“智慧政务”中的文本挖掘

摘要

近年来，随着科技的进步以及互联网的广泛应用，网络问政平台逐步成为政府与群众建立联系的重要渠道，但是，随着文本数据量的增大，给相关部门对文本的整理带来了巨大挑战。因此，运用网络文本分析和数据挖掘技术对自然语言进行处理具有重大意义。

对于问题一，我们要建立关于留言内容的一级标签分类模型。首先要对给定的留言内容进行预处理，通过 *jieba* 中文分词工具进行分词和去停用词,其次， 利用 *Word*2*vec* 算法将留言主题以及三级标签词语转化为可识别向量，通过特征选择及特征提取对数据进行降维处理，接着，利用 $K-means$ 聚类与最近邻算法算出相似度，最后，根据一级标签与三级标签的对应关系，将留言内容分到对应一级标签下完成第一步。在对分类方法进行评价中，本文先计算各类的查全率与查准率，再利用公式求解最终的 F 值。

对于问题二，对于问题二，综合考虑影响热点问题的热度排序的主要因素有：某段时间内留言数、浏览关注度某地点问题的集中反应等，这些因素是来自留言所包含的词条内的基本属性，在一定程度上决定了问题的热度。基于以上分析，我们提出基于留言关注度（点赞数和反对数）和留言频数的热度评价指标。首先，进行文本内容预处理，选取特征词作为有效词，转化为词向量；接着，利用 $TF-IDF$ 对有效词加权，对词的权重进行调整，得到某一时段内反应特定地点的留言；最后，基于群众的留言量和关注度对问题集合提出合理的热度评价方法。

对于问题三，通过对现有资料的查找，分析，从多角度分析答复意见的质量，对多角度进行逐一分析，选出有价值的指标。建立基于给出指标得出质量评价方案。首先，对留言答复内容进行数据预处理，其次，根据答复内容是否贴近主题、答复内容是否完整，是否可解释的有理有据等角度对答复数据进行量化，最终得到一套评价方案。

关键词：分词　*Word*2*vec* 算法　$K-means$ 聚类　热点问题挖掘　评价指标　$TF-IDF$

Abstract

In recent years, with the progress of science and technology and the wide application of the Internet, the network political platform has gradually become an important channel for the government to establish contact with the masses. however, with the increase of the amount of text data, it has brought great challenges to the relevant departments in sorting out the text. Therefore, it is of great significance to use network text analysis and data mining technology to process natural language.

For question 1, we need to establish a first-level label classification model about the content of the message. First of all, we need to preprocess the content of the given message. We use jieba Chinese word Segmentation tool for word segmentation and de-deactivation of words. Secondly, the topic of the message and the three-level tag words are transformed into recognizable vectors by using the word2vec algorithm. Dimensionality reduction of data by feature selection and feature extraction. Then, the similarity is calculated by using K-Means clustering and nearest neighbor algorithm. Finally, according to the corresponding relationship between the first-level tag and the third-level tag, the content of the message is divided into the corresponding first-level tag to complete the first step. In the evaluation of the classification method, this paper first calculates the recall and precision of all kinds, and then uses the formula to solve the final F value.

For question 2, the main factors that affect the heat ranking of hot issues are: the number of messages in a certain period of time, the concentrated response of browsing to a certain location, and so on. These factors come from the basic attributes in the entries contained in the message, which determine the heat of the problem to a certain extent. Based on the above analysis, we propose a hot evaluation index based on message attention (likes and dissents) and message frequency. First of all, the text content is preprocessed, and the feature words are selected as valid words, which are transformed into word vectors. Then, the effective words are weighted by TF-IDF, and the weight of the words is adjusted to get the messages that reflect the specific location in a certain period of time. Finally, a reasonable heat evaluation method for the problem set is proposed based on the number of messages and attention of the masses.

For question 3, through the search and analysis of the existing data, analyze the quality of the responses from multiple angles, analyze them one by one, and select

valuable indicators. Establishing the quality evaluation scheme based on the given index. First of all, the reply content of the message is preprocessed. Secondly, according to whether the content of the reply is close to the topic and whether the content of the reply is complete. Quantifying the response data from the point of view of whether it can be explained or not, and finally we get a set of evaluation scheme.

Keywords: Word Segmentation Word2vec algorithm TF-IDF K-Means clustering Hot issues Mining Evaluation Indexes

目录

1.挖掘目标

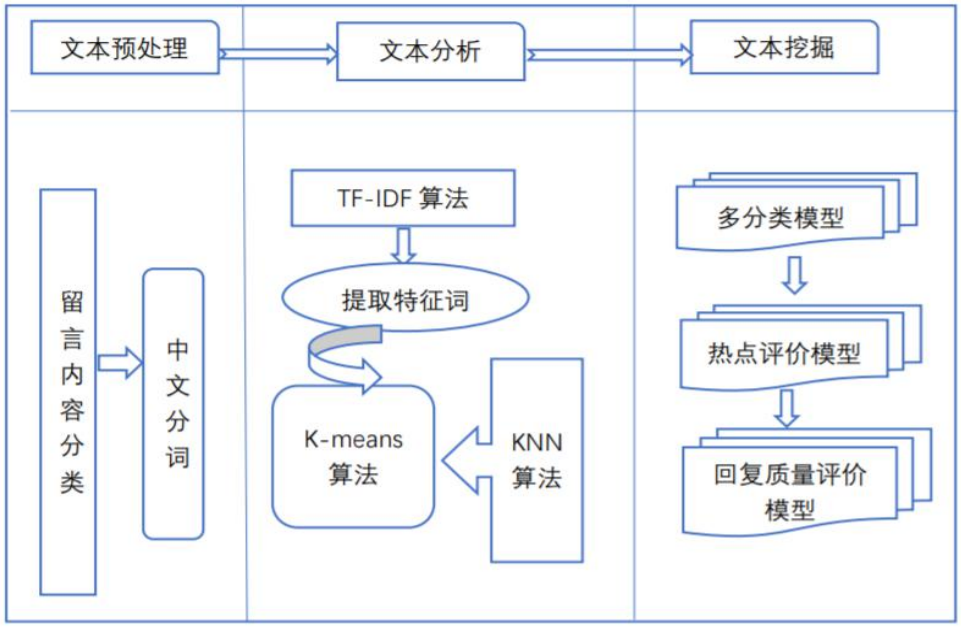
本次建模目标是建立自然语言处理技术的智慧政务系统，为提升政府的管理水平和施政效率达到以下三个目标：

1）利用文本分词和文本聚类对已有数据进行文本挖掘，根据留言信息，将其与所给出的内容分类三级标签建立联系，识别留言类型，给出留言分类。

2）将留言进行分类后，为提高服务效率，需根据在某一时间段内反映特定地点或特定人群问题的频数，对问题进行归类，及时发现热点问题，定义热度评价指标，给出留言排名及留言详情，以助于相关部门进行针对性处理。

3）针对给出的政府相关部门对留言内容的答复，对相关部门答复数据进行优劣量化，多角度的对答复意见给出评价方案。

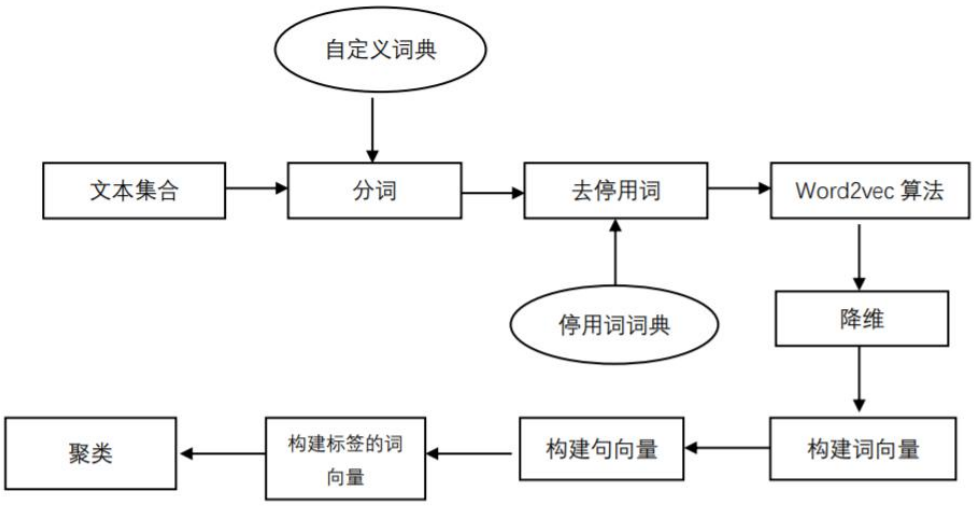
2.分析方法及过程



图表 1 总体流程图

2.1 问题一分析方法及过程

2.1.1 流程图呈现：



图表 2 问题一流程图

2.1.2 数据预处理

通过对已有的数据分析，发现了留言信息有以下特点：

1）留言内容文本有长有短，有大量铺垫性语句，造成数据量过大，不利于后期分析。这就需要进行关键词提取或者长文本转换。

2）重要内容出现在留言标题中或者在留言详情中会出现总结性的语句。

3）在表述同一事件时，群众可能在同一文本中对同种含义的事件有多种表达方式。

2.1.2.1 对留言详情中文分词

在附件二中，以中文的方式给出了留言主题和留言详情等信息。为了方便我们分类，本文将这些内容采用一定的规则进行分词处理，使得句子转化为多个词，词是文本数据中含有语义的最小单位，进行分词处理后，文本的一句话转换为多

个词来表示。利用 *python* 的中文分词工具 *jieba* [1]， *jieba* 分词结合了基于规则和基于统计这两类方法，本文采用 *jieba* 中最适合文本分析的精确模式进行处理。基于前缀词典进行词图扫描，前缀词典是指词典中的词按照前缀包含的顺序排列，形成一种层级包含结构。生成句子中汉字所有可能成词情况所构成的有向无环图；采用了动态规划查找最大规律路径，找出基于词频的最大切分组合。

2.1.2.2 去停用词

由于分词之后文本量的增大，为节省存储空间并提高搜索效率，搜索引擎在索引页面或处理索引请求时会自动忽略某些字或词。这些字或者词被称为停用词。停用词包括两类，第一类是助词、语气词等没有实际含义的词。第二类是副词，介词，连接词等出现次数均匀，但是对文本分类是无效的。

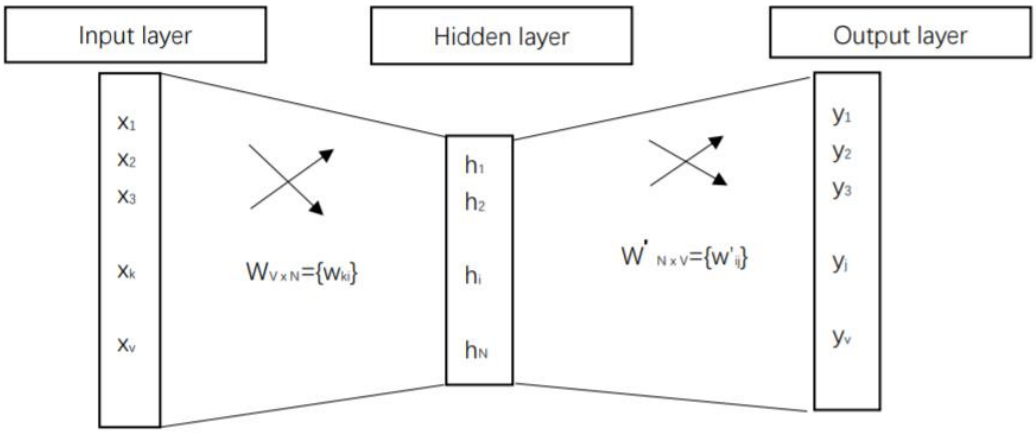
2.1.3 文本到数据转化

2.1.3.1 *Word2vec* 算法构建词向量

在自然语言处理（ *NLP* ）过程中，文本是由句子组成的，而句子又是由词语组成的。所以，在 *NLP* 的处理过程中，首先要考虑词语，词语是符号形式的（比如中文、英文等），所以需要把他们转换成数值形式，或者说——嵌入到一个数学空间里，这种嵌入方式，就叫词嵌入[13]（ *word embedding*).

词嵌入，是指把一个维数为所有词的数量的高维空间嵌入到一个维数低的多的连续空间中，每个单词或词组被映射为实数域上的向量。而 *Word2vec* 算法，就是词嵌入（ *word embedding*)的一种。简单点来说就是把一个词语转换成对应向量的表达形式，来让机器读取数据。机器学习模型，可以归结为 $f(x)->y$,在 *NLP* 中，把 x 看作一个句子里的一个词语， y 是这个词语的上下文词语，那么这里的 f ，便是 *NLP* 中经常出现的语言模型，这个模型的目的，就是判断 (x,y) 这个样本，是否符合自然语言的法则。而 x 的原始输入形式是用 *one-hot encoder*

表示的，所谓*one – hot encoder*，本质上是用一个只含一个 1，其他都是 0 的向量来唯一表示词语。要想得到一个词的向量表达方法，并且这个向量的维度很小，而且任意两个词之间是有联系的，可以表示出在语义层面上词语词之间的相关信息。我们就需要训练神经网络语言模型，这个模型的输出我们不关心，我们关心的是模型中第一个隐含层中的参数权重，这个参数矩阵就是我们需要的词向量。它的每一行就是词典中对应词的词向量，行数就是词典的大小。下图表示其原理（如图表 3）



图表 3 原理图

2.1.3.2 特征降维

对数据进行以上两步处理后，如果直接将处理结果作为特征词，通过向量空间模型构建文本特征向量，会导致文本特征向量数据量过大，并且文本特征向量维度会过高纬度过大会使文本特征向量数据过大，并且维度会过高，而维度的变大并不会使得数据的价值变大，反而会降低数据处理的效率，并且获得真正有价值信息的难度也会变大。所以，再进行聚类分析之前，我们要剔除掉与任务不相关的冗余特征项。文本数据的特征降维方式主要有特征选择以及特征抽取[3]。特

征选择原理（图表 4）

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}\right)$$

图表 4 特征选择原理

特征选择是基于某一准则函数，从原始特征集合中挑选出一个最具有代表性的子集来进行文本表示，从而达到降低维度的目的。其通常是根据所设定的评估函数来计算文本集中每个特征项的值，然后将不属于阈值中的特征项进行剔除。特征抽取如下（图表 5）

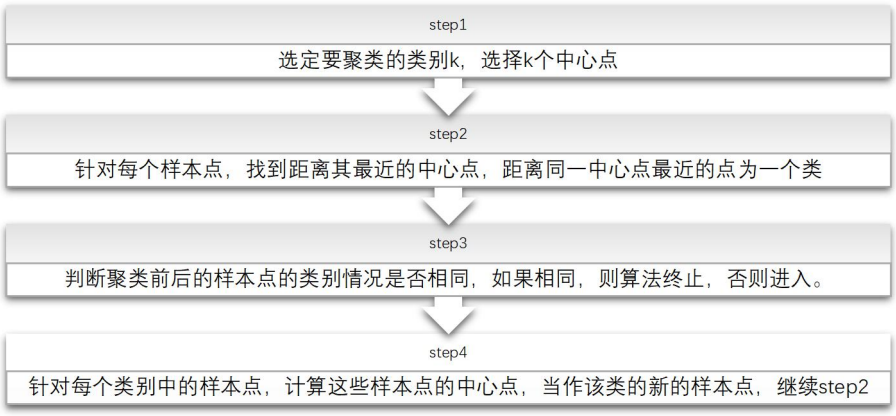
$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{im} \end{bmatrix}$$

图表 5 特征抽取原理

特征抽取是指应用线性或者非线性的映射方式，将初始特征按照一定方式，映射至维度较低的特征空间上，去实现文本的降维。

2.1.4K-means 算法

$K-means$[3]是最常见的聚类算法。算法的输入为一个样本集，通过该算法可以将样本聚类，具有相似特征的样本聚为一类。针对样本集中的每一个点，计算这个点距离所有中心点最近的那个中心点，然后将这个点归为这个中心点代表的簇。一次迭代结束后，针对每个簇类，重新计算中心点，然后针对每个点，重新寻找距离自己最近的点。如此循环，直到前后两次迭代的簇类没有变化。其基本步骤为：



假设有一个包含 n 个 d 维数据点的数据集 $X=\{x_1,x_2,\cdots,\ x_{i,}\cdots,\ x_n\}$,其中 $x_i\in R^d$，$K-means$ 聚类将数据集 X 组织为 K 个划分 $C=\{c_k\}$,每一个划分代表一个类 c_k，每个类 c_k 都一个类别中心 $u_k(k=1,2,...,K)$。选取欧氏距离作为相似性和距离判断准则，计算该类每个点到聚类中心的距离平方和，公式如下

$$J(C_k)=\sum_{x_i\in C_k}\|x_i-u_k\|^2$$

聚类目标是使各类总的距离平方和最小，其公式如下

$$J(C)=\sum_{k=1}^{K}J(C_k)=\sum_{k=1}^{K}\sum_{x_i\in C_k}\|x_i-u_k\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}\|x_i-u_k\|^2$$

当 $x_i\in c_i$ 时，$d_{ki}=1$，否则 $d_{ki}=0$，所以根据最小二乘法和拉格朗日原理，聚类中心 u_k 应该取为类别 c_k 类各数据点的平均值。

2.1.5 *Knn* 最近邻分类算法[4]

由 $K-means$ 分类得到聚类中心，利用 *Knn* 算法找出与各中心相似的元素，根据个数多的判定所属类别。根据向量空间模型，将每一类别文本训练后得到该类别的中心向量记为 $C_j(W_1,W_2,...,W_n)$,将待分类文本 T 表示成 n 维向量的形式 $T(W_1,W_2,...,W_n)$，则文本内容被形式化为特征空间中的加权特征向量，即

$D = D(T_1, W_1; T_2, W_2; ...; T_n, W_n)$。对于一个测试文本，计算它与训练样本集中每个样本的相似度，找出 K 个最相似文本，根据加权距离和判断测试文本所属的类别。具体算法步骤如下：

(1)对于一个测试文本，根据特征词形成测试文本向量

(2)计算该测试样本与训练集中每一个文本的文本相似度，计算公式如下

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{\sum_{k=1}^{M} W_{ik}^2} \sqrt{\sum_{k=1}^{M} W_{jk}^2}}$$

(3)按照文本相似度，在训练问本纪中选出与测试文本最相似的 k 个文本。

(4)在测试文本的 k 个邻近中，以此计算每类的权重，计算公式如下

$$P(X, C_j) = \begin{cases} 1, \text{若} \sum_{d \notin knn} Sim(x, d_i) y(d_i, C_j) - b \geq 0 \\ 0, \text{其他} \end{cases}$$

(5)比较类的权重，将文本分到权重最大的那个类别中

2.1.6 对分类方法的评价

为便于解释，以二分类问题为例，对查准率和查全率进行解释，将样本根据其真实类别与学习器预测类别的组合划分为真正例(true positive)，假正例(flase positive)，真反例(true negative)，假反例(flase negative)四种情形，令 TP, FP, TN, FN 分别表示其对应的样例数，则明显有 $TP + FP + TN + FN =$ 样本总数。分类结果的混淆矩阵为（如表格 1）

表格 1

真实情况	预测结果	
	正例	反例
正例	*TP*(真正例)	*FN* (假反例)
反例	*FP* (假正例)	*TN* 真反例)

查准率，是针对预测结果而言的表示的是预测为正的样例中有多少是真正的正样例，其公式为

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

查全率又称为召回率，是针对原来的样本而言的，表示的是样本中的正例有多少被预测准确，其公式为

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

查全率与查准率是一对矛盾的度量。一般来说，查准率高时，查全率往往偏低；而查全率高时，查准率往往偏低。

2.2 问题二分析方法及过程

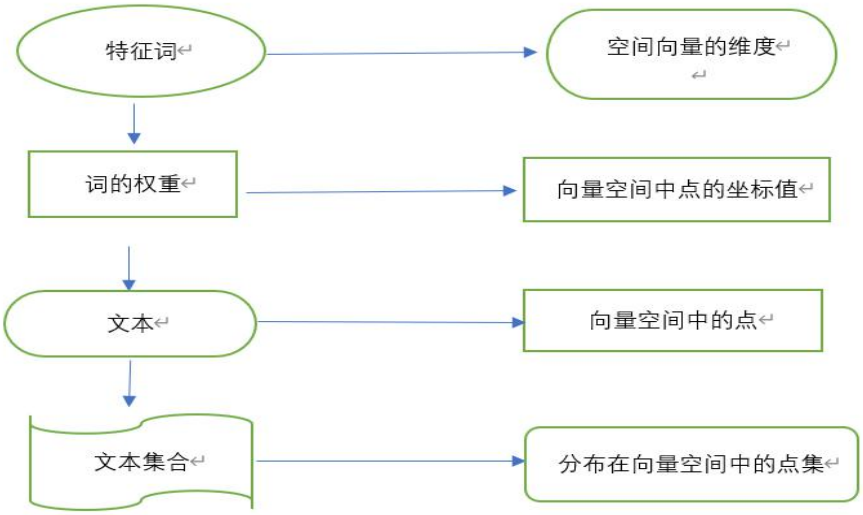
2.2.1 问题分析

因群众反映的问题内容繁多，需要对反映问题中的热点问题进行挖掘与分析，并运用信息分析技术，对留言中出现的热点问题及时做出回应，掌握处理问题的最佳时机，提高政府的服务效率。因此，如何在众多的留言问题中进行热点问题的汇集和挖掘，对于提高服务质量具有十分重要的现实意义。

在第一问的基础上，我们需要进一步确定在一段时间内群众反映很多的，具有突出影响的问题，将这些问题作为热点问题进行呈现。

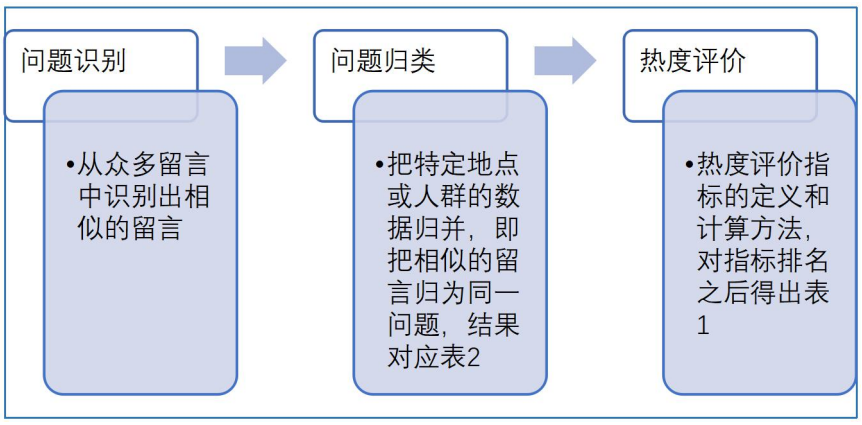
对于热点问题的计算方法，主要是从留言的关注度和时效度进行量化。留言的关注度是从群众的点赞数和反对数来考虑的，时效度是对在某一特定的时间内群众所反映的问题的度量。本文将从留言的关注度对热点问题进行挖掘。

对于问题二，本文采用向量空间模型[6](Vector Space Model)来进行文本处理，在*VSM* 模型中，每一条文本记录都可以表示成由特征词权重组成的*n* 维向量，文本与空间向量的关系（图表 6）



图表 6 文本与向量对应关系

2.2.2 流程图：



2.2.3 问题识别

在空间向量模型中，留言的文本应该是由正交特征向量组成的多维空间中点的集合。*VSM* 的实现需要以特征词作为向量表示的基本单位，它的维度的权重代表了它的重要性，因此，每一个留言对应着空间中的一个向量，而权重体现出特征值是否能表示出此条留言与其他留言的区别。在分词，去停用词的基础上，将留言映射到空间中，空间向量模型的最终目的就是将所有的特征此作为列，留言内容在特征词上的权重作为行，建立起一个矩阵。之后的运算基于此矩阵。

2.2.3.1 特征词的抽取

文本经过了对与特征词抽取和向量化处理后，才可以用数学模型的方式包含所代表的信息。在对留言文本进行语句分析、词频统计及词性判断的条件下，实现特征值的精确选取，来降低向量空间的维度。

要获得的特征词应该具有以下几点特征：

1）尽可能多的包含我们的留言信息

2）选区的特征词可以将文本明显区分开来

3）特征词的数目不能过大

4）特征词的选取要便于计算机计算

而特征词的选取方法有以下几种：

通过变换或者映射将原有的词语进行降维处理

直接在文本词语中选取特征词

3）通过数学方法选取

本文选用第二种方法选取，基本思想为从原有的词语中利用某一种算法选取具有代表性的词语作为特征词。选取权值最高的前 n 个词语得到留言的特征词流程如图

2.2.3.2 $TF-IDF$ 算法[2]

如果一个词在好多留言中都出现过，那么它包含的信息熵就低，反之，就越高。其中，词频（Term Frequency，TF）表示特征词在某一条留言中的重要程度。TF 越大权重越大。逆向文档频率（Inverse Document Frequency，IDF）反映的是特征词在整个留言内容中的出现情况，如果此词语在总体中出现的次数较少，表明，此词语可以很好地代表其属性，因此 IDF 越大权重越大；反之，如果此词语在总体中出现的次数较多，表明，此词语可以不能很好地代表其属性，因此 IDF 越大权重越小。由此得出，对于特征词语 T_i 在留言文本 x_j 中的权重可以表示为式

$$w(T_i,x_j)=TF(T_i,x_j)*IDF(T_i)$$

文档频率 DF 表示的是留言集合中出现特征词 T_i 的留言条数，记为 DF_i，通常情况下，DF 与 IDF 的值是成反比关系的，如式

$$IDF_i=\log(\frac{N}{DF_J})$$

其中，N 代表留言集合中的所有留言条数。IDF_i 值越大，表示 T_i 对留言内容的区分越明显。如果特征词只在一个留言内容中出现，那么则有公式

$$IDF=IDF_i=\log(\frac{N}{DF_i})=\log(\frac{N}{I})=\log(N)$$

反之，如果特征词在所有留言内容中都出现，则有公式

$$IDF = IDF_i = \log(\frac{N}{DF_i}) = \log(\frac{N}{N}) = \log(1) = 0$$

文本特征词的权重值一般按下式计算

$$w_i(x_j) = TF_{ij} * \log(\frac{N}{N_i} + 0.01)$$

其中，TF_{ij} 表示第 i 个特征值 T_i 在留言 x_i 中的出现次数，N 表示留言集合的总数量，$N_i = DF_i$ 表示留言集合中出现特征词 T_i 的留言条数。为便于计算，一般要将向量进行归一化，则最后得到的 $TF-IDF$ 权值计算公式为

$$w_t(x_j) = \frac{TF_{ij} * \log(\frac{N}{N_i} + 0.01)}{\sqrt{\sum_{k=1}^{N}(TF_{ij})^2 * \left[\log(\frac{N}{N_k} + 0.01)\right]^2}}$$

2.2.4 问题归类

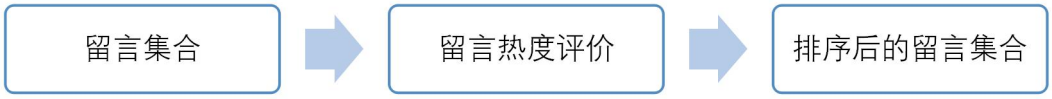
根据向量，计算相似度，可以得到相似矩阵，我们可以将问题聚类，可以得到问题的排名，考虑到时间问题，首先我们将问题对时间进行排序，给定一个时间周期，通过计算某一时间段内问题的排名，我们对热点问题进行总结。

2.2.5 热度评价

2.2.5.1 热度评价模块

由于留言的内容文本量较大，对于捕捉热点问题比较困难，本文考虑在第一问基础上，针对每一个一级分类标签，进行热度评分，区别不同一级分类下的活跃度，采用的热度评价方法，分为基于用户关注度和时效性的留言热度评价。

热度评价模块的流程图



2.2.5.2 留言的浏览关注度

热点问题往往是在一段时间内会受到群众的反映。考虑到影响留言的关注度的主要参数有：点赞数和反对数。这些参数在附件中已经给出，通过这些数据我们可以计算出该留言的浏览关注度。计算浏览关注度的公式为：

$$Attention = \alpha \Pr aise + \beta Opposition \quad （\alpha + \beta = 1）$$

2.2.5.3 留言的频数

基于特定的时间、特定人群的留言频数，进行筛选、归类、排序，得出群众某一时间段内所集中反映的问题。

由于热点问题是在某一特定的时间段内所反映的，我们所总结的留言问题，应该是是近期内收到的重点关注的问题，否则我们所发现的问题将不具代表性，不能被称为热点话题。

$$Frequency = \sum_{i=1}^{n} f$$

其中，n 为留言总数，$f = \begin{cases} 1 & f \in F \\ 0 & f \notin F \end{cases}$，F 为同一问题所组成的集合

2.2.5.4 留言热度算法

在介绍完留言的浏览关注度和留言的频数，留言的热度会综合考虑以上两种因素，对留言进行最终的考察。

留言 N 的浏览关注度是在留言期间内针对归类后同一类问题的所有留言其浏览关注度（点赞数、反对数）进行加权求和，计算公式为

$$Attention = \sum_{n} attention_n$$

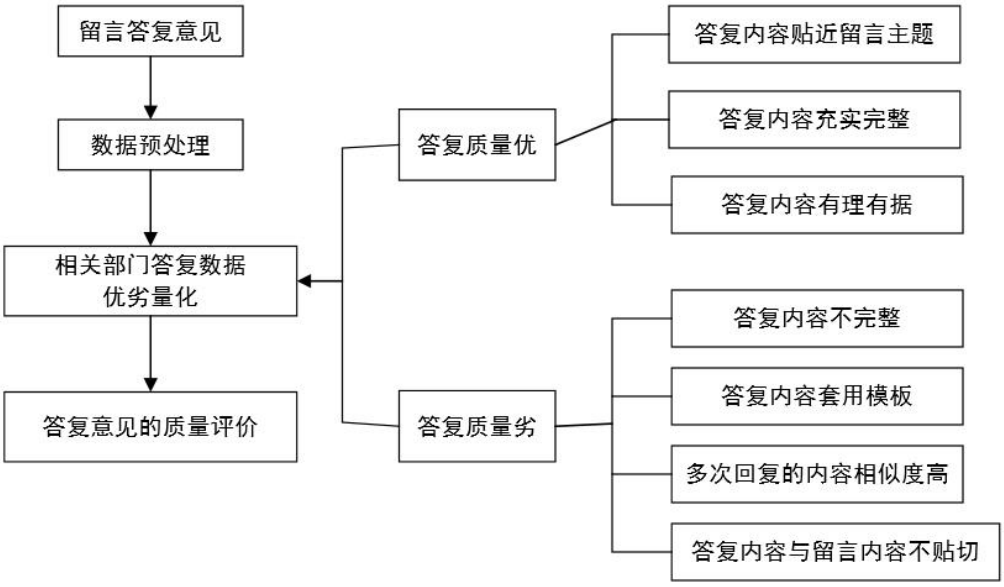
根据以上分析，得出留言 N 的热度公式为

$$Hotness_n = Attention_n^{\lambda} \cdot Frequency^{1-\lambda}$$

其中 $\lambda(0 < \lambda < 1)$ 的值会根据实验结果进行调整，以达到最佳效果。

2.3 问题三分析方法及过程

2.3.1 流程图：



2.3.2 回复数据优劣量化[3]

2.3.2.1 回复内容是否有理有据

回复内容需要有准确的表达，即有理有据。将回复内容中的关键词与留言的主题进行语义匹配。若在回复的文本中匹配到相对应的关键字，则认为该条回复相关性较大，假定算出相匹配的留言条为 N_0，总的留言数为 N，则两者的比值，

频率值，可以作为评价项F_1

$$F_1=\frac{N_0}{N}$$

2.3.2.2 回复内容是否充实完整

回复内容是否充实，回复内容的详细程度与文本长度有着直接的关系，一般情况下，文本长度较短，信息量也就较少，评分也就较低。因此，可以考虑利用对数函数来量化回复的文本长度与评分之间的关系，建立“回复内容是否完整”的评价项

$$\frac{1}{N}\sum_{i=1}^{N_0}\log_m L_i$$

其中，L_i为为针对第i个问题回复的文本长度，m为常数

2.3.2.3 回复内容贴近主题

留言是由很多关键词组成，利用这些关键词可以建立向量，设$X_K=\{x_{K1},x_{K2},...,x_{Ki},...,x_{KK}\}$表示由$K$个词语组成的词集，$Y_K=\{y_{K1},y_{K2},...,y_{Ki},...,y_{KK}\}$其中，$x_{Ki}$，$y_{Ki}$,分别表示$X_K$，$Y_K$中的第$i$个词语，留言与回复的相似矩阵为$S=(s_{ij})$

$$s_{ij}=F(x_{ki},y_{ki})，其中 i,j=1,2,...,k$$

由于文本进行两两对比过程计算量较大，因此我们只比较回复中与留言中与该词相似度最高的。

2.3.2.4 答复内容不完整

答复内容不完整与答复内容完整相对应

2.3.2.5 答复内容套用模板

有些部门在回复过程中，会出现固定的词语，本文建立了一个回复较差的关键词组成的集合。首先对回复内容进行关键词匹配，若一条回复中出现这个集合，则认为该条回答较差。

2.3.2.6 回复内容不贴近主题

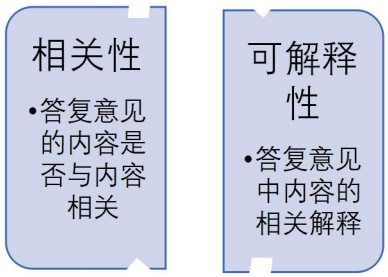
答复内容不贴近与答复内容相贴近相对应。

2.3.2.7 多次内容回复的内容相似度高

若多次回复内容相似度高，可以理解为相关部门回复过程中，出现多个文本较相似，解释为内容为固定的词语，与回复内容套用模板可以归为一类。

2.3.3 多角度评价

结合以上的分析，对相关指标进行合并，最终我们得到以下两点评价指标



2.3.3.1 相关性的度量

根据上述回复内容是否贴近主题来计算相似度。

2.3.3.2 可解释性

回复的可解释性理解为回复是否及时以及回复内容是否对所提出的问题进行全面地解释，附件已经给出留言时间及对应的回复的时间，我们利用两者的差

算出时间差，可以对比出回复是否及时，另外，在对相关性的分析中，我们已经给出计算方法。

2.3.4 答复质量评价模型

上面我们已经从多角度给出了评价的指标，为便于计算，我们只给出基于回复相关性和可解释性的评价模型。

3.结果分析

3.1 问题一结果分析

3.3.1 预处理

分词部分结果图（如图表 7）

```
[['市', '西湖', '建筑', '集团', '占', '道', '施工', '安全隐患'],
 ['市', '在水一方', '大厦', '人为', '烂尾', '多年', '安全隐患'],
 ['投诉', '市', 'A1', '区苑', '物业', '违规', '收', '停车费'],
 ['A1', '区', '蔡锷', '南路', 'A2', '区华庭', '楼顶', '水箱', '长年', '不洗'],
 ['A1', '区', 'A2', '区华庭', '自来水', '好大', '一股', '霉味'],
 ['投诉', '市', '盛世', '耀凯', '小区', '物业', '无故', '停水'],
 ['咨询', '市', '楼盘', '供暖', '一事'],
 ['A3', '区', '桐梓', '坡', '西路', '可可', '小城', '长期', '停水', '得不到', '解决'],
 ['C4', '市', '收取', '城市', '垃圾处理', '费', '平等'],
 ['A3', '区', '魏家坡', '小区', '脏乱差']]
```

图表 7 分词部分结果

3.1.2 文本数据化

通过去停用词对文本内容进行分词后，考虑将文本数据化转化成数值形式，通过*Word2vec*算法构建词向量，下图展示事故处理相关语句的词向量。（图表 8）

```
[-0.0025768573, -0.17998976, -0.015775379, -0.2674951, 0.22186975, -0.1961228, 0.03703919, -0.11123164, 0.070154175, 0.018356668, 0.012064987, 0.13217436, -0.04476746, 0.22259097, 0.18373734, -0.18002304, -0.2879216, -0.15052281, 0.15147895, -0.12556757, -0.11043545, 0.21313915, 0.16592333, -0.120208815, 0.09102462, -0.15228832, -0.0665827, -0.045857284, -0.2660031, -0.03486605, 0.089766495, 0.12014196, 0.078606, 0.18038237, 0.111875154, 0.023788067, -0.32209957, 0.053725813, 0.016242577, -0.12623483, -0.06751512, -0.08864673, -0.20107819, 0.11131455, -0.07312154, 0.056863133, 0.11073609, -0.072177656, -0.027421014, -0.026418261, -0.021211963, 0.049049795, 0.10187054, -0.00038123142, -0.21090457, 0.007873054, 0.20620936, 0.18470164, 0.17027108, -0.007903423, 0.22847818, -0.13045798, -0.11064421, -0.053438917, -0.13814998, -0.023753818, 0.16127324, -0.034947924, -0.21450983, -0.06808145, 0.0061902953, -0.21657081, 0.10362011, -0.067003705, -0.11692973, -0.050296083, -6.782797e-05, -0.21868475, -0.014559896, 0.14573877, 0.005361319, -0.04869418, 0.049298815, -0.00874092, -0.12904447, -0.3627968, 0.2052994, -0.019155584, 0.045403693, 0.12039683, 0.09338372, -0.0066664675, -0.0073448587, 0.067360386, 0.033256575, 0.13345577, 0.08220598, 0.029212184, -0.15819299, 0.07243101]
```

图表 8 部分词向量

3.1.3 聚类中心分类结果

通过前期的处理准备，根据特征词测试文本向量，计算该测试文本与数据集中每一条词条的文本相似度，选择出最相近的文本，比较权重，将该词条反映的权重最大的一级标签定为预测标签，下图展示部分词条分类后结果（程序、详细分类见附录）：

表格 2 **此条分类**

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
城乡建设	党务政务	国土资源	环境保护	纪检监察	交通运输	经济管理	科技与信息	民政	农村农业	商贸旅游	卫生计生	政法	教育文化	劳动和社会保障	
11	0	0	0	0	1	0	1	0	1	0	0	1	0	0	
12	0	0	0	0	0	0	0	2	0	1	0	0	0	0	
13	0	0	0	0	0	0	0	0	0	0	0	0	2	0	
11	0	0	0	0	0	0	1	0	1	0	0	2	0	0	
10	0	0	0	0	0	0	1	0	1	0	0	2	1	0	
6	0	2	0	0	4	0	1	0	1	0	0	1	0	0	
3	0	3	0	0	1	0	0	5	0	0	0	0	2	1	
10	0	1	0	0	0	0	0	1	1	0	0	2	0	0	
7	3	0	0	0	2	0	0	0	0	0	0	0	3	0	
6	0	0	0	0	0	0	1	4	1	1	0	2	0	0	
7	0	0	0	0	4	0	1	2	0	1	0	0	0	0	
11	0	0	0	0	0	0	1	0	1	0	0	2	0	0	
10	0	0	1	0	0	0	0	0	0	0	0	2	0	0	
12	0	0	0	0	0	0	0	0	2	0	0	1	0	0	
10	0	0	0	0	0	0	0	0	1	0	0	2	0	0	
11	0	0	0	0	0	0	0	2	0	0	0	2	0	0	
9	0	0	0	0	0	0	0	0	2	0	0	2	0	0	
14	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
8	3	0	0	0	4	0	0	0	0	0	0	0	0	0	
7	0	0	0	0	4	0	1	0	1	0	0	2	0	0	
10	0	1	0	1	0	1	0	0	0	1	0	0	0	1	
7	0	0	0	0	0	0	1	0	3	0	0	2	0	0	

分析上表：对于第一个词条聚类处理后得到与城乡建设距离最近，因此将该词条分类到城乡建设的标签中，以此类推。

表格 3

A	B	C	D	E	F
留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
24	A00074011	建筑集团占道施工有没	20/1/6 12:09:	围墙内。每天尤其	城乡建设
37	U0008473	大厦人为烂尾多年，安	20/1/4 11:17:	着，不但占用人行道	城乡建设
83	A00063999	市A1区苑物业违规收伙	9/12/30 17:06	多次向物业和社区	城乡建设
303	U0007137	南路A2区华庭楼顶水箱	19/12/6 14:40:	，霉是一种强致癌	城乡建设
319	U0007137	2区华庭自来水好大一股	19/12/5 11:17:	，霉是一种强致癌	城乡建设
379	A00016773	市盛世耀凯小区物业无	19/11/28 9:08:	业不是为业主服务的	城乡建设
382	U0005806	询A市楼盘集中供暖一	9/11/27 17:14	月亮岛片区近年规划	城乡建设
445	A00019209	西路可可小城长期停水	9/11/19 22:39	帮助至今没有找到	城乡建设
476	U0003167	取城市垃圾处理费不合	9/11/15 11:44	在的物业公司也未给	城乡建设
530	U0008488	A3区魏家坡小区脏乱差	9/11/10 18:59	让人好好休息一下	城乡建设
532	U0008488	A市魏家坡小区脏乱差	9/11/10 12:30	让人好好休息一下	城乡建设
673	A00080647	四届非法业委会涉嫌侵	9/10/24 11:29	责令B4区有关部门把	城乡建设
994	U0005196	梅溪湖壹号御湾业主用	19/9/18 22:43:	别的城市都已经一门	城乡建设
1005	U0006509	翠湾强行对入住的业主	19/9/18 13:36:	地产公司和金晖物业	城乡建设
1110	A00099772	市锦楚国际星城小区三	19/9/9 11:07:	是无通知，突然断电	城乡建设
1309	U0005083	和紫都用电的问题能不	19/8/21 15:12:	之后，我们的用电	城乡建设
1440	A0003288	际新城从6月份开始停电	19/8/6 10:28:	的生活，而且我们	城乡建设
1775	U0002150	区南西片区城铁站设立	19/7/4 18:52:	A市，并且规划有	城乡建设
1783	U0004763	府加大对滨水新城的投	19/7/4 14:25:	或者几个半大小树	城乡建设
1887	U000613	慧府锐月会小区经常停	19/7/1 20:14:	停电三次，说是	城乡建设

与实际情况比较发现前十条数据聚类较好，分析结果理想

3.1.4 对分类方法评价

使用 $F-Score$ 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

通过 Excel 计算得出 F_1 的结果为 0.798。

3.2 问题二结果分析

3.2.1 对附件 3 数据点赞数进行降序排列

表格 4

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
208636	A00077171	汇金路五矿万境K9县存在一	2019/8/19 11:34:04	狗咬人，请问有人对	0	2097
223297	A00087522	反映A市金毛湾配套入学的问	2019/4/11 21:02:44	纳入配套入学，A3	5	1762
220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	案情消息总是失望，	0	821
217032	A00056543	A市58车贷特大集资诈骗案保	2019/2/25 9:58:37	小股东、苏纳弟弟苏	0	790
194343	A000106161	A市58车贷案警官应跟进关注	2019/3/1 22:12:30	经侦并没有跟进市领	0	733
263672	A00041448	小区距长赣高铁最近只有30米	2019/9/5 13:06:55	复到我如下问题：1、	0	669
193091	A00097965	富绿物业丽发新城强行断业主	2019/6/19 23:28:27	提供地摊上买的收据	0	242
284571	A00074795	省尽快外迁京港澳高速城区	2019/1/10 15:01:26	、长浏高速出口，这	0	80
200667	A00079480	要把和包支付作为任务而不让	2019/1/16 17:01:25	层工作者也不理解，	0	78
262052	A00072424	月亮岛路沿线架设110kv高压线	2019/3/26 14:33:47	从上电力线路，应采	0	78
226723	A00040222	一大道全线快速化改造何时启	2019/9/15 15:31:19	改造，打通机场北通	0	66
272089	A00061602	A6区月亮岛路110kv高压线的	2019/4/9 17:10:01	地省体操学校、西地	2	55
281898	A00096623	号云时代多栋房子现裂缝，质	2019/2/25 15:17:38	检站处理，陷入了与	5	55
239595	A00057814	回东六路恒天九五工厂地块，	2019/11/8 15:48:07	地区，这里潜力巨大	0	44
267630	A000100648	铁3号线松雅湖站点附近地	2019/5/22 23:37:38	东四路和东四路西侧	0	42
279062	A00027836	加大A7县东六线榔梨段拆迁	2019/1/17 19:25:45	有土地（正钢机械厂	1	42
209742	A00012969	区郝家坪小学什么时候能改扩	2019/3/24 21:07:12	小学、滦湾路小学	0	41
239670	A00080329	东六线以西泉塘昌和商业中心	2019/1/11 15:46:04	置业有限公司厂房，	0	41

3.2.2 对附件 3 数据反对数进行降序排列

表格 5

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
203187	A00024716	咨询A9市高铁站选址的问题	2019/8/1 13:48:57	城区东进的步伐，为	53	10
215654	A00013092	来漫城物业不作为，还能评为	2019/11/21 11:40:50	移动公司，并要求业	15	2
275406	A00089627	县东一路公交首末站是临时的	2019/2/13 10:08:14	且不用增加线路。星	14	10
248495	A00056543	县泉塘中学领导对待老师不公	2019/4/4 11:15:52	不能补课，只能不	11	12
198975	A00051906	咨询A7县道路规划的问题	2019/8/1 17:13:39	划的，会从东七路	9	4
217233	A00050376	壹城小区开放式入户通道存在	2019/6/25 14:19:30	词采用“不宜”。	9	0
360114	A0182491	经济学院体育学院变相强制	2017-06-08 17:31:20	了合同，并且公司也	9	0
236295	A00053343	县广圣大酒店4楼瑞生堂足道	2019/1/18 15:35:48	99，188，888，133，	7	8
232887	A00088606	妇幼医院连儿童常见病都无法	2019/1/23 15:30:02	所有的家庭带来了极	7	7
237034	A00061821	绿地香树花城小区公积金贷款	2019/6/19 16:19:08	3月底已经完成现场	7	0
262850	A00073469	屋定制个体老板恶意拖欠工资	2019/5/25 13:43:32	物流运费，商场房租	6	6
202622	A000113231	投诉A7县保利香槟国际开发商	2019/3/28 10:25:30	，随时准备复工。但	6	4
260093	A00050376	解决A市奥克斯缔壹城连廊的	2019/6/9 13:33:27	政府、还是怪业主	6	0
360108	A0283523	小区一楼的夜宵摊严重污染附	2019-08-01 16:20:02	护社会和谐稳定，	6	0
223297	A00087522	反映A市金毛湾配套入学的问	2019/4/11 21:02:44	纳入配套入学，A3	5	1762
281898	A00096623	号云时代多栋房子现裂缝，质	2019/2/25 15:17:38	检站处理，陷入了与	5	55
288398	A00053962	出让星沙滨湖路以南，特立	2019/2/11 14:09:40	间有大量的闲置土	5	40

3.2.3 热点问题部分表

基于不同留言内容的相似度和留言关注度，计算相似留言的频数、点赞数、反对数和关注度，对得到的结果进行加权，通过留言热度算法得出留言 N 的热度公式，将结果以文本形式呈现。（如表格 6）

表格 6

热度排名	热度指数	时间范围	问题描述
1	53	2019/3/2-2019/4/2 00	A2 区李丽发新城附近无资质混凝土搅拌站为何禁而不止？
2	31	2019/3/2-2019/4/2 00	A2 区丽发新城小区云塘路还没有铺沥青
3	29	2019/3/2-2019/4/2 00	万科魅力之城小区底层门店深夜经营，各种噪音扰民
4	24	2019/2/20-2019/3/19	A3 区 A3 区街道惟盛园安置小区收费不合理
5	23	2019/3/2 002019/4/2 00	投诉 A 市伊景园滨河院捆绑销售车位
6	22	2019/6/20-2019/7/17	请履行承诺把申请的 A 市 A1 区廉租房退还给我
7	20	2019/4/17-2019/8/16	对 A 市经开区泉星公园项目规划再进一步优化的建议
8	19	2019/3/2-2019/4/2	万科魅力之城小区底层门店深夜经营，各种噪音扰民
9	17	2019/12/3-2019/5/18	请 A 市 A5 区公安分局和候家塘派出所认真对待聚利网诈骗一案
10	15	2019/11/22-2019/3/12	A 市 A3 区金岭小区内开 KTV，怎么回事？

3.3 问题三结果分析

文本匹配度较好，能达到 79.36%相似度，说明答复较好

相关部门对留言的答复时应注意以下几点：

1）在回复过程中，应注意语言的逻辑性，问题的针对性，以及说话的方式，做到简洁

2）答复具有时效性，应做到迅速、及时。

3）相关部门在答复一段时间后，应该通过走访，电话等调查方式对群众进行调查，查看问题是否解决。

4.参考文献

[1]杨开平,李明奇,覃思义.基于网络回复的律师评价方法[J].计算机科学,2018,45(09):237-242.

[2]时志芳. 移动投诉信息中热点问题的自动发现与分析[D].北京邮电大学,2013.

[3]李思洋. 面向热点话题发现的聚类算法研究[D].东北师范大学,2016.

[4]王岩. 面向金融领域 BBS 的话题发现和热度评价[D].哈尔滨工业大学,2010

[5]马婵媛. 基于 K-Means 的分布式文本聚类系统的设计与实现[D].西安电子科技大学,2018.

[6]张海涛. 基于文本降维和蚁群算法的文本聚类研究[D].安徽大学,2016.

[7]王素格,李书鸣,陈鑫,穆婉青,乔霈.面向高考阅读理解观点类问题的答案抽取方法[J].郑州大学学报(理学版),2018,50(01):54-59.

[8]王烟.自然语言处理技术在建筑使用后评价中的应用[J].南方建筑,2019(01):82-87.

[9]胡菊香,吕学强,刘克会.利用类别引导词的投诉文本分类[J].现代图书情报技术,2015(Z1):97-103

[10]方小飞,黄孝喜,王荣波,谌志群,王小华.基于 LDA 模型的移动投诉文本热点话题识别[J].数据分析与知识发现,2017,1(02):19-27.

[11]郭建永. 聚类分析在文本挖掘中的应用与研究[D].江南大学,2008.

[12]吴夙慧,成颖,郑彦宁,潘云涛.K-means 算法研究综述[J].现代图书情报技术,2011(05):28-35.

[13]词嵌入，百度百科 https://baike.baidu.com/item/词嵌入/