C 题

摘要

网络问政平台是当前政府了解民意，关注民生的一个重要渠道。但随着数据量不断地攀升，给以往人工进行留言划分和热点整理带来了极大挑战。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管理水平和施政效率有极大的意义。本文基于自然语言处理和文本数据挖掘技术，对网络问政平台的数据进行深度挖掘和分析，实现了群众留言分类、热点问题挖掘，并给出一套对答复意见质量的评价方案。

针对问题一，本文首先通过将附件 2 的数据进行预处理，实现数据的去重去空；然后使用中文分词工具 jieba 对附件 2 的数据进行分词以及停用词过滤，并将分词后的文本转化为词向量构建词袋模型。全部数据被划分为训练集和测试集两部分。基于大量的分析对比，本文采用对于短文本分类效果较好、精度较高的朴素贝叶斯算法对附件 2 的留言主题进行文本分类，同时我们还进行了数据增强，将多次出现且对分类没有用处的词加入到停用词表中，从而达到更好的分类效果，进一步优化了模型。最后通过 F1-score 评价方法对模型分类效果进行评价。经过多次测试，测得模型分类的平均准确率为 0.86（86%），模型评价指标 F1-score 的平均分值为 0.83。

针对问题二，首先对数据进行命名实体识别，将数据按地区进行分类。随后以地区为单位，将该地区的数据进行预处理，并使用词频-逆向文件词频（TF-IDF）计算出问题以及留言文本的词频矩阵。再利用基于奇异值分解（SVD）的 LSI 方法将其转化为奇异矩阵，根据余弦相识度计算出不同文本之间的相似度。当相似度达到设定值时归类为一个热点，根据改进的 Reddit 热度排名算法得出每个留言的热度，最后将每个热点中的留言热度求和得出排名前五的热点问题。最终得到热点问题排名的 excel 表，排名前五的热点问题与所给数据基本吻合。

针对问题三，对相关部门的留言回复进行预处理后，本文从留言回复的相关性、完整性、可解释性、答复效率四个指标进行分析建模，实现了一套对留言回复质量的完整评价方案。我们进一步基于自然语言处理和文本数据挖掘技术，将留言回复的相关性、完整性、可解释性、答复效率四个指标进行量化，实现由计算机自动对留言回复的质量进行评价，并给出评价分数。分析结果表明，当留言回复所得的评价分数越高，参照其具体的答复内容，表明其答复意见的质量确实更高。因此，该

评价方案是可操作的、可借鉴的。

另外，为了实现上述模型，本文所有程序均在 pycharm 环境下运行。

关键词：自然语言处理；朴素贝叶斯算法；TF-IDF 模型；LSI 模型

Abstract

The online political inquiry platform is an important channel for the government to understand public opinion and pay attention to people's livelihood. However, with the increasing amount of data, it has brought great challenges to the manual division of messages and the organization of hot spots in the past. Therefore, the establishment of a smart government system based on natural language processing technology is of great significance for improving the level of government management and governance efficiency. Based on natural language processing and text data mining technology, this paper deeply mines and analyzes the data of the network questioning platform, realizes the classification of mass messages, and mining of hot issues, and gives a set of evaluation schemes for the quality of reply opinions.

For problem one, this article first preprocesses the data of Annex 2 to achieve deduplication and emptying; then uses the Chinese word segmentation tool jieba to perform word segmentation and stop word filtering on the data of Annex 2 and convert the text after word segmentation Build a bag of words model for word vectors. All data is divided into training and test sets. Based on a large number of analysis and comparisons, this paper uses a simple Bayesian algorithm that has a good effect on short text classification and high accuracy to classify the text of the message topic in Annex 2, and we have also performed data enhancement, which will appear multiple times and classify Words that are not useful are added to the stop word list to achieve better classification results and further optimize the model. Finally, the model classification effect is evaluated by F1-score evaluation method. After many tests, the average accuracy of the model classification is 0.86, and the average score of the model evaluation index F1-score is 0.83.

For problem two, first identify the named entities of the data and classify the data by region. Afterwards, the data of the area is pre-processed in units of regions, and the word frequency matrix of the question and message text is calculated using the word frequency-reverse file word frequency (TF-IDF). Then the LSI method based on singular value decomposition (SVD) is used to convert it into a singular matrix, and the similarity

between different texts is calculated according to the cosine acquaintance. When the similarity reaches the set value, it is classified as a hotspot, and the hotness of each message is obtained according to the improved hotness Reddit ranking algorithm. Finally, the hotness of the messages in each hotspot is summed to obtain the top five hotspot issues. Finally, an excel sheet for ranking of hot issues is obtained. The top five hot issues are in line with the data given.

In response to question three, after preprocessing the message replies of the relevant departments, this article analyzes and models the four indicators of relevance, completeness, interpretability, and response efficiency of the message replies, and realizes a complete set of quality of message replies. Evaluation plan. Based on natural language processing and text data mining technology, we further quantified the four indicators of relevance, completeness, interpretability, and response efficiency of the message reply, so that the computer can automatically evaluate the quality of the message reply and give an evaluation. fraction. The analysis result shows that when the comment score obtained by replying to the message is higher, referring to the specific content of the reply indicates that the quality of the reply is indeed higher. Therefore, the evaluation scheme is operable and can be used for reference.

In addition, in order to achieve the above model, the programs used in this article are run in the pycharm environment.

Keywords: natural language processing; Naive Bayes algorithm; TF-IDF model; LSI model

目录

一、挖掘目标

1、根据附件 1 的一级分类标签，建立分类模型对附件 2 的留言进行分类以及对该分类模型进行评价。

2、对附件 3 的留言进行归类，提取在特定时间特定地点下，排名前五的热点问题。

3、通过对附件 4 相关部门的答复意见进行分析评价，给出一套评价方案。

二、数据预处理

2.1 数据描述

通过观察所给数据，可以发现数据量较大，且以汉语语言形式呈现，词语之间没有明显的区分标记，不能被机器直接读取。需要将文本进行分词，将文本词向量化才能进行分析，且文本中有较多的无用词等，如果不做处理，将会对后续分析造成影响，因此需要先对所要数据进行预处理。

2.2 文本预处理

2.2.1 中文分词

从文本中提取词语时需要分词，由于中文文本之间没有词与词的明显界限，难以被机器直接读取，因此本文采用 python 中的中文分词模块——jieba 分词，将附件二、三的留言以及附件四的答复意见进行切分，将非结构化的文本信息转化为计算机能够识别的结构化信息，同时，利用 jieba 分词系统进行词性标注，为后续步骤做好准备。

jieba 分词用到的算法：

1、基于前缀词典实现词图扫描，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），采用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；

2、对于未登录词，采用了基于汉字成词能力的 HMM 模型，采用 Viterbi 算法进行计算；

3、基于 Viterbi 算法的词性标注；

4、分别基于 TF-IDF 和 textrank 模型抽取关键词。

jieba 分词主要实现三个模块：1、分词 2、词性标注 3、关键词提取。

部分分词结果如图 1 所示:

```
['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']
['投诉', 'A', '市', 'A1', '区苑', '物业', '违规', '收', '停车费']
['A1', '区', '蔡锷', '南路', 'A2', '区华庭', '楼顶', '水箱', '长年', '不', '洗']
['A1', '区', 'A2', '区华庭', '自来水', '好', '大', '一股', '霉味']
['投诉', 'A', '市', '盛世', '耀凯', '小区', '物业', '无故', '停水']
['咨询', 'A', '市', '楼盘', '集中', '供暖', '一事']
['A3', '区', '桐梓', '坡', '西路', '可可', '小城', '长期', '停水', '得不到', '解决']
['反映', 'C4', '市', '收取', '城市', '垃圾处理', '费不', '平等', '的', '问题']
```

图 1 部分数据 jieba 分词结果图

以上图是还没采用停用词的分词结果，可以看出存在大量符号以及无意义的词，这将会对后续分析造成较大的影响，因此，需要采用停用词进行过滤，提高后续分析的准确率。

2.2.2 停用词过滤

为了提高机器搜索速率以及腾出更多储存空间，将很多对于分类任务没有效果的标点符号、数字以及词，如文本中的形容词、连接词、副词等，生成一个 Stop words（停用词表），在文本处理过程中，如果遇到这些词将会立即停止处理，将其剔除，减少索引量，增加索引效率，提高索引效果。同时，在实际操作时，我们结合问题本身，对数据进行增强，将影响模型效果的词加入停用词表，优化了停用词表，使用停用词过滤后的部分结果如图 2 所示：

```
['西湖', '建筑', '占', '施工', '安全隐患']
['在水一方', '大厦', '人为', '烂尾', '安全隐患']
['区苑', '物业', '违规', '停车费']
['蔡锷', '南路', '区华庭', '楼顶', '水箱', '洗']
['区华庭', '自来水', '一股', '霉味']
['盛世', '耀凯', '物业', '停水']
['咨询', '楼盘', '供暖']
['桐梓', '坡', '可可', '小城', '停水']
['收取', '城市', '垃圾处理', '费不', '平等']
```

图 2 停用词过滤后的部分结果图

三、符号说明

为方便文本数据挖掘模型建立和求解，我们定义本文的符号说明如下表 1 所示：

表 1 符号说明

符号	说明
TP	表示真正例，将正类正确预测为正类数
TN	表示真负例，将负类正确预测为负类数
FP	表示假正例，将负类错误预测为负类数
FN	表示假正例，将正类错误预测为负类数
i	表示文本中的词
j	表示语料库
A_{mxn}	表示经过奇异值分解之后的 m 行 n 列的矩阵 A
U_{m*n}	表示经过奇异值分解之后的 m 行 n 列的矩阵 U
Σ_{mxn}	表示经过奇异值分解之后的 m 行 n 列的矩阵 Σ
V^{T}_{nxn}	表示经过奇异值分解之后的 n 行 n 列的矩阵 V 的转置
A_{ij}	表示第 i 个文本的第 j 个词的特征值
U_{il}	表示第 i 个词和第 l 个词义的相关度
V_{jm}	表示第 j 个文本和第 m 个主题的相关度
Σ_{lm}	表示第 l 个词义和第 m 个主题的相关度
t	表示留言在过去某一特定时间点的时间间隔
y	表示对留言的整体看法
d	表示点赞数与反对数之差
z	表示点赞数超过反对数的数量
Score	表示热度值
Z	表示不同答复区间的分值
T	表示答复时间间隔
F	表示答复质量的综合评分值

四、群众留言分类

4.1 文本分类

如今，常用的文本分类算法有朴素贝叶斯算法、支持向量机算法、决策树、KNN 最近邻算法等，KNN 最近邻算法的原理简单，但分类精度一般，速度慢；朴素贝叶斯算法对于短文本分类的效果最好，精度很高；支持向量机算法支持线性不可分的情况，在精度上取中，综合比较，本文采用朴素贝叶斯算法进行文本分类。

经过上述对留言文本预处理后，将全部文本数据以 4：1 的比例分为训练样本和

测试样本，通过朴素贝叶斯分类将训练样本构建为分类器模型，再将测试样本放入分类器模型中，由 F1-Score 评价指标对模型分类效果进行评价，根据评价对模型算法、数据处理等方面对模型进行反馈优化。建立文本分类模型的实现流程如图 3 所示：

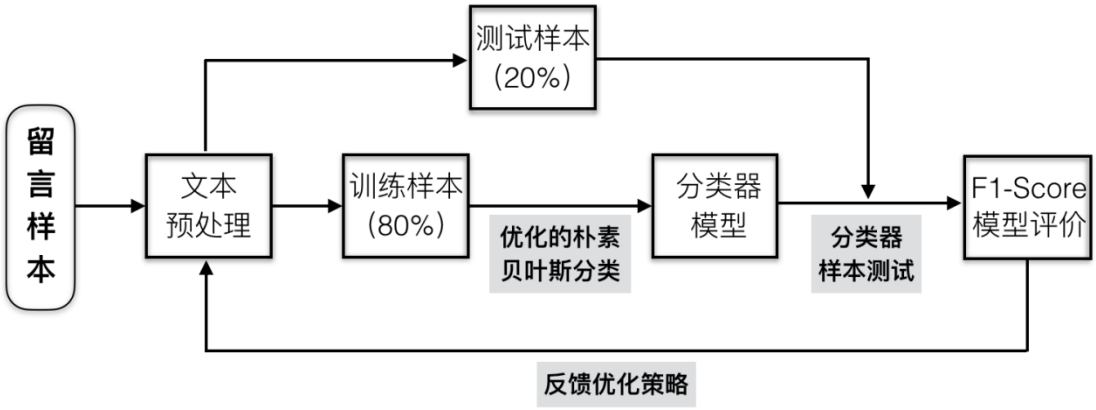

图 3 建立文本分类模型流程图

4.1.1 朴素贝叶斯算法

朴素贝叶斯分类基于条件概率、贝叶斯公式和独立性假设原则。

（1）条件概率公式

已知事件B发生的条件下事件A发生的概率称为事件A关于事件B的条件概率，记为 P(A|B)，对于任意事件 A 和 B，若 P(B)≠0，则“在事件 B 发生的条件下事件 A 发生的条件概率”，记为 P(A|B)，定义为：

$$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{P(A|B)}{P(B)} \tag{1}$$

（2）贝叶斯公式

设组$(\mathrm{A}_1, \mathrm{A}_2, \ldots \mathrm{A}_n)$是样本空间 Ω 的一个划分，B为任一事件，则有:

$$\mathrm{P}(A_i|\mathrm{B}) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n} P(A_j)\,\mathrm{P}(\mathrm{B}|A_j)} \tag{2}$$

（3）独立性假设原则

基于概率论的方法可知，当只有两种分类时:

如果$\mathrm{P1(x,y)}>\mathrm{P2(x,y)}$，那么分入类别 1；

如果$\mathrm{P1(x,y)}<\mathrm{P2(x,y)}$，那么分入类别 2。

这里的$\mathrm{P1(x,y)}$和$\mathrm{P2(x,y)}$分别表示$\mathrm{P(c1|x,y)}$和 $\mathrm{P(c2|x,y)}$。

（4）朴素贝叶斯分类的定义

设$\mathrm{x}=\{a_1,a_2,\dots,a_m\}$为一个待分类项，$a_i$为x的一个特征属性，有类别$\mathrm{C}=\{y_1,y_2,\dots,y_m\}$。

1. 计算$\mathrm{P}(\mathrm{y}_1|x),\mathrm{P}(\mathrm{y}_2|x),\dots,\mathrm{P}(\mathrm{y}_n|x)$。

2. 如果$\mathrm{P}(\mathrm{y}_1|x)=max\{\mathrm{P}(\mathrm{y}_1|x),\mathrm{P}(\mathrm{y}_2|x),\dots,\mathrm{P}(\mathrm{y}_n|x)\}$，则$\mathrm{x}\in y_k$。

对第二步中各个概率的算法可以采取以下方法：

找到一个已知分类的待分类项的集合，这个集合称为训练样本集。

统计得到各类别下各个特征属性的条件概率估计，即$\mathrm{P}(a_1|y_1),\mathrm{P}(a_2|y_1),\dots,\mathrm{P}(a_m|y_1);\dots;\mathrm{P}(a_1|y_n),\mathrm{P}(a_2|y_n),\dots,\mathrm{P}(a_m|y_n)$。

如果各个特征属性是条件独立的，则根据贝叶斯公式有如下推导：

$$\mathrm{P}(\mathrm{y}_i|\mathrm{x})=\frac{P(x|y_i)P(y_i)}{p(x)} \tag{3}$$

因为分母对所有类别相同，所以只要将分子最大化即可，又因为各特征属性是条件独立的,所以有：

$$\mathrm{P}(\mathrm{x}|y_i)\mathrm{P}(\mathrm{y}_i)=\mathrm{P}(a_1|y_i)\mathrm{P}(a_2|y_i)\dots\mathrm{P}(a_m|y_i)\mathrm{P}(y_i)=\mathrm{P}(y_i) \tag{4}$$

（5）朴素贝叶斯分类的流程

总的来说，朴素贝叶斯分为三个阶段，可用图 4 表示。

1、**准备阶段：**主要是根据具体应用情况确定特征属性，并对每个特征属性进行适当的划分，然后由人工对一部分待分类的项进行分类，形成训练样本集合。输入附件二的全部留言信息，输出的是相对应的特征属性和训练样本集合，本文根据附件一的一级分类标签给予它们不同系数进行分类，如表 2 所示。

2、**分类器训练阶段：**主要工作是计算每个类别在训练样本中的出现概率以及每个特征属性划分对每个类别的条件概率估计，并将结果记录，其输入是特征属性和训练样本，输出是分类器。

3、**应用阶段：**这个阶段主要是使用分类器对分类项进行分类，其输入是分类器和待分类项，输出的是待分类项与类别的映射关系。

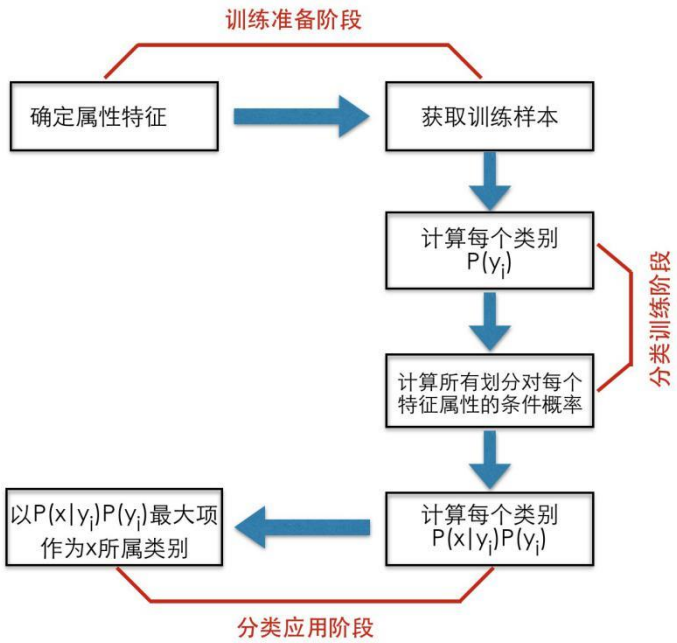

图 4 朴素贝叶斯分类的流程图

表 2 一级分类标签对应系数表

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

（6）算法改进

根据模型测试结果，考虑到可能存在一些词出现的概率为 0，那么会导致$\mathrm{P}(a_1|y_i)\mathrm{P}(a_2|y_i)\ldots\mathrm{P}(a_m|y_i)$的值也为 0，使贝叶斯分类器对文档分类时即下面式子中的分子为 0，无法比较大小，最终无法分类。

$$\mathrm{P}(\mathrm{y}_i|\mathrm{x}) = \frac{P(x|y_i)P(y_i)}{p(x)} \tag{5}$$

为了避免这种现象，经过查询相关资料以及反复探讨，本文将所有单词的出现次数由初始化为 0 改为 0.0000001，并将每个类别下每个词出现的概率分数中的分母初始化为 2.0，确保分母不为 0。另一方面，当计算乘积$\mathrm{P}(a_1|y_i)\mathrm{P}(a_2|y_i)\ldots\mathrm{P}(a_m|y_i)$

时，由于大部分因子都非常小，乘积将会变得更小，导致可能会产生下溢出或者得到不正确的答案，本文采取对乘积取自然对数的做法，在不影响结果的前提下有效解决了上述问题。

4.2 留言分类结果与评价

我们通过将原始数据分为训练数据和测试数据，并把它们转化为相应的文本词频矩阵，城乡建设对应的部分文本词频矩阵如图 5 所示：

```
[-9.092682228507956, -25.210777979466272, -25.210777979466272, -9.092682228507956]
[-25.210777979466272, -25.210777979466272, -25.210777979466272, -25.210777979466272]
[-9.092682228507956, -25.210777979466272, -25.210777979466272, -25.210777979466272]
[-8.399535097948005, -25.210777979466272, -25.210777979466272, -25.210777979466272]
[-25.210777979466272, -9.092682228507956, -8.399535097948005, -9.092682228507956]
[-25.210777979466272, -25.210777979466272, -8.399535097948005, -9.092682228507956]
[-25.210777979466272, -25.210777979466272, -25.210777979466272, -25.210777979466272]
[-25.210777979466272, -25.210777979466272, -25.210777979466272, -25.210777979466272]
[-9.092682228507956, -25.210777979466272, -25.210777979466272, -25.210777979466272]
```

图 5 城乡建设对应的部分文本词频图

其中由于上文为了优化模型对乘积采用的是自然对数，出现的次数设置为 0.0000001，得出该分类未出现的词的词频取对数为-25.184566，随着词在该分类的出现次数增多，词频的值也会增加。

然后利用训练数据来建立模型，利用测试数据来验证，本文采用的评价指标主要包括准确率（Accuracy）、F1-score 以及总体的宏平均（marco average）和加权平均（weighted average），准确率是指对于给定的测试数据集，模型正确分类的样本数占总的样本数的比值，但是在样本不均衡的情况下，得到的高准确率没有任何意义，此时准确率就会失效。

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

为了解决上述现象，我们引入了 F1-score 作为上述模型新的评价指标，它是 precision 和 recall 的调和函数，它的值越大说明该模型越好，其具体算法如下。

$$\text{precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\mathrm{F1-score}=\frac{2\times TP}{2\times TP+FP+FN} \tag{9}$$

为了更全面的评价该模型的优劣，我们还采用了宏平均和加权平均作为该模型的新指标，宏平均是指所有类别的每一个统计指标的算术平均值，加权平均数（weighted average）是以测试集中每一类的样本数作为相对应的权重。

针对每一类预测结果及预测评估的结果如表 3。

表 3 每一类的预测结果及预测评估表

	precision	**recall**	**F1-score**	**support**
城乡建设	0.76	0.89	0.82	409
环境保护	0.87	0.79	0.83	176
交通运输	0.94	0.74	0.83	125
教育文体	0.86	0.85	0.85	308
劳动和社会保障	0.82	0.93	0.87	388
商贸旅游	0.84	0.72	0.78	253
卫生计生	0.91	0.74	0.82	183
Accuracy	/	/	0.83	1842
Macro veg	0.86	0.81	0.83	1842
Weighted avg	0.84	0.83	0.83	1842

由以上结果可知，城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生它们的 F1-score 分别为 0.82，0.83，0.83，0.85，0.87，0.78，0.82，所有类别的平均 F1-score 为 0.83，参照其他指标结果可得出，此模型的 F1-score 分值相对较高。

五、热点问题挖掘

在进行热点问题挖掘前，需要用命名实体识别将留言数据按地区分类后。再将描述相似问题的留言进行识别归类为一个热点问题，本文采用 TF-IDF（词频-逆向文件频率）模型以及 LSI 模型（潜在语义索引）获取文本主题矩阵，若两文本之间的余弦相似度达到设定值将其归类。最后将全部地区的热点问题汇总，通过改进后的 Reddit 热度排名算法进行排名。热点问题挖掘的实现过程如图 6 所示：

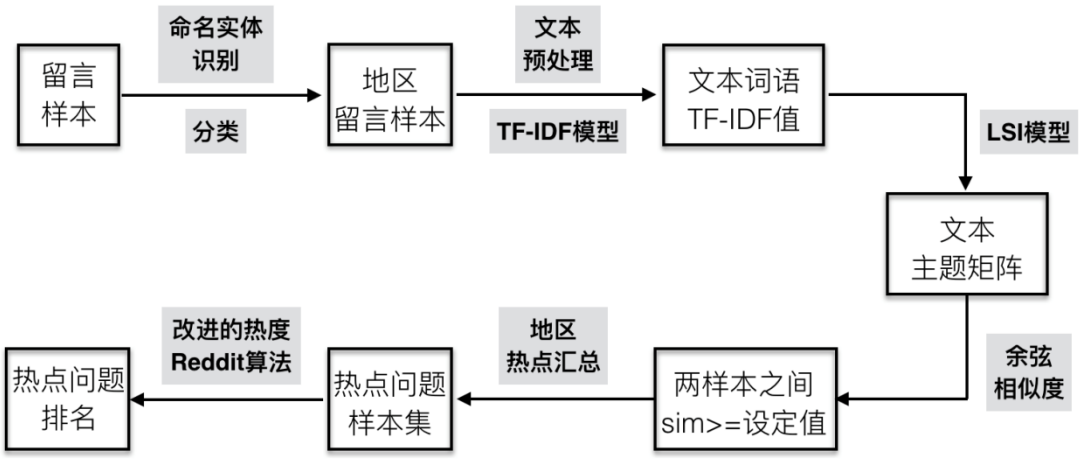

图 6 热点问题挖掘流程图

5.1 命名实体识别

由于留言信息众多，不同地区的人所反馈的问题就难免会出现相同的问题，又因后续进行要进行相似性处理，而问题是占文本的主体，所以将附件 3 的问题按地区进行分类，避免出现热点问题出现在不同地区的情况。本文采取命名实体识别，作为一个提取留言中的特定地点的有效工具，将有效提取所给数据中的地点，从而提取不同地点的热点问题提供便利。

实体是指有特定属性集合的物体，命名实体一般包括三大类（实体类、时间类、数字类）以及七小类（人名、机构名、地名、时间、日期、货币和百分比），而命名实体识别的过程主要分为：将文本进行分词→词性标注→实体识别三大步骤，而词性标注就是将实体分类。

本文利用哈工大的 Pyltp 库来完成命名实体识别，其中的 NE 识别模块的标注结果采用 O-S-B-I-E 标注形式，识别三种 NE：人名（Nh）、机构名（Ni）、地名（Ns），其余的标注为 0，最终识别的效果如图 7 所示：

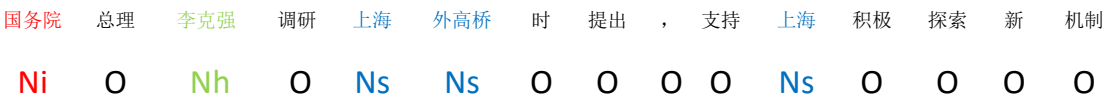

图 7 O-S-B-I-E 标注效果图

但由于哈工大的 Pyltp 库是针对中文命名实体识别，对于本次数据出现的中文英文数字合为一体的地点，如 A 市、A1 区等等，在分词阶段就把地名拆分成多个分

词无法识别，所以我们将文本数据中的英文数字替换为不常用的中文后才进行命名实体识别，如‘A’替换成‘艾’，‘1’替换成‘壹’。最后将问题数据处理后的结果如表 4：

表 4 部分数据的命名实体识别表

留言主题	地区
咨询 A6 区道路命名规划初步成果公示和城乡门牌问题	A6 区
反映 A7 县春华镇金鼎村水泥路、自来水到户的问题	A7 县 春华镇 金鼎村
A2 区黄兴路步行街大古道巷住户卫生间粪便外排	A2 区 黄兴路
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	A 市 A3 区 中海 国际
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	A3 区 麓泉
A2 区富绿新村房产的性质是什么？	A2 区 富绿 新村

5.2 热点问题识别

5.2.1 词频-逆向文件频率模型

一般来说，如果某个词在一条语料库中出现的频率较高，且在整体预料库中很少出现，则该条语料与其他的预料区分开的能力，适宜作为关键词。词频-逆向文件频率模型（Term Frequency-lnverse DocumentFrequency, 简称 TF-IDF）是基于此思想下的一种算法，由 TF（词频）和 IDF（反文档频率）组成，用以评估一个文件集或一个预料库中的其中一份文件的重要程度。

词频 TF 算法：

$$TF_{i,j} = \frac{\text{词}i\text{在语料}j\text{中的出现次数}}{\text{语料}j\text{的总词数}} \tag{10}$$

反文档词频 IDF 算法：

$$IDF_i = \log\left(\frac{\text{包含词}i\text{的语料数}}{1 + \text{语料库中语料总数}}\right) \tag{11}$$

其中，为避免 IDF 分母为 0 的情况，将其加 1，确保其可行性。

则 TF-IDF 的算法为：

$$\text{TFIDF} = \text{TF} * \text{IDF} \tag{12}$$

在基于文本的词语重要性与词语在文本中出现的位置不相关的假设下，根据上述公式可以获取文本中每个词语的 TF-IDF 值。

5.2.2 潜在语义索引模型

潜在语义索引简称 LSI，是一种基于奇异值分解（SVD）的简单使用的主题模型，并使用 TF-IDF 方法计算得出词频矩阵转化为奇异矩阵，再将词语和文本映射到一个新空间进行降维。

SVD 是指：一个m × n的矩阵A，可以分解为下面的三个矩阵：

$$A_{m*n} = U_{m\times m}\sum_{m\times n}V^{T}_{n\times n} \tag{13}$$

其中，U_{m*m}为左奇异矩阵，V_{n*n}为右奇异矩阵，Σ_{m*n}为对角矩阵，对角线上是矩阵 A 的奇异值从大到小排列，n 为矩阵 A 的秩数。

有时为了降低矩阵的维度到 k，SVD 的分解可以近似的写为：

$$A_{m*n} = U_{m\times k}\sum_{k\times k}V^{T}_{k\times n} \tag{14}$$

如果把上式用到一个主题模型，则 SVD 可以解释为我们输入 m 个留言文本，每个文本有 n 个词，而A_{ij}则对应第 i 个文本的第 j 个词的特征值。这里A_{ij}最常用的是通过预处理后的标准化 TF-IDF 值矩阵，通过 SVD 分解后，U_{il}对应第 i 个词和第 l 个词义的相关度。V_{jm}对应第 j 个文本和第 m 个主题的相关度。Σ_{lm}对应第 l 个词义和第 m 个主题的相关度。

这样我们通过一次 SVD，就可以得到文档和主题的相关度，词和词义的相关度以及词义和主题的相关度。

5.2.3 文本相似度计算

因为 LSI 模型训练是建立在 TF-IDF 之上的，所以我们先建立 TF-IDF 模型，再通过 LSI 模型得到的文本主题矩阵可以用于文本相似度计算，计算方法是通过余弦相似度。

$$\text{similarity} = \cos(\theta) = \frac{X\cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n}X_i\times Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\times\sqrt{\sum_{i=1}^{n}(Y_i)^2}} \tag{15}$$

这里的X_i，Y_i分别表示向量 X 和 Y 的各分量。

给出的相识度范围从-1 到 1：-1 意味着两个文本内容截然不同，1 表示两个文本内容完全相同，0 通常表示两个文本是独立的，而在这之间的值表示中间的相似

度和相异性。某一文本与其他文本之间的相似度如图 8 所示：

```
1.00000024e+00 -6.80078869e-04 -5.98990475e-04 -1.43953075e-04
 2.16511238e-04  8.62324450e-05 -2.26099772e-04 -2.04148309e-04
 3.36544399e-05  5.13142324e-04 -2.38513283e-04 -5.80298365e-04
-7.27598730e-04  3.39249673e-05  4.87521727e-04 -1.71024687e-04
-9.84634156e-04 -4.83377989e-05 -9.27696310e-05 -1.05838093e-03
-6.59544836e-04  6.13644021e-04 -1.20461220e-03  4.54460445e-04
-2.36254025e-04  2.36776556e-04  2.60424014e-04 -2.07335706e-05
 0.00000000e+00 -1.09798799e-03 -1.48304782e-04  9.15767159e-05
-2.47747317e-04  1.97982474e-04  6.56757446e-04 -3.91331036e-04
-4.25071805e-04  1.73580993e-04 -8.07261677e-05  1.30308836e-04
-2.00242430e-04  2.66601169e-03 -3.57141369e-04  1.53050362e-03
 1.37833843e-03  7.13660615e-04 -1.23200898e-05  2.29698286e-04
 3.97936301e-03 -4.58584764e-05 -1.12621235e-02  1.34964520e-03
-1.30423272e-04 -1.10609712e-04  1.03769603e-03 -2.30155172e-04
-3.29967501e-04 -2.87886098e-04  2.34977202e-03 -1.28574873e-04
```

图 8 某一文本与其他文本余弦相似度图

因为第一个相识度是与自身比配，故为 1，剩下的则是该文本与其他文本的相识度。

在测试的过程中发现，LSI 模型由于文本数据的数量的不同，将相似的留言问题匹配在一起形成热点所需要的相识度也不同，文本数据越多，所需要的相识度就越大。根据上文的命名实体识别，如 A 市、西地省、A1 区等出现 200 个留言问题以上，将相识度调为 0.85478 以上，相似匹配效果最佳，如魅力之城、体育经济学院等出现几十次留言问题，将相识度调为 0.0785123，相似匹配效果最佳。最后，将相识度高于某个值得留言问题归为同一热点问题。

5.3 热度评价及结果

所谓热点问题是指在某一时段内群众集中反映某个问题。因此，我们根据群众反映某个问题的数量以及群众对该问题的的看法，如点赞和反对来反映群众对该问题的关注度。利用改进后的 Reddit 热度排名算法来对每一个留言进行热度分析，然后将反映相似问题的留言的热度求和，其值即为该热点问题的总热度，最后对每个热点问题进行热度排名。

Reddit 热度排名算法如下:

$$\text{Score} = \log_{10} z + \frac{yt}{31536000} \tag{16}$$

其中：

t 指该留言在过去某一特定时间点的时间间隔，本文的过去某一特定时间点取得是全部数据中最早的时间的前一年。t 越大，得分越高，即新留言的得分超过老留言，它会自动将老留言的排名往下拉的作用。（单位：秒）

$$\mathrm{t} = t_{留} - t_{过} \tag{17}$$

y 指对留言的整体看法，如点赞数居多，表示该留言问题确实影响着很多人，y 就等于 1，如果反对数居多，表示该留言问题在其他人观点里可能没有发生。y 就等于-1，如果点赞数和反对数一样，表示该留言问题无人关注，没有影响，y 等于 0。其中 d = 点赞数 - 反对数。

$$\mathrm{y} = \begin{cases} 1 & if\ \ d > 0 \\ 0 & if\ \ d = 0 \\ -1 & if\ \ d < 0 \end{cases} \tag{18}$$

z 表示点赞数超过反对数的数量。$\log_{10} z$这部分表示，点赞数超过反对的数量越多，得分越多。这里是以 10 为底的对数，意味着 z=10 可以得到 1 分，z=100 可以得到 2 分，简单地说，就是前 10 个投票人与后 90 个投票人的权重是一样的，即如果一个留言特别受关注，那么越到后面点赞数对得分不会产生影响，而当反对数超过或等于点赞数，也就是不产生得分。

$$\mathrm{z} = \begin{cases} |d| & if\ \ |d| \geq 1 \\ 1 & if\ \ |d| \leq 1 \end{cases} \tag{19}$$

由于 Reddit 热度排名算法只考虑留言的新旧程度和留言的整体看法这两个指标，假如没人点赞或反对，则留言的发表时间对排名有很大影响，违背了群众想要政府关注此留言所反映问题的主旨，故我们对 Reddit 热度排名算法进行算法改进，将其原本权重减少，再加入一个固定值，表示一条留言出现一次，其热度加 0.6 分。得出的算法如下：

$$\text{Score} = 0.15 * \left(\log_{10} z + \frac{yt}{31536000}\right) + 0.6 \tag{20}$$

得出每个留言的热度后，将每个热点出现的留言的热度进行求和，再对总热度

进行排名，得出来的排名如表 5 所示：

表 5　热点问题排名表

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	35.04589	2019/06/25 至 2020/01/26	A 市 A2 区丽发新城小区	附近搅拌站扰民、污染环境
2	2	28.14456	2019/07/07 至 2019/09/01	A 市伊景园滨河苑	违法捆绑车位销售
3	3	12.06556	2018/11/15 至 2019/12/02	A 市	人才租房购房补贴问题
4	4	11.05008	2019/02/21 至 2019/12/28	A 市 A7 县星沙凉塘路	旧城改造什么时候可以进行
5	5	9.242182	2019/01/30 至 2019/09/16	A 市 A4 区绿地海外滩小区	距离高铁太近，严重受噪音困扰

根据表 5 可知，排名前五的热点问题都发生在近期的一年之内，热度最高的是 A 市 A2 区丽发新城小区附近的搅拌站扰民、污染环境，热度紧跟其后的是 A 市伊景园滨河苑违法捆绑车位销售，第三名到第五名的热度就很接近，分别为 A 市人才租房购房补贴问题、A 市 A7 县星沙凉塘路旧城改造什么时候可以进行、A 市 A4 区绿地海外滩小区距离高铁太近，严重受噪音困扰。

六、答复意见评价

6.1 评价方案

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，方便用户及时处理相关事务，越来越多用户通过智慧政务系统查询办理相关事务。因此，相关部门的答复意见的质量将反映出智慧政务系统能否有效解决民众所反映的问题。

本文从答复意见与相对应留言的相关性、答复意见的完整性、答复意见的可解释性以及答复效率这四个指标来对答复意见的质量进行评价，并基于自然语言处理以及文本数据挖掘技术对这四个指标进行量化，给出每条答复意见对应的四个指标分别得到的分数，并赋予四个指标一定的权重，最终得到每一条答复意见的综合评分，综合评分越高，答复意见的质量越佳。故答复意见评价的实现过程如图 9 所示：

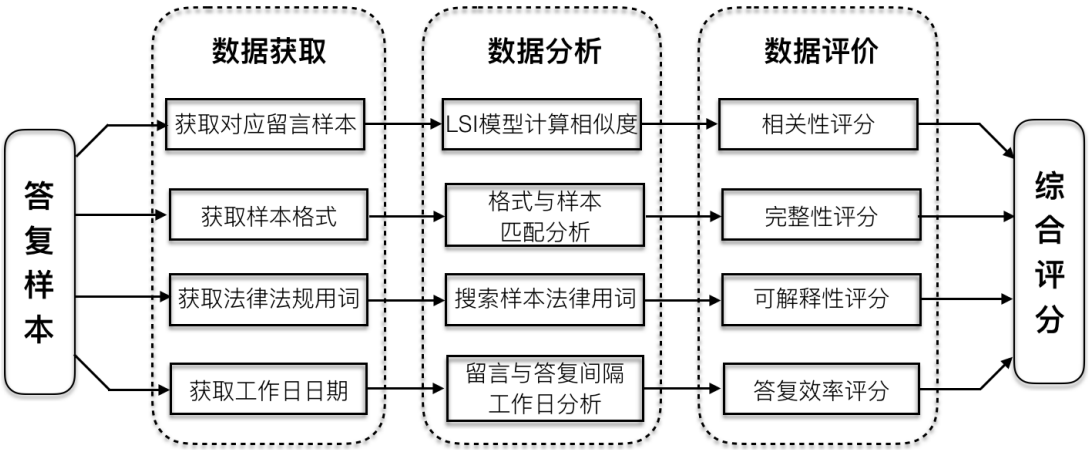

图 9 答复意见评价流程图

6.2 评价指标

6.2.1 相关性

这里的相关性是指附件四里面的留言主题和答复意见的相关程度，即将两者结合起来进行关联性分析，并在 LSI 算法的基础上，通过对留言主题和答复意见的内容进行演变分析，分析两者的文本余弦相似度，最终以一个数字呈现出来其相关性。部分回复的相关性如图 10 所示：

留言主题	留言回复	相关性
关于A市公交站点名称...	网友“A000923...	0.20073761
A3区含浦镇马路卫生...	网友“A000775...	0.45279858
A3区教师村小区盼望...	网友“A000100...	0.62277853
反映A5区东澜湾社区...	网友“UU00812...	0.18524623
反映A市美麓阳光住宅...	网友“UU00879...	0.46212846
反映A市洋湖新城和顺...	网友“UU00868...	0.5729753

图 10 部分回复相关性图

6.2.2 完整性

留言答复的完整性主要是指留言答复的规范性。通常政务系统的答复具有一定的格式规范性，因此答复的规范用语在全部数据中出现的次数会非常多，故我们将全部留言回复分词后做成词云 WordCloud，通过观察选出部分用来作为答复规范性的词语，以下图 11 为词云的效果图。

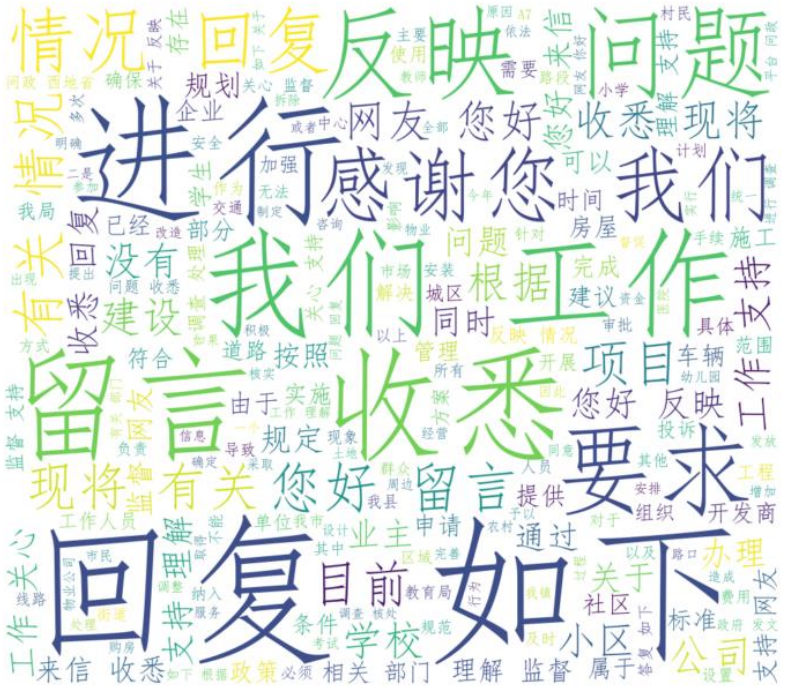

图 11 回复词云图

在得到规范性词语后，将其分为三类用语：问候用语，留言答复用语，感谢用语。如您好、网友等分为问候用语，回复、如下分为留言答复用语，感谢您、工作分为感谢用语。又由于中文博大精深，一种意思有不同的表达方式，因此，我们在分类用语中加入同义词替换的成语。然后对留言进行匹配各类用语，每个分类中的词在回复中出现一次为 1 分，最高为 3 分。其次，根据问候用语，留言回复用语，感谢用语在文章中出现的顺序，若以该顺序出现，则在给 1 分。最后得出每条回复在完整性得分。

6.2.3 可解释性

可解释性在此是指答复的内容是否有依据可言，政府部门对留言问题进行认真调查核实时，针对不同类别的问题的答复依据不同，例如对于一些具体问题，需要进行实地调查，对于一些咨询性问题，则根据一些法律法规和文件进行答复等，我们根据这些特点来对答复意见的可解释性进行评价，首先将众多的法律法规和搭配法律法规一起频繁运用的词语收纳到词袋当中，通过词袋分析识别出回复中有关法律的文本，若识别出一次，可解释性则为 1 分，识别出两次为 2 分，最高为 2 分。最后得出每条回复在可解释性得分。

6.2.4 答复效率

“为民服务”是我国政府的宗旨，“对人民负责”是我国政府的基本原则，政府效率反映一个政府的整体功能水平，而政府对留言的答复效率也包括在政府效率中。根据资料可知，一般性投诉问题在 5-15 个工作日内办结回复。因此，本文先得出用户留言与相关部门的答复时间间隔（以日数为单位），判断是否在规定工作日内答复，根据时间间隔的长短赋予不同的分值,将答复时间间隔分成五个不同的区间，不同的区间对应的分值如下：

$$Z=\begin{cases}5，\ 0\le T\le 5\\ 4,5<T\le 15\\ 3,15<T\le 30\\ 1,30<T\le 365\\ 0,T>365\end{cases} \tag{21}$$

最后结果以数字形式来表示答复效率的高低。

6.2.5 综合评分

根据相关性、完整性、可解释性、答复效率四个指标所给的分数，再给各个指标相应的权重，由于相关性反映的是相关部门的答复意见与用户留言之间的关联程度，这很大程度上决定了答复意见是否能有效解决用户反映的问题，因此，我们赋予相关性 0.5 的权重；完整性、可解释性以及答复效率这三个指标反映的是答复意见格式是否规范、可信度以及答复是否及时，因此赋予它们相对较小的权重。因此，综合评分公式如下所示：

$$\mathrm{F}=0.5\times sim_i+\frac{0.1\times x_i}{10}+\frac{0.2\times y_i}{2}+\frac{0.2\times z_i}{5} \tag{22}$$

其中$\mathrm{sim_i}$是第 i 个留言的相识度分数，$\mathrm{x_i}$是第 i 个留言的完整性分数，$\mathrm{y_i}$是第 i 个留言的可解释性分数，$\mathrm{z_i}$是第 i 个留言的答复效率分数。所有数据在不同分值区间对应的留言个数如图 12 所示。部分答复的评分情况如表 6 所示。

分析结果可得，如表 6 第一条答复所示，当答复意见与所对应的留言存在一定的关联、在文本上具有一定的格式规范性、存在可解释性的法律法规，且答复及时，故其综合评分为 0.974。又如表 6 第五条答复所示，其答复意见的内容对用于留言没有起实际帮助，且答复时间间隔过长，故其综合评分为 0.129。

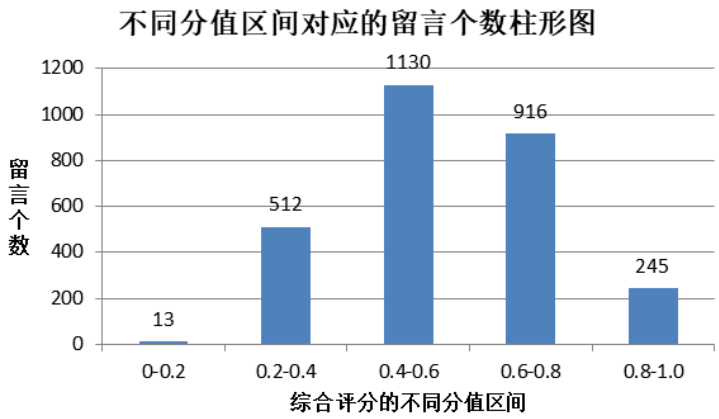

图 12 不同分值区间对应的留言个数柱形图

表 6 部分答复意见综合评分情况

留言主题	留言时间	答复意见	答复时间	综合评分
咨询 K5 县电动车...	2019/8/19	网民朋友：贵帖文已经收悉...	2019/8/23	0.974
咨询 A8 县农村医...	2016/10/26	网友“UU008824” 您好！...	2016/10/27	0.766
强烈反对 A7 县星沙...	2015/12/29	您好！首先感谢您的来信！...	2016/1/6	0.478
希望 F 市规划建师...	2019/1/29	已转市教育体育局调查处理...	2019/3/8	0.283
A5 区体育新城...	2014/7/14	网友：您好！留言已收悉	2014/8/27	0.129

七、总结

本文的主要目的是建立有效的算法，对网络问政平台上的用户留言进行分类、挖掘热点问题以及对相关部门的留言回复给予一套评价方案。总结本次比赛，本文主要运用基于自然语言处理和文本数据挖掘技术来解决“智慧政务”中的三个问题。我们在数据预处理阶段对文本进行分词，词性标注，去停用词等；我们基于朴素贝叶斯算法的文本数据挖掘技术对留言进行分类，并运用 F1-score 评价指标对所建立的模型进行评价；基于命名实体识别技术、TF-IDF 模型以及基于 LSI 模型（潜在语义索引）的奇异值分解法挖掘出民众反映的热点问题。对于相关部门对用户留言的答复意见，本文则根据其规律，深入挖掘其关联规则，从答复意见的相关性、完整性、可解释性以及答复效率四个指标给出一套答复意见的评价方案，其评价结果有效反映出答复意见的质量。

但是，本文在进行热点问题挖掘时，同一热点问题的分类结果准确度需要提高，这涉及到当今中文文本挖掘模型的不足，例如面对一些同义词，算法很难将它们识

别出来并归类在一起，后期我们会进一步对文本挖掘进行探讨；同时进行文本分类时，由于数据较少，分类效果有一定的局限性，但是本文坚信随着数据的增加，其分类效果必定会大幅改善。

八、文献

[1] 陶伟.警务应用中基于双向最大匹配法的中文分词算法实现[J].电子技术与软件工程,2016(04):153-155.

[2] 赵琳瑛. 基于隐马尔科夫模型的中文命名实体识别研究[D].西安电子科技大学,2008.

[3] 胡小娟. 基于特征选择的文本分类方法研究[D].吉林大学,2018.

[4] 王美方, 刘培玉, 朱振方. 基于 TFIDF 的特征选择方法[J]. 计算机工程与设计,2007, 28(23):5795-5796.

[5]《python 项目案例开发从入门到实战》，郑秋生，夏敏捷等，清华大学出版社

[6] https://blog.csdn.net/zkq_1986/article/details/77478682

[7] http://www.ltp-cloud.com/intro

[8] https://www.cnblogs.com/pinard/p/6805861.html