基于机器学习的智慧政务系统

摘 要：近年来互联网的不断发展，网络平台成为了政府了解民意的重要渠道。随着网络平台中各类社情民意的数据量高速增长，人工处理数据将导致效率低、差错率高等问题。本文旨在利用机器学习实现文本处理。基于互联网公开来源的群众问政留言记录及相关部门对群众留言的答复意见，经过数据预处理、特征提取与数据分析，实现对留言信息的分类、热点问题的提取和答复意见的评价。

针对问题一，群众留言分类，对附件二中留言内容、留言主题进行**去重**，获取未重复的群众留言信息。利用 **Jieba** 中文分词工具对留言信息进行**分词**，分词后将停用词进行删除，即词频较高但信息含量少的字词。再利用 **TF-IDF 算法**将文本向量转化为数字向量。将向量化的数据输入 **KNN 分类模型**，对每个留言信息进行分类。引入 Precision、Recall、F-score 三项评价指标对模型的有效性进行评价排序，最终模型的 F-score 值为 0.9。

针对问题二，热点问题挖掘，对附件三中留言信息进行分词和向量化。采用 **K-means 聚类**算法对向量化的数据进行聚类，即将留言内容相似的数据归为一类，以**轮廓系数**为指标确定 **K-means 聚类**算法中类的个数。基于 **Reddit 算法**，提出了一种新的留言热度计算算法，称为 Jyreddit 算法，并计算每一类留言信息的热度值，最后提取热度最高的五类热点问题。

针对问题三，答复意见的评价，基于相关性、完整性、可解释性三个角度对附件四中答复意见进行质量评价。先将留言内容和答复意见进行数据预处理，采用 **TextRank 算法**提取留言内容中关键词，采用 **BM25 算法**计算关键字与答复意见的文本相似度。基于文本的相似度对答复意见的相关性进行评分。通过对样本进行数据分析得出文本的完整性与答复意见字数具有关联性，则完整性的评分以答复意见文本字数乘以 0.01 表示。对于两项评分总分低于 4.0 分的答复，其部分答复仍具有可解释性，通过关键字匹配筛选出具有解释性的答复并赋予可解释性加分 20。最终得到所有数据的评分结果。

关键词：TF-IDF 算法，KNN 分类模型，K-means 聚类，Reddit 算法，TextRank 算法，BM25 算法

Abstract

The rapid development of the Internet in recent years has made the online platform an important channel for the government to understand the thoughts of the people. However, with the rapid growth of the data of people's opinions on the network platform, manual data processing will lead to low efficiency and high error rate. So the purpose of this article is to implement text processing using machine learning. Based on the record of people's comments on government affairs and the replies on people's comments received by relevant departments from Internet public sources, the classification of message information, extraction of hot issues and evaluation of reply comments are realized through data pre-processing, feature extraction and data analysis.

In response to question one, we will classify the comments of the masses.In order to obtain unrepeated mass message information,we remove duplication of the message content and message subject in Annex II. By using Jieba Chinese word segmentation tool,we segment the message information and delete the stop words after the segmentation, namely, the words with higher frequency but less information content. Then, we convert the text vector into the digital vector by using TF-IDF algorithm.We classify each message by feeding vectorized data into the KNN classification model.Three evaluation indicators are introduced to rank the effectiveness of the model,including Precision, the Recall and F-score. The final F-score of the model is 0.9.

In response to question two, we'll dig out the hot topics in the comments. The message information in Annex III is segmented and vectorized. The K-means clustering algorithm is used to cluster the vectorized data. Namely, the data with similar messages fall into one category, and the number of categories in the K-means clustering algorithm is determined by the contour coefficient. A new message popularity calculation algorithm called Jyreddit algorithm is proposed based on the Reddit algorithm,which calculates the hot value of each category of message information, and finally extracts the top five issues with the highest values.

In response to question three, we will evaluate the quality of the replies in annex iv on the basis of relevance, completeness and interpretability.Firstly, the message content and reply comments will be pre-processed. The TextRank algorithm is used to extract keywords in the message content. The BM25 algorithm is used to calculate the text similarity between the keywords and the reply comments. The relevance of the reply opinions is scored based on the similarity of the text. Secondly,through the data analysis of the sample, it is concluded that the completeness of the text is related to the number of words in the reply opinions, and the score of completeness is expressed by multiplying the number of words in the reply opinion text by 0.01. For the two responses with a total score of less than 4.0 points, some of the responses are still interpretable, and the answers that are interpreted through keyword matching are filtered out and given an additional 20 points for interpretability. Finally get the score results of all data.

Keywords: word segmentation; TF-IDF algorithm; KNN algorithm; K-means algorithm; Reddit algorithm.

目 录

引言

近年来随着科技不断发展，互联网已经逐渐成占据中国民众的政治、经济和社会生活中重要地位，同时也成为了民众行使知情权、参与权、表达权和监督权的重要渠道。政府以互联网为媒介通过网络问政的方式做宣传、做决策，了解民意，问政于民，从而以达到取之于民，用之于民的目的。

相比于其他媒体，互联网独特的时效性、便捷性以及互动性等优势和特点为网络问政技术支持。“网络问政”作为当前的新事物，其方便快捷的方式引起党和政府的高度重视。作为一种便捷的互动方式，“网络问政”有利于社会矛盾的释放，对于建立健全社会矛盾释放机制具有重要意义[1]。但随着大数据时代的到来，互联网中通过“网络问政”方式收集到的群众留言信息量十分庞大，采用人工处理数据的方式将导致效率低、差错率高等一系列问题。所以，如何有效的将群众留言信息有效的进行分类，提取热点问题并送至相关部门处理成为一大难题。

为解决上述问题，本次建模利用网络信息平台系统发布的网络招聘信息数据，利用 jieba 中文分词工具对职位描述进行分词、K-means 聚类的方法及 KNN 算法，达到以下三个目标：

（1） 利用文本分词和文本向量化的方法对文本数据进行预处理。基于机器学习的方法对群众留言内容进行分类。

（2） 对据群众反应的问题进行归类，统计一段时间内留言集中反应的问题。

（3） 根据研究相关部门对群众留言内容的答复，从相关性、完整性、可解释性三个角度设计一套对答复内容的评价体系。

一、群众留言分类

1.1 问题分析

随着网络的日益普及，互联网在中国民众的政治、经济和社会生活中扮演着日益重要的角色。中国公民以网民的身份通过互联网行使知情权、参与权、表达权和监督权，即网络问政。在处理网络问政平台的群众留言时，需要先根据留言内容进行分类分派至相应的职能部门进行处理。但目前大多电子政务系统的留言分类均依赖人工进行处理，存在着存在工作量大、效率低，且差错率高等问题。

为解决上述问题，本文将根据附件 2 给出的留言数据建立基于留言内容和留言内容的分类模型。其主要步骤为：

（1）数据预处理:对数据进行去重和分词，并对重复且不具有意义的字词进行删除。

（2）文本向量化：为便于文本分类，需要将文本数据转换为数字向量。

（3）分类：将转为数字向量的数据作为输入，对样本进行分类。

（4）结果分析：采用 F-score 对分类方法进行评判。

1.2 数据去重

附件一中所给数据主要包含留言编号、留言用户、留言主题、留言时间、留言详情、一级分类共六类信息。问题一中要求对根据文本信息进行分类，考虑数据中可能存在同一用户针对同一问题多次留言的情况，为防止此类冗余数据对模型训练时产生干扰，故先将重复数据进行去重，即删除 159 条内容相同的留言。对于分类问题而言，将根据留言的主题和详情所表述的文本信息对文本内容进行分类并去除留言时间、用户、编号等干扰信息。

1.3 中文分词

分词目的在于将一句话或者一个段落拆分成许多独立个体的词，便于后续将词转化成向量。例如实例数据中表 2 的数据：“建议增加 A 小区快递柜”，我们期望语料库统计后分词的结果是：″建议/增加/A 小区/快递柜″。但由于中文的语义分歧，一定概率会使分词结果为“建议/增加/A 小/区/快递/柜”。为了能够正确的分词，″建议/增加/A 小区/快递柜″这个分词后的句子出现的概率要比“建议/增加/A 小/区/快递/柜”大。即如果有一个句子, 它有 m 种分词选项如下：

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n_1} \\ A_{21} & A_{22} & & A_{2n_2} \\ \vdots & & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn_m} \end{bmatrix} \tag{1}$$

其中 n_i 表示使用第 i 种分词方法后字词个数。若第 r 种分词方法为最优的分词方法，则该分词方法的统计分布概率应最大，即：

$$r = \arg max P(A_{i1},A_{i2},...,A_{in_i}) \tag{2}$$

但上式并不易求出，由于涉及到计算 n_i 个字词的联合分布 $P(A_{i1}, A_{i2},..., A_{in_i})$ 。故为了简化计算，通常使用马尔科夫假设，即每一个字词出现的概率仅与上一个词有关，即：

$$P(A_{ij} \mid A_{i1}, A_{i2},..., A_{i(j-1)}) = P(A_{ij} \mid A_{i(j-1)}) \tag{3}$$

则联合分布的求解可写为：

$$P(A_{i1}, A_{i2},..., A_{in_i}) = P(A_{i1})P(A_{i2} \mid A_{i1})P(A_{i3} \mid A_{i2})...P(A_{in_i} \mid A_{i(n_i-1)}) \tag{4}$$

通过标准语料库计算出所有字词间的二元条件概率，如任意两个字词 $w1, w2$ ，其条件概率分布可近似表示为：

$$P(w_2 \mid w_1) = \frac{P(w_1, w_2)}{P(w_1)} \approx \frac{freq(w_1, w_2)}{freq(w_1)} \tag{5}$$

式中 $freq(w1, w2)$ 表示 $w1, w2$ 在语料库中相邻一起出现的次数，其中 $freq(w_1), freq(w_2)$ 分别表示 $w1, w2$ 在语料库中出现的次数。基于语料库建立的统计概率，当出现新句子时通过计算各种分词方法对应的联合分布概率，找到最大概率对应的分词方法，即为最优分词。

本文为解决分词带来的语义分歧问题，采用了结巴分词。结巴分词基于自带的词典生成 Trie 树从而实现高效的词图扫描，并将句子划分成所有可以成词的情况后生成有向无环图（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；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。其过程如图 1 所示：

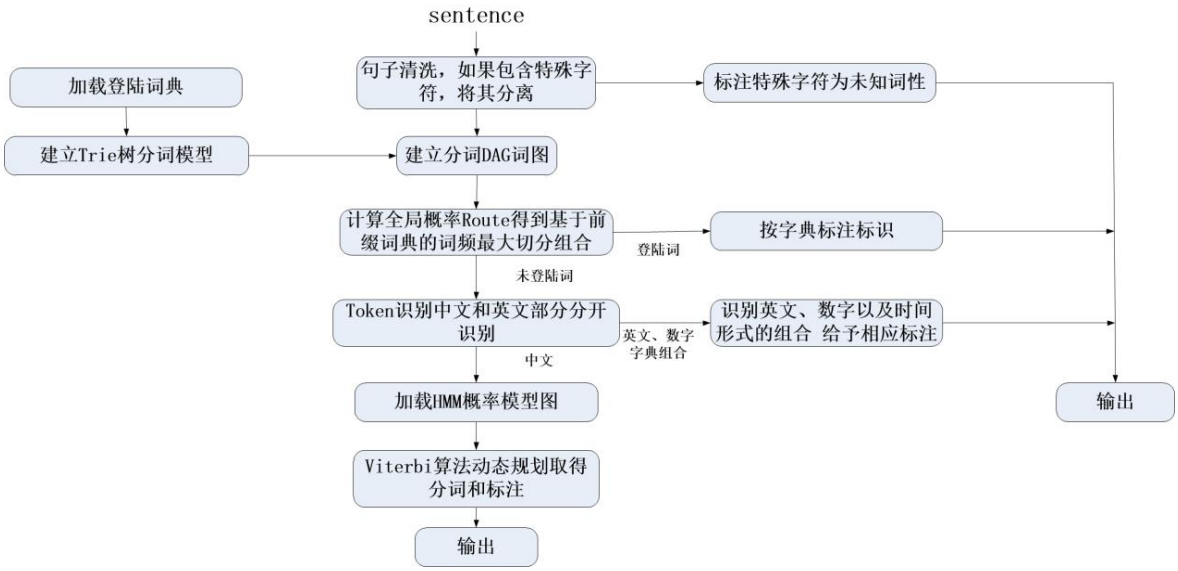


图 1 数据预处理流程图

1.3.1 Trie 树

分词过程中，将根据自带的语料库列出文本数据中所有可能成词的情况。结巴分词将语料库生成 Trie 树结构，通过对 Trie 树进行词图扫描，将句子划分成所有可以成词的情况并生成有向无环图（DAG）。Trie 树是一种树状数据结构，其节点存储字母。基于特定方式构造节点，可以通过遍历树的分支路径从结构中检索单词和字符串。Trie 的核心思想是利用字符串的公共前缀来降低查询时间的开销以达到提高效率的目的。

Trie 树查找过程如图 2 所示。从根结点开始逐次往后搜索，如搜索“海淀区”；先查找定位到第一个关键字“海”；当第一个关键字查找完毕后，根据第一个关键字的子树以相同的方式继续进行查找关键字“淀”；并以此方式进行迭代直到判断树节点的 isEnd 节点为 True,即图中红色的点，并返回“区”isEnd=True，则查找结束。

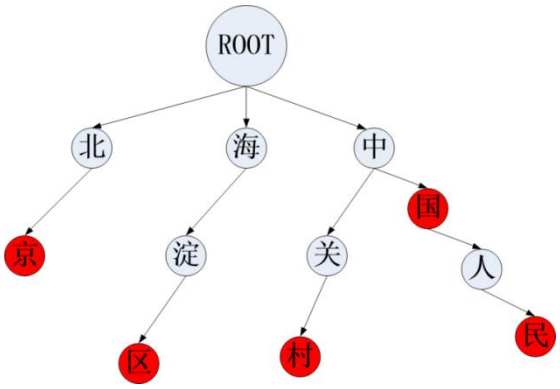


图 2 Trie 树查抄字词过程

结巴分词基于自带的词典生成 Trie 树从而实现高效的词图扫描，并将句子划分成所有可以成词的情况后生成有向无环图（DAG）。

1.3.2 动态规划

在基于自带词典生成 Trie 树后，将文本数据划分为一个有向无环图（DAG），从而将分词问题转化为动态规划（Dynamic programing）解决最优路径问题[2]。

最优路径算法的实质是在一个带权有向图中寻求最短路径,可将各路段的最优目标函数看作路段的权值，有效的联通路径集合配合路段的权值构成基础的带权有向图。假设有向图如图 3 所示,B 为起点,E 为终点，在起点与终点间存在多个路径节点。

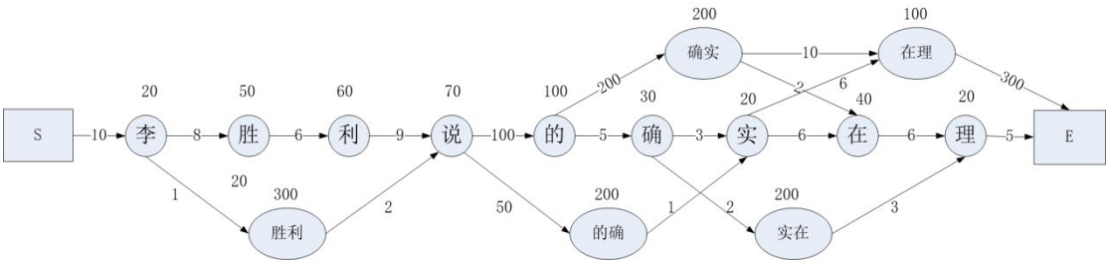


图 3 有向无环图

若在路径节点中存在孤立节点，即从起点到终点的必经节点。在有向图寻优的过程中，可将“李”、“说”、“实”点作为中间节点，对有向图进行划分，即将全局最优问题划分成多个局部最优问题。如句子“李胜利说的确实在理”划分成子问题：“李/胜/利”和“李/胜利”的最优选择、“说/的/确实”和“说/的确/实”的最优选择等。找每一个子路段的最优路径,将找寻到的各个子路段的最优路径顺序连接后即可得到全局最优路径，即结果“李胜利/说的/确实/在理”。特别的，对于子路段中只存在一条路径的情况下,直接默认其为选择的路径。

1.3.3 未登录词处理

在分词过程中，当出现某些词并未出现在结巴自带的词典中时，将采用隐式马尔科

夫模型对其进行划分[3]。采用 HMM 模型的分词结果详情见附录（一）。

隐式马尔可夫模型模型（HMM）一般用于处理含有以下特点的问题：

（1）问题是基于序列的，如状态序列等。

（2）问题中有两类数据，一类序列数据是可以观测到的，即观测序列；而另一类数据是不能观察到的，即隐藏状态序列，简称状态序列。

HMM 模型用于处理文本数据时，划分的一系列字词表示为观测序列，而每次分词时所具有的所有可能性则是隐藏序列，而 HMM 模型的任务则是预测分词的结果，即将可能性最大的情况选择出来。下面，我们将用数学符号来描述 HMM 模型的工作过程

隐式马尔可夫模型用于表示观察序列上的概率分布如图 4 所示。变量 Y_t 表示时间 t 处的观测值，即最后呈现出的分词结果；变量 S_t 表示时间 t 处的状态分布，即每次分词时可选状态的概率分布。HMM 的目的是定义观测值 Y 的概率分布 S。我们假设观测值是离散、等距时间间隔采样的，则 t 为整数并表示时间索引值，故 Y 表示分词后的序列。

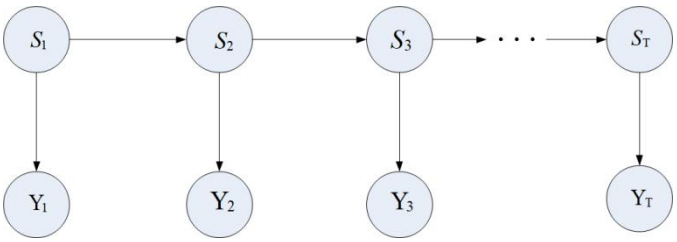


图 4 隐式马尔科夫链

隐式马尔可夫模型具有齐次性和观测独立性假设。其中观测独立性表示假设在时间 t 的观测值 Y_t 仅由该时刻的状态 S_t 所决定。齐次性表示在给定状态 S_{t-1} 的情况下，当前状态 S_t 独立于 $t-1$ 时刻之前的所有状态。隐式马尔可夫模型可通过状态和观测值的联合分布进行表示：

$$P(S_{1:T},Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T}P(Y_t|S_{t-1})P(Y_t|S_t) \tag{6}$$

式中 T 表示观测序列长度，$1:T$ 表示。HMM 模型由初始概率分布 Π 、状态转移矩阵 A 和发射矩阵 B 三部分组成。Π、A 决定状态序列， B 决定观测序列。HMM 模型可以表示如下：

$$\lambda = (\Pi, A, B) \tag{7}$$

HMM 模型的输入为 λ 和观测序列长度 T，输出是观测序列 $O = \{o_1, o_2, ...o_T\}$，其步骤如表 1 所示：

表 1 HMM 模型观测序列生成步骤

输入：初始概率分布 Π 、状态转移矩阵 A 、发射状态矩阵 B 和观测序列长度 T
输出：观测序列 $O = \{o_1, o_2, ...o_T\}$
1. 根据初始状态概率分布 Π 生成初始状态 i_1

2. for t from 1 to T

a.根据隐藏状态 i_t 的观测状态分布 $b_{i_t}(k)$ 生成观察状态 o_t

b.根据隐藏状态 i_t 的状态转移概率分布 a_{i_t} 产生隐藏状态 i_{t+1}

3. 所有的 o_t 一起形成观测序列 $O=\{o_1,o_2,...o_T\}$

1.3.4 分词结果展示

采用结巴分词后的部分结果如图 5 所示：

```
0      [A3，区，大道，西行，便，道，，，未管，所，路口，至，加油站，路段...
1      [位于，书院，路，主干道，的，在水一方，大厦，一楼，至，四楼，人为，拆...
2      [A，市政府，、，市，交警支队，、，市，安监局，、，市，环保局，、，...
3      [胡书记，，，您好，，，感谢您，百忙之中，查看，这份，留言，。，我，的...
4      [  ，  ，K8，县，丁字街，的，商户，乱，摆摊，，，前段时间，丁字街,...
                                  ...
490    [邓，局长，：，?，?，?，?，  ，  ，B9，市，卫生局，卫生，监...
491    [  ，  ，山门，镇，人民，医院，挂号费，这几年来，一直，都，是，高额,...
492    [我，是，E7，县，人民，医院，的，一个，普通，医务人员，，，我，想...
493    [张，厅长，：，  ，  ，您好，！，我，是，一名，医务，工作者，。，...
494    [书记，：，  ，  ，你好，，，我们，是，G8，县乡镇，卫生院，的，一...
```

图 5 结巴分词结果

根据图 5 显示采用结巴分词后的结果可看出，文本已经被分成单个的字词，但数据中仍然存在大量的高频字词和标点符号，如“，”、“。”、“我们”、“的”等，这次字词出现的频率很高但是蕴含的信息量非常的少。为了便于后续对数据的分析和处理，我们将对这些冗余字词进行删除处理。其删除后的结果如图 6 所示：

```
14    [市区，朝晖路，锦楚，国际，新城，三区，月份，一共，停电，于次，每次，...
15    [西地省，地区，常年，阴冷，潮湿，气候，近年，气候，恶劣，地处，月亮，...
16    [胡书记，冬天，市，湿冷，冬天，受不了，诶，太冷，被子，感觉，潮潮，洗...
17    [尊敬，市委，市政府，市是，一座，历史，名城，一座，幸福感，城市，幸福感...
18    [县城，更新，公交线路，新，公交车，试运行，中，市民，出行，一项，民生,...
19    [希望，县，路路，公交车，延迟，晚上，点，晚上，点，路路，公交车，沿线...
20    [县，公交车，破旧不堪，这是，最让人，愤怒，车人，监控，看似，监控，插卡...
21    [上周，提交，请求，迎丰，公园，人性，关怀，角度，延后，清晨，路灯，熄...
```

图 6 删除停用词结果

1.4 TF -IDF 算法

对文本数据进行去重、分词后需将文本向量转换为数字向量以供模型对数据进行分析计算，此处我们采用 TF -IDF 算法对文本数据进行处理。

TF -IDF(Term Frequency-Inverse Document Frequency, 词频-逆文件频率)是一种统计方法，用于评估字词与文档集/语料库中单个文档的重要性。字词的重要程度与其在单个文档中出现的次数成正比，但与其出现在文档集中的频率成反比。因此，对于一些文档中常出现的词汇，例如“的”、“我们”、“如果”等字词，由于它们对于该文档的意义并不大，因此即使出现多次，TF -IDF 值也会很低。但是，如果“错误”一词在单个文档中高频出现，则可能意味着它对于该文档重要程度较高。因此，TF -IDF 值通过

计算两个量来完成：一个字词在文档中出现的次数——词频(term frequency,TF)、单词在一组文档中的反向文档出现频率——逆向文件频率(inverse document frequency, IDF)。

TF -IDF 值的计算由如下两部分组成：

（1）词频（TF）：字词出现在文档中的次数。为了防止该值偏向于文本内容较多的文档，故进行归一化，即词频除以文章总词数。鉴于一些高频常用词与主题关联性不强, 但部分出现频率较小的词能够表达文章的主题, 所以仅使用 TF 是不合适的。字词权重的设计必须满足：一个词主题表征的能力越强则权重越大，反之，权重越小。所有统计的文章中，一些字词仅在较少文章中高频出现，则该词课可高度表征该文章的主题，所以此类字词权重应该设计的较大。为解决该问题，引进 IDF 值来衡量字词的重要性

（2）逆文档频率（IDF）：某个含有指定字词的文档在文档集中出现的次数，表示单词在整个文档集中有的普遍性，IDF 值与单词出现的频率成反比。可以通过以下方法来计算该指标：将文档总数除以包含某个指定单词文档的数量，然后计算对数。因此，如果该单词非常普遍并且出现在许多文档中，则该数字将接近 0。否则，它将接近 1。

将上述两个量相乘得出文档中字词的 TF -IDF 值。值越高，该字词与该特定文档相关性越强，即字词的重要性越高。TF -IDF 值数学表达式如下：

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{8}$$

式中 D 为文档集数量，d 表示文档集中标号为 d 的文档，t 表示 d 文档中的标号为 t 的字词。其中：

$$tf(t,d) = \log(1 + freq(t,d)) \tag{9}$$

$$idf(t,D) = \log\left(\frac{N}{count(d \in D; t \in d) + 1}\right) \tag{10}$$

式中 *freq* 表示第 t 个单词在第 d 个文档的出现频率，N 表示文档集中的文档总数，$count(d \in D; t \in d)$ 表示包含字词 t 的文档个数。

1.5 分类模型介绍

1.5.1 KNN 近邻算法

在采用 TF -IDF 算法对文本数据进行向量化后，我们采用 K 近邻法（K-nearest neighbor, KNN）对数据进行分类。KNN 是一种分类与回归方法，其主要思想为：给定已知类别的数据集 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，其中 $x_i \in X \subseteq R^n$ 为样本的特征向量，$y_i \in Y = \{c_1, c_2, \cdots, c_k\}, i = 1, 2, ..., N$，$y_i$、$k$ 分别表示样本的类别和类别种类个数。计算训练集中与未知类别的测试数据距离并找出最靠近的 k 个样本，然后基于这 k 个最近邻的信息来进行预测。

K 近邻法计算步骤如表 2 所示：

表 2 KNN 计算步骤

输入：近邻个数 K、已知类别的数据 D、测试数据 T

输出：测试数据类别 Y

For j from 1 to J

1.计算测试数据 t_j 与所有训练样本的距离 L

2.基于距离值 L 对其进行升序排序

3.选择对应的前 K 个训练样本

4.K 个样本中出现次数最多的类别为测试样本的类别。

End

KNN 最邻近分类算法的实现原理：为了判断未知类别的测试数据 $T=\{t_1,t_2,\cdots,t_J\}$，以所有已知类别的样本作为参照，计算未知样本与所有已知样本的欧式距离：

$$L_2(x_i,t_j)=(\sum_{l=1}^{n}|x_i^{(l)}-t_j^{(l)}|^2)^{\frac{1}{2}} \tag{11}$$

式中 $x_i^{(l)}$、$t_j^{(l)}$ 分别表示已知类别数据和测试数据的第 l 个特征。选出与未知样本距离最近的 K 个已知样本，根据投票法则（majority-voting）将未知样本与 K 个最邻近样本中所属类别占比最多的归为一类。

判别阶段在分类任务中使用“投票法”，在回归任务中使用“平均法”，即将 k 个实例的实值输出标记的平均值作为预测结果。且在计算距离时除了可以使用欧式距离，还可以使用 Lp 距离或曼哈顿距离。

1.5.2 支持向量机（SVM）

支持向量机是机器学习中主流的分类模型之一。其基本模型是定义在特征空间上间隔最大的线性分类器。SVM 包括核技巧，这使它成为实质上的非线性分类器，对于输入空间中的非线性分类问题，可以通过非线性变换将它转化为高维特征空间中的线性分类问题[4-5]。模型的输入为训练数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中，$x_i\in\mathbb{R}^N$，$y_i\in\{+1,-1\},i=1,2,...,N$。$x_i$ 为第 i 个特征向量，y_i 为类标记，当它等于+1 时为正例，等于-1 时为负例。输出为分离超平面和分类决策函数。算法主要流程如下：

（1）选择惩罚参数 C>0，构造并求解凸二次规划问题：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i,x_j)-\sum_{i=1}^{N}\alpha_i\\&s.t\sum_{i=1}^{N}\alpha_i y_i=0,0\le\alpha_i\le C,i=1,2,...,N\end{aligned} \tag{12}$$

（2）得到最优解：$\alpha^*=(\alpha_1^*,\alpha_2^*,...,\alpha_N^*)$

（3）计算：$w^*=\sum_{i-1}^{N}\alpha_i^* y_i x_i$，选择 α^* 的一个分量 α_j^* 满足条件 $0<\alpha_j^*<C$。

（4）计算出 $b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i, x_j)$

（5）求得分离超平面： $w^* x + b^* = 0$ 分类决策函数： $f(x) = sign(w^* w + b^*)$

其中， $wx + b = 0$ 为超平面； (x_i, y_i) 为超平面样本点； α_i 为拉格朗日乘子。

1.5.3 XG Boost

XG Boost 实现的是一种通用的 Tree Boosting 算法[6]，本节将基于决策树弱分类器总结 XG Boost 的算法主流程。模型的输入是训练集样本 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ 、最大迭代次数 T、损失函数 L 和正则化系数 λ, γ 。输出是强学习器 $f(x)$ 。

对迭代轮数 $t = 1, 2, ...T$ 有：

计算第 i 个样本 $(i-1, 2, ..m)$ 在当前轮损失函数 L 基于 $f_{t-1}(x_i)$ 的一阶导数 g_{ti} ，二阶导数 h_{ti} ,计算所有样本的一阶导数之和 $G_t = \sum_{i=1}^{m} g_{ti}$,二阶导数之和 $H_t = \sum_{i=1}^{m} h_{ti}$ 。

基于当前节点尝试分裂决策树，默认分数 score=0，G 和 H 为当前需要分裂的节点的一阶二阶导数之和。

对特征序号 $k = 1, 2...K$:

（1） $G_L = 0, H_L = 0$

（2）将样本按特征 k 从小到大排列，依次取出第 i 个样本，依次计算当前样本放入左子树后，左右子树一阶和二阶导数和：

$$G_L = G_L + g_{ti}, G_R = G - G_L \tag{13}$$

$$H_L = H_L + h_{ti}, H_R = H - H_L \tag{14}$$

（3）尝试更新最大的分数：

$$score = max(score, \frac{1}{2}\frac{G_L^2}{H_, + \lambda} + \frac{1}{2}\frac{G_R^2}{H_„ + \lambda} - \frac{1}{2}\frac{(G_L + G_R)^2}{H_, + H_„ + \lambda} - \gamma) \tag{15}$$

基于最大 score 对应的划分特征和特征值分裂子树。

如果最大 score 为 0，则当前决策树建立完毕，计算所有叶子区域的 w_{tj} ，得到弱学习器 $h_t(x)$ ，更新强学习器 $f_t(x)$,进入下一轮弱学习器迭代.如果最大 score 不是 0，则转到第（2）步继续尝试分裂决策树。

1.5.4 贝叶斯分类器

贝叶斯分类器是各种分类器中分类错误概率最小或在预先给定代价的情况下平均风险最小的分类器。它的设计方法是一种最基本的统计分类方法。其原理是利用贝叶斯公式通过先验概率计算出后验概率，选择具有最大后验概率的类作为该对象所属的类。

贝叶斯文本分类算法是一个经典的文本分类算法，其在预测一个未知类别的可能属性中有着较为详细的理论和实践基础。事件 B 发生的条件下事件 A 发生的概率可通过条件概率推导出：

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{16}$$

其一般化我们可以得到，其中集合$\{A_i\}$表示事件集合里的部分集合：

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_i P(B \mid A_j)P(A_j)} \tag{17}$$

针对文本分类主要存在着三种不同的贝叶斯模型：高斯模型、多变量的伯努利模型和多项式模型，根据以往的学者的研究经验，本文选取了后者，即多项式模型进行实验。该多项式贝叶斯分类模型算法的通用公式如下式：

$$P(w_k \mid c_i) = \frac{N_{ki} + 1}{\sum_{k=1}^{|V|} N_{ki} + |V|} \tag{18}$$

式中N_{ki}是w_k类别C_i的所有文档中出现的总次数，$|V|$是训练数据集的总单词数。

1.6 模型评价指标

分类模型中，经常使用查准率和查全率来对分类结果进行评价。查准率表示对于给定的测试数据集，分类模型正确分类的样本数与总样本数之比；查全率表示分类器实际正确分类的样本数与实际应该被分类的样本数之比。对于二分类问题其混淆矩阵如表 3 所示：

表 3 二分类问题混淆矩阵

真实值	预测值	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

根据表 3 可以得出查准率计算公式为：

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} \tag{19}$$

查全率的计算公式为：

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{20}$$

理想情况下，我们希望查准率和查全率都尽可能高，但实际上这两个指标呈现负相关，无法达到两个指标的同步最高。因此我们引入了 F-score 评价指标来对精确值和召回率的均值进行调和，F-score 表示为：

$$F\text{-score} = (1 + \beta^2) \times \frac{\mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{(\beta^2 \times \mathrm{Pr}\,ecision) + \mathrm{Re}\,call} \tag{21}$$

式中，β 是用来平衡 Precision，Recall 在 F-score 计算中的权重。本文中我们把 β 取为 1，表示 Precision 和 Recall 一样重要。综上所述，对于本文中的多分类问题，其 F-score 表示为：

$$F\text{-score} = \frac{1}{n}\sum_{i=1}^{n}\frac{2\mathrm{P}_i \cdot \mathrm{R}_i}{\mathrm{P}_i + \mathrm{R}_i} \tag{22}$$

式中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

1.7 结果分析

根据上述介绍的模型评价指标，将本文采用的 KNN 分类模型与其他几个主流的分类模型进行对比。首先，对附件 1 提供的数据进行预处理后，按照 10:1 的比例随机地分成训练集和测试集，即得到 4940 条训练样本，494 条测试样本。其中，4940 条数据作为训练样本，494 条数据作为测试样本。测试样本中各类样本数量分别为“城乡建设”1010 条、“环境保护”330 条、“交通运输”550 条、“教育文体”960 条、“劳动和社会保障”1030 条、“商贸旅游”480 条、“卫生计生”580 条。

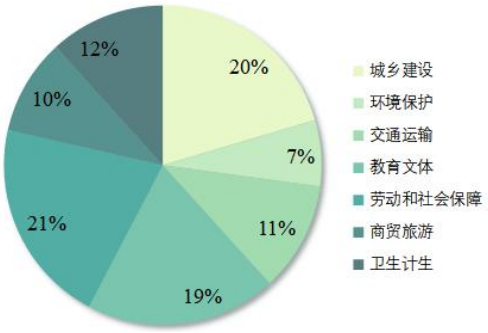


图 7 训练样本数据分布

本文采用 KNN 模型对文本进行分类，并分别与 SVM 模型、多项式贝叶斯分类模型、XG Boost 模型进行对比。并使用上文中提及的查准率、查全率和 F-score 对分类结果进行分析评价。实验中，KNN 模型的近邻个数 k=30；XG Boost 模型树最大深度设置为 8，样本特征随机采样比例为 0.7，学习率为 0.1，损失函数采用了均方误差（MSE）；SVM 模型采用线性核函数，最大迭代次数为 8；多项式贝叶斯分类模型拉普拉斯平滑值设为 1，不考虑先验概率。

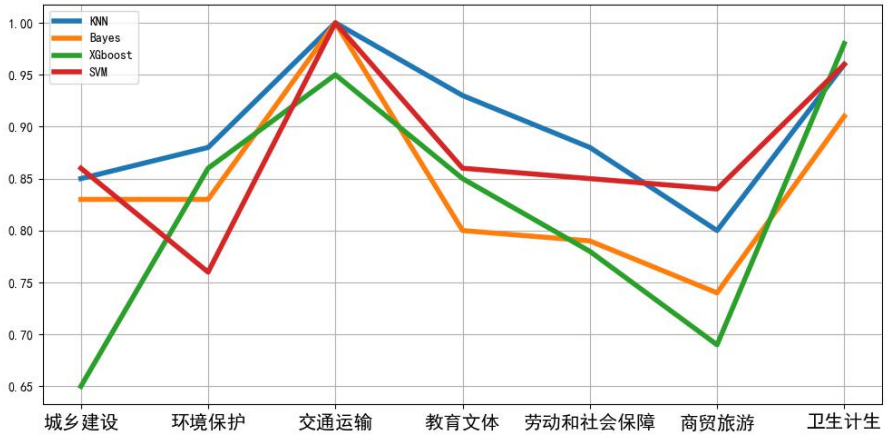


图 8 模型查准率比较

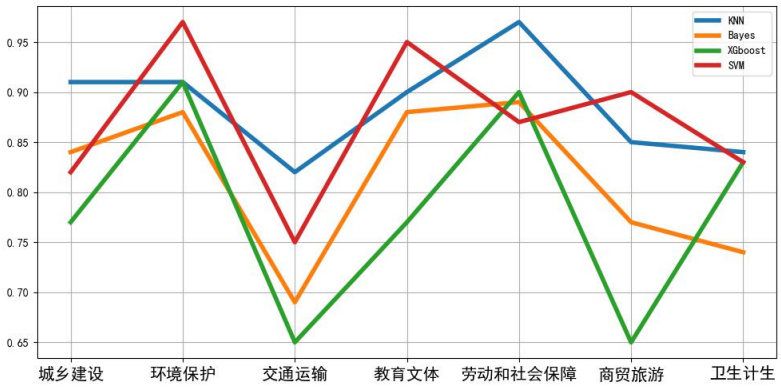


图 9 模型查全率比较

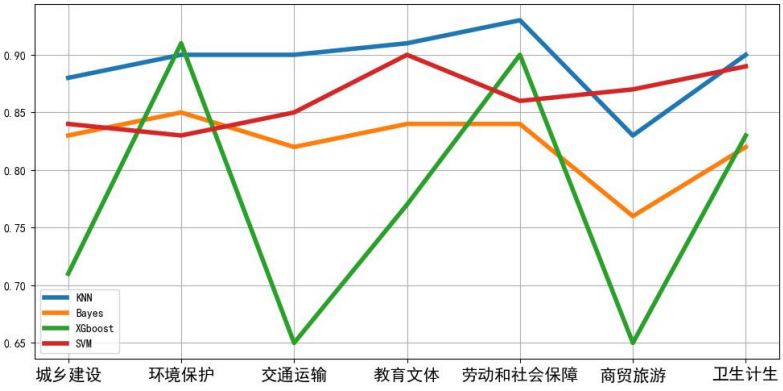


图 10 模型 F-score 比较

三项指标的四种模型比较折线图如图 8、9、10 所示。图 8、9 中可以看出，由于数据转化为 TF -IDF 数字向量后出现了极大的稀疏性，使 XG Boost 在分类时结果极其不稳定。在样本数量较少情况下，如“交通运输”，除 XG Boost 外，各分类模型的查准率均为 1，但 KNN 模型的查全率仍能保持较高的数值。在图 10 中，KNN 模型、SVM 模型和贝叶斯分类模型均具有稳定性，但 KNN 分类模型在大多类别的 F-score 值均最高，分类效果最好。

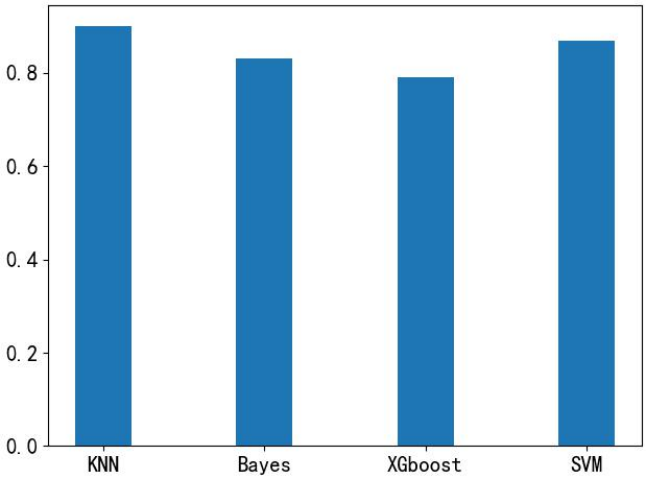
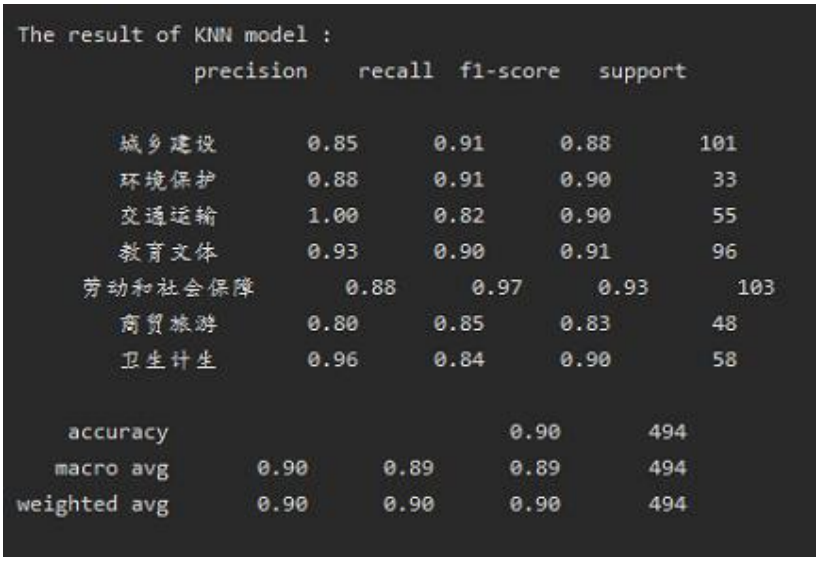


图 11 四类模型精度结果比较

综合对比折线图，KNN 模型作为本题选用分类模型，其分类效果最优。全部样本数据分为七类：城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生。精度为 0.90，查准率均值为 0.9，查全率均值为 0.89，F-score 均值为 0.9。用 KNN 模型进行分类结果如图 12 所示：



The result of KNN model :

	precision	recall	f1-score	support
城乡建设	0.85	0.91	0.88	101
环境保护	0.88	0.91	0.90	33
交通运输	1.00	0.82	0.90	55
教育文体	0.93	0.90	0.91	96
劳动和社会保障	0.88	0.97	0.93	103
商贸旅游	0.80	0.85	0.83	48
卫生计生	0.96	0.84	0.90	58
accuracy			0.90	494
macro avg	0.90	0.89	0.89	494
weighted avg	0.90	0.90	0.90	494

图 12　KNN 模型分类结果

二、热点问题挖掘

2.1　热点定义

“热门话题”是一个比较模糊的概念，题设要求某一时段内群众集中反映的某一问题可称为热点问题。故首先对热点问题进行定义。通过分析，热门问题均具备以下特征：

（1）话题代表性强。有多个留言问题的情况下，若有同类型人群多次重复地提及该问题，或者是同一地区反复出现该问题，我们都能够认为该问题是热度较高。

（2）话题讨论度高。在本题中，我们认为点赞数和反对数都对话题热度有着积极作用，均能够代表话题的讨论度。

（3）话题时间新颖。本题中我们用话题发表时间来量化话题新颖度。更具体的说说，在问题留言主题和内容有差异的前提下，发表时间越短、越接近当前时刻，表示问题越新颖。

综上所述，根据所给数据，本文用留言时间、留言主题、留言内容、留言点赞数与留言反对数等指标来描述热度问题排名情况。

对于话题代表性问题，我们利用问题 1 的数据预处理方法对留言主题、留言内容进行数据清洗，采用文本聚类方法计算留言的文本相似度，按照人名或者地名标准对文本

内容进行聚类分析，以总结同地区或同类人群前提下提及较多的问题。

对于话题讨论度问题，我们将附件 3 数据中留言的点赞数和反对数都看作能够提高话题讨论度的指标。计算每个聚类后的问题簇的话题讨论度，来对问题簇进行热度排名。

对于话题时间的问题，我们采用 Reddit 热度排名算法，将留言时间指标与当前时刻进行组合分析，将靠近当前时刻的问题热度排名适当靠前，最终利用公式计算总热度排名，并总结热点问题留言明细表。

2.2 文本聚类

2.2.1 聚类分析综述

聚类分析(cluster analysis)是将样本个体或指标变量按其具有的特性进行分类的一种统计分析方法。分为对样本进行聚类和对指标进行聚类两种。本题我们考虑对样本聚类，采用 K-means 聚类模型。

首先对附件 3 中的数据进行预处理，利用 TF -IDF 算法将文本向量转化为数字向量。并计算各个样本向量间的欧式距离，即各变量差值的平方和，用于度量文本的相似度。将相似度较高的留言聚类，减少留言内容的重复率，也为提取热度问题进行很好的数据处理。

聚类分析完全是根据数据情况来进行的。以 n 个样本、k 个特征变量组成的数据文件为例，对样本进行聚类分析，等价于对 k 维坐标系中的 n 个点进行分组，所依据的是它们的距离；当对变量进行聚类分析时，相当于对 n 维坐标系中的 k 个点进行分组，所依据的也是距离。所以距离或相似性程度是聚类分析的基础。

2.2.2 聚类思路

附件 3 的数据中，首先对留言内容和留言主题进行数据预处理并聚类，将文本相似度较高的问题归为一类；再对问题聚类簇中的地点和人群进行关键词提取，以总结热点问题的地点和人群。

2.2.3 K-means 聚类

（1）欧氏距离[7]：

令 $x_i=\left(x_{i1},\ldots x_{it},\ldots,x_{ik}\right)$ 是第 i 个样本观察值，$x_j=\left(x_{j1},\ldots,x_{jt},\ldots,x_{jk}\right)$ 是第 j 个样本观察值，两者之间的欧式距离定义为：

$$d_{ij}=\sqrt{\sum_{t=1}^{k}\left(x_{it}-y_{jt}\right)^2} \tag{23}$$

（2）K-means 聚类原理[8]

K-means 集群算法，1967 年由学者 J. B. MacQueen 所提出，也是最早的组群化计算技术。因为其简单易于了解使用的特性，对于球体形状(spherical-shaped)、中小型数据库的数据挖掘有不错的成效，可算是一种常被使用的集群算法。

K-means 聚类算法步骤：

1）任意选择 k 个对象作为初始的类的中心。

2）计算其余所有点到 K 个中心点的距离，并把每个点到 K 个中心点最短的聚簇作为自己所属的聚簇。

3）根据类中文档的平均值，将每个文档(重新)赋给最相近的类。

4）更新类的平均值。

5）直到分类结果不再发生变化时，即没有对象进行被重新分配时过程结束。

6）该算法试图找出使平方误差值最小的 k 个划分，可扩展性较好，对大数据集处理有较高的效率。

2.2.4 聚类结果分析

根据上文所述，采用 K-means 聚类时选择合适的类别数量 k 将至关重要。本文中将轮廓系数（Silhouette method）为指标对 k 值进行选取。轮廓系数衡量簇的密集与分散程度，其值在[-1, 1]范围内。Silhouette 值接近 1，说明对象与所属簇之间有密切联系，则选取轮廓系数最大时 k 的值。轮廓系数与 k 值的变化如图所示:

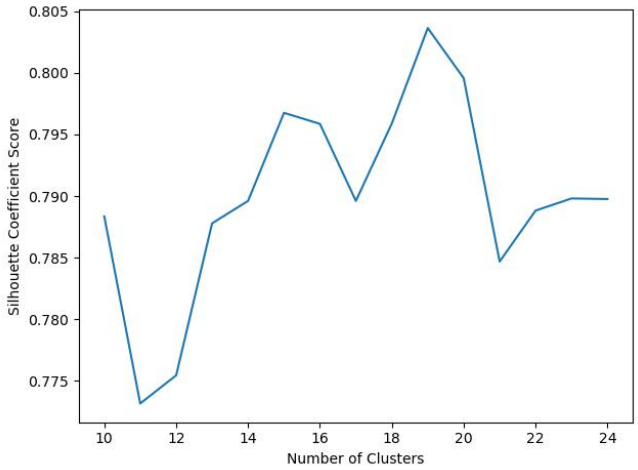


图 13 轮廓系数与 k 值变化图

本次测试从附件 3 中选取部分数据进行聚类，从图中可以在 k=19 时，模型的轮廓系数达到最佳，样本聚类合理。其聚类结果的三维可视化如图 14 所示：

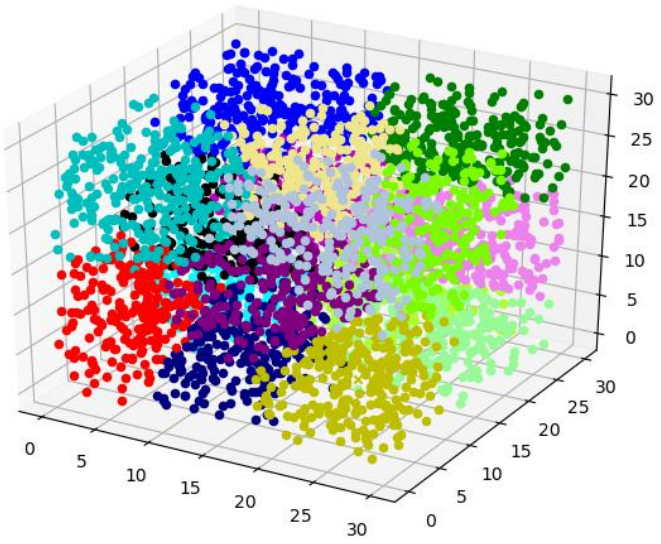


图 14 k=19 时聚类结果三维图

图 14 中相同颜色的点表示相同类中的样本。可以看出 k=19 时，各颜色的点交叉现象并不明显，即簇内的密集度较高。当类的个数为 19 时，其聚类结果如表 4 所示：

表 4 部分聚类结果

问题 ID	时间范围	问题描述
188416	2019-01-01 至 2020-01-08	A 市 X115 等待时间太久了在 A 市的居住证未满一年
188031	2019-01-08 至 2020-01-08	A7 县山水人家附近酒吧涉嫌酒托骗局 A7 县自来水公司费用收取太高了 A7 县黄花镇合心村将灌溉山林用的水塘用来堆土 A7 县楚龙路附近一餐馆的音乐声超级大 A7 县北山镇青田村徐家老屋旁火车声音特别大 A7 县开发商将消防通道当车位卖
188409	2019-01-02 至 2020-01-07	请 A 市加快高铁西站的建设力度 A 市地铁 1 号线北延二期有规划建至丁字湾街道
188073	2019-01-02 至 2019-12-17	咨询 A 市绿地海外滩二期与长懿高铁问题
.....		
189587	2019-01-06 至 2020-01-07	对 A8 县高新区复兴-艾家冲Ⅰ、Ⅱ线 500kV 线路杆迁工程临近居民区的质疑 A8 县科目三补考费要 600 一次
200610	2019-01-14 至 2020-01-02	还小区居民一个美丽的生活环境
188475	2019-01-09 至 2020-01-03	让 A6 区的孩子享受市区同样的教育资源 A6 区时代倾城万婴幼儿园多收取学费

2.3 热度排名算法

根据题中要求，为挑选出热度最高的五类问题，本节引入 Reddit 算法并对其进行改良。Reddit 是美国最大的网上社区，社区里每个帖子前面都有向上和向下的箭头，分别表示“赞成”和“反对”。用户通过点击进行投票，Reddit 根据投票结果计算出最新的“热点文章排行榜”。

2.3.1 Reddit 热度排名算法

Reddit 的热度得分计算公式如下：

$$Score=\log_{10} z+\frac{yt}{45000} \tag{24}$$

（1）文章的发表时间 t

$$t=\text{发表时间}-2005\text{ 年 }12\text{ 月 }8\text{ 日 }7\text{ 时 }46\text{ 分 }43\text{ 秒} \tag{25}$$

式中 t 用来标注文章新旧程度的单位为小时，而 Reddit 算法的单位为秒，其使用 Unix 时间戳（从 1970 年 1 月 1 日到当前时间的秒数）进行计算。通过上式可以看出 t 为固定值，即不随时间改变，且帖子越新，t 值越大。

发表时间和话题排名的影响可以被概括如下:发表时间对排名有较大影响，该算法使得新的话题比旧的话题排名靠前。话题的得分不会因为时间的流失而减少，但是新的话题会比旧的话题得分高。

（2）赞成票与反对票的差 z

$$z=\text{赞成票}-\text{反对票} \tag{26}$$

由于 Reddit 提供了投反对票的功能，所以可以使一些具有争议的话题会排的较后。z 表示赞成票超过反对票的数量。如果赞成票少于或等于反对票，则 z 等于 1。z 表示问题的受肯定程度。

得分计算公式中的 $\log_{10} z$ 表示赞成票超过反对票的数量越多，得分越高。这里用以 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。更具体的，前 10 个投票人与后 90 个投票人（乃至再后面 900 个投票人）的权重是一样的，即如果一个帖子特别受到欢迎，那么越到后面投赞成票，对得分越不会产生影响。而当反对票超过或等于赞成票，z=1，因此这个部分等于 0，也就是不产生得分。

（3）投票方向 y

y 是一个符号变量，表示对话题的总看法，用来产生正分和负分。当赞成票超过反对票时，得分为正；当赞成票少于反对票时，得分为负；当两者相等，得分为 0。y 是话题评价的一种定性表达，0 表示没有倾向，大于 0 表示正面评价，小于 0 表示负面评价，保证了得到大量净赞成票的文章，会排在前列；得到大量净反对票的文章，会排在

最后。

得分计算公式中 yt 与 45000 的比值则表示 t 越大得分越高，即新话题的得分会高于老话题。它起到自动将老话题的排名往下拉的作用。分母的 45000 秒，等于 12.5 个小时，也就是说，后一天的话题会比前一天的话题多得 2 分。

2.3.2 Jyreddit 热度排名算法

结合本题具体情况，我们利用附件 3 数据中的留言时间、点赞数与反对数三要素进行排名指标。需注意，本题假设不管是点赞数还是反对数都对该问题的热度提高有着正向作用，而不进行抵消，故本文对 Reddit 热度排名算法进行改进，并将改进后的算法称为 Jyreddit 热度排名算法。

（1）文章的发表时间 t

$$t = \text{发表时间-2020 年 5 月 4 日} \tag{27}$$

t 用来标注文章新旧程度的单位为小时，而 Jyreddit 的单位为秒，其使用 Unix 时间戳进行计算。通过上面的公式可以看到一旦话题发表，t 就是固定值且是负值，不会随时间改变，而且话题越新，t 值越大。

（2）赞成票与反对票的和 d

$$d = \text{赞成票+反对票} \tag{28}$$

相比于 Reddit 算法，本题中我们不讨论反对数对问题产生的消极影响，只考虑反对数对问题热度所产生的影响，因此我们认为不管是点赞数还是反对数，都是对热度排名有着正向作用。从而我们用 d 表示留言点赞数和留言反对数之和。

（3）投票方向 y

y 是一个符号变量，由于本题将点赞数还是反对数视为一体，因此投票均为正向，在这里默认 y 固定为 1。

综上所述，Jyreddit 热度排名算法公式为：

$$Score = \log_{10} d - \frac{t}{45000} \tag{29}$$

2.4 热点问题汇总与分析

首先，使用 Python 中 jieba 分词对附件 3 中部分数据的留言内容进行分词并统计词频，得到词频最高的前 20 个词，如图 15 所示。

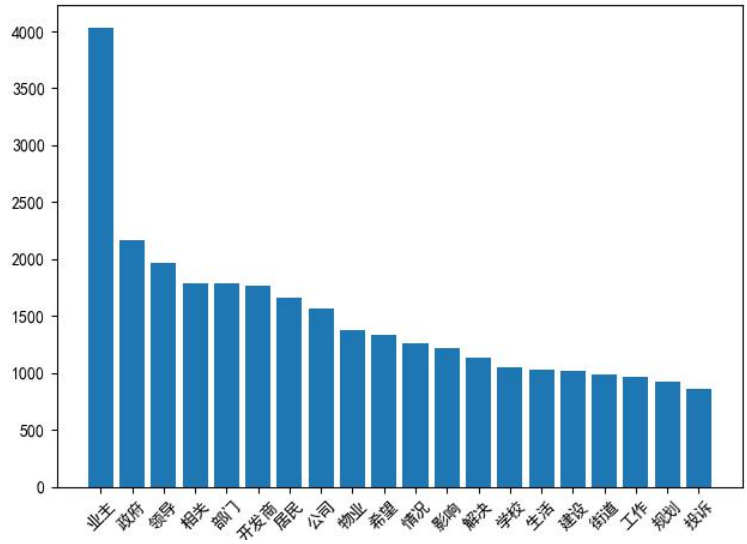


图 15 词频

从得到的前 20 个词频上看，留言内容涉及的话题主要围绕“业主”、“居民”、“生活”、“学校”、“街道”、“工作”等关键词展开。图中，关键词“业主”、“政府”的出现频率分别达到了 4030 次和 2168 次，表明留言中大部分谈论内容可能与城乡建设中的房屋拆迁、物业保障等住房保障与房地产问题相关。关键字“学校”、“生活”、“居民”出现频率分别达到了 1051、1028 和 1664 次，占总数 12%。表明群众留言的内容中，劳动、教育问题也占据了一定比重。

基于代表性、讨论度、新颖性三个角度，采用 K-means 聚类算法对上述样本数据进行聚类提取热点问题。根据 Jyreddit 算法对每个样本进行热度值的计算，排名前 5 的热点问题。根据题目要求，为了获取留言所反映问题的发生地点，将每一类热点问题的问题描述进行分词，即把一句话划分为单个字词。将分词后的每一个字词与“市”、“县”、“区”等关键字进行匹配，从而获取问题发生地点。其结果如表 4 所示。

表 4 排名前 5 的热点问题

热度排名	问题 ID	热度值	时间范围	地点人群	问题描述
1	200302	10	2019-01-03 至 2020-01-01	A 市 2 号社区	A 市东山湾路什么时候修？
2	200610	9. 39	2019-01-14 至 2020-01-02	A6 区金悦花园	还小区居民一个美丽的生活环境
3	200316	9. 26	2019-01-03 至 2020-01-04	A 市慈济医院	A 市住建委不能认定我与开发商签定商品房买卖合同
4	201329	8. 69	2019-01-22 至 2019-12-11	A 市 A4 区	咨询 A 市绿地海外滩二期与长懿高铁问题

5	200611	8.34	2019-01-02 至 2019-12-26	A 市 A5 区	请问关于六号路的相关验收数据已经移交交警部门了吗？

为验证热点问题评判系统的鲁棒性，从附件 3 中随机抽取 2486 条数据，并根据第一问中 KNN 分类模型建立关于留言内容的一级标签分类模型，其结果如图 16 所示。其中城乡建设、 环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生的数量分别为 568、168、280、502、580、186、202。显然，城乡建设问题、劳动和社会保障和教育文体问题出现的次数最多，与上述分析相符。

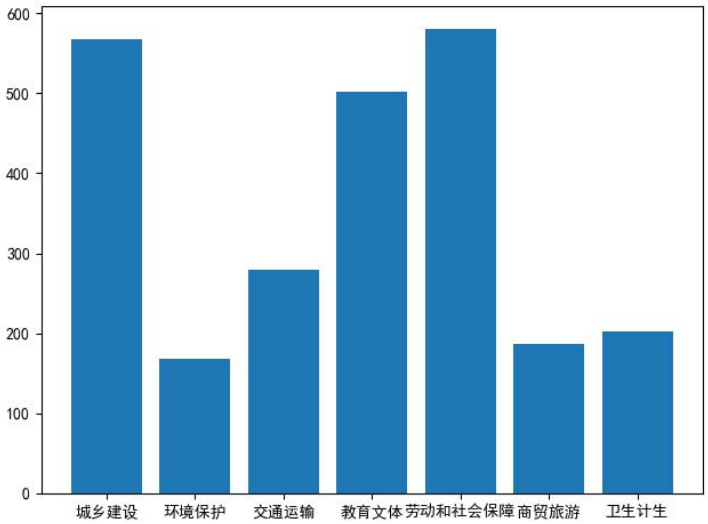


图 16 各类别样本数量

三、答复意见质量评价方案

3.1 答复意见质量分析

根据附件 4 相关部门对留言的答复意见，我们从答复的相关性、完整性、可解释性、这三个角度对答复意见的质量给出评价方案[9-11]。通过对数据答复意见的分析可知三个角度的评判思路为：

（1）相关性：留言答复的相关性指的是用户留言内容与答复意见的关联程度，即答复意见能够与留言内容匹配，且做到对留言内容中相关意见和问题都有对应的回复。

（2）完整性：完整性一方面体现在答复内容句式结构的完整，没有文本的缺失；另一方面体现在对用户留言的答复全面且详尽。

（3）可解释性：可解释性体现在基于相关性和完整性无法评判的答复意见，却合理且具有可解释性。例如：用户留言内容不属于回复部门处理范围，需要转交给其他部门处理的答复意见、已经回复过留言内容的答复意见。

3.2 数据预处理

为了方便对留言和答复意见进行对比分析得到质量评价方案，对所有数据中的留言主题、留言内容以及答复意见进行中文分词，这里用到的是 Python 中文分词包 Jieba 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。

3.3 相关性分析

相关性从句子和关键词两个方面进行评价。首先，由于留言内容中提出的问题一般有多个，故采用 BM25 算法对留言内容和答复意见进行文本相似度计算，确保留言内容中所提问题答复意见都有涉及，每个答复都没有跑题。相似度越高则说明答复意见与留言内容相关性越好。由于计算时可能会由于一些无意义的词，如“的”、“我们”等字词大量的重复而导致相似度的值比较高，故仅靠这一点评判相关性欠缺说服力。为解决该问题，引入关键词匹配。即利用 TextRank 算法对留言答复和留言内容进行关键字提取，并根据 BM25 算法计算留言内容和答复意见中提取关键词的相关性，结果表示答复是否抓住留言核心问题。下面两小节将介绍相关性评价用到的提取留言内容关键词的 TextRank 算法、以及计算留言内容与答复意见之间文本相似度的 BM25 算法。

3.3.1 TextRank 关键词提取算法

本项目调用了 python 写的类库 snowNLP 提取常规答复数据集里留言内容关键词。其中用到的是关键词提取 TextRank 算法。TextRank 算法是一种用于文本的基于图的排序算法[12]。其基本思想来源于谷歌的 PageRank 算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘。TextRank 算法提取关键词步骤如下：

（1）把给定的文本 T 按照完整句子进行分割，即对于每个句子进行分词和词性标注处理，并过滤掉停用词，只保留指定词性的单词，如名词、动词、形容词。这些词为保留后的候选关键词。

（2）构建候选关键词图 G = (V,E)，其中 V 为节点集，由步骤 1 生成的候选关键词组成。采用共现关系（co-occurrence）构造任两点之间的边，两个节点之间存在边仅当它们对应的词汇在长度为 K 的窗口中共现，K 表示窗口大小，即最多共现 K 个单词。

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{30}$$

（3）根据上面公式迭代传播各节点的权重，直至收敛。

（4）对节点权重进行倒序排序，从而得到最重要的 T 个单词，作为候选关键词。

（5）由步骤 4 得到最重要的 T 个单词，在原始文本中进行标记，若形成相邻词组，则组合成多词关键词。如果文本中出现多个关键词相邻的情况则这些关键词可以构成一个关键短语。

TextRank 算法提取关键字结果见附录（二）所示。

3.3.2 BM25 文本相似度算法

BM25 算法通常用来作搜索相关性评分[13]。其主要思想是对 Query 进行语素解析，生成语素 q_i；然后，对于每个搜索结果 D，计算每个语素 q_i 与 D 的相关性得分，最后，将 q_i 相对于 D 的相关性得分进行加权求和，从而得到 Query 与 D 的相关性得分。

BM25 算法的一般性公式如下：

$$Score(Q,d)=\sum_{i}^{n}W_i\cdot R(q_i,d) \tag{31}$$

式中 Q 表示 Query，q_i 表示 Q 解析之后的一个语素（对中文而言，我们可以把对 Query 的分词作为语素分析，每个词看成语素 q_i）；d 表示一个搜索结果文档；W_i 表示语素 q_i 的权重；$R(q_i, d)$ 表示语素 q_i 与文档 d 的相关性得分。定义 W_i 来判断一个词与一个文档的相关性的权重，方法有多种，较常用的是 IDF。这里以 IDF 为例，公式如下：

$$IDF(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \tag{32}$$

式中 N 为索引中的全部文档数，$n(qi)$ 为包含了 q_i 的文档数。根据 IDF 的定义可以看出，对于给定的文档集合，包含了 q_i 的文档数越多，q_i 的权重则越低。也就是说，当很多文档都包含了 q_i 时，q_i 的区分度就不高，因此使用 q_i 来判断相关性时的重要度就较低。语素 q_i 与文档 d 的相关性得分为 $R(q_i, d)$。BM25 中相关性得分的一般形式：

$$R(q_i,d)=\frac{f_i(k_i+1)}{f_i+K}\cdot\frac{qf_i\cdot(k_2+1)}{qf_i+k_2} \tag{33}$$

$$K=k_1\cdot(1-b+b\cdot\frac{dl}{avgdl}) \tag{34}$$

式中，$k1$，$k2$，b 为调节因子，设置 $k1=2$，$b=0.75$；f_i 为 q_i 在 d 中的出现频率，f_i 为 q_i 在 Query 中的出现频率。dl 为文档 d 的长度，$avgdl$ 为所有文档的平均长度。由于绝大部分情况下 q_i 在 Query 中只会出现一次，即 $q_i=1$，因此公式可以简化为：

$$R(q_i,d)=\frac{f_i(k_i+1)}{f_i+K} \tag{35}$$

从 K 的定义中可以看到，参数 b 的作用是调整文档长度对相关性影响的大小。b 越大，文档长度的对相关性得分的影响越大，反之越小。而文档的相对长度越长，K 值将

越大，则相关性得分会越小。即当文档较长时，包含 q_i 的机会越大。因此，同等 f_i 的情况下，长文档与 q_i 的相关性应该比短文档与 q_i 的相关性弱。BM25 算法的相关性得分公式可总结为：

$$Score(Q,d)=\sum_{i}^{n} IDF(q_i)\cdot\frac{f_i(k_1+1)}{f_i+k_1(1-b+b\cdot\frac{dl}{avgdl})} \tag{36}$$

从 BM25 的公式可以看到，通过使用不同的语素分析方法、语素权重判定方法，以及语素与文档的相关性判定方法，可以衍生出不同的搜索相关性得分计算方法，这对设计算法提供了较大的灵活性。

3.3.3 相关性评价

相关性的评分将从两方面评判：留言内容与答复意见的关键词匹配分数；留言内容与答复意见的文本相似度分数。将两个分数相加即得到相关性的得分。

3.4 完整性分析

通过数据分析，发现答复意见不完整有以下两种情况，如图 17 所示：i）答复意见内容中只出现日期、抬头的文本缺失情况。ii）答复意见只说明已经收悉用户留言，但没有具体的回复。文本的不完整会丢失答复所表达的意义，导致答复没有实际意义。从不完整的答复意见中可以发现，文本的完整性与答复意见字数有一定的相关性，即答复意见的文本字数越少，则答复的内容越不具体且不能很好地表达答复意见，文本的完整性也就相应的较差。

留言编号	答复意见
37459	2019年1月14日
25918	“UU0081182”
94701	国网西地省J9县供电公司
125462	L1区网宣办　2015年9月18日
130767	“UU0082103”　　感谢您对L2县的关注与支持。
166919	A00085038：您在《问政西地省感谢您的理解与支持！2018年8月6日
117388	“UU008582”　　你好！2019年2月18日
122596	经我办协调，您反映的情况L2县下坪乡回复如下：
30019	已收悉
12189	网友：您好！留言已收悉

图 17 不完整答复意见样例图

对 3.3 节中所述的文本相关性评分进行排序，不难发现：相关性分数较低的答复意见字数一般都较少，在完整性上也较差。因此本文中将以答复意见的文本字数为标准对文本完整性进行评价，评分将以答复意见文本字数乘以 0.01 表示。如下图 18 所示，相

关性评分和完整性评分较为契合。

图 18 相关性和完整性评分对比图

3.5 可解释性分析

前两节对数据已经进行了相关性评分及完整性评分，两项评分相加排序后再对数据进行分析发现：即使两项评分很低，但是部分答复意见是具有可解释性，如抽取的样例图 19。也就是说，对于部分答复意见的评价不能仅依靠于相关性和完整性，只要答复意见合理且具有可解释性，这类答复意见也是质量较好的答复。

留言编号	答复意见	评分总和
123860	网友：您好！您反映的问题已转L市职院调查、核处。	0.52
88291	您的留言已收悉！我们已将您反映的问题转相关部门进行处理，敬请关注后续回复，谢谢！	0.79
120364	网友：您反映的问题已转红十字会。如有问题，请电话咨询市红十字会，电话：0745-2235116。	0.93
33670	“UU008572”您好！您通过平台《问政西地省》的留言收悉，已交B2区及有关部门调查核处，如有相关情况将及时反馈，谢谢！2018年12月28日尊敬的“邹建礼”网友：您好！您通过平台《问政西地省》的留言收悉，已交我区有关部门调查核处，如有相关情况将及时反馈，谢谢！2019年1月2日	3.87
51521	网友：您好！您所反映的问题，已进行过回复。	0.80
126023	网友“UU0081524”您好！您反映的问题相关部门已经答复。链接：https://baidu.com/	1.25

图 19 样例相关性和完整性评分总和图

对相关性和完整性评分分析可知，具有可解释性的答复意见分为两类：一类是需要转交给其他部门处理的答复意见；另一类是已回复过留言内容的答复意见。其他答复仍以相关性和完整性来评价。由于评分总和在 4.00 分之后基本上都是常规答复，则从评分总和中低于 4.00 分的答复意见分别以“转”、“交”、“咨询”、“已”“回复”为关键词筛选出具有可解释性的答复意见。其可解释性得分为 20.00 分。

3.6 质量评价体系

评价质量将以百分制分数来体现，为了将所有数据的评分放在 0-100 的范围内，对分数做归一化处理，即将分数放缩至 0-100 范围内。整个质量评价流程如下图 20 所示。

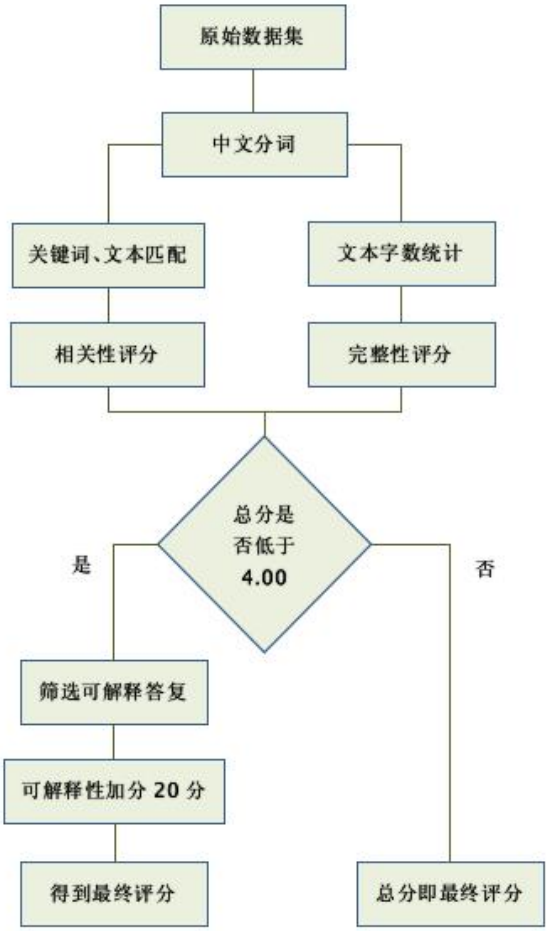


图 20 质量评价体系流程图

原始数据集预处理后经过留言内容与答复意见关键词匹配、留言详情与答复意见文本匹配得到相关性评分，再通过答复意见字数缩放 100 倍得到完整性评分，两项评分和为答复意见质量的基础得分。从低于 4.00 基础得分的答复意见中筛选出具有可解释性的答复意见，对其进行可解释性加分 20 分。最后整理出所有数据中答复意见的质量评价得分。如下图 21 所示为部分样例的最终评分。

留言编号	答复意见	最终评分
6556	已收悉	0.34
51521	网友：您好！您所反映的问题，已进行过回复。	20.80
58419	网友：你好！你反映的问题，我们已转交电力部门。2019年3月28日	21.03
76633	您的留言已收悉。关于您反映的问题，已转F7县调查处理。网友：你好！您在2019年1月17日反映的关于吴根福反映的有关“桃林村低保补助问题”诉求的信访件，我镇党委、政府高度重视，配合县相关部门单位，立即成立调查组对信访人反映的问题进行调查处理。现将有关情况汇报如下： 一、基本情况 信访人吴根福，男，身份证号码：**********，现年58岁，系我镇桃林村村民。其户籍人口为4人。。。	78.83
96762	“UU0082390”首先感谢您对J9县扶贫工作的支持与关注！针对您咨询的关于加快推进原生态有机农产品品牌建设的决定和农产品品牌建设扶持政策，现回复如下：为充分发挥我县的生态优势，打造有机农产品生产基地，助力精准扶贫，2016年7月，县委县政府出台了《关于加快推进原生态有机农产品品牌建设的决定[政府发文]5号，以下简称：决定），现已实施三年，并取得了一定成效。下一步，县委政府根据《决定》的实施情况将进行修订，明年起，扶持政策待文件出台后才能确定。[政府发文]5号文件。2018年12月24日附[政府发文]5号中共J9县委J9县人民政府关于加快推进原生态有机农产品品牌建设的决 定（2016年7月7日）推进农产品品牌建设，是贯彻落实市委、市政府“产业主导、全面发展”理念的重要内容，是大力“兴产业、强实体、提品质、增实效”的有效举措。为加快我县积连片1万平方米以上，每平方米另奖补10元。小水果。新种、全垦整地基地连片50亩市级金奖的产品设作出突出贡献的乡镇、单位、新型农业经营主体、新型农业服务组织和新型农产品销售主体。本决定自印发之日起施行。（此件发至乡镇级）。。。	95.00

图 21 样例的最终评分结果图

四、总结与展望

4.1 总结

为解决群众留言问题处理低效、工作量大的问题，我们首先对留言信息进行分类。我们首先对文本向量进行数据预处理，即去除重复项，对文本内容进行分词在通过 TF-IDF 算法转化为数字向量便于计算。然后在 PyCharm 开发平台上构建 KNN 分类模型并对样本进行分类，通过十折交叉验证对算法准确性进行测试。并引入 Precision、Recall、F-score 三项评价指标对模型的有效性进行评价排序，最终模型的分类精度为 0.9。

紧接着采用 K-means 算法对问题进行聚类，即把内容相似的问题归为一类，找出一段时间的热点问题。基于群众留言的点赞数、反对数和留言发表时间，并运用 Jyreddit 算法计算出每条留言的热度值，统计每一类问题的热度，将热度排名前五的问题提取出来。

最后从基于相关性、完整性、可解释性这三个角度对相关部门所作答复进行质量评价。利用 TextRank 算法提取关键字，在用 BM25 算法计算文本的相似程度。根据数据分析结果采用答复字长来初步判断完整性和可解释性，并通过相应关键字判断答复字数较少的数据是否具有可解释性。

4.2 展望

4.2.1 文本向量化改进——矩阵分解

文本数据进行去重、分词后采用 TF -IDF 算法将文本向量转化为数字向量。TF -IDF 算法根据字词出现在文档中的次数和某个含有指定字词的文档在文档集中出现的次数将文本向量转化为数字向量。显然，随着文本数量的增加，出现的字词种类呈正比增长，而部分文本中的字词数量却可能很少，从而导致矩阵具有极大的稀疏性。数据的极大稀疏性将使得许多算法的精度不佳。

为解决上述问题，可对稀疏矩阵进行分解，如奇异值分解（SVD）、几何均值分解（GMD）等，提取出主要信息。然后在用分类模型进行分类，其计算精度和计算时长均会有提升。

4.2.2 地点/人名提取

对于地点/人名的提取文中采用了分词后匹配关键字的方法提取。但该方法需要人工对数据进行分析，工作量较大，且在大数据中并不能有效的提取出准确的地点和人名。目前随着深度学习技术的发展，可尝试与深度学习技术相结合，进而有效的提取地名和人名。

参考文献

[1] 郭鹏杰.“网络问政”是健全社会矛盾释放机制的新途径[J]. 兰州大学.2010.07.

[2] MOSHE SNIEDOVICH,"Dynamic Programming and Principles of Optimality,"Journal of Mathematical Analysis and Application,65, 586-606 (1978).

[3] Ghahramani,"An Introduction to Hidden Markov Models and Bayesian Networks," International Journal of Pattern Recognition and Artificial Intelligence 2001,15(1):9-42.

[4] 周志华. 机器学习[M]. 北京:清华大学出版社. 2016.01.

[5] 李航. 统计学习方法[M]. 北京:清华大学出版社. 2012.03.

[6] Tianqi Chen,Carlos Guestrin,"XG Boost: A Scalable Tree Boosting System," in arXiv:1603.02754v3 [cs.LG] 10 Jun 2016.

[7] 张振亚, 王进等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005, 32(9):160-163.

[8] 任远航，面向大数据的 K-means 算法综述[J], 计算机应用研究, 2020.

[9] 刘高军, 马砚忠, 段建勇. 社区问答系统中“问答对”的质量评价[J]北方工业大学学报,2012,24(3).
[10] 蒋楠, 王鹏程, 社会化问答服务中用户需求与信息内容的相关性评价研究——以“百度知道”为例[J]. 信息资源管理学报.2012(03).
[11] 李晨, 巢文涵, 陈小明, 李舟军. 中文社区问答中问题答案质量评价和预测[J]. 计算机科学.2011(06).
[12] P. Wongchaisuwat,"Automatic Keyword Extraction Using TextRank,"
2019 IEEE 6th International Conference on Industrial Engineering
and Applications (ICIEA), Tokyo, Japan, 2019, pp. 377-381.
[13] M.Murata,H.Nagano,R.Mukai,K.Kashino and S.Satoh,"BM25 With
Exponential IDF for Instance Search," in IEEE Transactions on Multimedia,vol.16,no.6,pp.1690-1699,Oct.2014.

附录

（一） 分词结果

对于使用 Jieba 分词时，是否采用 HMM 模型的分词结果对比如下表所示：

原始文本	2019 年 4 月以来，位于 A 市 A2 区桂花坪街道的 A2 区公安分局宿舍区（景蓉华苑)出现了一番乱象，该小区的物业公司美顺物业扬言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了(原水电在小区买，水 4.23 一吨，电 0.64 一度）所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交 20 万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪？

精准模式 （采用 HMM 模型） 分词后的结果	2019/年/4/月/以来/，/位于/A/市/A2/区/桂花/坪/街道/的/A2/区/公安分局/宿舍区/（/景蓉华苑/)/出现/了/一番/乱象/，/该/小区/的/物业公司/美顺/物业/扬言/要/退出/小区/，/因为/小区/水电/改造/造成/物业公司/的/高昂/水电费/收取/不了/(/原/水电/在/小区/买/，/水/4.23/一吨/，/电/0.64/一度/）/所以/要/通过/征收/小区/停车费/增加收入/，/小区/业委会/不知/处于/何种/理由/对/该/物业公司/一再/挽留/，/而/对/业主/提出/的/新/应聘/的/物业公司/却/以交/20/万/保证金/，/不能/提高/收费/的/苛刻/条件/拒之门外/，/业委会/在/未/召开/全体/业主大会/的/情况/下/，/制定/了/一/高昂/收费/方案/要/各/业主/投票/，/而/投票/不/采用/投票箱/只/制定/表格/要/物业公司/人员/这一/利害关系/机构/负责/组织/，/对/投票/业主/隐私权/没有/任何/保护/，/还/对/投/反对票/的/业主/以/领导/做/工作/等/方式/要求/改变/为/同意/票/，/这种/投票/何来/公平/公正/公开/，/面对/公安干警/采用/这种/方式/投票/合法性/在/哪/？/
搜索引擎 （不采用 HMM 模型） 分词后的结果	2019/年/4/月/以来/，/位于/A/市/A2/区/桂花/坪/街道/的/A2/区/公安/安分/分局/公安分局/宿舍/宿舍区/（/景蓉华苑/)/出现/了/一番/乱象/，/该/小区/的/物业/公司/物业公司/美顺/物业/扬言/要/退出/小区/，/因为/小区/水电/改造/造成/物业/公司/物业公司/的/高昂/水电/电费/水电费/收取/不了/(/原/水电/在/小区/买/，/水/4.23/一吨/，/电/0.64/一度/）/所以/要/通过/征收/小区/停车/车费/停车费/增加/加收/收入/增加收入/，/小区/委会/业委会/不知/处于/何种/理由/对/该/物业/公司/物业公司/一再/挽留/，/而/对/业主/提出/的/新/应聘/的/物业/公司/物业公司/却/以交/20/万/保证/保证金/，/不能/提高/收费/的/苛刻/条件/之门/门外/拒之门外/，/委会/业委会/在/未/召开/全体/业主/大会/业主大会/的/情况/下/，/制定/了/一/高昂/收费/方案/要/各/业主/投票/，/而/投票/不/采用/投票/票箱/投票箱/只/制定/表格/要/物业/公司/物业公司/人员/这一/利害/关系/利害关系/机构/负责/组织/，/对/投票/业主/隐私/隐私权/没有/任何/保护/，/还/对/投/反对/反对票/的/业主/以/领导/做/工作/等/方式/要求/改变/为/同意/票/，/这种/投票/何来/公平/公正/公开/，/面对/公安/干警/公安干警/采用/这种/方式/投票/合法/合法性/在/哪/？/

（二）关键字提取结果展示

TextRank 算法提取部分留言内容关键字结果如下所示：

留言内容	关键字提取
2019 年 4 月以来，位于 A 市 A2 区桂花坪街道的 A2 区公安分局宿舍区（景蓉华苑)出现了一番乱象，该小区的物业公司美顺物业扬言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了(原水电在小区买，水 4.23 一吨，电 0.64 一度）所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交 20 万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪？	物业/公司/小区/投票/业主/不/方式/公安/业/高昂
潇楚南路从 2018 年开始修，到现在都快一年了，路挖得稀烂用围栏围起，一直不怎么动工，有时候今天来台挖机挖两几下，过几天又来挖几下，对当地的交通和店面的生意带来很大影响，里面的车出去和外面的车进来要绕很大一个圈，很不方便，请有关部门对此监管一下，这路修的时间也太长了，至少可以一段一段的修好，方便街上的老百姓出行。	很/挖/路/大/修/绕/年/生意/带来/南
地处省会 A 市民营幼儿园众多，小孩是祖国的未来，但民营幼儿园教师一直都是超负荷工作且收入又是所有行业最低，甚至连养老和医疗金都没交，在国家大力倡导普惠型幼儿园的同时更是加大了教师的工作压力，在降低成本的同时还增加了学生数量，让本来就喘不过气的教师更是雪上加霜，希望市委市政府加快提高民办幼儿园教师工资待遇水平和降低工作压力有何具体政策和行动？	幼儿园/教师/工作/都/降低/更是/压力/超负荷/提高/民办
尊敬的书记：您好！我研究生毕业后根据人才新政落户 A 市，想买套公寓，请问购买公寓能否享受研究生 3 万元的购房补贴？谢谢。	研究生/公寓/3/新/享受/万/政/人才/尊敬/您好
建议将“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”	马坡/岭/更名/口/小学/竹坡路/白/原/取消/建议
601 小区钻石新村 20 栋楼下快乐休闲网吧扰民：1、私自凿开楼道墙壁开门。2、网吧空调污水直接排放到过道。3、网吧 24 小时营业，对外排放噪音高达 80 分贝严重影响居民生活和休息。	网/排放/贝/分/达/80/凿/20/私自/栋

近段时间夜晚和国假日醴官公路有大量的外地牌照超载超高大货车快速行驶，醴官公路本来刚好两台车的宽度，而且已经到处坑坑洼洼，超载大货车严重超载车速快，致使当地道路险象环生，希望相关部门能够严查！	公路/官/超载/醴/大/货车/刚好/超高/国假日/牌
想了解渔民机动船柴油补贴标准，今年只发了 900 元每条船，觉得太少不合理！	机动船/补贴/渔民/柴油/条/900/元/只发/船/太
B 市 B3 区田心街道田红路中段一直以来是 B 市公交发展公司 T35 路公交线“九方中学”首末站。为了便于公交车辆调头，公交站区域的路面修得格外宽阔。可就是因为此处路面宽、上下车人流量大，随着近段时间天气越来越热，各路水果商贩都纷纷盯上了这块做生意的黄金宝地。每到傍晚时分或双休日，各路水果商贩们或推着三轮电动车、或开着小货车争前恐后地在公交站区域抢占有利位置，横七竖八地摆开售卖各种水果。	公交/站/开/路面/水果/市/区域/商贩/B/这块