“智慧政务”中的文本挖掘应用

摘要

为了更好地适应大数据时代要求，更好地提高政务处理的效率，本文对留言文本进行数据处理，并采用机器学习的方法，用朴素贝叶斯的分类方法对留言进行分类；用 K-Means 聚类的方法研究热点问题；最后本文建立答复评价指标体系，用 LSH 解决答复相关性问题，PMI 解决答复完整性问题，TextRank 解决答复可解释性问题。结果证明，朴素贝叶斯的分类方法的准确率高达 85.81%，是一种对文本分类较为高效的分类方法。除此之外，K-Means 也能有效提取热点问题，帮助相关部门优先解决紧急、呼声高的民生问题。最后，我们建立指标体系解决文本中的相关性、完整性和可解释性，在一定程度上提高了管理效率，使得政务评价更科学、更公平公正。

"Political wisdom" in the text mining applications

Abstract

In order to meet the requirements of big data era and improve the efficiency of government affairs processing, this paper uses machine learning method and Naive Bayesian Classifier method to classify the message In the end, this paper establishes the Evaluation Index System of reply, uses LSH method to solve the reply correlation problem, PMI method to solve the reply integrity problem, TextRank method to solve the reply interpretability problem. The result shows that the accuracy of Naive Bayesian Classifier method is up to 85.81% , thus it is an efficient method for text classification. In addition, K-Means can effectively pick up hot issues and help relevant departments prioritize urgent, high-volume livelihood issues. Finally, we establish an index system to solve the relevance, integrity and interpretability of the text, to a certain extent, improve the management efficiency, make the government evaluation more scientific, more fair and just.

目录

一、问题重述

1.1 问题背景

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，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的工作带来了极大挑战。同时随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的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。

1.2 需要解决的问题

（1）在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。根据给出的数据，建立关于留言内容的一级标签分类模型。

（2）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，并给出评价结果，给出 排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

（3）针对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，并尝试实现。

二、模型假设

根据目前已获得的数据以及软件技术限制，本文做出以下假设：

1、假设所获得的数据是真实的、可靠的

2、假设群众留言是正常反应城市问题而不是恶意刷留言

3、假设留言中的英文地名不会对结果造成影响

4、数据的相似度是保持最大可能的相似度，其中的误差忽略不计

三、符号说明

名称	说明
D	文档总数
N_i	出现词语 i 的文档个数
t_k	该文档中出现过的单词
$\|V\|$	训练样本中包含多少种单词
V	训练样本的单词表
Accuracy	所以的预测正确（正类负类）的占总的比重
Precision	正确预测为正的占全部预测为正的比例
Recall	正确预测为正的占全部实际为正的比例
A	判断矩阵
Z	答复意见的质量
O	欧几里得衡量距离

四、问题一的分析与求解

4.1 文本挖掘基础理论知识

4.1.1 分词

我们的数据中的评论是以句子组成的若干段落，我们对这一部分的文本进行研究时，对整个段落或者整个句子进行判别显然是不现实的，故而我们先把句子化整为零：把评论数据的段落分成句子，句子分成短语，短语再分成词汇。对于中文文本分词，根据相关文献论述，目前已经有几种现成的分词模型产生。

- 最大概率模型：根据 Trie 树构建有向无环图和进行动态规划算法
- 隐马尔科夫模型：根据基于某些语料库构建的 HMM 模型进行分词
- 混合模型：结合使用最大概率模型和隐马尔可夫模型进行分词

（1）词典

词典就是若干词语组成的库，在文本数据里面，最基本的单位就是词，所以我们对文本的分析都是基于词汇的。在文本分析里面，我们可能会用到的词典有停用

词词典、副词词典、否定词词典、负向情感词典以及正向情感词典等等。一般词典都可以从网络上面下载得来。

（2）停用词

在自然语言处理中，停用词是一些人类语言中包含的功能词，这些功能词极其普遍，与其他词语相比，功能词不包含什么信息，没什么意义，而且出现频率特别高，比如“什么”、“其他”、“其实”、“的”、“单单”、“同样”等。

（3）副词

副词是用以修饰动词的词语，是加权描绘词组或者整个句子的词。常见的副词有“最”、“太”、“完全”、“过于”、“十分”、“稍微”

（4）否定词

顾名思义，既是起否定意义的词汇。常见的否定词有：“不”、“没有”、“切莫”、“不要”

（5）自定义词典

由于文本中包含一部分比较特殊的词语，故而要自定义词典进行分词识别，提高分词的准确率。自定义词典是为了避免机器自动分词而造成的语义错误，使得原本重要的词语被分成毫无意义的词语，使得错判率升高。

4.1.2 文本特征提取

（1）词袋模型

词袋模型即为建立一个词典库，该词典库包含训练数据集语料库的所有词语，每个词语对应一个唯一识别的编号，利用 one-hot 文本表示。

文档的词向量维度与单词向量的维度相同，每个位置的值是对应位置词语在文档中出现的次数，即词袋模型（BOW)），使用词袋模型的时候，我们舍弃了输入文本中的大部分结构，比如章节、段落、句子和格式，只计算语料库中的每个文本的出现频次。

引入词袋模型，，我们可以使得数据保存为稀疏矩阵这种数据格式只保存非零元素。之所以是稀疏矩阵，是因为大多数文档中都只包含词表中的一小部分的单词，也就是特征数组中的大部分元素都为 0.

（2）TF-IDF 文本特征提取

利用 TF 和 IDF 两个参数来表示词语在文本中的重要程度。

TF 是词频，指的是一个词语在一个文档中出现的频率，一般情况下，每一个文档中出现的词语的次数越多词语的重要性更大，例如 BOW 模型一样用出现次数来表示特征值，即出现文档中的词语次数越多，其权重就越大，但是在长文档中的词语次数普遍比短文档中的次数多，导致特征值偏向差异情况。

TF 体现的是词语在文档内部的重要性；IDF 是体现词语在文档间的重要性。

即如果某个词语出现在极少数的文档中，说明该词语对于文档的区别性强，对应的特征值高，IDF 值高，$IDF_i = \log(\frac{|D|}{N_i})$，D 指的是文档总数，$N_i$ 指的是出现词语 i 的文档个数，很明显 N_i 越小，IDF 的值越大。

最终 TF-IDF 的特征值的表达式为：

$$TF - IDF(i, j) = TF_{ij} \times IDF_i$$

4.1.3 朴素贝叶斯分类器相关理论知识

（1）先验概率

在贝叶斯统计推断论中，一个未确定数目的先验概率分布（一般简称为先验）是一种表达了某人对于该数目的推断的一种概率分布，这种推断是没有考虑到一些（当前的）证据的。

（2）后验概率

在贝叶斯推断中，一个随机事件的后验概率是指：当与事件相关的一些证据或背景也被考虑进来时的条件概率。“后验”在这个语境下即指的是在考虑了与要被检验的特定事件相关的证据。

（3）贝叶斯定理

在概率论与统计学中，贝叶斯定理（或称贝叶斯法则、贝叶斯规则）描述了一个事件的可能性，这个可能性是基于了预先对于一些与该事件相关的情况的知识。举例来说，如果癌症和年龄有关，那么使用贝叶斯定理的话，相比根本不了解关于此人的任何其他信息，知道了它的年龄的话就可以用来更准确地帮助评估它得癌症与否的概率。

对于随机试验 E 有两个随机事件 A，B，A、B 两个事件独立且 $P(B) > 0$ 那么在 B 事件发生的条件下 A 发生的概率为:

$$P(A|B)=\frac{P(AB)}{P(B)}$$

其中 $P(AB)$ 为 A，B 两个事件的联合概率。对上式利用乘法公式可以变形为：

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

这样就得到了贝叶斯公式。贝叶斯文本分类就是基于这个公式，利用先验概率来得到文本的分类。

我们假设其中 $P(C_i)$ 为第 i 个文本类别出现的概率，$P(w_1,w_2,\cdots,w_n|C_i)$ 为文本类别为 C_i 时出现特征向量 $(w_1,w_2,\cdots,w_n)$ 的概率，$P(w_1,w_2,\cdots,w_n)$ 为特征向量出现的概率。一般的会假设特征——词，在文本中出现的概率是独立的，也就是说词和词之间是不相关的（虽然这个不一定成立，但是为了简化计算往往又不得不这么做），那么这时候的联合概率就可以表示为乘积的形式，如下：

$$P(C_i|w_1,w_2,\cdots,w_n)=\frac{P(w_1|C_i)P(w_2|C_i)\cdots P(w_n|C_i)}{P(w_1)P(w_2)\cdots P(w_n)}$$

对于特定的训练集合来说，上式中 $P(w_1)P(w_2)\cdots P(w_n)$ 是一个固定的常数，那么在进行分类计算的时候可以省略掉这个分母的计算，如是得到：

$$P(C_i|w_1,w_2,\cdots,w_n)=P(w_1|C_i)P(w_2|C_i)\cdots P(w_n|C_i)$$

（4）朴素贝叶斯

朴素贝叶斯（Naive Bayesian Classifier,NBC）是贝叶斯算法中最简单的算法，朴素贝叶斯之所以叫做“朴素”贝叶斯是因为它完全按照贝叶斯计算公式来分类，假设各个属性是条件独立的，互不影响、互不依赖。朴素贝叶斯虽然简单，但是其性能可以达到神经网络，决策树的水平。是文本分类中最常用的算法之一。

在使用朴素贝叶斯进行分类的时候，需要特别注意，它所选取的特征向量是相互独立的。因为，这里要满足划分定义的第一个条件。在实际应用中，往往达不到这个要求，这些变量之间往往有着连续，因此在进行数据挖掘之前可以进行特征向量的选取。这个条件好像是限制了朴素贝叶斯的使用范围，但是由于特性向量的压缩，减少了计算量，降低了构建贝叶斯网络的复杂性。使得贝叶斯分类器应用到更广的领域中。

朴素贝叶斯分类器是一种有监督学习，常见有两种模型。

- 多项式模型(multinomial model)即为词频型。
- 伯努利模型(Bernoulli model)即文档型。

二者的计算粒度不一样，多项式模型以单词为粒度，伯努利模型以文件为粒度，因此二者的先验概率和类条件概率的计算方法都不同。计算后验概率时，对于一个文档 d，多项式模型中，只有在 d 中出现过的单词，才会参与后验概率计算，伯努利模型中，没有在 d 中出现，但是在全局单词表中出现的单词，也会参与计算，不过是作为“反方”参与的。

多项式模型和伯努利模型的关键区别：

① 伯努利模型是以“文档”为统计单位，即统计某个特征词出现在多少个文档当中（最大似然方法估计条件概率 $P(x(i)|c)$ 的时候），当某个特征词在某一个文档中出现多次的时候，贝努利模型忽略了出现的次数，只算作一次。而多项式模型是以“词”为统计单位，当某个特征词在某个文档中多次出现的时候，与贝努利模型相反，它算作多次——这个更符合做 NLP 人的想法。

② 对特征向量中 0 值维度的处理。对于某个新的输入样本，当某个维度的取值是 0 的时候（也就是该特征词不出现在输入样本中），贝努利模型是要计算 $P(x(i,0)|c(0))$ 的值。而多项式模型直接忽略这样的特征，即只用 $P(x(i,1)|c(0))$ 的值来进行计算，从而区分各个类别。

（5）多项式模型

在多项式模型中，设某文档 $d=(t_1,t_2,\cdots,t_k)$，t_k 是该文档中出现过的单词，允许重复，则

$$\text{先验概率 } P(c)=\frac{\text{类}c\text{下单词总数}}{\text{整个训练样本的单词总数}}$$

$$\text{类条件概率 } P(t_k|c)=\frac{\text{类}c\text{下单词}t_k\text{在各文档中出现过的次数之和}+1}{\text{类}c\text{下单词总数}+|V|}$$

V 是训练样本的单词表（即抽取单词，单词出现多次，只算一个），$|V|$ 则表示训练样本包含多少种单词。

$P(t_k|c)$ 可以看作是单词 t_k 在证明 d 属于类 c 上提供了多大的证据，而 $P(c)$ 则

可以认为是类别 c 在整体上占多大比例(有多大可能性)。

（6）伯努利模型

以文件为粒度，或者说是采用二项分布模型，伯努利实验即 N 次独立重复随机实验，只考虑事件发生/不发生，所以每个单词的表示变量是布尔型的类条件概率。

$$P(c)=\frac{\text{类}c\text{下文件总数}}{\text{整个训练样本的文件总数}};$$

$$P(t_k\mid c)=\frac{\text{类}c\text{下包含单词}t_k\text{的文件数}+1}{\text{类}c\text{下文件总数}+2}$$

4.2 问题一挖掘实践

4.2.1 问题一挖掘目标

本次建模的目的是利用提供的群众留言表，采用自然语言处理（NLP）技术，建立一级分类指标和群众留言的相互关系，通过 python 软件进行分词，利用词袋模型使得群众留言转换成稀疏矩阵，计算一级分类与留言详情的关联度。从而去知道一个留言是与什么类别有关，正确把留言分类，提高相应的政府部门工作效率。

挖掘流程示意图如下所示：

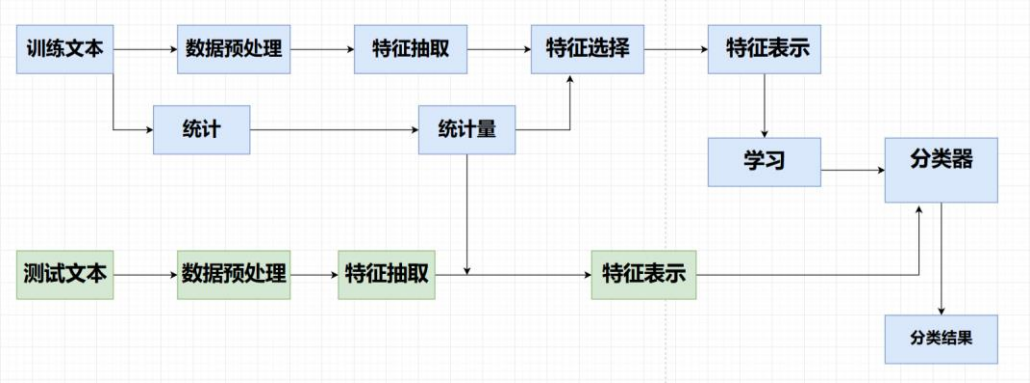

图 1 挖掘流程示意图

由上图我们可以看到，本次建模我们包含三个部分的内容，分别是群众留言数据预处理、模型建模、以及分类模型评估分析

4.2.2 群众留言数据预处理

（1）缺失值处理

在群众留言表中，有一些用户的留言是空值或者是缺失值，对于这种情况，选择了删除这一用户数据的处理方法。

（2）重复值处理

在原始数据中，同样的留言群众可能留言多次，我们这次建模挖掘是研究计算一级分类与留言详情的关联度，重复的留言对与代码的运行以及结果出现有影响，所以把重复出现的留言只保存其一。

（3）数据的清洗

利用 python 中的 re 库的函数，并采用正则表达式把群众留言数据中的英文字母、特殊符号、换行符、特殊符号、换页符号等等清除。用相应的函数使得数据录入进 python 的时候的数据类型始终为字符型，完成以上的内容后就可以开始我们的分词处理了。

（4）分词处理

利用 python 的 jieba 库，我们将的留言详情表文件形式更改为 csv 模式，然后进行数据的读入以及分词。停用词文档文件我们选择了在网上下载，利用 python 中 jieba 库的函数参数设置把留言详情中的停用词等等进行删除。

同时我们还自定义了一个词典，包含了特殊地名、三级标签、二级标签、一级标签、否定词以及特殊事件名字等等。用以方便机器正确地进行分词，提高分类的正确率。至于副词本次就暂时不做处理。

分词前后对比图如下所示：

Out[3]: \n\t\t\t\t\t\n\t\t\t\t\t　　干部职工的医疗保险费能否像住房公积金那样，让人民能在网上查询，知道自己的医保费到底有多少？ \n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t\t
2
\n\t\t\t\t\t\n\t\t\t\t\t　　八路军抗日写下的借条经E7县文物局（专家）和西地省文物鉴定中心人员认定提出解放前没用简体字，没有使用一九四五年公历年号，祖国统一，爱国人士等都要说成是疑点，全盘否认借条为假。持有人认为借条经历史验证为真，中央党史证明和43份历史证据证明，疑点都要是当时八路军部队正常使用的词句，并非什么疑点，疑点不成立借条为真。持有人认为文物人员违反背历史，认定不准，定性错误，特提出意议。为还原历史真实性，本人多次申请西地省文物管理局对这有争议的文物认定作出裁定，依予以纠正，但未果，有关人员确说没有这个权力，职责，义务来处理这方面的事。按照中华人共和国文件化部2009《46》号令（文物认定管理暂行办法）“第三条　认定文物，由县级以上地方文物行政部门负责。认定文物发生争议的，由省级文物行政部门作出裁定”。请依法作出客观公正尊重历史真假裁定，为了不让这份具有政治、历史、文物、教育意义珍贵史料不灭失，不让爱国人士误为诈骗犯，这份抗战七十年借条是中国人民抗日胜利七周年的见证证据，它的政治意义大于经济意义。可作为西地省人民献给中国人民抗日胜利周年的一份礼物，八路军借条的出现反映出多项抗日革命历史真相。为此敬请西地省文件化厅领导督促省文物局裁定认定（不是鉴定，双方对鉴定结果有争议，现按照历史证据予以裁定真假）习总书记说“让历史说话，用史实发言”。\n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t\t
1
\n\t\t\t\t\t\n\t\t\t\t\t戴局长：　　您好！现在在G市进行两工地设备检测的人员（向某，原为C市特检所辞退人员，仅有室内电梯检验员证）根本不具备塔吊及施工电梯的检验资格，却代表C市一家检测公司在G市检测了几年！其检验报告的合法性都有问题！ \n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t\t
1
\n\t\t\t\t\t\n\t\t\t\t\t　　感谢西地省交通运输厅下文明令西地省高速公路管理局清退借调人员！但是通过关系从西地省高速公路管理局K市管理处属收费站借调到路政队担任内勤人员的收费员依然没有被清退，这些借调人员在路政队呆得稳若金汤，目前为止完全没有要被清退的迹象，貌似西地省交通运输厅所下发文件完全没有被当成一回事，没有人去执行。难道上级的文件都可以不执行的吗？ \n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t\t

图 2 分词前

```
Out[4]: 0     [区，大道，西行，便道，未管，路口，加油站，路段，人行道，包括，路灯，...
        1     [位于，书院，路，主干道，在水一方，大厦，一楼，四楼，人为，拆除，水电,...
        2     [市政府，市，交警支队，市，安监局，市，环保局，区政府，市区，杜鹃，文苑...
        3     [胡书记，您好，感谢您，百忙之中，查看，这份，留言，父亲，区，金星，北路...
        4     [县，丁字街，商户，乱，摆摊，前段时间，丁字街，交通，好，几天，丁字街,...
        5     [南门，街，前段时间，整改，劝阻，摆摊，占道，情况，改善，很多，情况，...
        6     [现县，冷，江东，路，蓝波，旺，酒店，外墙，装修，搭，架子，无人，施...
        7     [九亿，广场，城区，人民，休闲，娱乐，场所，景观，点，很漂亮，每到，晚...
        8     [石期，市镇，农贸市场，旁边，公厕，旱厕，脏乱差，臭气熏天，老百姓，厕所,...
        9     [李，书记，您好，感谢您，阅读，十二五，期间，非，省会，地级市，轨道交通...
        10    [易，市长，您好，感谢您，阅读，十二五，期间，非，省会，地级市，轨道交通...
        11    [媒体报道，市，公交，地铁，爱心卡，一卡通，残疾人，爱心卡，乘坐，地铁，...
        12    [地铁，号线，施工，导致，万家，丽路，锦楚，国际，星城，小区，三期，一...
        13    [尊敬，领导，你好，区润，紫，郡，业主，今年年初，小区，周边，竖起，一...
        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，更加，恶劣，地处，...
        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，重大,...
        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，明显，最让人，愤怒，车人，无，监控，看似,...
        21    [你们好，上周，提交，请求，迎丰，公园，人性，关怀，角度，考虑，延后，...
        22    [市，城北，中坡山，国家森林公园，一座，自然，风光秀丽，市民，休闲，好去处...
        23    [尊敬，彭，书记，公园，里，门，球场，修好，修，二个，门，球场，停工...
        24    [雅礼洋湖，实验，中学，家长，代表，雅礼洋湖，实验，中学，西地，省内，办...
        25                [英祥，春天，后及，万科，城边，脏乱，不堪，环境，改变]
        26    [区，阳光，丽城，垃圾场，白天，偷偷，垃圾，运，晚上，夜色，进行，焚烧...
```

图 3 分词后

4.2.3 模型建模

（1）文本向量化表示

在使得机器可以正确识别我们的数据这一研究问题中，我们采用词袋模型的技术方法，我们首先把分词得到的文件进行创建词表，里面包含出现在留言详情中的所有词，并对他们进行编号。对于每个留言详情，计算词表中每个单词在该留言详情的出现频次。

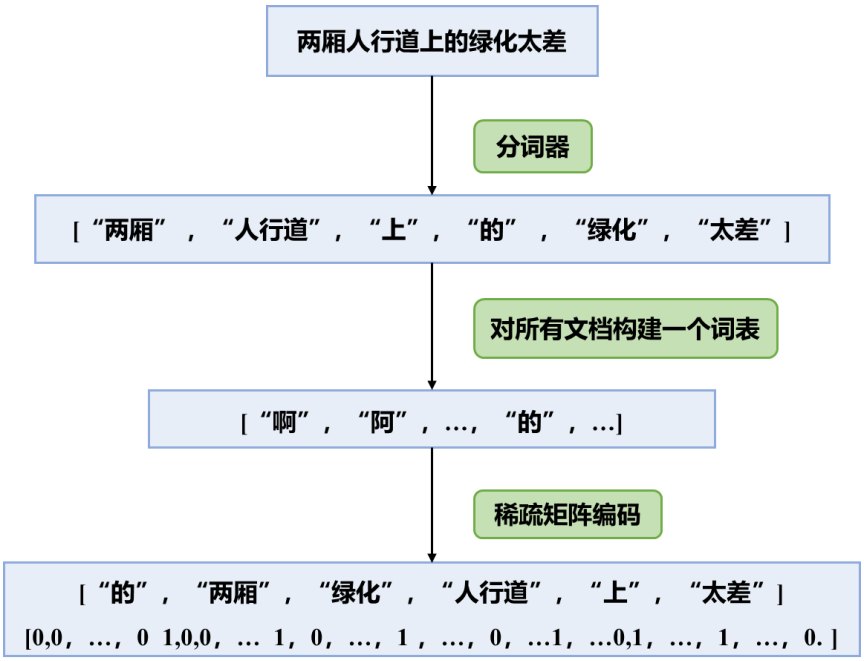

图 4 词袋模型

在这一个问题中我们经过了文本词袋处理后，得到了数据的稀疏矩阵，是一个9210×77720 的一个数据矩阵。

（2）数据分集

训练集（Training Set）：帮助我们训练模型，简单的说就是通过训练集的数据让我们确定拟合曲线的参数。

测试集（Test Set）： 为了测试已经训练好的模型的精确度。当然，test set 这并不能保证模型的正确性，他只是说相似的数据用此模型会得出相似的结果。因为我们在训练模型的时候，参数全是根据现有训练集里的数据进行修正、拟合，有可能会出现过拟合的情况，即这个参数仅对训练集里的数据拟合比较准确，这个时候再有一个数据需要利用模型预测结果，准确率可能就会很差。

由于本次建模的用户留言详情的数据观测值有 9210 个，相对来说比较大，所以在分层抽样还是随机抽样分训练集和测试集我们选择了随机抽样分组。所以本次建模利用 python 的 scikit-learn 库把分词后的数据进行随机抽样分组分为训练集和测试集，训练集和测试集的比例为 7：3，

（3）基于朴素贝叶斯的分类

P(“属于哪个一级标签”｜“具有某特征”)=在已知某样本“具有某特征”的条件下，该样本“属于哪个一级标签”的概率。

P(“具有某特征”｜“ 属于哪个一级标签”)=在已知某样本“属于哪个一级标签”的条件下，该样本“具有某特征”的概率。

P(“属于哪个一级标签”)=（在未知某样本具有该“具有某特征”的条件下，）该样本“属于哪个一级标签”的概率。所以叫做『先验概率』。

P(“具有某特征”)=(在未知某样本“属于哪个一级标签”的条件下，)该样本“具有某特征”的概率。

而我们这一个文本分类问题只要判断 P(“属于哪个一级标签”｜“具有某特征”)是否大于 1/6 就够了。贝叶斯方法把计算 “具有某特征的条件下属于哪个一级标签”的概率转换成需要计算“属于哪个一级标签的条件下具有某特征”的概率，而后者获取方法就简单多了，我们只需要找到一些包含已知特征标签的样本，即可进行训练。而样本的类别标签都是明确的，所以贝叶斯方法在机器学习里属于有监督学习方法。

我们在解决这个问题的时候利用了 python 中的 scikit-learn 库的 naive_bayes 函数进行朴素贝叶斯模型的构建与拟合。

4.2.4 分类模型评估

在分类模型一块中，我们依旧利用的是 python 中的 scikit-learn 库，使用的评估参数是 F-Score。

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

在我们最后的结果中，我们的贝叶斯分类器 F-Score 为 0.8581。

（1）TP、FP、FN 和 TN

以文本分类的一个常见例子—垃圾短信识别来说，毫无营养的短信便诊断为有垃圾短信，无骚扰并且含有重要信息的短信就会被诊断为不是垃圾短信。二元分类模型的个案预测有四种结局：

1、真阳性（true positive， TP）：诊断为是， 实际上是垃圾短信；

2、伪阳性（false positive，FP）：诊断为是，实际却不是垃圾短信；

3、真阴性（true negative，TN）：诊断为不是，实际上也不是垃圾短信；

4、伪阴性（false negative，FN）：诊断为不是，实际上却是垃圾短信。

这四种结局可以画成 2 × 2 的混淆矩阵：

表 1 混淆矩阵

		真实值		总数
		p	n	
预测输出	P1	真阳性	伪阳性	P
	N1	伪阴性	真阴性	N
总数		P	N	

准确率（accuracy），所有的预测正确（正类负类）的占总的比重

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

精确率（也叫查准率，precision)，即正确预测为正的占全部预测为正的比例，（真正正确的占所有预测为正的比例）

$$Precision = \frac{TP}{TP+FP}$$

召回率（recall），即正确预测为正的占全部实际为正的比例（真正正确的占所有实际为正的比例）

$$Recall = \frac{TP}{TP + FN}$$

F-score 值，F1 值为算数平均数除以几何平均数，且越大越好，将 Precision 和 Recall 的上述公式带入会发现，当 F1 值小时，True Positive 相对增加，而 false 相对减少，即 Precision 和 Precision 都相对增加，即 F1 对 Precision 和 Recall 都进行了加权。

$$F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}$$

在本题中的结果为：F-Score 为 0.8581，各个类别的 Precision、Recall、F-score 如下图和表格所示

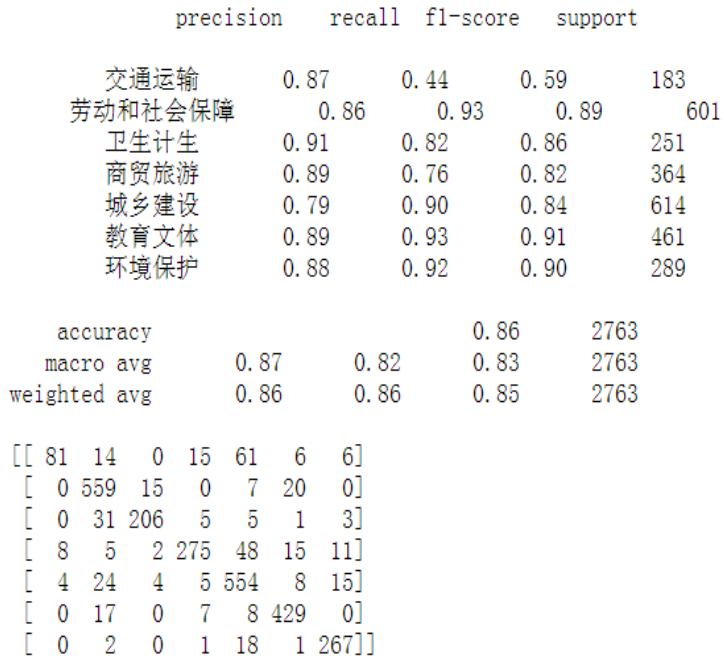

图 5 各个类别的 Precision、Recall、F-score

根据下表我们可以知道 7 个类别的 precision 和 recall 的值近乎相等，交通运输、商贸旅游、城市建设这三个的两个值相差比较大。汇总加权后的 precision 和 recall 相等，所以我们可以认为模型拟合得不错。

表 2 七个类别的结果

标签	Precision	Recall	F-score	Support
交通运输	0.87	0.44	0.59	183
劳动和社会保障	0.86	0.93	0.89	601

卫生计生	0.91	0.82	0.86	251
商贸旅游	0.89	0.76	0.82	364
城乡建设	0.79	0.90	0.84	614
教育文体	0.89	0.93	0.91	461
环境保护	0.88	0.92	0.90	289

五、问题二的分析与求解

5.1 问题二分析

问题二要求我们根据附件 3 将某一时间段内反映特定地点或特点人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。则是要求我们用机器学习的分类方法对留言进行分类。

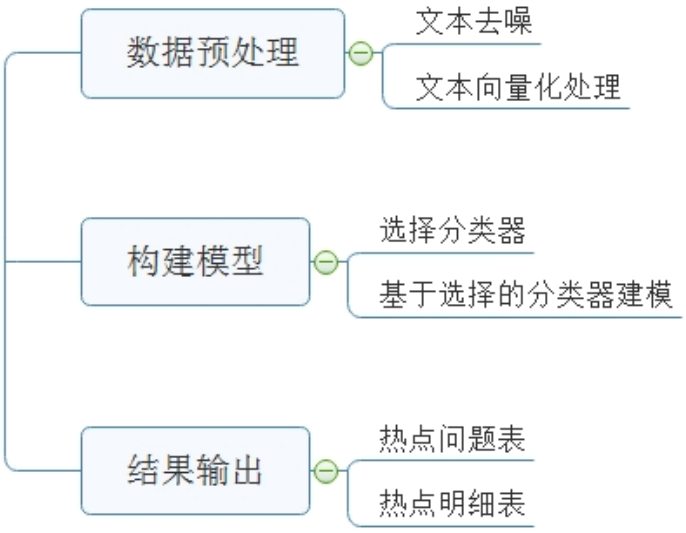

图 6 第二题思路框架

5.2 问题二挖掘实践

5.2.1 数据预处理

（1）文本去噪

• 因为“留言主题”已经包含有效信息，因此选取“留言主题”作为文本分析对象

• 建立语料库，加载分词文本。将有意义又可能被 python 的 jieba.lcut()函数分

割的词语留起来

- 移除无意义的停用词，包括特殊符号、阿拉伯数字、中文感叹词等。
- 移除低频词，避免文本特征过大
- 输出分词结果

（2）文本向量化处理

- 将词语转化为词频矩阵
- 计算 TF-IDF。计算出来的文本特征向量相当于把文本映射到向量空间，可以用来做文本相似度的计算。也就是在没有标签的情况下，可以做文本聚类。

5.2.2 选择分类方法

（1）常见文本分类器

表 3 常见文本分类器

算法分类	算法说明	算法举例
划分法	给定一个有 N 个元组或者纪录的数据集，分裂法将构造 K 个分组，每一个分组就代表一个聚类，K<N。而且这 K 个分组满足下列条件： （1）每一个分组至少包含一个数据纪录； （2）每一个数据纪录属于且仅属于一个分组；对于给定的 K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都较前一次好，而所谓好的标准就是：同一分组中的记录越近越好，而不同分组中的记录越远越好。	K-Means 算法、K-MEDOIDS 算法、CLARANS 算法
层次法	这种方法对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案，即合并聚类（由下而上）和分裂聚类（由上而下）。合并层次聚类是将语料库中的任一数据都当作一个新的簇，计算所有簇相互之间的相似度，然后将相似度最大的两个簇进行合并，重复这个步骤直到达到某个终止条件，因此合并聚类方法也被称为由下而上的方法。分裂聚类恰好与合并聚类进行相反的操作，	BIRCH 算法、CURE 算法、CHAMELEON 算法等

	它是一种由上而下的方法，该方法先将数据集中所有的对象都归为同一簇，并将不断地对原来的簇进行划分从而得到更小的簇，直到满足最初设定的某个终止条件。	
密度法	基于密度的方法与其他方法的一个根本区别是：它不是基于各种各样的距离的，而是基于密度的。这样就能克服基于距离的算法智能发现“类圆形”的缺点。这个方法的指导思想就是，只要一个区域中的点的密度大过某个阈值，就把它加到与之相近的聚类中去。	DBSCAN 算法、OPTICS 算法、DENCLUE 算法等
网格法	这种方法首先将数据空间划分成为有限个单元（cell）的网格结构，所有的处理都是以单个的单元为对象的。这么处理的一个突出的优点就是处理速度很快，通常这是与目标数据库中记录的个数无关的，它只与数据空间分为多少个单元有关。	STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法

（2）K-means 方法说明

本文拟用 K-means 方法进行文本分类。

K-means 算法是一种典型的基于划分的聚类算法，该聚类算法的基本思想是在聚类开始时根据用户预设的类簇数目 k 随机地在所有文本集当中选择 k 个对象，将这些对象作为 k 个初始类簇的平均值或者中心，对于文本集中剩余的每个对象，

根据对象到每一个类簇中心的欧几里得距离，划分到最近的类簇中；全部分配完之后，重新计算每个类簇的平均值或者中心，再计算每篇文本距离这些新的类簇平均值或中心的距离，将文本重新归入目前最近的类簇中；不断重复这个过程，直到所有的样本都不能再重新分配为止。

K-means 算法优点：

（1）对待处理文本的输入顺序不太敏感

（2）对凸型聚类有较好结果

（3）可在任意范围内进行聚类

K-means 建模步骤：

（1）从 N 个文档随机选取 K 个文档作为质心

（2）对剩余的每个文档测量其到每个质心的距离，并把它归到最近的质心的类。

用欧几里得距离衡量距离：

$$O=\sqrt{(x_1+y_1)^2+(x_2+y_2)^2}$$

（3）重新计算已经得到的各个类的质心

对于 n 个点的数据集，迭代计算 k from 1 to n，每次聚类完成后计算每个点到其所属的簇中心的距离的平方和，可以想象到这个平方和是会逐渐变小的，直到 k==n 时平方和为 0，因为每个点都是它所在的簇中心本身。但是在这个平方和变化过程中，会出现一个拐点也即“肘”点，下图可以看到下降率突然变缓时即认为是最佳的 k 值。

如下图所示：

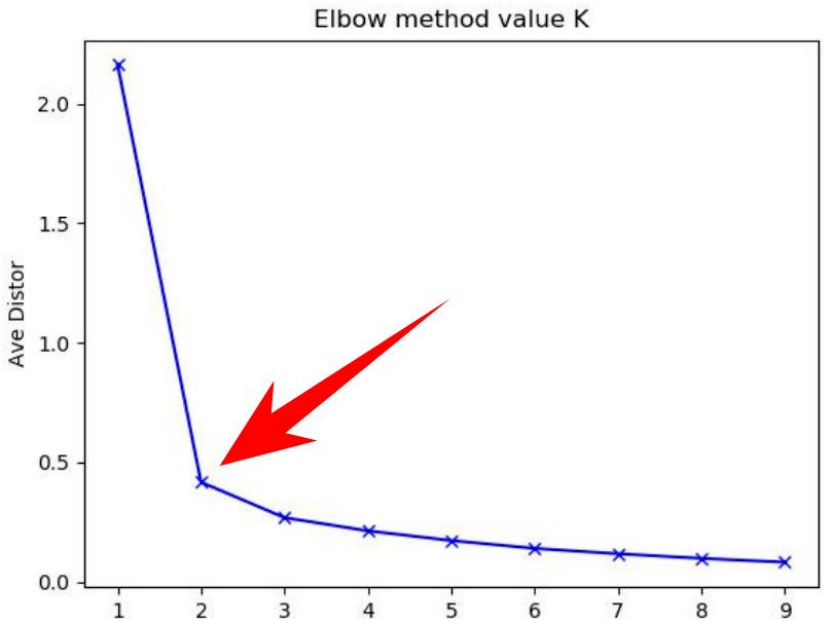

图 7 迭代计算 K 值

（4）迭代 2～3 步直至新的质心与原质心相等或小于指定阀值，算法结束

5.2.3 基于 K-mean 分类建模

（1）建立模型

- 将文本向量化处理后的数据储存给一个对象
- 定义测试集，根据测试结果寻找拐点，选择 K 值。结果如下图所示：

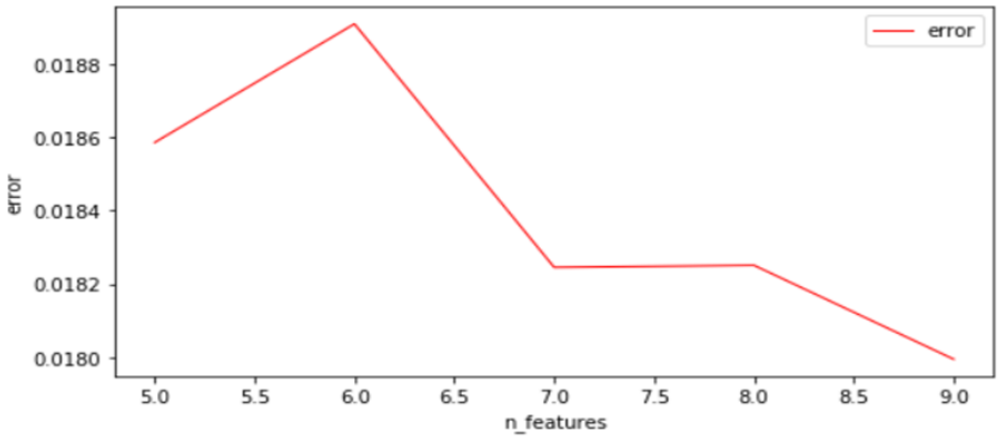

图 8 选择 K 值

可以看到在 7 那里有一个明显的拐点，因此选择 7 为最佳的 K 值

- 输出聚类结果，前五个聚类结果截图如下：

表 4 部分分类结果

簇数	分类结果
Cluster 1	大厦 公开 市经 扰民 严重 规划
Cluster 2	建议 对市 优化 地铁 增设 公交
Cluster 3	问题 咨询 解决 小曲 质量 房屋
Cluster 4	小区 扰民 投诉 街道 严重 西地省
Cluster 5	购房 补贴 人才 资格 咨询 问题

5.2.4 建立评价指标

（1）根据前五个分类结果分别命名为 id1-id5

（2）分别对 id1-id5 包含的留言个数进行求和，并由高到低地进行权重赋值

（3）分别统计 id1-id5 的点赞数和反对数，并从高到低进行排序，进行赋值，点赞数与反对数相互抵消

（4）输出最终分数，并起名为热度

5.3 结果输出

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
188059	A0002857	A市A3区中	2019/11/	A市A3区中	0	0
188073	A909164	A3区麓泉	2019/3/1	作为麓泉	0	0
188119	A0003502	对A市地铁	2019/5/2	我是一名	0	0
188249	A0008408	A3区保利	2019/9/1	保利麓谷	0	0
188260	A0005348	A3区青青	2019/5/3	还我宁静	0	0
188396	A0004758	关于拆除	2019/4/1	桐梓坡58	2	1
188399	A0009793	A市利保壹	2019/7/3	您好，我	0	0
188414	A0009684	A4区北辰	2019/8/1	您好！我	0	0
188467	A0005018	投诉A市温	2019/3/2	退费之日	0	1
188475	A0005581	A6区乾源	2019/12/	A市A6区月	0	0
188546	A0006817	A2区佳兆	2019/1/2	敬爱的领	0	0
188592	A0003945	A市长房云	2019/6/1	长房云时	0	0
188774	A0004879	A2区政府	2019/6/1	多年来A2	0	0
188799	A0001073	A7县橄榄	2019/5/2	我是橄榄	0	1

图 9 结果截屏

统计其条数为 1912，可以看到聚类结果大大地减少了留言的个数，从而更好地从众多无规律的留言中更好地聚焦到热点问题，优先解决呼声高、时间急的问题，从而更好地提高政务效率。

六、问题三分析与求解

6.1 问题三分析

问题三要求对答复意见的质量给出一套评价方案。根据前面的分词模型，可以把此问看作是问题一的延伸，问题一解决了文本分词，将模型代入附件 4 中的数据进行分词。为解决分词文本中的相关性、完整性和可解释性，要把分词文本进行数字化处理，处理后再对文本数字进行答复意见和留言详情的比对，再进行文本相关性等分析。

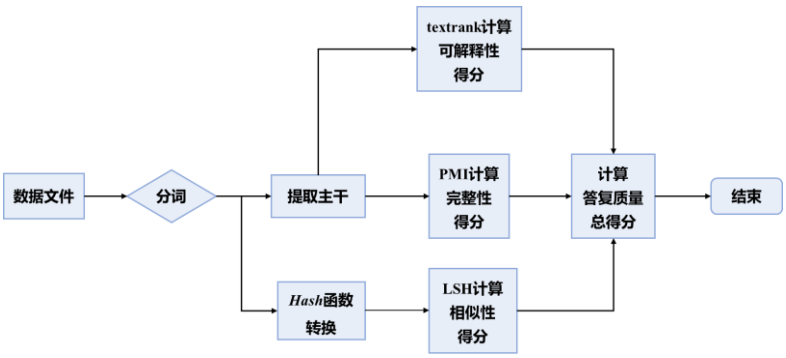

图 10 问题三思路框架

6.2 模型求解

6.2.1 *LSH* 解决答复相关性的问题

根据问题的分析，需要将附件 4 中答复意见和留言详情代入问题一中的模型进行分词，如果两个文本在原有的数据空间是相似的，那么分别经过 hash 函数转换以后的它们也具有很高的相似度；相反，如果它们本身是不相似的，那么经过转换后它们应仍不具有相似性。hash 是相对的，数据的相似度是保持最大可能的相似度，其中的误差忽略不计。

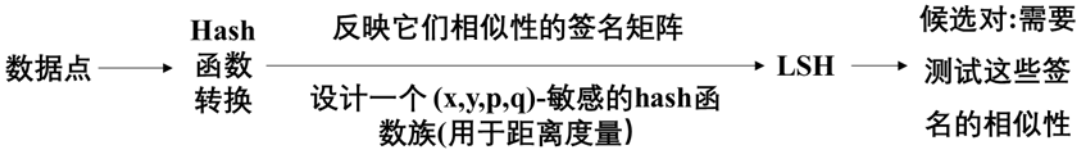

图 11 LSH 计算过程

利用问题一的模型对文本数据进行分词，把分词后的数据文件组成矩阵，组成矩阵之后再对数据进行 *hash* 函数的转换。

hash 成降维后的签名矩阵，用 *minhashs* 函数来做，*minhashs* 是一个比较有效的降维功效，同时也不会损失太多信息量，原来相似的文本表现的还是相似。

在衡量文本的相似度中，我们可以利用欧式距离、余弦距离、*Jaccard* 距离等来进行相似度的度量。在本题中我们运用 *Jaccard* 相似度，即用 *minhash* 计算距离。接下来我们就要去找一种 *hash* 函数，使得在 *hash* 后尽量还能保持这些文档之间的 *Jaccard* 相似度，即：

If sim(C1,C2)is high, then with prob.h(C1)=h(C2)

If sim(C1,C2)is low, then with prob.h(C1)≠h(C2)

需要找到这样一种 *hash* 函数，如果原来文档的 *Jaccard* 相似度高，那么它们的 *hash* 值相同的概率高，如果原来文档的 *Jaccard* 相似度低，那么它们的 *hash* 值不相同的概率高。两列的最小 *hash* 值就是这两列的 *Jaccard* 相似度的一个估计。

先要对数据进行常规化的 *hash* 处理，处理之后进行两两比较，然后再进行降维和局部寻找匹配对的 *LSH*，可以在这两者基础上更快的找到相似、可匹配的对象，而且继承了 *minhash* 的优点，相似文档 *LSH* 计算之后还是保持相似的。

Jaccard 系数用来度量两个集合的相似度，设有两个文本即答复意见和留言详情为 S_1 和 S_2，它们之间的 *Jaccard* 系数定义为：

$$S=\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Jaccard 系数值越大越相似，计算出来的 *Jaccard* 系数范围在 $0 \le S \le 1$，因此把文本数据代入进而可以得到两个文本的相关性。

6.2.2 *PMI* 解决答复完整性的问题

要讨论文本的完整性，即答复的内容是否完整回答留言详情所提出的问题，即可以简化为将留言详情分词中的内容精简，提取出每个问题的主干跟答复的内容进行词语之间的相似性比较。

点间互信息（*PMI*）主要用于计算词语间的语义相似度，基本思想是统计两个词语在文本中同时出现的概率，如果概率越大，其相关性就越紧密，关联度越高。两个词语 S_1 与 S_2 的 PMI 值计算公式如下式所示为：

$$PMI(S_1,S_2)=\log_2(\frac{P(S_1 \& S_2)}{P(S_1)P(S_2)})$$

$$pmi(x;y) \equiv \log\frac{p(x,y)}{p(x)p(y)}=\log\frac{p(x|y)}{p(x)}=\log\frac{p(y|x)}{p(y)}$$

$$MI(X,Y)=\log\frac{2p(x,y)}{p(x)p(y)}$$

$P(S_1 \& S_2)$ 表示两个词语 S_1 与 S_2 共同出现的概率，即 S_1 与 S_2 共同出现的文档数，$P(S_1)$ 与 $P(S_2)$ 分别表示两个词语单独出现的概率，即 S 出现的文档数。若两个词语在数据集的某个小范围内共现概率越大，表明其关联度越大；反之，关联度越小。

$P(S_1 \& S_2)$ 与 $P(S_1)P(S_2)$ 的比值是 S_1 与 S_2 两个词语的统计独立性度量。其值可以转化为 3 种状态：

$P(S_1 \& S_2)>0$：两个词语是相关的，值越大，相关性越强；

$P(S_1 \& S_2)=0$：两个词语是统计独立的，不相关也不互斥；

$P(S_1 \& S_2)<0$：两个词语是不相关的，互斥的。

在信息论中常用互信息 *MI* 来衡量两个词的相关度：

$$MI(X,Y)=\log\frac{2p(x,y)}{p(x)p(y)}$$

当 X，Y 关联大时，$MI(X,Y)>0$；当 X 与 Y 关系弱时，$MI(X,Y)=0$；当 $MI(X,Y)<0$时，X 与 Y 称为“互补关系”。MI 越大，表示两个词之间的结合越紧密。这个算式看起来很直观，但计算还是有些麻烦，因为计算概率值 $p(x)$，$p(y)$ 都需要在语料中进行分词。

因此采用更加简便直观的算法进行计算，假设文本集合为$\{C\}$，总行数为 N，其中含有单词 X 的文章总数为N_x，含有单词 Y 的文章总数是N_y，含有$\{x+y\}$的文章总数是N_{xy}，那么完整性可以用以下公式计算：

$$Corr(x,y)=Math.\log_{10}(\frac{N}{N_x})\times Math.\log_{10}(\frac{N}{N_y})\times\frac{N_{xy}}{N_x+N_y-N_{xy}}$$

$Corr(x,y)$值越大越相似，计算出来的$Corr(x,y)$范围在$0\le Corr\le 1$，因此把文本数据代入进而可以得到两个文本的完整性。

6.2.3 *TextRank* 解决答复可解释性的问题

文本的可解释性即字词、语句之间是可以说通的，整个句子的主谓宾都是完整的，因此可以简化为利用已经分好词的答复意见，再精简提取出每个句子的主干，迭代计算每个文本的 *TextRank* 值，最后把排名高的文本抽取出来，作为这段文本的关键词或者文本摘要，如若说得通即表明该句子的可解释性高。

附件 4 的分词文件是以词作为切分的时候，构成词与词之间是否连接的，是词之间是否相邻。相邻关系可以分为 *n* 元，不过在中文中，我们认为 2 元关系已经非常足够了。如果是以句子切分的，那么得到的称之为文本摘要。

文本摘要其实就是从文档中提出我们认为最关键的句子。我们用 *textrank* 包的 *textrank_sentences* 函数，这要求我们有一个分句的数据框，还有一个分词的数据框。这次分词必须以句子为单位进行划分。对任意的一个长字符串，我们要能够切分成多个句子，然后按照句子分组，对其进行分词。然后我们会得到一个句子表格和单词表格。 其中，我们切分句子的标准是，切开任意长度的空格，这在正则表达式中表示为“$[:space:]+$”。

如果要得到文本的关键句子，还是要对每句话进行分词，得到每句话的基本词要素。根据句子之间是否包含相同的词语，我们可以得到句子的相似度矩阵，然后再根据相似度矩阵来得到最关键的句子。

最后将提取到答复意见最关键的句子与留言详情最关键的句子进行相关性的比较可以得到可解释性的系数，我们利用关键句子的文本机械相似性来得出结果，在这题中我们选用余弦相似度来计算文本之间的可解释性。

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。假定留言详情和答复意见是两个分别是 n 维和 m 维向量， a 与 b 的夹角的余弦等于：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i\times y_j)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{j=1}^{m}(y_i)^2}}$$

6.2.4 层次分析求权重

（1）建立递阶层次结构模型

根据问题所选取的指标，构造出一个有层次的结构模型。在这个模型下，复杂问题被分解为元素的组成部分。这些元素又按其属性及关系形成若干层次。上一层次的元素作为准则对下一层次有关元素起支配作用。

（2）构造出各层次中的所有判断矩阵

设答复意见的质量为 Z，答复意见与留言详情的相关性、完整性和可解释性这三个因子为 $X=\{x_1,x_2,x_3\}$，采取对因子进行两两比较建立成对比较矩阵的办法。即每次取两个因子 x_i 和 x_j 对 Z 的影响大小之比，全部比较结果用矩阵 $A=(a_{ij})_{n\times n}a_{ij}a_{jk}=a_{ik}CR=0.062n$ 表示，称 A 为 Z-x 之间的成对比较判断矩阵。

容易看出，若 x_i 与 x_j 对 z 的影响之比为 a_{ij}，则 x_j 与 x_i 对 z 的影响的之比应为 $a_{ji}=\frac{1}{a_{ij}}$。

根据层次分析判断矩阵的评判标度，确定 a_{ij} 的值为

$$A=\begin{pmatrix} 1 & \frac{1}{3} & 2 \\ 3 & 1 & 3 \\ \frac{1}{2} & \frac{1}{3} & 1 \end{pmatrix}$$

判断矩阵 A 对应于对大特征值 $\lambda_{\max}$ 的特征向量 W，经归一化处理后即为同一层次相应因素对于上一层次某因素相对重要性的排序权值。

上述构造成对比较判断矩阵的办法虽能减少其他因素的干扰，较客观地反映出一对因子影响力的差别。但综合全部比较结果时，其中难免包含一定程度的非一致性。如果比较结果是前后完全一致的，则矩阵 A 的元素还应当满足：

$$a_{ij}a_{jk}=a_{ik}，\ \forall i,j,k=1,2,3$$

因此需要检验构造出来的判断矩阵 A 是否严重地非一致，以便确定是否接受 A。因此需要计算一致性指标 CI

$$CI=\frac{\lambda_{\max}-n}{n-1}$$

查找相应的平均随机一致性指标 RI，对 $n=1,\mathrm{L}\ ,9$，给出 RI 的值，如表 1 所示。

$$RI=\frac{\lambda_{\max}^{'}-n}{n-1}$$

表 5 一致性指标 RI 取值

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

计算一致性比例 CR

$$CR=\frac{CI}{RI}$$

经过 R 语言计算得 $CR<0.10$ 即我们所选取的判断矩阵 A 通过一致性检验，此判断矩阵可用，继续代入数据得到层次总排序的权值如表 2 所示，,

表 6 各个因子的权值

准则	相关性	完整性	可解释性
权值	0.249	0.594	0.157

即有得到的评价答复意见的质量的关系式为：

$$Z = 0.249 \times 相关性 + 0.594 \times 完整性 + 0.157 \times 可解释性$$

6.3 结果分析

用以上三种方法分别代入文本数据可以计算出答复意见的指标相关性、完整性和可解释性系数，然后用层次分析法计算得出的权重和评价质量的关系式代入每个各个系数得到最终的结果。最后得到 Z 的范围在 0~1 之间，Z 越大则表示答复意见的质量越高，Z 越小表示答复意见的质量越低，得到的结果选取前十行如下所示：

表 7 答复质量结果

编号	答复意见	相似性	完整性	可解释性	答复质量
2549	《问政西地省》栏目向胡华衡…	0.5860	0.4887	0.5174	0.4986
2554	针对您反映 A3 区潇楚南路洋…	0.2689	0.3884	0.3531	0.3318
2555	“请加快提高民营幼儿园教师…	0.4655	0.3535	0.3866	0.3743
2557	《问政西地省》上的留言已…	0.4156	0.2652	0.3096	0.3025
2574	来信人建议“白竹坡路口”…	0.6658	0.0786	0.2521	0.2674
2759	针对您反映 A3 区含浦镇马路…	0.2353	0.4872	0.4128	0.3828
2849	针对您反映 A3 区教师村小区…	0.2837	0.5460	0.4685	0.4355
3681	关于小区附近幼儿园的问题…	0.2558	0.4167	0.3692	0.3453
3683	美麓阳光项目位于 A3 区枫林…	0.4317	0.2333	0.2919	0.2878
3684	您所反映的地点为洋湖新城…	0.4275	0.3325	0.3604	0.3485

七、参考文献

[1]夏小娜,邹麒.基于兴趣相似度传递的增强 LSH 统计预测算法[J].计算机应用与软件,2020,37(03):286-291.

[2]赵春红,刘国华,王柠,何玲玲.外包数据库模型中文本数据的完整性检测方案[J].小型微型计算机系统,2010,31(09):1790-1796.

[3]杨俊闯,赵超.K-Means 聚类算法研究综述[J].计算机工程与应用,2019,55(23):7-14+63.

[4]邓林培.经典聚类算法研究综述[J].科技传播,2019,11(05):108-110.

[5]赵文. 基于朴素贝叶斯算法的不良文本过滤技术研究及应用[D].长安大学,2018.

[6]邓丁朋,周亚建,池俊辉,李佳乐.短文本分类技术研究综述[J].软件,2020,41(02):141-144.

[7]曹鲁慧,邓玉香,陈通,李钊.一种基于深度学习的中文文本特征提取与分类方法[J].山东科学,2019,32(06):106-111.

[8]方秋莲,王培锦,隋阳,郑涵颖,吕春玥,王艳彤.朴素 Bayes 分类器文本特征向量的参数优化[J].吉林大学学报(理学版),2019,57(06):1479-1484.