所选题目：c 题

综合评定成绩：

评委评语：

评委签名：

网络问政平台留言的挖掘与分析

摘要

近年来，网络问政平台上反应民意的文本数据不段增加，建立基于自然语言处理技术来处理文本信息具有重要意义。本文对问政平台的群众留言进行文本分类，建立一级分类模型，按一定标准提取热点问题，对答复意见的质量给出了一套评价方案

对于问题一，第一步，我们对文本数据进行预处理，运用的方法为 jieba 分词和停用词过滤。第二步，我们利用 IF-IDF 算法提取每条留言的前 5 个关键词，再利用 IF-IDF 将词语转换为权重向量。第三步，我们构建朴素贝叶斯分类模型，并对模型进行测试。最后，利用 F1 分数评估模型。本题运用最适合用于基于词频的高维数据分类器——朴素贝叶斯分类器，算法原理简单，准确率高，可行性较强，旨在真正达到减少工作量、提高准确率的目的。可以广泛应用于问政平台，对留言类别进行划分。

对于问题二，第一步，对留言文本进行分类，依然采用问题一的方法，通过分词对文本进行预处理、IF-IDF 算法构建词空间向量、利用朴素贝叶斯分类模型分类三个步骤，将留言分类，并对每一类留言进行关键词提取，加以概括。第二步，我们以层次分析法为基础，建立留言热度衡量标准模型。由留言数量比例、点赞量、反对数构成第二层结构。利用 matlab 软件构造判断矩阵，并完成特征向量归一化处理、判断矩阵的一致性检验的步骤，最后算得各留言热度指数并排序，得出热点问题。本题模型优点在于不割断各个因素对热度评价结果的影响，且影响程度量化。

对于问题三，我们基于层次分析模型建立了一套评价方案：第一步，将相关性、完整性、可解释性和及时性定义为准则层变量，对每一变量评价指标给出一定标准；第二步，基于层次分析法建立评价模型，利用软件构造判断矩阵，进行特征向量归一化处理、判断矩阵的一致性检验；第三步，计算留言答复意见的质量指数。本题模型从群众的角度考虑留言回复的质量，综合多方面因素建立质量函数，旨在提供客观、完整、易行的评价方案。

综上，本文所建立的模型、函数及运用的方法基本遵循易于理解、结果精准、结论中肯、可行性强的宗旨，在进行充分的理论分析的基础上，充分利用数学计算软件，在输出计算结果、介入人工检验、各个计算过程衔接的方面均获得较显著效果。

关键词 jieba 分词、 TF-IDF 算法、朴素贝叶斯分类模型、层次分析法

Abstract

In recent years, the number of text data that responds to public opinion on the Internet questioning platform has increased, and it is of great significance to establish natural language processing technology to process text information.
This article classifies the texts of the masses on the questioning platform, establishes a first-level classification model, extracts hot questions according to certain standards, and gives a set of evaluation plans for the quality of the answers.

For question one, the first step, we preprocess the text data, using the method of jieba word segmentation and stop word filtering. In the second step, we use the IF-IDF algorithm to extract the first 5 keywords of each message, and then use the IF-IDF to convert the words into a weight vector. In the third step, we build a naive Bayes classification model and test the model.
Finally, use the F1 score to evaluate the model. This question uses the most suitable for high-dimensional data classifier based on word frequency-Naive Bayes classifier. The algorithm is simple in principle, high in accuracy and strong in feasibility. It aims to truly reduce the workload and improve the accuracy. It can be widely used in the political inquiry platform to classify message types.

For question two, the first step is to classify the text of the message. The method of question one is still used. The text is preprocessed by word segmentation, the IF-IDF algorithm is used to construct the word space vector, and the naive Bayes classification model is used to classify the three steps Classify messages, and extract keywords for each type of message to summarize.
In the second step, based on the analytic hierarchy process, we establish a standard model for measuring the message popularity. The second layer structure is composed of the number of messages, the number of likes and the number of objections. The matlab software is used to construct the judgment matrix, and the steps of normalizing the feature vectors and checking the consistency of the judgment matrix are completed. Finally, the hot index of each message iscalculated and sorted, and hot issues are obtained. The advantage of this model is that it does not cut off the influence of various factors on the evaluation results of heat, and the degree of influence is quantified.

For question three, we established a set of evaluation schemes based on the analytic hierarchy process model: In the first step, the correlation, completeness, interpretability, and timeliness were defined as criteria-level variables, and certain criteria were given for each variable evaluation index; In the second step, an evaluation model is established based on the analytic hierarchy process, a judgment matrix is constructed using software, the feature vectors are normalized, and the consistency check of the judgment matrix is performed. In the third step, the quality index of the message reply opinion is calculated. The model of this question

considers the quality of the message reply from the perspective of the masses, and integrates various factors to establish a quality function, aiming to provide an objective, complete and easy evaluation plan.

In summary, the models, functions, and methods used in this article basically follow the purposes of easy to understand, accurate results, fair conclusions, and strong feasibility. On the basis of adequate theoretical analysis, full use of mathematical calculation software, output calculation results, interventions in manual inspections, and convergence of various calculation processes have all achieved significant results.

Keywords： jieba word segmentation, TF-IDF algorithm, naive Bayes classification model, analytic hierarchy process

目录

一、挖掘目标

本次挖掘的目标：利用一定的分类模型，对附件二留言文本进行分类，建立一级标签，此模型旨在减少人工工作量，提高效率和准确率。创建合理的热度评价标准，快速、准确地提取热点留言，有助于相关部门进行有针对性地处理，提升服务效率。答复意见的质量给出一套评价方案，并尝试实现，帮助相关部门更好的服务群众。

二、分析方法与过程

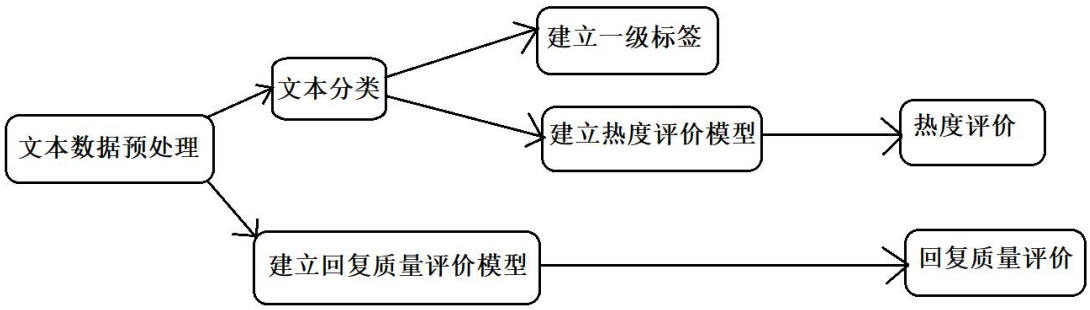
1.总体流程

问题一的主要目标为建立文本分类的模型。我们首先进行文本数据处理得到分词后的文本数据，在此基础上构建词空间向量，用向量间的相似度来反应文本相似度，采用朴素贝叶斯分类模型，完成文本分类，最后对分类的结果进行评价分析。

问题二的重点在于建立热度评价标准。在热度分析前，同样需要对文本进行分类，方法与问题一基本一致，分类结束后，建立层次分析模型。我们将留言数量、点赞数和反对数作为热度标准的几项指标，进而对每一类问题的热度进行计算，最后得出热点文题表和热点问题明细表。

问题三的主演目的为对答复意见的质量给出一套评价方案。这里我们选取了相关性、完整性、可解释性和及时性几个指标作为影响因素，建立层次分析模型，计算留言回复意见的质量指数。

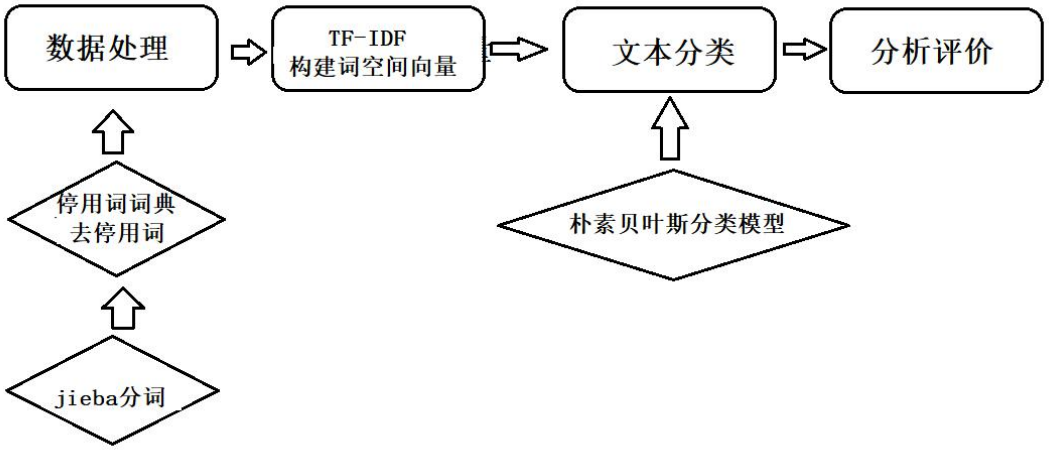
C 题总体流程图如下：



2. 具体步骤

2.1 问题一分析方法与过程

（1）流程图



（2）文本数据处理：

（a）文本分词。

我们对附件 1 中的文本数据进行进行分词处理，即将连续的字序列按照一定的规范重新组合成词序列。常用的文本分词方法有 jieba 分词、HanLP 分词、pyltp 分词。针对本题的特点，我们采用 jieba 分词的方法。

Jieba 分词原理：以前缀词典为基础，实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)；采用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，使用了 Viterbi 算法。优点：Jieba 分词可以将句子最精确地切开，适合文本分析；同时有全模式的优点，把句子中所有的可以成词的词语都扫描出来，速度非常快，搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

对于本题附件 2 的文本数据，进行 jieba 分词的处理。

（b）去停用词。经过 jieba 分词这一步骤，将初始的文本处理成为词的集合，即 $d=(w_1, w_2, w_3\cdots\cdots)$ 。但是文本中含有对文本含义表达无意义的词语，应进行删除，以消除它们对文本挖掘工作的不良影响，此类词称为停用词。停用词的两个特征为:一是极其普遍、

出现频率高:二是包含信息量低，对文本标识无意义。在特征选取的过程中，停用词的介入可能会造成选出的特征几乎都是停用词，从而影响结果的分析。但是在停用词的去除中，应注意要保留其中的否定词,可以对停用词表进行人工筛选相结合的方式,对停用词进行处理。

文本采用停用词词典的文本停用词过滤方式，将分词结果于停用词表中的词语进行匹配，若匹配成功，则进行删除处理。

经过预处理后，得出每个分类中前 100 个高频词。

部分高频词结果：

城乡建设：解决 问题 住房 严重 治疗 公司 公园 公积金 小区 改造 房产证 居民 建议 建设 请求 业主 咨询。

交通运输：客运 建议 运输车 收费 快递公司 请 公路 收费 严重问题 交通 打表 邮政 出租 乱收费 路面 A 区 出行 物流 司机

劳动和社会保障 员工 社保 退休咨询 工资 医保人员 职工 解决 企业 问题 公司 工作 报销 政策 有限公司 公务员 工商 事业单位 待遇 养老金 新农

卫生计生：独生子女 医院 请问 再婚 超声 批的手 家庭问题 人民 政策 计划生育 二胎 生育 谢谢 卫生院办理 咨询 一流 医生 准生证 超声

环境保护：污水 公司 扰民 严重 养猪场 排放 抖音 污染 居民 污染环境 问题 环保局 影响 有限公司 生产

教育文体:补课 学校 教育局 教育 小学 中学 学生 幼儿园 违规招生 反应 咨询 文化 乱收费

商贸旅游;收费 传销 市场 问题 小区 质量 垄断 公司 电梯 景区 乱收费 涉嫌 有限公司 价格 旅游景区 存在 故障 违规 举报 严重

(2)构建词向量空间

在对留言信息分词后，需要把这些词语转换为向量,以供挖掘分析使用。用数学向量来代替文本数据，向量的相似度一定程度上反应了文本的相似度。这里采用 TF-IDF 算法，把职位描述信息转换为权重向量。TF-IDF 算法的具体原理如下:

第一步，计算词频，即 TF 权重(Term Frequency)。

词频(TF). =某个词在文本中出现的次数

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$\text{词频(TF)} = \frac{\text{某个词在文本中出现的词数}}{\text{文本的总次数}}$$

或

$$\text{词频(TF)} = \frac{\text{某个词在文本中的出现次数}}{\text{该文本出现次数最多的词的出现次数}}$$

第二步，计算 IDF 权重，即逆文档频率(Inverse Document Frequency)，需要建立一个语料库(corpus)，用来模拟语言的使用环境。IDF 越大，此特征性在文本.中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率(IDF)} = 1 \frac{\text{语料库的文本总数}}{\text{包含该词的文本数+1}})$$

生成 TF-IDF 向量的具体步骤如下:

首先，使用 TF-IDF 算法，找出每个留言信息描述的 5 个关键词;

然后，对每个留言提取的 5 个关键词，合并成一个集合，计算每个岗位描述对于这个集合中词的词频，如果没有则记为 0;

最后，生成各个留言的 TF-IDF 权重向量，计算公式如下:

$$\text{TF- IDF=词频(TF) X 逆文档频率(IDF)}$$

得到评价数据 9210 条后，全部评论中的所有词语数和词语对(相邻两个单词的组合)的总数是 56783。

部分 TF-IDF 的特征值结果如下：

(1,40793)　0.3107737611568129

(1,9310)　　0.310773761 1568129

(1,22149)　0.310773761 1568129

(1,20730)　0.3107737611568129

(1,7046)　0.l 409376617026666

(1,22000)　0.21649752181 061938

(1,40791)　0.26950978760106314

(1,9307)　0.2879426209905127

(1,22144)　0.244406 13452037493

(1,20729)　0.3107 737611568129

(1,23690)　0.20474922957592534

(1,52506)　0.349683346382 76066

(2,41 194)　0.3346558638327778

(2,15162)　0.3496833463827 6066

（3）实现分类

(a)构建朴素贝叶斯分类模型

完成词空间向量的构建,我们使用朴素贝叶斯分类模型对留言文本进行分类朴素贝叶斯分类器最适合用于基于词频的高维数据分类器，本文使用的 sklearn 的朴素贝叶斯分类器 MultinomialNB

朴素贝叶斯分类器是一系列以假设特征之间强（朴素）独立下运用贝叶斯定理为基础的简单概率分类器。该分类器模型会给问题实例分配用特征值表示的类标签，类标签取自有限集合。

样本空间有个 m 类别 $\{C_1, C_2, \cdots C_m\}$，数据集有 n 个属性 $A_1, A_2, \cdots A_n$，给定未知样本 $X = (x_1, x_2, \cdots x_n)$，其中 x_i 表示第 i 个属性的取值，即 $x_i \in A_i$，则可用贝叶斯公式计算样本 $X = (x_1, x_2, \cdots x_n)$ 属于类别的 $C_k (1 \le k \le m)$ 概率。

由贝叶斯公式，有 $p(C_k/X)=\frac{p(C_k)p(X/C_K)}{p(X)}\propto p(C_k)p(X/C_k)$，即得到的 $p(C_k/X)$ 值，把未知类别样本 X 指派给类别 C_i，再由朴素贝叶斯分类器的属性独立性假设，假设各属性 $x_i(1,2,\cdots,n)$ 间相互类条件独立，则：

$$p(X/C_k)=\prod_{k=1}^{n}p(x_k/C_i)$$

$p(C_i)$ 为先验概率，可通过 $p(C_i)=d_i/d$ 计算得到，其中 d_i 是属于类别 C_i 的训练样本的个数；d 是训练样本总数。

（b）训练分类模型

预测自定义文本分类训练，同时测试朴素贝叶斯分类模型。这里，我们自定义文本：反映 A 市城市垃圾处理不及时的问题。得到预测结果如下，预测正确。

预测结果：

预测:反映 A 市城市垃圾处理不及时的问题

{0:’城乡建设，1:’环境保护’，2:’交通运输’，3:’教育文体’，4:’劳动和社会保障’5:’商贸旅游’，6:‘卫生计生’}

预测一级标签为: 卫生计生

model_ name RandomFores tClassifier

accuracies [0.34652928 0.36008677 0.3559414 0.36175991 0.36126224]

model_ _name LinearSVC

accuracies[0. 78850325 0. 82104121 0. 79978296 0. 84030418 0. 79107726]

model_ name Multinomia1NB

accuracies [0. 63069414 0. 6664859 0. 64243082 0. 65181966 0. 6218716]

nodel name Logi sti cRegression

（4）模型评价

本文利用 F1 分数评估模型，公式如下：

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

其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率，得到评估结果如下表:

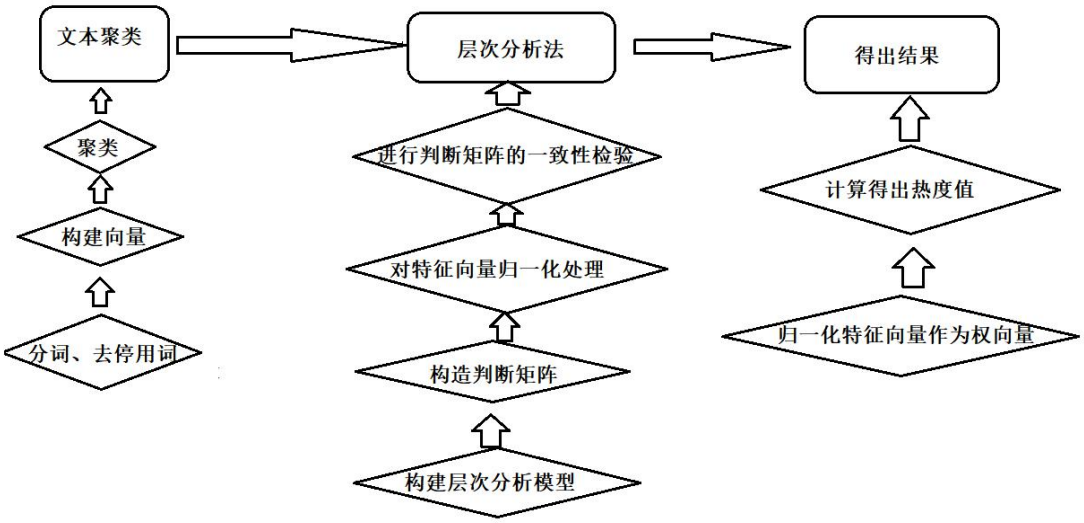
评估结果表

	精确率	召回率	F_1 指标	数据个数

城乡建设	0.7	0.89	0.78	1206
环境保护	0.88	0.78	0.82	563
交通运输	0.93	0.72	0.81	368
教育文体	0.88	0.86	0.87	953
劳动和社会保障	0.84	0.9	0.87	1181
商贸旅游	0.86	0.71	0.77	729
卫生计生	0.9	0.74	0.81	526

2.2.问题二分析过程与方法

（1）流程图



（2）文本分类

与问题一相类似，我们首先对附件 3 中的留言文本进行预处理。

（a）文本分词、去停用词

分词和去停用词等文本数据预处理是构建向量和后续聚类的基础，对聚类的效果有直接的影响。关于分词和去停用词的原理，在问题一中已经有详细的介绍，这里不再赘述。

选取附件 3 中“留言主题”一栏的文本数据，利用 jieba 进行文本分词。

经过分词，初始的文本处理成为词的集合，接下来，采用停用词词典的文本停用词过滤方式，进行去停用词的处理。

（c）构建空间向量 权重策略

采用 TF-IDF 算法，把留言信息转换为权重向量。以数学空间的表达代替语言文本的表达，向量间的相似度。具体的原理和过程已经在问题一中说明，此处不再重复。

（3）实现分类

(a)构建朴素贝叶斯分类模型

样本空间有个 m 类别 $\{C_1, C_2, \cdots C_m\}$，数据集有 n 个属性 $A_1, A_2, \cdots A_n$，给定未知样本 $X = (x_1, x_2, \cdots x_n)$，其中 x_i 表示第 i 个属性的取值，即 $x_i \in A_i$，则可用贝叶斯公式计算样本 $X = (x_1, x_2, \cdots x_n)$ 属于类别的 $C_k (1 \le k \le m)$ 概率。

由贝叶斯公式，有 $p(C_k/X) = \dfrac{p(C_k)p(X/C_K)}{p(X)} \propto p(C_k)p(X/C_k)$，即得到的 $p(C_k/X)$ 值，把未知类别样本 X 指派给类别 C_i，再由朴素贝叶斯分类器的属性独立性假设，假设各属性 $x_i(1, 2, \cdots, n)$ 间相互类条件独立，则：

$$p(X/C_k) = \prod_{k=1}^{n} p(x_k/C_i)$$

$p(C_i)$ 为先验概率，可通过 $p(C_i) = d_i/d$ 计算得到，其中 d_i 是属于类别 C_i 的训练样本的个数；d 是训练样本总数。

(b)训练分类模型

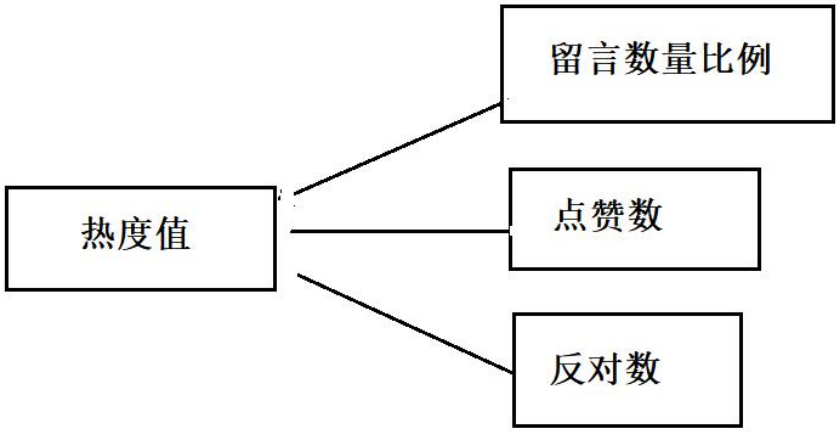
训练朴素贝叶斯分类模型，得到分类结果。

主要内容提要
A 市 A5 区魅力之城小区 小区临街餐饮店油烟噪音扰民
A 市经济学院体育学院变相强制实习
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
A 市金毛湾配套入学的问题
A 市 A4 区 58 车贷案
A 市 58 车贷特大集资诈骗案保护伞
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到

关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉
建议加大 A7 县东六线榔梨段拆迁力度

（3）构建层次分析模型

第一步：对于本题目，我们构建层次分析模型，选取热度值为目标层，留言数量比例（p）点赞数，反对数，准则层。层次分析模型如图：



第二步：进行构造判断矩阵，比较它们对目标的影响程度，确定在该层中相对于某一准则所占的比重。比较时取 1~9 尺度。用 a_{ij} 表示第 i 个因素相对于第 j 个因素的比较结果，所以：

$a_{ij}=1/a_{ji}$

$A=(a_{ij})_{n*n}=$

建立准则层到目标层的判断矩阵

利用 MATLAB 求判断矩阵的特征向量、特征值以及最大特征值

```
>> a=[1 1/2 1/3 1/2;2 1 1/2 1;3 2 1 2;]

a =

    1.0000    0.5000    0.3333    0.5000
    2.0000    1.0000    0.5000    1.0000
    3.0000    2.0000    1.0000    2.0000
```

```
>> [x,y]=eig(a)

x =

    0.2243 + 0.0000i   -0.0873 - 0.2458i   -0.0873 + 0.2458i    0.0000 + 0.0000i
    0.4163 + 0.0000i   -0.1562 + 0.2293i   -0.1562 - 0.2293i   -0.7071 + 0.0000i
    0.7766 + 0.0000i    0.8821 + 0.0000i    0.8821 + 0.0000i   -0.0000 + 0.0000i

y =

    4.0104 + 0.0000i    0.0000 + 0.0000i    0.0000 + 0.0000i    0.0000 + 0.0000i
    0.0000 + 0.0000i   -0.0052 + 0.2038i    0.0000 + 0.0000i    0.0000 + 0.0000i
    0.0000 + 0.0000i    0.0000 + 0.0000i   -0.0052 - 0.2038i    0.0000 + 0.0000i

>> diag(y)

ans =

    4.0104 + 0.0000i
   -0.0052 + 0.2038i
   -0.0052 - 0.2038i

>> max(diag(y))

ans =

    4.0104
```

第三步：对特征向量归一化处理：

```
x=x/norm(x)

x =

    0.1123 + 0.0000i   -0.0593 - 0.1672i   -0.0593 + 0.1672i    0.0000 + 0.0000i
    0.2270 + 0.0000i   -0.1062 + 0.1559i   -0.1062 - 0.1559i   -0.4809 + 0.0000i
    0.4236+ 0.0000i    0.5998 + 0.0000i    0.5998 + 0.0000i   -0.0000 + 0.0000i
```

所以最终 A 的最大特征值 λ=4.0104

对应的归一化后的特征向量为 ω=（0.123，0.2270，0.4236，）

第四步：进行判断矩阵的一致性检验

表 1: 随机一致性指标表

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

定义一致性指标：（λ是判断矩阵的最大特征根（该题中λ=4.0104），n=3）CI 越大，不一致越严重，引起的判断误差越大。定义随即一致性指标 RI。随即模拟得到 a_{ij} 形成 A，计算 CI 即得 RI。(其中 CI=（λ−n）/(n-1)=0.0035)

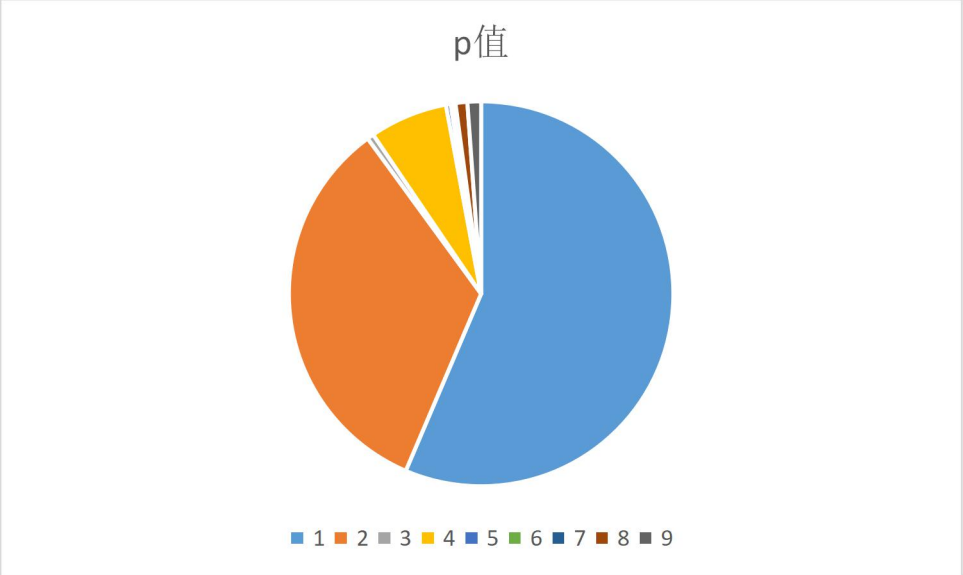
定义一致性比率：当 CR=CI/RI<0.1 时，能通过一致性检验，认为 A 的不一致程度在容许范围之内，可用其可用其归一化特征向量作为权向量，因此，进行一致性检验：CR=CI/RI=0.0035/0.9=0.0039<0.1,因此一致性检验通过，此时的ω可作为权向量带入。

（4）计算得出热度值

对附件 3 中的留言文本进行通过数量比例、点赞数、反对数、时间跨度几个方面进行等级排序，利用计算得到的权向量，计算出各个留言的热度值，最后提取出热点问题。

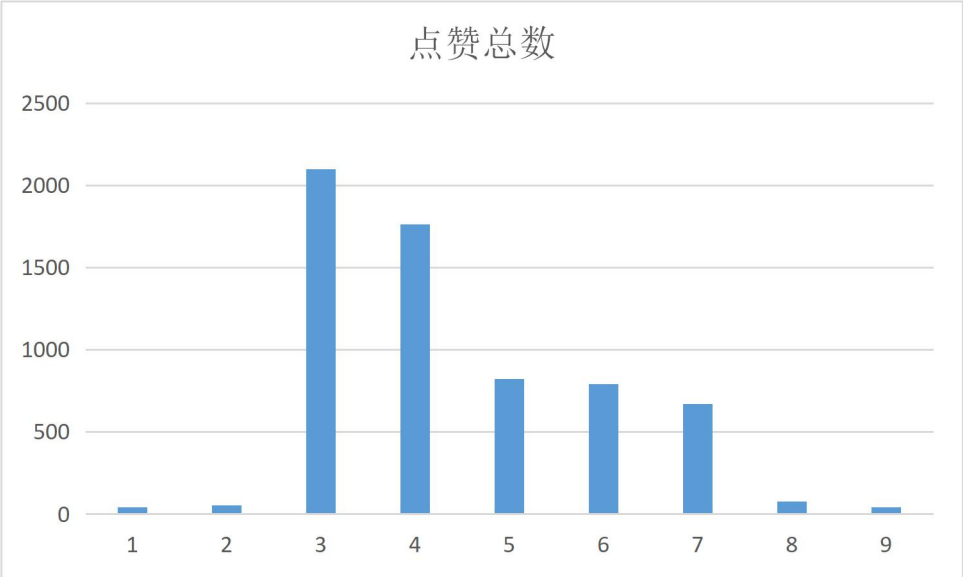
p 值评分表

主要内容提要	p 值	评分
A 市 A5 区魅力之城小区 小区临街餐饮店油烟噪音扰民	0.074	95
A 市经济学院体育学院变相强制实习	0.044	75
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	0.00073	16
A 市金毛湾配套入学的问题	0.00086	19
A 市 A4 区 58 车贷案	0.00052	15
A 市 58 车贷特大集资诈骗案保护伞	0.00031	13
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到	0.00022	10
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	0.0013	9
建议加大 A7 县东六线榔梨段拆迁力度	0.0015	9



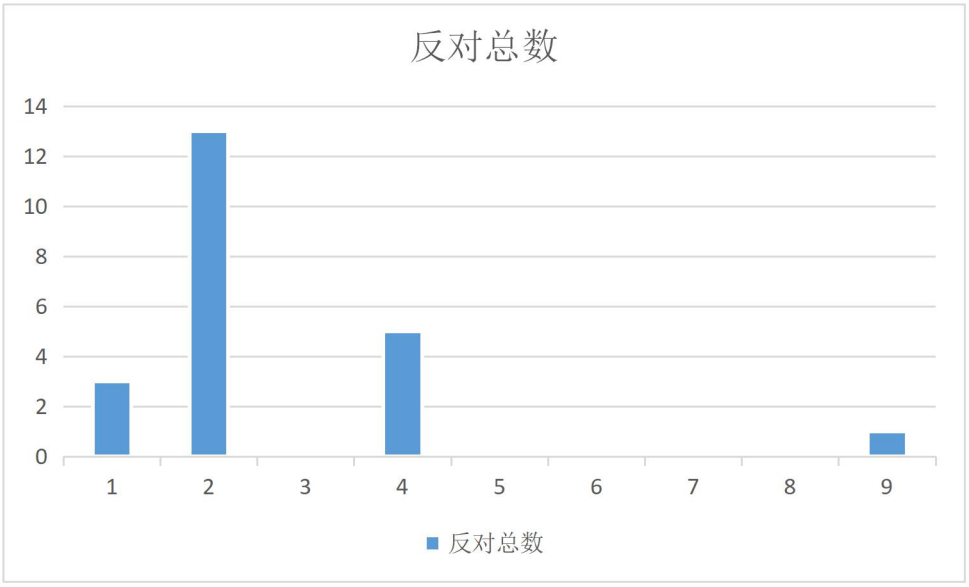
点赞数评分表

主要内容提要	点赞总数	评分
A 市 A5 区魅力之城小区 小区临街餐饮店油烟噪音扰民	43	4
A 市经济学院体育学院变相强制实习	52	4
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2097	100
A 市金毛湾配套入学的问题	1762	90
A 市 A4 区 58 车贷案	821	65
A 市 58 车贷特大集资诈骗案保护伞	790	63
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到	669	60
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	78	5
建议加大 A7 县东六线榔梨段拆迁力度	42	4



反对数评分表

主要内容提要	反对总数	评分
A 市 A5 区魅力之城小区 小区临街餐饮店油烟噪音扰民	3	40
A 市经济学院体育学院变相强制实习	13	100
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	0	0
A 市金毛湾配套入学的问题	5	50
A 市 A4 区 58 车贷案	0	0
A 市 58 车贷特大集资诈骗案保护伞	0	0
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到	0	0
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	0	0
建议加大 A7 县东六线榔梨段拆迁力度	1	20



热度评价汇总表（热度值降序）

主要内容提要	p 值 评分	点赞 评分	反对 评分	热度值
A 市 A5 区魅力之城小区 小区临街餐饮店油烟噪音扰民	95	4	40	95.31194
A 市经济学院体育学院变相强制实习	75	4	100	87.84917
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	16	100	0	83.48672
A 市金毛湾配套入学的问题	19	90	50	69.29204
A 市 A4 区 58 车贷案	15	65	0	58.38298
A 市 58 车贷特大集资诈骗案保护伞	13	63	0	56.16019
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到	10	60	0	54.50347
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	9	16	0	44.59563
建议加大 A7 县东六线榔梨段拆迁力度	9	7	0	41.73827

2.3.问题三分析过程与方法

（1）评定指标的分析

（a）相关性：

相关性指的是留言的答复意见，与原留言文本的相似度。

推荐 BM25 算法计算相关性。BM25 是二元独立模型的扩展，其得分函数有很多形式，最普通的形式如下：

$$\sum \frac{(ri+0.5)/(R-ri+0.5)}{(ni-ri+0.5)/(N-ni-R+ri+0.5)} \cdot \frac{(k1+1)fi}{K+fi} \cdot \frac{(k2+1)qfi}{k2+qfi}$$

其中，k1,k2,K 均为经验设置的参数，fi 是词项在文档中的频率，qfi 是词项在查询中的频率。K1 通常为 1.2，通常为 0-1000。

K 的形式较为复杂

$$\sum \frac{(N-R+0.5)}{(R+0.5)} \cdot \frac{(k1+1)fi}{K+fi}$$

上式中，dl 表示文档的长度，avdl 表示文档的平均长度，b 通常取 0.75

BM25 具体实现：由于在典型的情况下，没有相关信息，即 r 和 R 都是 0，而通常的查询中，不会有某个词项出现的次数大于 1。因此打分的公式 score 变为

$$\sum \frac{(N-R+0.5)}{(R+0.5)} \cdot \frac{(k1+1)fi}{K+fi}$$

（b）完整性：

完整性指的是留言回复意见能否完整的回答留言中提出的问题。提取出留言与答复意见的关键字，找到两者相同的关键词后，计算两者相同关键词的个数占留言详情的关键词的个数的比例。

（c）可解释性：

可解释性指的是答复问题提出的方案，对于问题的解决效果如何。包括答复意见语言的可读性、处理方案的可行性、以及解决问题效果的持久性。

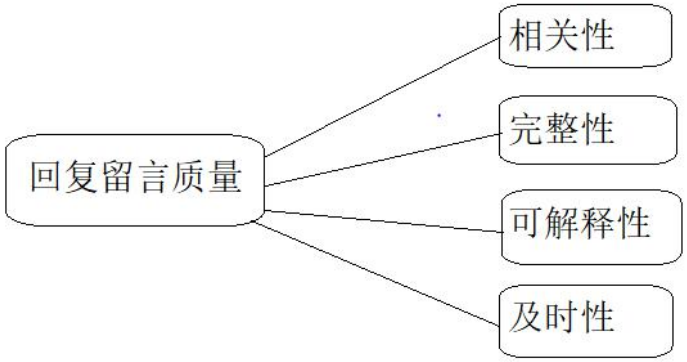
（d）及时性：

及时性指的是对于留言问题的回答是否及时。我们的时间和答复时间的跨度来判断答复意见的及时性。

（2）建立回复质量评价模型

（a）第一步：

基于层次分析法，建立评价模型。留言回复质量指数为目标层，将相关性、完整性、可解释性和及时性设为第二层准则层，层次分析法图示如下：



（b）第二步：

进行构造判断矩阵，比较每一项指标对留言回复质量的影响程度，确定在该层中相对于某一准则所占的比重。比较时取 1~9 尺度。用 a_{ij} 表示第 i 个因素相对于第 j 个因素的比较结果，

（c）第三步：

进行判断矩阵的一致性检验：

表 1: 随机一致性指标表

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

定义一致性指标：（λ是判断矩阵的最大特征根（该题中λ=4.0104），n=4）CI 越大，不一致越严重，引起的判断误差越大。定义随即一致性指标 RI。随即模拟得到 a_{ij} 形成 A，计算 CI 即得 RI。(其中 CI=（λ-n）/(n-1)=0.0035)

定义一致性比率：当 CR=CI/RI<0.1 时，能通过一致性检验，认为 A 的不一致程度在容许范围之内，可用其可用其归一化特征向量作为权向量，因此，进行一致性检验：CR=CI/RI=0.0035/0.9=0.0039<0.1,因此一致性检验通过，此时的ω可作为权向量带入。

（e）第四步：

利用计算得到的权向量，计算附件四中的留言答复意见的质量指数，完成质量评价。

3. 结论

通过本次数据挖掘，在解决问题的经验上，我们认识到：在进行文本数据分析时，应选用适当的分类模型，朴素贝叶斯分类模型具有原理简单、精确度高的优点。而建立评价标准时，采用层次分析模型，可以得到全面而客观的结果。

4. 参考文献

1.孙海峰 郑中枢 杨武岳《网络招聘信息的数据挖掘与分析》2017
2.周涛 吴家舜 《基于电商平台家电设备的消费者的消费者评论数据挖掘分析》2016
3.杨震 段立娟 赖英旭 《基于字符串相似性聚类的网络短文本舆情热点发技术》2010
4.赵琳瑛.基于隐马尔科夫模型的中文命名实体识别研究.西安电子科技大
5.朱志远.基于数据挖掘的网络招聘系统是设计与实现.电子科技大学.硕士学位论文.2013