“智慧政务”中的文本数据挖掘与综合分析

摘要

近年来，随着互联网的广泛应用和智慧政务的提出，政府把眼光投放到了政务微博、政务公众号和政务网站等各大政务平台上，通过获取公众反映的信息、对政府事件的评论等文本内容，利用网络文本分析技术和数据挖掘技术对文本进行深入研究，以此获得更多有用信息，提高政府办事效率。因此，技术的选择对研究具有重要的影响力和重大的意义。

针对问题 1： 通过 Factorize 函数将六个一级标签对应转换成 1，2，3，4，5 和 6 的 id,方便以后分类模型的训练。接着定义删除除字母、数字、汉字以外的所有符号函数，并利用 jieba 中文分词工具对留言详情进行分词处理，紧接着利用 TfidfVectorizer 提取每个一级标签中最相关的组合和最相关的二元组。数据预处理后，通过 train_test_split 进行数据集划分，利用 Count Vectorizer 类将文本中的词语转换为词频矩阵，并使用 TfidfTransformer 统计 vectorizer 中每个词语的 TF-IDF 值。为了优化实验结果，实验选用逻辑回归（Logistic Regression）、随机森林分类（RandomForestClassifier）、朴素贝叶斯分类器（MultinomialNB）、线性支持向量机（LinearSVC）、决策树分类器(DecisionTreeClassifier)和 K-神经网络分类器（KNeighborsClassifier）六种分类器模型进行实验结果比较。最后，通过比较 f1-score 值，本文发现线性支持向量机（LinearSVC）分类模型适合实验分析。

针对问题 2：本题按照由果溯因的思路进行实验。第一步先分别采用 Hacker News 热度算法、Reddit 热度算法、魔方秀热度计算每一条留言记录的热度进行比较，然后从结果中筛选出一种比较适合的热度计算方法。确定完算法后，将实验数据按照从大到小顺序进行排序，此举目的可以有效地将样本量控制在一定范围内，有效地缓解了实验的测试。第二步，本文利用 K-Means 对整体留言主题进行聚类分析。此举目的是验证第一步的准确性。首先为了获取精确的簇数，本文采用肘部法进行粗略观察以便确定 k 值；接着利用 K-Means 对文本进行聚类分析；最后采用轮廓系数检验实验结果。本题主要是以热度为主，文本聚类分析为辅进行实验。

针对问题 3：同样地利用 jieba 中文分词工具对留言详情信息以及答复信息

进行分词，并通过 jieba 分词词性标注来剔除没有意义的词。首先利用文本相似度来衡量两个文本之间的相关性，主要是通过主流的谷歌 word2vec 算法建立的高级词向量模型，通过 word2vec 模型建立留言和对应答复的词向量，随后利用余弦相似度计算两个向量之间的相似度以表示两个文本的相似度。其次利用主题相似度来衡量答复信息是否在主题层面上与留言信息类似，以表示答复信息是否完整地回应了留言详情。为了更好地提取主题信息，我们采用了 LDA(Latent Dirichlet Allocation)用来推测文档的主题分布，并且我们对获得的留言和相应答复的主题概率分布进行 js 散度(Jensen-Shannon)距离计算来表示它们的主题相似度。同时，本文将可解释性用情感分析替换，所以我们利用百度自带的情感分析 API 进行回复内容的情感分析，0 代表消极情绪，1 代表中性情绪，2 代表积极情绪。

关键词：线性支持向量机（LinearSVC）、Hacker News 热度算法、LDA(Latent Dirichlet Allocation)、K-Means

Text Data Mining and Comprehensive Analysis in "Smart Government Affairs"

Abstract

In recent years, with the widespread use of the Internet and the introduction of smart government affairs, the government has set its sights on various government affairs platforms such as government affairs microblogs, government affairs public accounts and government affairs websites, by obtaining information reflected by the public and commenting on government events, etc. Text content, using network text analysis technology and data mining technology to conduct in-depth research on text, in order to obtain more useful information and improve the efficiency of government affairs. Therefore, the choice of technology has important influence and great significance for research.

Aiming at Problem 1: The six first-level labels are converted into ids of 1, 2, 3, 4, 5, and 6 through the Factorize function, which facilitates the training of classification models in the future. Then define and delete all symbol functions except letters, numbers, and Chinese characters, and use the jieba Chinese word segmentation tool to perform word segmentation on the message details, and then use TfidfVectorizer to extract the most relevant combinations and most relevant binary groups in each first-level label. After data preprocessing, the data set is divided by train_test_split, the words in the text are converted into word frequency matrices using the Count Vectorizer class, and the TF-IDF value of each word in the vectorizer is counted using TfidfTransformer. In order to optimize the experimental results, the experiment selected Logistic Regression, Random Forest Classification (RandomForestClassifier), Naive Bayes Classifier (MultinomialNB), Linear Support Vector Machine (LinearSVC), Decision Tree Classifier (DecisionTreeClassifier) and K-Neural Six

classifier models of network classifier (KNeighborsClassifier) were used to compare the experimental results. Finally, by comparing the f1-score values, this paper finds that the Linear Support Vector Machine (LinearSVC) classification model is suitable for experimental analysis.

Aiming at question 2: This question is experimented with the idea of traceability. The first step is to use the Hacker News heat algorithm, Reddit heat algorithm, Rubik's cube show heat to calculate the heat of each message record for comparison, and then select a more suitable heat calculation method from the results. After the algorithm is determined, the experimental data is sorted in order from large to small. This purpose can effectively control the sample size within a certain range and effectively ease the experimental test. In the second step, this article uses K-Means to cluster the overall message topics. The purpose of this move is to verify the accuracy of the first step. First, in order to obtain the exact number of clusters, this article uses the elbow method to make a rough observation in order to determine the k value; then uses K-Means to perform cluster analysis on the text; and finally uses the contour coefficient to test the experimental results. This question is mainly based on the heat, and the text clustering analysis is supplemented by the experiment.

Aiming at question 3: Similarly, use jieba Chinese word segmentation tool to segment message details and reply information, and use jieba word segmentation tag to remove meaningless words. First, the text similarity is used to measure the correlation between the two texts, mainly through the advanced word vector model established by the mainstream Google word2vec algorithm, and the word vector of the message and the corresponding response is established by the word2vec model. The similarity between two vectors represents the similarity of

two texts. Secondly, the topic similarity is used to measure whether the response information is similar to the message information at the topic level to indicate whether the response information completely responds to the message details. In order to better extract topic information, we use LDA (Latent Dirichlet Allocation) to speculate the topic distribution of the document, and we use the js divergence (Jensen-Shannon) distance calculation to calculate the topic probability distribution of the received messages and corresponding responses Indicates the similarity of their subjects. At the same time, this article replaces interpretability with sentiment analysis, so we use Baidu's own sentiment analysis API for sentiment analysis of reply content. 0 represents negative emotions, 1 represents neutral emotions, and 2 represents positive emotions.

Keywords: Linear Support Vector Machine (LinearSVC), Hacker News popularity algorithm, LDA (Latent Dirichlet Allocation), K-Means

目录

1 挖掘目标..... 8
2 分析方法与过程..... 9
2.1 问题 1 分析方法与过程..... 10
2.1.1 数据流程图..... 10
2.1.2 数据预处理..... 10
2.2 问题 2 分析方法与过程..... 13
2.2.1 数据流程图..... 13
2.2.2 热度计算..... 13
2.2.3 文本聚类..... 14
2.3 问题 3 分析方法与过程..... 15
2.3.1 数据流程图..... 15
2.3.2 数据预处理..... 16
3 结果分析..... 19
3.1 问题 1 结果分析..... 19
3.2 问题 2 结果分析..... 24
3.3 问题 3 结果分析..... 28
3.3.1 分析相关性指标..... 28
3.3.2 分析完整性指标..... 28
3.3.3 分析可解释性指标..... 30
4 结论..... 31
5 参考文献..... 32

1 挖掘目标

本次建模目标是利用政府收集的互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见数据，利用 jieba 中文分词工具对留言详情进行分词，利用逻辑回归（Logistic Regression）、随机森林分类（RandomForestClassifier）、朴素贝叶斯分类器（MultinomialNB）、线性支持向量机（LinearSVC）、决策树分类器(DecisionTreeClassifier)和 K-神经网络分类器（KNeighborsClassifier）六种分类器进行实验比较。同时针对热度问题采用 Hacker News、Reddit 和魔方秀热度算法统计每一条留言记录热度问题，接着通过 K-means 聚类方法对留言主题聚类，形成最终实验结果。此外，为了提高公众满意度，政府需要提供良好的服务态度。所以，本文利用留言主题、留言详情和回复内容三个模块，进行文本的相关性分析、完整性分析和可解释分析，以确保政府拥有良好的服务形象。

因此，本文主要达到以下三个目标：

a) 利用文本分词和分类器模型对大量留言详情进行文本挖掘，建立了一套关于留言内容的一级标签分类模型。

b) 根据热度排名算法将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，定义合理的热度评价指标；同时利用聚类算法、轮廓系数和肘部法检验实验结果，此举可以做到及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。

c) 利用文本相似度计算、文本主题相似度和情感分析计算来合理定义答复与留言信息的相关性、完整性和可解释性。

2 分析方法与过程

为了更好地开展实验活动，我们设计了相应的实验整体流程图如图 1 所示：

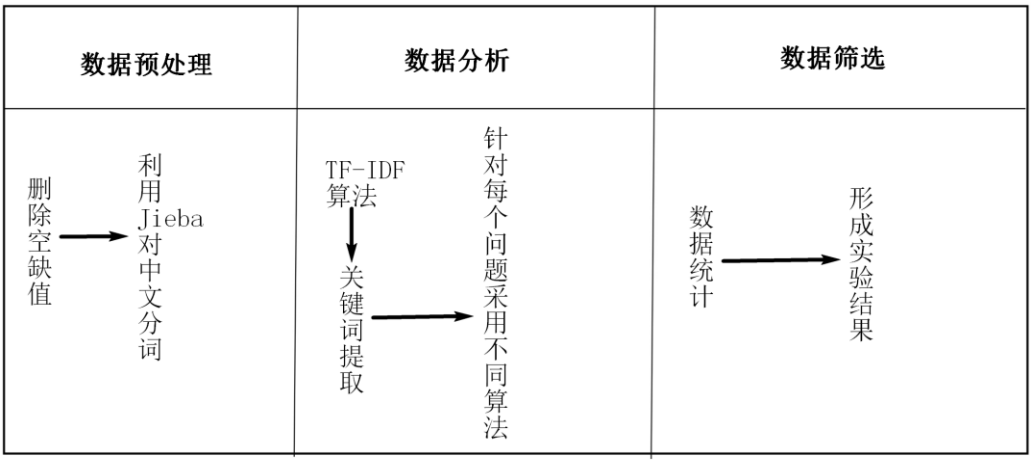

图 1 实验总体流程图

本用例主要包括如下步骤：

步骤一：数据预处理。在题目给出的数据中，先判断是否出现空缺值，如有将进行删除；否则进行下一步数据处理。完成空缺值判断后，将把相应字段进行中文分词。

步骤二：数据分析。在对留言详情、留言主题和回复内容分词后，把这些词语转换为向量，以供挖掘分析使用。本实验采用 TF-IDF 算法，找出每一条留言详情或留言主题的关键词，利用 LDA 模型进行主题挖掘。

步骤三：数据筛选及整理。统计相关数据，分类筛选汇总，预测留言热门问题等。将实验结果按照题目要求汇总成相应的表格。

2.1　问题 1 分析方法与过程

2.1.1数据流程图

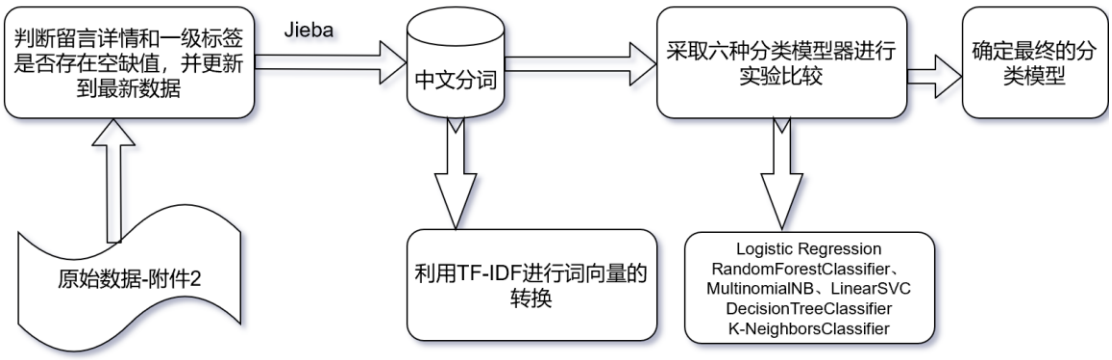

图 2. 一级标签分类模型建立流程图

2.1.2 数据预处理

2.1.2.1　判断留言详情空缺值情况

首先判断题目给出的数据中是否出现空缺值情况。因为考虑到实验的准确性，留言详情必须不为空。对数据判断空缺值的 python 程序见附件 ProblemOne.py。预处理过的数据保存到了原文件附件 2 中。

2.1.2.2　分词留言详情

在对留言详情进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件 2 留言详情字段中，以中文文本的方式给出了数据。为了便于转换，先要对这些留言详情描述信息进行中文分词。本文采用 Python 的中文分词包 Jieba 进行分词。Jieba 采用了基于前缀词典实现的高效词图扫描技术，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)，同时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，采用 Viterbi 算法进行计算和词性标注，更好地实现中文分词效果。在分词的同时，本文采用了 TF-IDF 算法，

将每个一级标签中最相关的组合和最相关的二元组提取出来。

2.1.2.3 TF-IDF 算法

在对留言详情信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。本文采用 TF-IDF 算法，把留言详情信息转换为权重向量。TF-IDF（Term Frequency－Inverse Document Frequency，词频-逆向文件频率）是一种用于信息检索（Information Retrieval）与文本挖掘（Text Mining）的常用加权技术。它是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 算法的具体原理如下：

(一) TF 是词频(Term Frequency)

$$词频 = \frac{某个词在文本中出现的词数}{文本的总词数}（1）$$

(二) IDF 是逆向文件频率(Inverse Document Frequency)

$$逆文档频率 = \log（\frac{语料库的文本总数}{包含该词的文本数+1}）（2）$$

(三) TF-IDF 实际上是：TF * IDF

$$TF - IDF = TF * IDF（3）$$

生成 TF-IDF 向量的具体步骤如下：

（1）使用 TF-IDF 算法，找到每个一级标签中的六个相关关键词；

（2）对每个一级标签的六个关键词，合并成一个集合，利用 Count Vectorizer 类会将文本中的词语转换为词频矩阵，计算每个一级分类描述对于这个集合中词的词频，如果没有则记为 0；

（3）使用 TfidfTransformer 统计 vectorizer 中每个词语的 TF-IDF 值，生成每个一级标签描述的 TF-IDF 权重向量，计算公式如下：

$$TF - IDF = TF * IDF(4)$$

2.1.2.4 对比六种分类模型

为了更好地实验结果，我们选取了六种分类器模型进行实验结果比较：逻辑

回归（Logistic Regression）、随机森林分类（RandomForestClassifier）、朴素贝叶斯分类器（MultinomialNB）、线性支持向量机（LinearSVC）、决策树分类器(DecisionTreeClassifier)和 K-神经网络分类器（KNeighborsClassifier）。六种分类器模型具体参数设置如图 3 所示：

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import cross_val_score
class RandomForestClassifierWithCoef(RandomForestClassifier):
    def fit(self, *args, **kwargs):
        super(RandomForestClassifierWithCoef, self).fit(*args, **kwargs)
        self.coef_ = self.feature_importances_
class DecisionTreeClassifierWithCoef(DecisionTreeClassifier):
    def fit(self, *args, **kwargs):
        super(DecisionTreeClassifierWithCoef, self).fit(*args, **kwargs)
        self.coef_ = self.feature_importances_


models = [
    RandomForestClassifierWithCoef(n_estimators=200, max_depth=3, random_state=0),
    LinearSVC(),
    MultinomialNB(),
    LogisticRegression(random_state=0),
    DecisionTreeClassifierWithCoef(max_depth = 2),
    KNeighborsClassifier()
]
```

图 3 六种分类器模型参数设置代码图

2.2　问题 2 分析方法与过程

2.2.1数据流程图

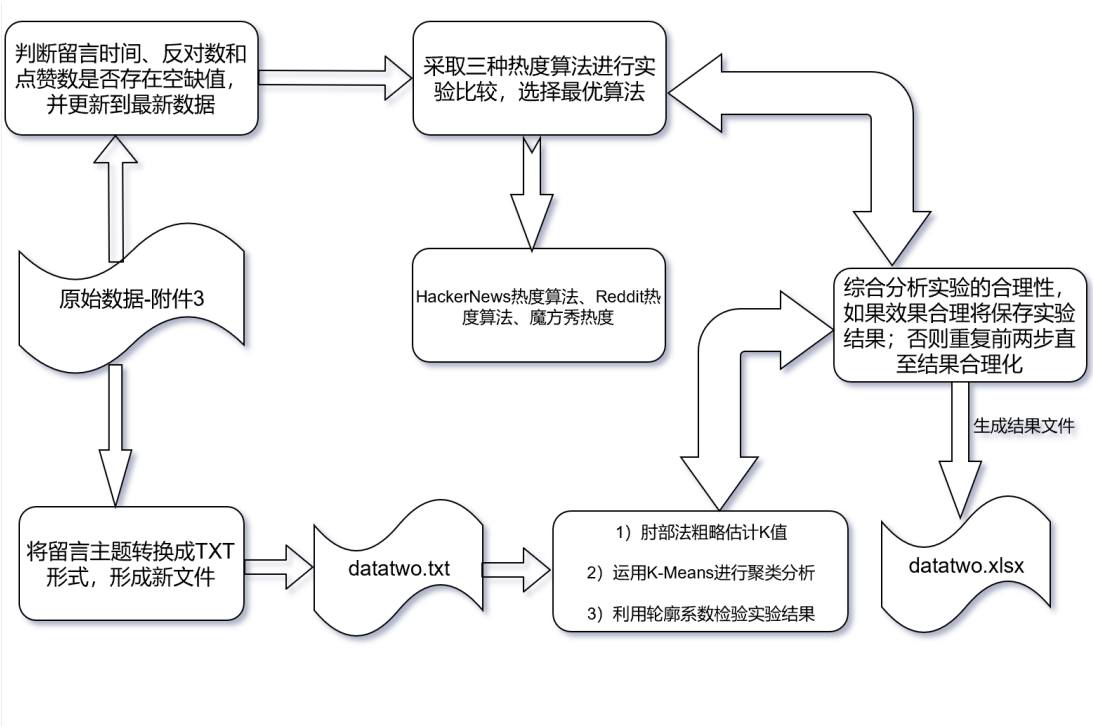

图 4 热度评价指标体系建立流程图

2.2.2热度计算

本文分别采用 Hacker News 热度算法、Reddit 热度算法和魔方秀热度算法计算每一条留言热度情况。三种算法介绍如下：

a) Hacker News 热度算法

$$Score = \frac{P-1}{(T+2)^G} (5)$$

其中 P 代表投票数(本文加和点赞数和反对数当为投票数)。-1 是把自己投的过滤掉。T 代表发布到现在的时间间隔，单位小时，+2 防止除数太小（本文利用测试当时时间-留言时间）。G 代表重力加速度，它的数值大小决定了排名随时间下降的速度快慢。本文 G=1.8。

Hacker News 热度算法能够根据浏览量过滤出同时发布的一批作品集中的热

度作品，因为他们的分母相同，而分子大的热度肯定大。同时能够过滤出不同时间段热度高的作品，避免误差的产生。

b） Reddit 热度算法

$$Score = \log_{10} Z + \frac{yt}{4500} (6)$$

其中 t 代表代表发布到现在的时间间隔，单位秒；y 是一个符号变量，表示对新闻内容的总体看法。如果赞成票居多，y 就是 +1；如果反对票居多，y 就是 -1；如果赞成票和反对票相等，y 就是 0。y 是对新闻内容评价的一种定性表达，0 表示没有倾向，大于 0 表示正面评价，小于 0 表示负面评价（本文采用点赞数-反对数作为 y 值取值的标准）；z 表示点赞数超过反对数的数量。如果赞成数少于或等于反对数，那么 z 就等于 1。

c） 魔方秀热度

$$Score = \frac{(a*0.7+s*0.3)*1000}{(T+2)^{1.2}} (7)$$

其中 a 代表点赞数；s 代表总数和(本文把点赞数和反对数加和)；T 代表发布到现在的时间间隔，单位小时，+2 防止除数太小（本文利用测试当时时间-留言时间）。1.2 代表重力加速度，它的数值大小决定了排名随时间下降的速度快慢。

2.2.3文本聚类

本文利用 K-Means 做文本聚类，其中采用的是欧几里得距离的平方，公式如下：

$$d(x,y)^2 = \sum_{i=1}^{n}(x_i - y_i)^2 = ||x - y||_2^2 (8)$$

实验首先加载第一步产生的语料，把关键词出现 1 次以上，转化成向量(tf)，idf 设为权重，把关键词转化成词篇矩阵。接着利用肘部法粗略估计 K 值。

同时获取 K-means 算法的簇内误差平方和 SSE（within-cluster sum of squared errors）（公式如下），通过最小化 SSE 值，优化聚类结果。

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{m} w^{(i,j)} = ||x^{(i)} - u^{(j)}||_2^2 （9）$$

为了实验结果的准确性，本文采用轮廓系数来判断实验结果的好坏。具体公

式如下：

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\{b^{(i)}, a^{(i)}\}} \quad (10)$$

以上两个步骤完成后，本文进行了综合分析，判断结果是否符合题目要求。如果不符合则重新设置值重新实验。

2.3 问题 3 分析方法与过程

2.3.1 数据流程图

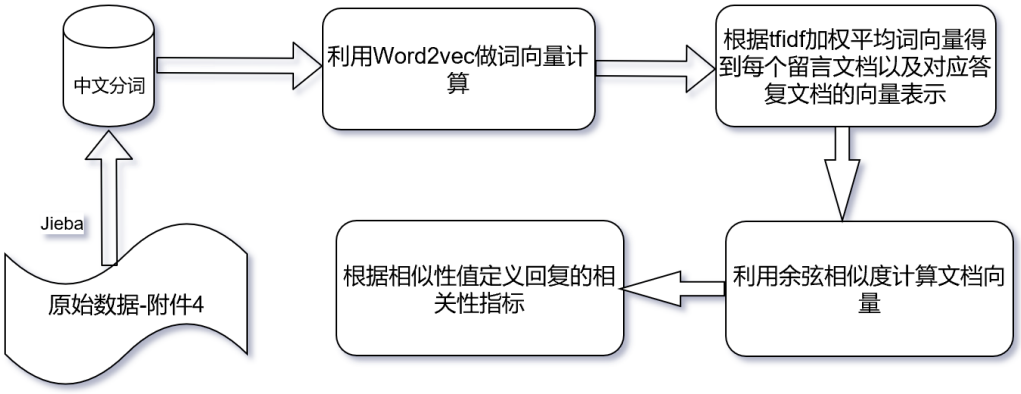

图 5 回复相关性指标建立流程图

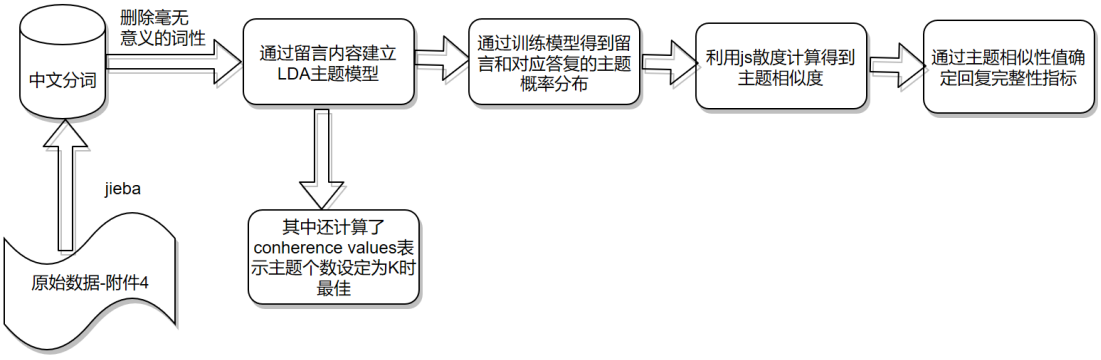

图 6 回复完整性指标建立流程图

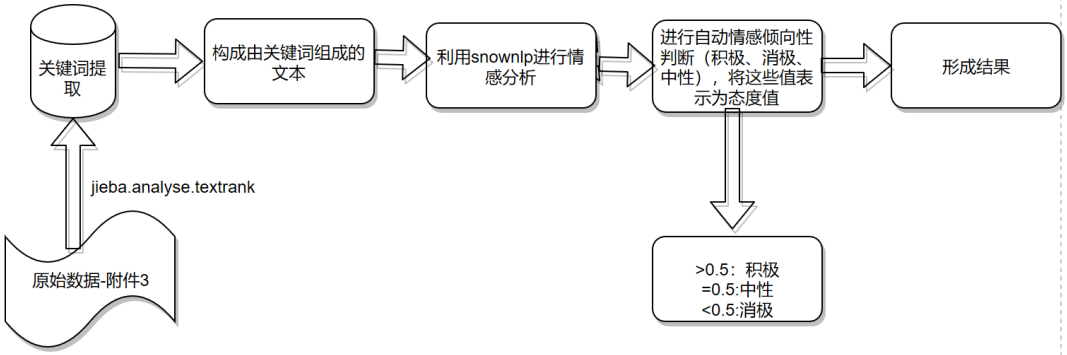

图 7 回复可解释性指标建立流程图

2.3.2数据预处理

2.3.2.1 对文本(留言信息和答复)分词

操作流程和前两问分词一样，所以这里就不再赘述。

2.3.2.2 文本相似度计算

留言详情和答复之间的相关性类似于智能问答系统主要解决问句的真实意图分析、问句与答案之间的匹配关系，理解以自然语言形式描述的用户提问，并通过检索异构语料库或问答知识库返回简洁、精确的匹配正确答案，那么我们可以从文本相似度指标来考察问答文本之间的关系。其中词向量可以用于测量单词之间的相似度，相同语义的单词，其词向量也应该是相似的。而想要从文本种提取特征，有各式各样的方法创建词向量。我们采用的是谷歌在 2013 年发布的基于神经网络实现的 word2vec 模型，其使用 CBOW（Continuous Bag of Words）和 skip-gram 两种结构学习单词的分布式表示。其中 CBOW 的输入是多个词向量，输出是所有词的 softmax 概率，可以通过一次 DNN 前向传播算法并通过 softmax 激活函数找到概率最大的词对应的神经元， Skip-Gram 模型和 CBOW 的思路是反着来的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量。

Word2vec 模型相当于与其他神经网络实现运行速度更快，而且不需要手工标签来创建单词的意义标识。在代码实现部分，我们采用的是 python 的 gensim 库，该库是 word2vec 的 python 实现。在此研究中输入的是留言详细信息和答复信息分词并处理后的文本数据。输出的结果得到词向量表示每一个词。其次我们使用单词的 TF-IDF 评分对每个匹配的词向量进行加权，对它们进行求和并处以文档中匹配的单词数量。随后可以得到每个文档的一个 TF-IDF 加权平均词向量。以上描述的数学公式可以表示为。

$$TWA(D)=\frac{\sum_{1}^{n} wv(w)\times \mathrm{tfidf}(w)}{n} \quad (11)$$

其中 TWA(D) 表示的是 TF-IDF 文档 D 加权平均词向量，文档 D 中包括多个中文词用 wi 表示，wv(w)表示的是中文词 w 的词向量，tfidf(w)是单词 w 的 TF-IDF 权重。

最后我们已经通过加权词向量来表示每个文档，利用文档的向量表示可以通过向量相似度计算方法来计算文本相似度。这里我们采用余弦相似度。所谓余弦相似度，是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。我们可以把两个文本你向量想象成空间中的两条线段，都是从原点（[0, 0, ...]）出发，指向不同的方向。两条线段之间形成一个夹角，如果夹角为 0 度，意味着方向相同、线段重合，这是表示两个向量代表的文本完全相等；如果夹角为 90 度，意味着形成直角，方向完全不相似；如果夹角为 180 度，意味着方向正好相反。因此，我们可以通过夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似。以下是余弦相似度的公式表示：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(X_i\times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\times\sqrt{\sum_{i=1}^{n}(Y_i)^2}} \quad (12)$$

2.3.2.3 主题相似度计算

隐含 Dirichlet 分布（Latent Dirichlet Alloction，LDA）技术是一种概

率生成模型，其中假定每个文档具有类型于概率隐含语义索引模型的主题组合，但是在此情况下，隐含主题包含它们的 Dirichlet 先验分布。算法过程：

1）初始化必要的参数主要包括是主题数 K 的设置，迭代次数等

2）对于每个文本，随机讲每个单词初始化为 K 个主题之一。

3）开始如下的迭代过程

4）对于每个文本 D 中的每个单词 w 计算 p(T|D),指 D 中分配给主题 T 的词的比例。再计算 P（w|D），指对于含有词 w 的所有文本分配给主题 T 的比例。在考虑所有其他词及其主题分配，用主题 T 和概率 p（T|D）x p(w|D) 重新分配词 w。

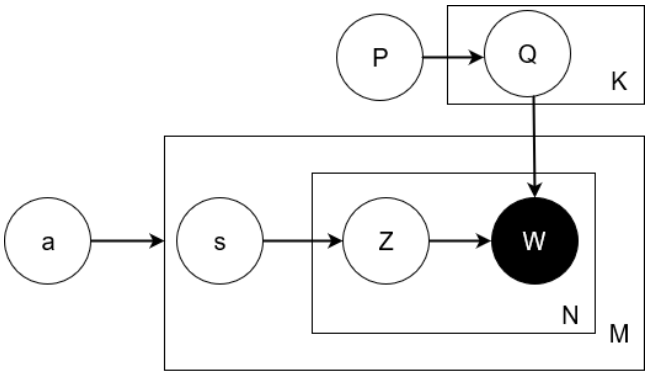

图 8 LDA 盘子表示法图

K 是主题数量，N 是文本种的词组数量，M 是待分析的文本数量，a 是每个文本主题分布的 Dirichlet 先验集中参数,p 是每个主题词分布的相同参数,Q(k)是对于主题 K 的词分布，S(i)是对于文本 i 的主题分布，z(i,j)是对于 w(i,j)的主题分配，w(i,j)是第 i 个文本的第 j 个词。

在使用 LDA 进行文本相似的计算时，其目标是找到每一篇文档的主题分布和每一个主题中词的分布。LDA 模型通过类似词聚类的办法将相似词聚类为一个主题，使得同一主题下的词具有近义词的特性，而不同主题之间的词具有多义词的特性，也就是此时我们能够得到主题-词项矩阵即表示每个主题下每个词的权重。此外我们也能得到一个文档-主题矩阵，从而是计算文本主题分布来计算文本间的相似度。计算两个文本相似度可以计算与之对应的主题概率分布来实现：KL 距离、JS 距离等，在我们的研究方法中采用了 JS 距离计算方式，因为 JS 散度度量了两个概率分布的相似度，基于 KL 散度的变体，解决了 KL 散度非对称的问题。以下是 JS 距离公式：

$$JS = \frac{1}{2} \times KL(P \| M) + \frac{1}{2}(Q \| M) \quad (13)$$

此外我需要考虑的主题模型构建的好坏，需要注重主题数 K 的设定，在我们的研究中我们考察模型的 conherence values 即连贯性来衡量模型。

2.3.2.4 回复可解释性计算

本文认为一个人在情绪积极时，所说的话是比较合理的，是可以解释的；但是当一个人拥有消极情绪，所说的话是有点带有个人情绪进去，有可能讲的话语序混乱，别人理解起来会有点困难。所以，本文采用分析回复内容的情感来分析文本可解释性，本文采取 0 代表消极情绪，1 代表中性；2 代表积极情绪。

本文采用 SnowNLP 包计算情感。它是国人开发的 python 类库，可以方便的处理中文文本内容，是受到了 TextBlob 的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的，于是写了一个方便处理中文的类库，并且和 TextBlob 不同的是，这里没有用 NLTK，所有的算法都是自己实现的，并且自带了一些训练好的字典。注意本程序都是处理的 unicode 编码，所以使用时请自行 decode 成 unicode。

3 结果分析

3.1 问题 1 结果分析

目前，电子政务系统大部分还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。为了更好地将留言进行分类，我们采取目前比较主流的六种分类模型进行实验结果分析。首先我们利用 is null().sum()函数判断空缺值存在情况，如图所示并没有存在空缺值情况。

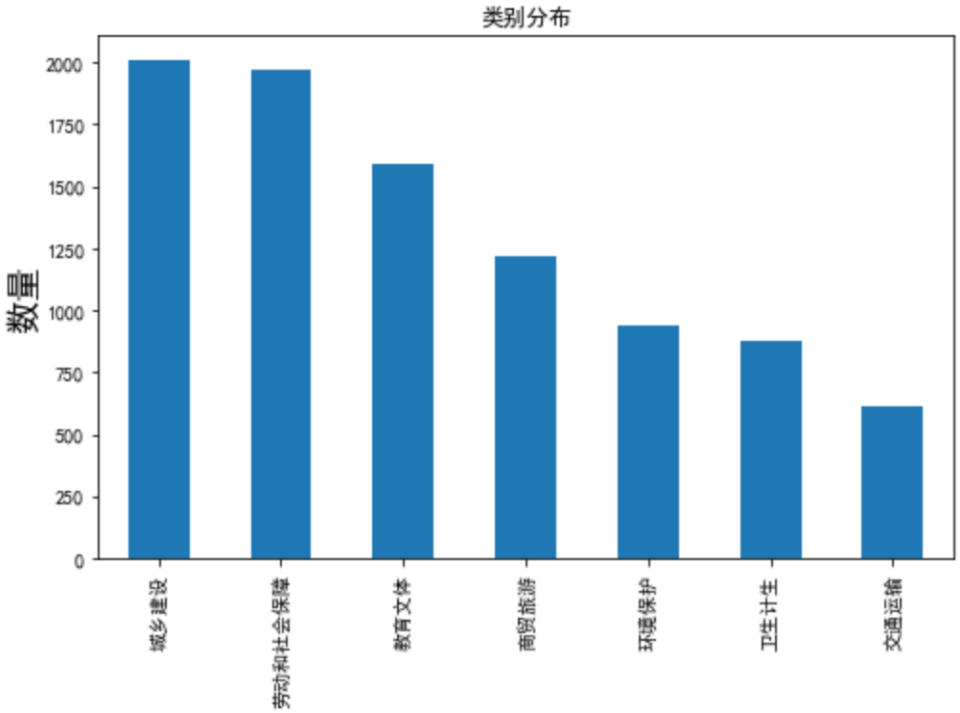

图 9 六个一级标签数据分布情况

接着利用 Factorize 函数将一级标签转换成 Id,这样便于以后的分类模型，效果图如图 10 所示：

	一级分类	留言详情	category_id
0	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	0
1	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	0
2	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA市政府、市交警支队、市安监局、市环保局、A...	0
3	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t胡书记，您好，感谢您百忙之中查看这份留言。我...	0
4	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t K8县丁字街的商户乱摆摊，前段时间丁字街...	0

图 10 一级标签转换成 Id 图

紧接着将留言详情分词后，通过分析六种模型的准确率（accuracy），具体公式如下所示：

$$准确率 = \frac{分类器正确分类的样本数}{总样本数} \quad (1)$$

来确定选择效果较好的分类器。结果图如下所示：

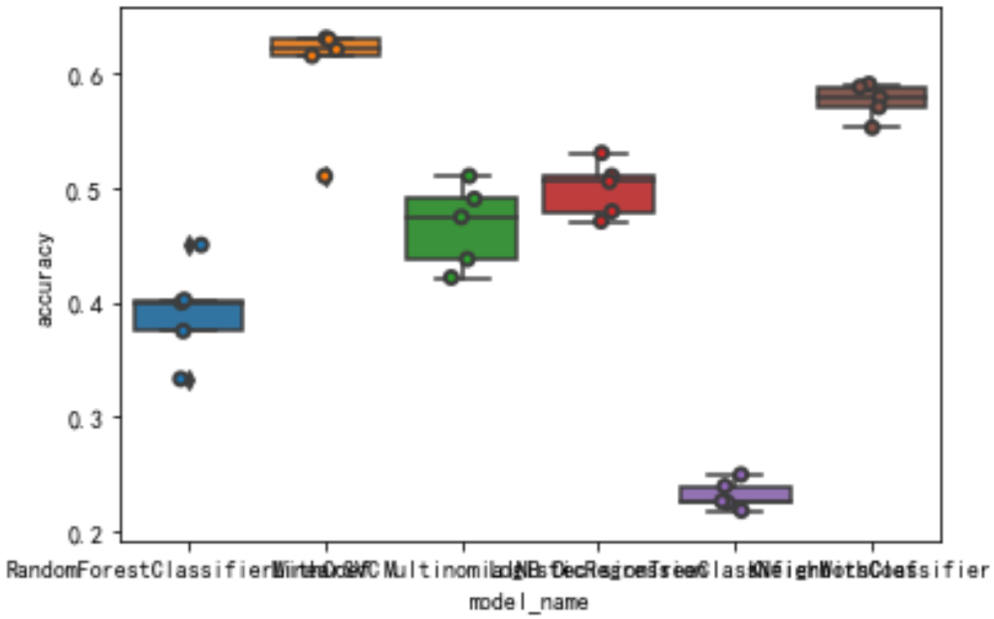

图 11 蓝色代表随机森林；橙色代表线性支持向量机；绿色代表朴素贝叶斯；红色代表逻辑回归；紫色代表决策树；红褐色代表 K-神经网络

从

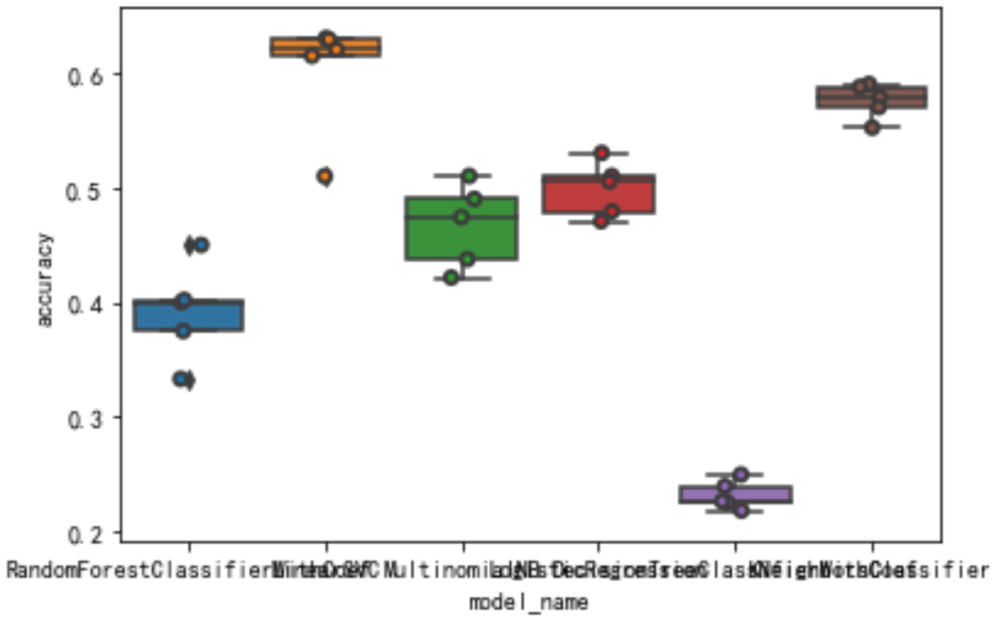

图 中我们可以看出线性支持向量机准确率最高，K-神经网络排名第二，朴素贝叶斯和逻辑回归差不多水平，最差的是决策树。具体准确率值如表 1 所示：

表 1 六种分类器模型准确率表

分类器模型	准确率
DecisionTreeClassifierWithCoef	0.232209
KNeighborsClassifier	0.575629
LinearSVC	0.600651
Logistic Regression	0.498982
MultinomialNB	0.466659
RandomForestClassifierWithCoef	0.392079

通过以上步骤，我们决定采取支持向量机来进行实验。首先利用支持向量机来画关于测试和预测的混淆矩阵，如图 12 所示：

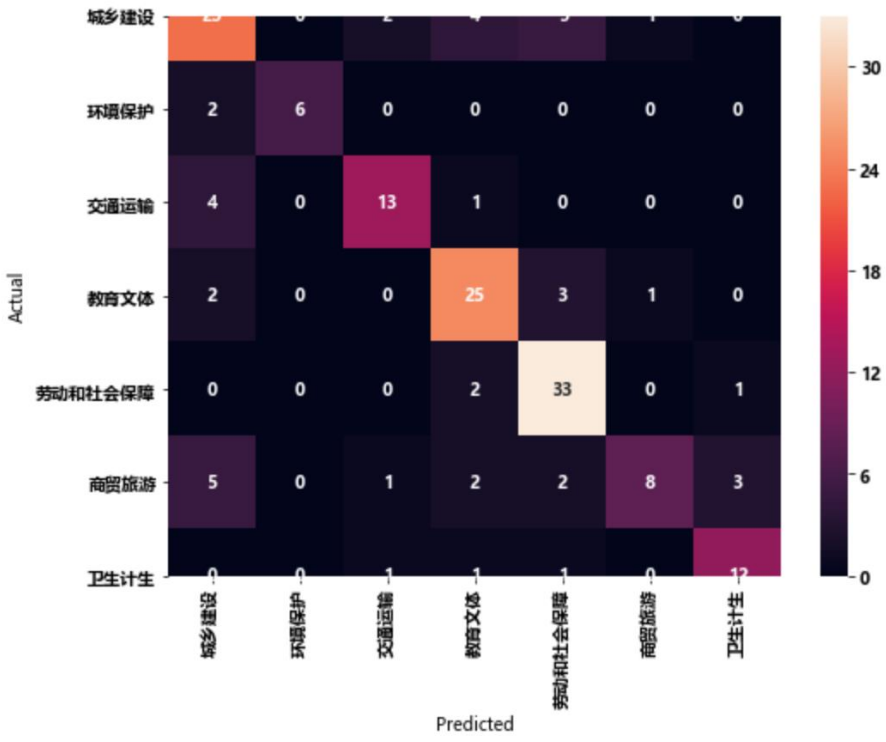

图 12 基于线性支持向量机的混淆矩阵图

接着通过 sklearn.feature_selection 中的 chi 模块，进行特征预测，获得六个一级标签下相关关键词，如表 2 所示：

表 2 六个一级标签关键词提取情况表

一级分类	Top unigrams	Top bigrams
交通运输	出租车	出租车 司机
	快递	的士 司机
劳动保障	社保	劳动 关系
	退休	退休 人员
卫生计生	医院	社会 抚养费
	医生	乡村 医生
商贸旅游	故障	监督 管理局
	景区	市场 监督
城乡建设	房子	棚户区 改造
	开发商	公积金 贷款
教育文体	学校	培训 机构
	教育局	教育局 领导
环境保护	污染	环评 手续
	环保局	环保局 投诉

从表格我们可以看出，大多数关键词的提取都是比较贴切一级分类的，只有少部的关键词提取优点问题，例如商贸旅游中提取到市场、监督、管理局等字眼，接着我们通过 Tokenizer 模块进行实验预测，同时利用 precision、recall 和 f1-score 三个指标进行实验结果展示，如表 3 所示：

表 3 基于线性支持向量机的三个指标表

	precision	recall	f1-score
城乡建设	0.64	0.66	0.65
环境保护	1.00	0.75	0.86
交通运输	0.76	0.72	0.741
教育文体	0.71	0.81	0.76
劳动和社会保障	0.75	0.92	0.83
商贸旅游	0.80	0.38	0.52
卫生计生	0.75	0.80	0.77
accuracy			0.73
macro avg	0.77	0.72	0.73
weighted avg	0.74	0.73	0.72

3.2 问题 2 结果分析

及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。我们分别采用 Hacker News 热度算法、Reddit 热度算法和魔方秀热度算法统计出前五条记录，如表 4、表 5、表 6 所示。通过三种热度算法实验结果对比，本文将采用 Hacker News 热度算法进行热度计算。同时通过三种热度算法我们可以发现 A 市 A4 区 58 车贷案问题、A 市富绿物业丽发新城强行断业主家水、A 市万科金域华府三期出行难，断头路何时能通和 A 市湖楚财富及两大国资委股东侵吞百姓血汗钱等问题都是公众比较关注的热门问题。

表 4 基于 Hacker News 热度算法前五条记录表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	433	0.287174589	2020-05-06	A 市湖楚财富及两大国资委股东	侵吞百姓血汗钱
2	249	0.060171077	2019-11-04	A 市 58 车贷案警官	应跟进关注留言
3	424	0.040051098	2020-04-27	A 市万科金域华府	三期出行难，断头路何时能通
4	383	0.01385616	2020-03-17	A7 县东四路远大路到人民路南	延拆迁进度等问题
5	205	0.013517829	2019-09-2	A 市富绿物业丽发新城	强行断业主家水

表 5 基于魔方秀热度算法前五条记录表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	249	1385.761341	2019-11-04	A 市 58 车贷案警官	应跟进关注留言
2	433	870.5505633	2020-05-06	A 市湖楚财富及两大国资委股东	侵吞百姓血汗钱
3	205	354.6026385	2019-09-21	A 市富绿物业丽发新城	强行断业主家水
4	432	267.5805206	2020-05-05	A2 区余易贷平台	涉嫌诈骗，群众合法维权被强行扣押
5	424	225.1068802	2020-04-27	A 市万科金域华府	三期出行难，断头路何时能通

表 6 基于 Reddit 热度算法前五条记录表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	843	3.230487319	2021-06-20	A 市 A5 区汇金路五矿万境 K9 县	存在一系列问题
2	1483	3.011438428	2023-03-22	A 市金毛湾	配套入学的问题
3	249	2.905992864	2019-11-04	A 市 58 车贷案警官	应跟进关注留言
4	1212	2.72451598	2022-06-24	A 市	58 车贷特大集资诈骗案保护伞
5	1383	2.703232046	2022-12-12	A 市 A4 区	58 车贷案

经过热度计算后，本文利用 K-Means 整体对留言主题进行聚类分析，以此来验证热度算法的合理性。首先本文采用肘部法则进行粗略估计，确定 K 值。由图可以观察到 K=效果图如下所示：

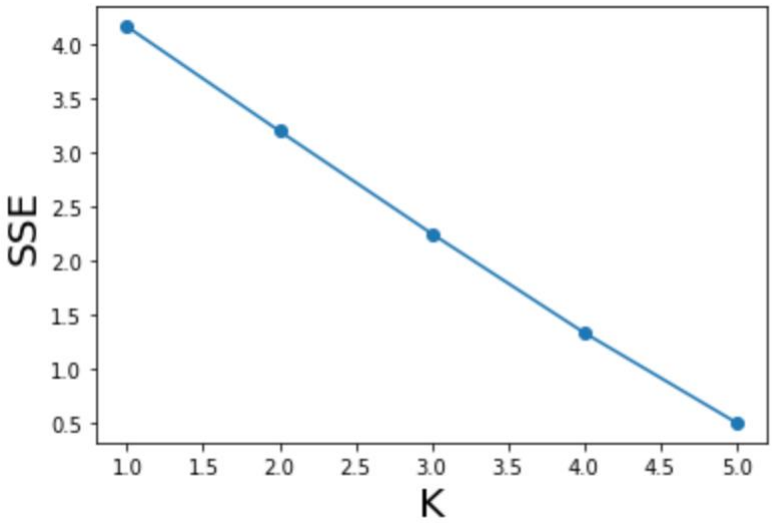

图 13 肘部法则效果图

接着计算 SSE 值，效果图如图 14 计算 SSE 效果图所示：

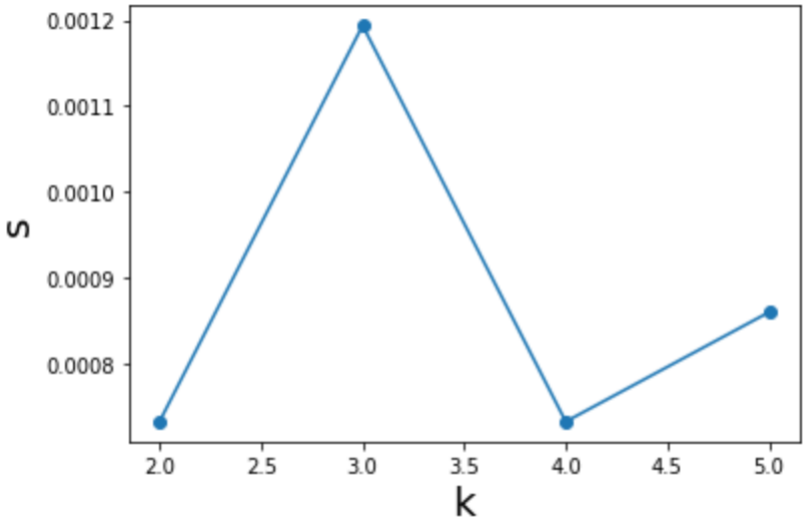

图 14 计算 SSE 效果图

3.3　问题 3 结果分析

3.3.1 分析相关性指标

本文采用 Word2vec 文本相似度计算，对数据进行测试，结果距离如下图所示：

```
Word2Vec(vocab=21808, size=50, alpha=0.025)
.model load time 0.1267
留言详情文本向量表示:[[ 0.434  1.349 -0.273 -0.258 -1.846  1.417  0.069  0.05  -0.679  0.886
  -0.1   -0.009  0.233 -0.694 -0.083 -0.182  1.216  0.017 -0.333 -0.496
  -0.967 -1.685 -1.033 -1.465  0.342 -0.725  0.361  1.744  1.023 -0.642
  -0.445  0.408  1.04  -2.135  0.429  0.149  0.295  2.136 -0.047 -0.771
  -1.431 -0.959  0.009 -1.372 -0.84   0.056 -0.038  0.464  0.793 -0.299]]
答复文本向量表示:[[-0.074  1.056 -0.017 -0.897 -1.26   1.015  0.406 -0.024 -0.449  0.498
   0.14  -0.48   0.399 -0.041  0.632  0.138  0.573  0.895 -0.267 -0.258
   0.15  -1.456 -0.92  -0.965 -0.246 -0.605  0.069  0.718  0.399 -0.576
  -0.5   -0.02   0.838 -1.451  0.356 -0.296  0.835  1.784 -0.287 -0.9
  -1.466 -0.225 -0.391 -0.999 -0.056 -0.229 -0.309 -0.021  0.963 -0.657]]
word2vec文本相似度:0.8592
```

图 15 文本相似度效果图

本文把相似度控制在 0-1 范围内，所以我们采取 0.5 做为分界线，小于 0.5 的认为回复和留言的相关性较差，大于 0.5 的认为相关性较好。通过实验结果不难发现样本数据中留言和答复相关性比较高，可见政府的服务态度很好。

3.3.2 分析完整性指标

本文认为答复的完整性就是留言和答复的主题是否相关，即主题点数目是否相同。所以本文采用 LDA 进行数据实验。实验结果如图所示：

```
留言详情文本处理后呈现内容:['以来 位于 区 桂花 坪 街道 区 公安分局 宿舍区 景
答复文本处理后呈现内容:['现 网友 平台 问政 西地省 栏目 胡华衡 书记 留言 反映
LDA主题相似度:0.982
留言详情文本处理后呈现内容:['潇楚 南路 开始 修 到 快 路 挖 稀烂 围栏 围 起
答复文本处理后呈现内容:['网友 您好 反映 区 潇楚 南路 洋湖 没 修好 问题 区洋
LDA主题相似度:0.838
留言详情文本处理后呈现内容:['地处 省会 民营 幼儿园 小孩 是 祖国 民营 幼儿园
答复文本处理后呈现内容:['市民 同志 你好 反映 请 加快 提高 民营 幼儿园 教师
LDA主题相似度:0.919
留言详情文本处理后呈现内容:['尊敬 书记 您好 研究生 毕业 后 人才 新政 落户 想
答复文本处理后呈现内容:['网友 您好 平台 问政 西地省 上 留言 收悉 市 住 建局
LDA主题相似度:0.762
```

图 16 文本主题相似度效果图

为了实验的准确性，本文采用 JS 距离计算方式，效果图如图所示：

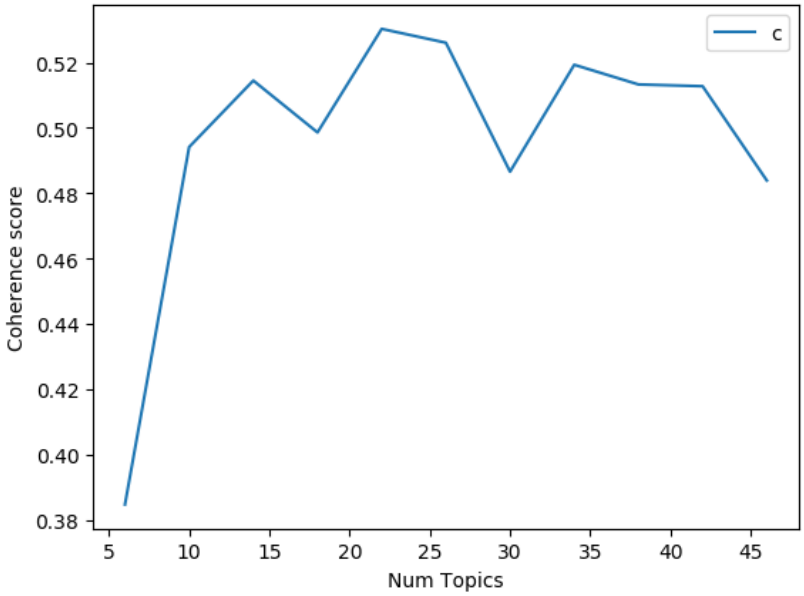

图 17 JS 距离计算方式效果图

由图可知，当 Num Topics = 22 时，coherence score 值最高，因此在我们的模型中设置主题数量为 k=22 时模型最佳。

3.3.3分析可解释性指标

由图 18，我们可以看出各个问题的态度值分布情况，发现几乎都是积极态度，说明政府都是比较重视公众留言内容的。

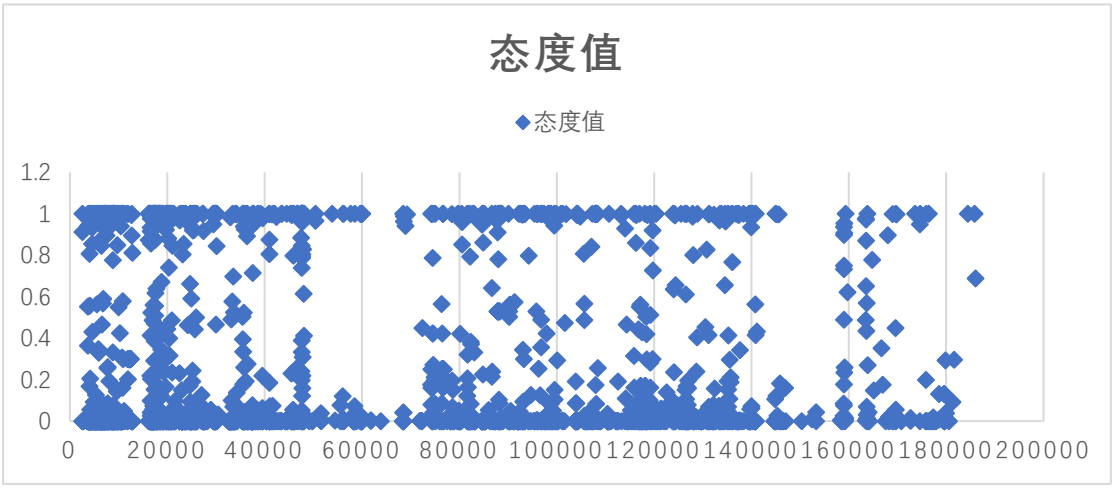

图 17 回复态度值分布图

```
由关键词构成的答复意见：业委会业主业主大会问题停车意见工作情况反映进行收费物业公司召开小区栏目调查答复书记街道理解
答复意见态度:0.012
由关键词构成的答复意见：施工项目道路处理渠道管线排水换填才能进行大道监督组织关心集镇调查问题通车土方坪塘
答复意见态度:0.010
由关键词构成的答复意见：待遇民办教师依法方面保障保险提高推进加强教职工监管确定规范市民支持意见养老保险实施同志
答复意见态度:0.979
由关键词构成的答复意见：购房补贴管理中心房屋交易购买商品毕业生西地省公寓博士住房交由情况建议事业单位问题办理机关还有试行
答复意见态度:0.025
```

图 18 回复态度值实验结果图

然后 ，本文将相关性、完整性和态度值作为三个指标利用改进的熵值法进行指标体系的建立。通过实验测试，三个一级指标权重对应值如下表所示：

表 7 三个一级指标权重分配表

	相关性	*完整性*	*态度值*
权重	0.087966	0.013067	0.898967

4 结论

对政府网站上的群众问政留言记录进行分析研究，了解群众关注社会实时热点，如何高效高质量答复群众留言意见对政府部门有重大意义，同时也是电子政务发展过程中要解决的一个问题。本文采用 jieba 中文分词工具对留言详情进行分词，利用逻辑回归（Logistic Regression）、随机森林分类（RandomForestClassifier)、朴素贝叶斯分类器（MultinomialNB)、线性支持向量机(LinearSVC)、决策树分类器(DecisionTreeClassifier)和 K-神经网络分类器（KNeighborsClassifier）六种分类器进行实验比较。同时针对热度问题采用 Hacker News、Reddit 和魔方秀热度算法统计每一条留言记录热度问题，并通过 K-means 聚类方法对留言主题聚类，统计群众实时关注的社会热点问题以及地区，并按照一级标签进行分类，深入了解不同领域内的群众关心的热点问题。

由分析结果可得，群众关注的社会热点问题主要分为交通运输、劳动保障、卫生计生、商贸旅游、城乡建设、教育文体、环境保护六个方面。统计某个时期内群众集中反映的热点问题，能帮助相关部门及时发现热点问题，进行针对性地处理，提升服务效率。

构建政府部门答复留言意见评价指标，能帮助相关部门有效回复群众留言意见，提升服务质量。

5 参考文献

[1]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.

[2]任远航.面向大数据的 K-means 算法综述[J/OL].计算机应用研究:1-7[2020-05-07].https://doi-org-s.vpn2.zufe.edu.cn:8088/10.19734/j.issn.1001-3695.2019.10.0581.

[3]艾楚涵,姜迪,吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术,2020,44(04):65-70.

[4]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.

[5]刘辉. 一种话题热度预测方法及系统[P]. 广东:CN106503209A,2017-03-15.