“智慧政务”中的文本挖掘应用

摘要

随着信息技术的不断发展，各行各业的信息化建设成为必然趋势，各类数据飞速增长，各类的网络问政平台和渠道应运而生，如微信、微博、市长邮箱、阳光热线等，这些都成为政府了解民意、汇聚民智以及凝聚民气的重要渠道[1]，如何在政府信息化建设中应用好这些大数据成为一项重要的研究课题。

本文基于已有的留言文本数据基础，通过数据特征工程、数据分析和深入的数据挖掘，对未分类文本数据进行文本聚类自动分类；此外，自定义了热度评价指标，挖掘文本中的热点问题并排行显示热度值；最后，探究了政府答复民众的质量性、可解释性、可靠性等评价方案。总之，本文结合量化和质化来处理“智慧政务”中文本数据并解决问题，主要研究内容有如下几点：

1. 针对每个问题提出不同的数据预处理方案：方案中包含缺省值、重复值、空值等处理，此外，对于问题一二中进行了特殊符号剔除预处理，问题三需从语法结构上进行分析，保留了一些句号、逗号等预处理。

2. 针对问题一，是典型的文本分类问题，本文方案流程采用构建文本特征、然后导入基于深度学习模型或者机器学习模型，最后根据混淆矩阵计算模型评分标准 F 值(F-Measure)评估模型质量。

3. 针对问题二，涉及自然语言处理的很多领域，本文解决思路方案为聚类、命名实体挖掘、语义解析等，在进行建模、聚类之后，对得到的结果利用层次分析法制定合理的热度评价指标，并赋予合理的权重然后量化得到分数进行排名。

4. 针对问题三，为一种决策类的问题，本文根据时间、语句结构、语句语义等为留言质量提出评价方案，将复杂的多决策问题逐一分解，分解为相关性、时效性、完整性、可解释性，在问题二热度评价模型的基础上，再次搭建层次分析模型对留言质量进行评价，分为很好，良好，一般，差，很差五个等级。

关键字： 智慧政务，文本分类，KMeans 聚类，层次分析法

Abstract

In recent years, with the update of the times, the rapid development of the society, the popularization of the Internet age, the data shows "explosive" growth, all kinds of text data continue to skyrocket, wechat, microblog, mayor's mailbox, sunshine hotline and other network platforms for politics are constantly born, and have become an important channel to understand the public opinion, gather people's wisdom, and gather people's spirit.

The text of this message is all in the existing text On the basis of data, through data feature engineering, data analysis and in-depth data mining, the unclassified text data is classified automatically by text clustering, and the heat evaluation index is defined, the hot issues in the text are mined and the heat value is displayed in the ranking. Finally, the government's quality, interpretability, reliability and other evaluation schemes in response to the public are explored. Combined with quantitative and qualitative text data, this paper explores the application of text mining in "smart government".

1. Different data preprocessing schemes are proposed for each problem, including default value, duplicate value, null value and so on. Some special solutions are proposed according to the analysis of the problem.

2. For problem one, it is a typical text classification problem. The basic process is to build text features, then import the model based on deep learning or machine learning model, and finally calculate the model scoring standard F-measure according to the confusion matrix to evaluate the model quality.

3. In view of problem 2, it involves many fields of natural language processing, and the general solution direction is clustering, named entity mining, semantic analysis, etc. after modeling and clustering, the results are used AHP to develop a reasonable heat evaluation index, and given a reasonable weight and then quantified scores for ranking.

4. For problem 3, it is a decision-making problem. According to various aspects,

an evaluation scheme is proposed for message quality. A complex multiple decision-making problem is decomposed one by one into relevance, timeliness, integrity and interpretability. On the basis of problem 2 heat degree evaluation model, the AHP model is built again to evaluate message quality, which is divided into good and good, General, poor, very poor five grades.

Key Words： Smart government Text categorization Clustering Analytic Hierarchy Process

目录

1 绪论

1.1 研究背景及意义

随着大数据、云计算、人工智能等技术和新媒体的迅速发展以及政府职能的不断转变，传统的电子政务模式已经无法跟上信息时代的脚步, 利用信息技术, 促进电子政务向智慧政务的转变已是必然的趋势。“智慧政务”是电子政务发展的高级阶段[2]。网络问政是电子政务发展实践中的新形式，成为了政府治理能力提升的新契机，亦是构建智慧政务的重要部分。微信和微博等网络问政平台，为政府加强了解社情民意塑造了一个全新的信息传播平台，并逐步替代政府传统方法的成为一个重要的渠道。大数据在网络问政方面的应用，一是能尽可能全面的获取相关的社情民意，从而反映更真实的民心民意，二是能更迅速、准确的完成数据的统计、分析和可视化，更直观的展示分析结果，为政府工作人员做出科学决策提供了可参考性，推进了治理现代化，促进了政府由电子政府向智慧政府的转变。

鉴于智能手机和移动端平台的便捷性以及网络的及时性与互动性，越来越多的普通民众参与到网络问政，各类社情民意相关的文本数据也随之不断攀升，能在众多数据中将文本分类并快速整理提取有用信息，对政府的工作效率提升起着重要的作用，同时若能将相关部门对民众留言的答复意见进行统计、分析，构建一个评估模型，有利于提高答复意见质量，进而提升政府的服务质量，推动服务型政府的建设。

智慧政务的需求日益增加，但是大数据为政府工作带来便利的同时也存在一定的挑战，怎么在海量数据中快速挖掘关键、有价值的信息？怎么找到各个数据中的关联性？若靠传统的人工方法来进行划分归类，需要耗费大量人力物力且不够准确，因此，迫切需要一种新的数据分析技术针对数据仓库海量的文本数据挖掘进行分析，并从中提取有价值的信息，文本挖掘技术应运而生[3]。

文本挖掘也称文本数据挖掘，一般是指文本在处理中获取高质量高价值的信息的这个过程。在文本挖掘当中需要用到自然语言处理（NLP），这项技术在今天随着人工智能等的子领域而一并飞速发展。自然语言处理可用于文本分词，亦可以用于实体识别等方面，而文本挖掘应用最多的几种技术有文档分类、文档聚

类和摘要抽取等。文本挖掘与自然语言处理技术在智慧政务系统的应用,为实现“智慧政务”带来的难题提供了有效的解决方案，推动了政府的管理水平和施政效率的提升。

1.2 本文相关工作

针对本次赛题，本文通过对基于自然语言处理的智慧政务系统的实现，主要做了以下的工作：利用提供的留言与相关答复意见等数据，将所要研究的内容分为如下四个问题并利用文本挖掘技术等方法解决。

（1） 基于群众留言的留言主题，进行文本分词，构建特征矩阵，并利用Word2Vec+LogisticRegression 获取词向量、构建文本分类模型将留言按照所提供的一级分类标签进行分类划分，并给出评价结果。

（2） 基于留言主题与留言详情，利用 word2vec 构建词向量，构建 K-Means 文本聚类模型，对聚类后的结果按照某一时段反映特定地点、人群的留言进行提取，作为每一类别的问题摘要。并对每一类别中的留言详情进行相似度计算，提取出每一类别中的与其余文本相似度最高的内容作为类别的主要内容，并利用 TextRank 对该文本进行提取作为每一类的详情摘要。

（3） 基于每一类别以及每一类，通过赞同数、反对数、文本语义、留言时间等方面分别构建评价指标，并给出自定义评价方案，利用层次分析法赋予合理的权重，并计算分数进行排行。

（4） 基于相关部门给留言的答复意见，分别在答复质量的相关性、时效性、完整性、可解释性方面制定出合理的评价方案，给出算法流程，利用层次分析法赋予合理的权重，为答复意见构建一个较为合理的评估模型，并计算得分，转换为很好，良好，一般，差，很差五个等级进行评估。

2 模型及算法理论

2.1 Word2Vec

无论是传统方法的机器学习，还是现在较流行的深度学习，在学习数据输入方面，必须要输入计算机能识别和认知的数据。而在自然语言处理（NLP）这个信息工程的子领域上，其所获得的数据皆为文本数据，并不是计算机可直接识别使用的数据。我们需要将这些数据转换为数值的形式后再进行数据的使用。这种嵌入方式称为为词嵌入(Word Embedding)，所以需要一款工具来对文本数据进行量化，为计算机提供合适的数据输入。

这款数据处理工具就是 Wrod2Vec，在自然语言处理领域相关受到广泛的应用，如聚类、同义词、语义分析、词性分析等等。其原理是借助了深度学习思想如图 2.1 所示，通过不断的对数据进行训练，把输入文本训练成 K 维向量空间的向量运算，而在向量空间上的距离代表着词与词之间的相似度。

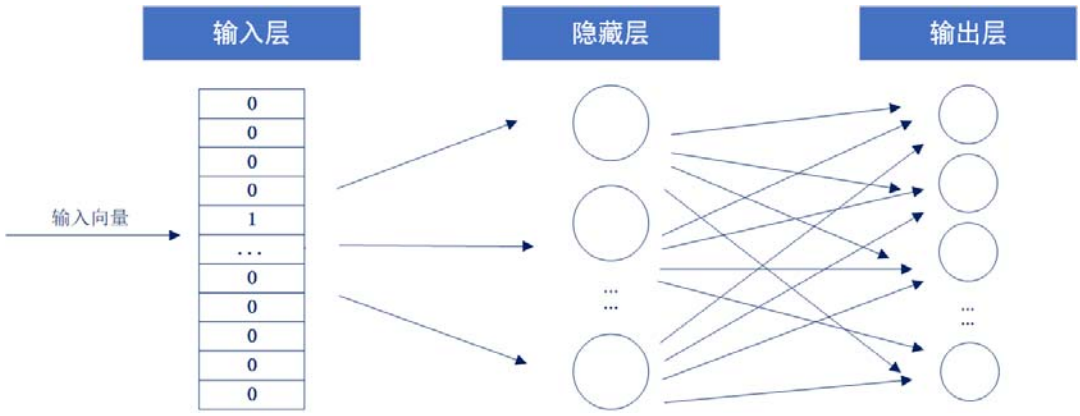

图 2.1　Word2Vec 模型基本思想

其中输出向量是 One-Hot Vector（独热编码矩阵），独热编码也是一种可以用来产生词向量的方法，其原理是根据语料库构建一个含有维度为 K 的向量，对于每一个文本内出现的词语，在相应维度下标上显示为 1，其余为 0，如图所示(图二)。

这样做的优点，是解决了文本在量化上的问题以及一定程度的扩充提取的特征；但缺点是当语料库非常大的时候，One-Hot 存在着维度爆炸的风险，从而给向量计算带来很大的困难，并且得到的特征十分稀疏，此时将该向量作为输入最后得到的结果差强人意，故利用 Word2Vec 得到的词向量代替 One-Hot。

对于定义数据的输入和输出，Word2Vec 利用两种模型体系结构之一来生成单词的分布式表示形式：连续词袋（CBOW）或连续 Skip-Gram。下面将对这两种模型进行简单的介绍。

2.1.1 CBOW 模型

CBOW(Continuous Bag Of Words)又称作为连续词袋模型，是一个三层神经网络，如图 2.2 所示，在连续词袋体系结构中，该模型从周围上下文词的窗口中实现预测当前词。输出对当前单词的预测(图三)。

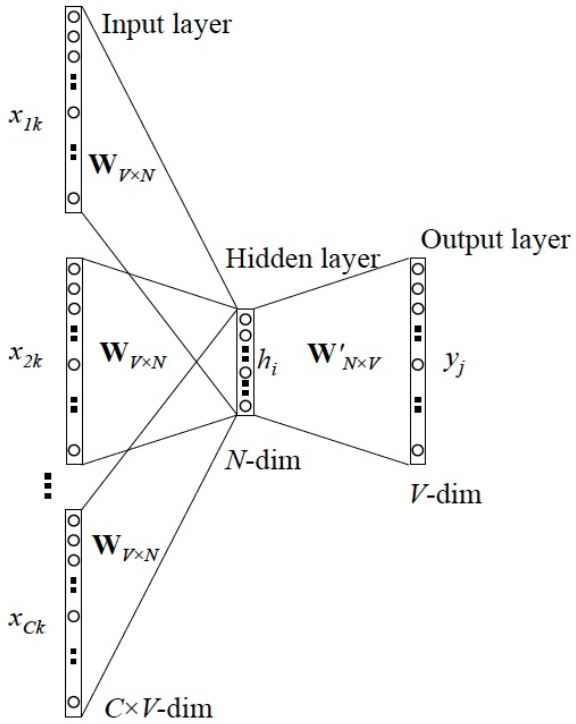

图 2.2　CBOW 模型原理详细图

图中输入层是由 One-Hot 编码输入的上下文矩阵$\{x_1, x_2, x_3, ..., x_C\}$构成，语料库维度为 V，隐含层是一个自定义的 N 维向量，最后输出层也是一个被 one-hot 编码的输出单词 y。

一、计算隐藏层 h 的输出输入向量的加权平均，如公式 2.1 所示

$$h = \frac{1}{C} W \cdot \sum_{i=1}^{C} x_i \tag{2.1}$$

其中W为输入层到隐藏层的权重矩阵，C为窗口大小。

二、计算在输出层每个节点的输入，如公式 2.2 所示

$$u_j = v_{wj}^{'T} \cdot h \tag{2.2}$$

其中$v_{wj}^{'T}$是输出矩阵$W^{'}$的第 j 列

三、最后计算输出层的输出，如公式 2.3 所示

$$y_{c,j} = p\left(w_{y,j} \mid w_1, w_2, w_3, ..., w_c\right) = \frac{\exp(u_j)}{\sum_{j^{'}=1}^{V} \exp(u^{'} j)} \tag{2.3}$$

四、定义损失函数，迭代更新权重，如公式 2.4 所示

$$\begin{aligned} E &= -\log\left(w_O \mid w_I\right) \\ &= -v_{W_O}^{T} \cdot h - \log \sum_{j^{'}=1}^{V} \exp(v_{w_{j^{'}}}^{T} \cdot h) \end{aligned} \tag{2.4}$$

五、对上述损失函数求偏导，反向传播，得到权重矩阵$W^{'}$的更新规则。如公式 2.5 所示

$$w^{'(new)} = w_{ij}^{'(old)} - \eta \cdot (y_i - t_j) \cdot h_i \tag{2.5}$$

其中，η 为学习率

六、同理可以得到权重矩阵W 的更新规则，如公式 2.6 所示

$$w^{(new)} = w_{ij}^{(old)} - \eta \cdot \frac{1}{C} \cdot EH \tag{2.6}$$

最终得到了每个词语的结果值和每个词相对应的词向量。

2.1.2 Skip-gram 模型

Skip-Gram 模型又称作为连续跳过语法架构，与 CBOW 模型恰恰相反，通过利用一个单词来预估上下文的，相对 CBOW 模型较慢，但是对不常用的单词来说表现更好，其原理模型图如下所示：

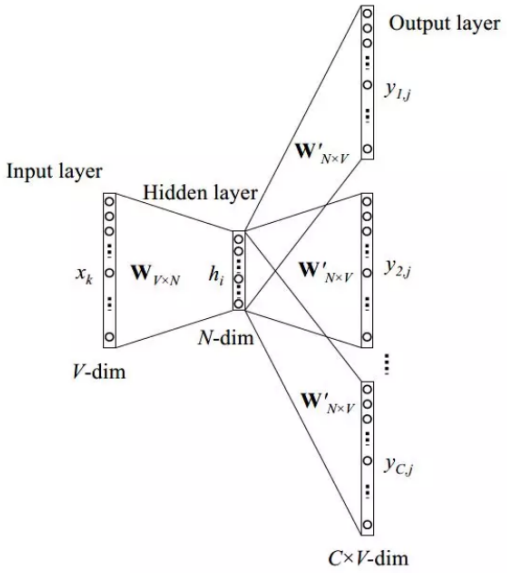

图 2.3　skip-gram 神经网络模型图

图 2.3 中，输入变量 x 为某个输入单词的 one-hot 编码，其中对应输出向量为 $\{y_1, y_2, ..., y_n\}$，语料库维度为 V，隐含层是一个自定义的 N 维向量，其中输入层和隐藏层中间的权重矩阵 W_i，表示第 i 行代表词汇表中第 i 个单词的权重，故 W 和 $W^{'}$ 权重矩阵为我们所需要的学习的目标，对于输出结果而言，每个单词向量有 $N \times V$ 维的输出向量 $W^{'}$，并且还有 N 个结点的隐藏层，就可以发现在隐藏层中的节点 h_i 的输入即为输入层输入的加权求和，又因为输入向量矩阵 x 为 one-hot 编码。因此只有向量中元素非零的才能对隐藏层产生影响，对于输入向量 x 有 $x_k = 1$，$x_{k^{'}} = 0, k \neq k^{'}$ 成立，故模型中的隐藏层的输出只与权重的第 k 行相关，并且可以得到如公式 2.7 所示：

$$h = x^T W = W_{k,.} := v_{wI} \tag{2.7}$$

对于每个输出，可计算得到公式 2.8 所示：

$$u_{c,j} = v_{wj}^{'T} h \tag{2.8}$$

由于输出层中的每个单词都是共享权重的，因此我们有 $u_{c,j} = u_j$。再根据输出结果，通过 softmax(激活函数)产生第 C 个单词的多项式分布，如公式 2.9 所示：

$$p(w_{c,j}=w_{O,c}\mid w_I)=y_{c,j}=\frac{\exp(u_{c,j})}{\sum_{j'=1}^{V}\exp(u_{j'})} \tag{2.9}$$

定义输出了 skip-gram 模型的输入向量以及输出的多项式分布，接下来就是定义损失函数进行不断的学习，通过似然函数定义损失函数如公式 2.10 所示：

$$\begin{aligned} E &= -\log p(w_{O,1},w_{O,1},...,w_{O,C}\mid w_{O,I}) \\ &= -\log\prod_{c=1}^{C}\frac{\exp(u_{c,j})}{\sum_{j'=1}^{V}\exp(u_j')} \end{aligned} \tag{2.10}$$

对上面的概率公式求偏导，可以得到输出权重矩阵W'的更新规则如如公式 2.11 所示：

$$w'^{(new)}=w_{i,j}'^{(old)}-\eta\cdot\sum_{c=1}^{C}(y_{c,j}-t_{c,j})\cdot h_i \tag{2.11}$$

同理对于输入权重矩阵W的更新规则如公式 2.12 所示：

$$w^{(new)}=w_{ij}^{(old)}-\eta\cdot\sum_{j=1}^{V}\sum_{c=1}^{C}(y_{c,j}-t_{c,j})\cdot w_{wj}'\cdot x_j \tag{2.12}$$

2.2 TF-IDF 算法

TF-IDF(term frequency–inverse document frequency，词频-逆向文件频率)，是一种用于信息检测、文本挖掘领域上的一种算法[4]，用来评估一个单词再一个文档集或者语料库中的重要程度，TF-IDF 的值会随着单词在文档中出现的次数增加而增加，同时也会因为语料库中出现的次数增加而减少，解决了一些无用词如：“的”，“吗”，“是”，影响整个算法，可用于构建特征工程，用与计算文本相似度、文本分类等。

TF（Term Frequency）词频，表示一个给定词语t在一篇文档中出现的频率，TF 越高表示词语t对于这篇文档来说很重要，越低则对于这篇文档来说不重要，但这只是一方面，不能作为唯一标准，例如一些停顿词：“的”，“吗”，“我”这类词对于文档来说是一些近似无用词，故需要 IDF 进行权重均衡化。

对于某一文档d_j里的词语t_i来说，t_i的词频可以表示为如公式 2.13 所示：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.13}$$

其中 $n_{i,j}$ 是词语 t_i 在文档 d_j 中的出现次数，分母则是在文件 d_j 中所有词语的出现次数之和。

IDF(Inverse Document Frequency)，逆向文件频率的主要思路是：对于词语t而言，如果包含该词语的文档数越少，则 IDF 的值越大，说明词语 t 在整个文档集层面上具有很好的类别区分能力。

给定某一特定词语的 IDF，计算公式如公式 2.14 所示：

$$IDF_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2.14}$$

其中 $|D|$ 是语料库中所有文档总数，分母是包含词语 t_i 的所有文档数。

2.3 K-Means 类模型算法

2.3.1 基础算法之 K-Mean

K-Means 算法是常见的无监督聚类算法之一，运用十分广泛，本文主要应用于文本留言的聚类。K-Means 的思想很简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为 K 个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。其主要算法流程如下：

①给定需要聚类的数据集合

$$[x^{(1)}, x^{(2)}, ..., x^{(m)}] \tag{2.15}$$

②随机选择 k 个聚类质点为：

$$\mu_1, \mu_2, ..., \mu_k \in R^n \tag{2.16}$$

即每一个样本元素都是 n 维向量。

③对于每一个样本 i，计算其应该属于的类，如公式 2.17 所示

$$c^{(i)} = \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2 \tag{2.17}$$

④对于每个一个类别，重新计算每一个类别的质心，如公式 2.18 所示

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \tag{2.18}$$

其中 $c^{(i)}$ 代表样例 i 与 k 个类中距离最近的那个类，

⑤重复③、④，并给出损失函数，直至收敛

$$J=\sum_{i=1}^{n}\sum_{j=1}^{k}r_{ij}\left\|x^{(i)}-\mu_j\right\|^2 \tag{2.19}$$

其中 r_{ij} 表示数据点 $x^{(i)}$ 被分类到 μ_j 的时候为 1，否者为 0

改算法最为重要的是 k 值得选择，不同得 k 值对聚类得结果造成很大得影响，文本利用手肘法来确认 k 值的大小，其主要思想是根据 KMeans 的核心指标：SSE(sum of the squared errors，误差平方和)[5]来判断的，其公式如 2.20 所示：

$$SSE=\sum_{i=1}^{k}\sum_{p\in C_i}\left|p-m_i\right|^2 \tag{2.20}$$

其中 C_i 是第 i 个簇，p 是 C_i 中的样本点，m_i 是 C_i 的质心，SSE 是所有样本的聚类误差，代表了聚类效果的好坏。

根据公式可以看出随着聚类数 k 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和 SSE 自然会逐渐变小。

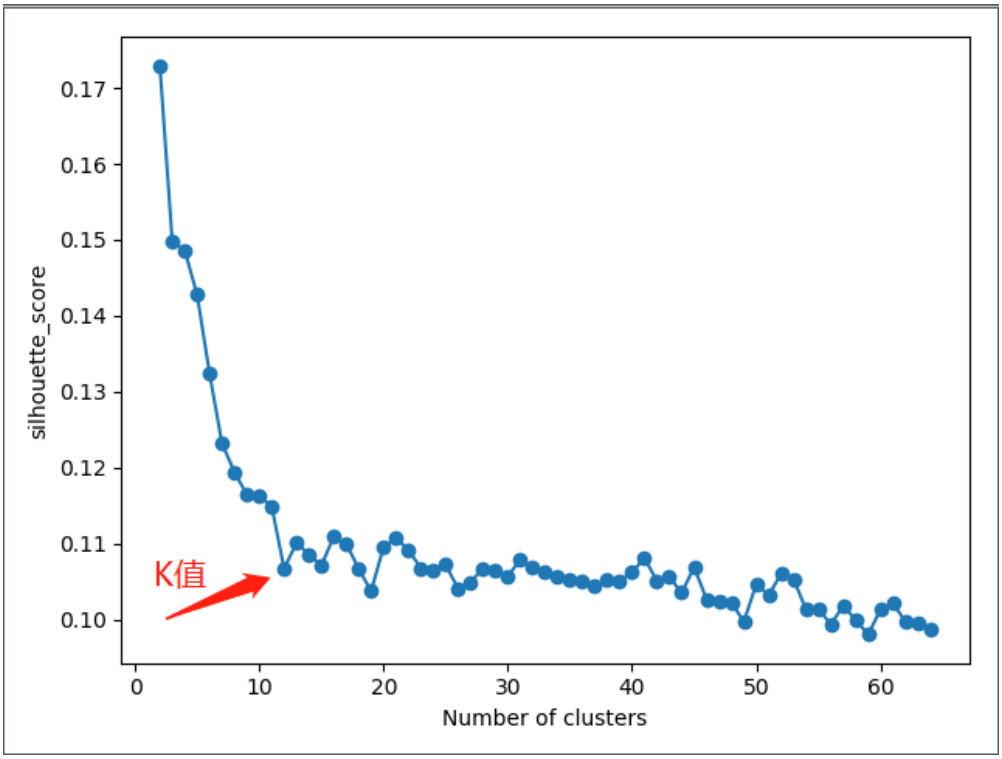

图 2.4 手肘法确定 k 值

当 k 小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故 SSE

的下降幅度会很大，而当 k 到达真实聚类数时，再增加 k 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓，也就是说 SSE 和 k 的关系图是一个手肘的形状，而这个肘部对应的 k 值就是数据的真实聚类数。

2.3.2 算法改进之 K-Means++

K-Means++算法是在 K-Means 算法的基础上进行的改进，是为了弥补 K-Mean 受到初始点选取的对聚类的效果照成的影响，显著的改善分类结果的最终误差。其主要的改进即为初始点的改进，让初始的聚类中心之间的相互距离要尽可能的远，其他步骤与 K-Means 算法步骤一致。

主要做法如下：

①：从数据集中随机选取一个样本点作为初始聚类中心 C_i

②：首先计算每个样本与当前已有聚类中心之间的最短距离，用 $D(x)$ 表示；接着计算每个样本被选为下一个聚类中心的概率 $\frac{D(x)^2}{\sum_{x\in k} D(x)^2}$，最后，按照轮盘法选出下一个聚类中心。

③：重复②，直到选出 K 个聚类中心

④：执行 K-Means 算法的③④步骤。

2.4 LogisticRegression 逻辑回归算法

2.4.1 算法描述

Logistic 回归模型是用概率估计来进行分类的[6]。一个文档分到某个类别中假如看做是一个事件，文档的特征即文档中出现的词语、概念看成影响这个分类事件发生与否的因素。给定一个文档集合，利用回归分析,研究文档特征和文档类别之间的内在关联，从而预测文本文档所属类别。大致的推算思路如下：

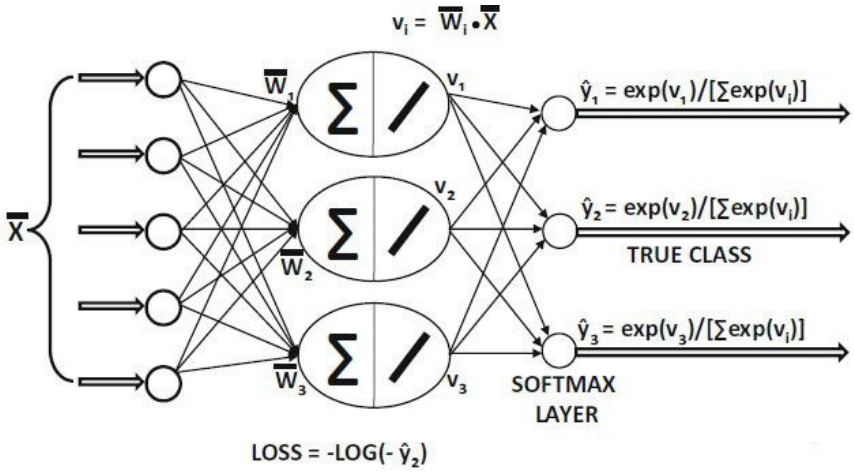

图 2.5　LogisticRegression 模型结构图

①：假设给定 n 个带标签的样本

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_3, y_3), ..., (\vec{x}_n, y_n),$$

其中对于 $\vec{x}_i = (x_1, x_2, x_3, ..., x_D)$ 是一个维度为 D 的特征向量，对于标签 $y_i \in \{1,2,3,...,K\}$ 表示 K 个类别中的一类，并且每个类别对应一个权重向量，并且总共对应 K 个权重向量，假设第 i 个类别权重为 $\vec{w}_i$，必然模型输出的是一个概率分布，利用 softmax 函数表示样本点 $\vec{x}_i$ 属于类别 r 的后验概率 $P(r \mid \vec{x}_i)$ 的表达式如公式 2.21 所示：

$$P(r \mid \vec{x}_i) = \frac{\exp(\vec{w}_r \cdot \vec{x}_i)}{\sum_{j=1}^{K} \exp(\vec{w}_j \cdot \vec{x}_i)} \tag{2.21}$$

②定义损失函数，利用对数似然建立函数表达式如 2.22 所示

$$\begin{aligned} \log L &= -\sum_{i=1}^{n} \log P(y_i \mid \vec{x}_i) \\ &= \sum_{i=1}^{n} \{-\vec{w}_{y_i} \cdot \vec{x}_i + \log \sum_{j=1}^{K} \exp(\vec{W}_j \cdot \vec{x})\} \end{aligned} \tag{2.22}$$

③利用梯度下降法迭代计算权重，此处假设 $v_{i,r} = \vec{w}_r \cdot \vec{x}_i$，表示特征向量 $\vec{x}$ 与第 r 类别对应权重 $\vec{w}_r$ 的向量积，得到梯度表达式如公式 2.23 所示

$$\frac{\partial \log L}{\partial \vec{w}_r} = \sum_{i=1}^{n}\sum_{j=1}^{K} \frac{\partial \log L_i}{\partial v_{i,j}} \frac{\partial v_{i,j}}{\partial \vec{w}_r}$$
$$= \sum_{i=1}^{n} \begin{cases} -\vec{x}_i(1-P(r|\vec{x}_i)) \\ \vec{x}_i P(r|\vec{x}_i) \end{cases} \tag{2.23}$$

④利用梯度表达式进行迭代计算概率。

2.4.1 逻辑回归中的正则化

正则化是为了防止过拟合的现象的发生，所谓的过拟合是指为了得到一致假设而使假设变得过度严格。

对于我们的分类任务，可能会训练出一个高阶多项式可能很好地拟合训练集，能够几乎拟合所有的训练数据，但这函数太过庞大，变量太多，如果没有足够多的数据去约束这个变量过多的模型，此时就会发生一种现象，模型在训练集（已知数据）上表现非常好，但是对于未知的数据表现得非常糟糕。这对于一个分类器而言是十分糟糕的。

故正则化在代价函数后添加正则项，保留所有的特征，减少参数θ的大小，能够使得获得的边界函数更加平滑。更好的模拟现实数据，而非训练样本仅此而已[7]。

2.5 相似性度量

在自然语言的文本分类中，模型需要对文本进行相似度计算，文本之间的相似度直接影响到了模型聚类的效果，现在公认的聚类模型有很多种，采用不同的相似性度量方法，聚类的效果自然不同。

文本聚类的主要原理就是利用对象与对象之间的相关或不相似程度来进行聚类，它直接反应的就是数据之间的距离，这种距离可以用于衡量数据之间的不相关或不相似程度，其中文本在文本聚类时，大多数情况下都是把距离进、相似度高、相关度高的数据集中到一个类别中，把距离远、相似度低、相关度低的数据集中到另一个类别中[8]。

对于文本之间的相似性，可以使用两个向量之间的距离来表示，例如闵可夫斯基距离(Minkowski Distance)，对于两个向量$a(x_{11}, x_{12}, ..., x_{1n})$和$b(x_{21}, x_{22}, ..., x_{2n})$

而言，其闵可夫斯基距离可表示为：

$$dist_{12}=\left(\sum_{k=1}^{n}\left|x_{1k}-x_{2k}\right|^{p}\right)^{\frac{1}{p}} \tag{2.24}$$

其中 p 为常数，对于 p 而言有以下几种情况：

①绝对距离

当 P 等于 1 时，得到绝对距离，也成为曼哈顿距离（Manhattan distance）。

②欧式距离

当 p 等于 2 时，得到欧几里德距离（Euclidean distance）距离，就是两点之间的直线距离（以下简称欧氏距离）。欧氏距离中各特征参数是等权的。

③切比雪夫距离

令 $p\to\infty$，得到切比雪夫距离。

同时也可以用余弦相似度来度量两个文本之间的相似度，余弦相似度利用向量空间的两向量的夹角的余弦值来表示，相比距离度量，余弦相似度更加注重两个向量在方向上的差异，一般情况下，用 Embedding 得到两个文本的向量表示之后，可以使用余弦相似度计算两个文本之间的相似度。

对于两个向量 $a(x_{11},x_{12},...,x_{1n})$ 和 $b(x_{21},x_{22},...,x_{2n})$ 而言，其之间的夹角余弦为：

$$\cos(\theta)=\frac{a\cdot b}{|a|\cdot|b|}=\frac{\sum_{k=1}^{n}x_{1k}\cdot x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k}^{2}}\cdot\sqrt{\sum_{k=1}^{n}x_{2k}^{2}}} \tag{2.25}$$

普遍来说大部分文本距离利用余弦相似度进行衡量。

3 问题解决思路及流程

3.1 问题一

在处理网络问政平台留言时，大部分电子政务系统仍处于由人工依据经验为留言进行处理分类，存在着工作量大，效率低的问题，故根据附件一、附件二为附件二的一级标签构建分类模型。

3.1.1 解决思路及算法流程

此问题为典型的分类问题，为 NLP 领域最经典的使用场景之一，利用分类器对目标文本进行分类，其中分类方法主要分为两种，一种为计语传统机器学习的文本分类，如贝叶斯、SVM 等，另一种为基于深度学习的文本分类如 FastText 文本分类，Text-CNN 文本分类。本文在此利用基于传统机器学习的文本分类模型 LogisticRegression 逻辑回归进行分类。LogisticRegression 逻辑回归具有计算代价不高、容易使用和解释的优点。

在进行分类器分类之后，使用 F-Score 对分类模型进行评价。其中 F-Score 是分类器的常见评价方法，用来评估分类模型的好坏，其主要的公式如 3.1 所示：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{3.1}$$

其中 P_i，R_i 分别为第 i 类的查准率和查全率。其中 $P=\frac{a}{a+b}$，$R=\frac{b}{a+c}$ 其中：a 表示分类正确的文档数；b 表示分入该类而实际不属于该类的文档数；c 属于该类的文档数而判为不属于该类别的文档。

3.1.2 数据预处理并分词

（1）数据处理

本问题的数据预处理问题主要是针对附件二中的数据进行预处理，数据预处理是指在主要的数据处理以前的一些数据处理。实际上，收集的部分数据一般为不完整、不合格的，对于这些数据不能直接对这些数据进行挖掘，否则挖掘效果不理想并且会对挖掘结果产生巨大的偏差，对于数据二而言其主要的数据预处理

为以下步骤：

①无用词替换，由于留言的不规范性，并非经过精心思考和审查，所以存在着很多符号、噪声等，例如图片表情、文字表情、空格、换行符等，因此需要对这些符号进行清除，利用 python 的 re 模块对留言主题、留言内容中存在的符号进行替换。

②缺省值处理，对于问题一而言，主要的任务是分类，那么对于数据而言，一级分类、留言主题和留言详细字段不能为缺省值，因为该字段作为重要的建模特征，故对于该字段为缺省值的数据进行删除。

③重复值处理，对于留言数据而言，出现重复数据一定为数据出错，对于文本数据而言，多特征的数据对模型的建立毫无意义，故对此数据进行删除。

通过对以上的分析，利用 python 编程对数据进行处理，运行 preprocessing 目录下的 preprocessingFileTwo.py 文件得到 data 目录下的 datafile2.csv

（2）jieba 分词

对文本数据进行文本分析、处理最关键的一步是分词，中文分词的效果直接影响到后续文本的特征值的提取以及特征向量的生成，经过分词器分词后的文档会变成词的列表，更进一步的方便文本特征的提取。

对于英文而言，每一个单词之间有间隔，十分容易分词，而对于中文而言，词语与词语之间并没有太明显的分隔符，因此必须要有一款工具对中文文本进行分词，本文采用堪称：做的最好的中文分词组件，jieba 分词器进行分词。

Jieba 分词有功能丰富、提供多种编程语言实现、使用简单等优点。在功能上，并不只是有分词这一功能，还提供了很多在分词之上的算法，如关键词提取、词性标注等，能在自然语言处理中取得较好的总体效果。

3.1.3 获取文本词向量

对于将文本预处理、分词之后的数据需要进行 Embeding，转换为数值矩阵，可采用的方法有 TF-IDF、Word2Vec、One-Hot 等方法，由于数据量较大，不宜采用 TF-IDF 模型，矩阵较为稀疏，故本文采用的是 Word2Vec。

通过导入每一篇文章的留言内容进行训练 word2vec 模型，构建一个与 word2vec 维度一致的矩阵，再将切分后的数据进行判断是否存在于模型中，对

于存在的词语为词语计数器加一，并将该词语在 word2vec 模型中的维度向量相加，遍历一片文章后取矩阵的均值，作为该篇文章的文档特征向量矩阵。

3.1.4 建模结果

通过以上方法经过数据统计，附件二中的分词类别一共有 7 种，将官方给定的数据分为训练集和测试集，训练集的目的是为了训练模型，而测试集是为了验证模型的有效性，从而来调整超参数的数量。

在传统的机器学习中，科学的安排训练集和测试集合的比例很重要，故此处利用普遍的比例，8:2 的比例可以很好的进行训练和测试。由表 3.1 可以很好的看得出来，每个类别之间的数量间隔较为普遍，没有出现类别之间间隔较大的情况。

表 3.1 文档总数据

类别	数据/篇
1.交通运输	613
2.劳动和社会保障	1969
3.卫生计生	877
4.商贸旅游	1215
5.城乡建设	2009
6.教育文体	1589
7.环境保护	938

利用 python 中的 sklearn 库可以很好的将数据进行切分，如表 3.2 所示，将数据按类别进行合理的分割，不会出现个别类别为空或者数量小的现象。

表 3.2 文档数据分布

类别	训练/篇	验证/篇
1.交通运输	491	122
2.劳动和社会保障	1575	394
3.卫生计生	702	175
4.商贸旅游	972	243
5.城乡建设	1607	402
6.教育文体	1271	318
7.环境保护	750	318

通过不断的对 word2vec 模型的参数调整，最终调整为词向量的维度 240，运行附件中问题一目录下的 Word2Vec.py 文件，对数据进行训练取得了 88.8%的正确率即 F1 值，利用测试集进行测试得到以下结果：

表 3.3　测试数据精确度,召回率,F1 值

类别	Precision	Recall	F1-score	数据量
交通运输	0.79	0.78	0.79	122
劳动和社会保障	0.91	0.95	0.93	394
卫生计生	0.92	0.89	0.90	175
商贸旅游	0.86	0.81	0.83	243
城乡建设	0.83	0.87	0.85	402
教育文体	0.93	0.92	0.92	318
环境保护	0.94	0.88	0.91	188
平均值	0.88	0.87	0.88	1842

由表三可以明显的看出分类结果较为准确，由于 word2vec 会考虑上下文、高效性的优点，使得分类效果更加明显。其中查准率(Precision)和查全率(Recall)的值几乎一致，且 F 值也达到了 88%的效果。

其中劳动和社会保障、卫生计生和教育文体的文本分类效果比其他的效果好一点，这可能是因为这几个类别的独立性更强，这些类别的特征向几乎都在本类，在其他类别上显示的更少的缘故。而其他的类别，例如交通运输则效果较差，这可能是该类别的特征更有可能包含了其他类别的特征，例如：A 市交警今日例行检查，维护道路秩序，保障市民的正常出入。那么这一篇类别可以被理解为是交通运输也可以被理解为是城乡建设，而建立的语料库只是做了但类别的判断，因此分类进行误判是有可能的，那么这本质上与语料库的质量密切相关。

3.2 问题二

对于在某一时间内的群众反应的问题称为热点问题，针对数据，将某一时段内的反应的特点地点/人群的留言进行归类，并制定合理的评价指标给出热度指数，并进行排行，提升部门分服务效率。

3.2.1 问题分析及算法流程

问题二意在挖掘热点问题以及制定热点问题的热度指标评价并做出热度的排名。其中的热点问题即为某一时段内群众集中反映的某一问题，并且根据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。

①那么针对留言归类而言，就需要将留言数据进行无监督的文本聚类。文本

聚类主要是依据著名的聚类假设：同类的文档相似度较大，而不同类的文档相似度较小。那么便可根据 Word2Vec 算法对留言数据进行量化，利用基于语义的词向量模型比基于词频统计的算法模型效果更加，从而进一步提高聚类效果，再利用 K-Means++聚类模型进行聚类。

②针对某一时间段或特定人群为基点进行聚类问题，在 NLP 领域存在着命名实体识别(NER)可以对其进行处理，经过对中文句子的观察以及留言数据的探究发现，大部分地名或者特定人群而言皆为名词，故以此进行解决，通过提取留言主题的名词作为类别的地点/人群。

③对于热度指标的制定而言，由于需要对类别中的留言进行排行以及对每一类的热度进行排行，两者的作用范围不同，自然指标也不同，故制定类别内的热度指标以及类别之间的热度排行分别进行评价。

对于建模方法的采用，可以使用 AHP(层次分析法)，并采用适当的方法对热度指标进行加权，其中加权方法包括主观权重评估方法和客观权重评估方法。

主观权重评估方法通过基于专家经验和意见的主观判断的定性方法获得权重值，然后对层次分析法，指标权重法和效能系数法等评价指标进行进一步评价。客观加权评估方法将根据评估指标之间的相关性或每个指标的变化系数来确定权重以进行综合评估。

权重分配是否合理在评价结果的科学合理性中，起着至关重要的作用： 当一个因素的权重发生变化时，极有可能会影响整个评价结果。所以，权重的赋 值需要要做到客观。这就会要求我们寻找合适的权重去确定方法。

3.2.2 数据处理

根据问题一的挖掘以及处理的方法进行借鉴之后，根据问题的具体要求，选取问题一的无用词替换、缺省值处理以及重复值处理，并且在这之上进行添加以下处理条件。

①缺省值，对于该数据集而言并且根据后面的热度分析，对于除留言编号外的其他字段均为必需字段，故对于除留言编号外的字段为空的数据均采取删除的措施。

②时间字段的处理，对于时间字段，由于对后面的热度排名有影响，故对该

字段进行统一化，对格式进行规范化，以便后面更好的处理。

通过对以上的分析，利用 python 编程对数据进行处理，运行 preprocessing 目录下的 preprocessingFileThree.py 文件得到同目录下的 datafile3.csv，得到空值数据导出到 emptyData 目录下的 file3.csv，并发现一条空数据如表 3.4 所示：

表 3.4 附件二空值表

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
189856	A00073717		2019/7/3 11:53	A 市保....	0	1

由此证明数据预处理算法可用，并且证明数据纯度在完整性上较为全面可以进行模型建立。

3.2.3 文本聚类

（1）词向量获取

根据官方提供的数据统计，附件三共有 4326 条数据，数据量中等，在词向量的获取方面上考虑 TF-IDF 和 Word2Vec 两种模型进行构建，利用 Jieba 分词器对附件的留言主题进行分词后文档共有 7391 个词语，利用 TF-IDF 构建矩阵后出现的“维度爆炸”、“维度稀疏”的现象（如图 3.1 所示），图中很明显看的出来几乎大部分向量都为 0，故此题采用 Word2Vec 对留言的主题进行构建模型。

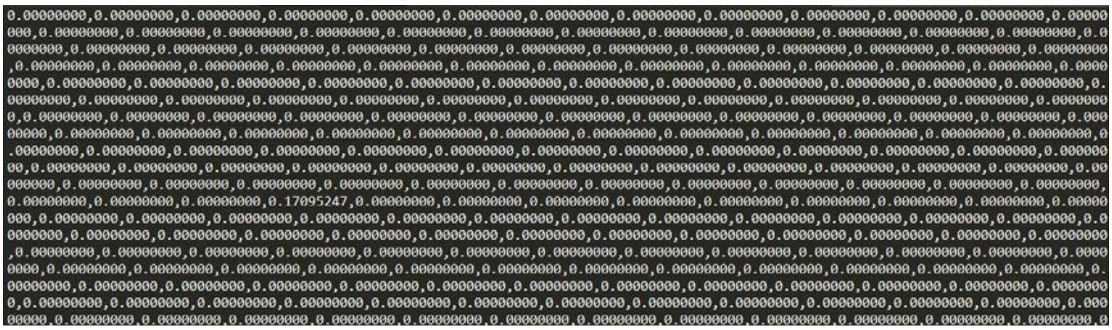

图 3.1　TF-IDF 矩阵稀疏图

（2）K-Means++算法聚类

由于 K-Means 算法的聚类结果严重依赖与初始簇中心的选择，如果初始簇中心选择不好，就会陷入局部最优解，故此处采用 K-Means 的改进算法 K-Means++算法对数据进行聚类。利用 python 的 sklearn 中的 kmean 对数据进行聚类，并利用“手肘法”来确定 K 的数值。

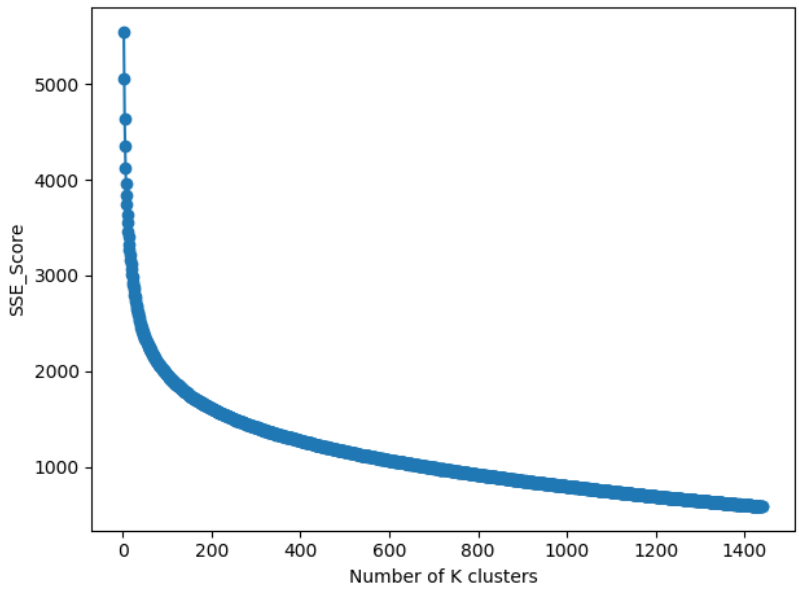

图 3.2 手肘法聚类效果图

显然，肘部对应的 k 值为 140(曲率最高)，故对于这个数据集的聚类而言，最佳聚类数应该选 140，运行附件问题二目录下的 KMeans.py 可确定 K 值，再根据观察，运行同目录下的 KMeanTest.py 得到文件 KMean.csv 为下一步做铺垫。

（3）聚类异常值、离群值处理

由于在 word2vec 模型的训练当中存在着某些词的向量不存在的情况，就会出现一些异常值，这些异常值表现为某些留言主题的词向量皆为 0（如图 3.3），对于这些数据进行文本聚类没有什么意义，即使聚类出来了，反而影响整体效果，则采取删除的措施。

67	0.03023	0.008323	0.019794	-0.0236	-0.02417	0.13049	-0.07093	-0.12224	0.270191	-0.06616	-0.03729	-0.21327	0.056358	0.11871	-0.00751
68	0.029969	0.005431	0.018963	-0.0279	-0.02441	0.127987	-0.0709	-0.11784	0.267131	-0.0667	-0.03574	-0.2156	0.057413	0.116919	-0.00829
69	0.03025	0.006096	0.019172	-0.02552	-0.02318	0.128966	-0.07174	-0.11998	0.270316	-0.06691	-0.03505	-0.21702	0.057036	0.117336	-0.00783
70	0.031985	0.005612	0.018025	-0.0272	-0.02444	0.129597	-0.07192	-0.1203	0.269894	-0.06663	-0.03577	-0.21988	0.055439	0.117476	-0.00718
71	0.030867	0.006767	0.019689	-0.02651	-0.0236	0.130105	-0.07128	-0.1198	0.270292	-0.06767	-0.03498	-0.21651	0.056146	0.118597	-0.00622
72	0.032798	0.004334	0.019088	-0.02827	-0.0284	0.127688	-0.07064	-0.11888	0.269253	-0.06949	-0.03534	-0.21971	0.062353	0.118889	-0.00511
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
82	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

图 3.3 词向量异常值

对于异常数据，在源数据中截取后发现，异常数据为 67 条(表 3.5)，总体上数量不多，不影响整个问题的解决，删除率为 1.5%,数据较为完善。

表 3.5 词向量异常数据表

留言编号	留言用户	留言主题	留言时间
190087	A00052076	A 市新奥燃气服务态度差	2019/7/22 0:01
190192	A00094992	试管做出基因缺陷女婴，不幸家庭雪上加霜	2019/9/1 22:30
193457	A00074672	A 市骧龙井湾子机动车检测站乱收费	2019/7/29 11:51
193835	A00031421	期盼 A 市星沙 15 路车在上班高峰期能准时发车	2019/5/14 11:08
196072	A00050651	A 市家四水厂乖乖兔电动车商家逃避责任	2019/8/5 18:50
196140	A00081622	A 市国储电脑城负一层存消防隐患	2019/3/19 16:56
...	...	.....	....
321736	A9992521	A 市能不能提高医疗门诊报销范畴	2019/6/12 8:23

在聚类完成之后，即使是聚成了一类，仍然会出现偏离聚类中心的点，这些点称为离群点。

对于离群点的判断采用正态分布图法[10]，由正态分布的定义(如图 3.4)可以知道，数据落在均值标准差正负一倍的距离内的概率为 68.2%，落在均值标准差正负二倍的距离内的概率为 95.4%，落在均值标准差正负三倍的距离内的概率为 99.6%。

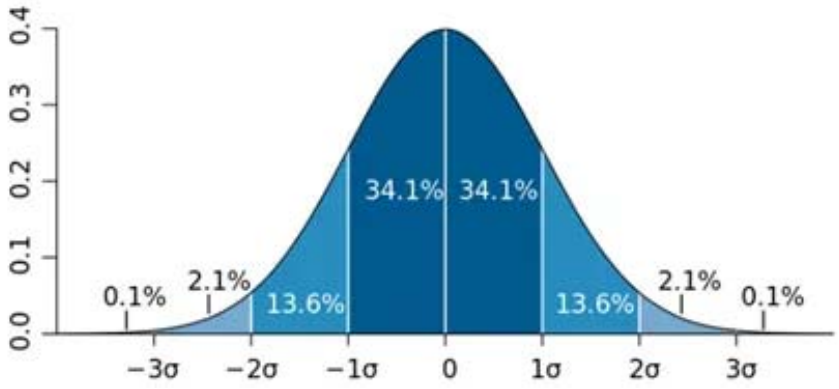

图 3.4 正态分布图

因此可以认为落在均值标准差正负二、三倍的范围内的数据为异常数据，也成为离群值(表)。对于这些离群点对于后续的排名、简介抽取等皆造成影响，对与这些数据也采取删除的措施。

表 3.6 异常值评判标准

判断标准	结论
$\bar{x}-2\sigma<x<\bar{x}+2\sigma$	异常点
$\bar{x}-3\sigma<x<\bar{x}+3\sigma$	极端异常点

聚类效果部分数据如下表所示，整体效果还算可以。

表 3.7 部分聚类效果

留言编号	留言用户	留言主题	类别
360114	A0182491	A 市经济学院体育学院变相强制实习	1
360113	A3352352	A 市经济学院强制学生外出实习	1
360107	A0283523	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	2
360108	A0283523	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理！	2
289588	A909183	投诉 A 市伊景园滨河苑开发商	3
289950	A00044759	投诉 A 市伊景园滨河院捆绑销售车位	3
....	....	.....	....

3.2.4 特定人名/地名获取

根据对留言数据的观察，留言数据的留言主题为留言概括内容，对于特定人名/地名的提取，通过对类内的留言主题进行分词后抓取名词，然后再统计类内的词频，将词频进行排序后提取排名前 8 个数据作为特定人名/地名代表类别。

对于热点问题表所需要的问题描述而言，由于每一类别的问题数量很多，需要从中进行语义分析，然后对留言内容进行抽取出关键句能符合留言主题的语义，故此处利用 TF-IDF 算法，分别对每一类别的留言内容与抽取的特定人名/地名进行相似度匹配，然后进行排序，对相似度最高的一类别进行关键句抽取，此处采用的是 TextRank 算法，利用 python 的 SnowNLP 库，SnowNLP 可以方便进行处理中文文本内容，进行中文分词、词性标注、提取文本摘要等。

3.2.5 层次分析法热度排名

对于层次分析法（AHP）而言，是指能够将决策分为有些学校的目标、准则、方案等层次，将数据进行量化，给出基于权重的多目标的综合评价方法[11]，针对

第三问而言，可以使用层次分析法，构建热度评价指标，由于需要对类别中和类别进行分别的热度排名，其相对的热度指标也自然不同，例如对于类别的留言数而言，留言数对类别之间的评价起到了至关重要的作用，但是在我们类中的热度评价却无法作为热度评判指标，故在此分别对簇内和簇之间构建层次分析模型、热度评价指标。

（1）簇内层次分析法模型构建及构建指标

对于簇内而言，主要针对留言客观评论数、留言内容以及留言时间构建模型。其中一个类别中的评价指标如表 3.5 所示。

表 3.8 类中评价指标表

一级指标	评价标准
留言客观评论数	点赞、反对数
留言内容	留言内容平均相似度
留言时间	留言时间距平均时间程度

标准一留言点赞数，在分为一簇的簇内，每一篇的留言热度固然与点赞数、反对数相关，间接着反应了留言的访问量，并直接影响了留言的热度性，得分标准计算公式如下所示，为确保数据的合理性，将最终数据归一化，并转化为百分制。

$$y_i = \frac{\partial_i + \lambda_i}{\sum_{k=1}^{n}(\partial_k + \lambda_k)} \times 100 \ \ (i = 0,1,2,.....n)(\partial\text{为点赞数},\lambda\text{为反对数})$$

标准二留言内容，在每一簇中，留言的热度自然与留言内容相关，设想前一步聚类的过程中，存在着稍微噪声的数据没有去除，那么在这一簇中自然与其他中心簇附件的留言相似度较低，那么在定义的这一簇内，自然热度降低，得分标准计算公式如下所示，其中ζ为留言相似度的平均值，循环簇内的每篇留言，利用 TF-IDF 算法，计算与其他留言之间的相似度，并取平均值，最后归一化，取百分制。

$$y_i = \frac{\zeta_i}{\sum_{i=k}^{n}\zeta_k} \times 100(i = 0,1,2,...,n)(\zeta_i\text{为簇内第}i\text{篇留言平均相似度})$$

标准三留言时间，在每一簇中，存在着分布散乱的数据，其中的一个指标是留言时间，设想一个类别中留言发布的主要时间集中在一个时间段，那么偏离这个时间段越远的留言固然热度降低，好比我们生活新闻中，“冠状病毒”的新闻报道固然每年都有，那么集中爆发的时候热度固然很高，但是没爆发之前的新闻大家固然的关注度会降低，所以将此作为评价指标，计算公式如下：

$$y_i=(1-\frac{\left|\lambda_i-\overline{\lambda}\right|}{\overline{\lambda}})\times 100(i=1,2,3,...,n)(\text{其中}\lambda\text{表示时间点},\overline{\lambda}\text{表示平均时间点},\overline{\lambda}=\frac{\sum_{k=1}^{n}\lambda_k}{n})$$

选取一个类别中的最早时间点，将最早时间点作为原点，计算每一篇留言到此原点的距离，并取平均值，最后再利用每一篇留言的时间点计算与此平均点之间的距离，保障了评价的公平性，最后再百分化。

得到类间的评价指标表(3.6)如下：

表 3.9 类中评价指标得分表

评价标准	得分标准
留言点赞数	$y_i=\frac{\partial_i+\lambda_i}{\sum_{k=1}^{n}(\partial_k+\lambda_k)}\times 100\ \ (i=0,1,2,.....n)(\partial\text{为点赞数},\lambda\text{为反对数})$
留言内容	$y_i=\frac{\zeta_i}{\sum_{i=k}^{n}\zeta_k}\times 100(i=0,1,2,...,n)(\zeta_i\text{为簇内第}i\text{篇留言平均相似度})$
留言时间	$y_i=(1-\frac{\left\|\rho_i-\overline{\rho}\right\|}{\overline{\rho}})\times 100(i=1,2,3,...,n)(\text{其中}\rho\text{表示时间点},\overline{\rho}\text{表示平均时间点},\overline{\rho}=\frac{\sum_{k=1}^{n}\rho_k}{n})$

对于构建完的得分指标，在确定每一评判指标各因素之间的权重时，不能根据人意进行判断，往往不容易被人接受，Saaty 等人提出一致矩阵法，即不把所有因素放在一起比较，而是两两相互比较，对此时采用相对尺度，以尽可能减少性质不同的诸因素相互比较的困难，以提高准确度。

如对某一准则，对其下的各方案进行两两对比，并按其重要性程度评定等级。

为要素与要素重要性比较结果，下表中列出 Saaty 给出的 9 个重要性等级及其赋值。按两两比较结果构成的矩阵称作判断矩阵，其中判断矩阵有以下性质：

$$a_{ij}=\frac{1}{a_{ji}} \tag{3.2}$$

其中判断矩阵元素 a_{ij} 的标度方法如下表（标度法）所示：

表 3.10 标度方法表

因素 i 比因素 j	量化值
同等重要	1
稍微重要	3
较强重要	5
强烈重要	7
极端重要	9
两相邻判断的中间值	2，4，6，8
倒数	因素 i 于 j 的比较 a_{ij},则因素 j 于 i 比较,$a_{ji}=\frac{1}{a_{ij}}$

根据标度法构建的判断矩阵如下：

$$D=\begin{bmatrix}1 & 3 & 5\\ 1/3 & 1 & 2\\ 1/5 & 1/2 & 1\end{bmatrix}$$

再接着求权重方法种的规范列平均法求取权重。首先对判断矩阵列向量归一化得到对应矩阵 E。再对此进行算术平均法求取并得到特征向量 w。

计算步骤[12]：

①D 的元素按列归一化，即求取 $\frac{a_{ij}}{\sum_{k=1}^{n}a_{kj}}$

②将归一化的各列相加；

③将相加后的向量除以 n 即得到权重向量 w。

$$W_i=\frac{1}{n}\sum_{j=1}^{n}\frac{a_{ij}}{\sum_{k=1}^{n}a_{kj}},\ i=1,2,...,n$$

得到归一化、权重向量矩阵如下：

$$E=\begin{bmatrix}0.6521 & 0.6666 & 0.6250\\ 0.2173 & 0.2222 & 0.2500\\ 0.1304 & 0.1111 & 0.1250\end{bmatrix}$$

$$w=\begin{bmatrix}0.6479 & 0.2298 & 0.1221\end{bmatrix}$$

当$CR<0.1$时，认为判断矩阵的一致性在可接受的范围内，若$CR>0.1$时，则判断矩阵不符合一致性要求，则对该判断矩阵进行重新修改批正

表 3.11 随机一致性指标 RI

n	1	2	3	4	5
RI	0	0	0.58	0.90	1.12
6	7	8	9	10	11
1.24	1.32	1.41	1.45	1.49	1.51

并且计算可得$CI=0.0018$，$CR=0.0031$，一致性比例 CR 小于 0.1，故权重值合理，权重表如下表所示。

表 3.12 类中指标权重表

评价标准	权重
留言点赞数	0.6479
留言内容	0.2298
留言时间	0.1221

（2）簇间层次分析法模型构建及构建指标

在进行类别之间的热度排行时，需要另外从每一类中提取指标，作为每一类的特征，并合理考虑留言的特点，制定了评价指标表如下。

表 3.13 类间评价指标表

一级指标	评价指标
留言客观评论数	点赞、反对数总数
留言持续热度	评价留言平均时间间隔
留言重复率	用户留言重复率
留言数	类别留留言总数

对于留言客观评论数而言，评价指标为点赞、反对总数，数值应为类别之间的点赞、反对总数取总和之后归一化并百分化，给出计算公式如下:

$$y_i=100\frac{\sum_{n=1}\partial_{in}+\lambda_{in}}{\sum_{k=1}\sum_{n=1}\partial_{kn}+\lambda_{kn}}(i=1,2,3,...,n)(\partial\text{为赞同数},\lambda\text{为反对数})$$

对于持续热度而言，一个类别中的留言带有留言时间，假设两个类别中，均

有两条数据，一个类比中时间相差一天，另一个类别中留言时间相差一年，那么固然相差一天的类别热度比相差一年的类别热度高，故给出此标准，并给出量化公式：

$$y_i = 100\frac{\overline{\rho_i}}{\sum_{n=1}\overline{\rho_n}}(i = 1,2.3....,n)(\text{其中}\overline{\rho_i} = \frac{\sum \rho_k - \rho_0}{n}, \rho_0\text{为最早的时间点})$$

计算每个类别的平均留言时间相差数，并且取均值后对每一个类别取均值并百分化得出评分。

对于用户重复率而言，热度评价与类别中的用户数量自然相干，假如两类别中各留言数量均为 100，且 A 类中 95 篇重复用户留言，B 篇中 5 篇重复用户留言，那么相对于热度而言，B 类别中的热度固然比 A 类别热度高，故选用户重复率作为指标，并给出数据量化公式如下：

$$y_i = 100(1 - \frac{\omega_i}{\sum_{k=1}\omega_k})(i = 1,2,3,...,n, \omega_i\text{为第}i\text{类的留言重复率})$$

对于热度而言，最最基础的就是留言数量，不考虑其他因素，留言数量的多少直接影响热度的评判，故给出此标准并给出数据量化公式：

$$y_i = \frac{100\phi_i}{\sum_{k=1}\phi_k}(i = 1,2,3,...,n)(\phi_i\text{表示在第}i\text{类中的留言数量})$$

以上位各个指标的具体解释，并给出汇总表如下：

表 3.14 类间评价指标得分表

评价指标	得分标准
留言点赞反对数	$y_i = 100\frac{\sum_{n=1}\partial_{in} + \lambda_{in}}{\sum_{k=1}\sum_{n=1}\partial_{kn} + \lambda_{kn}}(i = 1,2,3,...,n)(\partial\text{为赞同数}, \lambda\text{为反对数})$
留言平均时间间隔	$y_i = 100\frac{\overline{\rho_i}}{\sum_{n=1}\overline{\rho_n}}(i = 1,2.3....,n)(\text{其中}\overline{\rho_i} = \frac{\sum \rho_k - \rho_0}{n}, \rho_0\text{为最早的时间点})$
用户重复率	$y_i = 100(1 - \frac{\omega_i}{\sum_{k=1}\omega_k})(i = 1,2,3,...,n, \omega_i\text{为第}i\text{类的留言重复率})$

留言数量	$y_i = \frac{100\phi_i}{\sum_{k=1}\phi_k}(i = 1,2,3,...,n)(\phi_i$表示在第$i$类中的留言数量)

同上，利用标度法构建判断矩阵如下:

$$D = \begin{bmatrix} 1 & 3 & 5 & 1/3 \\ 1/3 & 1 & 2 & 1/5 \\ 1/5 & 1/2 & 1 & 1/5 \\ 3 & 5 & 5 & 1 \end{bmatrix}$$

构建列向量归一化、权重矩阵

$$E = \begin{bmatrix} 0.2205 & 0.3157 & 0.3846 & 0.1923 \\ 0.0735 & 0.1052 & 0.1538 & 0.1153 \\ 0.0441 & 0.0526 & 0.0769 & 0.1153 \\ 0.6617 & 0.5263 & 0.3846 & 0.5769 \end{bmatrix}$$

$$w = \begin{bmatrix} 0.2783 & 0.1120 & 0.0722 & 0.5374 \end{bmatrix}$$

当$CR < 0.1$时，认为判断矩阵的一致性在可接受的范围内，若$CR > 0.1$时，则判断矩阵不符合一致性要求，则对该判断矩阵进行重新修改批正

并且计算可得$CI = 0.0440$，$CR = 0.0489$，一致性比例 CR 小于 0.1，故权重值合理,权重表如下表所示。

表 3.15 类间评价权重表

评价标准	权重
留言点赞数	0.2783
留言平均时间间隔	0.1120
留言重复率	0.0722
留言数量	0.5374

通过以上合理的构建权重表，比人为的确认权重表具有很大的优点，更加合理的保障了热度值的公平性以及保障性。通过运行附件中问题二目录下的 RankScore.py 可以获得通过计算得分后的排名数据等。

3.2.6 小结

聚类为无监督学习，合理的选择聚类的类数至关重要，类数影响到了整个问题的后续研究以及解决，本文利用手肘法确认通过类数后进行聚类，得到一个较好的结果，并利用相似度度量方法进行估算，得到每一项指标的分值，合理的赋予分值权重能够保障数据的公平性，更加的保障了解决的可行性，本文利用层次

分析法合理构建得到权重矩阵，并通过一致性校验，最终得到一个合理的权重，赋予指标项，得到合理的结果。

3.3 问题三

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合理的评价方案。

3.3.1 问题解决思路

“智慧政务”的本质是提升政府的管理水平和施政效率，那么针对每一个留言政府的回应应该给一套评价标准，对政府人员的办公效率进行评估，本题主要从、回复的相关性、时效性、完整性、可解释性四个方面对质量进行评估对于每个方面，均采用量化的形势构建层次分析模型，确定权重，并且将评价等级分为很好，良好，一般，差，很差共五个等级来对答复的质量进行评价[13]。

3.3.2 数据处理

对于附件四而言，借鉴以上问题二、三的数据处理方案，在缺省值、数据噪声方面上做出改变外其余不做出改变。

①缺省值，对于本问题而言，无关字段有留言编号、留言用户，其余不能为缺省值，故对于其余字段为空的数据进行删除。

②噪声数据，由于需要对答复质量的完整性做出评估，答复质量的完整性依靠每个句子，那么需要对每个句子进行分句，故不能对“！”，“.....”，“，”，“。”等符号做出删除，所以在计算完整性的时候不对数据进行噪声处理，而对于质量的时效性、相关性、可解释性则需要进行噪声数据处理。

根据以上分析，利用 python 编程对数据进行处理，运行 preprocessing 目录下的 preprocessingFileFourFirst.py 文件得到同目录下的 datafile41.csv 作为时效性、相关性、可解释性的实验数据，运行同目录下的 preprocessingFileFourSecond.py 文件得到同目录下的 datafile42.csv 作为完整性的实验数据。

3.3.3 质量的相关性评估方案

相关性指的是数据之间的相关程度，对于答复质量的相关性而言，如果留言内容与答复内容的相关性高，那么其语义上必然相近，故在此可以采用基于word2vec 模型的语义相似度计算评估。

先对数据进行模型训练生成 word2vec 模型，并对每一篇留言 id 的留言内容与答复内容进行计算相似度，先对留言内容、答复内容进行获取文本均值词向量矩阵(图 3.3)。然后对俩文本进行相似度度量，采用的方法为余弦相似度计算，最后对整个数据集进行均值化，得到的均值相似度百分化作为该指标的得分(图 2)，其主要的算法流程图如下：

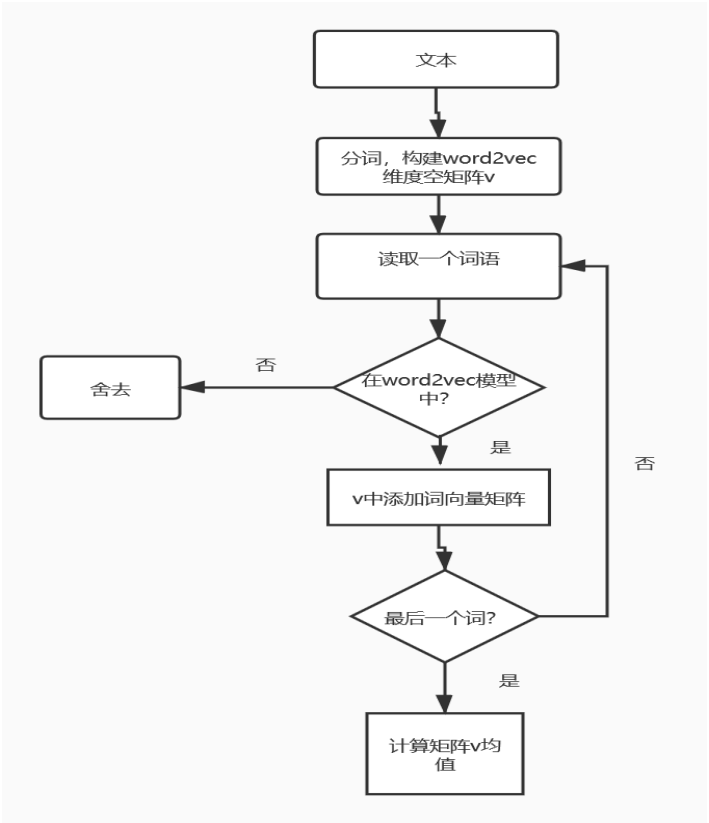

图 3.5 文本的词向量矩阵均值算法图

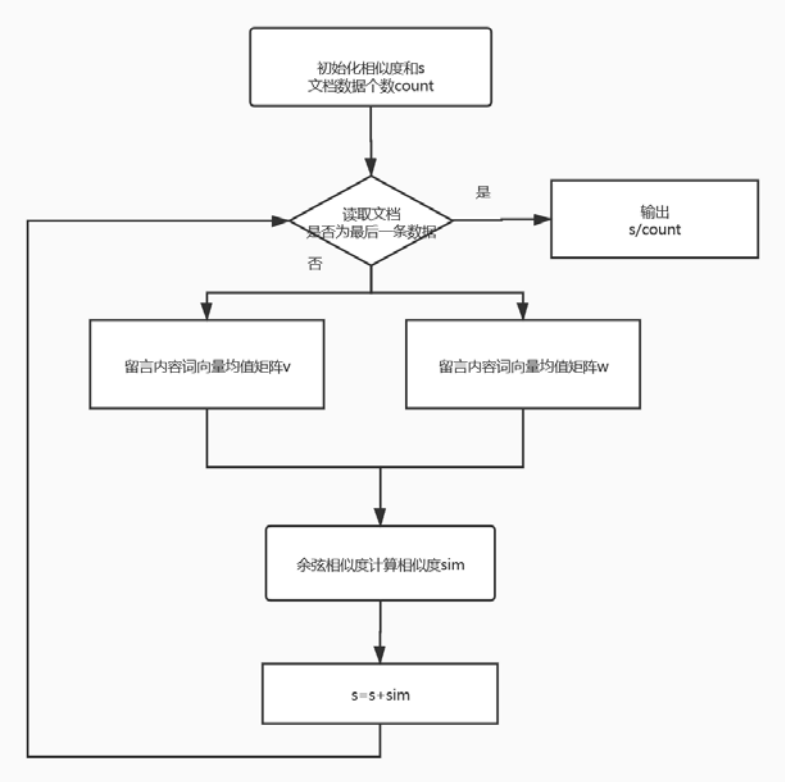

图 3.6 相关性计算分数算法

通过不断遍历每个留言与答复得到相似度，最后取均值得到文档的相关性指数。

3.3.4 质量的时效性评估方案

对于政府的回复质量的时效性而言，可以通过政府的回复时间距离留言时间的间隔来进行评价，由于每个省份、市区政府留言的答复时间不统一，有 3 天、7 天、15 天甚至一个月的也有，本文选取中间值 15 天作为答复极限值，通过计算时间间隔来计算得分，对于 15 天以及 15 天之后的留言答复的单条留言质量的时效性取 0 分，对于 15 天之前的答复对其进行距离化并百分化得到分数。其主要算法流程图如下：

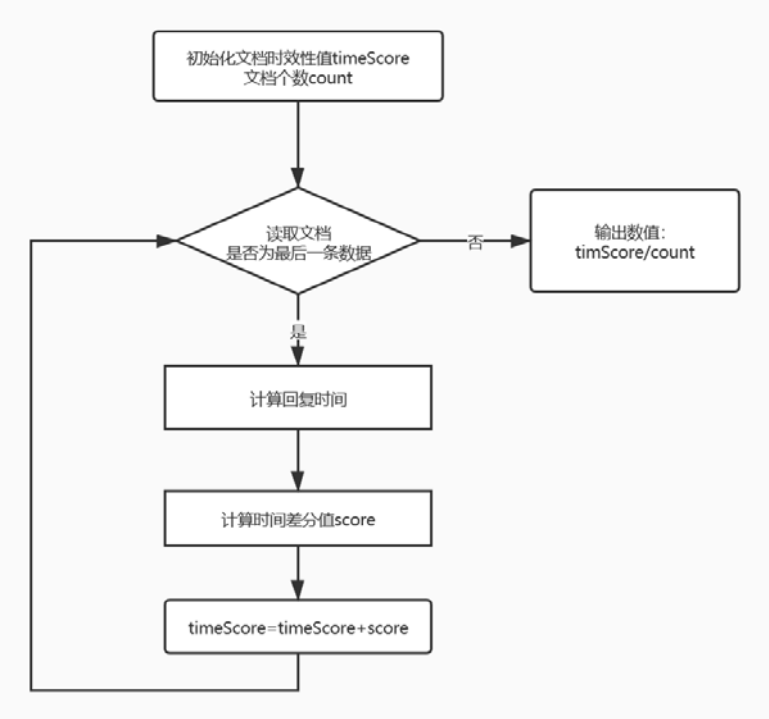

图 3.7 时效性算法流程图

3.3.5 质量的完整性评估方案

完整性是指信息具有一个实体描述的所有必需的部分。不完整的数据对数据分析会产生影响，对于每一条答复内容而言，如果每个句子的结构都为完整句子，那么该答复内容为完整的，由于中文的结构复杂，且运用场景不用则句子的结构固然不同，故此处选取句子的最基本组成成分主语、谓语、宾语作为评价指标。

其中对于利用 Stanford CoreNLP 官方 API 对每个句子进行成分分析，对于主语和宾语而言其主要是根据在句中的语法成分来判断，而针对谓语而言，主要是根据词性进行判断，其代表表格如下：

表 3.16 主语结构代表表

主语				
nsubj	nsubjpass	subj	npsubj	xsubj
名词主语	被动的名词主语	主语	被动型主语	x 主语

表 3.17 宾语结果代表表

宾语			
pobj	obj	dobj	iobj

介词宾语	宾语	直接宾语	非直接宾语

表 3.18 谓语结果代表表

谓语						
cop	auxpass	infmod	prt	VP	VV	VRD
动词	被动词	动词	动词短语	动词短语	动词	动补复合词

对于缺少的任一成分的句子进行扣分，每缺少一个成分扣除 33.3 分，然后对留言内容百分制，循环所有数据集后再进行平均化，主要的算法流程图如下：

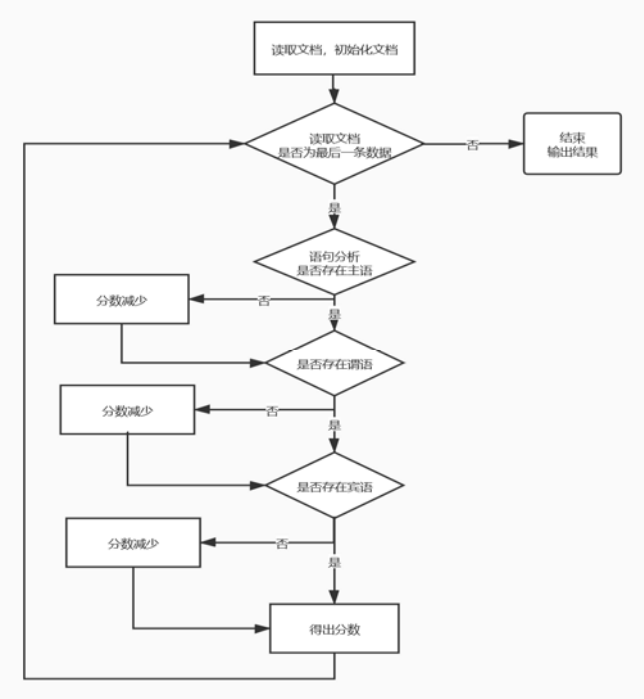

图 3.8 完整性算法流程图

3.3.6 质量的可解释性评估方案

可解释性，也称为可读性，旨在衡量政府“答复内容”的信息对公众诉求问题的解决程度，是政府网络回应效度的最重要的指标。则对于答复文本而言，将写信词频比例与回信词频比例进行皮尔逊相关系数(Pearson correlation coefficient）检验，相关系数越大表示政府回应越具针对性，那么其答复质量的可解释性自然高。

皮尔逊相关系数(Pearson correlation coefficient），又称皮尔逊积矩相关系数（Pearson product-moment correlation coefficient，简称 PPMCC 或 PCCs），是用于度量两个变量 X 和 Y 之间的相关（线性相关），其值介于-1 与 1 之间，其计

算公式如下：

$$k=\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}$$
$$=\frac{E[(X-\mu\alpha)(Y-\mu y)]}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \tag{3.3}$$

其中 E 为数学期望值，cov 表示协方差

当系数 k>0 时，表明两个变量正相关；

当系数 k<0 时，表明两个变量负相关；

当系数 k=0 时，表示两个变量不是线性相关；

当系数 k=±1 时，表示两个变量可以用很好的直线方程来描述。

则可以取系数 k 的绝对值，然后进行百分制，最后取整个数据集的分数作为答复质量的可解释评估分数，为下一步做铺垫，主要的算法流程图如下：

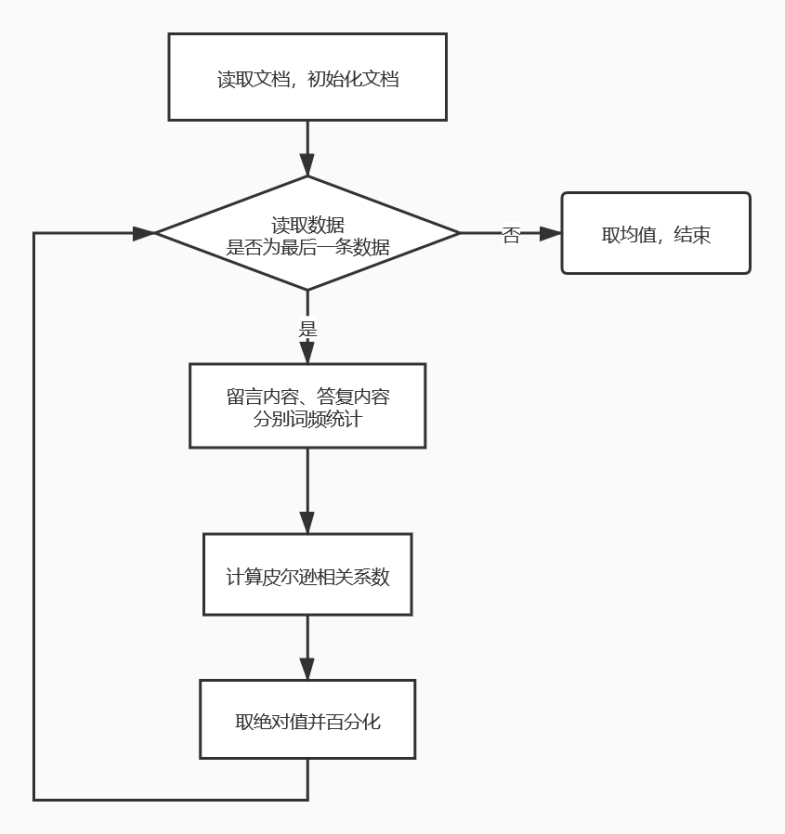

图 3.9 可解释性算法流程图

3.3.7 层次分析法模型构建及总评估

在对答复质量的相关性、时效性、完整性、可解释性搭建完评价方案后，根据问题二的层次分析法，构建权重矩阵为：

$$D=\begin{bmatrix} 1 & 3 & 3 & 1/3 \\ 1/3 & 1 & 1/3 & 1/5 \\ 1/3 & 3 & 1 & 1/2 \\ 3 & 5 & 2 & 1 \end{bmatrix} \tag{3.4}$$

再接着求权重方法种的规范列平均法求取权重。首先对判断矩阵列向量归一化得到对应矩阵 E。再对此进行算术平均法求取并得到特征向量 w，如下：

$$E=\begin{bmatrix} 0.2142 & 0.2500 & 0.4736 & 0.1639 \\ 0.0714 & 0.0833 & 0.0526 & 0.0983 \\ 0.0714 & 0.2500 & 0.1578 & 0.2458 \\ 0.6428 & 0.4166 & 0.3157 & 0.4918 \end{bmatrix} \tag{3.6}$$

$$w=\begin{bmatrix} 0.2754 & 0.0764 & 0.1813 & 0.4667 \end{bmatrix} \tag{3.7}$$

当$CR<0.1$时，认为判断矩阵的一致性在可接受的范围内，若$CR>0.1$时，则判断矩阵不符合一致性要求，则对该判断矩阵进行重新修改批正

并且计算可得$CI=0.0841$，$CR=0.0934$，一致性比例 CR 小于 0.1，故权重值合理，权重表如下表所示。

表 3.19 答复质量指标权重表

评价标准	权重
相关性	0.2754
时效性	0.0764
完整性	0.1813
可解释性	0.4667

导入数据并带入权重最终可得到质量的总分数为：68.46，评价质量为良好，其中各指标的平均得分值如下表：

表 3.20 答复质量指标得分表

	相关性	时效性	完整性	可解释性
分数	51.0	34.0	97.0	35.0

3.3.8 结果分析

由得分表可知到答复质量的时效性和可解释性较为差，这是因为答复数据在回复的时长上较长所导致的，在可解释性上较为差，在相关性上效果一般，答复质量的针对性一般，但是整个预料在句子的完整性上基本上完整，几乎很少出现缺少主谓宾语的现象。

4 总结与展望

4.1 总结

作为高速发展的信息时代的 21 世纪，各行各业的高速发展，数据的爆炸性增长，知识工程对人类社会的发展起到了巨大的作用。“智慧政务”中的文本挖掘是知识工程的一个重要的分支，能够让政府了解民意、汇聚民智以及凝聚民气的重要渠道，这具有极高的实际意义以及研究价值。

本文通过对自然语言的深入学习，利用传统机器学习构建分类器，word2vec 模型构建文本特征，通过 LogisticRegression 模型训练数据，对数据进行自动化文本分类，构建混淆矩阵，并计算 F1 值进行评价，对模型结果进行评估。

接着对留言内容利用 K-Means++进行文本聚类，聚类类别数通过手肘法确定，在聚类之后利用正态分布图法对数据进行离群值处理，并截取命名实体类并进行留言详情的文本摘要，最后通过层次分析法构建簇内、外评价指标，生成对应的权重，导入数据计算分数生成结果。

对于答复数据，分别从文本语义、文本结构、答复时间等对答复质量的相关性、时效性、完整性、可解释性等进行分析，同样利用层次分析法，合理的构建四个评价指标，并赋予权重量化生成数据，给定很好，良好，一般，差，很差五个评价等级，并判定答复质量。

4.2 展望

本文利用传统机器学习的方法，使用 LogisticRegression+Word2Vec 构建分类器为文本数据进行分类，利用 KMeans 文本聚类算法进行聚类和特定人名、地名提取，以及构建评价指标以及热度指标计算，并从语义、结构等为答复质量构建评价指标，为答复质量给出综合评价

在本文中，利用 LR 逻辑回归算法进行分类，虽然准确率达到了 88%但是仍存在局限性，无法继续上调，且是在特定的语料库上建立的分类，可以适当调整，提高分类器的通用性。在利用 KMeans 进行文本聚类的时候消耗大量内存，计算量大并不太具有解释性。在进行特定人名、地名的挖取的时候只进行名词的挖取，存在误差，应该可以考虑使用命名实体识别算法。在进行句子完整性的探究的时候，只考虑了三种结构，导致分数过高，应该考虑多一些句子结构进行探究。

参考文献

[1] 江胜尧.论提高网络执政能力推进社会管理创新[J].电子政务,2013(03):79-85.

[2] 阚璇. 我国智慧政务建设问题与对策研究[D].长春工业大学,2018.

[3] 张青云.基于文本数据挖掘技术的图书馆地方文献资源开发利用研究[J].河南图书馆学刊,2019,39(10):107-109.

[4] TF-IDF(term frequency–inverse document frequency)[Z](2011-03-28)，[Online] https://blog.csdn.net/allenshi_szl/article/details/6283375(2020-05-01)

[5] kmeans 聚类选择最优 K 值 python 实现[Z](2018-09-05)，[Online] https://blog.csdn.net/xyisv/article/details/82430107(2020-05-01)

[6] 李新福,赵蕾蕾,何海斌,李芳.使用 Logistic 回归模型进行中文文本分类[J].计算机工程与应用,2009,45(14):152-15

[7] 机器学习基础---过拟合问题及正则化技术[Z](2020-05-01)，[Online]https://www.cnblogs.com/ssyfj/p/12811577.html(2020-05-02)

[8] 机器学习中的相似性度量[Z](2011-03-08)，[Online] https://www.cnblogs.com/heaad/archive/2011/03/08/1977733.html(2020-05-02)

[9] 朱雪梅. 基于 Word2Vec 主题提取的微博推荐[D].北京理工大学,2014.

[10] Python 数据清洗--异常值识别与处理 01[Z](2019-04-24)，[Online] https://mp.weixin.qq.com/s?__biz=MzIxNjA2ODUzNg==&mid=2651438025&idx=1&sn=8da92f5d1ee161be6314964f881e2b0b&chksm=8c73a31ebb042a081fe444a1227d9cbab1166525d08902414551c39e7d8648b5382fb2fdfcc4&scene=21#wechat_redirect(2020-05-03)

[11] 夏杰. 基于道路运输企业安全生产管理数据的驾驶行为安全与节能评价方法[D].北京交通大学,2016.

[12] 层次分析法权重计算方法分析及其应用研究_邓雪[J]

[13] 刘妍东.大数据中数据的质量问题探析[J].现代商贸工业,2020,41(04):193.