“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，利用自然语言处理和文本挖掘的方法，对社会治理创新发展以及提升政府的管理水平和施政效率具有极大的推动作用。

针对问题一：通过 mbedding 方式将每个词成映射成一个 64 维词向量，再将所有的词向量拼接构成一个二维矩阵，通过卷积操作，让 max-pooling 保持主要特征,从中选取最大值作为输出,降低参数的数目，最后将输出外接 softmax 来做 n 分类。根据预测标签以及实际标签来计算损失函数和参数需要更新的梯度, 依次更新 softmax、max-pooling、激活以及卷积核这四个函数中的参数，完成每一轮训练，得到相应的精确度和损失，最后用各类的精确率, 查全率和 f1-score 对 CNN 中文文本分类进行评价。

针对问题二：将附件 3 中的数据通过去除停用词和 jieba 分词处理后，用 gensim 基于文本建立词袋模型转换为语料库，使用 LsiModel 模型算法处理语料库，将字典映射到向量空间进行相似度计算。相似的文本归类完，运用 Reddit 的话题排名算法对话题的点赞反对数以及发布时间得出合理的热度评价指标，计算热点问题的热度排序。

针对问题三：对附件 4 相关部门对留言的答复意见，正则来删除特殊字符，并用 jieba 分词后，对答复意见进行特征词抽取，计算出答复意见和留言的相似性，同时统计出主题关键词词频和回复时间间隔以及回答长度。然后用因子分析做主成分分析，最后通过建立 SVM 模型得到对答复信息的评价。

关键词：文本分类，卷积神经网络(CNN)，Gensim，Reddit 排名算法，Jieba 分词，因子分析，SVM

Text mining application in "smart government"

Abstract

In recent years, with the development of wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms, the government has gradually become aware of public opinion It is an important channel to gather people's wisdom and gather people's spirit. The amount of text data related to various social situations and public opinions keeps increasing. Using natural language processing and text mining methods, it has a great role in promoting the innovation and development of social governance and improving the management level and efficiency of the government.

Aiming at the problem of the first: Each word is mapped into a 64 dimensional word vector by mbedding method, and then all word vectors are spliced to form a two-dimensional matrix. Through convolution operation, Max pooling keeps the main features, selects the maximum value as the output, reduces the number of parameters, and finally, the output is connected with softmax for N-classification. According to the prediction tag and the actual tag to calculate the loss function and the gradient of parameters to be updated, update the parameters of softmax, Max pooling, activation and convolution kernel in turn, complete each round of training, get the corresponding accuracy and loss, and finally evaluate the CNN Chinese text classification with all kinds of accuracy, recall and F1 score.

Aiming at the problem of the second: after removing the stop words and Jieba word segmentation from the data in Annex 3, we use gensim to build a word bag model based on the text and transform it into a corpus, use lsimodel model algorithm to process the corpus, and map the dictionary to the vector space for similarity calculation. After the similar text is classified, we use reddit's topic ranking algorithm to get a reasonable heat evaluation index for the number of likes and objections and the release time of topics, and calculate the heat ranking of hot issues.

Aiming at the problem of the third: for the reply opinions of relevant departments in Annex 4 to the message, delete special characters regularly, and use the Jieba word segmentation to extract the characteristic words of the reply opinions, calculate the similarity

between the reply opinions and the message, and at the same time count the frequency of the subject keywords, the reply time interval and the reply length. Then factor analysis is used as the principal component analysis, and finally the response information is evaluated by establishing SVM model.

Keywords: text classification, convolutional neural network (CNN), gensim, reddit ranking algorithm, Jieba segmentation, factor analysis, SVM.

目录

1、挖掘目标

本次建模目标是利用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的文本数据，利用 jieba 中文分词工具对文本进行分词、gensim 和 CNN 中文文本分类的方法以及 Reddit 的话题排名算法，达到以下 3 个目标：

1） 利用中文文本分词和文本分类的方法对留言进行分类，建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。

2） 根据附件将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，给出评价结果，并按给定的格式制作出排名前 5 的热点问题表和相应热点问题对应的留言信息的热点问题明细表，有助于及时发现热点问题，相关部门进行有针对性地处理，提升服务效率。

3） 对于附件中相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等多个角度对答复意见的质量做分析并给出一个评价。

2、总体流程图

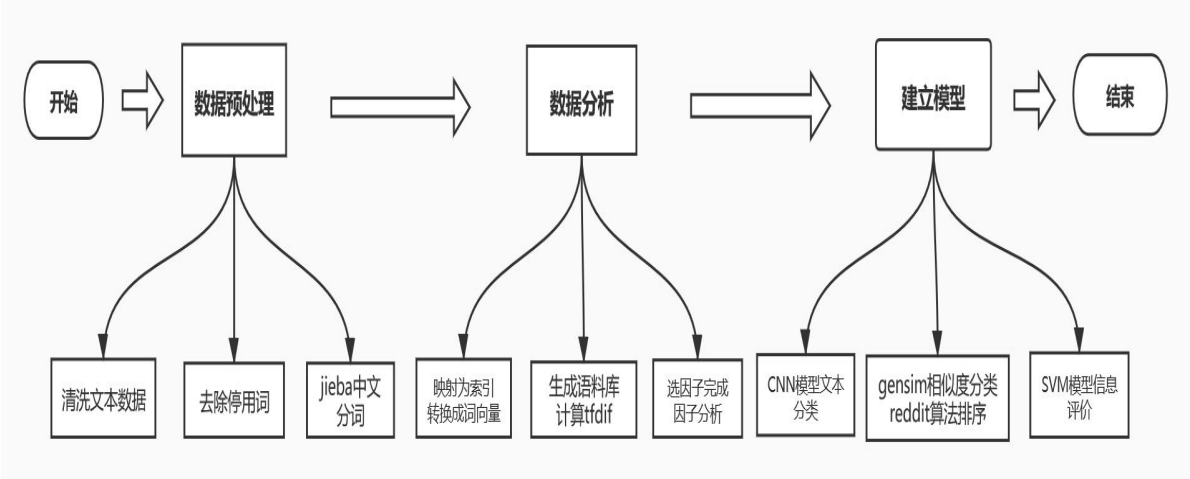


图 1：总体流程图

图 1 的总体流程图主要包括如下步骤：

步骤一：数据预处理。给出的文本数据中，出现了一些特殊字符，需要在原始的数据上进行数据清洗处理，去掉重复的信息，再将处理好的信息去除不必要的停用词并进行中文分词。

步骤二：数据分析。分词完后，对 CNN 模型要将词映射成索引表示，从预训练的词向量模型中读取出词向量，作为初始化值输入到模型中，并将数据集分割成训练集和测试集；对 gensim 模型将分词转换为语料库，并将语料库计算出 Tfidf

值。对信息的好坏做相关性、评价等操作并选择特征词，因子分析完成主成成分分析。

步骤三：建立模型。面对不同的问题和附件，对处理好的数据用 CNN 和 gensim 两种不同的模型进行分类，分别得到一级标签的分类和热点问题分类，然后计算热点问题的热度排序;建立 SVM 模型对处理好的信息给出评价。

3、分析方法与过程

3.1 问题 1 分析方法与过程

3.1.1 流程图

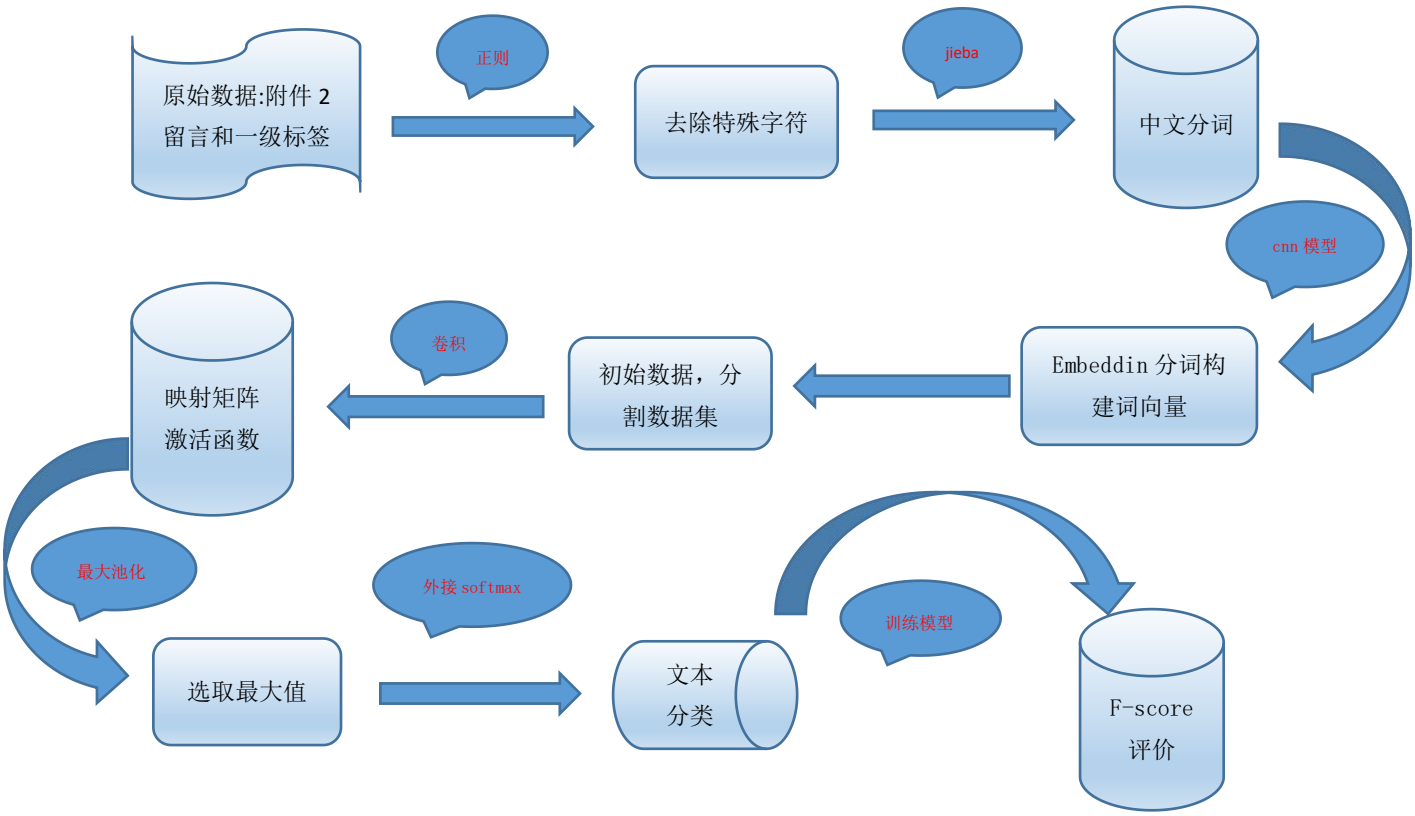


图 2：问题 1 流程图

3.1.2 数据预处理

在题目给出的附件 2 数据中，出现了一些特殊字符的文本数据。例如读取文本内容时，留言详情前面会有\n\t 等转义字符，考虑到执行分类和运行程序时产生异常错误，需要用正则表达式对附件内容的特殊字符进行数据清洗。还有一些重复的数据，也需要去除。读取通用停用词表，使那些无明确的意义的语气助词、连接词、副词和常见词“不是，一个”等词不拆分和去除的操作，避免对有效信息造成干扰，还能让所优化的关键词更集中、更突出并且节省存储空间和提高了效率。然后采用 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。

3.1.3 CNN 模型

3.1.3.1 CNN 结构

基于 TensorFlow 在中文数据集上的实现，使用了字符级 CNN 对中文文本进行分类，且 CNN 的优势为网络结构简单,使参数数目少，计算量少，训练速度快。如下图 3 所示，就是一个常见的 CNN 的基本结构。

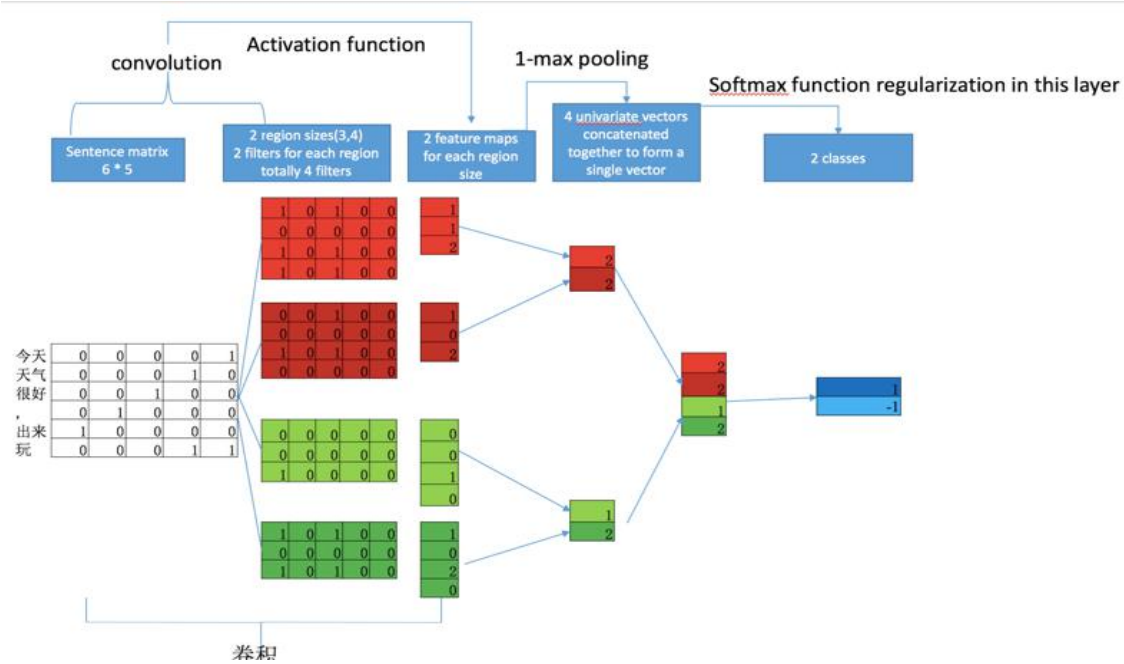


图 3：CNN 结构图

图中是一个图形识别的 CNN 模型。可以看出最左边的就是我们的输入层，计算机输入若干个矩阵，接着是卷积层（Convolution Layer）。卷积层的激活函

数使用的是 ReLU。在卷积层后面是池化层(Pooling layer)，CNN 在网络结构上没有任何变化(甚至更加简单了)，从图 3 可以看出 CNN 其实只有一层卷积,一层 max-pooling，最后将输出外接 softmax 来 n 分类，达到信息分类的作用。

3.1.3.2 卷积

首先，我们去学习卷积层的模型原理，我们需要了解什么是卷积(Convolution)，以及 CNN 中的卷积是什么样子的。学习数学时都有学过卷积的知识，微积分中卷积的表达式为：$S(t)=\int x(t-a)\omega(a)da$ (公式 3.1.1)。

离散形式是：$S(t)=\sum_{a} x(t-a)\omega(a)$ (公式 3.1.2)。

这个式子如果用矩阵表示可以为：$s(t)=(X*W)(t)$ (公式 3.1.3)。其中星号表示卷积。如果是二维的卷积，则表示式为：

$$S(i,j)=(X*W)(i,j)=\sum_{m}\sum_{n} x(i-m,j-n)\omega(m,n)$$ (公式 3.1.4)

在 CNN 中，虽然我们也是说卷积，但是我们的卷积公式和严格意义数学中的定义稍有不同,比如对于二维的卷积，定义为：

$$S(i,j)=(X*W)(i,j)=\sum_{m}\sum_{n} x(i+m,j+n)\omega(m,n)$$ (公式 3.1.5)

后面讲的 CNN 的卷积都是指的上面的最后一个式子。其中，我们叫W为我们的卷积核，而X则为我们的输入。如果X是一个二维输入的矩阵，而W也是一个二维的矩阵。但是如果X是多维张量，那么W也是一个多维的张量。

3.1.3.2 卷积层和池化层

要完成信息的分类，主要的就是卷积层和池化层，只要把卷积层和池化层的原理理解了，那么搞清楚 CNN 就容易很多了。

CNN 中的卷积，假如是对图像卷积，参考卷积公式 3.1.5，其实就是对输入的图像的不同局部的矩阵和卷积核矩阵各个位置的元素相乘，然后相加得到。举个例子，图若输入是一个二维的 3x4 的矩阵，而卷积核是一个 2x2 的矩阵。这里我们假设卷积是一次移动一个像素来卷积的，那么首先我们对输入的左上角 2x2 局部和卷积核卷积，即各个位置的元素相乘再相加，得到的输出矩阵S的 S00 的元素，值为$a\omega+bx+ey+fz$。接着我们将输入的局部向右平移一个像素，现在是

(b,c,f,g)四个元素构成的矩阵和卷积核来卷积，这样我们得到了输出矩阵 S 的 S_{01} 的元素，同样的方法，我们可以得到输出矩阵 S 的 S_{02}， S_{10}， S_{11}， S_{12} 的元素。

相比卷积层的复杂，池化层则要简单的多，所谓的池化，就是对输入张量的各个子矩阵进行压缩。假如是 2x2 的池化，那么就将子矩阵的每 2x2 个元素变成一个元素，如果是 3x3 的池化，那么就将子矩阵的每 3x3 个元素变成一个元素，这样输入矩阵的维度就变小了。要想将输入子矩阵的每 $n \times n$ 个元素变成一个元素，那么需要一个池化标准。常见的池化标准有 2 个，MAX 或者是 Average。即取对应区域的最大值或者平均值作为池化后的元素值，降低了过拟合的风险，使参数减少，进一步加速计算。

3.1.4 留言内容的一级标签分类

用 CNN 模型对附件 2 的留言内容进行一级标签分类，步骤如下：

1. 通过 embedding 方式将文本中的留言内容的每个词映射成一个 64 维的词向量，并将所有的词向量拼接起来构成一个二维矩阵，作为最初的输入。
2. 通过卷积操作，将输入的 600*64 的矩阵映射成一个 596*1 的矩阵，这个映射过程和特征抽取的结果很像，最后提取出 256 个特征。
3. 用 max-pooling 方法在保持主要特征的情况下，降低了参数的数目，从多个值中取一个最大值。
4. 将 max-pooling 的结果拼接起来，送入到 softmax 当中，得到各个类概率。
5. 根据预测标签以及实际标签来计算损失函数,通过每一轮训练数据，最后计算出每个类别的准确率、召回率和 F1 值、混淆矩阵的值。

3.2 问题 2 分析方法与过程

3.2.1 模型和算法介绍

3.2.1.1 gensim 模型

Gensim 是一个用于从文档中自动提取语义主题的 Python 库，足够智能。Gensim 可以处理原生，非结构化的数值化文本(纯文本)。Gensim 里面的算法，比如潜在语义分析 LSA，LDA，随机投影，通过在语料库的训练下检验词的统计共生模式来发现文档的语义结构。这些算法是非监督的，也就是说你只需要一个语料库的文档集。当得到这些统计模式后，任何文本都能够用语义表示来简洁的表达，并得到一个局部的相似度与其他文本区分开来。

在 gensim 模型中运用了词袋 doc2bow，LsiModel 模型算法，一下是对两种算法的简单理解：

1、**词袋模型：**词袋模型首先会进行分词，然后通过统计每个词在文本中出现的

次数，得到文本基于词的特征，若将各个文本的词与对应的词频放在一起，就是将文本完成向量化。即把一篇文本想象成一个个词构成的，所有词放入一个袋子里，没有先后顺序、没有语义。

2、LsiModel：LSI 是从文本潜在的主题来进行分析，是概率主题模型的一种，LSI 通过奇异值分解的方法计算出文本中各个主题的概率分布。假设有 5 个主题，那么通过 LSI 模型，文本向量就可以降到 5 维，每个分量表示对应主题的权重。

3、TF-IDF:表示一个词在这个文档中的重要程度。如果词 w 在一篇文档 d 中出现的频率高，并且在其他文档中很少出现，则认为词 w 具有很好的区分能力，适合用来把文章 d 和其他文章区分开来。

3.2.1.2 Reddit 的话题排名算法

Reddit 是全球化最知名的 Digg 类社区，Reddit 是一个社会化新闻类网站，Reddit 内的用户能对各个帖子以投票的方式进行赞成或反对，发布时间和票数将作为一种评价关系来决定帖子的排名。第二问的热点指数就是通 reddit 的话题排名算法基于时间和点赞反对数完成的。

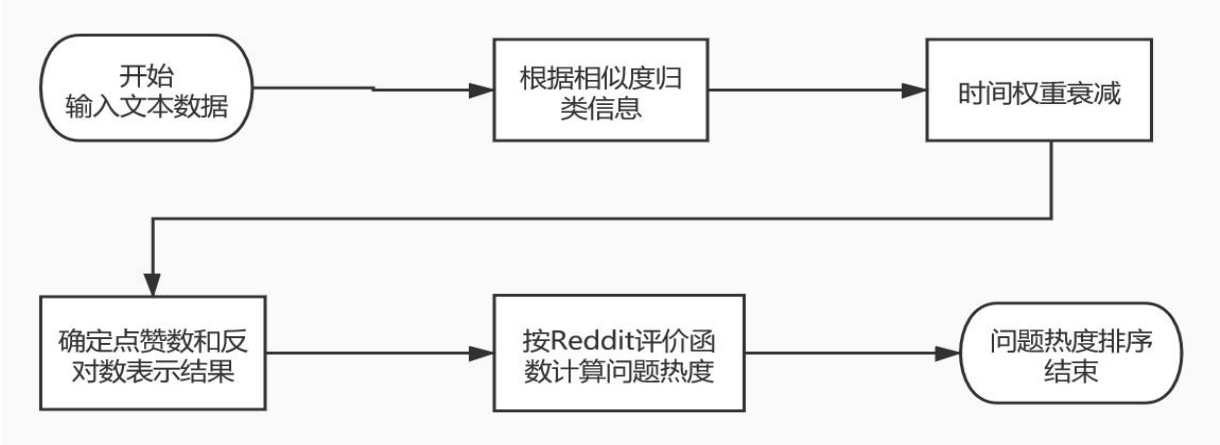


图 4：热点问题排名算法流程图

采用 Reddit 的话题排名算法对问题进话题热度排序，其算法描述如下 ：

Reddi 话题评价函数 $f(t_s,y,z)$ 数学表达式：

$$f(ts,y,z)=\log_{10}z+\frac{yts}{45000} \qquad \text{(公式 3.2.1)}$$

其中 t_s 表示差值时间，　　$t_s=A-B$　　(公式 3.2.2)

给出文档发表的时间与 2015 年 12 月 8 日 07:46:43 这个 Reddit 网站成立上线时刻之间经过的时间，并用 t_s 表示差值时间的秒数。 并且 x 是一个表示点赞数 U 和反对数 D 之间的差值：$x=U-D$ (公式 3.2.3)

在公式中，$y\in-1,0,1$ 且 y 是对 x 的符号函数值 $y=sign(x)$，即：

$$y=\begin{cases}1 & if \quad x>0\\ 0 & if \quad x=0\\ -1 & if \quad x<0\end{cases} \qquad \text{（公式 3.2.4）}$$

z 是一个限制最大优化值的限制值，在 x 的绝对值与 1 之间：

$$z=\begin{cases}|x| & if\ |x|\geq 1\\ 1 & if\ |x|\leq 1\end{cases} \qquad \text{（公式 3.2.5）}$$

其公式的具体意义可以看作为以下两点：

(1) 新话题比旧话题更受关注，因此发布时间对排名影响大

(2) 话题得分随时间衰减，且新话题评价得分更高一些

在 Reddit 的话题热度排序中使用了对数阶来限制投票反差的增长，这让算法更关注于人们的评价是否呈现两极分化，而不是考虑具体差得是不是太多，让权重的评价更加归一化。

3.2.2 问题 2 解决过程

问题 2 流程图如下图 5 所示。

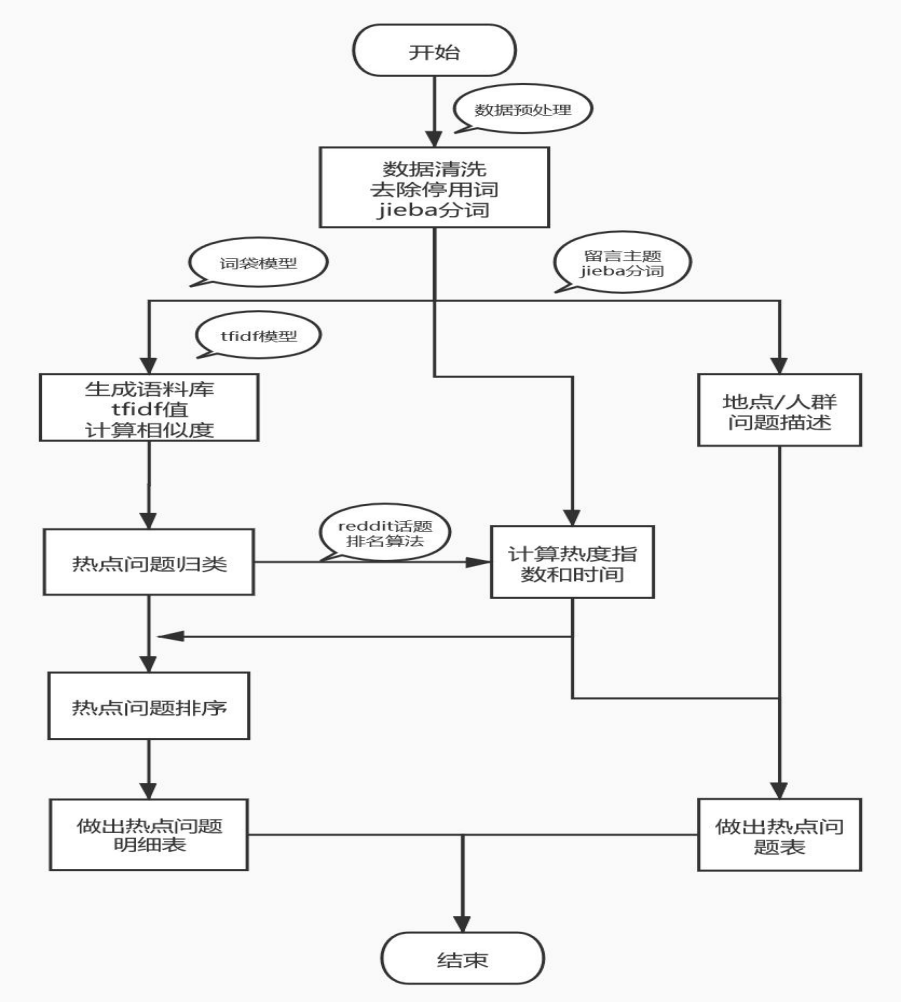


图 5：问题 2 流程图

对图 5 的流程图详解如下：
1. 对附件 3 中的数据进行数据预处理，如第一问的 2.1.2 一样，对一些含有特殊字符的、有重复的数据进行数据清洗，再去除停用词后完成 jieba 中文分词。
2. 然后就是建模完成第二问的两张表格：
（1）针对热点问题明细表，开始使用留言主题进行分词，但分类效果一般，后改进为对留言详情分词。分完词后通过词袋模型的 doc2bow 稀疏向量，形成语料库，接着运用 LsiModel 模型算法，将语料库计算出 Tfidf 值。然后获取词典 token2id 的特征数，计算稀疏矩阵相似度，建立一个索引并读取 excel 行数据，通过 jieba 进行分词处理通过 doc2bow 计算稀疏向量，求得相似度。取相似度大于 0.5 的归为一类，完成热点问题的归类。然后计算进行热点问题排序，将排好序的数据按照“问题 ID”，“留言编号”等一一写入 Excel 文件中，从而完成了第二张表格“热点问题明细表”。
（2）针对热点问题表，对留言主题进行 jieba 分词，提取特征词，将地点/人群和热点问题描述的信息逐一写入表格中。通过相似度对热点问题归类后，运用 Reddit 的话题排名算法，统计出对归类的问题的点赞数和反对数，再结合 2.2.1.2 所介绍的 reddit 排名算法，计算出每类热点问题的热度指数，用 unix 时间戳计算热点问题的时间范围时，t 的单位为秒。热度指数计算完后就可以对热点问题进行排序，最后将问题 ID，热度指数和时间范围等写入热点问题表”。

3.3 问题 3 分析方法与过程

3.3.1 评价方法介绍

通过已有的研究成果，参考文献的基础上，认为评论可以提取回答的社会性情感、准确性、完整性、相关性等方面的指标对回答进行评价。通过这些指标，挖掘出对回答评价关联程度最大的特征，并建立基于挖掘出的特征的自动化评价模型。

通过已有资料，初步筛选答案文本和实效性作为挖掘特征，采用因子分析法对各个挖掘出答复间隔、问答相关性、回答长度、主题关键词频为主要分析因子，并对挖掘出的特征进行主成分分析与关联性分析。各特征向量通过数据降维后，并通过 SVM 进行模型训练，得到评价模型。

3.3.2 信息评价过程

3.3.2.1 数据预处理

通过对附件 4 数据的研究分析，可以找到答复间隔、问答相关性、回答长度、主题关键词频、文本情感为主要的研究特征。并通过已有的模型对特征进行量化。

下表给出各个特征的量化过程。

特征	量化方法
答复间隔	留言时间与答复时间差
问答相关性	问题 1 模型进行相关性评价
回答长度	答复意见文本长度
主题关键词频率	答复意见中留言主题关键词频
文本情感	情感分析模型评分

表 1：特征量化表

通过编程实现，并对数据进行标准化处理，得到如下的特征矩阵：

	情感得分	主题关键词频	回答长度	问答相关性	答复间隔
0	0.042821	0.006250	0.059416	1.0	0.151515
1	0.039258	0.015625	0.037906	1.0	0.141414
2	0.041337	0.059375	0.046931	0.0	0.141414
3	0.044956	0.050000	0.035199	1.0	0.141414
4	0.034589	0.012500	0.018803	0.0	0.151515
5	0.033338	0.018750	0.023315	0.0	0.313131
6	0.033969	0.034375	0.027076	1.0	0.404040
7	0.066768	0.021875	0.075963	1.0	0.282828
8	0.059221	0.028125	0.062726	1.0	0.161616
9	0.037389	0.025000	0.026925	1.0	0.161616
10	0.050526	0.003125	0.057912	0.0	0.707071

图 6：特征矩阵图

3.3.2.2 模型构建

（1）皮尔逊相关性检验

需要从已有的特征矩阵进行关联性分析，分析选取的特征是否耦合，采用皮尔逊相关系数检验法。对各个特征进行相关性计算，得到特征相关性热力图。并从相关系统计算中可以看出各个成分之间耦合度，是否可以采用上述特征向量构建评价模型。

（2）因子分析

确定好研究的特征之后，还需要挖掘各个特征对结果的具体影响，采用因子分析法可以对已有的特征挖掘出影响最大的因子，然后对数据进行可视化后得到热力图。然后通过因子方差可以看出文本特征因子对整体方差贡献率，相关性因

子、时效因子次之。从整体因子分析的结果来看，判断所选取的特征能否作为构建评价自动化回答评价模型的指标。

(3)数据可视化

采用常规的高维数据可视化方法 T-SNE 可以将降维高维数据，并进行可视化，从而容易观察出数据的分布规律。对评论特征矩阵进行降维可视化.

(4)SVM 学习模型构建

通过降维操作聚合得到的训练标签，参与有监督的模型训练，得到自动化评价模型。可以将回答分为满意回答 1 与一般回答 0。

4、结果分析

4.1 问题 1 结果分析

```
Training and evaluating...
Epoch: 1
Iter:      0, Train Loss:    1.9, Train Acc:  15.62%, Val Loss:    1.9, Val Acc:  14.29%, Time: 0:00:02 *
Epoch: 2
Iter:    100, Train Loss:    1.0, Train Acc:  65.62%, Val Loss:   0.82, Val Acc:  72.14%, Time: 0:00:47 *
Epoch: 3
Epoch: 4
Iter:    200, Train Loss:   0.66, Train Acc:  73.44%, Val Loss:   0.48, Val Acc:  82.86%, Time: 0:01:31 *
Epoch: 5
Epoch: 6
Iter:    300, Train Loss:   0.19, Train Acc:  93.75%, Val Loss:   0.36, Val Acc:  87.14%, Time: 0:02:15 *
Epoch: 7
Epoch: 8
Iter:    400, Train Loss:  0.059, Train Acc:  96.88%, Val Loss:   0.38, Val Acc:  85.00%, Time: 0:02:58
Epoch: 9
Epoch: 10
Iter:    500, Train Loss:  0.048, Train Acc:  98.44%, Val Loss:   0.35, Val Acc:  86.43%, Time: 0:03:42
```

图 7：训练数据图

在建立 CNN 模型之后，对验证集上的数据进行训练，结果是在验证集上的最佳效果为 87.14%。

```
Testing...
Test Loss:   0.38, Test Acc:  86.43%
Precision, Recall and F1-Score...
               precision    recall  f1-score   support

         城乡建设       0.70      0.68      0.69        80
         环境保护       0.94      0.93      0.93        80
         交通运输       0.74      0.91      0.82        80
         教育文体       0.94      0.99      0.96        80
      劳动和社会保障       0.93      0.93      0.93        80
         商贸旅游       0.87      0.72      0.79        80
         卫生计生       0.96      0.90      0.93        80

    accuracy                           0.86       560
   macro avg       0.87      0.86      0.86       560
weighted avg       0.87      0.86      0.86       560

Confusion Matrix...
[[54  3 14  1  2  5  1]
 [ 4 74  2  0  0  0  0]
 [ 5  1 73  1  0  0  0]
 [ 1  0  0 79  0  0  0]
 [ 1  1  2  1 74  0  1]
 [12  0  7  2  0 58  1]
 [ 0  0  0  0  4  4 72]]
Time usage: 0:00:03
```

图 8：评价准确率图

如上图 8 所展的示就是对测试集中的数据进行测试，在测试集上的准确率达到了 86.43%，与验证集所训练出的准确率相差不多。对于卫生计生类、环境保护类等类别的 precision，recall 和 f1-score 都超过了 0.9，准确率还是很高的，然而像商贸旅游类、城乡建设类仅仅超过 0.7，效果较为一般。从混淆矩阵也可以看出分类效果，商贸旅游和城乡建设分类一般，其余类比较好。

对于文本未能很好的关于留言内容的一级标签分类，认为有可能的原因是：

（1）数据预处理时可能未处理的好。

（2）超参数未调节到最好。

（3）影藏网络层数较多。

（4）文本数据量少，未能更好地训练数据。

4.2 问题 2 结果分析

问题 2 是让我们根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，最后按照格式保存到“热点问题表”和“热点问题留言明细表”的 Excel 的文件里。下图 9 和下图 10 就是按照要求的格式完成的表格。

问题ID	热度指数	时间范围	地点/人群	问题描述
4	148.51015	2019-01-18 16:09:10至2020-01-02 16:41:42	A市居民	A市居民私自违章建设引发的各种问题
5	140.93824	2019-01-18 07:34:15至2020-01-03 20:08:05	A市	A市噪音扰民问题
10	139.48591	2019-01-13 10:01:58至2020-01-06 11:29:11	A市	A市公共设施建设及商店营业释放噪音影响居民生活
6	136.55726	2019-01-02 16:12:58至2019-12-30 12:07:45	A市业主	A市业主反应城市治安问题
15	131.60351	2019-01-01 10:26:56至2020-01-06 14:14:51	A市业主	A市各小区业主反应物业在小区建设中存在的各种问题

图 9：热点问题表

图 9 按照格式用 reddit 排名算法通过问题出现次数和总点赞反对数对热点问题计算出了一个热度指数，排序后并对排名前 5 的提取到表中，反映交通和基层建设，商户相关问题等 5 个热点问题在一众问题中排名前五，并且从时间范围可以看出，群众对社区产生的问题反映时间的跨度很大，如排名第一的热点问题，从 2019/01/18 至 2020/01/02，将近一年的时间仍然还存在问题，这可能代表着中途有来解决过问题，但没处理完善导致热点问题尚且存在。

通过建立热点问题表，希望相关部门利用表格清楚的了解在某一段时间内群众所遇到的问题和麻烦，能尽快的处理好这件事，为人民群众做好服务，给予百姓便利和更好的生活环境和氛围。

对于热点问题表出现的问题做以下的详解：

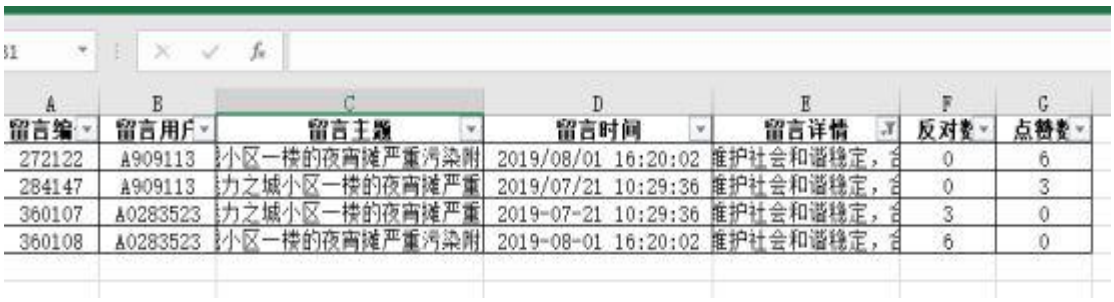
（1）地点/人群一列，并未出现人群信息，有可能存在运用 jieba 对主题分词时，主题并不存在人群信息。

（2）问题描述可能并未像题目给出的表格中将事件描述的很具体。

（3）对于问题 ID 一列，由于一开始归类是将群众留言数量的多少来排的序，再进行 Reddit 加权后，就可能不是第一了。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	[188006]	A0001029	A3区一米阳	2019/2/28	座落在A市A3	0	0
1	[191150]	A0001042	A4区大道专	2019/6/25	A市A4区大道	0	0
1	[198659]	A0005424	A市万科金	2019/2/25	尊敬的领导：	4	0
1	[201463]	A0001275	A3区莱茵城	2019/10/12	A市A3区莱茵	0	0
1	[201789]	A0002382	A市楚雅宿	2019/6/26	感谢政府201	0	0
1	[206733]	A0008049	咨询A市小	2019/11/18	尊敬的卢局长	1	0
1	[219542]	A0005160	A3区省委党	2019/3/28	A市A3区东方	0	0
1	[230618]	A0004365	A3区东方红	2019/1/9 5	您好，请问A	0	0
1	[241892]	A0002636	A5区万科金	2019/2/26	A3区东方红路	0	0
1	[242614]	A0004667	A5区左家城	2019/12/6	A市A5区万科	0	0
1	[244529]	A0001001	A3区枫林三	2019/7/25	位于A市A5区	6	0
1	[255199]	A0001622	A6区桑梓路	2019/9/20	A3区枫林三路	0	0
1	[261918]	A0006348	A3区莲花钱	2019/4/5 1	我是老A市人	0	0
1	[276992]	A0008049	咨询小孩转	2019/11/27	我是A市A3区	0	0
1	[277136]	A0005772	A3区阳光1	2019/1/30	尊敬的市委领	0	0
1	[287566]	A0007639	A3区岳北社	2019/6/25	年关降至，2	0	0
1	[287978]	A0001403	A3区科教新	2019/6/20	A市A3区岳北	0	0
1	[188007]	A0007479	咨询A6区道	2019/2/14	A市A3区科教	1	0
2	[195005]	A0004162	A6区二中弘	2019/7/12	A市A6区道路	0	0
2	[197487]	A0001201	2019年A3区	2019/8/23	20119年暑假	0	0
2	[203070]	A0001757	对A市罐子	2019/8/19	西地省公务员	2	0
2	[203115]	A0009403	希望A市能	2019/2/15	4号线地铁的	0	0
2	[234004]	A0003161	A6区银杉路	2019/3/15	尊敬的书记：	6	0

图 10：热点问题留言明细表



留言编	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
272122	A909113	小区一楼的夜宵摊严重污染附	2019/08/01 16:20:02	维护社会和谐稳定，各	0	6
284147	A909113	力之城小区一楼的夜宵摊严重	2019/07/21 10:29:36	维护社会和谐稳定，各	0	3
360107	A0283523	力之城小区一楼的夜宵摊严重	2019-07-21 10:29:36	维护社会和谐稳定，各	3	0
360108	A0283523	小区一楼的夜宵摊严重污染附	2019-08-01 16:20:02	维护社会和谐稳定，各	6	0

图 11：部分数据展示

图 10 为热点问题的留言明细表。在数据预处理时，删除特殊字符再存入表格读取后，留言详情看起来更为方便和清楚。通过图 10 的留言明细表可以清楚的知道许多人反映了某一段时间社区产生的同一个问题，群众反映的留言详情很充分的表达了问题所在以及问题带来的影响，这些问题在很长一段时间困扰和影响着他们。例如图 11 所展示的留言详情以及发送时间都相同，猜测是一个人有两个账号的情况，都向上级表述的最近一段时间所受到的影响，用两个账号来反馈事实，像此类情况希望能让上级和相关部门注意到并及时派人解决。

对于归类热点问题有一些不足之处，一个问题被归类到另一个问题中去，产生这种问题的可能有：

（1）分词后选择特征词时没有选到关键的词语，导致未能有效地归类。

（2）选择的模型可能归类效果并不是很好。

（3）数据预处理时未删除相似的且不重要的词，导致两个问题之间相似度较高从而归类错误。

4.3 问题 3 结果分析

(1)皮尔逊相关性检验结果

Pearson correlation coefficient:

	情感得分	主题关键词频	回答长度	问答相关性	答复间隔
情感得分	1.000000	0.610053	0.853926	0.065697	0.021551
主题关键词频	0.610053	1.000000	0.703110	0.137182	0.037382
回答长度	0.853926	0.703110	1.000000	0.106541	0.066769
问答相关性	0.065697	0.137182	0.106541	1.000000	0.009805
答复间隔	0.021551	0.037382	0.066769	0.009805	1.000000

图 12：皮尔逊系数图

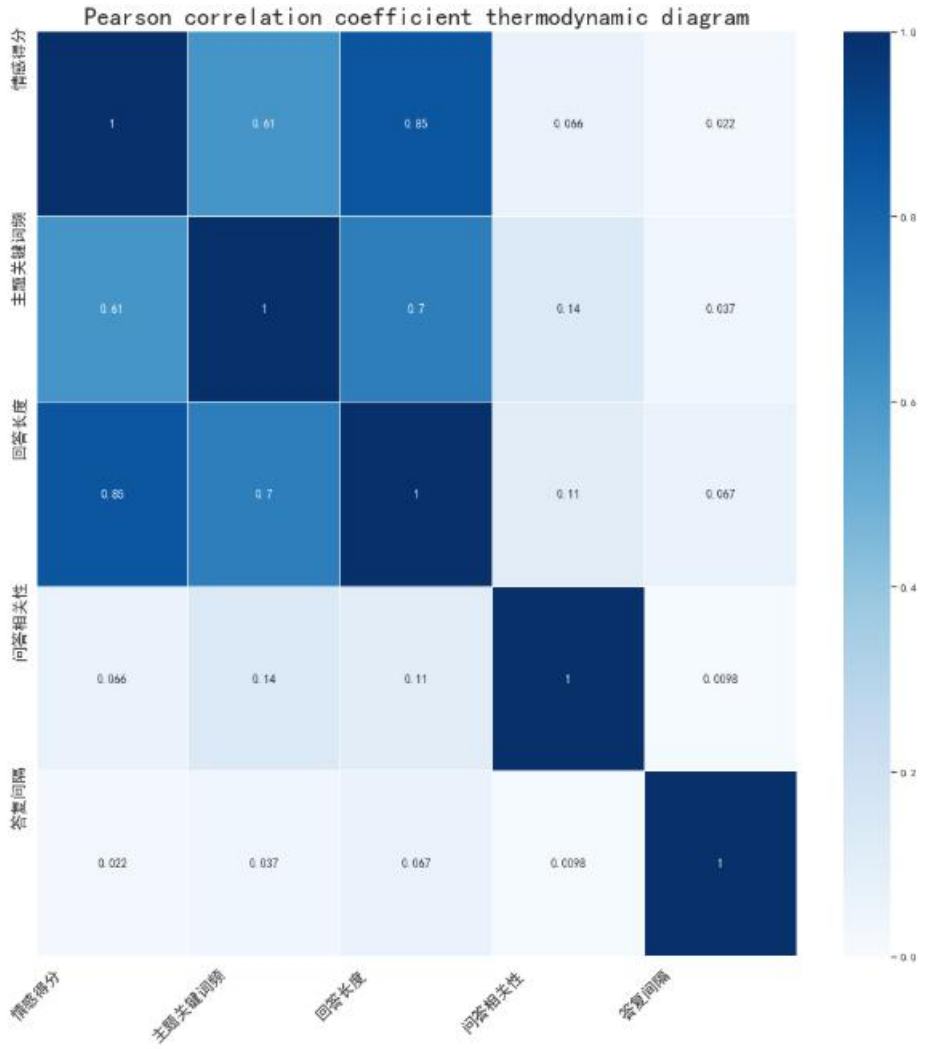


图 13：特征相关性热力图

从上图 12 和图 13 相关系统计算中可以看出各个成分之间耦合度不高，可以采用上述特征向量构建评价模型。

（2）因子分析结果

通过因子分析计算库，得到数据的负荷矩阵如下所示：

	factor3	factor2	factor1
0	0.986908	-0.155089	-0.090056
1	0.589223	0.282114	-0.009144
2	0.954177	-0.024768	0.120328
3	-0.018647	0.264818	-0.002421
4	-0.001420	-0.002260	0.196950

图 14：负荷矩阵图

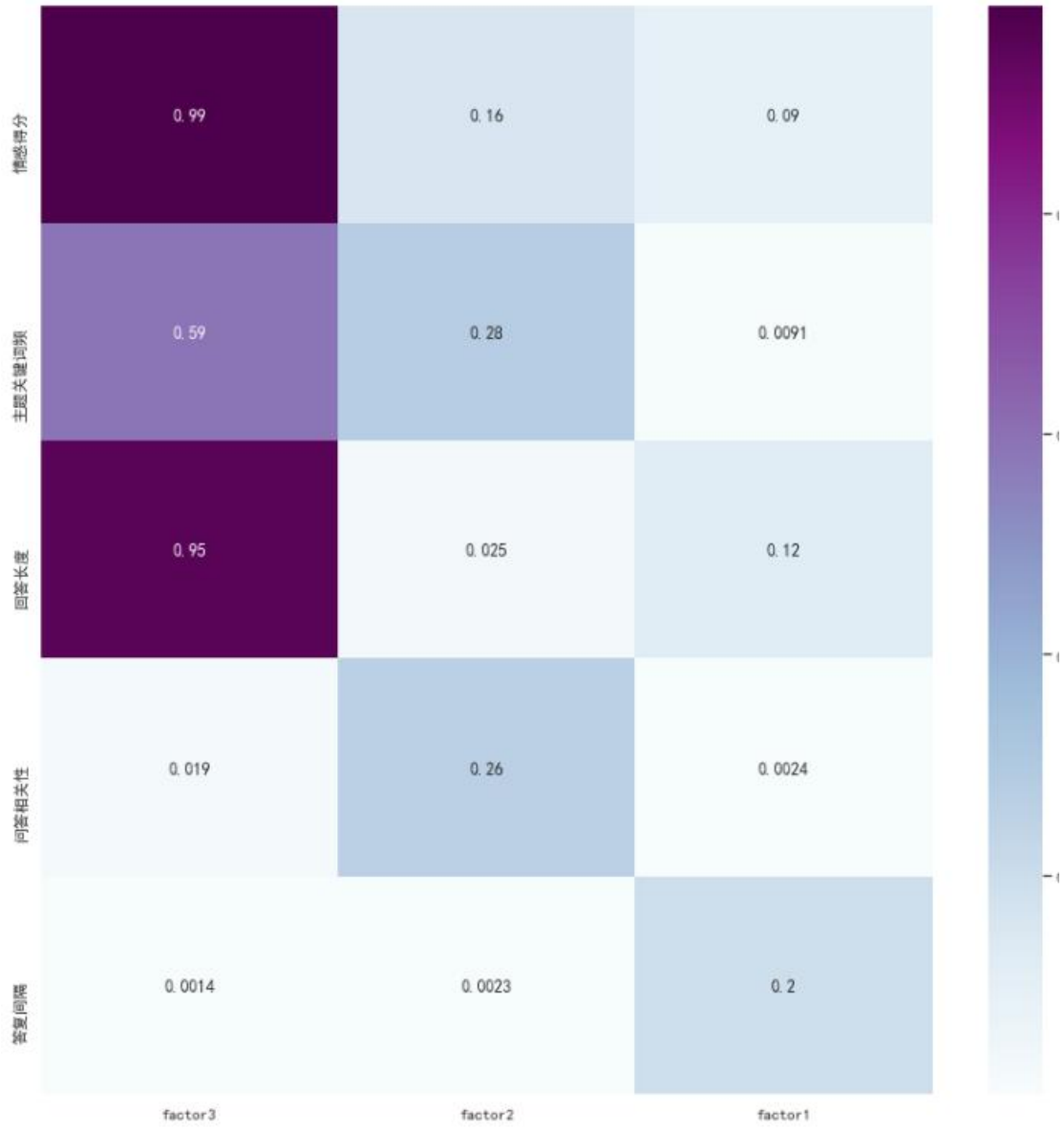


图 15：热力图

通过观察上面图，可以看出分析得到的三个主成分因子中，基于文本特征的主题关键词频、回答长度、情感得分的文本特征因子占比较大，另外两个因子主要为问答相关因子和时效性因子。
通过计算因子的方差信息，可以更具体得看出各个因子对结果的贡献率。

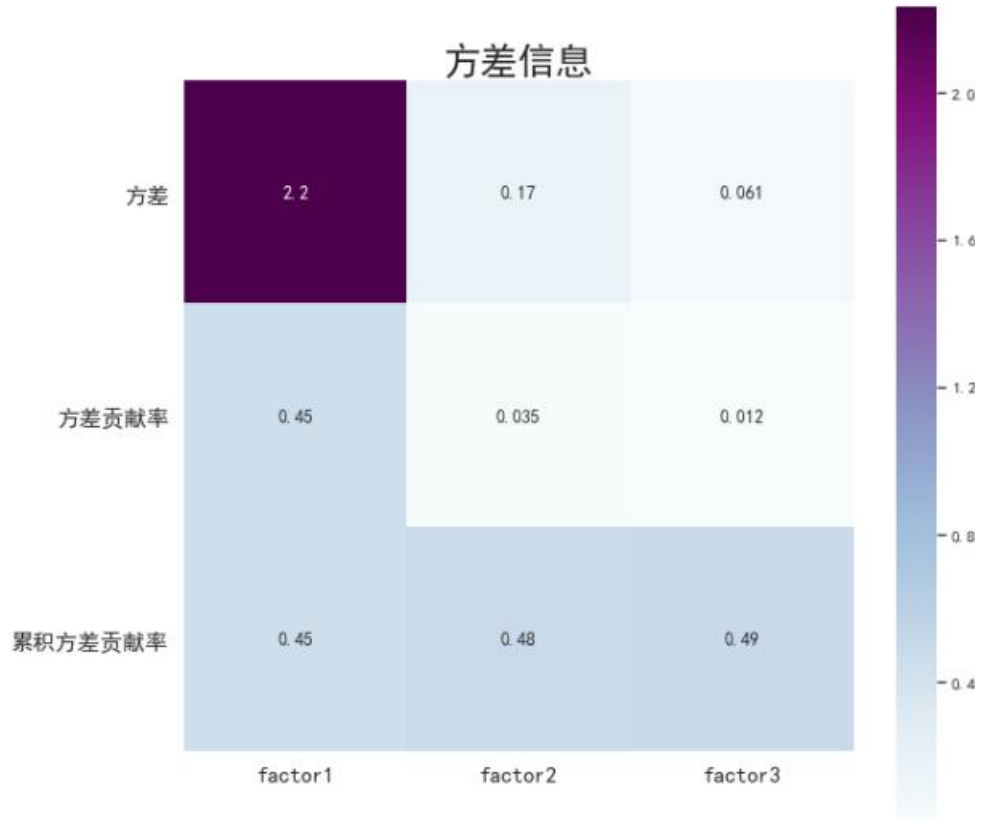


图 16：因子方差信息图

通过因子方差可以看出文本特征因子对整体方差贡献率较大，相关性因子、时效因子次之。从整体因子分析的结果来看，所选取的特征能够作为构建评价自动化回答评价模型的指标。

（3）数据可视化结果

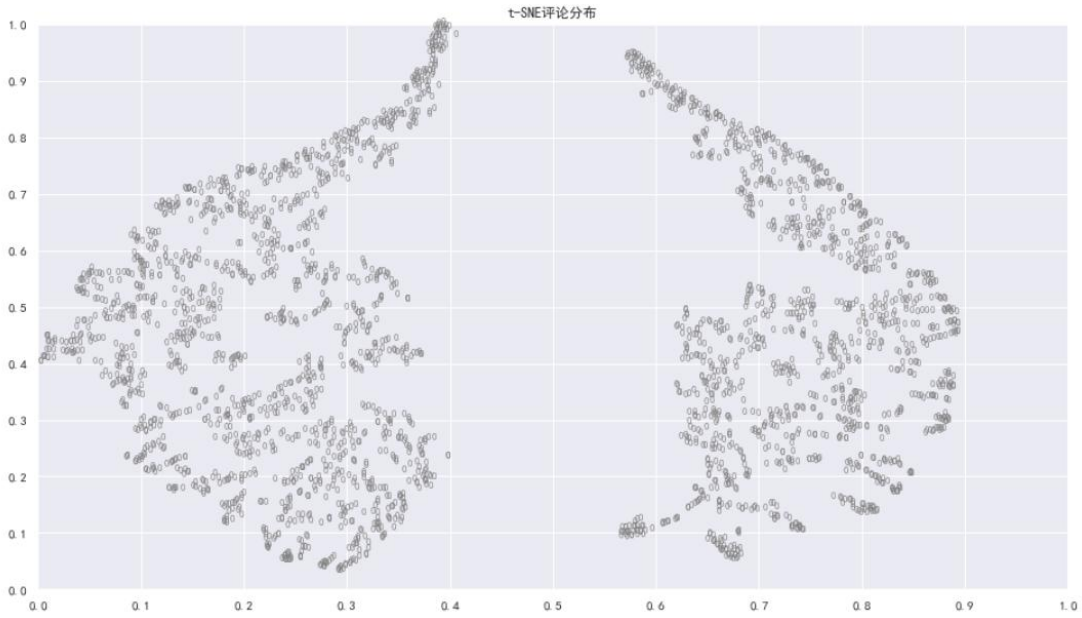


图 17：矩阵降维可视化图

从可视化结果来看，数据分布比较集中，并且明显呈现出两边聚合的特点。分别对左右两侧数据进行抽取验证。

在左侧聚合分布部分数据集如下：

54	4090	UU0082048	反映A市阳光100小区没有公办幼儿园的问题	2018/11/19 12:00:38	的书记，你好:据了解，今年1月，市政府办公...	网友"UU0082048"您好！您的留言已收悉。现将有关情况回复如下：阳光100小区位于岳路...	2018/12/5 14:00:13
58	4134	UU0081325	咨询2018年执业医师资格证的相关问题	2018/11/14 15:17:17	\n\t\t\t\t\n\t\t\t\t\t您好，2018年执业医师成绩已经出来了，请问...	网友"UU0081325"您好！您的留言已收悉。现将有关情况回复如下：执业医师资格证是由省卫...	2018/11/20 9:41:09
...	...	...	...	...	...	...	...
2757	167540	UU0082379	咨询G市驾驶证异地检验换证问题	2019/9/21 15:33:14	\n\t\t\t\t\n\t\t\t\t\t我是2014年1月在G市拿的驾驶证，驾驶证6...	UU0082379:您好，您咨询的"咨询G市驾驶证异地检验换证问题"的问题，我局已收悉，现将...	2019/9/23 16:06:30
2759	168110	UU008482	G1区婴儿医保卡办理的咨询	2019/11/14 11:15:21	\n\t\t\t\t\n\t\t\t\t\t尊敬的领导您好！本人想咨询下6个月的宝宝在G...	您好，我办已将您反映的情况直接转至区医保局，您也可直接咨询区医保局，联系电话0736-717...	2019/11/27 10:45:14
2762	168272	A00074619	G1区朝阳市场占道经营，严重扰民	2019/3/26 14:25:51	\n\t\t\t\t\n\t\t\t\t\tG1区朝阳市场内明显占道，私搭乱宰、修改公用...	收悉后，我局派分管局长和市场管理责任人于4月3日再次到市场进行了现场调查，发现该市场脏乱的问...	2019/4/9 14:43:40

右侧聚合分布数据集如下：

56	4111	UU0081139	关于A3区岳银欣苑小区停车问题的反映	2018/11/16 12:58:21	\n\t\t\t\t\n\t\t\t\t\t尊敬的A市市委领导，您好！A3区银杉路559...	网友"UU0081139"您好！您的留言已收悉。现将有关情况回复如下：一、关于小区室外停车车...	2018/12/7 14:32:05
57	4133	UU0081467	咨询小孩上A市的小学是否为一类优先录取问题	2018/11/14 15:36:55	\n\t\t\t\t\n\t\t\t\t\t本人小孩明年即将读小学，孩子和母亲的户口在A...	网友"UU0081467"您好！您的留言已收悉。现将有关情况回复如下：根据《关于进一步做好义...	2018/11/27 14:47:14
59	4156	UU008684	强烈建议放开集体户口的博士研究生首套房购买资格	2018/11/12 20:42:39	\n\t\t\t\t\n\t\t\t\t\tA市出台了人才新政是好事，但不允许集体户口博...	网友"UU008684"您好！您的留言已收悉。现将有关情况回复如下：根据《A市人民政府办公厅...	2018/12/4 16:37:07
...	...	...	...	...	...	...	...
2756	166919	A00085038	为何EMS快递身份证收费这么贵?	2018/7/11 14:44:22	\n\t\t\t\t\n\t\t\t\t\t朱局长您好，我之前在派出所办理新身份证。工作...	A00085038：您在《问政西地省感谢您的理解与支持！2018年8月6日	2018/8/7 15:34:12
2758	168031	UU0082326	咨询l市2012年两个数据	2013/7/24 23:15:33	长：	\n\t\t\t\t\n\t\t\t\t\t张局你好！打扰了。我因研究洞庭湖生态...	UU0082326： 您好！您需要的两个数据现说明如下： 1、2012年https:...	2014/7/2 10:12:39

从两侧数据集分布来看，位于左边的回答相较于右侧较差。根据聚合的情况为不同文本特征向量添加训练标签，以训练学习模型。

上面三个步骤和结果皆是为了通过对回答文本特征的抓取，抓取到有效的文本特征，答复间隔、问答相关性、回答长度、主题关键词频、文本情感得分。通过相关性检测和因子分析法确定了所研究的特征能够较好的解释回答的完整性、相关性、可解释性。通过对附件 4 所给数据得到的文本特征数据进行降维可视化，可以明显看出满意回答和一般回答呈现两个集群。通过对 SVM 机器学习方法学习两个集群的特征作为自动化评价是否为满意回答的模型，较好的基于答复意见给出了一套合适的评价方案。

对于集群数据的分类模型选择仍有不足之处，由于两个集群仍有边界模糊的回答，产生问题主要有：

(1) 对特征进行量化的过程仍然存在不精准的情况，数据预处理不够充分。

(2) 可能仍有部分特征不能更好地解释回答的相关性、可解释性、完整性。

(3) 训练模型对部分数据分类情况不佳，可能是高维数据在映射到二维下丢失了部分信息导致。

5、结论

由于大数据、云计算、人工智能等技术的发展，利用自然语言处理和文本挖掘的方法来整理微信、微博、市长信箱等网络问政平台所产生的互联网公开来源的群众问政留言，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极大的便利，对提升政府的管理水平和施政效率具有极大的推动作用。

对问题一建立了一级标签分类模型能较好地对留言分类，且有较高的准确率，减少工作量和差错率的同时、提高了效率；对问题二建立归类和热度指数排序模型，将留言进行归类，定义合理的热度指标，最后按给定格式保存两张表格；对于问题三建立模型从答复的相关性、完整性、可解释性等角度对答复意见给出一套较好地评价方案，保证了答复信息的质量。

对一个问题长期反复地投诉和留言，耗费了群众太多的时间和精力，并且也占用了平台过多的时间和资源，让工作人员的工作量增加了好几倍。对此我提出一个建议：可采取此市民监督机制，评价栏应分几个内容组成：市民投诉原因、职能部门已完成的处理结果、没完成的原因、预期完成时限，市民对处理结果是否满意，不满意的原因是什么，同一类同一事件投诉设置同一编号可追溯，累计追溯投诉达三次，由各行业各部门组成的监督管理委员会对此处理单进行审定，如果确实是职能部门原因导致没有处理完成的将计入绩效，也可视情况利用媒体进行监督。

政府服务是一个综合性的事务，需要不断转变作风，创新工作理念，改进工作模式，进一步研究和完善热线管理办法及考核机制来提高办件的处理效率和成效，以科学有效的方式推动该地区公共服务水平的提高，让人民生活水平不断得到提升，并坚持为人民服务的根本宗旨，真正做到为人民造福。

6、参考文献

[1] 王伟，冀宇强，王洪伟，郑丽娟。中文问答社区答案质量的评价研究: 以知乎为例 国书情报工作.第 61 卷第 22 期 2017 年

[2] Kalchbrenner，N.，Grefenstette，E.和＆Blunsom，P.（2014）。用于句子建模的卷积神经网络

[3]张翔，俊波赵，亚·莱卡。字符级卷积网络的文本分类 ，2015 年 9 月

[4]郑忠明，江作苏，网络用户劳动与媒介资本价值——基于美国社交新闻媒体 Reddit 的案例分析 ，2015 年

[5]李连、朱爱红、苏涛，一种改进的基于向量空间文本相似度算法的研究与实现，2012 年 2 月
[6] 楼海淼，孙秋碧.基于因子分析的我国各省经济活力评价研究[J].福州大学学报(哲学社会科学版)，2005 年
[7] 郭锐，基于 LDA 主题樽型的电商客户评论情感分析 ，2017 年
[8]侯小培、高迎,卷积神经网络 CNN 算法在文本分类上的应用研究 ,2019 年