“智慧政务”中的文本挖掘应用

摘 要

近年来，网络问政平台已成为政府了解社情民意的重要渠道之一，各类反映社情民意的文本数据量也在不断攀升. 因此，针对问政平台的文本数据，对问题的分类、热点问题的筛选以及答复质量的评价，利用自然语言处理技术和文本挖掘技术，为智慧政务系统的建立和完善提供重要的数据支持.

针对问题一，首先对问政平台的留言数据，进行数据清洗、结巴中文分词、停用词过滤等预处理工作，以减少特征提取中的错误；然后利用 TF-IDF (Term Frequency–Inverse Document Frequency) 算法对文本进行特征抽取，获得每个一级标签的权重向量；最后，建立基于朴素贝叶斯算法的留言内容的一级标签分类模型，并进行测试和优化模型，实现留言的分类.

针对问题二，首先分析热点问题数据特征，对其进行数据清洗、中文分词等预处理；其次建立基于 TF-IDF 的 LSI (Latent Semantic Indexing) 模型，实现留言信息的相似度计算，进而实现对不同留言的分类；再使用舆情热度计算方法以及结巴中文分词获得每一类留言的热度指标和主题词；最后，将排名前 5 的热点问题以及概括性信息对应地保存至题目要求的两个文件中.

针对问题三，首先分析问题的要求，通过查阅相关资料，提取出影响答复意见质量的四个主要指标；其次，构建词袋模型，利用余弦相似度计算实现评价指标相关性的量化；再通过自定义模板以及可读性指数 ARI (Automated Readability Index)，实现对指标完整性、可解释性和及时性的量化；最后，利用独立性权系数法确定各指标的权重，给出关于评价答复意见质量的方案.

本文建立了基于朴素贝叶斯算法的分类模型，并通过多次将留言主题关键词映射到向量空间以及用欠抽样的抽样方法平衡样本数据，提高了模型训练的准确率；利用基于 TF-IDF 的 LSI 模型，不同于传统 LSI 模型，其可实现对不同留言更加精确地相似度计算，准确提取出热点问题；采用独立性权系数法，降低量化后评价答复意见质量的四个指标之间的相关性. 因此，本文利用 Python 中的自然语言处理技术和工具，高效、精准地完成了问政平台中文本数据的挖掘和处理，为有关部门对群众留言信息的及时和准确处理提供了重要的数据参考.

关键词：文本数据处理，评价指标，数据挖掘，TF-IDF 算法， LSI 模型

目录

1．挖掘目标

本次建模基于网络问政平台上的群众留言相关数据，针对留言分类、热点问题挖掘、答复意见的评价等问题，利用 Python 中的自然语言处理技术，研究了有关数据预处理、TF-IDF 算法、标签分类模型、LSI 模型、独立性权系数法、余弦相似度计算等相关算法和模型，以达到以下三个目标：

（1）利用中文分词文本特征向量化方法对非结构化的数据进行数据处理，构建关于留言内容的一级标签分类模型.

（2）利用基于 TF-IDF 的 LSI 模型对不同留言的自然文本信息进行相似度的量化处理，实现根据留言内容对留言信息的分类，进而量化得到不同类信息的热度指标，筛选出当下急需政府有关部门处理的热点问题.

（3）利用余弦相似度计算和独立性权系数法提取评价答复意见的相关指标并确定各指标权重，给出评价答复意见质量的方案.

2．分析方法与过程

2.1 问题一分析方法与过程

围绕着留言分类问题，本节首先对问政平台的留言数据进行预处理，提高数据质量，减少特征提取中的错误；然后进行文本特征的提取，将特征进行量化；再建立特征分类模型、算法实现和程序设计，对留言内容的一级标签进行分类；最后对分类模型和算法进行测试，调整和优化模型设计的有关参数.

2.1.1 流程图

本题采用的分析方法与过程的流程图如图 1 所示.

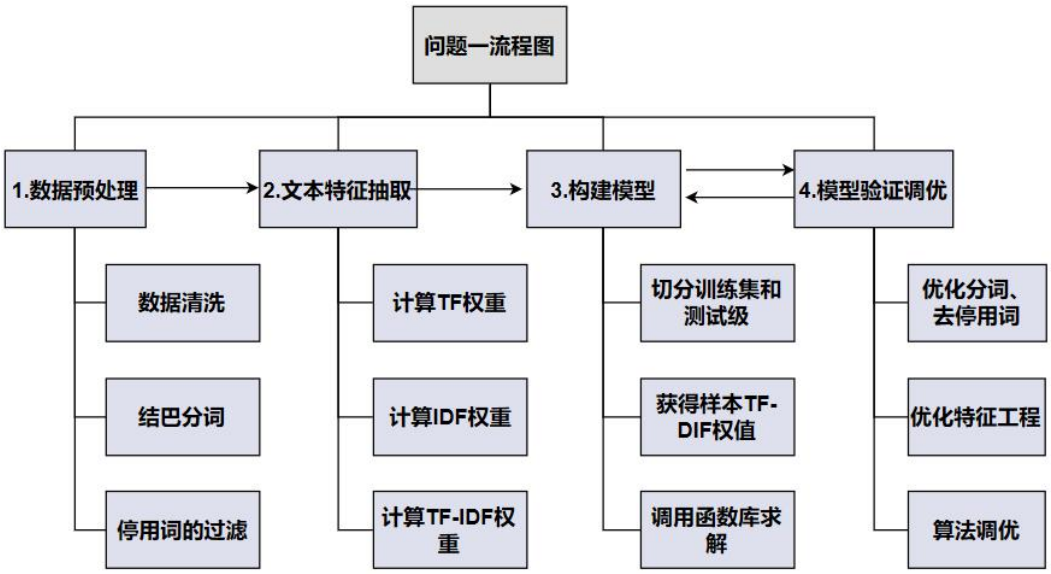


图 1. 问题一分析过程的流程图

2.1.2 数据预处理

为了让群众留言数据更适合做挖掘和分析，我们需要对附件 2 中的文本数据进行预处理，包括以下 3 个步骤：

（1） 数据清洗

观察可以发现数据中包含大量与问题研究无关的空格、X 序列（主要是数字）等，并且每条数据中的留言详情都存在大量噪声，影响对后续数据分析、分词、词频统计、模型建立的效率和质量，因此将“留言主题”单独生成一个数据帧，利用 Python 中的 Numpy 库对数据进行去重，删除无关字符等处理从而完成数据的清洗工作.

（2）留言主题的中文分词

为了后续对留言数据的信息挖掘，我们需要把非结构化的文本信息转换成计算机能够识别的结构化信息，即对清洗后的“留言主题”数据帧进行分词.本文采用了 Python 中的中文分词模块—Jiebe 分词，其运用了动态规划来找出词频最大切分组合，可以实现基于前缀词典的高效词图扫描，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)；对于未登录词，采用了基于汉字成能力的 HMM 模型和 Viterbi 算法，因为 Jieba 分词系统提供分词、词性标注、未登录词识别，支持用户自定义词典，从而可以获得更高的分词准确率，为特征提取和分类模型的构建提供了可靠的、准确的信息.其中部分分词结果如表 1 所示.

从表 1 可以看出分词后仍有大量标点以及无实质性无意义的字词，为了避免对后续数据分析的影响，还需要对停用词等进行过滤，以获得更纯净的数据.

表 1. 中文分词结果

386	[退休政策，调整，的，问题，有话，向，市长，说]
349	[A,市，交警总队，逼迫，孕妇，辞职]
379	[L,县，关庄镇，内退，员工，合法权益，被，侵占]
359	[西地省，采丰商贸有限公司，好润佳，I，市电，违反，劳动法，非法]
355	[A，市，和顺洋湖壹号壹号小区，建，千伏，……，不顾，老百姓]
	……
115	[M,区，环保局，为，蛇形山镇，破坏，生态环境，重叠，保护伞]
105	[关于，L，县，被，砖厂，拆除，相关，政策，的，咨询]
109	[L，县，沙溪，炼油厂，环评，合格，吗，？]
131	[泸阳镇，下坪区，与，壮稻村，采石场，严重破坏，生态环境]
117	[A，县，格力空调，生产产家，不许，员工，辞职，请假]

（3）停用词的过滤

针对停用词有两个特征：一是包含信息量低，对文本标识无意义；二是很普遍、出现频率很高.收集相关资料，用中文分词常见过滤词（具体见文本 Stopword 和文本 Stopword1），并分析了留言标题文本的特点，添加了一部分自定义停用词，提高了过滤效率和数据质量.

表 2. 停用词过滤结果

386	[退休政策，调整，有话，市长，说]
349	[交警总队，逼迫，孕妇，辞职]
379	[关庄镇，内退，员工，合法权益，侵占]
359	[西地省，采丰商贸有限公司，好润佳，违反，劳动法，非法]
355	[和顺洋湖壹号壹号小区，建，千伏，……，不顾，老百姓]
	……
115	[环保局，为，蛇形山镇，破坏，生态环境，重叠，保护伞]
105	[关于，砖厂，拆除，相关，政策，咨询]
109	[沙溪，炼油厂，环评，合格]
131	[泸阳镇，下坪区，壮稻村，采石场，严重破坏，生态环境]
117	[格力空调，生产产家，员工，辞职，请假]

2.1.3 基于 TF-IDF 算法的文本特征抽取

本小节对预处理之后的留言主题数据进行向量化，其是分类模型建立的基础和关键.我们采用可以评估一个字词对于一个文件集重要程度的 TF-IDF 算法.该算法中，字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降.如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类.该算法的具体实现步骤如下：

（1）计算词频 TF

为了实现 TF-IDF 算法，首先需要计算词频，词频简称为 TF，它表示某个词在一段或一篇文本中出现的次数. TF(x)表示词 x 的词频，文本总词数记为 N，则

$$\mathrm{TF}(x) = \frac{x}{N}$$

（2）计算逆文档频率 IDF

逆文档频率通常简称为 IDF，为了计算 IDF，我们需要计算整个语料库. 语料库由多个文本组成. IDF(x)表示词 x 的逆文档频率，语料库的文本总数记为 A，包含该词的文本数记为 B，则有：

$$\mathrm{IDF}(x)=\log(\frac{A}{\mathrm{B}+1})$$

由上式可见，如果一个词越常见那么分母就越大,IDF(x)就越小越接近 0.分母之所以加 1，是为了避免分母为 0（即所有文档都不包含该词）.IDF(x)越大，则此特征性在文本中的分布越集中，说明词 x 区分该文本内容的属性能力越强.

（3）计算 TF-IDF

将求得词 x 的词频和逆文档频率，计算词 x 的 TF-IDF

$$\mathrm{TF}\text{-}\mathrm{IDF} = \mathrm{TF}(x)\times\mathrm{IDF}(x)$$

可以看到，TF-IDF 与一个 x 在文档中的出现次数成正比，与词在整个语言中的出现次数成反比. 所以，自动提取关键词就是计算出文档的每个词的 TF-IDF 值，然后按降序排列，取排在最前面的几个词即可. 基于以上原理，借助 Python 中的 sklearn 库获得训练集样本的 TF-IDF 权值.

2.1.4 建立朴素贝叶斯分类模型

基于上面的给出的留言主题的 TF-IDF 权重向量，本小节通过建立统计模型对留言内容进行分类. 这里我们采用了朴素贝叶斯算法进行建模，其算法原理可表述如下：朴素贝叶斯分类（简称 NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y. 设有样本数据集 $D=\{d_1,d_2,\cdots d_n\}$ ，对应样本数据的特征属性集为 $X=\{x_1,x_2,\cdots x_d\}$ 类变量为 $Y=\{y_1,y_2\cdots y_m\}$ ，即 D 可以分为

y_m 类别. 其中 $x_1, x_2, \cdots x_d$ 相互独立且随机，则 Y 的先验概率 $P_{prior} = P(Y)$ ，Y 的后验概率 $P_{post} = P(Y \mid X)$ ，根据朴素贝叶斯算法，后验概率可以由先验概 $P_{post} = P(Y)$ 、证据 $P(X)$ 、类条件概率 $P(X \mid Y)$ 计算出

$$P(Y \mid X) = \frac{P(Y)P(X \mid Y)}{P(X)}.$$

朴素贝叶斯算法基于各特征之间相互独立，在给定类别为 y 的情况下，上式可以进一步表示：

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(x_i \mid Y = y).$$

由以上两式可以计算出后验概率为：

$$P_{post} = P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{d} P(x_i \mid Y)}{P(X)}$$

由于 $P(X)$ 的大小是固定不变的，因此在比较后验概率时，只比较上式的分子部分即可. 因此可以得到一个样本数据属于类别 y 的朴素贝叶斯计算如下：

$$P(y_i \mid x_1, x_2, \cdots, x_d) = \frac{P(y_i)\prod_{i=1}^{d} P(x_i \mid y_i)}{\prod_{i=1}^{d} P(x_j)}$$

朴素贝叶斯分类模型有高斯模型、多项式模型和伯努利模型三种. 其中高斯模型适用于多个类型变量，特征为连续型变量且假设特征符合高斯分布；多项式模型用于离散计数. 如一个句子中某个词语重复出现，我们视它们每个都是独立的，所以统计多次；伯努利模型用于特征向量是二进制（即 0 和 1）的构建. 由于本题的特征是离散的，所以我们选择多项式模型. 构建多项式模型的逻辑步骤如下：

1) 将处理好的数据切分为训练集和测试集；

2) 将训练集样本转换为 TF-IDF 向量；

3) 对每个类别计算其先验概率，对每个特征属性计算所有划分的条件概率；

4) 对每个类别计算 $P(x|y_i)$和 $P(y_i)$.

5) 以 $P(x|y_i)$、$P(y_i)$的最大项作为 x 所属的类别

其中有公式：

$$P(\text{所属类别}y_i \mid \text{某种特征}x) = \frac{P(\text{所属类别}y_i) * P(\text{某种特征}x \mid \text{所属类别}y_i)}{P(\text{某种特征}x)}$$

基于朴素贝叶斯模型实现文本分类的流程如图 2：

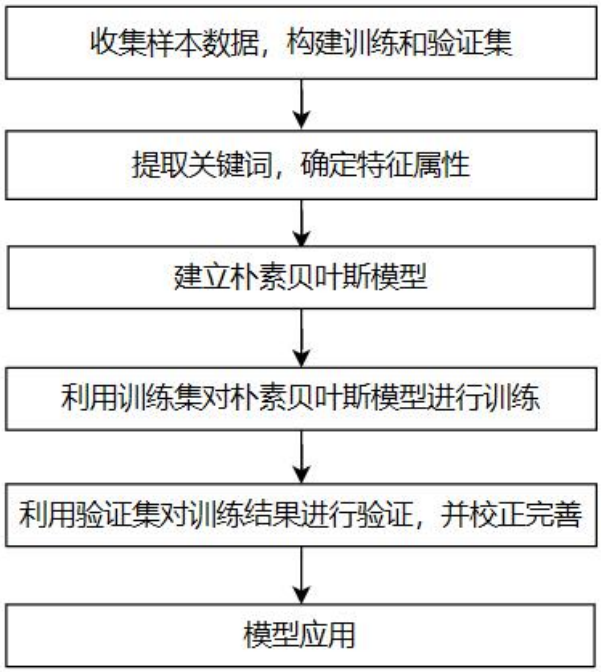


图 2. 朴素贝叶斯模型实现流程

由于 Python 中的 scikit learn 库中有 MultinomialNB 函数，可直接调用用于模型建立，最后对模型进行测试.

2.1.5 模型的优化

（1）样本数据的优化

由于留言主题内容较少且留言主题有时候不能准确反映其所属类别，提取到的留言主题关键词较少即能够用于区分留言类别的较少，因此，严重降低了模型的准确率.

为了减小由于留言主题关键词数目不足和留言主题不能充分反映其所属类别对留言分类造成的影响，我们将留言主题和留言内容合并后作为一个数据帧，对其进行数据预处理、文本特征抽取操作，然后再构建朴素贝叶斯模型，最后对模型进行测试.

（2）抽样方式的优化

将上述模型进行测试验证后发现模型的准确率较低（具体结果见问题一结果分析）.查阅相关资料得知，在构建分类模型时，数据的不平衡性会对分类器带来负面影响从而影响结果的准确性.对样本数据做统计分析，结果如图 3

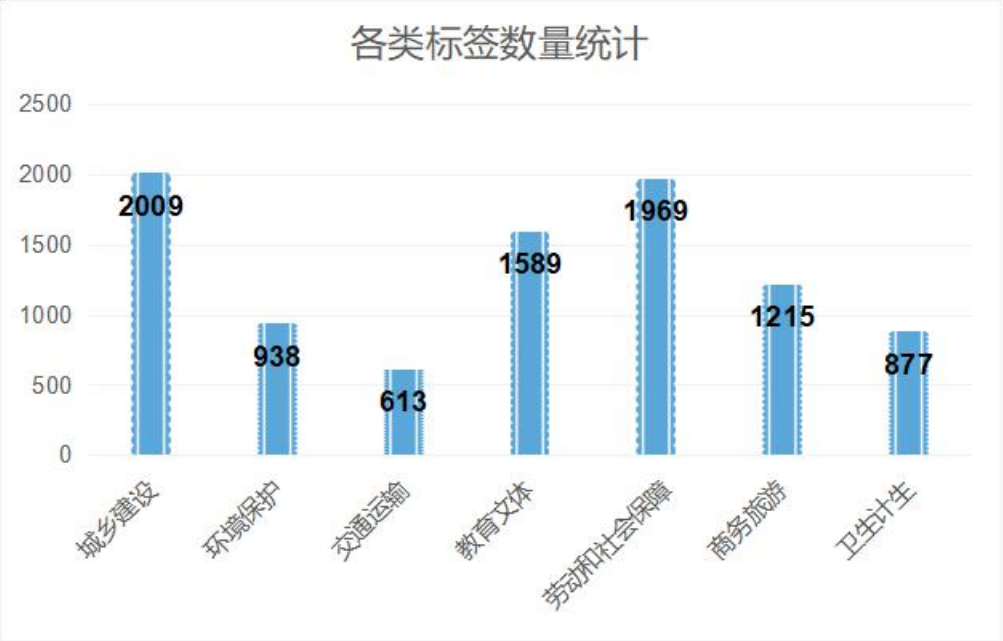


图 3. 各类标签数量统计

由图 3 可见，各类标签的数量不均衡，其中，交通运输类留言有 613 条，数目最少；城乡建设类留言有 2009 条，数目最多，因此，在用 Python 建立朴素贝叶斯模型对数据集进行抽样学习时，为了降低数据的不平衡对分类结果造成的影响，采取欠抽样的抽样方法对样本进行抽样学习.

欠抽样技术是将一些数据从原始数据集中移除从而使数据达到平衡状态，本题中，选取交通运输类留言的数量作为每类留言的抽样总数，欠抽样后，每类留言在数目上是相等的，因此朴素贝叶斯分类器可以实现对数据的均衡学习.

2.2 问题二分析方法与过程

本问题要求在海量留言信息中将留言信息按内容分类，并定义热度评价指标找出其中的热点问题.本节首先对留言进行数据预处理，再用 jieba 中文分词工具对全部留言信息进行分词，通过关键词和核心词抽取模型，而后使用 TF-IDF 实现文本向量化，利用 LSI 模型实现留言信息两两之间的相似度度量计算，从而实现对不同留言的分类处理，最后使用舆情热度计算方法对不同类别的留言进行热度指标量化，使用 jieba 中文分词提取每一类留言的主题词，并将全部结果以及时间范围输出至文件中.

2.2.1 流程图

本问题的分析方法和过程的流程如图 4：

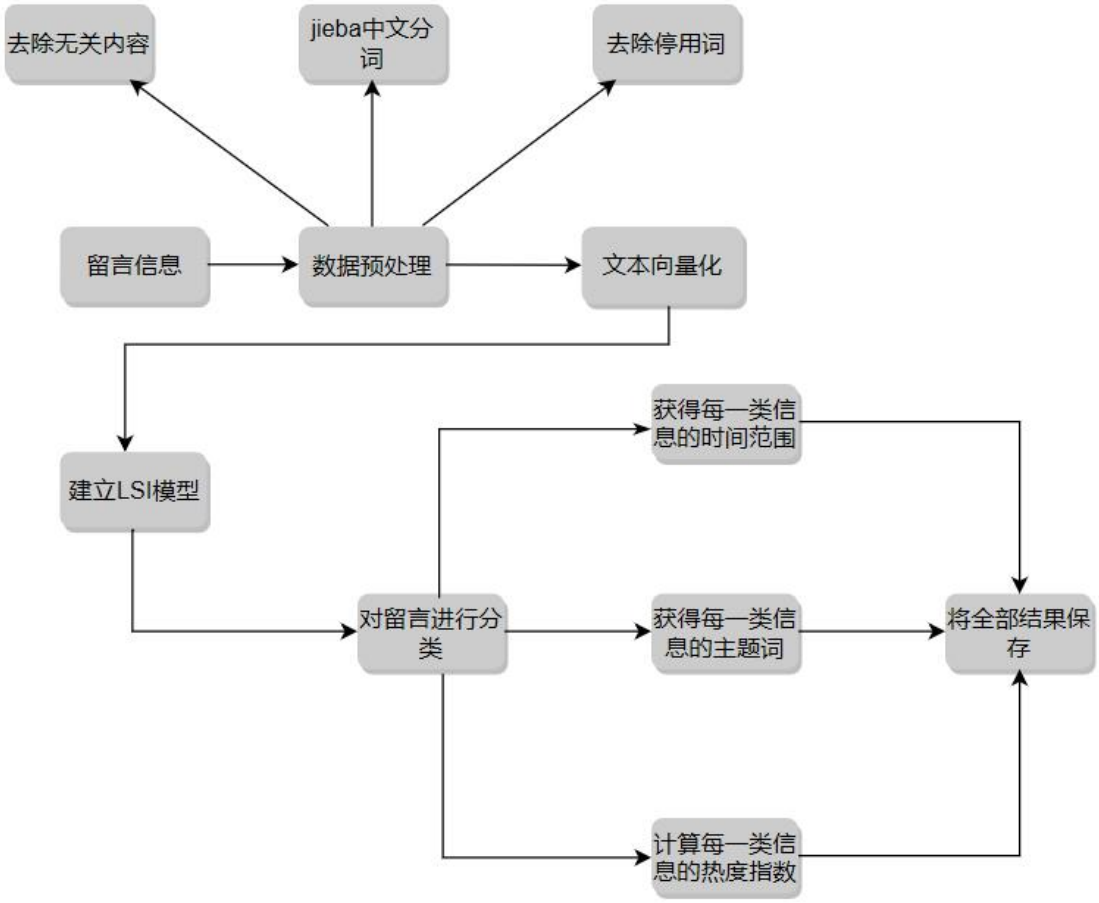


图 4. 问题二流程图

2.2.2 研究方法及理论基础

（1）LSI 模型

LSI(Latent Semantic Indexing)通过海量文献找出词汇之间的关系. 基本理念是当两个词或一组词大量出现在一个文档中时，这些词之间就是语义相关的.

LSI 模型是一种无监督的学习算法，它的原理是先把句子分词，用 BOW 方法对句子进行特征提取，组成句子的特征矩阵. 把矩阵通过奇异值分解(SVD)的方法生成一个降维并去噪的新矩阵，以此来表示原来的矩阵. 特征向量组成的矩阵降维，可以理解为把原来特征映射到其他低维空间. 通过把词语映射到其他的向量空间，使得 LSI 可以捕捉到句子潜在的语义，如同义词，近义词，不同表达方式的相似性.

（2）舆情热度计算方法[2][3]

每一类留言的热度指数应该由以下三个方面因素共同决定：

1）此类中含有的留言条数. 热度应该与留言条数呈正相关，留言条数越多，说明反映此问题的市民越多，此问题也应当被首先考虑解决.

2）市民间的互动行为（即点赞数和反对数）. 点赞数越多，说明认同此留言的市民越多，反对数反之也成立. 通过点赞数、反对数可以灵活、准确地反映出单条（某类）留言的受关注程度，从而使得为留言分类后的留言集合的热度提供准确数据基础

3）留言时间. 由于群众留言热度具有时效性，留言集合热度须随时间流逝而衰减，并且衰减趋势应该是越来越快，直至趋近于零. 否则，会出现曾经的某个留言集合一直处于靠前位置，影响留言问题的时效性和准确性. 因此某一类留言的热度值应该随时间衰减而不是线性.

由以上三方面因素，制订如下具体的热度评价指标：

群众交互热度分：　　S=K1*A-K2*D+...

热度衰减公式：　　Ti=e ^ (k*(t1 – t0))

单个留言热度分：　　Hi=(S0+S)/Ti

该留言类的热度分：　　H=H1+H2+...+Hn

符号说明：

H	热度值
n	群众反映数量
A	点赞数
D	反对数
S0	初始热度分
S	群众交互产生的热度分
Ti	某留言集合中第 i 个留言的热度衰减值
t0	留言发布时间
t1	当前时间
k	Ti 公式中指数参数
Ki	交互得分权值

2.2.3 数据预处理

对于包含留言信息的附件 3，先对全部留言信息进行时间格式统一，再对大量数据先进行数据预处理工作，将留言语句切割为词语，去除可能会对分类产生影响的无关内容，调整一条留言不同部分的权重，以期望不同部分在相似度度量计算时发挥不同的作用.

首先，由对附件内容的初步观察，可以看到留言时间在文件中有两种保存形式，为了方便程序的统一化处理，避免因格式不同造成的错误，所以利用 excel 统一将全部时间格式调整为 yyyy/mm/dd hh:mm:ss.

然后，提取每条留言的留言详情和留言主题，将两者合并为一条留言信息作为分类的初始语料库，同时通过提高留言主题在一条留言信息中的出现次数实现提高留言主题在留言信息中的权重.

其次，将未登录的新词加载至分词库中，利用 jieba 分词器的精确分词模式对初始语料库进行中文分词.

最后，加载自定义词典，对语料库中干扰词语，如停用词、无意义前缀、与内容无关的字符串，进行过滤. 经过预处理后的部分数据见表 3，全部数据保存在 dataprocess/adata.txt 中. 数据预处理代码见附件 dataprocss.py.

表 3. 预处理结果

市	学院	体育	强制	实习	市	学院	体育	学院
体育	学院	变相	市	学院	体育	学院	变相	强制
尔夫	分配	系	说	不准	提前	回	做	满个
市	app	申请	通不过	市	app	申请	购房	补贴
西地省	抗癌	药品	西地省	抗癌	药品	纳入	医保	西地省
西地省	抗癌	药品	西地省	抗癌	药品	纳入	医保	西地省
摊	区	东路	小区	临街	门面	烧烤	夜宵	摊
区	东路	魅力	一楼	夜宵	摊	污染	空气	处理
城轨	公交站	市	公交站	市	南塘	城轨	公交站	市
请求	市	地铁	溪湖	CBD	处	增设	站	请求
增设	站	请求	地铁	建成	消息	请问	这方面	有
市	学院	寒假	过年	期间	学生	工厂	市	学院

2.2.4 文本向量化

在随着数据挖掘与分析中，TF-IDF 可以将文本在计算机中的表示转化为文本向量，从而实现将无法量化的文本转化为可量化计算指标的问题. 而在此之前需要建立关于语料库的词典并计算每条留言的中关键词的词频.

首先，利用 Python 中的 gensim 库建立起关于语料库的词典，而后对每条留言进行词频统计.

在我们所建立的词典中包含 N 个词语，每个词语有唯一的索引，那么对每个留言信息，我们可以使用一个 N 维的向量来表示. 举例如下所示：

[1，2，1，1，1，0，0，0，1，1 ……]

[1，1，1，1，0，1，1，1，0，0 ……]

部分词典见表 4，其中第一列为编号，第二列为语料库中的关键词. 词典数据保存在 dataprocess/dictionary.txt 中.

表 4.词典

0	一名
1	一码事
2	不准
3	书记
4	体育
5	体适

6	做
7	儿童
8	几万块
9	分配
10	前
11	前台
12	变相
13	合同
14	商量
15	回
16	大四
17	天天
18	学期结束
19	学生
20	学院
21	实习

部分词频见图 5，其中每一行对应一条留言信息，对其中的每一个元组而言，前者表示词语在词典中对应的编号，后者表示词语在本条留言中出现的次数. 全部数据保存在 dataprocess/doc_vectors.txt 中.

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 101), (5, 1), (6, 4), (7, 2), (8, 1),  (10, 1), .....(62, 3)]
[(29, 103), (63, 101), (64, 1), (65, 1), (66, 1), (67, 2), (68, 1), (69, 1),  ..... (106, 1)]
[(67, 101), (139, 101), (140, 1), (144, 101), (146, 101), (151, 101), .....(446, 1)]
[(19, 103), (20, 101), (29, 101), (38, 1), (50, 1), (58, 1), (76, 1), .....(195, 1)]
[(29, 114), (148, 101), (149, 1), (175, 6), (190, 1), (227, 1), (304, 2),  .....(551, 1)]
[(29, 100), (241, 1), (552, 1), (553, 1), (554, 1), (555, 1), (556, 1), (557, 2), .....(564, 1)]
[(29, 100), (51, 1), (76, 1), (144, 100), (146, 102), (148, 1), (158, 1), (159, 1),..... (607, 2)]
[(51, 2), (71, 1), (86, 102), (136, 1), (194, 102), (562, 1), (608, 1), (610, 102),.....(616, 1)]
[(67, 100), (144, 101), (146, 103), (148, 2), (151, 101),  (168, 101), .....(650, 1)]
[(10, 1), (67, 1), (76, 1), (99, 1), (101, 1), (169, 1), (238, 2), (427, 1), .....(694, 1)]
[(29, 101), (144, 101), (146, 1), (148, 1), (159, 1), (161, 1), (162, 1),.....(584, 101), ]
[(31, 1), (136, 1), (140, 1), (144, 101), (145, 1), (146, 103), (151, 100),.....(721, 1)]
[(0, 1), (19, 106), (20, 101), (21, 104), (29, 101), (30, 100), (38, 1), .....(820, 1)]
```

图 5. 词频

通过观察发现，在建立的字典中，有一些高频率出现的词并不会在分类中起到较大左右，而只出现在某一条留言中的词语可能具有强烈的感情色彩，即较高的区分度.

因此，在此基础上引入 TF-IDF（Term Frequency-Inverse Document Frequency，词频和逆向文件频率）对每条留言做进一步考量. 通过 TF 计算重要词语的词频，IDF 凸显出现的少但占有强烈感情色彩的词语，从而得到留言信息的文本特征值，即完成了文本向量化，同时我们使得由“词频向量表示一句话”变换成为用“词的重要性向量表示一句话”.

部分向量化数据见表 5，全部数据保存在 dataprocess/tfidf_vectors.txt 中. 文本向量化代码见附件 model.py 中的 GetVector()函数.

表 5. 向量化数据

[(0,0.0034765569125723),(1,0.005135296108803111),(2,0.005135296108803111),...
[(29,0.08680910317875722),(63,0.4675566926176),(64,0.0036308173027491428),...
[(50,0.196743853943517),(107,0.00506292360519),(108,0.005062695236051519),...
[(6,0.003963553656545895),(67,0.1820477272281066),(139,0.359849813677548),...
[(51,0.001779648375763014),(67,0.136756158012592),(80,0.0029832206498029),..
[(20,0.0053390968612561),(29,0.105308567635096),(50,0.002192374869660294),...
[(29,0.066327973342015),(67,0.0009477155427652),(99,0.00183514145537528), ...
[(29,0.1202651947953483),(50,0.00247919487736536),(67,0.0017186708920992),...
[(67,0.2660870340822923),(139,0.52031953000397),(140,0.00579326230935911),...
[(19,0.2613915650117563),(20,0.230234412644352),(29,0.08992415573607272), ...
[(51,0.0027937956332339),(71,0.0031419890878073),(86,0.3204828869563494), ...
[(29,0.094878735532723),(144,0.165818499677244),(146,0.00149933569372807),...

2.2.5 建立 LSI 模型[4][5]

LSI 模型的建立可以分为三个步骤：模型构建与优化、降维、文本相似度计算.

模型的构建与优化：在词典建立和词频统计的基础上，首先用词-文档矩阵表示整个语料库，其中词-文档矩阵是由不同词在不同留言中的数量构成的矩阵. 从而建立起关于语料库具有高维空间的 LSI 模型.

降维：在上一步中，建立了高维度的 LSI 模型，但该模型的维度太高，若直接使用，会造成维度灾难，因此在这一步中要对矩阵进行奇异值分解实现对所建立模型的降维. 通过降维，把高维空间中的词和文档向量投影到更低维的空间使之更易于得到两条留言之间的关系.

文本相似度计算：对经过降维处理后的矩阵使用余弦相似度的方法就可以实现留言信息两两间文本相似度的计算.

对于计算得到的文本相似度，若值达到人为设定的某个阈值，便认为这两条信息为反映同一问题的一类留言. 文本向量化代码见附件 model.py 中的 LSI()函数.

2.2.6 计算留言热度指数

在对留言信息进行分类后，继续利用 Jieba 分词器提取每一类留言信息中的主题词即问题描述关键词和特定的地点/人群. 对于问题描述关键词的获得，利用 Jieba 分词器获得此类留言中权重排名前二十的名词或动词；而对于特定地点/人群，利用分词器获得权重排名前五的地名和人名. 获得主题词和特定地点/人群代码见附件 get_key_words.py.

通过舆情热度计算方法，可以对每一类留言进行热度量化，然后对所有类留言热度进行排序后便可找到问题二中所要求的热点问题. 获得某一类留言热度代码见附件 get_degree_hot.py.

2.2.7 参数优化

问题求解中涉及到的参数有两条留言间区分度的阈值、舆情热度计算方法中各参数.

通过多次调整，发现区分度阈值在 0.35 时，会得到较为理想的结果.通过将舆情热度计算方法中的 S0（初始热度分）、k（Ti 公式中指数参数）、Ki（交互得分权值）分别设置为 1000、$0.1*10^{10}$、40 时，可以使不同类留言之间热度指数具有较好区分性，便于识别当下急需处理的热点问题.

2.3 问题三分析方法与过程

问题三要求从答复的相关性、完整性、可解释性等角度给出一套评价答复意见的方案. 其中通过计算两文本之间的相似度来衡量留言答复和留言内容的相关性；完整性评价通过给出的一个完整性模板，采取满分值，按点得分；可解释性评价通过答复意见中是否含有可以用来支撑答复的相关法律条文或者政府文件以及答复意见是否在一定程度上具有可读性来衡量；最后，本文增加及时性这一评价角度，通过给出的及时性得分公式的计算值来衡量.

2.3.1 流程图

本题流程图如图 6 所示

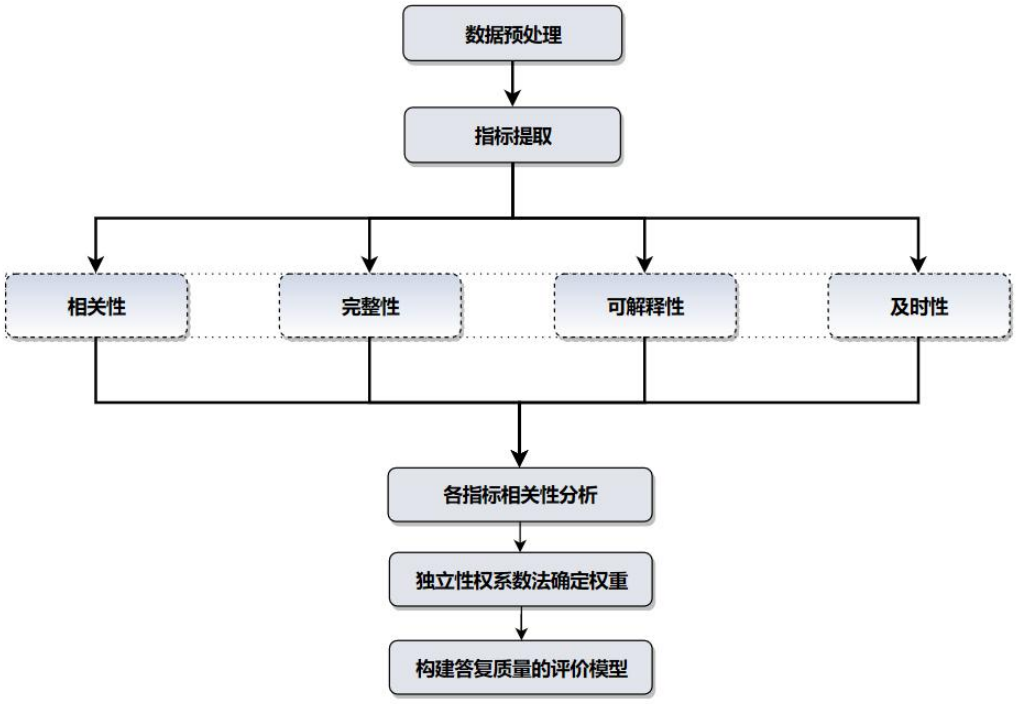


图 6. 问题三流程图

2.3.2 研究方法及相关理论

(1) 词袋模型

词袋模型（Bag-of-words model）是个在自然语言处理和信息检索(IR)下被简化的表达模型. 词袋的含义就是说，像是把一篇文档拆分成一个一个的词条，然后将它们扔进一个袋子里. 文档内容中出现频率越高的词项，越能描述该文档（不考虑停用词）. 因此统计每个词项在每篇文档中出现的次数，即某词项的频率，记为$tf_{t,d}$

其中，t 为词项，d 为文档，获得文档中每个词的 tf 权重，一篇文档则转换成了词—权重的集合，通常称为词袋模型. 该词袋模型可以来描述一篇文档. 在词袋模型中，词项在文档中出现的次序被忽略，出现的次数被通记. 将词项在每篇文档中出现的次数保存在向量中，这就是这篇文档的文档向量.

(2) 余弦相似度计算

获得文档中每个词的 TF-IDF 权重后，一篇文档就转换成了词——文档权重集合，其中每个分量代表词项在文档中的相对重要性. 一系列文档在同一向量空间的表示称为 VSM（词袋模型）.

如果要计算不同文档之间的相似度，可以将得到的文本和向量数据的映射关系，通过计算向量数据的差异大小，来计算文本相似度. 余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性，其范围为[-1,1]，且结果与向量长度无关. 余弦相似性最常用于高维正空间. 例如在信息检索中，每个词项被赋予不同的维度，而一个维度由一个向量表示，其各个维度上的值对应于该词项在文档中出现的频率. 余弦相似度因此可以给出两篇文档在其主题方面的相似度.

对于文本匹配，给定属性向量 A 和 B（A 和 B 通常是文档中的词频向量），由于一个词的频率（TF-IDF 权）不能为负数，所以这两个文档的余弦相似性范围从 0 到 1. 并且，两个词的频率向量之间的角度不能大于 90°. 计算公式如下：

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

这里 A_i， B_i 分别代表向量 A 和 B 的各分量.

(3) 独立性权系数法

独立性权系数法[7]独立性权系数法是根据各指标与其他指标之间的共线性强弱来确定指标权重的. 设有指标项 $X_1, X_2, \cdots X_m$，若指标 X_k 与其他指标的复相关系数越大，则说明 X_k 与其他指标之间的共线性关系越强，越容易由其他指标的

线性组合表示，重复信息越多，因此该指标的权重也就应该越小. 若指标 X_k 与其他指标的复相关系数越大，该指标的权重越小. 其中 $R=\frac{\sum(y-\bar{y})(\hat{y}-\bar{y})}{\sqrt{\sum(y-\bar{y})^2\sum(\hat{y}-\bar{y})^2}}$. 取 R 的倒数作为得分，再经归一化处理得到权重系数.

2.3.3 指标提取

对答复意见的质量进行评价，首先需要提取影响答复意见质量的指标. 根据题目要求查阅相关资料[6]，我们以答复的相关性、完整性、可解释性、及时性作为评价答复意见的四个特性指标.

(1) 相关性

相关性，是指两个变量的关联程度. 本题中，探究答复意见和留言内容的相关性，即两个文本之间的相关性. 本文通过计算两文本之间的相似度来衡量留言答复和留言内容的相关性.步骤如图 7 下：

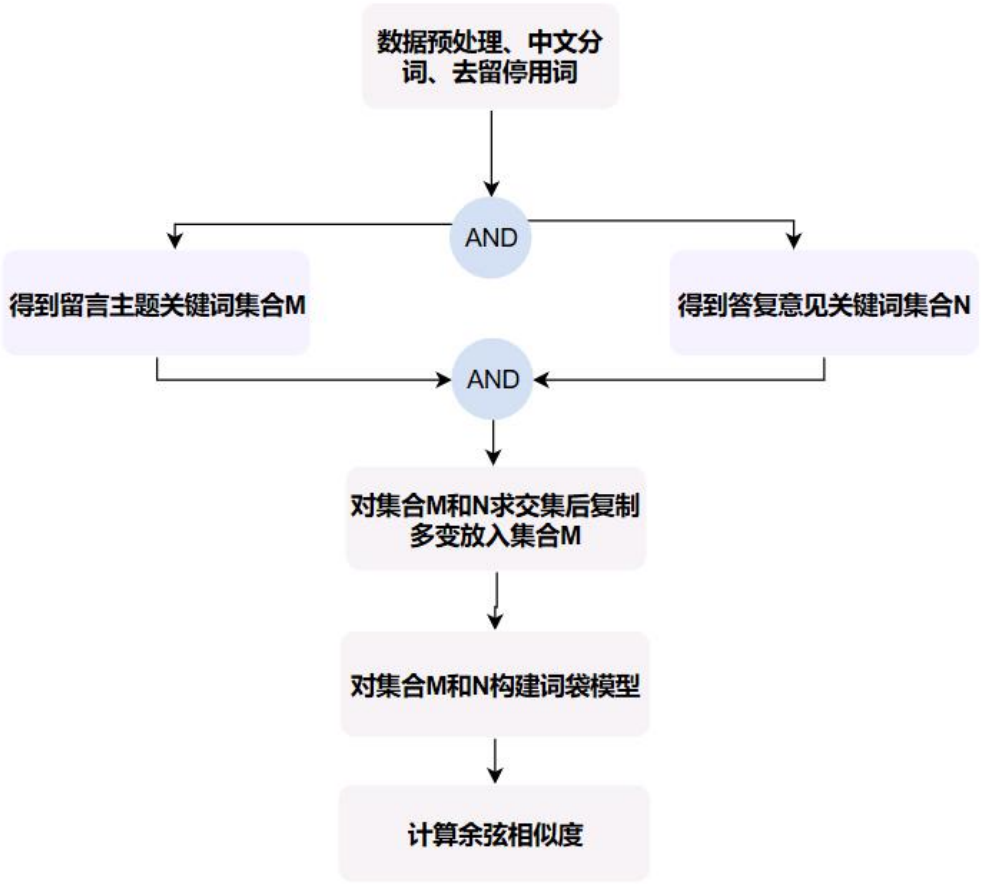


图 7. 余弦相似度计算流程

其中，多次复制留言主题关键词到答复意见关键词集合的目的是为了在一定程度上消除无关词语对相似度的影响，使集合 M 和集合 N 的向量表示更具比较意义.

(2) 完整性

为了简化模型，本文自定义了一个评价答复意见完整性的模板，采取满分制，按点得分. 自定义模板的六个得分点如下：含有尊敬的、您好（或你好）、谢谢（或感谢）、重视、答复（或回复），字数不少于 100 字.

(3) 可解释性

可解释性我们通过答复意见中是否含有可以用来支撑答复的相关法律条文或者政府文件以及答复意见是否在一定程度上具有可读性来衡量. 其中，可读性可以用自动化可读性指数 ARI（Automated Readability Index）来表示.搜集相关文本我们发现，ARI 常用于类似的研究中，计算可以直接采用公式：

$$ARI=4.71*（总字符数/总字数）+0.5*（总字数/总句数）-21.43$$

(4) 及时性

根据答复时间和留言时间的差值作为计算及时性得分 P 的公式.

$$P=\begin{cases}100\text{ 分，}\ time \le 7day\\ 80\text{ 分，}\ 7day < time \le 14day\\ 60\text{ 分，}\ 14day < time \le 21day\\ 20\text{ 分，}\ 21day < time \le 30day\\ 0\text{ 分，}\ time > 30day\end{cases}$$

2.3.4 构建评价答复意见质量的评价方案

(1) 计算各项指标得分

根据上述各指标计算方法，我们得到附件 4 答复意见的评分. 部分结果如表 6 所示：

表 6. 答复意见评分结果

留言编号	相关性	完整性	及时性	可解释性
2549	0.998999499	66.66666667	60	18.65998442
2554	0.999499875	83.33333333	60	6.445654762
2555	0.961249187	66.66666667	60	15.99489362
2557	0.86890736	50	60	2.913584416
2574	0.888819442	66.66666667	60	1.700838926
2759	0.999499875	83.33333333	0	19.3411315
2849	1	83.33333333	0	5.850921053
33970	0.902773504	50	80	3.027281106

33978	0.930053762	50	80	1.872068452
33984	0	33.33333333	100	1.203333333
34239	1	50	0	1.34218126

(2) 相关性分析

为了探究各指标之间的关系，降低各指标之间的相关性对模型构建的影响，本文把各指标和其余三个指标做了相关性分析，结果如下：

系数 [a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	42.169	8.613		4.896	.000
	完整性	.555	.127	.382	4.354	.000
	及时性	-.003	.075	-.004	-.045	.964
	可解释性	.565	.233	.211	2.421	.017

a. 因变量：相关性

系数 [a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	31.684	5.670		5.588	.000
	及时性	-.058	.050	-.090	-1.144	.255
	可解释性	.464	.156	.252	2.984	.003
	相关性	.253	.058	.368	4.354	.000

a. 因变量：完整性

系数 [a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	77.689	9.193		8.451	.000
	可解释性	-.076	.295	-.026	-.258	.797
	相关性	-.005	.115	-.005	-.045	.964
	完整性	-.193	.169	-.124	-1.144	.255

a. 因变量：及时性

系数 [a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	-7.525	3.608		-2.086	.039
	相关性	.085	.035	.228	2.421	.017
	完整性	.154	.051	.283	2.984	.003
	及时性	-.008	.029	-.022	-.258	.797

a. 因变量：可解释性

(3) 构建评价方案[8]

由相关性分析可以看出，四个指标之间存在着不同程度的相关性.为了降低

各指标之间的相关性对权重确定的影响，本文采用根据各指标与其他指标之间的共线性强弱来确定指标权重的独立性权系数法.将得到的各指标权重代入答复意见的各项评分，即可得到每条答复意见的综合评分.

3. 结果分析

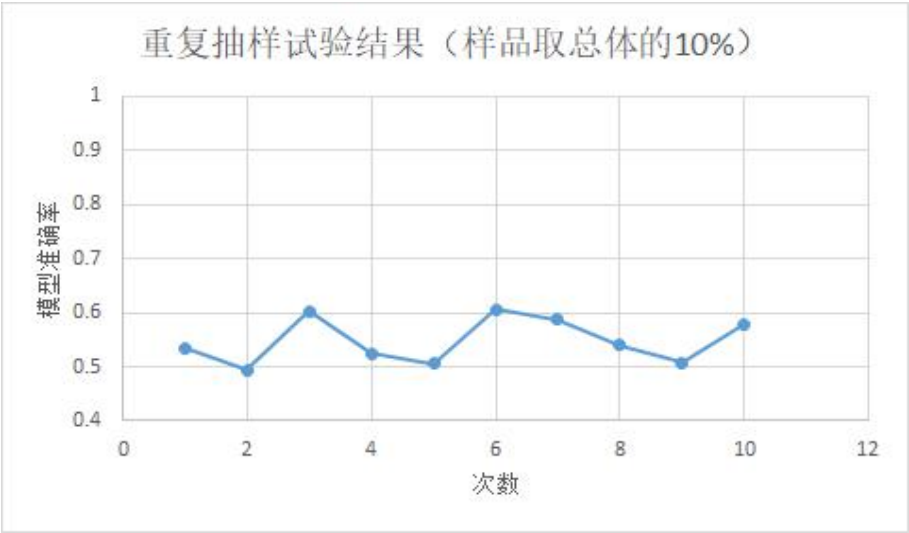
3.1 问题一的结果分析

3.1.1 朴素贝叶斯模型的测试结果

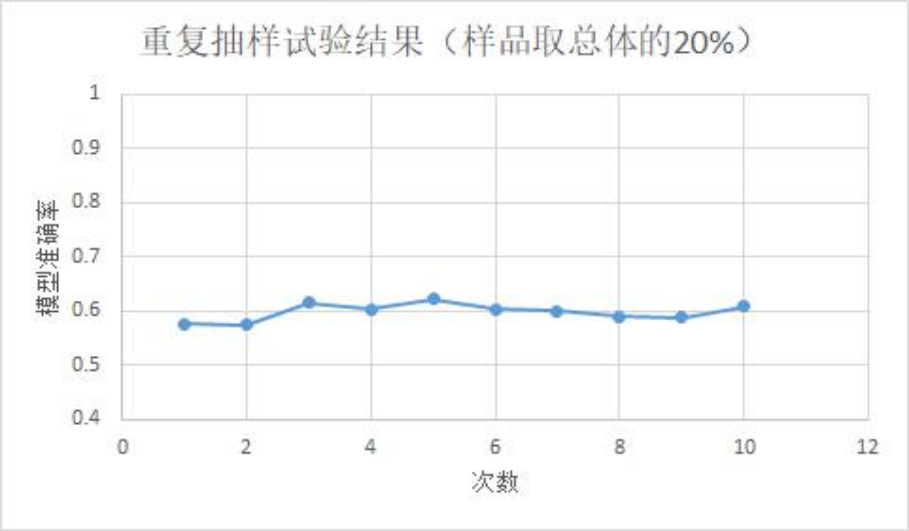
关于留言主题的一级标签分类模型准确率测试结果如表 7，图 8 所示，

表 7. 分类模型测试结果

次数	准确率（样品取总体的 10%）	准确率（样品取总体的 20%）
1	0.535135135	0.576368876
2	0.504565217	0.573487032
3	0.60326087	0.614942529
4	0.524324324	0.602305476
5	0.505434783	0.621776504
6	0.605405405	0.603448276
7	0.586956522	0.599423631
8	0.540540541	0.58908046
9	0.508108108	0.587896254
10	0.578378378	0.608069164



(a)



(b)

图 8. 重复抽样试验结果

由图 8 可见，测试的准确率最高不超过 62%,在 50%-62%之间上下波动.

3.1.2 模型优化后的测试结果

（1）样本数据优化结果

将留言主题和留言内容合并后作为一个数据帧，对其进行数据预处理、文本特征抽取操作，然后再构建朴素贝叶斯模型，测试模型的准确率结果如表 8，图所 9 示：

表 8. 优化测试结果

次数	准确率（样品取总体的 10%）	准确率（样品取总体的 20%）
1	0.81917808	0.767241379
2	0.777472527	0.793696275
3	0.81369863	0.75971223
4	0.832876712	0.780172414
5	0.812672176	0.779053085
6	0.760330579	0.785100287
7	0.799450549	0.787661406
8	0.798898072	0.803443329
9	0.782369146	0.78705036
10	0.793956044	0.771551724

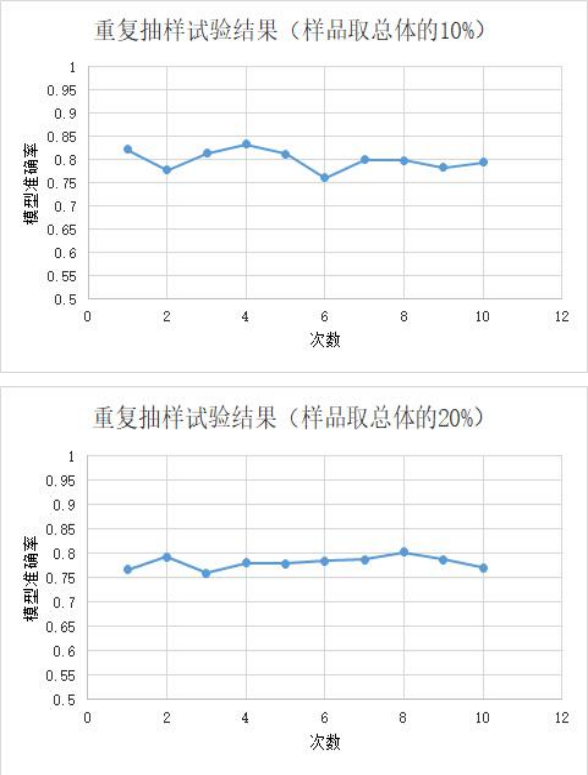


图 9. 分类模型测试结果图像

由图 9 可见，测试的准确率大大提高，在 75%-83%之间上下波动，比优化前有明前的提高.

（2）抽样方式优化结果

在对样本数据进行优化后，采用欠抽样方式对七类留言进行抽样，利用 F-score 对分类结果进行评价，多次测试求平均，结果如图 10 所示.

类别	precision	recall	f1-score
劳动和社会保障	0.78	0.93	0.85
城乡建设	0.85	0.90	0.88
教育文体	0.94	0.80	0.87
卫生计生	0.93	0.90	0.92
交通运输	0.92	0.85	0.89
商务旅游	0.89	0.88	0.89
环境保护	0.90	0.95	0.93
weighted avg	0.89	0.89	0.89

图 10 优化结果

其中，weighted avg 为 f1-score 加权求平均的结果. 由图 10 可见，经过样本数据和抽样方式优化后测试结果的准确率平均 89%，f1-score 的数值可达 89%，说明分类的精确度可达 89%，效果颇佳.

3.2 问题二的结果分析

将 2.2.4 节得到的每一类留言并对其赋予不同的 ID 编号、2.2.6 节得到每一类

的留言热度指数、2.2.5 节得到每一类留言的关键词以及每一类留言时间范围，对应地输出到 res/res.xls 文件中.部分输出结果如图 11 所示，文件输出代码见附件 print_file.py.

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	热度指数	问题描述关键词	地点人群关键词	时间范围
80	190337	A0009051	关于伊景园	2019/08/2	投诉伊景园	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	195511	A909237	车位捆绑违	2019/08/1	对于伊景园	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	195995	A909199	关于广铁集	2019/08/1	尊敬的市政	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	196264	A0009508	投诉A市伊	2019/08/0	A市伊景园	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	199190	A0009508	关于A市武	2019/08/0	武广新城为	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	204960	A909192	家里本来就	2019/08/2	我是广铁集	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	205277	A909234	伊景园滨河	2019/08/1	广铁集团强	0	1	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	205982	A909168	坚决反对伊	2019/08/0	我坚决反对	0	2	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	207243	A909175	伊景园滨河	2019/08/2	您好！A市	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	209571	A909200	伊景园滨河	2019/08/2	广铁集团钱	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	218709	A0001066	A市伊景园	2019/08/0	伊景园滨河	0	1	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	218739	A909184	A市伊景园	2019/08/2	A市伊景园	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0
80	220534	A0007509	投诉武广新	2019/08/1	投诉：武广	0	0	26753.6	景园、车位、捆	滨河、广铁集团、	2019/07/0

图 11. 输出结果

通过对留言热度指数进行降序排序，可以明显地看出，相同主题的留言被很好地集中在了一起，留言热度指数具有较好区分度，问题描述关键词以及地点人群关键词筛选效果良好，时间范围准确.

最后从经过热度降序排序的留言中选取热度指数最高的前五类留言，分别对应地添加到热点问题表.xls 以及热点问题留言明细表.xls 中.

3.3 问题三的结果分析

3.3.1 指标提取结果以及指标说明

指标	说明
相关性	答复意见与留言主题的相关性
完整性	衡量答复意见是否包含必要的关键词和是否符合字数要求
及时性	判断答复意见的回复是否及时
可解释性	用于检测答复意见的被理解成度以及是否含有相关依据

3.3.2 独立性权系数法确定权重结果

(1) 首先将附件四答复意见的评分结果做标准化处理，部分结果如下：

表 9. 标准化处理表

Z 相关性	Z 完整性	Z 及时性	Z 可解释性
0.84275	0.79465	-0.21093	1.06046
0.84427	1.52817	-0.21093	0.07074
0.72842	0.79465	-0.21093	0.84451

0.44873	0.06113	-0.21093	-0.21546
0.50904	0.79465	-0.21093	-0.31373
0.84427	1.52817	-1.89837	1.11566
0.84578	1.52817	-1.89837	0.02255
0.55131	0.06113	0.35155	-0.20625
0.63393	0.06113	0.35155	-0.29986
-2.18299	-0.67240	0.91403	-0.35404
0.84578	0.06113	-1.89837	-0.34279

(2) 利用 SPSS 计算各指标的复相关系数，最后输出各指标权重结果如下：

表 10. 独立性权重法计算结果表

独立性权重法计算结果			
项	复相关系数 R	复相关系数倒数 $1/R$	权重
Z 相关性	0.46	2.176	22.74%
Z 完整性	0.556	1.797	18.78%
Z 可解释性	0.504	1.985	20.75%
Z 及时性	0.277	3.611	37.73%

3.3.3 综合评价结果

(1) 利用独立性权重法确定的权重，计算附件四中答复意见的综合评分.本文以附件四中的前 120 条数据作为分析对象，计算答复意见的综合得分.部分结果如图 12 所示：

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	综合得分
179855	A000105818	家森林公园	18/10/24 22:27:	以后，路上的森林公		2018/10/30 9:42:29	79.37341972
35818	A00098677	生育险其配	18/11/9 14:58:	一直没有买生育证编		2018/11/15 9:42:34	79.15340172
175865	A000105046	文昌社区负	18/6/24 12:52:	官员到我家贫政策优		2018/6/26 14:37:18	76.91844172
168962	A00093003	4县户籍迁	18/11/14 12:46:	回原籍，转学学生到		018/11/20 11:53:0	75.08366654
176604	A000112959	干部独生子	013/8/1 22:41:2	直不发，不休干部需		2013/8/5 10:11:25	74.92831036
175800	A00022258	真高兴村养	019/5/6 10:29:1	一年，当初雨水沟20		2019/5/8 17:02:38	74.56852231
50412	A00072615	样的情况要	13/10/12 16:09:	怀第二胎者的双方		013/10/14 15:14:2	74.12275659
47751	A00088278	寒假时间边	18/2/24 15:54:	知道这是对告知学生		2018/2/26 15:44:42	73.65399287
176410	A00036931	字在"黑"加	016/6/7 15:39:0	费。再个，	在下达停	2016/6/8 17:40:15	72.59494719
175790	A00090006	家坪养猪场	19/6/26 23:47:	季遇到雨当新岗镇班		2019/7/2 11:06:28	72.53097525
41349	A000100583	田红路中段	16/5/18 15:43:	间天气越来摊点进行		2016/5/20 17:42:22	72.32240813

图 12. 综合评价得分

(2) 综合排名前二的答复意见具体内容如下：

留言编号	留言用户	留言主题	留言时间	答复时间	综合得分
179855	A000105818	建议 H 市国家森林公园安置路牌	2018/10/24 22:27:42	2018/10/30 9:42:29	79.37
35818	A00098677	B 市男职工生育险其配	2018/11/9	2018/11/15	79.15

		偶能报吗？	14:58:05	9:42:34	

图 13. 排名前二的数据信息

答复意见
您好，您所反映的问题，已转交相关部门调查处置尊敬的网友： 你好！就您在 10 月 24 日《问政西地省》上所反映的问题，我局高度重视，迅速安排专人调查了解情况，现回复如下.一是您反映的关于老木峪隧道岔路口无路灯的问题：您反映问题时间是在 10 月下旬，正值路灯和标牌安装施工期间，当时部分路段的路灯还未通电.截至 11 月 1 日，G1 区山大道景区段（含您在信中所提及的老木峪岔路口）的路灯安装和接线通电已全部通电亮灯. 二是您反映的森林公园标志牌的问题：按照 G1 区山大道的设计图纸和道路交通标志和标线国家标准及施工规范《GB-5768-2009》,G1 区山大道森林公园路口的交通标志设置是完全参照以上图纸和规范设置的，一般道路标牌只会设置地名及路名，而在路口还会设置一块棕色的 H 市国家森林公园旅游专用指示标志牌，是符合设计、国家标准及相关规范的.目前该路段的道路标牌和 H 市国家森林公园旅游专用指示标牌正在安装施工中. 感谢您对我部工作的支持. 2018 年 11 月 9 日
尊敬的网友：您好，来信收悉！根据西地省人民政府令第 179 号：第十三条 用人单位男职工的配偶生育第一胎，其配偶无工作单位的，从生育保险基金中支付一次性生育补助金，标准为统筹地区上年度平均生育医疗费用的 50%.男职工一次性生育补助金（男职工配偶无工作单位且生育第一胎，产后 6 个月至 1 年以内由单位统一办理，申领期间不得停保）办理资料： 1、生育证原件、复印件（生育证编号 1 开头的） 2、女方无工作证明（村、居委会出具）原件、复印件 3、婴儿出生证原件、复印件 4、单位盖公章的申请表 2 份(在 B 市人力资源和社会保障网下载） 5、单位盖财务专用章的收据（背面写清单位开户名、开户银行、账号、单位联系人及电话）备注：在 B 市本级参加了基本医疗、生育保险的参保职工在市本级正常连续缴费满 10 个月的次月起方可享受生育保险相关待遇.感谢来信！

图 14. 排名前二的答复意见

由排名前二的答复意见可见，这两条答复和留言主题关联度大，具有较高的相关性；具有必要的敬辞和关键词语，符合字数要求，符合完整性要求；含有必要的政府文件或法律条文，便于阅读，具有较强的可解释性；答复时间和留言时间的差值均少于 7 天，较为及时.

(3) 综合排名后二的答复意见如下：

留言编号	留言用户	留言主题	留言时间	答复时间	综合得分
172654	A00076496	咨询 I4 县农村危房改造资金	2014/12/14 21:10:38	2015/1/6 10:21:18	16.09
166919	A00085038	为何 EMS 快递身份证收费这么贵？	2018/7/11 14:44:22	2018/8/7 15:34:12	10.70

图 15. 排名后二的数据信息

图 16. 排名后二的答复意见

答复意见
尊敬的网友： 您好！您所反映的的事情已收悉，现回复如下：
A00085038：您在《问政西地省感谢您的理解与支持！2018 年 8 月 6 日

由排名后二的答复意见可见，答复意见和留言主题相关性不大；内容不完备，字数不符合要求，完整性欠缺；没有真正解决市民的问题，答非所问，不具备可解释性；答复时间和留言时间间隔太长，不具备及时性.

4. 结论

总结本次比赛，我们基于 TF-IDF 权重法提取特征词，建立基于朴素贝叶斯算法的留言内容的一级标签分类模型，并通过多次将留言主题关键词映射到向量空间以及用欠抽样的抽样方法平衡样本数据，提高了模型训练的准确率；使用基于 TF-IDF 的 LSI 模型实现留言信息的相似度度量计算，使用舆情热度计算方法以及结巴中文分词获得每一类留言的热度指标和主题词，得到排名前 5 的热点问题；提取影响答复意见质量的四个主要指标，通过独立性权系数法确定各指标的权重，给出关于评价答复意见质量的方案.

5.参考文献

[1] 梁昌明,李冬强. 基于新浪热门平台的微博热度评价指标体系实证研究[J]. 情报学报,2015,34(12),1278-1283.

[2] liudongdong19. 热度算法，基于内容，用户个性化推荐[J/OL].https://blog.csdn.net/liudongdong19/article/details/80141994.2018-04-29.

[3] 夏火松.基于特征提取改进的在线评论有效性分类模型[J].情报学报.2015,34(5):493-500.

[4] 聂卉.基于内容分析的用户评论质量的评价与预测[J].图书情报工作,2014,58(13):83-89.

[5] 陈涛,谢丽莎.在线评论文本信息质量等级的测量探析——基于模糊综合评价法[J], 科技创业月刊,2012(7):50-52.

[6] 吴秋琴,徐元科,梁佳聚等.互联网背景下在线评论质量与网站形象的影响研究[J].科学管理研究,2012,30(1):81-88.

[7] 刘自远,刘成福.综合评价中指标权重系数确实方法探讨[J].中国卫生质量管理,2006,13(69).

[8] 郭银灵.基于文本分析的在线评论质量评价模型研究[D].内蒙古:内蒙古大学,2017.
[9] 寒若雪.Python 自然语言处理---TF-IDF 模型[J/OL].
https://www.cnblogs.com/no-tears-girl/p/6430737.html.2017-02-2.