“智慧政务”中的文本挖掘应用

摘要

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

对于问题 1，首先对“留言主题”以及“留言详情”中的内容进行去重，并且去除缺失值。利用 jieba 分词对于留言内容进行分词，使用哈工大停用词表去除停用词，通过 TF-IDF 算法[1]提取每个文档中的 16 个关键词。对于每个文档中的关键词使用 one-hot 编码[2]，使用朴素贝叶斯算法[3]，先计算出各个类别关键词的先验概率，再利用贝叶斯定理计算出文档中各关键词属于某个类别的后验概率，通过选出具有最大后验概率估计值的类别即为最终的类别。

对于问题 2，由于计算文本相似度的时间复杂度非常大，因此使用 LDA 主题模型[4]对于热点问题进行归纳，聚类。通过 TF-IDF 算法提取关键词，并且使用 LDA 主题模型，发现文档集中的潜在语义结构，它是由文档、主题和词语组成的三层贝叶斯生成模型。由自己规定一个 70 主题数量去训练模型，得到规定数量的主题。由每一个主题下的 10 个高频词人为的规定其主题类别，再通过计算文档特征词属于哪一类别的概率最大，就将文档归为哪一类别，其中主题是根据 LDA 主题模型得到的 10 个高频词汇来人工指定主题属于哪一个类别。并且将热点问题生成表格。

对于问题 3，通过计算留言内容与答复的相似度来计算相关性，根据答复内容相关的话术，构建专家词典，计算专家词典与答复内容的相似度来判断答复内容的可解释性与完整性，其中相似度使用余弦相似度进行计算。由于 one-hot 编码的词向量只包含了 0 和 1 数据，无法对其进行相似度计算，且 one-hot 编码没有考虑文章分词的顺序，忽略了整句话的语义信息，在本问题中，使用 word2vec[5]对于留言内容和答复内容进行向量化，对于每一个分词构成一个 80 维的向量空间，通过词向量的平均值来表达句向量。对于分词不再提取关键词，只需要去除停用词，因为 word2vec 向量化需要结合上下文语境，才能给分词较为完整清晰的向量。通过计算所有答复内容的相关性、完整性、可解释性，对于得到的所有相似度相加，并且归一化，通过相关性高低来判断答复内容的质量。

关键词： 中文分词，TF-IDF 算法，朴素贝叶斯算法，LDA 主题模型，word2vec 词向量，相似度计算

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that mainly rely on manual work to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and othe technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in

promoting the management level and efficiency of the government.

Aiming at the problem of the one, firstly, the content in "message subject" and "message details" is de duplicated and the missing value is removed. In this paper, we use the Jieba segmentation to segment the message content, use the halting words list of Harbin University of technology to remove the halting words, and extract 16 keywords in each document by TF-IDF algorithm. For each keyword in the document, one hot coding is used, and naive Bayesian algorithm is used to calculate the prior probability of each category of keywords, and then the posterior probability of each keyword in the document belonging to a certain category is calculated by Bayesian theorem. The final category is selected by selecting the category with the maximum a posteriori (MAP) estimation value.

Aiming at the problem of the two, Because the time complexity of computing text similarity is very large, LDA topic model is used to summarize and cluster hot issues. TF-IDF algorithm is used to extract keywords, and LDA topic model is used to discover the potential semantic structure of document set. It is a three-layer Bayesian generation model composed of documents, topics and words. By setting a 70 topic number to train the model, we can get a set number of topics. The subject category is defined by 10 high-frequency words under each topic, and then by calculating which category the document characteristic words belong to, the document will be classified into which category. The topic is to manually specify which category the subject belongs to according to the10 high-frequency words obtained from LDA subject model. And generate tables of hot issues.

Aiming at the problem of the three, The relevance is calculated by calculating the similarity between the message content and the reply. According to the relevant scripts of the reply content, an expert dictionary is constructed, and the similarity between the expert dictionary and the reply content is calculated to determine the interpretability and integrity of the reply content. The similarity is calculated by cosine similarity. Because the word vector of one hot code only contains 0 and 1 data, it can't calculate the similarity, and the one hot code doesn't consider the word segmentation order of the article, and ignores the semantic information of the whole sentence. In this problem, word2vec is used to quantify the message content and the reply content. For each word segmentation, an 80 dimensional vector space is constructed, and the average value of the word vector is used To express the sentence vector. For word segmentation, we don't need to extract key words any more, we only need to remove the stop words, because word 2vec vectorization needs to combine with context to give a more complete and clear vector of word segmentation. By calculating the relevance, integrity and interpretability of all the replies, adding all the similarity, and normalizing, the quality of the replies can be judged by the relevance.

Keyword: Chinese word segmentation, TF IDF algorithm, Naive Bayesian algorithm, LDA topic model, word2vec word vector, similarity calculation

目录

1. 挖掘目标

本次建模目标是利用各类社情民意相关的文本数据，利用中文分词、特征词提取、朴素贝叶斯算法、LDA 主题模型、word2vec 词向量化对于文本数据进行挖掘，达到以下三个目的：

1）利用文本分词、特征词提取以及朴素贝叶斯算法对非结构化的数据进行文本分类，通过已有的一级标签训练模型，从而对各类社情民意相关的文本数据进行分类，以便后续将群众留言分派至相应的职能部门处理，进而提升政府部门进行对口问题处理，提升工作效率。

2）根据群众社情民意留言内容，利用 LDA 主题模型挖掘其潜在语义，即文本的主题内容，再通过 LDA 主题模型对于留言内容进行分类，寻找出热点问题。其能反映出某一时段内群众集中反映的一些问题。及时发现热点问题，有助于相关部分进行有针对性的处理，提高服务效率，提升人民的幸福感。

3）利用 word2vec 对于留言内容和答复意见的分词向量化，再利用词向量平均值构建句向量，自己构建专业话术的专家字典，通过计算留言内容与答复意见余弦相似度得到相关性的数值，通过计算答复意见与专家字典中的专业话术预想相似度得到答复意见完整性与可解释性数值，两个数值相加并且归一化，得到答复内容的质量评价标准。

2. 分析方法与过程

2.1 问题 1 分析方法与过程

2.1.1 流程图

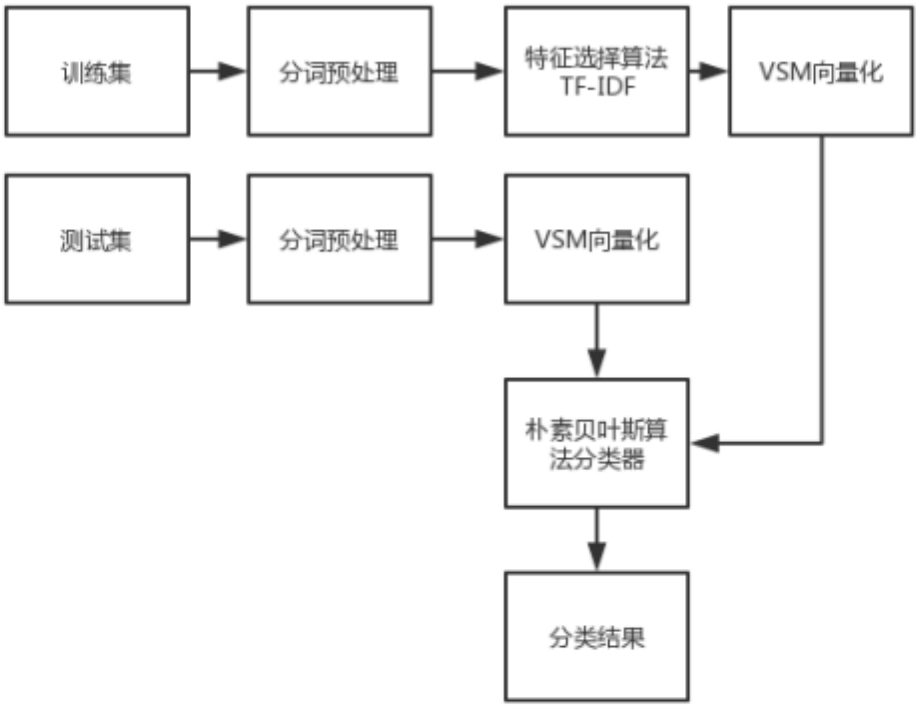

图 1.问题 1 流程图

2.1.2 数据预处理

2.1.2.1 留言内容的去重去空

附件 2 中给出的文本数据，可能会有留言内容为空或者重复现象存在。利用 python 自带函数对于空值以及重复值进行删除，避免空值以及重复值影响到朴素贝叶斯算法中的先验概率，以达到精度更高的目的。

2.1.2.2 对于留言主题以及留言详情进行中文分词

由于文字内容为非结构化的数据，因此需要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。附件 2 中给出的数据，不仅含有留言内容还有留言详情，都对反映的问题进行了描述，因此建立模型的时候需要将两个内容都考虑进去，本文之后将这两块内容统称为留言内容。采用 python 中的 jieba

库对于留言内容进行分词，其中 jieba 库中使用贪婪算法对于留言内容进行分词，是为了达到最大匹配效果，避免精准分词造成没有必要的语义损失。分词过程中一些没有意义的虚词和标点符号，包括如“的”，“啊”，“不但”等之类的词，对于文本语义并没有贡献，需要除掉（注：去除停用词原本需要将’0-9‘，’a-zA-Z’这样数字和字母去掉，但是考虑到留言内容的特殊性，比如“A5 区”表示一个特定的地名，因此停用词中去除数字以及字母）。正如上文所说，含有特殊意义 的词 jieba 分词并不能识别，可能将一个完整的词分为若干个，因此需要人为指定专有名词，防止专有名词被 jieba 分词的时候分开，如 “A5 区”避免被分为’A5 ‘，’ 区 ‘。停用词表以及专有名词详情请见 stopword.txt 和 dict.txt。

2.1.2.3 TF-IDF 算法

词频（Term Frequency，TF），即特征词在文档中出现的频数，权重的计算即为特征词在文档中出现的频数，所以对于特征词 t_k ，其在文档 D_i 中的权重为式 2.1。

$$Wt_k = TF(t_k, D_i) \tag{2.1}$$

反文档频率法（Inverse Document Frequency，IDF）也称为逆文档频率，其思想为如果特征词在某一类文档中出现较多，则其对于该类文档具有更高的分类能力。反文档频率的计算公式如式 2.2 所示。

$$IDF(t_k) = \log(\frac{N}{n(t_k)} + 0.01) \tag{2.2}$$

TFIDF 权重是在文本分类中运用比较多的计算方法，是由 Salton 提出。TFIDF 算法的思想：特征单词在某特定文本中出现的频数越大，其对于该文本的分类作用越大，特征单词在大多数文档中出现的频数越大，对于文本的分类作

用越小，其结合了词两者的优点。TFIDF 算法将词频和反文档频率结合作为特征的权重，计算方法公式 2.3，2.4 所示。

$$IDF(t_k)=\log(\frac{N}{n(t_k)}+0.01) \tag{2.3}$$

$$W_k=TF(t_k)*IDF*(t_k)=\frac{TF(t_k)*IDF*(t_k)}{\sqrt{\sum_{i-1}^{m}(TF(t_k)*IDF*(t_k))^2}} \tag{2.4}$$

其中 TF(t_k）为特征 t_k 在训练集中出现的频数，IDF(t_k)是反文档频率，N 表示训练集的总文档数，n(t_k）表示出现特征 t_k 的文档数，TF-IDF 算法考虑了特征词的局部和全局的分布特性。在本文，通过 TF-IDF 算法选出权重最大的 16 个特征词，进行向量化，并且建立模型。

2.1.2.4 文本向量化

对于提取出来的特征词，使用 one-hot 编码，即将所有出现过的特征词构建为一个词典，该词典中，每个特征词只出现过一次，对于每一篇文档中的特征词在词典中出现该特征词的位置记为 1，没有出现过的位置记为 0，由此将每一篇文档向量化为一个长度很长只包括 0 和 1 元素的列表，以便于模型的构建。

2.1.3 构建模型并分类

2.1.3.1 朴素贝叶斯模型

将文本向量化之后，就可以利用朴素贝叶斯模型对于留言内容进行分类。朴素贝叶斯有很多优势，主要包括其独特形式、丰富的概率表达能力、综合先验知识的增量学习特性，其简洁性和有效性都要优于其他算法[19，32-35]。所以问题 1 重点使用朴素贝叶斯算法。朴素贝叶斯算法的思想是首先计算出各个类别的先验概率，再利用贝叶斯定理计算出各特征属于某个类别的后验概率，通过选出具有最大后验概率（maximum a posteriori, MAP）估计值的类别即为最终类

别（注：其本质含义是为了使损失函数达到最小）。

本文使用的是伯努利朴素贝叶斯算法模型。伯努利模型[64]（Bernoulli Naïve Bayes，简称 BNB）认为一个事件有两种可能性，发生或者不发生。当进行 n 次独立重复的伯努利试验，会产生一个新的分布称为二项分布。对于一篇文档而言，，词典中的每个典中的一个特征词可以看作是进行一次伯努利试验，而词典中的所有特征词可以看作 n 重伯努利试验，就是二项分布。对于一篇文档 d,我们已经将其向量化表示为[t1,t2,t3,t4...tm],$tk \in \{0,1\}$，其中 tk=1 表示为该特征词在文档中出现过，反之则未出现，m 表示词典的大小 t.为了处理文本数据，朴素贝叶斯的一个主要假设是在给定文档类别的情况下，每个单词条件概率计算是相互独立的，在此假设下，BNB 模型可以使用公式 2.5 来预测文档 d 的类：

$$c(d)=\arg\max_{c\in C} p(C_j)\prod_{i=1}^{m}\left(t_k p(Dt_k|C_j)+(1-t_k)\left(1-p(Dt_k|C_j)\right)\right) \quad (2.5)$$

其中 $p(C_j)$表示先验概率，$p(Dt_k|C_j)$表示条件概率，可以采用频数计数近似估计值，计算公式如 2.6，2.7：

$$p(C_j)=\frac{tf(D,C_j)}{\sum_{j=1}^{N} tf(D,C_j)} \quad (2.6)$$

$$p(Dt_k|C_j)=\frac{tf(Dt_k,C_j)}{tf(D,C_j)} \quad (2.7)$$

其中 $tf(Dt_k,C_j)$表示含有单词 t_k 的文档在 C_j 类出现的文档数，tf（D,C_j）为 C_j 类中所有文档数，为避免概率计算时出现 p（$Dt_k|C_j$）为 0 的情况采用拉普拉斯平滑，具体计算公式如 2.8：

$$P\left(Dt_k \mid C_j\right)=\frac{\sum_{i-1}^{n} t_{ki}\delta\left(C_j, C\right)+1}{\sum_{i-1}^{n}\delta\left(C_j, C\right)+2} \tag{2.8}$$

其中δ（)表示二值函数公式如 2.9：

$$\delta(x, y)=\begin{cases}1, x=y\\ 0, x\neq y\end{cases} \tag{2.9}$$

2.1.3.2 根据留言主题和留言内容加权计算后验概率进行分类

由于使用的伯努利朴素贝叶斯算法模型，文档中的特征词在词典中只会出现一次，而包含了文档特定含义的词可能会多次出现，导致算法在 F1 上表现不是很理想。因此在本文中，没有将留言主题和留言内容合并在一起提取特征词，而是分别使用留言主题和留言内容的文档在朴素贝叶斯模型中计算最大后验概率，以解决伯努利模型所存在的不足，两部分内容都表示了归于哪一类的后验概率，对于得到的两个后验概率进行加权计算后再取每一个文档的最大后验概率，从而得到文档具体属于哪一类别，经实验证明，本文使用的加权算法使计算结果在 F1 上得到了提升，并且经实验证明，当权重为 1：1 时，F1 的值最大。流程如下图所示。

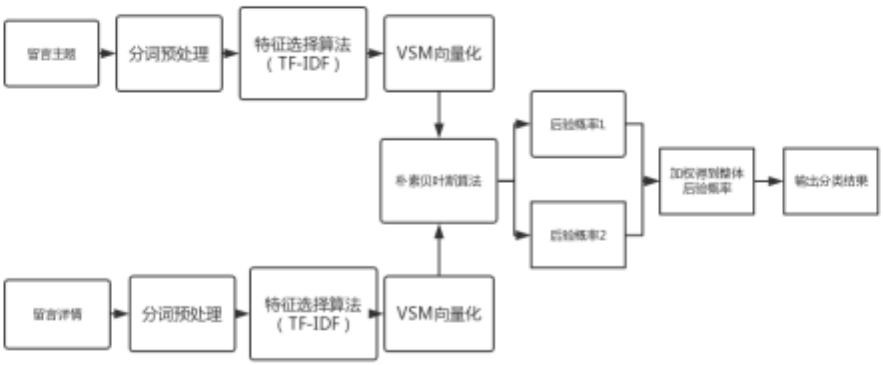

图 2.加权分类流程图

2.2 问题 2 分析方法与过程

2.2.1 流程图

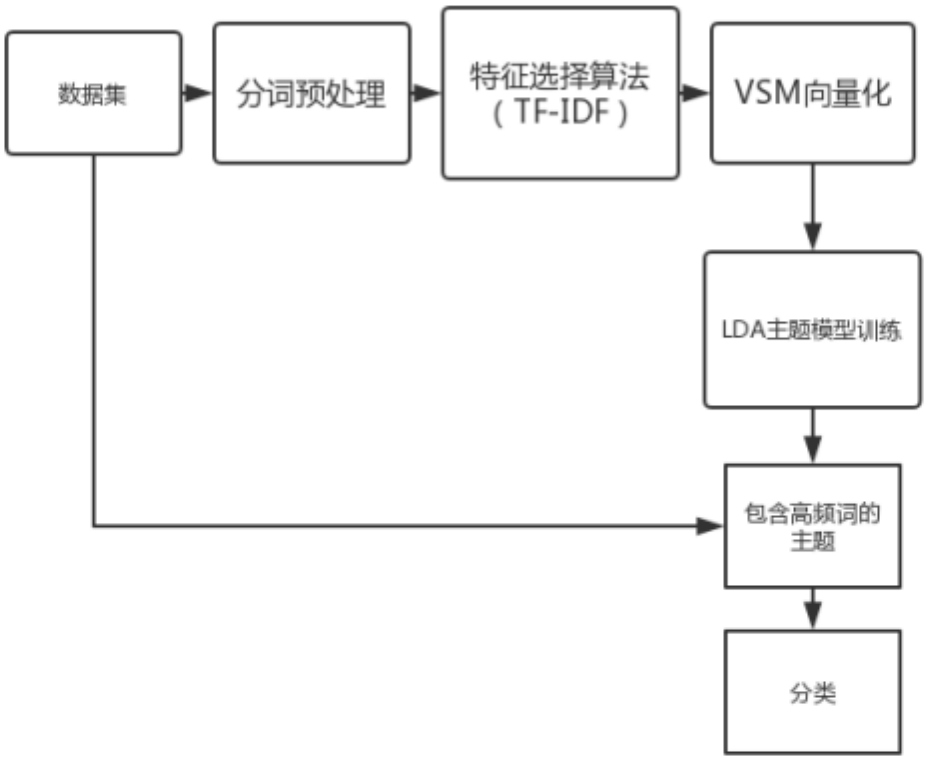

图 3.问题 2 流程图

2.2.2 数据预处理

同步骤 2.1.2.1-2.1.2.3 一致将留言主题中的文档数据向量化表示为 one-hot 编码，对于 LDA 主题模型而言需要将输入数据变为[[(0, 1), (1, 1), (2, 1)]形式，其中每一个元组中的第一元素个代表该关键词的位置信息，第二个元素表示该关键词出现次数，因为使用 one-hot 编码，第二个元素一致为 1，即出现一次。

2.2.3 构建模型并分类

2.2.3.1 LDA 主题模型

由于问题 2 中需要对热点问题进行挖掘，其中留言内容高达 5000 条，逐条

计算文档相似度时间复杂度较高，因此该问题考虑使用文档主题模型进行分类。文档主题模型（Topic Model）是挖掘大规模文档集或语料库中隐藏的潜在主题的一种无监督机器学习统计模型，在电商推荐系统、社交网络话题识别和新闻信息主题聚类等自然语言处理领域中应用广泛[14]。LDA(Latent Dirichlet Allocation）是一种生成式模型，也是一个三层的贝叶斯概率模型，由词、主题及文档三层结构构成。其核心公式如公式 2.10 所示

$$P(词|文档) = P(词|主题)P(主题|文档) \quad (2.10)$$

LDA 主题模型隐含狄利克雷分布（Latent Dirichlet Allocation,LDA），其原理是基于词袋模型，认为文档 d 与文档中词语 W 之间存在中间层主题 Z，且文档是主题的概率分布，主题又是词的概率分布，由此可将高维度的文档-词项向量空间模型映射为低维度的文档-主题和主题-词项空间，进而挖掘文档中潜在蕴含的若干主题。文档的层级关系见图 4。

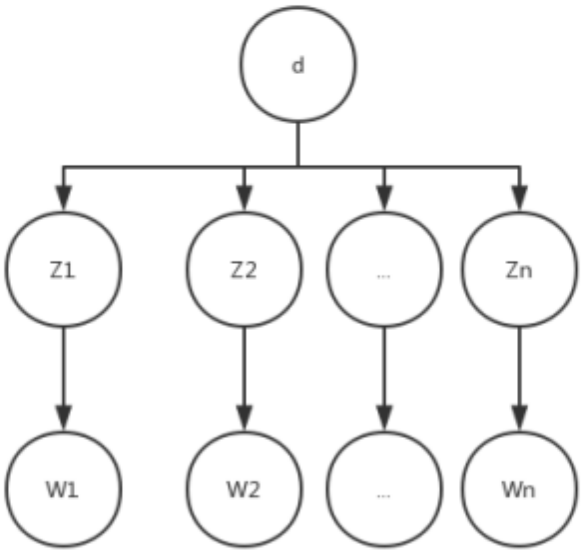

图 4.主题模型的文档结构

文档的矩阵转换关系见图 5。其中，矩阵 C 表示文档中的词语概率分布，矩阵Φ表示主题下的词语概率分布，矩阵θ表示文档下的主题概率分布，而分析主题模型的目的在于通过解析文档 C 得到矩阵Φ和矩阵θ。

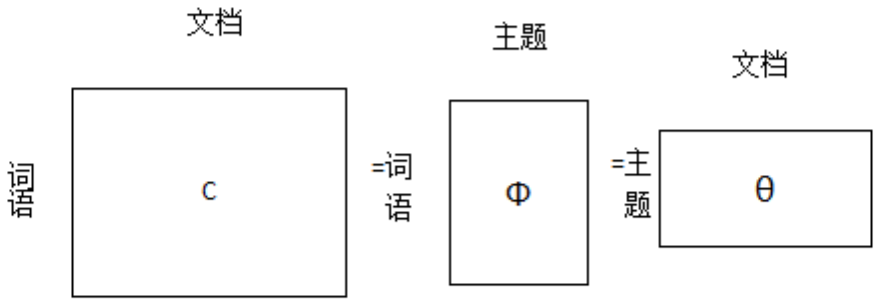

图 5.主题模型的矩阵转换关系

2.2.3.2 根据 LDA 主题模型进行分类

通过 python 中的 gensim 库对于我们提取到的关键词进行 LDA 主题模型训练，指定 150 个主题，训练之后对于每一个主题提取 10 个高频词，根据高频词来确定该主题的类别，通过模型训练获得每个文档的主题分布和每个主题的词分布。

1) 利用训练好的模型对于留言内容进行分类。

2) 根据该主题的高频词确定该类别，也就是热点问题的类别。

3) 选出数量最高的五个类别即为集中性最高的五个热点问题。

2.3 问题 3 分析方法与过程

2.3.1 流程图

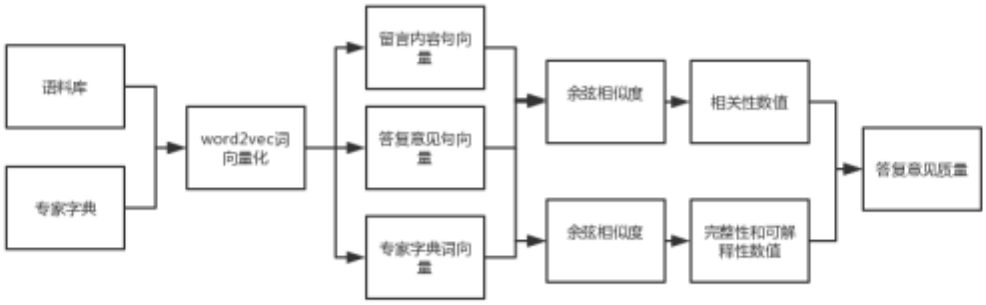

图 6.问题 3 流程图

2.3.2 问题 3 分析过程

2.3.2.1word2vec 分布式表示

使用 gensim 的 word2vec 对于分词进行向量化表示。选用 CBOW 模型，如图 7 所示。CBOW 模型是给定上下文来预测输入分词。Word2vec 模型实际上分为两部分，第一部分为建立模型，第二部分是通过模型获取嵌入词的词向量。假设词向量的维数为 dk，每条评论文本可以表示为一个行数是词向量的维度 dk，列数是评论文本长度 N 与主题特征词的个数 l 之和的文本矩阵 W。其中 w 为评论文本的词向量表示，wz 为通过 LDA 获得该评论文本的主题特征词的向量表示。CBOW 模型损失函数如下公式 2.11 所示[6]：

$$L(W) = 1N+|\Sigma|c|\leq s, c\neq 0 lbP\ (wi\ |\ wi\text{-}s,\ldots,\ wi+s) \qquad (2.11)$$

其中 wi 为某个中心词，s 为中心词左右窗口大小，P（wi|wi-s,...,wi+s）一直上下文中心词为 wi 的概率大小计算方法为 2.12：

$$P(wi\ |\ wi-s,\ldots,wi+s) = \exp(w0Twi)\sum w\in dictexp\ (w0Tw) \qquad (2.12)$$

其中 w0 是 wi 上下文词向量的均值，dict 为字典。

$$w0 = 12s\sum j=i-s,\ldots,i+s, j\neq iwj \qquad (2.13)$$

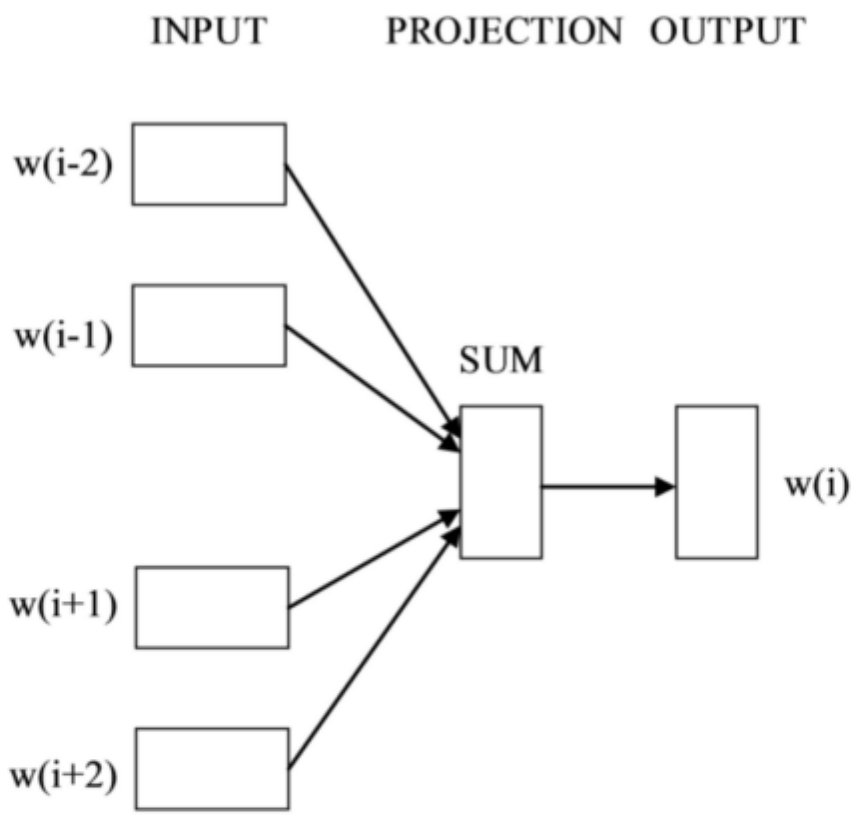

图 7.CBOW 模型图

3. 结果分析

3.1 问题 1 结果分析

3.1.1 分类分析

通过去重去空后对留言内容和留言详情进行分词，提取 29 个关键词作为特征值，利用朴素贝叶斯算法进行分类。其中使用 5 倍交叉验证法进行验证，取五分之四为训练集，五分之一为测试集验证算法准确性。需要调整的参数如下：

1）特征词提取

通过实验证明当特征词为 29 时，算法在 F1 上表现最好，因此特征词选用为 29，如 8 图所示：

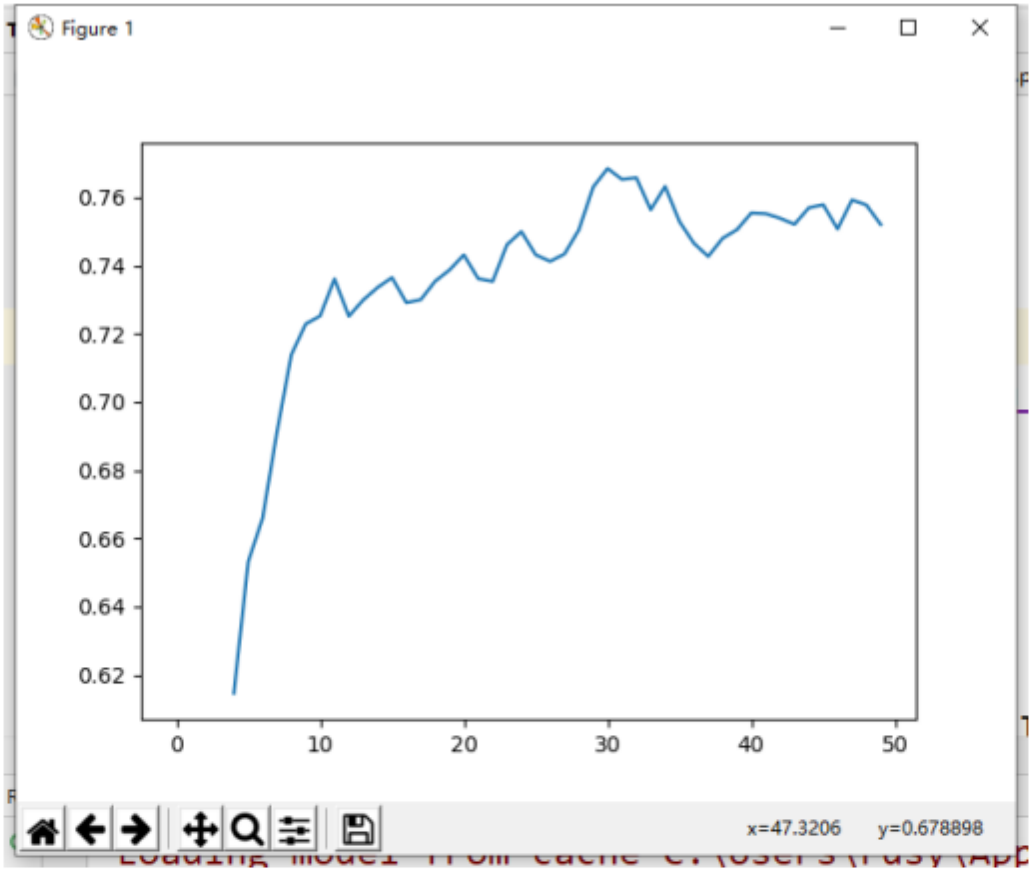

图 8.不同特征值数量对结果的影响

2）拉普拉斯平滑参数

为避免出现 0 值导致算法不能较好的建模，需要在分子以及分母上加入拉普拉斯算子，经本实验整明当分子的拉普拉斯算子取为 0.3，分母取为 0.7 时，算法效果达到最优，拉普拉斯平滑如下式 3.1 所示：

$$P\left(Dt_k \mid C_j\right)=\frac{\sum_{i-1}^{n} t_{ki}\delta\left(C_j, C\right)+1}{\sum_{i-1}^{n}\delta\left(C_j, C\right)+2} \tag{3.1}$$

3）留言主题与留言详情的加权系数

为避免伯努利模型忽略多次出现的关键词影响准确性，将留言详情和留言主题分别计算后验概率，并且加权得到最终的后验概率，经本实验证明，当系数为 1：1 时算法效果达到最好。

3.1.2 分类结果

选定好参数之后，使用五倍交叉验证计算查准率查全率的平均值，以及 F1 的值，实验结果如表 1 所示：

名称	P 查准率	R 查全率	F1
数据	78.7%	67.4%	72.3%

表 1.测试结果

其中前三个类别在 P 上表现不好，后三个类别在 R 上表现不好，其中城乡建设、交通运输、环境保护类多数关键词一致，基于朴素贝叶斯的算法对于相似词的分别能力较低，导致正确率偏低，城乡建设中多涉及环境保护和交通运输类，建议相关部门对于这三类可以做好协调工作。该三类的词云图如图所示，也可以直观看出关键词相似度较高。

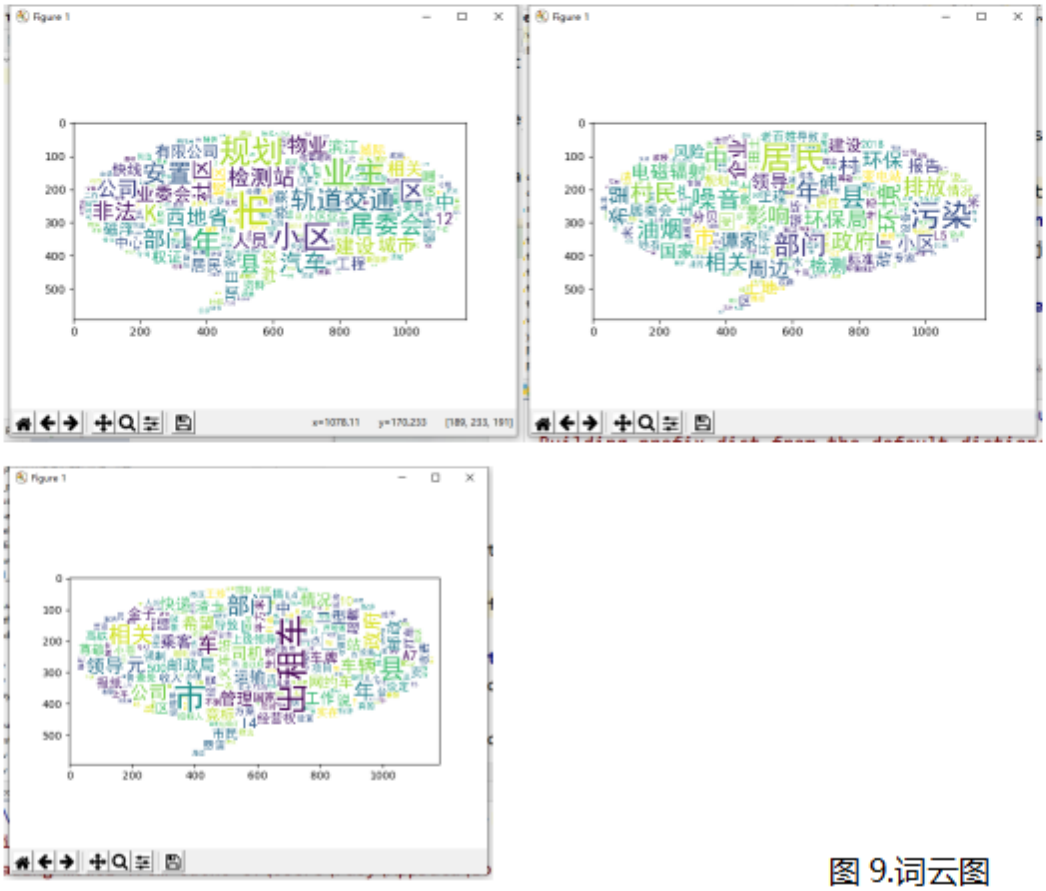

图 9.词云图

3.2 问题 2 结果分析

3.2.1 分类分析

使用 LDA 主题模型，依然使用留言主题的 29 个关键词进行训练，得到不同的主题，以及高频词。

1）主题数量

在测试数据阶段，50 个文档，使用十个主题达到很好的效果，对于全部数据达到 5000+文档，通过实验证明当指定主题数量达到 70 时，高频词语义较为清晰，可以明确指定出主题为哪一类别。

2）高频词数量

当指定 70 个主题数量之后，通过实验发现，通过频率最高的 10 个高频词，可以明确得到地点以及热点问题的描述。如图 10 所示，再根据高频词指定热点问题的地点以及类别：

```
"'), (3, '0.055*"A市" + 0.037*"核查" + 0.034*"兰亭" +
0.023*"手续" + 0.023*"车辆" + 0.022*"相关" + 0.022*"民警" +
 0.021*"职能部门" + 0.020*"司机" + 0.020*"50"'), (4,
'0.054*"小区" + 0.042*"夜间" + 0.042*"居民" + 0.027*"A市" +
 0.027*"二期" + 0.027*"开窗" + 0.026*"夜宵" + 0.022*"停车位"
```

图 10.主题高频词

3.2.2 分类结果

通过 LDA 主题模型对于文档进行分类使用字典保存分类类别，以及每一个类别共有多少文档，如下图 11 所示，其中第 34 类，46 类，36 类，4 类，31 类的文档数最多，即为热点问题，通过索引值将文档提取出并保存为热点问题留言明细表，并且将最多的五个类别，通过归纳总结，保存为热点问题表。

```
Counter({34: 109, 46: 109, 36: 108, 4: 98, 31: 94, 61: 86,
 30: 82, 42: 82, 19: 76, 6: 76, 62: 75, 67: 74, 40: 73, 18:
  73, 26: 73, 41: 72, 16: 72, 8: 72, 21: 72, 56: 71, 14:
 71, 0: 69, 50: 68, 29: 68, 13: 67, 2: 66, 5: 65, 38: 65,
 59: 65, 37: 63, 20: 63, 25: 63, 53: 62, 27: 61, 1: 61, 52:
  60, 11: 60, 15: 60, 66: 59, 57: 59, 32: 58, 47: 58, 54:
 57, 45: 57, 69: 56, 24: 55, 63: 54, 12: 54, 33: 54, 68:
 53, 48: 50, 22: 50, 60: 50, 10: 50, 51: 49, 35: 49, 28:
 48, 44: 48, 3: 47, 39: 47, 49: 45, 43: 43, 55: 41, 65: 41,
  23: 40, 58: 36, 17: 35, 64: 29, 7: 27, 9: 23})
```

图 11.热点问题数量以及索引字典

3.3 问题 3 结果分析

3.3.1 词向量化

对于留言内容以及答复意见进行停用词去除之后，通过 jieba 分词将文本信息分别放入 word2vec 中进行训练，指点向量维度为 80，上下滑动窗口为 2，低频词过滤设置为 0，即不过滤低频词。得到每个单词的词向量，并且将训练所得到的词向量保存以便后续调用，再通过词向量的平均值得到句向量，如下图 12 所示:

```
[array([0.16943664, 0.16943664, 0.16943664, 0.16943664,
 0.16943664,
       0.16943664, 0.16943664, 0.16943664, 0.16943664,
 0.16943664,
       0.16943664, 0.16943664, 0.16943664, 0.16943664,
 0.16943664,
       0.16943664, 0.16943664, 0.16943664, 0.16943664,
 0.16943664,
```

图 12.句向量

其中对于答复意见的分词来讲，前 10 个分词不放入模型进入训练，因为是

官方回答的话术，所有答复意见的开头都是使用的同一个话术，因此对于该部分内容进行去除操作，避免影响词向量化的准确性。

3.3.2 专家字典

通过对留言内容的分析，比如一些有转折性的词，或者一些特殊的术语，如果出现则说明其较为完整或可解释性较高，对于这部分词不需要再放入模型中训练，可以直接从上面已经保存的 word2vec 模型中调用该部分词的向量，进而和句向量进行相似度计算。详情请见 three_dict.txt。

3.3.3 质量评价

通过对答复意见和留言内容的相似度以及和专家字典的相似度求和，并且归一化得到答复意见的评价指标，归一化之后再将评价指标乘 100 倍，变为百分制，在附件 4 最后一列填加答复意见质量评价标准，0~60 评价为标准，60~90 评价为良好，90~100 评价为详细。效果如图 13 所示：

答复意见	答复时间	回复评价
车管理费，在业主大会结束后	2019/5/10 14:56:53	标准
换填，且换填后还有三趟雨	2019/5/9 9:49:10	详细
聘任教职工要依法签订劳动	2019/5/9 9:49:14	良好
岁以下（含），首次购房后	2019/5/9 9:49:42	详细
坡岭”的问题。公交站点的	2019/5/9 9:51:30	标准
中没有说明卫生较差的具体	2019/5/9 10:02:08	详细
人民政府办公室下发了《关	2019/5/9 10:18:58	详细
完成教学任务，但须征得学	2018/8/20 9:25:34	标准
当事人双方以及茨菇塘办事	2018/8/17 9:49:43	良好
您拨打12345市长热线，可	2018/8/7 9:13:42	详细
。二、厂房并没有加长，老	2018/4/16 11:33:45	标准

图 13.附件评价内容

4. 结论

在文本分类中，朴素贝叶斯虽然效果比较明显，但是对于相似程度较高的文本无法完成正确的分类效果，仍需结合上下文语境来进行分类，相关部门在阅读留言过程中也需要认真细致。对于热点问题，有一些称不上是热点的问题，但是

仍然是社会学校上面的主要矛盾，相关部门也需要针对社会上已存在的主要矛盾对于问题进行排查管理，提升政府部门管理全面性，主动性。政府部门的回复内容比较及时，相关性和完整性较强，但是有的问题依然需要和群众多沟通多交流，增加政府与群众面对面的机会，更加把握民意民情，创造一个和谐美满的社会大家庭。

5. 参考文献

[1]. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[M]. Pergamon Press, Inc. 1988.

[2]. Salton G , Wong A , Yang C S . A Vector Space Model for Automatic Indexing[J]. Communications of the Acm, 1974, 18(11):613-620.

[3]. 贺鸣, 孙建军, 成颖. 基于朴素贝叶斯的文本分类研究综述[J]. 情报科学, 2016, V34(7):147-154.

[4]. SwapnilHingmire, SandeepChougule, GirishK, etal. Documentclassification by topic labeling[C]//In SIGIR, 2013:877-880.

[5]. Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[6]. 宁建飞, 刘降珍. 融合Word2vec与TextRank的关键词抽取研究[J]. 现代图书情报技术, 2016(6):20-27.