基于贝叶斯分类的“智慧政务”文本挖掘应用

摘要

心系民生，倾听民声，对提高社会生活质量、推动社会发展是及其重要的。近年来，各类社情民意的文本数量不断增多，这使得部门及时了解民意成为了当前紧要问题。考虑到针对不同方面的留言数据量庞大，对处理、分类留言的工作人员而言，这无非是一项巨大而耗时的工作；鉴于此，本文将针对留言分类、热点问题的挖掘等不同方面进行建模解析、提出解决方案。提高处理问题的效率。

针对问题一，本题使用朴素贝叶斯算法训练分类器。在大量留言文本数据的背景下，对留言中文文本数据首先进行数据预处理，使用$jieba$对留言文本进行分词处理，再使用 TF-IDF 的统计方法对某个词条在该类型中的重要程度，最后使用朴素贝叶斯算法训练分类器。统计种类型问题类型中的重要词频，可以更有效的对留言问题类型进行分类。

针对问题二，本题采用LDA方法进行建模，结合$ESIM$实现文本相似度检测；最后结果是以热点问题排名模式展现，提出一种基于词频数排名的热点留言问题的挖掘方法，加以$K-means$算法进行聚类、聚簇关注度等方面进行热点问题的聚焦；基于该问题，建立好向量空间模型后采用$SinglePass$算法聚类挖掘出热点问题的排名。

针对问题三，先从答复意见的相关性出发，利用留言用户的留言详情和留言的答复意见计算其相似性，相似性求出来答复意见的相关性就可解决。根据答复意见的准确性和可靠性说明其完整性。答复意见与用户的留言内容相关，再根据答复意见内容，有理有据，说明其可解释性。

关键词：$jieba$　$K-means$算法　$SinglePass$算法聚类　文本相似度检测

The text mining application of "intelligent government" ba sed on bayesian classification

Abstract

It is very important to pay attention to people's livelihood and listen to their voi ces in order to improve the quality of life and promote social development.In recent years, t he number of texts on various social situations and public opinions has been constantly incr easing, which makes it an urgent issue for departments to timely understand public opinion s.Considering the huge amount of message data for different aspects, it is nothing more tha n a huge and timeconsuming work for the staff to process and classify the message;In view of this, this paper will focus on message classification, hot issues mining and other different aspects of modeling analysis, proposed solutions.Improve the efficiency of problem solving.

Aiming at problem one, this paper USES the naive bayes algorithm to train the classifi er.Under the background of a large number of message text data, we firstly preprocessed th e message Chinese text data, then used jieba to process the message text word segmentati on, then used tfidf statistical method to the importance of an entry in this type, and finally u sed naive bayes algorithm to train the classifier.The statistics of the important word frequenc y in the question types can classify the message question types more effectively.

For problem two, LDA method is adopted for modeling, and text similarity detection i s realized in combination with ESIM.The final result is presented by the ranking model of hot issues. A mining method of hot topic comment based on word frequency ranking is propos ed, and kmeans algorithm is used to focus hot issues in clustering and clustering attention.B ased on this problem, the SinglePass algorithm was used to mine the ranking of hot issues a fter the vector space model was established.

Aiming at question three, starting from the relevance of reply comments, the similarit y is calculated by using the message details of the message users and the reply comments.T he completeness of the replies is based on their accuracy and reliability.The reply is related t o the message content of the user, and then according to the reply content, reasonable and according to explain its interpretability.

Key words: jieba kmeans algorithm SinglePass algorithm clustering text similarity detection.

目录

摘要.....1
1 问题重述.....4
1.1 问题背景.....4
1.2 需解决的问题.....4
2 问题分析.....4
2.1 问题一的分析.....4
2.2 问题二的分析.....5
2.3 问题三的分析.....5
3 分析总结.....5
4 模型的建立与求解.....5
4.1 问题一：建立关于留言内容的一级标签分类模型.....5
4.2 问题二：热点问题挖掘.....9
4.3 问题三：答复意见的评价.....13
参考文献.....19

1 问题重述

1.1 问题背景

“数据化”时代的发展，“数据”已经不仅仅是局限于单纯的字数字符，而且还包括了“文本数据”“信息数据”。所以现如今处理“文本数据”成为了帮助人们提高工作效率的一个不错的帮手。正如为了解“民意”，相关部门需去处理通过各种途径写下的留言，这样才有助于去处理在生活中遇到的问题。但由于留言数量的庞大，人工处理已经力所不能及。这就需要运用“大数据”里的文本挖掘技术来将各类留言进行分类，将针对同一问题的留言归一化，这样在解决留言问题时，就大大提高了效率。不同类别的问题得到了分类，便分放到处理该类问题的部门去进行处理，这将给文本处理技术带来了广大的应用前景。

正是这样的情景极力需要文本挖掘应用，充分运用了多个技术来实现留言的分类，以及热点问题的挖掘。

1.2 需解决的问题

1. 群众的留言分类，针对于不同地点、不同人群会有不同的问题。

2. 热点问题的挖掘，对同一问题，提出比较多的，也许就是大家更关注的问题。

3. 答复问题的评价，对问题的回复，尽可能达到完整，此问是给出一个评价方案。

2 问题分析

2.1 问题一的分析

对分词后的文本数据进行重要词频的统计，再训练朴素贝叶斯分类器。

问题一旨在对分词后的词条进行词频的统计，再训练分类器。充分的利用重要词频来对作为每种留言类型的特征。但在数据预处理阶段需要对中文文本数据进行更详细的处理，以便在统计词频阶段能够更准确的统计出具有代表性的词条。

2.2 问题二的分析

对于热点留言问题的挖掘，其实质在于将相似度高的留言用文本挖掘技术提取，按照针对同一问题的留言数量进行排序，得出热点问题的排序，筛选出前 5 个热点问题排名；最后再将该类热点留言问题的明细根据聚类的结果来展开

2.3 问题三的分析

问题 3 要求根据附件 4 中相关部门的答复意见，我们需要解决三个方面的问题:是答复意见的相关性;二是其完整性;第三是其可解释性;然后通过一组方案来评估回复的质量。

3 分析总结

在基于传统的贝叶斯聚类算法基础上，结合 ESIM 实现文本相似度检测，运用 jieba 将评论主题中的主要词频提取，这样即可清晰的分析出主要的留言是更倾向于哪一地点、哪一类问题。这就对解决问题起到了引导性的作用，而且贝叶斯分类模型是基于贝叶斯定理的分类算法，它是适合于高维属性的分类模型的；将问题分类后需做出一个较为完整的答复，这将结合答复的准确性和可靠性来进行说明

4 模型的建立与求解

4.1 问题一：建立关于留言内容的一级标签分类模型

4.1.1 流程图

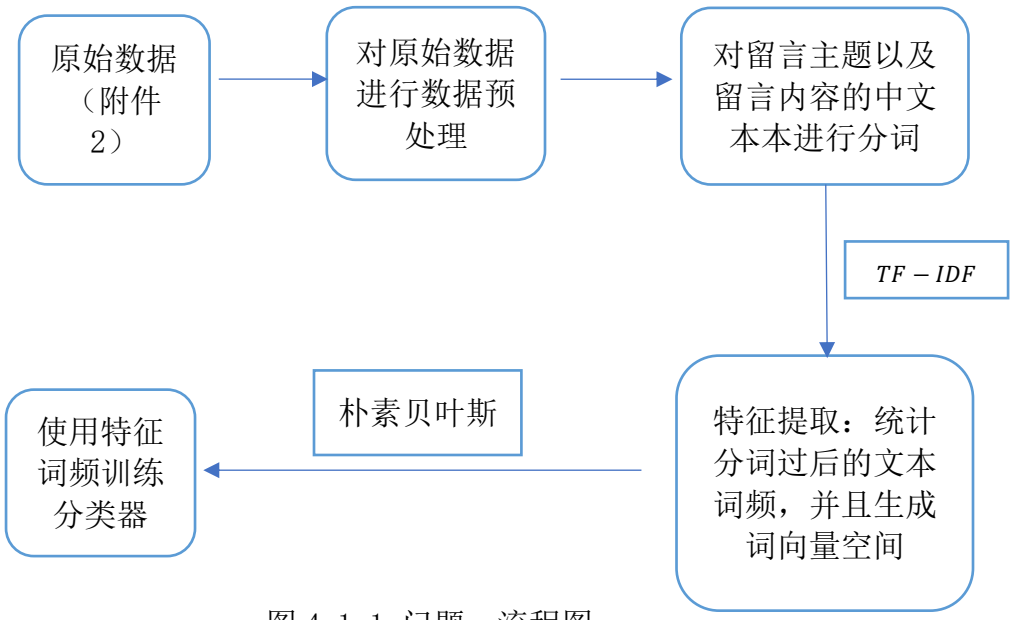

图 4.1.1 问题一流程图

4.1.2 数据预处理

首先对附件二中的留言主题以及留言内容这两个中文文本数据进行数据的预处理。因为在留言数据中仍存在部分没有价值的信息，如果将这些没有价值的信息计算入词频的话，会对分析结果过造成影响。文本数据的预处理有两步：第一步数据清洗，例如文本数据缺失值的处理，第二部分为机械压缩语料。

4.1.3 数据清洗

这里对文本数据缺失值的处理一般采用人工手段进行处理。

4.1.4 机械压缩语料

通过文本对重对数据预处理是不够的，还有一部分文本需要进行处理。比如在日常表达中常见的重复语料，但是对文本的分析并没有多大的价值，需要对这部分重复语料进行删除。

表 4.1.4 压缩语料

留言详情	进行机械语料压缩后的留言详情
尊敬的张市长：A 市的市政道路维护质量与水平实在是太太太差劲了，经常是上午补下午烂...	尊敬的张市长：A 市的市政道路维护质量与水平实在是太差劲了，经常是上午补下午烂.....

4.1.5 使用*jieba*对留言文本数据进行分词

因为留言信息都为中文文本，需要对文本进行分词处理。中文分词就是将一个句子或是段落分成一个一个的词语，这里使用*jieba*进行中文分词，它是用*python*专门开发的中文分词系统。使用*jieba*分词系统，要对*jieba*分词包进行安装，直接使用*pip*来进行安装：*sudo pip install jieba*。

表 4.1.5 分词示例表

分词前的文本	分词后的文本
A 市西湖建筑集团占道施工有安全隐患	A/市/西湖/建筑/集团/占/道/施工/有/安全/隐患

4.1.6 使用$TF-IDF$算法提取关键词

$TF-IDF$是一种统计方法，用来估计一个词在一个语料库中的重要程度。这里采用$TF-IDF$对每种类型的留言进行关键字的提取。在上一步对留言文本进行分词后，需要将这些分词之后得到的词语转化为向量，供挖掘分析使用。

TF代表词条在语料库中的出现频率。如果包含词条t的留言越少，则n越小，IDF越大，说明词条t具有很好的类别区分能力。

如果某一类问题类型C中包含词条t的留言数为m，而其它问题类型中包含t的留言总数为k，显然所有包含t的留言数$n=m+k$，当m大的时候，n也会大，IDF的值会小，就说明该词条t类别区分能力不强。

如果一个词条t在一个问题类型的留言中出现得很频繁的时候，那么就说明该词条t充分的代表了这中问题类型的文本特征，需要给词条t赋予比较高的权重，并用词条t作为区分该类别与其他类别的关键词。

$TF-IDF$算法的原理如下：

1.TF(词频) = 某个词条 t 在文本中出现的次数(1)

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

2.IDF(逆文档频率) = log[留言文本数据总数/(包含某词条t的留言数)+1]

$$idf_i=lg\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{2}$$

（分母一般情况下需要加一）

3.$TF-IDF$ = TF(词频) * IDF(逆文档频率)

$$tfidf_{i,j}=tf_{i,j}*idf_{i,j} \tag{3}$$

计算每个词条t的$TF-IDF$，然后进行排序，次数最多的则选取作为描述此种问题类别的关键词。

4.1.7 训练朴素贝叶斯分类器：$MultinomialNB$

$MultinomialNB$分类器的特征值是采用出现的次数，$TF-IDF$符合这类分布。朴素贝叶斯分类器是一种概率分类器。设现有的留言文本数据$D=\{d1，……dm\}$，给定一个留言文本数据，留言属于哪一类问题类型呢？用公式表示为：

$$c^{\wedge}=argmaxP(c|d) \quad c\in C \tag{4}$$

表 3 公式(4)符号说明

符号	说明
$c^{\wedge}$	问题类别D中，条件概率$P(c\|d)$得到最大值时候的问题类别。

$c^{\wedge}$为在所有问题类别$D=\{d1，……dm\}$中，条件概率$P(c|d)$得到最大值时候的问题类别。则上述公式可以转化为：

$$c^{\wedge}=argmaxP(c|d) \quad c\in C \tag{5}$$

$$c^{\wedge}=argmax\frac{P(d|c)P(c)}{P(d)} \quad c\in C \tag{6}$$

类别C中的每个问题类型，可以计算 $[p(d|c)*p(c)]/p(d)$ 的值，选取最大值对应的问题类型C_i，那么C_i就是最优解$c^{\wedge}$，所以在这里可以忽略掉分母 $p(d)$，

得到

$$c^{\wedge} = argmaxP(c|d) \quad c\in C \tag{7}$$

$$c^{\wedge} = argmaxP(d|c)P(c) \quad c\in C \tag{8}$$

$p(d|c)$ 为似然函数，$p(c)$ 为先验概率。训练朴素贝叶斯的过程就是计算先验概率和似然函数的过程。设留言文本数据中一共有N条留言，可以数出有多少条留言属于某个类别c，然后可以得出先验概率的公式如下：

$$P^{\wedge} = \frac{N_c}{N_{doc}} \tag{9}$$

这里用词袋模型表示某种问题的留言d，关于d中的每一个词条w_i，找到训练数据集中所有类别为c的留言，数一数词条w_i在为c中出现的次数：$count(w_i, c)$，再计算训练数据集中问题类别为c的留言一共有多少个词条。就可以得到似然函数。似然函数的公式为：

$$P^{\wedge}(w_i|c) = \frac{count(w_i,c)}{\sum_{w\in V} count(w,c)} \tag{10}$$

4.2 问题二：热点问题挖掘

致力于热点问题的挖掘，不仅有助于部门进行有针对性的处理，而且在处理特定地点特定人群问题时大大的提高了效率。使得部门对居民的服务效率提升，对城市整体的改观也起到了一定的作用。

该题针对热点问题的挖掘会考虑到：不同地点、不同人群的留言并不是采用统一的留言，或者对同一个地方的地名也会出现不同的描述这给建模型带来了难点；所以针对此难点文本采用LDA方法进行建模，对留言问题进行筛选、清洗；筛选出相似度较高且按照词频条数进行。结合$ESIM$实现文本相似度检测；最后结果是以热点问题排名模式展现，这将提出一种基于词频数排名的热点留言问题的挖掘方法，加以$K-means$算法进行聚类、聚簇关注度等方面进行热点问题的聚焦；基于此，建立好向量空间模型后采用$SinglePass$算法聚类挖掘出热点问题的排名

热点问题的挖掘将会涉及到数据清洗、识别相似留言、相似问题归并、给出热度评价指标的模型定义和计算方法等的综合因素；大致可分为 4 个步骤解决问题，下文将首先对数据进行清洗分析后，再做进一步的处理。

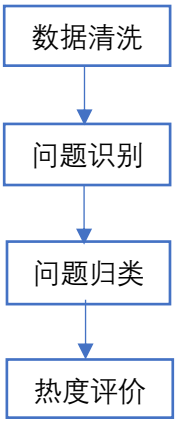

图 4.2 第二问步骤流程图

4.2.1 文本相似度检测

1.为了能够将热点留言从众多文本中提取出，我们将从热点问题的共同点出发，热点问题的本质是针对于同一问题，很多居民来提出，说明这是一个大家认为需亟待解决的问题，从而导致留言中针对该类问题的描述会大量存在，当然，在描述中会有些许的字词不同，但大致意思还是指向同一个问题；基于此，这里采用文本相似度检测的方法结合$ESIM$方法来筛出针对同一问题你的文本、计数，便可得出热点问题的排名。

2.模型的建立

将众多输入文本采用$BiLSTM$转化成不同的输入变量，即不同的文本一一对应不同的变量：

$$\bar{m}_i = BiLSTM(m,i) \tag{11}$$

$$\bar{m}_j = BiLSTM(m,j) \tag{12}$$

……

表 4.2.1.2 公式(11)符号说明

符号	说明
$\bar{m}_i$	变量i
$\bar{m}_j$	变量j
$i\ j$	序号

按照如此将所有的留言变量化，这样对下一步分析句子之间的关联性有了数据化的基础，将留言的数据化就相当于得到了数据的向量化。接下来可以通过计算得出两两向量之间的距离和夹角（这可根据留言的语境进行综合分析），然后比较夹角值，便可得出不同留言之间的相似度，相似越大，即可归为一类；这里的夹角值可转化为词向量间相乘，若相乘结果大，可发现它们之间的相关性也增大。两两之间成正相关关系，

$$e_{ij} = \bar{m}_i^{\,T} \bar{m}_J \tag{13}$$

$$\widetilde{m_i} = \sum_{j=1}^{l_m} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_b} \exp(e_{ik})} \overline{m_J} \tag{14}$$

$$\widetilde{m_j} = \sum_{j=1}^{l_m} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_b} \exp(e_{ik})} \overline{m_i} \tag{15}$$

表 5 公式(13) (14)符号说明

符号	说明
e_{ij}	$\bar{m}_J$与$\bar{m}_i^{\,T}$的乘积
$\exp(e_{ij})$	对e_{ij}求指数
$\bar{m}_j$	m_j向量

不同向量间分析内容的相似性来建立向量空间模型，上述计算出来的结果可得出权重，继而计算其他各个向量的权重，进行对比分析；将相关性存储于一个序列中，在序列中分析差异，就能看出留言联系大小。

$$p_a = [\bar{m}\,\tilde{m}\,(\bar{m} - \tilde{m})\,(\bar{m} \odot \tilde{m})] \tag{16}$$

$$p_a = [\bar{n}\,\tilde{n}\,(\bar{n} - \tilde{n})\,(\bar{n} \odot \tilde{n})] \tag{17}$$

表 6 公式(16)符号说明

符号	说明
p_a	m留言的存储序列
$(\bar{m} \odot \tilde{m})$	$\bar{m}$与$\tilde{m}$同或

这样可得出来序列中的文本分类。

4.2.2 聚类化已分类的文本

1. $K-means$层次聚类算法思想

基于上文的文本分类，现在进行分类文本的聚类。该类算法最大的特点是在不同层次对数据集进行划分，将数据集中的每个样本看作一个初始聚类簇，接着找出最近的簇进行合并，并不断重复，直到达到某种条件；每个簇都是一个集合，从而需要技术它们之间的距离。

2. $K-means$层次聚类模型建立

分别计算它们之间的min距离、max距离、$average$距离：

$$d_{min}(m_i, m_j) = min_{x\in m_i, z\in m_j} dist(x,z) \tag{18}$$

$$d_{max}(m_i, m_j) = min_{x\in m_i, z\in m_j} dist(x,z) \tag{19}$$

$$d_{ave}(m_i, m_j) = \frac{1}{|m_i||m_j|}\sum_{x\in m_i}\sum_{z\in m_j} dist(x,z) \tag{20}$$

最小距离由两个最近的簇样本来决定。现在序列中选定一部分测试分析；计算指标DBI和D的变化趋势，可得出一个距离阈值；将该阈值作为一个结束聚类的条件。

从结果可看出，DBI值呈下降趋势，DI呈上升趋势，基于此结果，可将k值另外设置值，会发现会有一些相同的样本被分到其他不同的类中。这样导致聚合结果的离差较大，不利于热点问题的挖掘。所以在理想情况下，尽可能将相似的留言在一个类中聚合。这里可给予k值适当的冗余（适当降低k值），然后根据上述公式再次计算；根据上述过程得出相似问题的分类。

3. 词云统计

从统计的词云可大致看出，在众多留言中，有关于“魅力小区的油烟严重”问题是最多的，次重要的就是学生强制实习等问题。这样就对下一步的统计聚合做了进一步的分析。词云效果图如下：

图 4.2.2 词云展示图

4.2.3 统计聚合留言文本的数量

1.热点问题统计

聚合了相似类文本，接下来就是统计文本，数量居多者，就会成为热点问题；然后一次按照排名来展现出排名表；

聚合的结果是将同类文本聚合到一起，在统计出热点问题后，可根据此得出热点留言问题明细表；

4.3 问题三：答复意见的评价

1．数据预处理

根据附件 4 所给的数据，找出异常数据并剔除。

表 4.3.1 剔除数据表

留言主题	答复意见
请问 B9 市带小孩打疫苗要带什么证件	2019 年 1 月 14 日
建议 B9 市规划一个校车接送计划	2018 年 12 月 12 日

根据表 4.3.1 显示，发现只给出“2019 年 1 月 14 日”这种答复意见，没有给出相关留言详情的回答，认为该种数据为异常数据。

2、名词解释说明

表 4.3.2 名词解释说明表

名词	解释说明
相关性	选取的标准必须与事项时相关的，相关的标准有助于得出结论，便于使用者作出决策

完整性	指精确性和可靠性
可解释性	在我们需要了解或解决一件事情的时候，我们可以获得我们所需要的足够的可以理解的信息

4.3.1 相关性分析

1．基本思路

通过向量之间的夹角来判断向量的相似度，角度越小，越相似。如果两个句子中的单词比较相似，那么它们的内容也应该比较相似，这样我们就可以计算这两个变量的相似度。

2．余弦相似性

余弦相似性是由向量空间中两个向量夹角的余弦值来度量的。其取值范围[-1,1]。在二维坐标上，设$a(x1,y1),b(x2,y2)$。

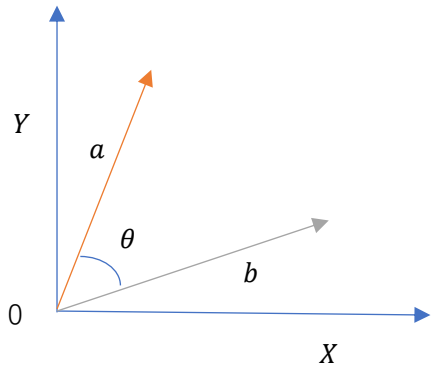

图 4.3.1 余弦角示意图

根据图 4.3.1，余弦相似度是通过计算$a(x1,y1),b(x2,y2)$两个向量夹角θ的余弦值来测量的，推导公式如下：

$$\cos(\theta)=\frac{a^2+b^2-c^2}{2ab} \tag{21}$$

$$\cos(\theta)=\frac{a\cdot b}{\|a\|\times\|b\|} \tag{22}$$

$$\cos(\theta)=\frac{(x1,y1)\cdot(x2,y2)}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} \tag{23}$$

$$\cos(\theta) = \frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} \tag{24}$$

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{25}$$

3．一般步骤

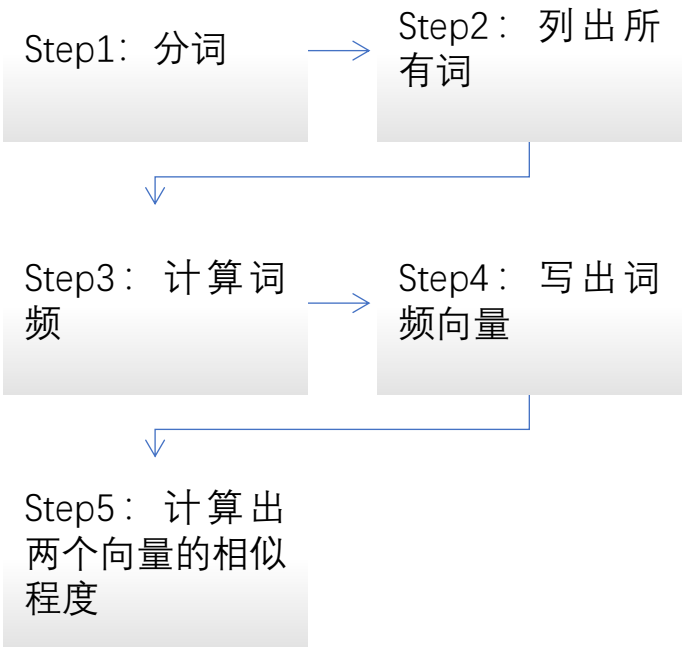

图 4.3.2 流程示意图

4．求解

编写相关代码，运行程序，得出留言详情和留言答复意见的相似度，数据部分显示如下：

表 4.3.3 留言详情和留言答复意见的相似度表

留言编号	留言详情和留言答复意见相似度
2549	0.639253
2554	0.453948
2555	0.609644
2557	0.352217
2574	0.611995
2759	0.180214
2849	0.453476
33970	0.386261
33978	0.461605
33984	0.204653

5．结果分析

根据表 4.3.3 结果显示，答复意见和留言用户所留言的问题的相关性还是普遍很高的，说明很多留言用户所反映的问题情况都得到了相关部门有效的答复。根据表 4.3.4 所示，可以知道用户的留言都会在相应的时间内得到答复，即使有些留言已经间隔时间比较长，但还是会得到答复。

表 4.3.4 留言时间和答复时间表

留言时间	答复时间
2019/4/25 9:32:09	2019/5/10 14:56:53
2019/4/24 16:03:40	2019/5/9 9:49:10
2019/4/24 15:40:04	2019/5/9 9:49:14
2019/4/24 15:07:30	2019/5/9 9:49:42
2019/4/23 17:03:19	2019/5/9 9:51:30
2019/4/8 8:37	2019/5/9 10:02:08
2019/3/29 11:53:23	2019/5/9 10:18:58
2018/8/12 10:56:10	2018/8/20 9:25:34
2018/8/8 13:15:50	2018/8/17 9:49:43
2018/8/3 21:26:53	2018/8/7 9:13:42
2018/2/3 16:27:45	2018/4/16 11:33:45
2018/1/25 12:01:13	2018/4/12 10:10:40
2018/1/24 11:12:25	2018/4/12 11:09:34
2019/12/2 19:34:28	2019/12/10 15:16:08
2019/11/28 15:41:22	2019/12/2 16:37:17
2019/11/19 9:32:36	2019/12/2 16:43:23
2019/11/5 21:51:41	2019/11/12 14:53:48
2018/12/5 23:49:06	2018/12/21 9:25:42
2018/12/4 10:09:51	2018/12/21 9:27:31

4.3.2 完整性

根据表 4.3.3 相关性的结果来分析，可以发现，用户的留言主题和留言详情与留言的答复意见的相关性还是普遍高的。这就说明，每一条留言都得到内容性的答复（即相关部门根据用户的留言主题和留言详情的内容进行相应的答复）。从而说明答复意见的数据还是有很高的可靠性的，从而也就说明答复意见的完整性还是很高的。

4.3.3 可解释性

针对可解释性，列出附件 4 的第一条数据来分析。

表 4.3.5 一条数据示例表

留言主题
A2 区景蓉华苑物业管理有问题
留言详情
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？
答复意见
现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2 区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于 2019 年 4 月 10 日至 4 月 27 日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反

馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5 月 5 日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。2019 年 5 月 9 日

表 4.3.5 为附件 4 一条数据部分内容，用户留言的主题是整个留言内容的关键部分，留言详情是整个留言内容的情况，答复意见给出了留言问题的相关内容答复。这一条数据的留言详情与留言答复意见的相关性为 0.639253，说明该条意见答复和留言主题的相关性强；答复意见针对用户所提到的问题进行逐一答复，说明答复意见的准确性和可靠性也是挺强的，进而说明答复意见是有一定的完整性的；答复意见和留言问题相关，答复的也有理有据，说明答复意见具有其可解释性。

参考文献

[1]刘福刚.一种适用于中文博客自动分类的贝叶斯算法[J].长春师范大学学报,2019,38(12):36-43.
[2]曾小芹.基于 Python 的中文结巴分词技术实现[J].信息与电脑(理论版),2019,31(18):38-39+42.
[3]范庆春.基于中文分词技术的文本相似度检测研究[J].池州学院学报,2019,33(03):19-20.
[4]阎令海.分析 Python 语言的中文文本处理[J].中国新通信,2019,21(02):147.
[5]齐丽花,张妮妮,秦晓梅.基于 K-means 的专利文本聚类分析[J].电脑知识与技术,2018,14(22):206-207+214.
[6]董婧灵. 基于 LDA 模型的文本聚类研究[C]. 中国中文信息学会.中国计算语言学研究前沿进展（2009-2011）.中国中文信息学会:中国中文信息学会,2011:463-469.
[7]任美睿. 基于朴素贝叶斯方法的自动文本分类系统的实现[C].