

---

# “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。因此，运用网络文本分析和数据挖掘技术对群众留言信息的研究具有重大的意义。

问题一，附件二的群众问政留言记录的原始数据可能存在缺失值、异常值、重复值，为了保证其数据的质量和全面性，需要对数据预处理和清洗。首先对留言编号、留言用户，留言主题等信息进行异常排除、去空、去重，过滤得到不重复的留言信息。由于我们对留言信息处理时需要找出关键词进行分析，所以需要留言记录进行中文分词、过滤停用词。首先选择特征词并构造矩阵文本，将构造好矩阵文本进行 Word2Vec 训练词向量，再结合群众问政留言详情特征来选取适合的合成方式形成每一个主题的特征矩阵表示。利用支持向量机算法对合成矩阵输入，其次通过训练集进行训练模型，测试集评估分类效果。最后根据具体的映射规则实现留言详情里的特征词分类到一级分类里的城乡建设、国土资源、党务政务、环境保护等标签，从而实现留言到一级标签的自动匹配。

问题二，对附件3的群众留言主题利用 jieba 进行分词处理，然后统计词频，构建高词频，再根据留言主题对特定人群或者地点构成留言索引，利用高频词和索引构建特征矩阵，然后用聚类的机器学习算法分类，再次找出类别计数最多的留言做清洗，最后可以筛选出排名前五的热点问题。

问题三，对附件4的留言详情、答复意见分别进行数据预处理，进行类别划分和关键词提取，得到相应的文本数据，而相关性分析模型的数据为数据型数据，采用自然语言处理领域中的词向量作为表达词语的数值型特征。用皮尔逊相关系数得出留言主题与相关部门对留言的答复意见的相关程度。用实体完整性约束结合实际情况将答复留言的开端与事件处理过程设为主属性，最后判断答复意见是否完整。

关键词：文本分类、支持向量机、匹配模型、特征矩阵

## Abstract

In recent years, with wechat, Weibo, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand public opinion, gather people's wisdom and gather people's spirit. Therefore, the use of network text analysis and data mining technology is of great significance to the research of the mass message information.

Question 1: the original data recorded in the mass political message in Annex 2 may have missing value, abnormal value and repetitive value. In order to ensure the quality and comprehensiveness of the data, it is necessary to preprocess and clean the data. First of all, the message number, message user, message subject and other information are excluded, emptied and de duplicated, and non repetitive message information is filtered. Because we need to find out key words to analyze when we process message information, we need to segment Chinese words and filter stop words for message records. First, select the feature words and construct the matrix text, and then construct the matrix text for word2vec training word vector, and then combine the characteristics of the masses' political message details to select the appropriate synthesis method to form the feature matrix representation of each theme. Support vector machine algorithm is used to input synthesis matrix, then training model is carried out through training set, and test set is used to evaluate classification effect. Finally, according to the specific mapping rules, the feature words in the message details are classified into the tags of urban and rural construction, land and resources, party affairs and government affairs, environmental protection, etc. in the first level classification, so as to realize the automatic matching of the message to the first level tags.

The second problem is to use Jieba to segment the message subject in Annex 3, then count the word frequency and build a high word frequency, then build a message index for specific people or places according to the message subject, build a feature matrix with high frequency words and index, then use clustering machine learning algorithm to classify, find out the message with the most category count again for cleaning, and finally filter out the row Top five hot issues.

Question 3: preprocess the message details and reply opinions in Annex 4, classify them and extract keywords to get the corresponding text data. The data of correlation analysis model is data-based data, and the word vectors in natural language processing field are used as the numerical characteristics of expression words. Pearson correlation coefficient is used to get the correlation degree between the subject of the message and the response of relevant departments to the message. Based on the entity integrity constraint and the actual situation, the beginning of reply message and the process of event processing are set as the main attributes, and finally the integrity of reply opinion is judged.

Keywords: text classification, support vector machine, matching model, feature matrix

# 目录

一、 第一道题的分析与求解.....	1
1.1 用正则表达式对特殊字符进行处理 .....	1
1.1.1 对附件二的数据进行去空去重处理.....	1
1.2 对群众问政留言进行中文分词并过滤停用词 .....	1
1.2.1 分词 .....	2
1.2.2 停用词.....	4
1.3 特征矩阵构造.....	6
1.4 留言主题分类.....	6
1.5 留言特征词与类别匹配.....	7
二、 第二题的分析与处理.....	10
2.1 对附件 3 的分词处理 .....	10
2.2 对附件 3 词频统计与绘制词云图 .....	11
2.3 热度评价指标.....	12
2.4 聚类的机器学习算法 .....	12
三、 第三题的分析.....	14
四、 参考文献 .....	15

# 一、 第一道题的分析与求解

## 1.1 用正则表达式对特殊字符进行处理

### 1.1.1 对附件二的数据进行去空去重处理

由语言的特点可知，在大多数情况下，不同民众之间的留言都不会出现完全重复，否则，这些留言一般都是毫无意义的。所以这种评论显然只有最早的评论才有意义（即只有第一条有作用）。

部分留言相似程度极高，可是在某些词语的运用上存在差异。此类留言可归为重复评论，若是删除文字相近留言，则会出现误删的情况。由于相近的留言也存在不少有用的信息，去除这类留言显然不合适。因此，为了存留更多的有用语言，本节针对完全重复的留言下手，仅删除完全重复部分，以确保尽可能保留有用的文本留言信息。同时字与字之间，段首与段尾等处存在多余且不必要的空格，我们需要对其进行去空。

对于附件二的群众问政留言共有 9210 条，经过文本去空去重以及只保留有价值的信息，共删除重复评论 305 条，剩余留言 8905 条。其结果如下图 1.1-1 所示：

```
----- RESTART: E:\python\1.1.py -----
原始数据的形状为: (9210, 6)
删除异常记录后数据的形状为: (8905, 6)
      留言编号  ...  一级标签
0          24  ...  城乡建设
1          37  ...  城乡建设
2          83  ...  城乡建设
3         303  ...  城乡建设
4         319  ...  城乡建设
...         ...  ...  ...
9204    9605390  ...  卫生计生
9205    11763690  ...  卫生计生
9206    11977590  ...  卫生计生
9208    15397690  ...  卫生计生
9209    16135190  ...  卫生计生
[8905 rows x 6 columns]
```

图 1.1-1 数据清洗

## 1.2 对群众问政留言进行中文分词并过滤停用词

1.2.1 分词

分词是文本信息处理的基础环节,是将一个单词序列切分成一个一个单词的过程。准确的分词可以极大的提高计算机对文本信息的是被和理解能力。相反,不准确的分词将会产生大量的噪声,严重干扰计算机的识别理解能力,并对这些信息的后续处理工作产生较大的影响。

汉语的基本单位是字,由字可以组成词,由词可以组成句子,进而由一些句子组成段、节、章、篇。可见,如果需要处理一篇中文语料,从中正确的识别出词是一件非常基础且重要的工作。

然而,中文以字为基本书写单位,词与词之间没有明显的区分标记。中文分词的任务就是把中文的序列切分成有意义的词,即添加合适的词串使得所形成的词串反映句子的本意,例子如表 1.2-1 所示。

表 1.2-1 中文分词例子

操作	内容
输入	我帮小明打饭
输出	我 帮 小明 打饭

当使用基于词典的中文分词方法进行中文信息处理时不得不考虑未登录词的处理。未登录词指词典中没有登录过的人名、地名、机构名、译名及新词语等。当采用匹配的办法来切分词语时,由于词典中没有登录这些词,会引起自动切分词语的困难。常见的未登陆词有命名实体,如“张三”、“北京”、“联想集团”、“酒井法子”等;专业术语,如“贝叶斯算法”、“模态”、“万维网”;新词语,如“卡拉 OK”“美刀”、“啃老族”等。

另外,中文分词还存在切分歧义问题,如“当结合成分子时”这个句子可以有以下切分方法:“当/结合/成分/子时”,“当/结合/成/分子/时”,“当/结/合成/分子/时”,“当/结/合成分/子时”。

可以说,中文分词的关键问题为:切分歧义的消解和未登录词的识别。

词典匹配是分词最为传统也最为常见的一种办法。匹配方式可以为正向（从左到右）或逆向（从右到左）。对于匹配中遇到的多种分段可能性（segmentation ambiguity），通常会选取分隔出来词的数目最少的。

很明显，这种方式对词表的依赖很大，一旦出现词表中不存在的新词，算法是无法做到正确的切分的。但是词表匹配也有它的优势，比如简单易懂，不依赖训练数据，易于纠错等等。

还有一类方法是通过语料数据中的一些统计特征（如互信息量）去估计相邻汉字之间的关联性，进而实现词的切分。这类方法不依赖词表，特别是在对生词的发掘方面具有较强的灵活性，但是也经常会有精度方面的问题。

分词最常用的工作包是 jieba 分词包，jieba 分词是 python 写成的一个分词开源库，专门用于中文分词，其有三条基本原理，即实现所采用技术。

基于 trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情所有构成的有向无环图（DAG）。jieba 分词自带了一个叫做 dict.txt 的词典，里面有 2 万多条词，包含了词条出现的次数（这个次数是作者自己基于人民日报图 1.2.-2 “有意见分歧” 切分生成的有向无环语料等资源训练得出来的）和词性。Trie 树是有名的前缀树，若一个词语的前面几个字一样，表示该词语具有相同的前缀，可以使用 trie 树来存储，trie 树存储方式具有查找速度快的优势。后一句的“生成句子中汉字所有可能成词情况所构成的有向无环图”意思是给定一个待切分的句子，生成一个如图 1.2-2 所示的有向无环图。

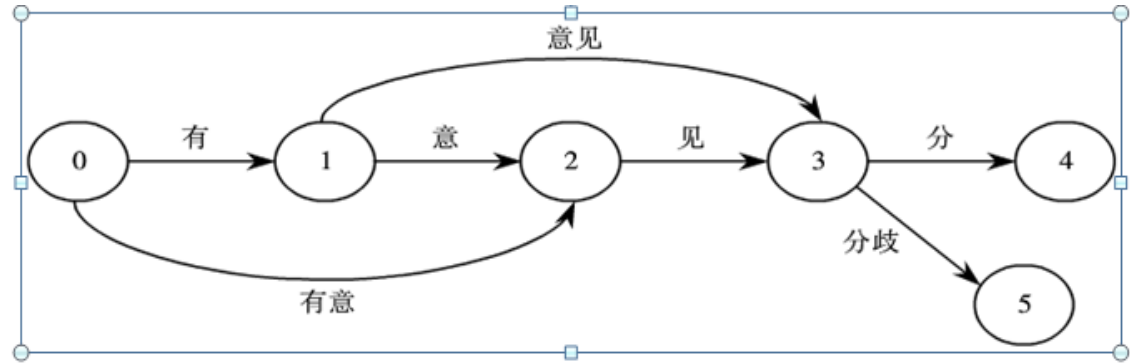


图 1.2-2 有向无环图

采用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。先查找待分词句子中已经切分好的词语，再查找该词语出现的频率，然后根据动态规划查找最大概率路径的方法，对句子从右往左反向计算最大概率（反向是因为汉语句子的重心经常落在右边，从右往左计算，正确率要高于从左往右计算，这个类似于逆向最大匹配），最后得到最大概率的切分组合。

于未登录词，采用 HMM 模型，使用了 Viterbi 算法，将中文词汇按照 BEMS 四个状态来标记。其中 B 是 begin，表示开始位置；E 是 end，表示结束位置；M 是 middle，表示中间位置；S 是 single，表示单独成词的位置。HMM 模型采用 (B, E, M, S) 这四种状态来标记中文词语，比如北京可以标注为 BE，即北/B 京/E，表示北是开始位置，京是结束位置，中华民族可以标注为 BMME，就是开始、中间、中间和结束。

下面是 jieba 分词的过程：

(1) 加载字典，生成 trie 树

(2) 给定待分词的句子，使用正则获取连续的 中文字符和英文字符，切分成短语列表，对每个短语使用 DAG(查字典)和动态规划，得到最大概率路径，对 DAG 中那些没有在字典中查到的字，组合成一个新的片段短语，使用 HMM 模型进行分词，也就是作者说的识别新词，即识别字典外的新词。

(3) 使用 python 的 yield 语法生成一个词语生成器，逐词语返回或者直接返回 list，得出的效果差不多一样。

### 1.2.2 停用词

留言文本在经过去重，去空、中文分词后，并非所有的剩下的词语都可以作为特征词，里面还有一些包含的信息量很低甚至没有信息量的词语，我们称之为停用词，需要将它们过滤掉，否则将会影响分析的正确率。

停用词 (Stop Words)，词典译为“电脑检索中的虚字、非检索用字”。在 SEO 搜索引擎中，为节省存储空间和提高搜索效率，搜索引擎在索引页面或处理搜索请求时会自动忽略某些字或词，这些字或词即被称为停用词。

停用词一定程度上相当于过滤词（Filter Words），区别是过滤词的范围更大一些，包含情色、政治等敏感信息的关键词都会被视做过滤词加以处理，停用词本身则没有这个限制。通常意义上，停用词大致可分为如下两类。

一类是使用十分广泛，甚至是过于频繁的一些单词。比如英文的“i”、“is”、“what”，中文的“我”、“就”等，这些词几乎在每个文档上均会出现，查询这样的词无法保证搜索引擎能够给出真正相关的搜索结果，因此无法缩小搜索范围来提高搜索结果的准确性，同时还会降低搜索的效率。因此，在搜索的时候，Google 和百度等搜索引擎会忽略掉特定的常用词，如果使用了太多的停用词，有可能无法得到精确的结果，甚至可能得到大量毫不相关的搜索结果。

另一类是文本中出现频率很高，但实际意义又不大的词。这一类词主要包括了语气助词、副词、介词、连词等，通常自身并无明确意义，只有将其放入一个完整的句子中才有一定作用的词语。常见的有“的”、“在”、“和”、“接着”等，例如“泰迪教育研究院是最好的大数据知识传播机构之一”这句话中的“是”、“的”就是两个停用词。

经过分词后，评论由一个字符串的形式变为多个由文字或词语组成的字符串的形式，可判断评论中词语是否为停用词。根据上述停用词的定义整理出停用词库，并根据停用词库去除评论中的停用词。利用代码实现如下：

```
import pandas as pd
import jieba
# 导入数据
z = pd.read_excel(r"C:/Users/Administrator/Desktop/2.xlsx")
print(原始数据的形状为: 'z.shape')
# 导入停用词，创建写入方法

output1 = open("C:/Users/Administrator/Desktop/分词后.txt", "w", encoding="utf-8")
stopwords_path = r"C:/Users/Administrator/Desktop/stopword.txt"
stop_list = []
with open(stopwords_path, "r") as f:
    for line in f.readlines():
        stop_list.append(line.replace("\n", ""))

z1 = z[z["一级标签"] == "城乡建设"] | z[z["一级标签"] == "环境保护"] | z[z["一级标签"] == "交通运输"] | z[z["一级标签"] == "教育文体"] | z[z["一级标签"] == "劳动和社会保障"] | z[z["一级标签"] == "商贸旅游"] | z[z["一级标签"] == "卫生计生"]
for row1 in z1.iterrows():
    print(row1)

for index, row1 in z1.iterrows():
    words = jieba.cut(row1["留言主题"])
    for word in words:
        if word not in stop_list:
            if word.strip() != "":
                output1.write(word + "\n")
```

图 1.2-3 代码

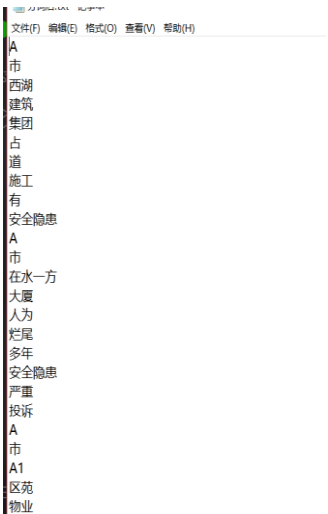


图 1.2-4 结果



### 1.3 特征矩阵构造

特征矩阵构造，就是提取每一位群众问政留言特征词构造成词向量的过程。首先利用模型训练得到每一个词的词向量表示，然后再通过词向量合成得到每条留言主题的词向量表示。其中，语料库的选取、训练参数设置、生成模型、模型测试构成了模型训练。由群众问政留言特征分析可得，维基百科等不适合表述较为正式的语料库作为模型训练语料库，因此采用上述去空去重、中文分词后的 9211 条留言文本作为训练语料库。采用 Skip-Gram 模型，将分词处理后的词采用 Word2Vec 训练成词向量。里面通过 Python 调用 gensim 库中 gensim.models.word2vec 包进行模型训练，模型训练部分参数设置如下：词向量维度 size=100，词向量上下文最大的距离 window=5，需要计算词向量的最小词频 min\_count=5，训练完毕后，每个词都训练成一个 100 维的向量，保存模型并测试。

模型训练完毕，进行词向量合成得到特征矩阵。合成的模式有直接进行词向量累加，去重处理之后进行词向量累加，以及对词向量进行累加后平均等方式。考虑到所处理的留言标题长短不一，为避免长度差距带来的对词向量合成的影响，采取累加后平均方式进行合成，即将每一条留言主题包含的每个词的词向量相加除以词的数量。处理过程中，由于设置词向量的最小词频为 5，其中 4 条留言所含词汇均不在训练得到的词汇表中，对其进行删除处理。处理完毕之后剩余 8905 条数据，每一条留言主题就表示成一个 100 维的向量形式。

### 1.4 留言主题分类

留言主题的自动分类属于典型的多分类问题，经常用经典的 SVM 算法是解决二分类问题。常见的解决思路是将多分类问题分解为多个两类问题求解，再通过决策函数确定分类结果。我们采用 python 并调用 sklearn 库中的 svm 包实现多类分类算法，其过程如下：

1) 对数据留言特征和标签进行预处理。首先对留言数据特征进行处理，即对上述得到的 100 维词向量进行归一化处理，通过进行线性变换对原始数据进行

0-1 标准化处理，让所有结果都在 $[0, 1]$ 范围内，然后得到归一化处理的 100 维词向量  $(x_1, x_2, x_3, \dots, x_{100})$ 。其次，对一级标签进行处理，利用 SVM 作为有监督的分类预测模型，并且将对应的标签  $y_i$  匹配到留言特征词。将城乡建设、环境保护、交通运输、教育文体等类别分别编号为 1、2、3、4 等，与归一化处理后的 100 维词向量形成每一条留言主题的词向量表示  $(x_1, x_2, x_3, x_4, \dots, x_{100}, y_i)$ ，其中  $y_i \in \{1, 2, 3, 4, \dots, 15\}$ 。

2) 分类模型的训练。根据 6:4 的比例将数据划分成测试集和训练集，采用 svm.SVC 训练模型，在训练过程寻找支持向量，就是要确定最大超平面的过程。具体寻找过程如下，样本集为  $(x_i, y_i), x_i = (x_1, x_2, \dots, x_{100}), y_i \in \{1, 2, 3, \dots, 17\}$ ，利用  $g(x) = w^T x + b$  线性判别函数， $w^T x + b = 0$  表示分类的超平面方程，其中 margin (b, w) 表示超平面  $w^T x + b$  离样本点的最小距离，我们再通过引入多项式核，高斯核等函数将样本 X 映射到一个高维特征空间，再将非线性问题转换成另一个空间的线性问题来寻求最优分类面。

3) 模型训练完毕后，利用优化参数设置进行分类效果评估。

## 1.5 留言特征词与类别匹配

实现自动分类后，每条留言特征对应某一标签，我们通过集合和映射表示来实现留言特征到标签的匹配。首先，将类别，留言主题特征词，留言详情特征词分别用集合 T、D、R 表示，集合的表示如公式所示：

$$T = \{T_1, T_2, \dots, T_i, \dots, T_M\}, i \in [1, M] \quad (1)$$

$$D = \{D_1, D_2, \dots, D_j, \dots, D_N\}, j \in [1, N] \quad (2)$$

$$R = \{R_{D_1}, R_{D_2}, \dots, R_{D_j}, \dots, R_{D_N}\} \quad (3)$$

其中， $T_i$  表示一级标签类别， $D_j$  表示留言主题的特征词， $R_{D_j}$  表示留言主题特征词的集合；M 为标签类别数，N 为留言特征词数。结合映射规则的需要，设  $n_j$  为某一类别特征词数，某一特征词采用留言编号加留言用户  $R_{jn_j}$  表示，如公式 (4) 所示。将公式 (4) 代入公式 (3)，公式 (3) 变成公式 (5)。

$$R_{D_j} = \{R_{j1}, R_{j2}, \dots, R_{jn_j}\} \quad (4)$$

$$R = \{R_{11}, R_{12}, \dots, R_{1n_1}, R_{21}, R_{22}, \dots, R_{2n_2}, \dots, R_{N1}, \dots, R_{Nn_N}\} \quad (5)$$

其次，我们将标签到留言特征词的匹配关系用映射的形式来表示，从标签集到留言集的映射用  $f: T \rightarrow R$  表示，某一标签  $T_i$  必定属于某一留言  $R_{jn_k}$ ，映射规则  $f$  是标签  $T_i$  包含于  $R_{jn_k}$ ，如公式（6）所示。其实现方式有两种，一是采用正则表达式进行标签与留言特征词的匹配，然后利用匹配结果进行逻辑判断。二是进行特征词相似度计算，选取相似度最高的特征词对应。留言主题特征词集到留言详情特征集的映射用  $g: R \rightarrow D$  表示。某一留言详情特征词必定属于某一留言主题，映射规则  $g$  为留言  $R_{jn_j}$  下标中第一个下标  $j$  与留言详情相等则对应。由公式（6）和（7）可以确定从标签到留言的对应关系  $g \cdot f$ ，从而留言到标签的匹配

$$f: T_i \subseteq R_{jn_k}, n_k \in [1, n_j] \quad (6)$$

$$g: R_{jn_k} \xrightarrow{\text{下标相等为 } j} D_j \quad (7)$$

下面表 1.5-1 简单描述所匹配的结果：

表 1.5-1

标签	类别	留言特征词 1	留言特征词 2
$T_1$	环境保护	污染	生产
$T_2$	交通运输	出租车	快递
$T_3$	教育文体	教师	问题
$T_4$	劳动和社会保障	单位	公司
$T_5$	城乡建设	小区	房产证
$T_6$	商贸旅游	电梯	垄断
$T_7$	卫生计生	医院	咨询

为了对上述分类设计进行验证，采用局部与整体评价结合的方式，利用查准率、查全率和 F1 值对实验结果进行测评（见表 1.5-2）。对于局部评价指标，对应每一类别的查准率、查全率和 F1 值。整体的分类性能采用宏平均方式来计算总体的查准率 Macro-P、查全率 Macro-R 和 Macro-F1 值。模型的自动分类结果为预测值，附件 2 中人工标注好的一级标签为真实值，通过算法得到的混淆矩阵进行计算，混淆矩阵示例见表 4，矩阵的每一列表达了分类器对于样本的类别预测，每一行则表达了版本所属的真实类别。

表 1.5-2 局部与整体分类指标公式

指标	局部	整体
查准率	$P = \frac{TP}{TP + FP}$	$Macro-P = \frac{\sum_{i=1}^N p_i}{N}$
查全率	$R = \frac{TP}{TP + FN}$	$Macro-R = \frac{\sum_{i=1}^N R_i}{N}$
F1 值	$F_1 = \frac{2PR}{P + R}$	$Macro-F_1 = \frac{2Macro-P * Macro-R}{Macro-P + Macro-R}$

其中，对于某一类别，TP 表示这一类的预测值与真实值一致的个数，FP 表示预测值误判为这一类的个数，FN 表示属于这一类被预测为其他类的个数，N 表示类别总数。

表 1.5-3 混淆矩阵示例（kernel='rbf'，gamma=0.4 时）

混淆矩阵		预测值		
		1	2	3
真实值	1	1965	77	14
	2	289	1926	20
	3	38	309	1544

选取离散取值方式对参数进行优化调整，一般以 F1 最高时对应的参数取值为最优，但会结合准确率、召回率和 F1 值综合考虑。具体调整过程说明如下：①参数训练集比例 C 的优化。参数训练集比例 C 一般取值为 0.6~0.9，在保证其他参数不变的情况下，分别选取 C=0.6, 0.7, 0.8, 0.9 看分类效果的变化，C=0.7 时查准率、查全率和 F1 值均优于其他取值，据此确定最优训练集比例为 0.7。②核函数和 gamma 值优化。通过固定核函数，调整 gamma 的大小来确定 gamma 值的设置。例如当核函数为 rbf 时，gamma=20，训练集查准率 99.73%，测试集的查准率只有 57.71%，出现了过拟合的情况。通过比较分类

效果，选定核函数为 rbf（即高斯核），当 gamma=0.4 时结果最优；选定核函数为 poly（即多项式核），当 gamma=0.1 时结果最优，不同核函数下总体分类宏平均指标统计如表 1.5-4 所示。由表 1.5-4 可知，核函数为 rbf( gamma=0.4) 时分类效果最优，此时，每一类的分类指标统计见表 1.5-5。

表 1.5-4 不同核函数下总体分类宏平均指标统计

参数	查准率 (%)	查全率 (%)	F1 值 (%)
Linear	83.56	83.26	83.18
rbf(gamma=0.4)	86.24	85.49	85.75
poly(gamma=0.1)	84.46	84.66	84.53

表 1.5-5 三个类目分类指标统计

类目	查准率 (%)	查全率 (%)	F1 值 (%)	预测数量	实际数量
教育文	92.78	94.61	93.70	2056	1965
劳动和社会保障	83.24	79.53	81.38	2421	1926
城乡建设	82.11	81.92	82.51	1891	1544

从表 1.5-4 可以看出，各指标均达到 83%以上，整体分类效果较为理想。从表 1.5-5 可以看出，三个类别的分类指标存在较大差异，其中城乡建设类别的各项指标均超过 92%，而劳动和社会保障类和教育文体类的各项指标均在 80%左右。经分析存在差异的原因可能有：①样本数据存在一定程度的不均衡，教育文体类的数量最多，使得超平面偏向数据少的一侧，导致其他类别的准确率会偏低。②通过人工查阅分词数据，发现教育文体类留言中包含“教师”“学校”“补课”的概率极高，分类指征明确。

## 二、第二题的分析与处理

### 2.1 对附件 3 的分词处理

利用 python 里一个分词开源库 jieba 分词包对附件 3 的群众留言主题进行分词处理，其原理如上述，例如：“A3 区一米阳光婚纱摄影是否合法纳税了”，经过分词处理后变成“A3/区/一米阳光/婚纱/艺术摄影/是否/合法/纳税/了”，其实现代码如图 2.1-1，结果见附件。

```

1 import pandas as pd
2 import jieba
3 # 导入数据
4 z = pd.read_excel(r"C:/Users/Administrator/Desktop/附件3.xlsx")
5 print('原始数据的形状为:', z.shape)
6 # 导入停用词, 创建写入方法
7
8 output1 = open("C:/Users/Administrator/Desktop/分词后.txt", "w", encoding="utf-8")
9 stopwords_path = r"C:/Users/Administrator/Desktop/stopword.txt"
10 stop_list = []
11 with open(stopwords_path, "r") as f:
12     for line in f.readlines():
13         stop_list.append(line.replace("\n", ""))
14
15 z1 = z[(z['反对数']==0) | (z['反对数']==1) | (z['反对数']==2) | (z['反对数']==3) | (z['反对数']==4) | (z['反对数']==5)]
16 for row1 in z1.iterrows():
17     print(row1)
18
19 for index, row1 in z1.iterrows():
20     words = jieba.cut(row1['留言主题'])
21     for word in words:
22         if word not in stop_list:
23             if word.strip()!="":
24                 output1.write(word + '\n')
25

```

图 2.1-1 代码

## 2.2 对附件 3 词频统计与绘制词云图

词频统计，就是对某一或某些给定的词语在某文件中出现的次数进行统计。用以评估一个词对于一个文件或者一个语料库中的一个领域文件集的重复程度。词频统计为学术研究提供了新的方法和视野。词的大小代表该词在文本中出现的频率，词越大，表示出现的频率越高。

在一份给定的文件里，词频（term frequency, TF）指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被正规化，以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否。）对于在某一特定文件里的词语  $t_i$  来说，它的重要性可表示为：

以上式子中  $n_{i,j}$  是该词在文件  $d_j$  中的出现次数，而分母则是在文件  $d_j$  中所有字词的出现次数之和。

进行数据预处理后，可绘制词云查看分词效果，词云会将文本中出现频率较高的“关键词”予以视觉上的突出。首先需要对词语进行词频统计，将词频按照降序排序，使用 wordcloud 模块的 WordCloud 绘制词云如下图所示：从结果可以知道，“扰民”、“噪音”等留言问题出现的频率最高。



图 2.2-1 词云图

### 2.3 热度评价指标

直接利用 Excel 表对留言主题进行特定地点或者特定人群组成留言索引，再次利用高频词和留言索引构建特征矩阵，进行分析与处理，然后利用聚类的机器学习算法把相似留言主题合并成一类。再次采用通过对关键特征词进行聚类分析来检测留言热度，并定义合理的热度评价。

热度指标的构建原则通常根据要求和用户留言的不同分为三个层面：指标选取层面一般采取客观性原则、系统性原则和敏感性原则。客观性原则是指热度指标体系的选择必须从客观实际出发，全面准确地反映留言的热度情况，克服因留言用户而异的主观因素的影响。系统性原则是指指标体系的设计应从系统整体出发，能够包络形成留言热度的各个因子，各指标间既相互独立又相互联系，共同构成一个有机整体。计算与操作层面一般采用数据的可得性和可操作原则，是指在设计指标体系时用较少指标反映较多的实质性内容，而且指标便于收集和量化。

针对留言热度这一特殊研究对象，在构建留言热度评价指标体系的过程中，除遵循以上基本性原则外，新增了趋势性原则和导向型原则。留言的热度是一个时刻变化的指标，趋势性原则就是体现留言的变化趋势；导向型原则是指该套指标体系的构建不仅要对留言进行检测，更是要为判断留言热度提供方向指导。

### 2.4 聚类的机器学习算法

聚类是一种无监督的分类的方法，我们可以对变量聚类或者样本聚类，从而达到将相似度大的变量或者样本分到一类，组内区分相似度较小，组间区分相似度大的最终目的，聚类的方法，也会根据聚类的目的分为若干种，比如层次聚类是一种基于变量的聚类，K-means 是一种基于样本的聚类。整个类聚算法的过程如下：

- (1) 计算整个的留言主题词与词之间的相似度
- (2) 初始化，词表中的每个词各代表一类，共 N 类（N 为词表中词的数量）
- (3) 找出具有最大相似度的两个词类，将这两个词类合并成一个新的词类
- (4) 计算合并词类与其它词类的相似度
- (5) 检查是否达到结束条件，也就是词类之间最大相似度小于某个预先决定的门槛值，或是词类的数目达到了要求。

我们了解到相似度定量标准有三种形式：相似度、左相似度和右相似度。根据这三种定量标准可以得到三种不同的聚类结果，不同的聚类结果可以用于不同的基于类的语言模型。再找出类别计数最多的留言进行清洗，如附件热点问题留言明细表所示，其中出现 56 个关于搅拌站扰民问题，25 个油烟直排扰民问题，10 个施工噪音问题，9 个安全隐患问题，7 个强制学生实习问题。最后可以筛选出热度前五的问题。结果如表 2.4-1：

表 2.4-1 热点问题表

热度排名	热度	时间范围	地点/人群	问题描述
1	56/4326	2019/07/至 2020/01/26	A 市万家丽南路丽发新城	A 市万家丽南路丽发新城居民区附近搅拌站扰民
2	25/4326	2019/03/至 2019/11/22	A5 区劳动东路 魅力之城	魅力之城小区临街门面油烟直排扰民
3	10/4326	2019/01/至 2019/11/22	A3 区中国国际社区	A3 区中海国际社区空地夜间施工噪音太大了
4	9/4326	2019/01/至 2019/08/28	A3 区青山镇青青家园	A3 青山镇青青家园管理不规范。扰民，存在安全隐患
5	7/4326	2017/06/至 2019/11/22	A 市经济学院	A 市经济学院强制学生实习



### 三、第三题的分析

对附件 4 的留言详情、答复意见分别进行数据预处理，在问题一的基础上进行类别划分和关键词提取，得到相应的文本数据，而相关性分析模型的输入数据为数据型数据。首先要将文本数据量化为数值型数据，然后利用相关性分析模型，对留言主题与答复意见的相互关系进行分析。本文采用自然语言处理领域中的词向量作为表达词语的数值型特征。对留言主题与答复意见进行分词后，在已有的大规模语料上对神经网络模型进行训练，得到词语对应的词向量，采用 Word2Vec 工具实现。

留言主题与答复意见相关性分析指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关程度，是机器学习中常用的技术。将留言主题数据记为  $x, x = (x_1, \dots, x_i, \dots, x_n)$  答复意见数据记为  $y, y = (y_1, \dots, y_i, \dots, y_n)$

其中  $n$  为留言总数量， $1 \leq i \leq n, x_i \in R, y_i \in R$ 。根据基础统计学理论，设  $x$  的数学期望为  $u^x$ ， $y$  的数学期末为  $u^y$ ，则  $x$  和  $y$  之间的相关系数如公式（1）所示：

$$Corr(x, y) = \frac{\sum_{i=1}^n (x_i - \mu^x)(y_i - \mu^y)}{\sqrt{\sum_{i=1}^n (x_i - \mu^x)^2} \sqrt{\sum_{i=1}^n (y_i - \mu^y)^2}} \quad (1)$$

该相关系数也称为皮尔逊相关系数，公式（1）中的  $\sum_{i=1}^n (x_i - u^x)(y_i - u^y)$  可理解为  $(x_i - u^x)$  和  $(y_i - u^y)$  的内积。 $\sqrt{\sum_{i=1}^n (x_i - u^x)^2}$  为  $(x_i - u^x)$  的 2 范数， $\sqrt{\sum_{i=1}^n (y_i - u^y)^2}$  为  $(y_i - u^y)$  的 2 范数，由公式（1）可推导出公式（2）：

$$Corr(x, y) = \frac{\sum_{i=1}^n (x_i - \mu^x)(y_i - \mu^y)}{\sqrt{\sum_{i=1}^n (x_i - \mu^x)^2} \sqrt{\sum_{i=1}^n (y_i - \mu^y)^2}} = \frac{\langle x - \mu^x, y - \mu^y \rangle}{\|x - \mu^x\|_2 \|y - \mu^y\|_2} \quad (2)$$

从公式(2)可以看出, 相关系数表达了两个变量之间的相关程度, 该值在0~1之间。当相关系数为1时, 二者具有最强的相关关系, 意味着留言答复回答的是相应的留言主题的内容。

经过分析关于相关部门对留言的答复意见中的完整性实行实体完整性约束, 实体完整性是指一个关系中所有主属性(即主码的属性)不能取空值。所谓“空值”就是“不知道”或“无意义”的值。如主属性取空值, 就说明存在某个不可标识的实体, 这与现实世界的应用环境相矛盾, 因此这个实体一定不是一个完整的实体。

留言的答复意见的完整性由三部分组成, 开头是某某市民或网友你好! 你反映的什么问题我们已获悉; 接下来说明处理过程与结果; 最后, 表明态度, 感谢网民或市民对我们工作的支持。这三个中只有最后表明态度才不是主属性, 因此缺主属性就表示答复意见不完整, 答复意见的质量较低。

## 四、参考文献

- [1] 汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究, 2014, 31(5): 1281-1286.
- [2] 刘晓亮, 丁世飞, 朱红, 等. SVM 用于文本分类的适用性[J]. 计算机工程与科学, 2010, 32(6): 106-108.
- [3] 杨槟泽, 李长军. 市长公开电话文本自动分类技术比较研究[J]. 中国海洋大学学报: 自然科学 版, 2017, 47(S1): 173-177.
- [4] 马宝君, 张楠, 谭棋天. 基于政民互动大数据的公共服务效能影响因素分析[J]. 中国行政管理, 2018(10): 109-115.
- [5] 郭德清, 廖祥文. 基于箱线图的微博客热点话题发现 [J]. 山西大学学报, 2014, 37(1): 19-25.

[6]梁昌明，李冬强. 基于新浪热门平台的微博热度评价指标体系实证研究  
(1. 山东师范大学历史与社会发展学院，济南 250014; 2. 北京科技大学图书馆，北京 100083)