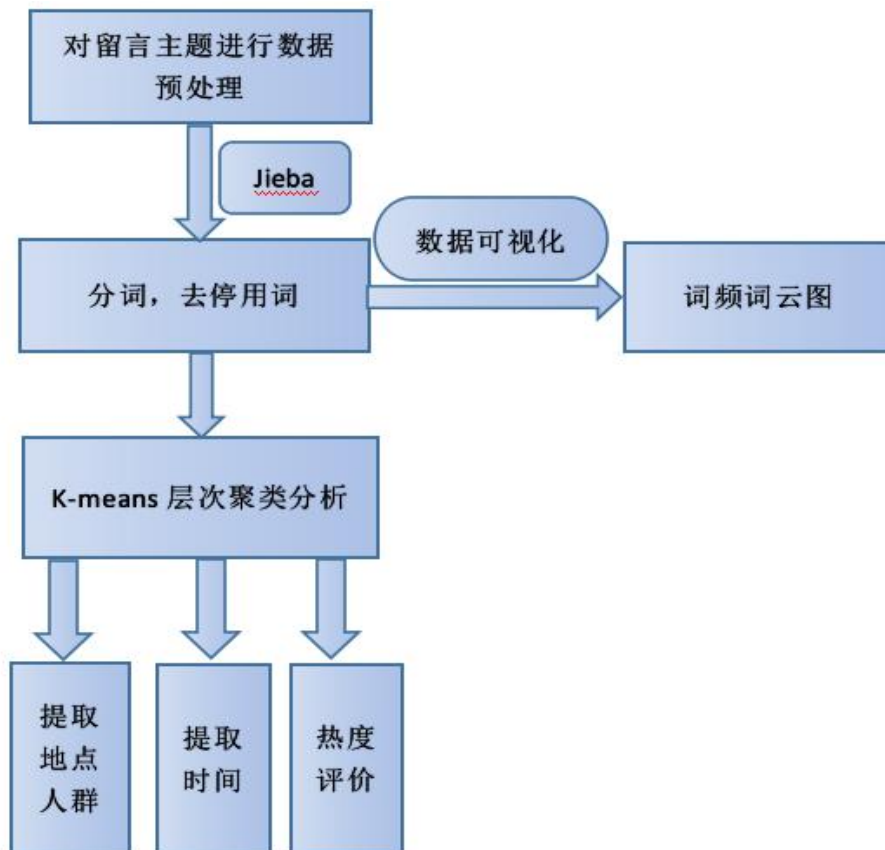


图 8 问题二流程图

3.2.2 具体解决过程

3.2.2.1 数据预处理

（1）文本去空去重

由于群众的留言信息数据较多，反馈情况种类复杂多样，则可能出现空值或重复值，若数据出现空值或重复值情况，则会导致后续进行聚类分析归类时出现偏差，进而影响热度计算的精度，所以先对数据进行去空、去重。

（2）中文分词

由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要分词，本文采用 Python 开发的一个中文分词模块——jieba 分词，对附件 3 中每一个留言主题描述进行中文分词。部分分词结果示例下图：

留言主题
[A3区, 一米阳光, 婚纱, 艺术摄影, 是否, 合法, 纳税, 了, ?]
[咨询, A6区, 道路, 命名, 规划, 初步, 成果, 公示, 和, 城乡, 门牌, 问题]
[反映, A7县, 春华, 镇金鼎村, 水泥路, , , 自来水, 到户, 的, 问题]
[A2区, 黄兴路, 步行街, 大, 古道, 巷, 住户, 卫生间, 粪便, 外排]
[A市, A3区, 中海, 国际, 社区, 三期, 与, 四期, 中间, 空地, 夜间, 施...
...
[A市, 经济, 学院, 寒假, 过年, 期间, 组织, 学生, 去, 工厂, 工作]
[A市, 经济, 学院, 组织, 学生, 外出, 打工, 合理, 吗, ?]
[A市, 经济, 学院, 强制, 学生, 实习]
[A市, 经济, 学院, 强制, 学生, 外出, 实习]
[A市, 经济, 学院, 体育, 学院, 变相, 强制, 实习]

图 9 分词结果

留言主题
[一米阳光, 婚纱, 艺术摄影, 合法, 纳税]
[咨询, 道路, 命名, 规划, 初步, 成果, 公示, 城乡, 门牌]
[春华, 镇金鼎村, 水泥路, 自来水, 到户]
[黄兴路, 步行街, 古道, 巷, 住户, 卫生间, 粪便, 外排]
[中海, 国际, 社区, 三期, 四期, 空地, 夜间, 施工, 噪音, 扰民]
...
[经济, 学院, 寒假, 过年, 期间, 组织, 学生, 工厂, 工作]
[经济, 学院, 组织, 学生, 外出, 打工]
[经济, 学院, 强制, 学生, 实习]
[经济, 学院, 强制, 学生, 外出, 实习]
[经济, 学院, 体育, 学院, 变相, 强制, 实习]

图 10 分词结果

（3）过滤停用词

上述分词结果是没有停用词过滤的结果（如图 9 所示），可以看到，其中有大量标点及表达无意义的字词，对后续分析会造成很大影响，因此接下来需要进行停用词过滤（过滤结果如图 10 所示）。

（4）制作词云图

词云图：由词汇组成类似云的彩色图形。“词云”就是对文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信

息，使浏览网页者只要一眼扫过文本就可以领略文本的主旨。

由高到低统计出文本中每个词出现的频数，利用 matplotlib 做数据可视化，运用 WordCloud 进行云图绘制（如图 11 所示）。根据云词图可以了解留言内容关键信息，以检测与后续热度评价判断是否一致。



图 11 词云图

3.2.2.2 聚类分析

聚类分析是一种无监督机器学习（训练样本的标记信息是未知的）算法，它的目标是将相似的对象归到同一个簇中，将不相似的对象归到不同的簇中。

（1）相似度计算

欧氏距离：

欧氏距离是一种常用的距离定义，指在 m 维空间中两个点之间的真实距离，对多维向量

$A=(A_1, A_2, \dots, A_n)$ ， $B=(B_1, B_2, \dots, B_n)$ ，欧氏距离的计算公式如下：

$$dist(A,B) = \sqrt{\sum_{i=1}^n (A_i - B_i)^2}$$

余弦相似度：

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体差异的大小。相比欧氏距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上的差异。余弦值的计算公式如下：

$$\cos \theta = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}}$$

相对于欧氏距离，余弦相似度更适合计算文本的相似度。首先将文本转换为权值向量，通过计算两个向量的夹角余弦值，就可以评估他们的相似度。余弦值的范围在 $[-1, 1]$ 之间，值越趋近于1，代表两个向量方向越接近；越趋近于-1，代表他们的方向越相反。为了方便聚类分析，我们将余弦值做归一化处理，将其转换到 $[0, 1]$ 之间，并且值越小距离越近。

(2) 性能度量

在选择聚类算法之前，首先来了解什么样的聚类结果是比较好的。我们希望同一个簇内的样本尽可能相似，不同簇的样本尽可能不同，也就是说聚类结果的“簇内相似度”高且“簇间相似度”低。

考虑聚类结果的簇划分 $\mathcal{C} = \{C_1, C_2, \dots, C_K\}$ ，定义：

$$avg(C) = \frac{2}{(|C|(|C| - 1))} \sum_{1 \leq i < j \leq |C|} dist(x_i, x_j)$$

$$diamC = \max_{1 \leq i < j \leq |C|} dist(x_i, x_j)$$

$$d_{min}(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} dist(x_i, x_j)$$

$$d_{cen}(C_i, C_j) = dist(\mu_i, \mu_j)$$

其中， μ_i 代表簇的中心点； $avg(C)$ 代表簇内样本的平均距离； $diamC$ 代表簇内样本间的最远距离； $d_{min}(C_i, C_j)$ 对应于簇 C_i 和 C_j 簇最近样本间的距离； $d_{cen}(C_i, C_j)$ 对应于簇 C_i 和 C_j 中心点间的距离。基于以上公式可导出下面两个常用的聚类性能度量内部指标：

DB 指数 (Davies-Bouldin Index, 简称 DBI)

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{avg(C_i) + avg(C_j)}{d_{cen}(C_i, C_j)} \right)$$

Dunn 指数 (Dunn Index, 简称 DI)

$$DI = \min_{1 \leq i \leq k} \left\{ \min_{j \neq i} \left(\frac{d_{min}(C_i, C_j)}{\max_{1 \leq l \leq k} diam(C_l)} \right) \right\}$$

DB 指数的计算方法是任意两个簇内样本的平均距离之和除以两个簇的中心点距离，并取最大值，DBI 的值越小，意味着簇内距离越小，同时簇间的距离越大；Dunn 指数的计算方法是任意两个簇的最近样本间的距离除以簇内样本的最远距离的最大值，并取最小值，DI 的值越大，意味着簇间距离大而簇内距离小。因此，DBI 的值越小，同时 DI 的值越大，意味着聚类的效果越好。

(3) 构建词袋模型

文本被切分成单词后，需要进一步转换成向量。先将所有文本中的词汇构建成一个词条列表，其中不含重复的词条。然后对每个文本，构建一个向量，向量的维度与词条列表的维度相同，向量的值是词条列表中每个词条在该文本中出现的次数，这种模型叫做词袋模型。

(4) 权值转换

TF-IDF 是一种统计方法，用来评估一个词条对于一个文件集中一份文件的重要程度。TF-IDF 的主要思想是：如果某个词在一篇文章中出现的频率 TF 高，并且在其他文件中很少出现，则认为此词条具有很好的类别区分能力，适合用来分类。将词袋向量转换为 TF-IDF 权值向量，更有利于判断两个文本的相似性。

TF(词频, term frequency):

$$tf_{i,j} = \frac{n_{i,k}}{\sum_k n_{k,j}}$$

分子是词条 t_i 在文件 d_j 中出现的次数，分母是文件中所有词条出现的次数之和。

IDF(逆向文件频率, inverse document frequency):

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

对数内的分子是文件总数，分母是包含词条的文件数，如果该词不存在，就会导致分母为零，因此一般使用 $1 + |\{j : t_i \in d_j\}|$ 作为分母。

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

(5) 使用层次聚类算法

层次聚类试图在不同的层次对数据集进行划分，可以采用“自底向上”的聚类策略，也可以采用“自顶向下”的分拆策略。一般采用“自底向上”的策略，它的思路是先将数据集

中的每个样本看作一个初始聚类簇，然后找出两个聚类最近的两个簇进行合并，不断重复该步骤，直到达到预设的聚类个数或某种条件。关键是如何计算两个簇之间的距离，每个簇都是一个集合，因此需要计算集合的某种距离即可。例如，给定簇 C_i 和 C_j ，可通过以下 3 种方式计算距离：

最小距离：

$$d_{\min}(C_i, C_j) = \min_{x \in C_i, z \in C_j} \text{dist}(x, z)$$

最大距离：

$$d_{\max}(C_i, C_j) = \max_{x \in C_i, z \in C_j} \text{dist}(x, z)$$

平均距离：

$$d_{\text{avg}}(C_i, C_j) = \frac{1}{|C_i| |C_j|} \sum_{x \in C_i} \sum_{z \in C_j} \text{dist}(x, z)$$

最小距离由两个簇的最近样本决定，最大距离由两个簇的最远样本决定，平均距离由两个簇的所有样本决定。

接下来要考虑如何确定一个合适的聚类个数或某种结束条件，具体思路是：

- (1) 选定一部分测试样本，对其进行层次聚类分析。
- (2) 记算性能度量指标 DBI 和 DI 的变化趋势，结合人工校验，得到一个合适的聚类个数和对应的距离阈值。
- (3) 将此距离阈值作为聚类结束的条件，对所有样本做聚类分析。此时无需再计算 DBI 和 DI 值，计算效率可以大幅提升。

3.2.2.3 提取地点人群，留言时间

(1) 提取地点人群

提取特定的地点和人群，采用命名实体识别。命名实体识别 (Named Entity Recognition, 简称 NER)，又称作“专名识别”，是自然语言处理中的一项基础任务，应用范围非常广泛。命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括 人名、地名、机构名、日期时间、专有名词等。通常包括两部分：实体的边界识别、确定实体的类型 (人名、地名、机构名或其他)。NER 所涉及的命名实体一般包括 3 大类 (实体类，时间类，数字类) 和 7 小类 (人名、地名、组织机构名、时间、日期、货币、百分比)。

(2) 提取时间

在聚类分析完成以后，对归类的每一簇中查找时间中最大最小值，得到的结果就是在每一类的问题的时间范围。

3.2.2.4 热度评价

由于群众反映的问题多种多样，聚类分析后得到的归类很多，为反映群众留言热点问题，对每一类问题的留言信息数做统计。统计结果由数值从高到小排列，最后抓取排名靠前的五类问题生成图中示例表格。

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	1.26%	2019/01/02至2019/12/27	A市A2区丽发新城	小区附近违建搅拌站噪音扰民和污染环境，危害健康
2	2	0.47%	2019/7/21至2019/09/25	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民，污染环境
3	3	0.28%	2019/2/21至2019/12/28	A市A7县星沙凉塘路	旧城区何时可以开始动工问题
4	4	0.19%	2019/04/28至2019/11/22	A市经济学院学生	学校强制学生外出实习
5	5	0.12%	2019/07/24至2019/11/24	A市A3区中海国际社区	小区夜晚施工，打扰休息

表 3 热度评价表

3.2.2.5 留言问题明细

返回上明面 3.2.2.4 中所抓取的热度排名前五的问题的类，在附件 3 中匹配每类问题在原表中所对应的数据，提取数据，按照热度排名将所有数据信息汇总。

问题ID	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
1	188809	A909139	丽发新城居民区附近	2019/11/19 18:07:54	建搅拌站，	0	1
1	189950	A909204	发新城附近建搅拌	2019-11-13 11:20:21	搅拌站。可	0	0
1	190108	A909240	新城小区旁边建搅	2019-12-21 15:11:29	千名学生	0	1
1	190523	A00072847	建搅拌站，彻夜施	2019/12/26 13:55:15	搅拌站几	0	0
1	190802	A00072636	区建搅拌站，噪音	2019/11/25 18:58:05	尘，严重	0	0
1	199379	A00092242	附近修建搅拌厂，	2019/11/25 10:17:56	得了疾病住	0	0
1	203393	A00053065	建设混凝土搅拌站，	2019/11/19 14:51:53	尘，严重	0	2
1	208285	A909205	小区附近搅拌站噪音	2019-12-15 12:32:11	是有很大噪	0	24
1	208714	A00042015	近修建搅拌站，污	2020-01-02 00:00:00	环境质量	0	4
1	213464	A909233	成小区附近违建搅	2019-12-10 12:34:21	。该搅拌	0	0
1	213930	A909218	近违规乱建混凝土	2019-12-27 23:34:32	呼吁政府和	0	0
1	214282	A909209	区附近搅拌站噪音	2020-01-25 09:07:21	烦死了不	0	0
1	215563	A909231	小区旁边的搅拌厂	2019-12-06 12:21:32	音和灰尘。	0	0
1	215842	A909210	丽发新城小区附近	2020-01-26 19:47:11	是怎么回事	0	0
1	216824	A909214	石料噪音污水影响	2019-12-25 12:15:57	严重扰民的	0	0
1	217700	A909239	小区旁的搅拌站严重	2019-12-21 02:33:21	到丽发新	0	1
1	222831	A909228	的A2区丽发新城附近	2019-12-22 10:23:11	改产生大	0	0
1	225217	A909223	成附近修建搅拌厂严	2019-11-15 09:17:36	拌站，每天	0	0
1	231136	A909204	发新城附近建搅拌	2019-12-02 11:20:21	上次投诉已	0	0
1	233158	A909242	小区旁建搅拌厂严	2019-12-05 08:46:20	我还能忍	0	0
1	234327	A909212	声不断的丽发新城小	2019-12-26 21:44:13	产生大量	0	0
1	235362	A909215	小区附近水泥搅拌	2020-01-06 20:45:34	重危害居民	0	0
1	238212	A909203	小区附近建搅拌站	2019-12-12 10:23:11	民区应是一	0	0
1	239336	A909213	发新城小区遭搅拌	2019-12-11 11:44:11	，离居民	0	0

表 4 热点问题明细表（部分）

4. 问题三分析与解决过程

4.1 问题三分析

充分分析附件四中相关部门对群众留言的答复意见，多角度地对答复进行分析。

分析附件四相关部门在回复群众留言时有哪些方面的优点与不足，从多个角度对答复意见进行分析，给予相关部门对群众留言的答复意见的修改意见。

4.2 问题三解决过程

4.2.1 流程图

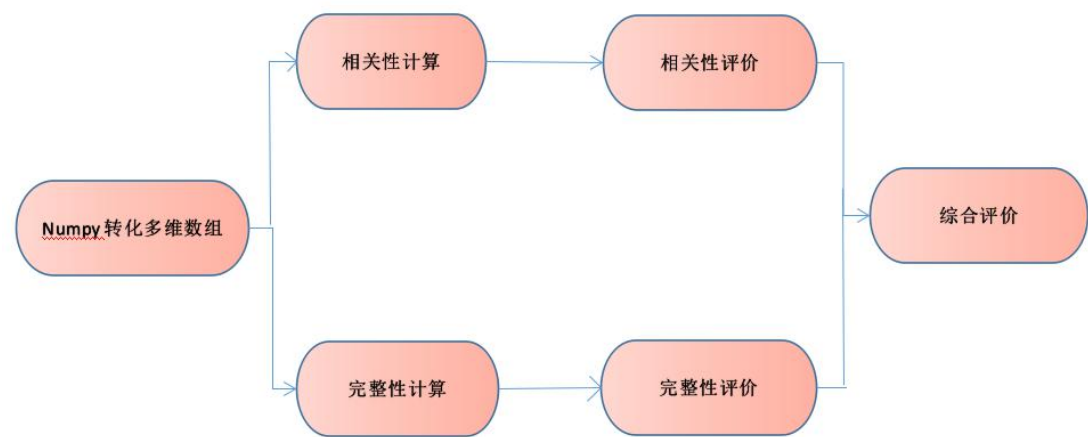


图 12 问题三流程图

4.2.2 解决过程

本文从相关性，完整性的角度对相关部门的留言答复意见进行评价，并予以实现

4.2.2.1 numpy 应用

NumPy 的主要对象是同质多维数组（ndarray，是 N-dimensional array 的缩写）。它是一个元素表（通常是数字），元素数量是事先准备好的，所有元素类型都相同，由非负整数元组索引。数组的维度和元素数量由数组的型（shape）来确定 shape 由 n 个正整数组成的元组来指定，元组的每个元素对应每个维度数组的大小，元素数量为元组的元素积。在 NumPy 中维度称为轴（axis），轴的数量称为秩（rank）。另外，ndarray 的大小是不可改变的。

利用 numpy，将附件 4 中的表格内容转为多维数组，那么 xlsx 文件中表格内的每一个值，都成为数组中的元素，方便后续对表中的值进行导出计算和插入评价后的结果。

4.2.2.2 相关性

从相关部门对群众留言的答复意见与群众留言内容是否相似度高,即相关性高来评价该回复。采用的主要方法为余弦相似度算法。

该方法原理为:通过测量两个向量之间的角的余弦值来度量它们之间的相似性。0角度的余弦值是1,而其他任何角度的余弦值都不大于1,并且最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向,所以,可以用来判断两个文本的相似度。


具体实现过程如下:程序见附件

- (1) 先给出两个要比较相似度的文本
- (2) 对文本数据进行 jieba 分词
- (3) 进行词频向量化:词频向量化函数,输入一个总字典和待向量化列表
- (4) 计算两个向量的余弦相似度,公式如下:

$$\cos\theta = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

- (5) 得出具体值(所有数据值均在 0-1),导入到 Excel 文件中用于后边的评价步骤

对附件 4 中的留言详情和答复意见做相关性计算,得到相关值数据后(部分数据值如图所示)储存在数组中。



```
Out[140]: [0.419617089869402,
0.3036656436742144,
0.31415288203945857,
0.11307783213313498,
0.19909289367330493,
0.18884542621819295,
0.4182594486952485,
0.39522103362650735,
0.26854604494407425,
0.49951860608847637,
0.21011515974645167,
0.1523531986874111,
0.3603393379160534,
0.3999317888467846,
0.18749437278670697,
0.517106909133079,
0.13031507602217382,
0.1442041755757817,
0.11776392994149894,
0.149922979150647909]
```

图 13 部分相关性数值

定义评价指标,所有的相关性分析后的值在[0,1]之间,以 0.2 为一个梯度,划分为“一星”“二星”“三星”...五类评价指标,将相关性分析后的结果按此类别判断,得到每一个数据相对应的评价,用 append 将每一个结果都储存在同一个数组中(部分结果如图所示)

```
Out[144]: ['三星',
           '二星',
           '二星',
           '一星',
           '一星',
           '一星',
           '三星',
           '二星',
           '二星',
           '三星',
           '二星',
           '一星',
           '二星',
           '二星',
           '一星',
           '三星',
           '一星',
           '一星',
           '一星',
           '一星']
```

图 14 部分数据相关性星级评价

4.2.2.3 完整性

对回复意见的完整性判断，判断回复意见是否完整，对这个问题我们从回复意见的文本长度进行判断。利用 len 函数统计每一个回复意见的文本长度，并将得到的每一个值都按文本顺序储存在同一个数组中（部分数据如同所示）

```
Out[141]: [454,
           305,
           357,
           310,
           161,
           232,
           245,
           624,
           505,
           224,
           489,
           427,
           139,
           101,
           210,
           651,
           145,
           141,
           192,
           122]
```

图 15 部分答复意见文本长度

定义评价指标，在统计完每个文本的长度以后对数据结果进行评价。评价指标定义为:长度 [0, 200] “较差”、[200, 300] “一般”、[300, 400] “良好”、[400, 500] “完整”、500 以上 “优秀”。数据相对应的评价，用 append 将每一个结果都储存在同一个数组中（部分结果如图所示）

```
Out[145]: ['完整',
           '良好',
           '良好',
           '良好',
           '较差',
           '一般',
           '一般',
           '优秀',
           '优秀',
           '一般',
           '完整',
           '完整',
           '较差',
           '较差',
           '一般',
           '优秀',
           '较差',
           '较差',
           '较差',
           '较差']
```

图 16 完整性评价结果

4.2.2.4 评价汇总

上述结果中得到的相关性数据，相关性评价，完整性数据，完整性评价四个数组用 `numpy.insert` 插入到原数据中，此时所有数据储存在一个二维数组中，最后用 `pandas.DataFrame` 将数组转为原表格形式（部分数据如图所示）

长度	完整性	相关值	相关性	留言编号	留言用户	留言内容	留言时间	留言详情	答复意见	答复时间
0	454	完整	0.419617	二星	2549	A00045581	A2区景善苑物业	2019/4/25 9:32:09	网友A00023583: 您好! 针对您反映A3区清善南路洋湖段怎么还没修好的问题,A2区桂花坪街道... 网友A00023583: 您好! 针对您反映A3区清善南路洋湖段怎么还没修好的问题,A2区桂花坪街道... 网友A00023583: 您好! 针对您反映A3区清善南路洋湖段怎么还没修好的问题,A2区桂花坪街道...	2019/5/10 14:56:53
1	305	良好	0.303666	二星	2554	A00023583	A3区清善南路洋湖段怎么还没修好?	2019/4/24 16:03:40	网友A00023583: 您好! 针对您反映A3区清善南路洋湖段怎么还没修好的问题,A2区桂花坪街道... 网友A00023583: 您好! 针对您反映A3区清善南路洋湖段怎么还没修好的问题,A2区桂花坪街道...	2019/5/9 9:49:10
2	357	良好	0.314153	一星	2555	A00031618	请加快提高A市民营幼儿园教师待遇	2019/4/24 15:40:04	市民同志: 您好! 您反映的请加快提高民营幼儿园教师待遇的问题,我们已收悉。现回复如下: 为了改善... 市民同志: 您好! 您反映的请加快提高民营幼儿园教师待遇的问题,我们已收悉。现回复如下: 为了改善...	2019/5/9 9:49:14
3	310	良好	0.113078	一星	2557	A000110735	在A市买公廉能享受人才新政吗	2019/4/24 15:07:30	网友A000110735: 您好! 您在平台《问政西地蜜》上的留言已收悉,市住建局及时将您反... 网友A000110735: 您好! 您在平台《问政西地蜜》上的留言已收悉,市住建局及时将您反...	2019/5/9 9:49:42
4	161	较差	0.199093	一星	2574	A0009233	关于A市公交站名称变更的建议	2019/4/23 17:03:19	网友A0009233: 您好, 您的留言已收悉, 现将具体内容答复如下: 关于来件人建议“白竹坡... 网友A0009233: 您好, 您的留言已收悉, 现将具体内容答复如下: 关于来件人建议“白竹坡...	2019/5/9 9:51:30

图 17 最终评价结果

参考文献

- [1] 邓三鸿, 傅余洋子, 王昊 基于 LSTM 模型的中文图书多标签分类研究《数据分析与知识发现》2017, 1 (7)
- [2] 陈倩, 齐林海, 王红 基于 LSTM 网络的谐波多标签分类 华北电力大学控制与计算机工程学院
- [3] 尉景辉, 何丕廉, 孙越恒 基于 K-Means 的文本层次聚类算法研究《计算机应用》, 2005 (10): 111-112
- [4] 黄志红 基于层次聚类的 k 均值算法研究《电脑开发与应用》, 2019, 022 (7): 1-2, 5
- [5] 赵巾帼, 徐德志, 罗庆云. 汉语句子相似度计算方法比对之研究[J]. 福建电脑, 2007, (10): 51-68
- [6] 骆亮 基于内容推荐算法和余弦相似度算法的领导决策辅助信息系统 知网 2018