

---

# “智慧政务”中的文本挖掘应用

## 摘要

近年来，随着互联网的发展，线上问政平台受众越来越大。网络上各类社情民意相关的文本数据量不断攀升，为提高相关部门受理线上平台的市民留言的效率，对这些线上问政平台的留言进行分析研究有重要意义。本文将基于文本挖掘和自然语言处理技术，对给出的 4 个附件中的民众留言和平台回复信息进行分析，从中提取我们所需的信息，并对此进行挖掘。

针对问题一：本文首先将附件 2 的留言详情信息进行去空、中文分词、停用词过滤的数据预处理，然后使用基于 VSM 和 TF-IDF 算法，将这些留言文本的分词结果作为文本的特征，按照词语的顺序和词频，构建出文本矩阵。虽然这种方法维度较高，但准确度较为可靠。然后将已有的数据每个类别分别抽取 20% 的样本作为测试集，剩下的 80% 作为训练集，将测试集、训练集的留言主题、留言详情内容都投影到同一个向量空间中，使用训练集中的文本向量训练 KNN 算法，用于预测测试集中文本内容的分类，并根据正确的分类结果和算法预测结果，给出这一分类方法的 F-Score 评价。

针对问题二：同样地，本文将附件 3 的留言主题进行去空、中文分词、停用词过滤的数据预处理，使用 TF-IDF 算法计算词频、词语权重构建文本矩阵。对得到的矩阵利用 K-Means 算法进行聚类，得到一时间段内给群众集中反映的同类问题，即热点问题，对这些问题进行合并，利用 TextRank 算法对这些问题进行摘要，得到留言描述。然后，利用层次分析法得到这些热点问题的热度指标，并利用此指标对问题进行排序。最后，利用 HMM 模型及 Viterbi 算法，筛选出排名前 5 的热点问题的人群或地点。

针对问题三：对附件 4 的答复意见内容进行分析，总结出民众比较可能会满意的答复内容一般含有以下特点：与留言主题及留言详情内容高度相似；含有比较详细的处理过程描述或者明确指出解决办法，这些内容一般包含相关的政府文件名、电话号码、时间描述等；答复中的名词、动词与留言内容含有对应性；答复与留言时间间隔较短。针对这些特点，给出合适的评分，并且给每个指标指定一定的权重，结合得到一套答复评价方案，使最终的评价具有区分度。

最后，还总结了本文的模型的不足之处，为以后的改进指明方向。

**关键词：**TF-IDF，文本分类，向量空间，K-Means，TextRank，评价方案

# 目录

一、挖掘目标 .....	1
二、分析方法与过程.....	1
2.1 问题 1 分析方法与过程 .....	2
2.1.1 流程图 .....	2
2.1.2 数据预处理.....	2
2.1.2.1 数据的去空 .....	2
2.1.2.2 中文分词、去停用词 .....	3
2.1.3 文本的向量化 .....	3
2.1.3.1 向量空间模型 .....	4
2.1.3.2 TF-IDF 权重 .....	4
2.1.3.3 生成向量 .....	5
2.1.4 选定测试集、训练集 .....	5
2.1.4 K 最近邻分类算法 <sup>[2]</sup> .....	6
2.1.5 F-Score 评价 .....	7
2.2 问题 2 分析方法与过程 .....	8
2.2.1 流程图 .....	8
2.2.2 数据预处理.....	8
2.2.2.1 去空，去除标点符号 .....	8
2.2.2.2 中文分词、去停用词 .....	8
2.2.2.3 TF-IDF 权重 .....	9
2.2.3 热点问题 .....	9
2.2.3.1 热点问题——用 K-means 进行文本聚类得到 <sup>[3]</sup> .....	9
2.2.3.2 描述热点问题——利用 TextRank 算法 <sup>[4]</sup> .....	10
2.2.3.3 热度指标制定——基于层次分析法 .....	11
2.2.3.4 词性标注——利用 HMM 及 Viterbi 算法 .....	12
2.3 问题 3 分析方法与过程 .....	14
2.3.1 评价方案大纲 .....	15
2.3.2 评价方案的制定 .....	15
2.3.2.1 相关性评价 .....	15
2.3.2.2 完整性评价 .....	18

2.3.2.3 可解释性评价 .....	20
2.3.2.4 及时性评价 .....	20
2.3.2.5 总体评价 .....	21
<b>三、结果分析 .....</b>	<b>22</b>
3.1 问题 1 结果分析 .....	22
3.1.1 K 值的选择 .....	22
3.1.2 分类方法的预测结果 .....	22
3.1.3 分类结果的 F-Score .....	23
3.2 问题 2 结果分析 .....	24
3.2.1 提取摘要 .....	24
3.2.2 热度指数 .....	25
3.3 问题 3 结果分析 .....	26
<b>四、结论 .....</b>	<b>27</b>
<b>参考文献 .....</b>	<b>28</b>

# 一、挖掘目标

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。

本次建模目标是利用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，其中包含结构化和非结构化的文本数据，通过对已有的文本数据进行预处理、中文分词、停用词过滤之后，利用特定的算法解决下面的问题：

- (1) 采用 KNN 算法，针对网络文章平台的群众留言，给出关于留言内容的一级标签分类方法，并且给出模型的 F-Score 评分。
- (2) 根据已有的群众反映的文本数据，利用 K-Means 聚类算法合并同类的文本得到给一时间段内给群众集中反映的某个热点问题，用 TextRank 算法计算热点问题的描述，使用层次分析法定义合理的热度评价指标，用 HMM 和 Viterbi 算法给出热点问题的人群或地点，然后将以此指标评价的排名前 5 的热点问题罗列出来，并且保存到热点问题留言明细表中。
- (3) 针对相关部门对留言的答复意见，从答复意见的相关性、完整性、可解释性等角度，给答复意见的质量作出评价。

# 二、分析方法与过程

总体流程图如下：

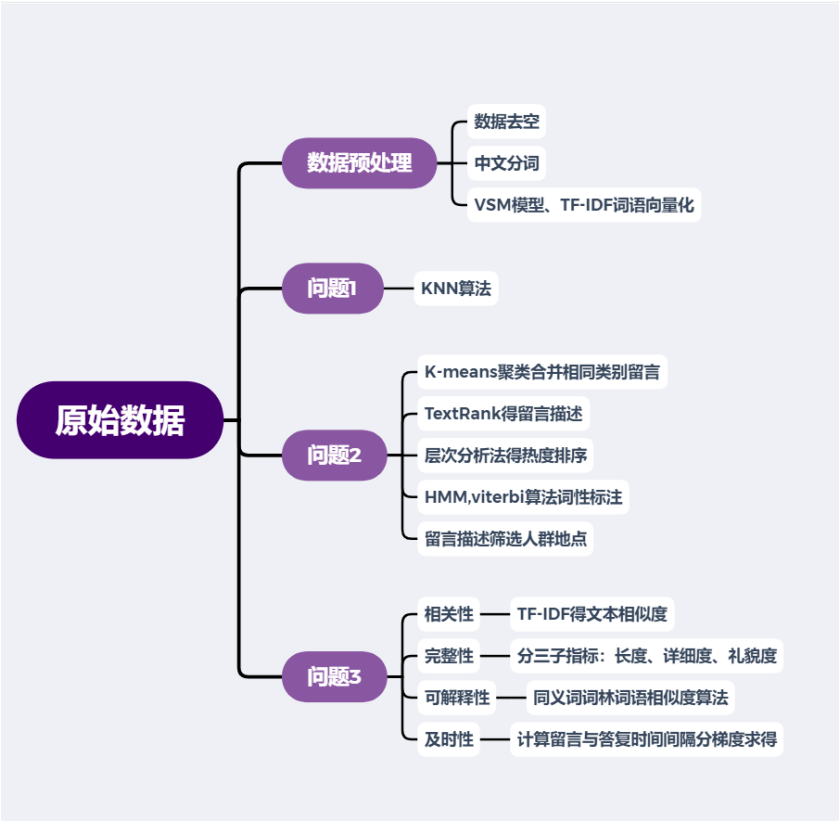


图 1 总体流程图

## 2.1 问题 1 分析方法与过程

### 2.1.1 流程图

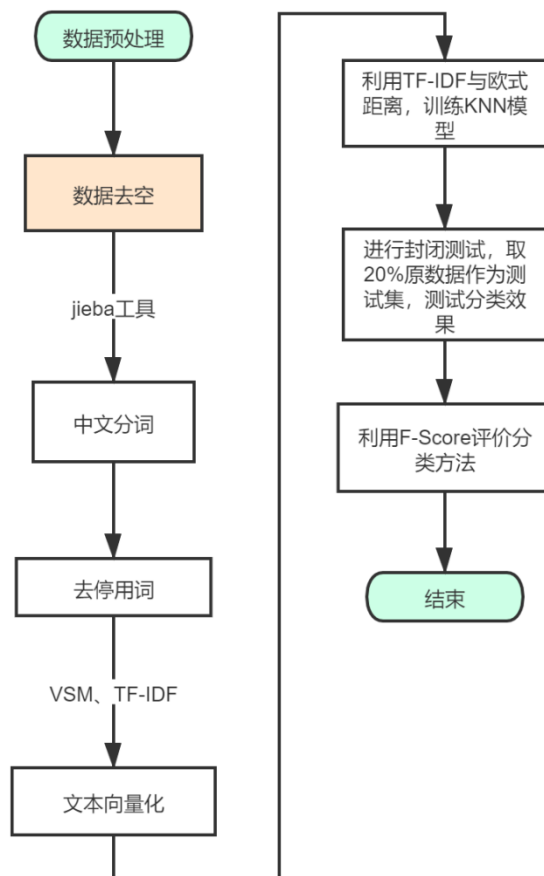


图 2 问题 1 处理流程图

上图为问题 1 的处理流程图，本文在处理该问题时主要分为以下步骤：

- (1)数据预处理。利用 jieba 分词工具处理原始留言文本，并去除停用词，结合 TF-IDF 算法，构建文本向量。
- (2)将已有的数据进行分割处理，取 20%的样本作为测试集，80%的样本作为训练集。
- (3)利用训练集训练 KNN 模型，利用测试集测试 KNN 算法。
- (4)利用 F-Score 给分类方法评价。

### 2.1.2 数据预处理

#### 2.1.2.1 数据的去空

在题目所给出的数据中，“留言详情”部分的文本头和尾都含有大量的空格，读取文本的时候会把这些空格当作文字读取进去，影响分析，因此需要对这些无用的数据进行去除。另外，虽然发现有部分重复数据，例如同一个用户在不同时间留相同的留言，但这部分只占少数，而且不影响问题 1 的分类过程，故不必进行去重操作。

利用 Python 第三方库 pandas 读取数据，对“留言详情”头尾的半角或者全角空格、半角双引号进行过滤，留下有用的文本，然后利用第三方库 openpyxl 将去空后的数据保存。对数据进行去空处理的程序请看附件 problem1 文件夹中的 delete\_space.py，去空后的数据保存为 problem1 文件夹下的附件 2\_处理后.xlsx。

### 2.1.2.2 中文分词、去停用词

在题目所给的附件 2.xlsx 中，数据以中文文本形式出现，需要将这种非结构化的数据转化成结构化数据。在此之前，需要对中文文本进行词语之间的分割。本文使用 python 的中文分词工具包 jieba 进行中文分词。jieba 工具包利用的算法有：

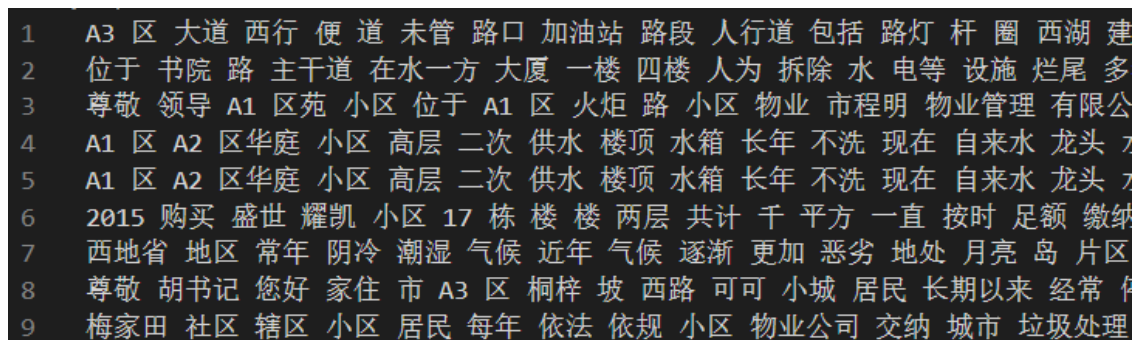
(1) 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG）。

(2) 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。

(3) 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

另外，由于文本中存在大量对分类要利用的文本特征没有贡献作用的字词，比如标点符号、语气词、人称代词、方位代词等等，因此需要对这些停用词作过滤处理。本文利用的是由开源社区上搜集的中文停用词表，在分词的过程中遍历停用词，如果发现分词结果中有词语是停用词，则不把它放入分词结果中。具体的停用词表保存在附件中的 problem1 文件夹下的 cn\_stopwords.txt 中。

部分的分词结果如图所示：



```
1  A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建
2  位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电等 设施 烂尾 多
3  尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公
4  A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头 水
5  A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头 水
6  2015 购买 盛世 耀凯 小区 17 栋 楼 楼 两层 共计 千 平方 一直 按时 足额 缴纳
7  西地省 地区 常年 阴冷 潮湿 气候 近年 气候 逐渐 更加 恶劣 地处 月亮 岛 片区
8  尊敬 胡书记 您好 家住 市 A3 区 桐梓 坡 西路 可可 小城 居民 长期以来 经常 住
9  梅家田 社区 辖区 小区 居民 每年 依法 依规 小区 物业公司 交纳 城市 垃圾处理
```

图 3 问题 1 部分分词结果

所有的分词结果保存在 problem1 文件夹下的 X.txt 文件中。

### 2.1.3 文本的向量化

分词工作完成以后，需要把这些词语表示为可供计算机挖掘的形式，一般而言是向量形式。本文采用的是常用的向量空间模型（Vector Space Model）<sup>[1]</sup>结合 TF-IDF 加权算法，

将所有的留言主题、留言详情信息作为一个语料库，将单一的留言的主题、详情信息作为一个文档，文档中的分词作为特征项，将文档转化成权重向量。

TF-IDF 是一种统计方法，用于评估一个词语对于一个语料库中的一个文本的重要程度。统计表示，一个词语出现在文本中的次数越多、在语料库中出现的次数越少，说明它对该文本的重要性越大。

### 2.1.3.1 向量空间模型

向量空间模型的主要思想是把一篇文章抽象成一个向量，每一个不同的词对应为向量空间中的一个维度，对应维度的值就是这个词在文章里的权重。下面做 VSM 模型的详细介绍。

用  $D$  表示文档，词语  $t_i$  ( $1 \leq i \leq n$ ) 作为特征项，那么文档  $D$  可以表示为：

$$D(t_1, t_2, t_3, \dots, t_n) \quad (1)$$

但不同的特征项在文章中具有不一样的重要性，因此需要赋予一定权重，来表示它的重要程度：

$$D = (t_1, w_1; t_2, w_2; t_3, w_3, \dots, t_n, w_n) \quad (2)$$

由于特征项的长度、顺序是固定的，因此可以简单记为：

$$D = (w_1, w_2, w_3, \dots, w_n) \quad (3)$$

这时， $D$  成为文档的向量表示， $w_i$  是特征项  $t_i$  ( $1 \leq i \leq n$ ) 的权重，这里的权重本文利用 TF-IDF 算法得到。

### 2.1.3.2 TF-IDF 权重

(1) 词频，即 TF，表示词语在文本中出现的频率。由于考虑到文本的长短，防止它偏向长文件，通常这一数值会被归一化。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (4)$$

式子中： $n_{i,j}$  是词  $t_i$  在文本  $d_j$  中出现的次数， $\sum_k n_{k,j}$  则是文本  $d_j$  中所有词汇出现的次数总和。

(2) 逆向文件频率，即 IDF，可以由语料库中所有文件的数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含该词语的文档越少，则 IDF 越大，说明词条具有很好的类别区分能力。

对于词语  $t_i$ ，它的  $idf_{t_i}$  基本公式为：

$$idf_{t_i} = \log \frac{|D|}{|\{d \in D: t_i \in d\}|} \quad (5)$$

式中： $|D|$  为语料库中所有文档的总数， $|\{d \in D: t_i \in d\}|$  是包含词语  $t_i$  的文档数。

但是这样的公式定义存在一定问题：如果有在语料库中不存在的生僻词，此时分母会为 0，IDF 没有意义。因此需要对公式进行平滑处理，一般使用如下的公式：

$$idf_{t_i} = \log \frac{|D|+1}{|\{d \in D: t_i \in d\}|+1} + 1 \quad (6)$$

(3) 知道了一个词的词频、逆文档频率之后，就可以得出一个词的 TF-IDF 权重：

$$tf-idf_{t_i} = tf_{i,j} \times idf_{t_i} \quad (7)$$

### 2.1.3.3 生成向量

得到了 TF-IDF 权重之后，文档 D 的向量表示即为：

$$D = (tf - idf_{t_1}, tf - idf_{t_2}, tf - idf_{t_3}, \dots, tf - idf_{t_n}) \quad (8)$$

本文利用 sklearn 库中的 CountVectorizer 方法，将文本中的词语转换为词频矩阵，再通过 fit\_transform 函数计算各个词语出现的次数，实现了文本特征提取，得到每个文本的 TF-IDF 权重，和以 TF-IDF 权重表示的文本向量。

### 2.1.4 选定测试集、训练集

为了方便评估分类方法，把已有的数据分为测试集、训练集，用训练集训练模型，测试集用以测试分类方法的准确性。

通过对附件 2.xlsx 的统计，不同类别的数量如下图所示。

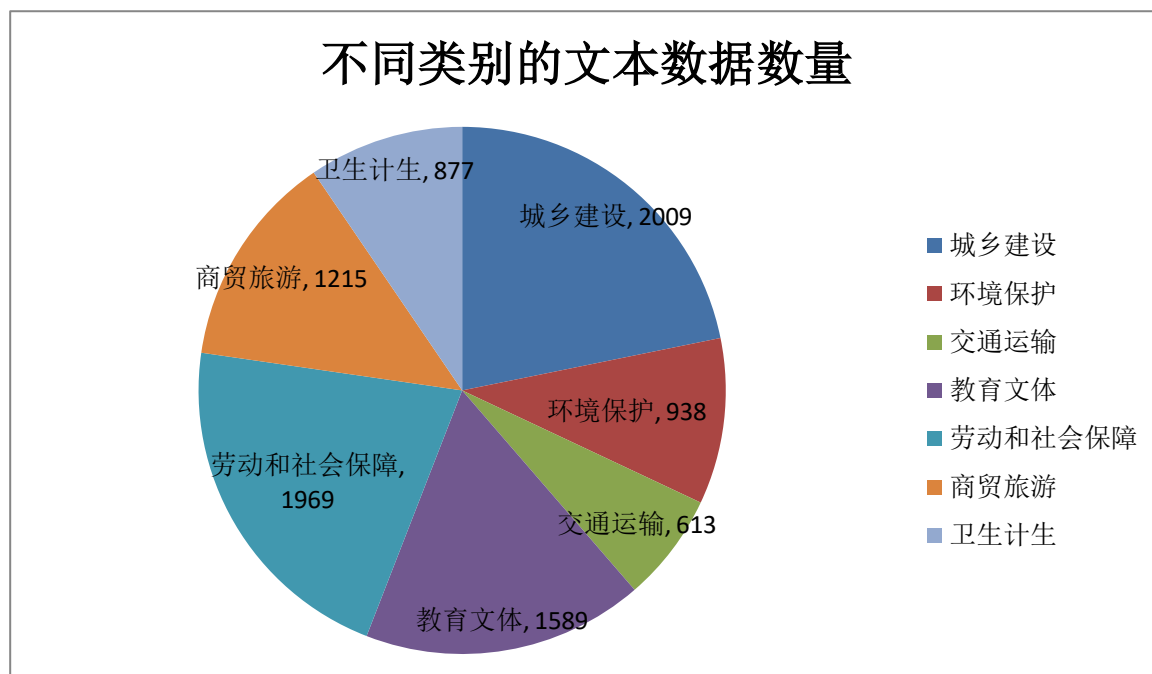


图 4 问题不同类别的文本数据数量

每个类别中抽取 20% 的样本，作为测试集，抽取的数量如下表所示：



表格 1 测试集中不同类别的文本数量

类别	数量
城乡建设	401
环境保护	187
交通运输	122
教育文体	317
劳动和社会保障	393
商贸旅游	243
卫生计生	175
总计	1838

按顺序从原数据中抽取相应数量的样本，保存为 problem1 文件夹下的“测试集.xlsx”，剩下的样本则保存为 problem1 文件夹下的“训练集.xlsx”，抽取的程序位于 problem1 文件夹下的 pick\_test\_and\_train.py。

#### 2.1.4 K 最近邻分类算法<sup>[2]</sup>

K 最近邻算法 (K Nearest Neighbors)，即 KNN 算法，其核心思想是如果一个样本在特征空间中的 k 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。KNN 是一个常用的统计方法。

KNN 算法的基本思路是：考虑在训练文本集中与新文本距离最近的 k 篇文本，根据这 k 篇已知分类的文本，判断新文本的类别。算法的具体步骤如下：

(1) 用已知分类的留言详情文本向量作为训练集，训练模型。

(2) 新文本的向量到达后，计算新文本与训练集中的文本之间的距离。本文利用的是基于距离度量的欧氏距离计算相似度，距离越近则表示文本越相似。两个文本的欧式距离的计算公式如下：

$$\text{dist}(d_i, d_j) = \sqrt{\sum (x_k - y_k)^2} \quad (9)$$

式子中：  $d_i = (x_1, x_2, \dots, x_n)$ ,  $d_j = (y_1, y_2, \dots, y_n)$ ,  $d_i, d_j$  表示两个不同的文本。

(3) 从训练集中选取 K 个与新文本最相似的文本。其中 K 的值的确定方法尚无定论，一般定一个初始值，然后使之在一定范围之内浮动，根据实验测试的结果，选出最合适的 K 值。

(4) 确定前 K 个文本所在类别的出现频率。计算公式如下：

$$P(X, C_j) = \sum_{i=1}^k \sum_{j=1}^m \text{dist}(x, d_i) \times y(d_i, C_j) \quad (10)$$

式子中：X 为新文本， $C_j$  为类别，m 为总的类别数； $d_i$  ( $1 \leq i \leq K$ ) 为 K 个文本中的第

i 个；如果  $d_i$  属于  $C_j$  类， $y(d_i, C_j)=1$ ，否则  $y(d_i, C_j)=0$ 。

(5) 返回前 K 个文本中出现频率最高的类别，作为新文本的预测分类结果。

进行分词、去停用词、进行向量化并且利用 KNN 算法进行预测的程序文件在 problem1 文件夹下的 KNN.py。

### 2.1.5 F-Score 评价

制定分类方法后，需要用一套评价方案来评估分类方法，为以后的改进指出方向。本文利用 F-Score 作为分类方法的评价标准。

分类模型有三个常用的评估指标：查准率（Precision）和查全率（Recall）。

查准率是所有预测文本中占分类正确的文本数比率，公式为：

$$\text{准确率 } P = \frac{\text{预测正确的文本数}}{\text{实际分类的文本数}}$$

查全率是一个已知分类下应有的文本数中分类正确的文本数所占的比率，公式为：

$$\text{查全率 } R = \frac{\text{预测正确的文本数}}{\text{该分类下应有的文本数}}$$

查准率、查全率反映了分类方法的两个方面的准确率，二者需要综合考虑，因此有 F-Score 这一新的指标，对于其中一个分类，其计算公式为：

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (11)$$

对于多个分类，计算其平均值即可：

$$F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_i R_i}{P_i + R_i} \quad (12)$$

式子中： $P_i$  为第 i 类的查准率， $R_i$  为第 i 类的查全率。

## 2.2 问题 2 分析方法与过程

### 2.2.1 流程图



图 5 问题 2 处理流程图

### 2.2.2 数据预处理

从附件 3 的 excel 中提取留言主题一行，将每行的信息提取到 txt 文件中。在进行热点问题挖掘前，对这些数据信息首先进行数据预处理操作。

#### 2.2.2.1 去空，去除标点符号

在程序中导入 re 库，对提取的数据信息首先使用正则表达式将数据中的标点符号以及一些特殊符号过滤掉，同时将文本中的 tab 符与空格进行去除处理，以便后续进行文本分析

#### 2.2.2.2 中文分词、去停用词

为了能够更准确地对留言主题进行分析，首先对这些主题信息进行中文分析。此处利用 jieba 分词首先进行中文分词。针对问题 2 的内容，需要提取出相对精确的内容，故在此使用 jieba 分词中的精确模式，将句子最精确地切开，以便于后续的文本分析。所有问题 2 的分词结果都保存在 problem2 文件夹中的 comb.txt 文件中。接着进行停用词去除操作，此次使用的是哈工大整理的停用词，其数据信息在 problem2 文件夹中的 stop\_words.txt 文件中。部分分词结果如下图所示：

```

A4 区 北辰 小区 非法 住 改商 问题 何时能 解决
end file:16
请 给 K3 县 乡村 医生 发 卫生室 执业 许可证
end file:17
A7县 春华 镇 石塘 铺村 有 党员 家开 麻将馆
end file:18
咨询 异地 办理 出国 签证 的 问题
end file:19
投诉 A市 温斯顿 英语 培训 学校 拖延 退费
end file:20
A6区 乾源 国际 广场 停车场 违章 乱建 现象 严重
end file:21
A7县 时代 星城 4 幢 有 非法经营 的 家庭旅馆
end file:22
A2 区佳兆业 水新 都 小区 垃圾 无人 处理

```

图 6 问题 2 分词、去停用词部分结果

### 2.2.2.3 TF-IDF 权重

在进行留言主题信息进行分词后，将文本信息转换为向量，以便于后续进行文本聚类。此处使用 TF-IDF 算法得到对应的权重向量。具体的方法在上文已有提及。

### 2.2.3 热点问题

在进行热点问题挖掘时，其思路首先是将附件 3 中留言主题一列进行文本聚类得到多个类别的主题，再通过这些相同主题的留言主题描述合并进行文本摘要算法，从而得到热点问题描述。而时间范围也根据合并在同一类的文本得到。利用层次分析法得到每个热点问题对应的热度排名，根据 HMM 及 Viterbi 算法，加载自定义词典和停用词，筛选出热点问题描述中的地点和人群。最后得到相应的热点问题表数据。

#### 2.2.3.1 热点问题——用 K-means 进行文本聚类得到<sup>[3]</sup>

利用 TF-IDF 得到文本的词语的 TF-IDF 权重，进行 K-means 文本聚类，将留言主题类似的合并在一起，再对这些合并完的文本进行进一步摘要，得到留言描述。

K-means 算法的思想主要是：对于给定的样本集合将其划分为 k 簇，从样本集中随机选取 k 个数据点作为质心，对样本集中每个样本计算与每个质心的距离，距离最近的则划分到对应的质心所在类中。接着又在新的类别中重新选出质心，一直循环，直到每个质心的距离小于某一设置的阈值则算法终止。

K-means 算法的原理则是：在给定的样本集中，将其划分为 K 簇，如果用数据表达式表示，假设将样本集中的簇划分为  $(C_1, C_2, \dots, C_k)$   $(C_1, C_2, \dots, C_k)$ ，则我们的目标是最小化平方误差 E：

$$E = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (13)$$

其中  $\mu_i$  为簇  $C_i$  的均值向量，即质心，表达式为：

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (14)$$

求上式的最小值只能采用启发式的迭代方法。

在此次运行过程中总共有 4326 条数据，将其划分为 787 类留言主题，并将对应的合并结果写入到 all\_txt4 文件夹中。

### 2.2.3.2 描述热点问题——利用 TextRank 算法<sup>[4]</sup>

在用 K-means 算法聚类之后，对相同类别的文本进行合并，并使用合并后的文本进行留言描述的实现。此次使用 text-rank 算法实现文本的摘要，从而得到对应的留言描述。

TextRank 算法是一种抽取式的文本摘要算法，也是一种基于图的排序算法，主要用文本分析，能够将文本分割成若干句子，通过构建节点连接图，用句子之间的相似度作为边的权重，接着循环迭代从而得到句子的 TextRank 值，抽取排名高的句子组成文本摘要。

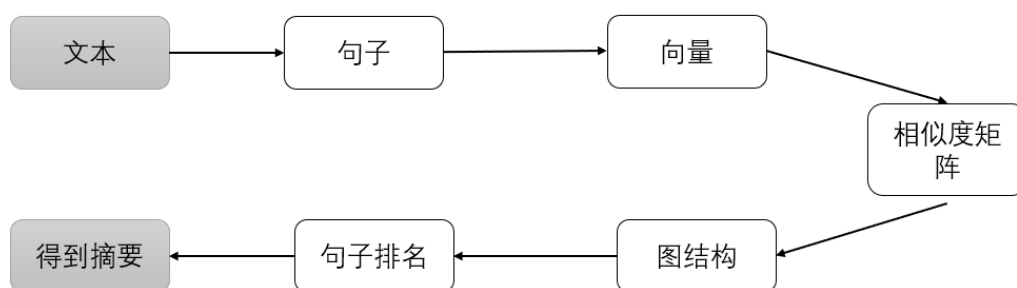


图 7 TextRank 算法流程图

TextRank 算法是由 PageRank 算法演变而来的，其算法流程为：

首先将文本分割成句子，接着将句子转换为词向量，通过计算句子间向量的相似性得到相似度矩阵，然后将相似度矩阵转换为以句子作为节点，相似性得分作为边的图结构，并用句子 TextRank 计算，最后将排名最高的句子作为最后的摘要。

此次摘要的部分结果如下图：

A7县海伦春天购房贷款进度
万科魅力城小区底层门店深夜营种噪音扰民
A1区农业学滨河小区麻馆扰民严重
A市梅溪湖嘉苑安置小区房屋质量差
西省聚利人普惠投资有限公司涉嫌诈骗巨额资金
投诉A4区沙坪街道卫生服务中心工程项目拖欠工资
A7县松雅湖小学噪音扰民严重
咨询A市人新政落户相关问题
A市绿城际空间站建筑质量太差

图 8 问题 2 部分摘要结果

2. 2. 3. 3 热度指标制定——基于层次分析法

由于熵权法对于有重复元素的数据处理效果并不好，因此我们选择了更简单有效的层次分析法。层次分析法原本用于解决多目标决策问题，同时可以确定各因素的权重，因此我们采用层次分析法，决策问题为热度指数，考虑的因素为各类主题的评价数、点赞数以及反对数。然后我们计算因素对于决策问题占的权重。

表格 2 符号说明

A	热度指数
B1	评论数
B2	点赞数
B3	反对数

首先建立递阶层次结构模型。

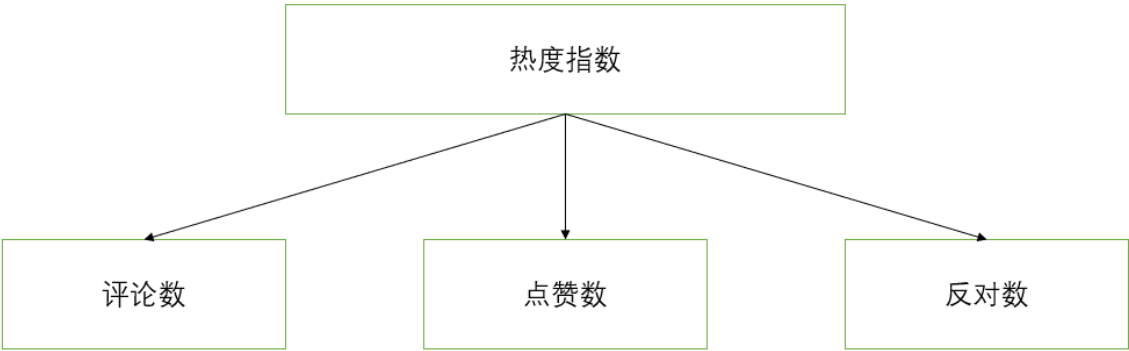


图 9 AHP 层次结构

其次设置标度。

表格 3 标度说明

标度	定义（比较因素 i 与 j）
1	因素 i 与 j 同样重要
3	因素 i 与 j 稍微重要
5	因素 i 与 j 较强重要
2、4	两个相邻判断因素的中间值
倒数	因素 i 与 j 比较得判断矩阵 $a_{ij}$ ，则因素 j 与 i 相比的判断为 $a_{ij} = \frac{1}{a_{ij}}$

显然，比值越大，则要素 i 的重要度越高。

构造判断矩阵：

表格 4 判断矩阵

A	B1	B2	B3
B1	1	5	3
B2	$\frac{1}{5}$	1	$\frac{1}{3}$
B3	$\frac{1}{3}$	$\frac{1}{5}$	1

计算判断矩阵的特征值，特征向量和归一化。用求和法计算特征值。

一致性检验通过，最大特征根为 3.0385，CI 值为 0.019, CR 值为 0.037。

表格 5 AHP 结果

因素	几何平均值	权重
B1	2.466	0.637
B2	0.405	0.104
B3	1	0.258

据此权重和原始数据，我们可以算出各个主题评论的热度指数如下。更多内容可见 Problem2 文件夹下 AHP\_Hot\_Point\_Origin.xlsx 文件。

热度指数	问题描述
534.18	A市五矿万境K9县存严重消防安全隐患
447.78	投诉A市A1区苑物业违规收停车费
393.91	请书记关注A市A4区58车贷案
204.59	请重视A5区砂子塘万境水岸小学路段交通安全问题
178.29	A7县星沙中贸城欺诈业主退业主购房资金
69.28	丽发新城小区旁搅拌站严重影响生
53.45	咨询A7县道路规划问题
48.49	关A6区月亮岛路线架设110kv高压电线杆投诉

图 10 热度指数部分内容

2.2.3.4 词性标注——利用 HMM 及 Viterbi 算法

这两个算法主要用于筛选文本中的地点/人群信息。jieba 分词中提供了词性标注功能，可以标注句子分词后每个词的词性，词性标注集采用北大计算所词性标注集，属于采用基于统计模型的标注方法。原理是基于 HMM 模型和 Viterbi 算法。由于题目要求我们筛选出地点/人群，所以词性设置为' ns' 及' n' ，再加载停用词及自定义词典即可。

具体步骤如下。

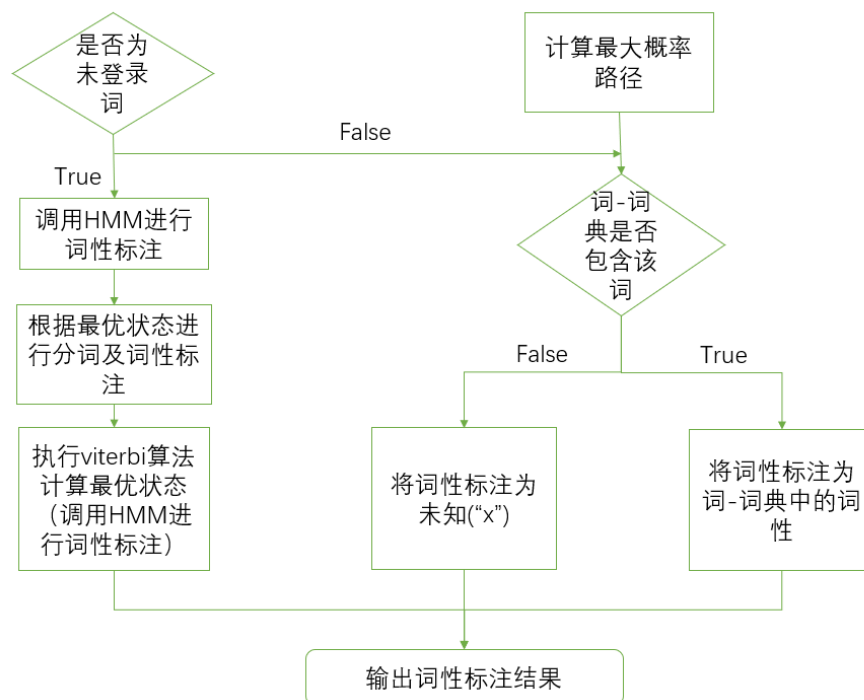


图 11 jieba 词性标注过程

隐马尔可夫模型（HMM）是一种统计分析模型，应用于大量领域。而 viterbi 算法是人们基于解码问题提出的解决算法。解码问题可以简单地理解为给定观测序列和模型参数，如何寻找某种意义上最优的隐状态序列。

HMM 模型的两个基本假设：

（1）齐次马尔可夫性假设：即假设隐藏的马尔可夫链在任意时刻  $t$  的状态只依赖于其前一时刻的状态，与其他时刻的状态及观测无关，也与时刻  $t$  无关

（2）观测独立性假设：即假设任意时刻的观测只依赖于该时刻的马尔可夫链的状态，与其他观测和状态无关。

HMM 模型中的五元组表示：

- （1）观测序列
- （2）隐藏状态序列
- （3）状态初始概率
- （4）状态转移概率
- （5）状态发射概率

Jieba 分词训练出每个词的初始状态概率如下式，每个概率值都是取对数之后的结果，其中  $-3.14^{100}$  表示负无穷，对应的概率值就是 0。在这个概率表中，说明了一个词中第一个字属于 {B、M、E、S} 这四种状态的概率。

$$P = \{B': -0.2626866, E' = -3.14^{100}, M' = -3.14^{100}, S' = -1.4652633\} \quad (15)$$



表格 6 HMM 符号说明

B	首字
M	词的中间位置
P	词的结束位置
S	单字成词

状态转移概率是指当前时刻的状态，只与当前时刻之前的状态有关。jieba 中的状态转移概率其实是嵌套的词典，数值是概率值求对数后的值。

$$P = \{ 'B': \{ 'E': -0.510256, 'M': -0.916290 \}, 'E': \{ 'B': -0.589714, 'S': -0.808525 \}, 'M': \{ 'E': -0.3334485, 'M': -1.2603623 \}, 'S': \{ 'B': -7211965, 'S': -0.665863 \} \} \quad (16)$$

P[B][E]是指从状态 B 转移到状态 E 的概率，如 P[B][E]=-0.589714，对应概率是 0.6，说明当我们处于一个词的开头时，下一个字是结尾的概率要远高于下一个字是中间字的概率。是因为双字词比多个字的词更常见。

状态发射概率，根据 HMM 模型中观测独立性假设，发射概率即观测值只取决于当前的状态值。

利用 Viterbi 算法，实际上就是使用动态规划来求解 HMM 解码问题，即用动态规划求概率路径最大。

jieba 分词实现 Viterbi 算法是在 viterbi(obs, states, start\_p, trans\_p, emit\_p) 函数中实现。viterbi 函数会先计算各个初始状态的对数概率值，然后递推计算，每时刻某状态的对数概率值取决于上一时刻的对数概率值、上一时刻的状态到这一时刻的状态的转移概率、这一时刻状态转移到当前的字的发射概率三部分组成。

由此我们得到的地点/人群效果如下。

表格 7 部分地点/人群结果

地点/人群	问题描述
A 市五矿万境 K9 县	A 市五矿万境 K9 县存严重消防安全隐患
A 市 A1 区物业	投诉 A 市 A1 区苑物业违规收停车费
A 市 A4 区书记	请书记关注 A 市 A4 区 58 车贷案
A5 区万境小学	请重视 A5 区砂子塘万境水岸小学路段交通安全问题
A7 县星沙业主	A7 县星沙中贸城欺诈业主退业主购房资金

筛选出来热度排名前 5 的热点问题表保存在“热点问题表.xlsx”中，这些热点问题的留言明细表保存于“热点问题留言明细表.xlsx”中。

2.3 问题 3 分析方法与过程

2.3.1 评价方案大纲



图 12 评价方案大纲图

2.3.2 评价方案的制定

2.3.2.1 相关性评价

一定程度上，两个文本相关度越高意味着两个文本的相似度越高。

传统的文本相似度量方法基本上都是将文本看作词的集合，分析每个词在文本中出现的次数、整个文本集合中出现的次数，利用词频将文本建模为向量。在此基础上，利用向量间的余弦相似度计算文本之间的相似度，但这种方法忽略了词语之间的语义相似性。

故本文使用的是一种结合词语语义信息和 TF-IDF 方法的文本相似度计算方法<sup>[5]</sup>，通过计算留言主题、留言详情、答复详情之间的相似性，来评估答复的相关性。相关处理的步骤如下。

(1) 首先需要处理文本中的地名、机构名等特殊名称。虽然它们在语义上判断文本的相似性并没有多大的帮助，但答复与留言中有相同的地名、机构名，说明这两个文本之间是有一定关联性的。这些词语在分词的时候，需要特别保留。

(2) 对每一篇留言的词语进行 TF-IDF 值的计算，并且将之转换为向量，具体的计算方法上文已有提及。一篇文章中含有多个词语，不是每个词语都对语义有贡献，因此可以从挑选出若干重要的词语，作为文本的关键词来代表文本，以减少文本特征向量的维度。具体的做法：把留言详情中词语的 TF-IDF 值排序，选择 TF-IDF 值较高的若干词作为关键词，加上文本的特征项，建立起留言文本的特征向量  $v$ 。特别地，对于留言文本，留言主题中的内容简洁且重要，因此需要保留留言主题的所有分词作为关键词，而留言详情、答复内容则用上述方法抽取关键词。本文在处理时，利用的是 jieba 工具中的基于 TF-IDF 算法的关键词抽取工具，默认抽取 30 个关键词。

(3) 计算留言文本、答复文本之间的相似度，作为留言与答复之间的相关性度量。

设 $v_i, v_j$ 是两篇文本 $d_i, d_j$ 的特征向量，其中 $v_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{im})$ ， $v_j = (w_{j1}, w_{j2}, w_{j3}, \dots, w_{jn})$ ，则文本相似度可以由以下式子定义：

$$\text{TextSim}(v_i, v_j) = w_f \times \text{VectSim}(v_i, v_j) \quad (17)$$

式子中： $w_f$ 为文本特征向量 $v_i, v_j$ 之间相似度的加权因子， $\text{VectSim}(v_i, v_j)$ 表示文本特征向量 $v_i, v_j$ 之间的相似度。如果 $d_i, d_j$ 中彼此相似度高的词语越多，则词语所占的 TF-IDF 值在各自的文档中比例越高，说明这些词语在文本中更重要。具体的加权因子 $w_f$ 计算方法如下：

$$w_f = 1 + \text{ave}(i, j) \times (\sqrt{\text{VectSim}(v_i, v_j)} - \text{VectSim}(v_i, v_j)) \quad (18)$$

$$\text{ave}(i, j) = \frac{1}{2} \left( \frac{\sum_{k \in \Lambda_i} \text{TFIDF}(w_{ik})}{\sum_{k=1}^m \text{TFIDF}(w_{ik})} + \frac{\sum_{k \in \Lambda_j} \text{TFIDF}(w_{jl})}{\sum_{j=1}^n \text{TFIDF}(w_{jl})} \right) \quad (19)$$

式中： $\text{TFIDF}(w_{ik})$ 表示词语 $w_{ik}$ 的 TF-IDF 值， $\frac{\sum_{k \in \Lambda_i} \text{TFIDF}(w_{ik})}{\sum_{k=1}^m \text{TFIDF}(w_{ik})}$ 表示 $d_i$ 向量中所有满足相似度阈值条件的词语的 TF-IDF 值在文本 $d_i$ 中的所有词语的 TF-IDF 值的比值。

如果 $d_i$ 中的一个词语 $w_{ik}$ 与另一向量 $d_j$ 中的一个词语 $w_{jl}$ 相似度超过阈值 $\mu$ ，则将该词语的位置放入 $\Lambda_i$ ， $\Lambda_j$ 中的元素同理。本文将阈值 $\mu$ 定为 0.8。

集合 $\Lambda_i, \Lambda_j$ 定义为：

$$\Lambda_i = \{k: 1 \leq k \leq m, \max_{1 \leq l \leq n} \{\text{Sim}(w_{ik}, w_{jl})\} \geq \mu\},$$

$$\Lambda_j = \{l: 1 \leq l \leq n, \max_{1 \leq k \leq m} \{\text{Sim}(w_{jl}, w_{ik})\} \geq \mu\} \quad (20)$$

$$\begin{aligned} \text{VectSim}(v_i, v_j) = & \frac{1}{2} \left( \frac{1}{m} \sum_{k=1}^m \max_{1 \leq l \leq n} \{\text{Sim}(w_{ik}, w_{jl})\} + \right. \\ & \left. \frac{1}{n} \sum_{l=1}^n \max_{1 \leq k \leq m} \{\text{Sim}(w_{jl}, w_{ik})\} \right) \end{aligned} \quad (21)$$

以上两个式子中： $\text{Sim}(w_{ik}, w_{jl})$ 表示词语 $w_{ik}, w_{jl}$ 之间的语义相似度。

其中的词语语义相似度是基于哈工大同义词词林扩展版的计算，词林来自于开源社区<sup>[6]</sup>。具体的计算方法如下<sup>[7]</sup>。

同义词词林词典分类采用层级体系，具备 5 层结构。随着级别的递增，词义刻画越来越细，到了第 5 层，每个分类里词语数量已经不大，很多只有一个词语，已经不可再分，可以称为原子词群、原子类或原子节点。

词典文件位于附件中的 problem3 文件夹中的 cilin.txt，部分词语如图所示：

Aa01A01= 人士 人物 人士 人氏 人选  
Aa01A02= 人类 生人 全人类  
Aa01A03= 人手 人员 人口 人丁 食指  
Aa01A04= 劳力 劳动力 工作者  
Aa01A05= 匹夫 个人  
Aa01A06= 家伙 东西 货色 厮 崽子 兔崽子 犊子 鼠辈 小崽子

图 13 同义词词林词典部分词语、编码

等于号前的是编码，编码含义见表：

表格 8 编码含义								
编码位	1	2	3	4	5	6	7	8
符号举例	D	a	1	5	B	0	2	=/#/@
符号性质	大类	中类	小类	词群	原子词群			
级别	第 1 级	第 2 级	第 3 级	第 4 级	第 5 级			

上表中的编码位按从左到右排序，第 8 位的标记中，“=”表示相等；“#”表示不等，属于相关词语；“@”表示词语在词典中没有同义词，也没有相关词。

了解了词典的具体结构，接下来说明具体的两个词之间的相似度计算算法。

首先判断在同义词林中作为叶子节点的两个义项在哪一层分支，即两个义项的编号在哪一层不同。从第 1 层开始判断，相同则乘 1，否则在分支层乘以相应的系数，然后乘以调节参数 $\cos(n \times \frac{\pi}{180})$ ，n 为分支层的结点总数。调节参数的功能是把词语的相似度控制在[0,1]之间。词语所在的树的密度、分支的多少直接影响到词语的相似度，密度较大的词语相似度应该比密度小的相似度更精确。因此需要再乘以一个控制参数 $\frac{n-k+1}{n}$ ，n 是分支层的节点数，k 是分之间的距离。

这样，得出来的词语相似度公式：

$$\text{Sim}(A, B) = \begin{cases} f, A \text{ 和 } B \text{ 不在同一棵树上} \\ 1 \times a \times \cos\left(n \times \frac{\pi}{180}\right) \times \frac{n-k+1}{n}, \text{若 } A \text{ 与 } B \text{ 在第 2 层分支} \\ 1 \times 1 \times a \times \cos\left(n \times \frac{\pi}{180}\right) \times \frac{n-k+1}{n}, \text{若 } A \text{ 与 } B \text{ 在第 3 层分支} \\ 1 \times 1 \times 1 \times a \times \cos\left(n \times \frac{\pi}{180}\right) \times \frac{n-k+1}{n}, \text{若 } A \text{ 与 } B \text{ 在第 4 层分支} \\ 1 \times 1 \times 1 \times 1 \times a \times \cos\left(n \times \frac{\pi}{180}\right) \times \frac{n-k+1}{n}, \text{若 } A \text{ 与 } B \text{ 在第 5 层分支} \\ 1, A \text{ 与 } B \text{ 编号相同} \\ e, A \text{ 与 } B \text{ 编号相同且只有末尾号为}\# \end{cases} \quad (22)$$

经过多次试验，人工评定后将层数的系数定义为： $a = 0.65, b = 0.8, c = 0.9, d = 0.96, e = 0.5, f = 0.1$ 。

基于同义词词林计算词语相似度的程序位于 problem3 文件夹中的 Cilin.py，计算文

本相似度的程序位于 problem3 文件夹中的 TextSim.py。

答复与留言的相关性度量如下：

$$\text{RelativeScore}(\text{ans}) = \text{TextSim}(\text{com}, \text{ans}) \quad (23)$$

式子中：ans 代表答复内容，com 代表留言内容。文本相似度即答复相关性最高评价为 1。

2.3.2.2 完整性评价

分析附件 4 中的答复意见内容，从中抽取出 3 条比较优秀的答复、3 条相对答复的回答，作为比较。

表格 9 部分优秀答复

优秀的答复
<p>现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2 区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于 2019 年 4 月 10 日至 4 月 27 日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真梳理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5 月 5 日下午，辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。2019 年 5 月 9 日</p> <p>网友“A000100804”：您好！针对您反映 A3 区教师村小区盼望早日安装电梯的问题，A3 区住建局高度重视，立即组织精干力量调查处理，现回复如下：为了完善住宅使用功能，提高我区既有多层住宅居民的宜居水平，2018 年 6 月 7 日，A 市 A3 区人民政府办公室下发了《关于 A 市 A3 区既有多层住宅增设电梯实施方案》的通知。该方案明确了增设电梯条件及流程。详情可咨询政务服务中心住建局受理申请老旧小区加装电梯窗口，咨询电话：0000-00000000。感谢您对我们工作的关心、监督与支持。2019 年 4 月 2 日</p> <p>尊敬的网友：你好，经安监、质监部门核实，您举报 B 市泰民米粉厂存在严重安全隐患与事实不符。一、举报锅炉有 8 公斤压力，实际生产时最高压力也只有 1.0 公斤，市质监局对锅炉已经做过质量检测，无质量问题，且在有效期检测范围内，不存在爆炸危险性。二、厂房并没有加长，老板黄云满在 2008 年租赁该厂房时到现在一直是维持原状。三、水井是打的钻井，并不是开掘的方井、圆井，井水并不用于生产，而是用于打扫卫生，用量并不多。四、B 市泰民米粉厂工商营业执照、食品生产许可证等证照齐全，不存在违法生产，属合法经营。</p>

表格 10 部分有缺陷的答复

不足的答复
<p>你的建议很好。我们已将你的建议转给有关部门研究。</p> <p>A00085038: 您在《问政西地省感谢您的理解与支持! 2018 年 8 月 6 日</p> <p>网友: 您好, 您反映问题已获悉并转职业中专学校按实际情况处置</p>

对比而言, 比较优秀的答复具有以下特点:

(1) 答复字数多。这里抽取出来的平均答复字数为 292.8 个, 而较差的答复平均答复字数为 21.4 个。

(2) 内容详细、有条理。具体体现在含有处理的进度日期或时间、含有相关部门的联系方式或者含有事件相关的数字信息、条目分点整理回答。

(3) 比较有礼貌, 含有问候语、感谢的短语表达。

因此, 可以根据以上特点制定完整性的评分公式:

$$\text{CompleteScore}(\text{ans}) = \text{LengthScore}(\text{ans}) + \frac{\text{DetailScore}(\text{ans})}{6} + \text{PoliteScore}(\text{ans}) \quad (24)$$

$$\text{LengthScore}(\text{ans}) = \log_l(\text{len}(\text{ans}) + 1) \quad (25)$$

$$\text{DetailScore}(\text{ans}) =$$

$$\begin{cases} 2, \text{ans 中含有时间、含有具体文件或联系方式、分点回答中的其中 1 个特点} \\ 4, \text{ans 中含有时间、含有具体文件或联系方式、分点回答中的其中 2 个特点} \\ 6, \text{ans 中含有时间、含有具体文件或联系方式、分点回答中的其中 3 个特点} \\ 0, \text{其他} \end{cases} \quad (26)$$

$$\text{PoliteScore}(\text{ans}) = \begin{cases} 1, \text{ans 中含有问候语或者感谢语} \\ 0, \text{其他} \end{cases} \quad (27)$$

上述的 4 条式子中: CompleteScore 表示完整性评分, ans 表示答复, LengthScore 表示长度评分, DetailScore 表示详细度评分, PoliteScore 表示礼貌评分, len(ans) 表示 ans 的字数, 1 表示界定优秀的回复的字数指标。

1 这里规定的值为附件 4.xlsx 中已有的答复意见的字数平均值 360。

利用 re 库和正则表达式匹配, 可以检查答复意见中是否含有时间和日期的表达, 比如“2018 年 12 月 12 日”、“1998 年”、“10 分钟”。因此可以用如下正则表达式:

`\d+年|\.日|\d+分钟|\d+小时|\d+秒|[今昨前]天|\.月|\.年|[早中午晚]|\d{4}\W`

用以匹配时间和日期的表达。

详细的答复意见中会提及具体文件、电话信息, 例如“楚财建〔2016〕74 号”、“《B 市知识产权战略推进专项资金管理办法》”、“文件”、“致电”、“拨打”。可用以下正则表达式匹配这些信息:

`\d+号|[《》]|办法|文件|拨打|致电`

分点分段回答则可以用以下正则表达式匹配“一、”“1、”等形式：

$\backslash d+[,、.]|一[,、.]$

问候语、感谢语用以下正则表达式匹配：

$[您你]好|尊敬|[感谢]谢$

### 2.3.2.3 可解释性评价

意见答复的可解释性体现在回复内容与留言内容的一致，即要让留言人明白为什么这样回复。可解释性高的答复能够提高市民对网络问政平台的信任度，进而提高用户对网络问政平台的依赖度。可解释性通常是有以下目的：

- (1) 向市民解释为什么这样答复。
- (2) 帮助市民解决提出的问题。
- (3) 增强市民的满意度。

文本中语义信息比较丰富的词语词性一般为名词、动词。如果留言和答复的名词、动词具有高的相关度，那么这样的答复可解释性就高。因此，本文采用 jieba 的基于 TF-IDF 权重的关键词提取功能，提取留言、答复中排名前 300 的词语，借助 jieba 词性标注功能，提取出关键词的名词、动词。对于这些词性相同的词语，采取上文所述的计算相似度的方法，计算出答复意见的可解释性评价。

$$\text{ExplainableScore}(\text{ans}) = \frac{\sum_{i=0}^n \sum_{j=0}^m \max\{\text{Sim}(w_{ik}, w_{jl})\}}{n} \quad (28)$$

式子中： $w_{ik}, w_{jl}$  分别为答复、留言中词性对应的词语， $n$  为答复关键词中名词和动词的总数， $m$  为留言关键词中名词和动词的总数，可解释性评价最高分为 1。

### 2.3.2.4 及时性评价

市民留言的时间和得到答复的时间通常有一定间隔，时间间隔的大小体现了答复的及时性，体现了网络问政平台的效率。因此，评价答复意见的时候及时性也是必须考虑的一环。

本文通过利用正则表达式抽取出留言时间、答复时间，并计算出二者的时间差，将结果添加到附件 4 的第 8 列中，并保存到 problem3 文件夹下的“附件 4\_时间间隔.xlsx”中。具体的程序文件在同一文件夹下的 time\_gap.py。

对时间间隔的大小及其出现次数分布进行分析，将数据可视化，得到下面的折线图。

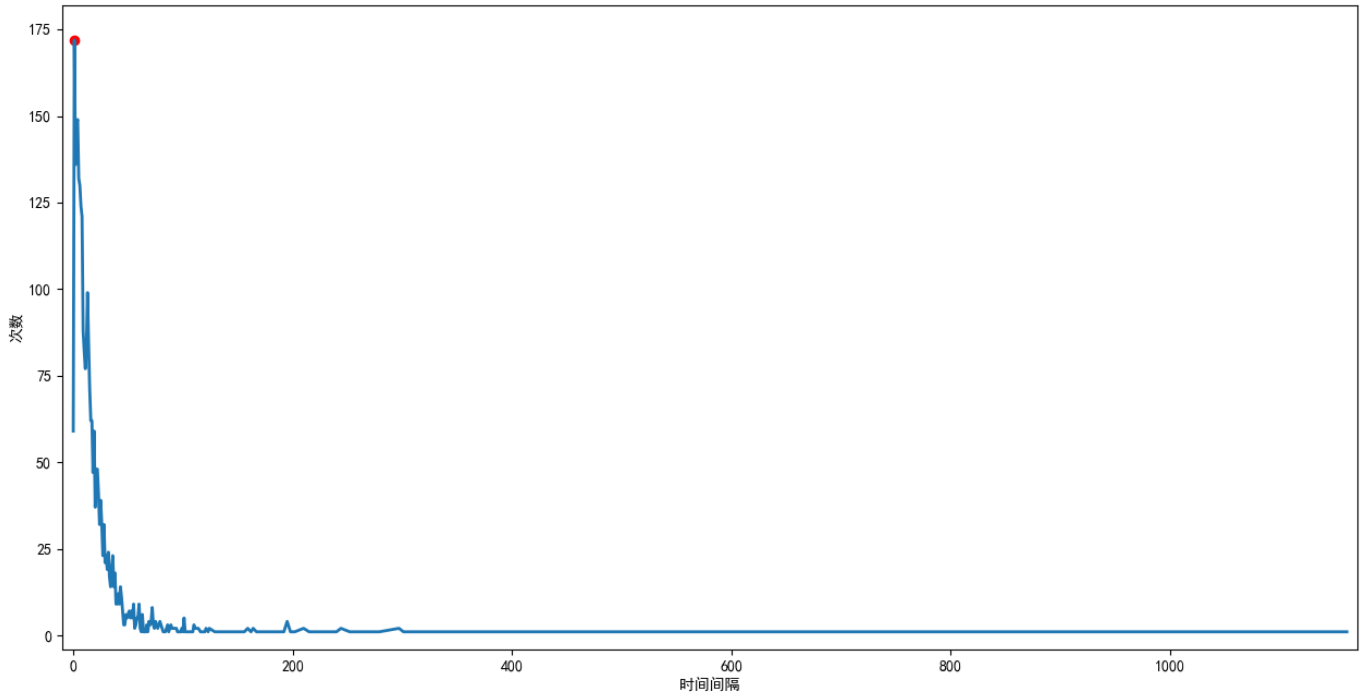


图 14 时间间隔与出现次数折线图

图的最高点为(1, 172)，即出现次数最多的时间间隔为 1 天，总共出现了 172 次。时间间隔平均值为 20.37 天，最小值是 0 天，最大值是 1161 天。大部分的时间间隔集中在 0 到 40 天之间，绝大部分的时间间隔出现在 0 到 3 天。

可以预见的，市民对于处理的及时性满意度随答复与留言之间的时间间隔的增长而降低。而一般市民能够接受的处理时长为一个星期，即 7 天。因此，就及时性评价，给出以下公式：

$$\text{TimelinessScore}(\text{gap}) = \frac{1}{\log_7(\text{gap}+2)} \quad (29)$$

式子中：gap 代表答复时间与留言时间的间隔，单位为天。

#### 2.3.2.5 总体评价

总体评价的公式为：

$$\begin{aligned} \text{TotalScore}(\text{ans}) = & \\ & a \times \text{TextSim}(\text{com}, \text{ans}) + b \times \text{CompleteScore}(\text{ans}) + c \times \text{ExplainableScore}(\text{ans}) + d \times \\ & \text{TimelinessScore}(\text{gap}) \end{aligned} \quad (30)$$

式子中：com 表示留言主题和留言详情的内容，ans 表示答复内容，a, b, c, d 为 4 个指标的权重，它们应满足  $a+b+c+d=1$  的条件。本文中经过多次试验，得出若要比客观的答复意见评分，它们的值应分别规定为：a=0.3, b=0.3, c=0.2, d=0.2。

总体评价 TotalScore 的值越大，说明答复越优秀，市民越有可能对这个结果表示满意。

计算总体评价的程序在 problem3 文件夹下的 Evaluate.py。部分的结果保存在附件 4\_部分结果.xlsx 文件中。



### 三、结果分析

#### 3.1 问题 1 结果分析

##### 3.1.1 K 值的选择

本文测试了分类方法中 K 值取 1 到 20 的情况，并且计算了这些情况下的 F-Score 指标，如图所示：

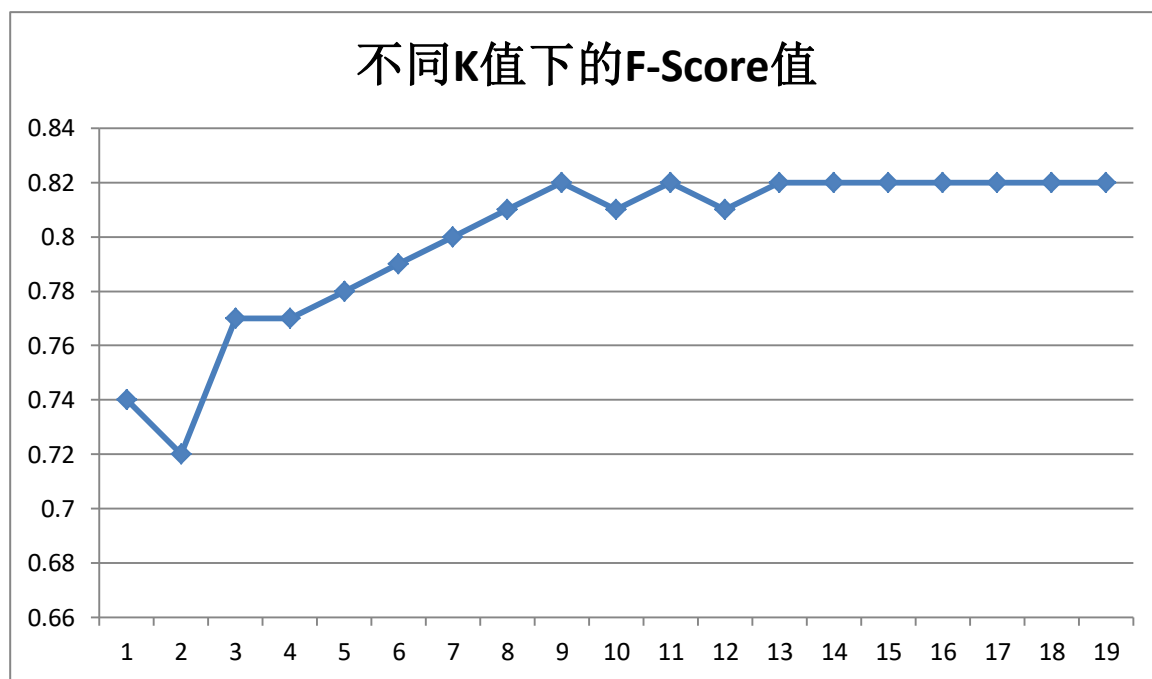


图 15 不同 K 值下的 F-Score 值

从折线图可以看出：F-Score 总体与 K 值成正相关关系。尽管在 K=9 和 K=11 时，F-Score 处于最高点，但却并不稳定，在 K=10 和 K=12 时都有浮动。但在  $K \geq 14$  时，曲线呈现出一条平行线，代表随着 K 值的增长 F-Score 趋于平稳。因此可以从实验中得知，K=14 时，分类的方法准确度、稳定度都比较可观，而且随着 K 值的增长算法的时间复杂度会增加，因此综合考虑正确率、效率以后可知，K 值应取 14。

##### 3.1.2 分类方法的预测结果

利用 KNN 算法对测试集进行预测，部分结果如下图：

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签	预测标签
24	A0007401	A市西湖建	2020/1/6	A3区大道	城乡建设	城乡建设
37	U0008473	A市在水一	2020/1/4	位于书院	城乡建设	城乡建设
83	A0006399	投诉A市A	2019/12/3	尊敬的领	城乡建设	城乡建设
303	U0007137	A1区蔡锷	2019/12/6	A1区A2区	城乡建设	城乡建设
319	U0007137	A1区A2区	2019/12/3	A1区A2区	城乡建设	城乡建设
379	A0001677	投诉A市盛	2019/11/2	我在2015	城乡建设	城乡建设
382	U0005806	咨询A市楼	2019/11/2	由于西地	城乡建设	城乡建设
445	A0001920	A3区桐梓	2019/11/1	尊敬的胡	城乡建设	城乡建设
476	U0003167	反映C4市	2019/11/1	我们是梅	城乡建设	城乡建设
530	U0008488	A3区魏家	2019/11/1	尊敬的A市	城乡建设	城乡建设
532	U0008488	A市魏家坡	2019/11/1	尊敬的A市	城乡建设	城乡建设
673	A0008064	A2区泰华	2019/10/2	请求依法	城乡建设	城乡建设
994	U0005196	A3区梅溪	2019/9/18	我住在梅	城乡建设	城乡建设
1005	U0006509	A4区鸿涛	2019/9/18	尊敬的领	城乡建设	城乡建设
1110	A0009977	地铁5号线	2019/9/9	地铁5号线	城乡建设	城乡建设
1309	U0005083	A6区润和	2019/8/21	尊敬的领	城乡建设	城乡建设
1440	A0003288	A市锦楚国	2019/8/6	A市A5区朝	城乡建设	城乡建设
1775	U0002150	给A9市城	2019/7/4	肯定是选	城乡建设	城乡建设
1783	U0004763	请A6区政	2019/7/4	尊敬的领	城乡建设	城乡建设

图 16 问题 1 部分预测结果

所有的预测的结果保存在 problem1 文件夹下的“预测结果.xlsx”。

### 3.1.3 分类结果的 F-Score

表格 11 分类方法对不同类别的 F-Score 详情

类别	查准率	查全率	F-Score	样本数
交通运输	0.84	0.66	0.74	122
劳动和社会保障	0.88	0.95	0.91	393
卫生计生	0.92	0.81	0.86	175
商贸旅游	0.82	0.73	0.77	243
城乡建设	0.73	0.82	0.77	401
教育文体	0.86	0.92	0.89	317
环境保护	0.88	0.73	0.80	187
平均值/总数	0.85	0.8	0.82	1838

不同的类别的查准率、查全率、F-Score 都如上表所示,总体的 F-Score 平均值为 0.82,劳动和社会保障类的 F-Score 最高为 0.91,最低的是交通运输的 0.74。从表中可以看出,F-Score 评价与样本数总体呈正相关,因此该分类方法比较依赖大量的数据。

对于错误的预测,取几个分类错误的例子作分析:

表格 12 分类错误的例子

留言	正确分类	算法预测
A 市大城市病之一——交通拥堵症将在三年后集中爆发	交通运输	城乡建设
楚府路快速化改造在 2016 年启动，说是 2018 年要全线通车的，2018 年还剩 10 多天了，该如何兑现呢？南二环天天堵车。要堵爆了。	交通运输	城乡建设
C 市金蕲遍远乡镇街道下雨污水严重，路面狭小，农贸市场没有建议，呼吁政府加大乡镇街道提示改造，基础设施和公益事业投入，	交通运输	劳动和社会保障

可以看出，分类错误的原因是留言中存在一定的混淆词条，而且在总的语料库中都比较常见，导致算法无法将“交通”、“通车”、“街道”等与交通运输密切相关的词条和交通运输这一分类相关联。

优化方案是通过构建自定义专业词库，分词之后遍历分词结果，与自定义专业词库进行比对，进而加重每一个分类特有的词条的权重，比如劳动和社会保障特有的“收入”、“工资”、“劳动”相关词条。

3.2 问题 2 结果分析

3.2.1 提取摘要

在求解问题描述前进行相同主题聚类，我们首先采用了层次聚类进行划分，但由于数据量大，计算复杂度较大，因而后续采取 K-means 算法进行聚类。如下表为进行聚类后的某一类别的结果：

表格 13 问题 2 聚类后某一类别的结果

A7 县龙楚社区余家珑组居民房屋前垃圾堆积危害居民健康
A 市暮云街道丽发新城社区搅拌厂危害居民健康
反应 A 市暮云街道丽发新城小区环境污染问题
A 市江山帝景新房严重安全隐患
A3 区江山帝景开发商拒承担房屋质量导致住户损失责
A 市江山帝景佛 4 期物业严重失导致业主财产重损失
A3 区江山帝景小区雅典五期 912 公交车停站时恢复
A3 区含浦学士路江山帝景小区施工造成安全隐患
A 市江山帝景新房脏乱差安全隐患
A 市江山帝景新房脏乱差安全隐患
A3 区江山帝景佛四期安全隐患
A 市江山帝景买房遇消费陷阱
A 市江山帝景新房严重安全隐患

## A 市江山帝景新房严重安全隐患

从聚类结果中仍存在小部分的文本不匹配对应主题的问题，由于该算法对异常值较为敏感，同时 K 值的不确定性，因而容易造成误差。

在使用 Text-Rank 算法对上表得到的类别文本抽取问题描述时，其结果为：“A 市江山帝景新房严重安全隐患”。

作为一种抽取式文本摘要算法，存在一定的用户留言表述不清楚等问题，其结果易受分词、文本清洗的影响。但由于其属于无监督方式，不需要构造数据集进行训练，其算法原理简单，同时利用了 TF-IDF 方法，充分利用文本之间的关系。

### 3.2.2 热度指数

热度指数的计算中，我们原本采用了熵权法进行计算，发现由于数据重复太多导致区分度并不高。

表格 14 熵值法计算热度指数

问题描述	热点指数	评论数	反对数	点赞数
请 K3 县乡村医生发卫生室执业许证	0.001996	2	0	0
A7 县东品小区业主违法圈占公绿化请查处	0.001996	1	0	0
A3 区阳烧烤城夜音响响	0.001996	7	0	0
A 市兴社区廉租房漏水导致电路火	0.001996	2	0	0
请求解决 A7 县唐田新村道路硬化问题	0.001996	4	0	0
A4 区金色溪泉湾小区业主反小区旁开设医院性质卫生	0.001996	2	0	0

所以选择了层次分析法，虽然较为主观，但是应用较简单。在实际中有点赞数十分高的评论，有可能反映该问题民众确实十分关心，也有可能是刷赞行为导致。该模型能对数据进行初步的区分，但是对于如刷赞行为等较难识别。

表格 15 AHP 计算热度指数

问题描述	热点指数	评论数	反对数	点赞数
A 市五矿万境 K9 县存严重消防安全隐患	534.18	11	0	2109
投诉 A 市 A1 区苑物业违规收停车费	447.78	6	5	1774
请书记关注 A 市 A4 区 58 车贷案	393.91	7	0	1558
请重视 A5 区砂子塘万境水岸小学路段交通安全问题	204.59	3	2	810
A7 县星沙中贸城欺诈业主退业主购房资金	178.29	8	0	693

在地点/人群的识别中，我们曾使用 pkuseg、jiagu 等库对数据进行实体识别及词性标注，由于准确度较低、速度较慢等问题而采用了 jieba 库。在该方面 jieba 库表现较好。jieba 库中我们采用了基于 HMM 和 Viterbi 算法的词性标识，通过加载停用词表和自定义词典，对数据进行了比较准确的划分。其中的缺点是需要人工筛选出一些不必要的词加入停用词表中，以减少数据的混杂。

表格 16 使用 jieba 库筛选地点/人群结果

地点/人群	问题描述
A 市五矿万境 K9 县	A 市五矿万境 K9 县存严重消防安全隐患
A 市 A1 区物业	投诉 A 市 A1 区苑物业违规收停车费
A 市 A4 区书记	请书记关注 A 市 A4 区 58 车贷案
A5 区万境小学	请重视 A5 区砂子塘万境水岸小学路段交通安全问题
A7 县星沙业主	A7 县星沙中贸城欺诈业主退业主购房资金

### 3.3 问题 3 结果分析

利用上文所述的评价方案，对附件 4 中的答复意见评价，部分结果如下表所示：

表格 17 问题 3 中部分答复意见的评分结果

答复意见	评分
网友： 您好！B 市渔船补贴标准现为 900 元/年/艘，此标准是根据《西地省财政厅、西地省畜牧水产局关于调整我省渔业捕捞和养殖业油价补贴政策促进渔业持续健康发展的通知》（楚财建〔2016〕74 号）确定的。该文件规定 2015 年-2019 年，无论渔船长度和功率大小，每艘渔船的补贴标准都为 900 元/年。 感谢您对 B 市财政工作的关注和支持！ 联系人：刘军毅 联系电话：0000-00000000	0.622177613
尊敬的网友：您好，来信收悉！根据西地省人民政府令第 179 号：第十三条 用人单位男职工的配偶生育第一胎，其配偶无工作单位的，从生育保险基金中支付一次性生育补助金，标准为统筹地区上年度平均生育医疗费用的 50%。男职工一次性生育补助金（男职工配偶无工作单位且生育第一胎，产后 6 个月至 1 年以内由单位统一办理，申领期间不得停保）办理资料： 1、生育证原件、复印件（生育证编号 1 开头的） 2、女方无工作证明（村、居委会出具）原件、复印件 3、婴儿出生证原件、复印件 4、单位盖公章的申请表 2 份（在 B 市人力资源和社会保障网下载） 5、单位盖财务专用章的收据（背面写清单位开户名、开户银行、账号、单位联系人及电话）备注：在 B 市本级参加了基本医疗、生育保险的参保职工在市本级正常连续缴费满 10 个月的次月起方可享受生育保险相关待遇。感谢来信！	0.72365013
2018 年 12 月 12 日	0.164045595
您好，你所反映的问题已转交相关单位调查处置。	0.256731409

对于不同的答复内容，由上述制定的 4 个指标构成的评价方案，可以给出基本吻合的评分。评分与答复可能的满意度总体呈现正相关关系，平均分接近 0.5，说明评价方案的制定相对是合理的。

然而，也存在着一定的问题。比如，对于一样的回复，这套评价方案可能会给出相差较大的评分：

表格 18 评分不理想的答复意见对比

答复意见	评分
您好，你所反映的问题已转交相关单位调查处置。	0.256731409
您好，你所反映的问题已转交相关单位调查处置。	0.430683438

这是由于评价方案主要依赖的相似度算法仍存在一定的缺陷。答复中的关键词语义较为笼统，与留言内容的关键词语义相似度较为接近，特别是对于长度较长的留言内容，与答复匹配出较高分数的可能性较高。

## 四、结论

本文基于 TF-IDF 算法计算出词语对于文本的重要性，结合 VSM 模型构造出文本特征向量，利用 KNN 算法对网民的留言内容进行了给出的一级标签分类预测；利用 K-Means 算法得到留言主题聚类，得到一段时间内在一定人群中发生的热点问题，再经过 TextRank 算法整合得到问题描述，进而利用层次分析法得到热度指标，对热点问题进行热度排序，再通过构建 HMM 模型与利用 Viterbi 算法，分析挖掘出热点问题的人群和地点；通过分析优秀答复内容的特点，针对答复的相关性、完整性、可解释性、及时性，按照梯度评分，最后结合四个子指标对总指标的贡献度制定权重，得到一套较为合理的评价方案。

最后得到的结果基本符合要求，但也有不足之处。例如，针对问题 3 制定的评价方案中，可解释性的评价对文本相似度或词语相似度的算法依赖较严重，无法得出更有区分度的评价。或许可以通过依存句法分析，进一步提取出答复与留言中的因果关系、逻辑联系，再给出相应的指标。我们后期会对这部分内容进行深入探讨与改进。

## 参考文献

- [1]G. Salton (1962), "Some experiments in the generation of word and document associations" *Proceeding AFIPS '62 (Fall) Proceedings of the December 4 - 6, 1962, fall joint computer conference*, pp. 234 - 250.
- [2]张宁, 贾自艳, 史忠植. 使用 KNN 算法的文本分类[J]. 计算机工, 2005(08):171-172+185.
- [3]姜园, 张朝阳, 仇佩亮, 周东方. 用于数据挖掘的聚类算法[J]. 电子与信息学报, 2005(04):655-662.
- [4]李娜娜. 基于 TextRank 的文本自动摘要研究[D]. 山东师范大学, 2019.
- [5]黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(05):856-864.
- [6]taozhijiang. chinese\_correct\_wsd[DB/OL]., 2015-08-19.
- [7]田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版), 2010, 28(06):602-608.