基于中文分词的“智慧政务”的文本挖掘应用

摘要

近年来，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的重要渠道，本论文通过对群众的留言主题、问政留言记录及相关部门对部分群众留言的答复意见，基于中文分词的方法，建议了 TF-IDF 和 F-score 模型，给出了群众留言的标签分类、热点问题挖掘、答复意见的评价等问题的解决方案。

针对问题 1，根据 Python 开发的一个中文分词模块——jieba 分词对附件 1 中的留言主题进行分词处理，并根据停用词 stop_words 过滤掉某些表达无意义的字或词，最终得到一个列表。然后列出每一个分类标签所包含的关键词，和前面分词处理过的列表进行对比，若列表元素包含关键词则贴上相应的分类标签。分类完成后利用 F-Score 对分类方法进行评价。通过分析附件所提供的留言，并带入查准率和查全率公式进行计算，最终得到每一个 P_i 和 R_i，再带入到 F-Score 模型公式，最终算出 F_1=69%.

针对问题 2，首先将附件 3 中的留言主题进行中文分词和去停用词处理。接着，对处理后的数据进行分析，采用 TF-IDF 算法将这些词语转换为向量，并利用 TF-IDF 值生成的向量，计算两两之间的余弦相似性，将相似度高于 0.5 的留言归类为同一问题。然后将总留言数较多的问题筛选出来，并进行归并。最后进行热度指数计算，得到筛选出来的每个问题的热度指数，并将热度指数排序，找出热度指数前 5 的问题，再对相关信息进行提取和汇总。

针对问题 3，利用 python 工具，对于答复的相关性，利用中文分词将留言与答复的关键词提取出来，以对比留言与答复的分词重叠频率为评价指标，根据频率的高低给出相关性高、中、低的评价；对于答复的完整性，首先创建一个列表存储“感谢”“理解”等关键词，然后利用 jieba 分词将答复的关键词全部提取出来进行一一配对，得到关键词重叠的次数，根据次数的多少给出相关性高、中、低的评价；对于答复的可解释性，首先创建一个列表，并存储“大会”“文件”“精神”等关键词，然后利用 jieba 分词将答复的关键词全部提取出来并与列表元素进行匹配，得到关键词重叠的次数，根据次数的多少给出可解释性高、中、低的评价。

关键词：中文分词；F-score 算法；TF-IDF 算法；余弦相似性

Abstract

This thesis provides solutions to the problems such as the classification of people's comments, the mining of hot issues, and the evaluation of comments.

For question 1, according to jieba, a Chinese word segmentation module developed by Python, the message subject in attachment 1 is segmented, and some meaningless words or words are filtered out according to stop_words. Then, the corresponding keywords in each message subject phrase are counted, and the frequency is counted. If the same message subject belongs to more than one tag, it is classified according to the frequency, and F-score is used to evaluate the classification method after the classification is completed.

For question 2, the message subject in annex 3 of the Chinese word segmentation and to stop processing. Then for data analysis, the TF - IDF algorithm converts the words to a vector, and using TF - IDF values generated vector, two cosine similarity calculation, will message classified as high similarity between the same problem. Then the total number of message filtering out of many questions, and to merge. Finally heat index calculation, calculation of heat index for each filter out will be ordered heat index, find out five problems before the heat index, and the information extraction and summary.

For question 3, with the python tool,the evaluation index is to compare the overlapping frequency of message and reply, and give the evaluation of high, medium and low correlation according to the frequency. For the integrity of the reply, first create a list to store the keywords such as "thanks" and "understanding", and then use jieba segmentation to extract all the keywords of the reply for one-to-one matching, get the number of overlapping keywords, according to the number of times to give a high, medium, low correlation evaluation; For the interpretability of replies, first create a list and store keywords such as "assembly", "file", "spirit", etc., then use jieba word segmentation to extract all the keywords of replies and match them with the list elements to get the number of overlapping keywords, according to the number of times to give a high correlation, in the evaluation.

Keywords:jieba segmentation,F-score,TF-IDF, Cosine Similarit

目录

一、问题分析

1.1 问题一分析

针对问题 1，根据题目给出的要求，对题目中所给的附件 2 进行一级标签分类模型的建立，其中，根据 Python 开发的一个中文分词模块——jieba 分词[1]对附件 1 中的留言主题进行分词处理，并根据停用词 stop_words 过滤掉某些表达无意义的字或词，紧接着统计出每一个留言主题短语中的相应关键词，并统计其频次，如果同一个留言主题属于多个标签，根据频次的高低对其进行标签的分类，分类完成后利用 F-Score[2]对分类方法进行评价。

1.2 问题二分析

针对问题 2，从众多留言中将相似问题识别出来，把某一时段内反映特定地点或特定人群的相似留言归类为同一问题，然后定义一个热度评价指标，对热点问题进行热度排名，找出热度指数最高的 5 个问题。

步骤 1：数据预处理，读取附件 3，将附件 3 中的留言主题进行中文分词和去停用词处理。

步骤 2：数据分析，在对附件 3 中留言主题进行数据预处理后，采用 TF-IDF 算法将这些词语转换为向量，并利用 TF-IDF 值生成的向量，两两计算余弦相似性，将相似性高的留言归类为同一问题。

步骤 3：数据筛选，将总留言数较多的问题筛选出来，并进行归并。

步骤 4：热度指数计算，计算筛选出来的每个问题的热度指数，将热度指数排序，找出热度指数前 5 的问题，并对相关信息进行提取和汇总。

1.3 问题三分析

针对问题 3，对于附件 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。

对于答复的相关性，利用 Python 开发的一个中文分词模块——jieba 分词将留言与答复的关键次提取出来，评价指标是对比留言与答复的分词重叠频率，根据频率的高低给出相关性高、中、低的评价。

对于答复的完整性：首先创建一个列表存储“感谢”“理解”“关心”“您好”“监督”“支持”“谢谢”“关注”等关键词，然后利用 jieba 分词将答复的关键词全部提取出来进行一一配对，得到关键词重叠的次数，根据次数的多少给出相关性高、中、低的评价。

对于答复的可解释性：首先创建一个列表，并存储“大会”“文件”“精神”“根据”“关注”“条例”“政府”等关键词，然后利用 jieba 分词将答复的关键词全部提取出来进行一一配对，得到关键词重叠的次数，根据次数的多少给出相关性高、中的评价。

二、问题假设

假设 1：评价体系创建的关键词列表涵盖留言与答复中所含大部分的关键词；
假设 2：假设文本向量化后余弦相似度大于 0.5 的留言皆为同一问题；
假设 3：假设每人仅一个账号，一个留言用户名。
注：文中其他未说明的假设具体参考使用处的说明

三、符号说明

表 1 符号说明表

符号	符号说明
$TF_{i,j}$	i 词的词频
$n_{i,j}$	i 词在所有留言主题中出现的次数
$\sum_k n_{k,j}$	所有留言主题包含的词数
IDF_i	i 词的逆文档频率
|D|	语料库中的文档总数
$\|j{:}t_i \in d_j\|$	包含 i 词的文档数目
H	问题 A 的热度指数
P1	针对问题 A 共 1 条留言人数
P2	针对问题 A 共 2 条留言人数
P3	针对问题 A 共 3 条留言人数
P4	针对问题 A 共 4 条留言人数

P5	针对问题 A 共 5 条留言人数
P0	针对问题 A 留言超过 5 条的人数
AP	问题 A 的所有留言的点赞数之和
AG	问题 A 的所有留言的反对数之和

注：文中其他未说明的符号具体参考使用处的说明

四、模型建立

4.1 问题一

4.1.1 数据预处理

4.1.1.1 数据描述

通过对附件 2 中的文本信息的观察，以及附件 1 中给出的内容分类三级标签体系，均为文本格式，需要将文本信息量化为数值形式，再对其进行分析。

其中，附件 2 的留言详情中有大量的空白行、句子中间出现空格，但是由于该问题只用留言主题对文本内容进行标签的分类，故不进行处理。

但是，由于附件 2 内容过多，影响运行效率，于是将留言主题所在列提出到新的 Excel 表格中进行分词等处理。

4.1.1.2 文本预处理

a.属性数值化

问题一主要根据附件 2 中的留言主题进行标签的分类，由于附件 2 源文件过大，将附件 2 中留言主题所在列单独提取出来生成的新的 Excel 表格，将附件 1 中一级标签的各个标签进行数值化：

表 2　数值化表格

一级标签	数值
城乡建设	0
党务政务	1
国土资源	2
环境保护	3
纪检监察	4

交通运输	5
经济管理	6
科技与信息产业	7
民政	8
农村农业	9
商贸旅游	10
卫生计生	11
政法	12
教育文体	13
劳动和社会保障	14

b.中文分词

由于中文分词的词与词之间没有特别明显的界限，从文本中提取词语时需要分词，使用的是 Python 中的中文分词模块——jieba 分词，对附件 2 处理后新生成的 Excel 文件的留言主题进行中文分词，利用 Python 中的切片功能对文本进行切片。

表 3 jieba 分词结果示例

jieba 分词结果示例
['{', ″″, ″″, ':', '', '', ″″, 'A', '市', '西湖', '建筑', '集团', '占', '道', '施工', '有', '安全隐患', ″″, '}']
['{', ″″, ″″, ':', '', '', ″″, 'A', '市', '在水一方', '大厦', '人为', '烂尾', '多年', '安全隐患', '严重', ″″, '}']
['{', ″″, ″″, ':', '', '', ″″, 'A3', '区', '杜鹃', '文苑', '小区', '外', '的', '非法', '汽车', '检测站', '要', '开业', '了', ″″, '}']
['{', ″″, ″″, ':', '', '', ″″, '民工', '在', 'A6', '区明发', '国际', '工地', '受伤', '工', '地方', '拒绝', '支付', '医疗费', ″″, '}']
['{', ″″, ″″, ':', '', '', ″″, 'K8', '县', '丁字街', '的', '商户', '乱', '摆摊', ″″, '}']
['{', ″″, ″″, ':', '', '', ″″, 'K8', '县', '南门', '街', '干净', '整洁', '了', '几天', '又', '是', '老', '样子', '了', ″″, '}']
['{', ″″, ″″, ':', '', '', ″″, 'K8', '县冷', '江东', '路', '蓝波', '旺', '酒店', '外墙', '装修', '无人', '施工', ″″, '}']
['{', ″″, ″″, ':', '', '', ″″, 'K8', '县', '九亿', '广场', '的', '公厕', '要', '安装', '照明灯', ″″, '}']

c.停用词过滤

在中文分词中使用 Python 中的切片功能对文本信息进行切割，得到的结果

中发现存在大量的无用的标点符号以及括号等信息，故而使用停用词库stop_words 对 jieba 分词的结果进行对无用词的过滤操作。

表 4　去停用词结果示例

去停用词结果示例
['A', '市', '西湖', '建筑', '集团', '占', '道', '施工', '有', '安全隐患']
['A', '市', '在水一方', '大厦', '人为', '烂尾', '多年', '安全隐患', '严重']
['A3', '区', '杜鹃', '文苑', '小区', '外', '的', '非法', '汽车', '检测站', '要', '开业', '了']
['民工', '在', 'A6', '区明发', '国际', '工地', '受伤', '工', '地方', '拒绝', '支付', '医疗费']
['K8', '县', '丁字街', '的', '商户', '乱', '摆摊']
['K8', '县', '南门', '街', '干净', '整洁', '了', '几天', '又', '是', '老', '样子', '了']
['K8', '县冷', '江东', '路', '蓝波', '旺', '酒店', '外墙', '装修', '无人', '施工']
['K8', '县', '九亿', '广场', '的', '公厕', '要', '安装', '照明灯']

4.1.2 评价分类方法

问题一采用 Python 中的中文分词模块——jieba 分词对所有的留言主题进行分词，后又通过停用词库对已经初步分词过的结果进行一些无意义的标点符号的过滤，然后再通过词频的计算对分级标签进行匹配得到标签分类的结果。

对该种分类方法最后的结果进行评价，评价该种分类方法主要从准确率（Accuracy）、精确率（Precision）、召回率（Recall）、P-R 曲线（Precision-Recall Curve）、F_1 Score、混淆矩阵（Confuse Matrix）等评价指标进行系列评价。

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真正例(true positive)、假正例(false positive)、真反例(true negative)、假反例(false negative)四种情形，令 TP、FP、TN、FN 分别表示其对应的样例数，则显然有 TP+FP+TN+FN=样例总数。分类结果的“混淆矩阵”如表 3 所示。

表 5　分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

查准率是针对我们预测结果而言的，它表示的是预测为正的样例中有多少是真正的正样例。定义公式为

$$P = \frac{TP}{TP + FP} \tag{1}$$

查全率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确。定义公式为

$$R = \frac{TP}{TP + FN} \tag{2}$$

得出每一类信息的 P 和 R 之后，再根据 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{3}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。假定查准率P_i和查全率R_i所占的权重相同，F_1表示的是P_i和R_i的调和平均。当F_1较高时，标签分类的模型的性能越好。

4.1.3 问题一求解

表 6　问题一求解

一级标题	P_i	R_i
城乡建设	0.65	0.58
党务政务	0.76	0.69
国土资源	0.71	0.54
环境保护	0.59	0.52
纪检监察	0.74	0.71
交通运输	0.81	0.76
经济管理	0.92	0.83
科技与信息产业	0.75	0.72
民政	0.84	0.79
农村农业	0.59	0.54
商贸旅游	0.62	0.58
卫生计生	0.66	0.62
政法	0.54	0.51
教育文体	0.75	0.66
劳动和社会保障	0.83	0.79
F_1	0.69	

如表可知，通过分析附件所提供的留言，并带入查准率和查全率公式进行计

算，最终得到P_i和R_i，再带入到 F-Score 模型公式，最终算出F_1=69%.

4.2 问题二

4.2.1 数据预处理

在对留言信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。为便于转换，先将附件 3 中的留言主题进行中文分词，这里使用 Python 的 jieba 中文分词包添加自定义的分词词典（附件：newdic1.txt）进行分词，然后利用停用词库（附件：stopword.txt）对分词后的留言数据进行去除停用词操作。(代码：附件 data processing.py)

表 7　分词结果示例

	留言主题
0	['A3 区', '一米阳光', '婚纱', '艺术摄影', '是否', '合法', '纳税', '了', '？ ']
1	['咨询', 'A6 区', '道路', '命名', '规划', '初步', '成果', '公示', '和', '城乡', '门牌', '问题']
2	['反映', 'A7 县', '春华镇', '金鼎村', '水泥路', '、 ', '自来水', '到户', '的', '问题']
3	['A2 区', '黄兴路步行街', '大古道巷', '住户', '卫生间', '粪便', '外排']
4	['A 市', 'A3 区', '中海国际社区', '三期', '与', '四期', '中间', '空地', '夜间', '施工', '噪音', '扰民']
5	['A3 区', '麓泉社区', '单方面', '改变', '麓谷明珠小区', '6', '栋', '架空层', '使用', '性质']
…	……
4323	['A 市', '经济', '学院', '强制', '学生', '实习']
4324	['A 市', '经济', '学院', '强制', '学生', '外出', '实习']
4325	['A 市', '经济', '学院', '体育', '学院', '变相', '强制', '实习']

表 8　去停用词结果示例

	留言主题
0	[’A3 区’, ’一米阳光’, ’婚纱’, ’艺术摄影’, ’合法’, ’纳税’]
1	[’咨询’, ’A6 区’, ’道路’, ’命名’, ’规划’, ’初步’, ’成果’, ’公示’, ’城乡’, ’门牌’]
2	[’A7 县’, ’春华镇’, ’金鼎村’, ’水泥路’, ’自来水’, ’到户’]
3	[’A2 区’, ’黄兴路步行街’, ’大古道巷’, ’住户’, ’卫生间’, ’粪便’, ’外排’]
4	[’A 市’, ’A3 区’, ’中海国际社区’, ’三期’, ’四期’, ’空地’, ’夜间’, ’施工’, ’噪音’, ’扰民’]

5	['A3 区', '麓泉社区', '单方面', '改变', '麓谷明珠小区', '栋', '架空层', '性质']
…	……
4323	['A 市', '经济', '学院', '强制', '学生', '实习']
4324	['A 市', '经济', '学院', '强制', '学生', '外出', '实习']
4325	['A 市', '经济', '学院', '体育', '学院', '变相', '强制', '实习']

4.2.2 数据分析

4.2.2.1 计算留言之间的相似度

a.TF-IDF 算法[5]

在对留言主题进行数据预处理后，为计算各留言的相似性，我们需要利用 TF-IDF 算法把这些词语转换为向量。TF-IDF 算法的细节如下：

第一步，计算词频。

$$TF_{i,j} = n_{i,j} \tag{2}$$

考虑到留言长短因素，为便于比较，进行词频标准化。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

第二步，计算逆文档频率。

$$IDF_i = \log \frac{|\mathrm{D}|}{(|\mathrm{j}:\mathrm{t_i} \in \mathrm{d_j}|+1)} \tag{4}$$

第三步：计算 TF-IDF。

$$TF - IDF = TF_{i,j} \times IDF_i \tag{5}$$

可以看出，TF-IDF 值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。

b.余弦相似性[6-7]

假设 A,B 是两个 n 维向量，$A = [x_1,x_2,...,x_n]$，$B = [y_1,y_2,...,y_n]$，则 A 与 B 的夹角的余弦值为：

$$cos\theta = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} = \frac{A \cdot B}{|A| \times |B|} \qquad i = 1,2,...,n. \tag{6}$$

余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似。

c.计算留言主题的两两相似性

使用 Python 将数据预处理后的的留言整理为一个留言列表，依次将列表中

的词语转换为词频矩阵、统计每个词语的 TF-IDF 值，并计算 TF-ITF 权重。然后用余弦相似性的算法计算各留言主题间的相似度，将相似度大于 0.5 的留言主题归并到一起，为同一问题。（代码：附件 xiangsi.py）

```
----这是第 46 条留言的相似性计算
----这是第 47 条留言的相似性计算
和第 64 条留言的相似性为 0.6674134327601475
和第 386 条留言的相似性为 0.9098140645249948
和第 1100 条留言的相似性为 0.7123849595364159
和第 1110 条留言的相似性为 0.9098140645249948
和第 1684 条留言的相似性为 0.6642361863743385
和第 2193 条留言的相似性为 0.82400703172019
和第 2757 条留言的相似性为 0.6342992089858905
和第 3308 条留言的相似性为 0.7473118492418885
和第 3636 条留言的相似性为 0.708526472058951
----这是第 48 条留言的相似性计算
----这是第 49 条留言的相似性计算
----这是第 50 条留言的相似性计算
和第 643 条留言的相似性为 0.503655097608374
和第 723 条留言的相似性为 0.5815385116020599
和第 1578 条留言的相似性为 0.5139373239585221
和第 2373 条留言的相似性为 0.5105809541359781
和第 3375 条留言的相似性为 0.5242850973164435
和第 3467 条留言的相似性为 0.7215763653522026
和第 4269 条留言的相似性为 0.5274494926590688
----这是第 51 条留言的相似性计算
----这是第 52 条留言的相似性计算
----这是第 53 条留言的相似性计算
```

图 1　余弦相似性计算代码运行结果示例

4.2.3 数据筛选

将包含五条及以上留言的问题筛选出来，并汇总，得到一系列具有一定热度的问题，为下一步找出热度指数最高的 5 个问题做准备。（附件：问题 0.xlsl）

4.2.4 热度指数计算

根据附件 3 中数据，结合常见热度评价标准，在我们的热度指数计算模型中主要考虑以下几个因素变量：留言条数，留言人数，点赞数，反对数。

4.2.4.1 自定义热度评价指标

（1）我们令问题 A 的热度指数初始值为 0；

（2）若有用户对问题 A 留言且是其对于问题 A 的第一条留言，则热度指数+5；

若有用户对问题 A 留言且是其对于问题 A 的第二条留言，则热度指数+9（该用户第一条留言使热度指数+5，第二条留言使热度指数+4）；

……

以此类推，若某用户针对问题 A 留言超过 5 条，则第 5 条以后留言数量增

加而热度指数不增加；

（3）若问题 A 的所有留言的点赞数之和+1，则热度指数+1；

若问题 A 的所有留言的反对数之和+1，则热度指数-1。

则根据我们自定义的热度评价体系，问题 A 的热度指数为：

$$H = 5 * P1 + 9 * P2 + 12 * P3 + 14 * P4 + 15 * (P5 + P0) + AP - AG$$

4.2.4.2 热度指数计算

将附件中“问题 0.xlsl”中的问题按照上述热度评价指标计算热度指数并排序，我们得到热度指数最高的 5 个问题。

表 9　热度指数最高的 5 个问题

热度排名	问题 ID	时间范围	地点/人群	问题描述
1	1	2019/11/2 至 2020/1/26	A 市 A2 区丽发新城小区	小区附近搅拌站粉尘和噪音污染严重
2	2	2019/7/7 至 2019/9/1	A 市伊景园滨河苑	小区捆绑销售车位
3	3	2018/11/15 至 2019/12/2	A 市人才新政引进人才	A 市人才购房租房补贴相关问题
4	4	2019/2/14 至 2019/10/11	A7 县星沙街道凉塘路	旧城改造何时才能动工
5	5	2019/7/21 至 2019/12/4	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民

表 10　热度指数计算

问题 ID	1	2	3	4	5
P1	47	38	20	9	13
P2	3	1	2	1	4
P3	0	1	0	0	0
P4	0	0	0	0	0
P5	0	0	0	0	0
P0	0	0	0	0	0
AP	48	23	28	68	18

AG	2	1	4	1	18
H	308	226	141	121	101

根据表 9、表 10，我们可以看到，热度指数排名前 5 的问题分别是 A 市 A2 区丽发新城小区的小区附近搅拌站粉尘和噪音污染严重问题、A 市伊景园滨河苑的小区车位捆绑销售问题、A 市人才新政引进人才购房租房补贴相关问题、A7 县星沙街道凉塘路旧城改造动工问题、A 市 A5 区魅力之城小区的小区临街餐饮店油烟噪音扰民问题，由我们定义的热度评价指标得到的这 5 个问题的热度指数分别为 308，226，141，121，101。

4.3 问题三

4.3.1 答复意见的评价以及评价方案

首先对于答复的相关性，利用 jieba 分词将留言与答复的关键次提取出来，评价指标是对比留言与答复的分词重叠频率，频率大于 0.02%则给出“相关性高”，频率大于 0.01%小于 0.02%给出“相关性中”，频率小于 0.01 则给出“相关性低”（代码：correlation.py）。

对于答复的完整性：首先创建一个列表存储“感谢”“理解”“关心”“您好”“监督”“支持”“谢谢”“关注”等关键词，然后利用分词将答复的关键词全部提取出来进行一一配对，得到关键词重叠的次数，若次数大于等于 4 次，则给出“完整性高”，若次数大于等于 2 次小于 4 次，则给出“完整性中”，若次数小于 2 次，则给出“完整性低”（代码：integrality.py）。

对于答复的可解释性：首先创建一个列表，并存储“大会”“文件”“精神”“根据”“关注”“条例”“政府”等关键词，然后利用分词将答复的关键词全部提取出来进行一一配对，得到关键词重叠的次数，若次数大于等于 2 次，则给出“可解释性高”，若次数小于 2 次，则给出“可解释性低”（代码：interpretability.py）。

表 11　评价方案

角度	第一步	第二步	第三步
答复的相关性	利用分词技术将留言和答复的所有关键词分出来	对比留言与答复的分词重叠频率	p>0.02%则给出“相关性高”
			0.01%<p≤0.02%给出“相关性中”
			p<0.01 则给出“相关性低”

答复的完整性	创建一个列表存储“感谢”“理解”“关心”“您好”“监督”“支持”“谢谢”“关注”等关键词	利用分词将答复的关键词全部提取出来进行一一配对	次数 n≥4，则给出“完整性高”
			次数 2≤n<4，则给出“完整性中”
			次数 n<2，则给出“完整性低”
答复的可解释性	创建一个列表，并存储“大会”“文件”“精神”“根据”“关注”“条例”“政府”等关键词	利用分词将答复的关键词全部提取出来进行一一配对	n≥2，则给出“可解释性高”
			n<2，则给出“可解释性低”

4.3.2 实现举例

按照上述问题三的评价方案，本论文从留言详情和答复意见中相应的各抽取29份文案进行测试，测试结果如下图。

留言编号	留言用户	留言详情	答复意见	相关性	完整性	可解释性
2549	A00045581	业公司却以交20万保证金，不	收取停车管理费，在业主大会结束后业委会	相关性低	完整性高	可解释性高
2554	A00023583	面的生意带来很大影响，里	需整体换填，且换填后还有三趟雨污水管	相关性高	完整性高	可解释性中
2555	A00031618	同时更是加大了教师的工作	办幼儿园聘任教职工要依法签订劳动合同，	相关性高	完整性中	可解释性高
2557	A000110735	落户A市，想买套公寓，请问	年龄35周岁以下（含），首次购房后，可分	相关性高	完整性中	可解释性高
2574	A0009233	“马坡岭小学”，原“马坡岭	保留“马坡岭”的问题。公交站点的设置需	相关性高	完整性高	可解释性中
2759	A00077538	每把泥巴冲到右边，越是上下	于您问题中没有说明卫生较差的具体路段，	相关性高	完整性高	可解释性中
2849	A000100804	为老社区惠民装电梯的规范	A市A3区人民政府办公室下发了《关于A市A3	相关性高	完整性高	可解释性高
3681	UU00812	好远，天寒地冻的跑好远，	修前期准备及设施设备采购等工作。下一步	相关性高	完整性高	可解释性高
3683	UU008792	也没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	相关性高	完整性高	可解释性高
3684	UU008687	立交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	相关性高	完整性高	可解释性中
3685	UU0082204	规划局审批通过《温室养殖	司支付一笔耕地征收补偿款给原大托村，但	相关性高	完整性高	可解释性中
3692	UU008829	区安置房地下室近两万平方	续，按长人防发[2014]7号文件要求，鄱阳	相关性高	完整性高	可解释性高
3700	UU00877	伞，大量从小区开车出去的业	分局配合进行具体选址，招标（邀标）进行	相关性高	完整性高	可解释性中
3704	UU0081480	是省相关政府部门的大力支持	映的相关警情，已由银盆岭派出所立刑事案	相关性高	完整性高	可解释性中
3713	UU0081227	小时以上！天寒地冻，其他公	正常。由于驾驶员工作时间长，劳动强度大，	相关性高	完整性高	可解释性中
3720	UU008444	址：https://baidu.com/。	的“披塘路路口两端各拆除20米中间花坛，	相关性高	完整性高	可解释性高
3727	UU0081194	便以各种理由拒绝退货，并	根据您提供的信息进行投诉信息的登记分送	相关性高	完整性高	可解释性中
3733	UU008706	称。建议在艺术中心先期借	营业。梅溪湖二期金菊路与雪松路东南角	相关性高	完整性高	可解释性中
3747	UU008201	上很早就施工，严重影响居民	查，施工单位由于需要夜间连续作业，已办	相关性高	完整性高	可解释性中
3755	UU0081681	希望可以尽快合一。让社保	以上不同机构，需三方或三方以上不同机构	相关性高	完整性高	可解释性中
3756	UU0081681	华为、苹果等手机都无法开通	具体上线时间请关注潇洒支付公司官网htt	相关性高	完整性高	可解释性中
3760	UU0081500	本农田。根据《土地管理法》	桥北组签订了土地补偿协议，并按协议达成	相关性中	完整性高	可解释性中
3762	UU0081057	直行车辆和行人通行，此路口	实施条例》第三十八条第一款第三项“红灯	相关性高	完整性高	可解释性高
3777	UU008162	，事故频发。如果8路线设立	9年1月15日您好，非常感谢您对于A市轨道	相关性高	完整性高	可解释性高
3788	UU0081604	金，是否能在A市办理商业住房	理中心不支持非本中心的缴存人以及异地商	相关性高	完整性高	可解释性中
3791	UU008694	在到A市国际会展中心非常不	2公里，已完成约800米路基，其余路段因涉	相关性高	完整性高	可解释性中
3797	UU008765	政府修A3区山景区西大门，挖	因政府投资计划调整，该项目已暂停。至	相关性高	完整性高	可解释性中
3838	UU0082119	是一个多亿好远，这笔大资金	二是村级举办的西地省洋兴置业公司。土地	相关性高	完整性高	可解释性中
3848	UU008233	店就是这样操作的。梅溪湖的	干洗店名称为A市A3区那么好干洗店，已办	相关性高	完整性高	可解释性中

图 2　评价指标举例

每一条留言和答复意见的后面都有相关性、完整性和可解释性的评价，其中答复意见的相关性普遍都是“高”，说明回答的内容和留言所提出的问题非常契合；答复的完整性普遍都是“高”，说明答复有开头和结尾，整体上是完整的；

可解释性有“高”有“中”，说明对于大部分问题的答复都有理有据，有迹可循。

五、模型评价

5.1 模型的优点

1.问题一中的 F-Score 评价模型在评价指标中为较为广泛使用的，F1-Score 模型能够较好地平衡查全率和查准率，从而更好地判断前期操作——Python 中文文本分词和停用词过滤，经过词频的计算去除掉对模型噪声较大的无用词，从而得到更好地分词结果，进一步得到更好地 F 值。

2.问题二中的模型很好了利用了所有的留言数据，将所有数据依次进行两两相似度计算，因此通过此模型能将所有相似问题归并到一起，即该模型所筛选出来的同一问题下的留言很全面。

3. 问题三中的模型利用了中文分词的优势，使得评价方案的可信度更高。

5.2 模型的缺点

1.问题二中由于留言数据过多，直接将全部留言分词后使用 TF-IDF 算法和余弦相似性进行相似度计算时，数据过多导致程度运行时间过长。

2.数据预处理时模型不能有效处理文本带有空格的问题。

3.问题三中模型的评价方案中，对于分词重叠概率以及分词相同次数的评价范围的确定不够严谨。

六、模型改进

6.1 改进步骤

6.1.1 缺点 1 的改进步骤

由于数据量过多，我们先进行一次地点的筛选，再进行相似度计算，具体步骤如下：

将所有留言数据分词后进行词频统计。（代码：附件 datafre.py）

按照词频统计中词频数量较多的特定地点分别提取数据，即依次提取留言主题中含有 A1 区，A2 区，A3 区，A4 区，A5 区，A6 区，A7 县，A8 县，A9 市的数据，进行相似度计算。

将各地点中，有 5 条以上留言的相似度大于 0.5 的留言数据提取出来，归于同一问题。

将各地提取出来留言合并到一起，与留言主题中含有 A 市的数据合并，再次进行留言相似度计算，并提取出有 5 条以上留言的相似度大于 0.5 的留言数据，归于同一问题。

将上述留言数据提取出来，与留言主题中不含 A1 区，A2 区，A3 区，A4 区，A5 区，A6 区，A7 县，A8 县，A9 市，A 市的数据合并，再进行留言相似度计算。

将相似度大于 0.5 的所有留言数据提取、归并为同一问题，则得到一系列具有一定热度的问题，再进行热度指数计算并排序，找到热度最高的五个问题并提取出相应数据。

6.1.2 缺点 2 的改进步骤

在文本预处理之前，使用 wps 软件打开 Excel 文件，将 Excel 表格的每一列的第一个格子里面的标题去掉，然后使用“文本处理”中的“去除空格”将剩下的文本中的空格全部删除并保存下来，在进行文本预处理时就可以直接导入了。

6.1.3 缺点 3 的改进步骤

将问题给出的所有数据测试一遍，并记录每一条留言及回复的分词重叠频率和分词相同次数。将分词重叠频率从高到底排序，记录前三分之一中的末尾频率和前三分之二中的末尾频率，这两个频率即为评价相关性高低的范围。

同理，对分词相同次数也用同样的方法得到评价完整性和可解释性高低的范围。

6.2 改进后的模型说明

6.2.1 缺点 1 改进后的模型说明

由于此模型先进行了一次地点的筛选，所以最终得到的热点问题下的留言数据可能有所遗漏，必要情况下可先提取筛选出来的某一热点问题下的留言数据的关键词，通过 Excel 再进行统计和筛选，得到更全面的相似留言数据，提取归并

为同一问题后再进行后续热度指数计算等操作。

6.2.2 缺点 2 改进后的模型说明

由于模型三不能处理带有空格的文本文件，所以需要人为的删除空格，步骤也很简单。删除空格之后，使用代码自动读入文本继而实现解决方案。

6.2.3 模型 3 改进后的模型说明

对所有的数据进行整体运行之后得到的评价范围更加准确合理，最后得到的评价指标的可信度也更高，但是相应的会增加一些工作量。

七、参考文献

[1] 葛日波,徐佳辉．Python中字符串切片技术在游戏开发中的应用研究[J]．大连理工大学城市学院，2017，11(4)：111-114．

[2]Raf Guns,Christina Lioma,Birger Larsen. The tipping point: F -score as a function of the number of retrieved items[J]. Information Processing and Management，2012，48(6).

[3]修驰. 适应于不同领域的中文分词方法研究与实现[D].北京工业大学,2013.

[4]秦赞. 中文分词算法的研究与实现[D].吉林大学,2016.

[5]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.

[6]张莉,夏佩佩,李凡长. 基于余弦相似性的供应商选择方法[J]. 山东大学学报(工学版),2017,47(01):1-6.

[7]王彬宇,刘文芬,胡学先,魏江宏. 基于余弦距离选取初始簇中心的文本聚类研究[J]. 计算机工程与应用,2018,54(10):11-18.