基于群众问政记录的数据挖掘分析

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。本文对“智慧政务”评论数据通过数据清洗、数据集成和融合、数据变换、数据规约等方法进行了预处理。利用 word2vec 和权重 TF-IDF 算法和借助卷积神经网络 CNN 使非结构化数据到结构化数据转化。利用半监督 RAE 深度学习模型对这些评论进行情感分析。主要进行的数据挖掘分析工作：一方面是根据“智慧政务”中文本挖掘的分析结果，对文本留言进行标签分类并找到相关热点问题；另一方面是根据不同评论数据情感分析结果，进行不同部门的工作比较，进而可以提升政府的管理水平和施政效率。

关键词：层次分析、文本分类、情感分析、词向量转化、清洗、权重 TF-IDF 算法、文本矩阵转化、卷积神经网络 CNN

Data mining analysis based on the records of mass inquiries and political affairs

Abstract：In recent years, with WeChat, micro-blogging, mayor's mailbox, sunshine hotline and other network political platform gradually become the government to understand public opinion, gather people's wisdom, rally the people's popularity of the important channels, all kinds of social and public opinion-related text data volume continues to climb, to the past mainly rely on manual to divide messages and hot-spot finishing of the relevant departments of the work has brought great challenges. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government system based on natural language processing technology has become a new trend of innovation and development of social governance, which has a great role in promoting the level of government management and the efficiency of governance. In this paper, the "smart government" comment data is pre-processed by data cleaning, data integration and fusion, data transformation, data specification and so on. Unstructured data is transformed into structured data using word2vec and weight TF-IDF algorithms and CNN with convolutional neural networks. These reviews are analyzed emotionally using the semi-supervised RAE deep learning model. The main data mining analysis work: on the one hand, according to the "smart government" Chinese the analysis of this mining analysis, to classify text messages and find related hot issues, on the other hand, according to the emotional analysis results of different comment data, the work of different departments to compare, and then can improve the level of government management and governance efficiency.

Key words: Hierarchical analysis, text classification, emotional analysis, word vector conversion, cleaning, Weight TF-IDF algorithm, Text matrix transformation,
Convolutional neural network CNN

目录

一、研究目标

本次建模目标是针对处理网络问政平台的群众留言数据，采用基于半监督 RAE 深度学习模型的数据挖掘方法，达到以下三个目标：

1) 对附件 1、附件 2 的数据利用多分类模型、评价模型按照一定的划分体系对留言分类，建立关于留言内容的一级标签分类模型，建立好模型后，通过 F1 值调整模型并选出最优的模型比较进行分析，根据分析结果得到各级对应的留言数据，从而可分析不同级的留言数据，提高和优化相关部门的工作。
2) 通过对附件 3 进行文本挖掘（针对相关问题的某一时间、特定地点、发生的问题的三要素进行数据清洗、数据集成和融合、数据变换、数据规约）进行归类，定义合理的热度评价指标并将问题归并后进行相似度计算得出对同一事件的留言数据，通过量化评价指标把相关问题归并后，分析得出相应热点问题的留言热度排名。
3) 针对附件 4 相关部分对留言的答复意见，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评价方案。

二、分析方法和过程

2.1. 总体流程

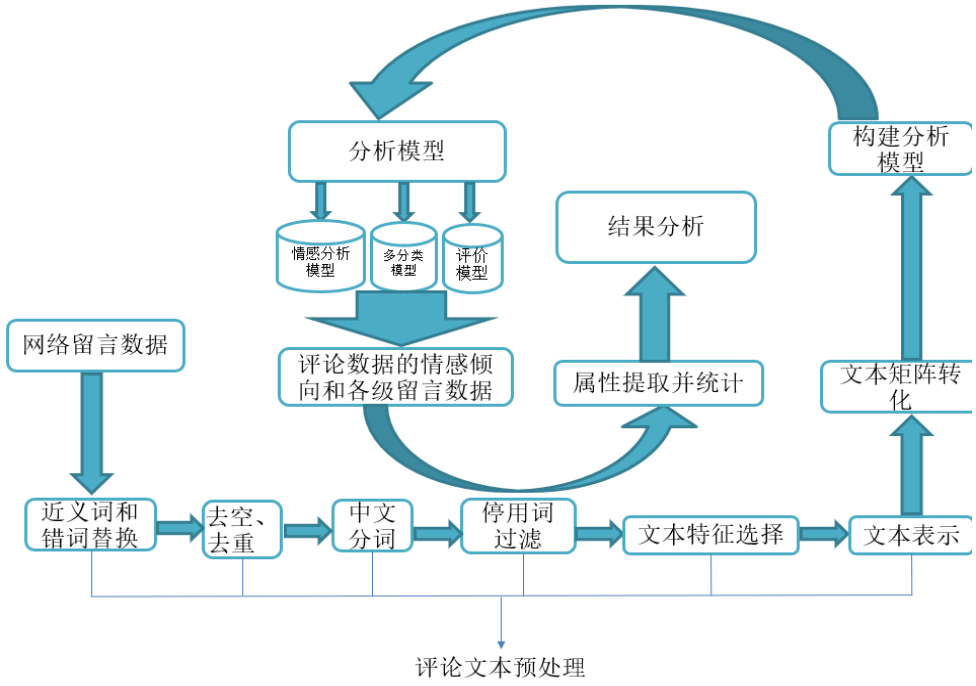

图 1 总体流程图

本用例主要包括以下几个步骤：

步骤一：下载网络留言数据，网络留言数据的获取是本次数据挖掘分析的第一步。本文中利用问题所给出获取数据的网址对附件 1、附件 2、附件 3、附件 4 留言文本进行下载保存，得到数据建立相应模型的实验数据。

步骤二：模型准备：对原始数据进行预处理。第一步要对留言的近义词和错词替换；第二步要“去空、去重”；第三步对评论数据进行中文分词，将一句评论分成多个词语进一步分析；第四步进行停用词过滤，去除掉留言中与情感判定不相关的词。第五步：文本特征选择，用数学方法（文档频率、信息增益、卡方校验和互信息等方法）选取最具分类留言信息的特征。第六步：文本表示（doc2vec、word2vec）对找到所有的特征词拉成一列做成列表。

步骤三：文本矩阵转化，使用基于半监督 RAE 深度学习模型进行情感分析，需要将文本词语全部转换为词向量，本论文中构建了一个词表和词向量表，词表中为全部文本词语和词语的编号，词向量表中为全部词语的词向量。

步骤四：情感分析和数据分析，构建基于半监督 RAE 的深度学习模型，利用选出的各级的留言数据集训练情感分析模型和多分类模型、评价模型，并进行测试，得到符合要求的模型。利用构建的模型分析得出评论数据的情感倾向和各级留言数据。

步骤五：属性提取并统计，分别将各级留言数据和热点问题留言数据从实验数据集中筛选出来，统计各级别的相关留言数据和热点问题的留言数据的数量，并进行热点问题留言的排序。

步骤五：结果分析，根据分析结果提取各级对应的留言数据和相应热点问题的留言热度排名。从而可分析不同级的留言数据，提高和优化相关部门的工作。并针对留言的答复意见，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评价方案。

2.2 具体步骤

随着微信、微博、市长信箱、阳光热线等网络问政平台的发展，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。

针对于问题一：群众留言分类

通过对留言内容数据文本进行分类和建立一级标签模型，对留言记录进行分类。然后建立评价模型，使用 F-Score 对分类方法进行评价。详细步骤如下：

步骤一：构建政务词向量库

1) 读取附件数据，获取网络问政平台的群众留言的文本数据。
2) 借助标准的语料库，基于词语的统计，采用 python 里的 jieba 工具进行政务文本的分词。对于政务文本进行文本过滤，过滤一些无用的文本数据，去除分词中存在的与文本分类无关词汇（如介词和谓词）。
3) 接着我们对文本过滤后的词汇，进行 one-hot 编码，将词汇数值化表示。根据编码后的文本词汇，构建政务文本词库的训练样本。
4) 输入神经网络模型，获取神经网络的层数、激活函数、分类方法、调优方法。根据政务文本词库的训练样本，训练神经网络模型，获取训练后的神经网络模型。

5） 根据构建的神经网络模型，获取政务文本分类领域的词向量库。

步骤二：建立文本分类模型

根据附件 2 的留言内容、留言主题和一级标签，结合附件 1，找到附件 1 相应二级标签和三次标签处理留言的问题。

步骤三：使用 F-Score 对分类方法进行评价

将得到处理后的分词利用向量转化最后拟合附件 1 的标签文本相似度得到 F-Score 中的查准率 P 和查全率 R。

通常使用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

图 2 F-Score 对分类方法评价公式

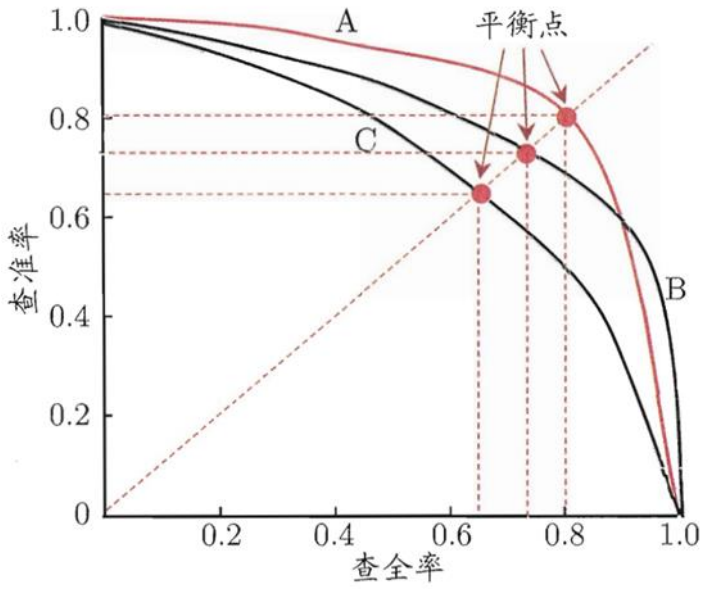

图 3 查准率和查全率关系图

针对问题二：热点问题挖掘

步骤一：下载网络留言数据

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。通过数据挖掘等技术手段实现对相关问题留言的智能分析，可以获得得到各级对应的留言数据，进而可以提升政府的管理水平和施政效率，获取网络留言数据中的有价值的信息，做出一套评价方案等。

1） 读取附件数据

```
读取附件数据
import panda as pd
Data=pd.read_csv(messagethree.csv)
Data.shape(查看数据结构)
```

图 4 查看数据结构

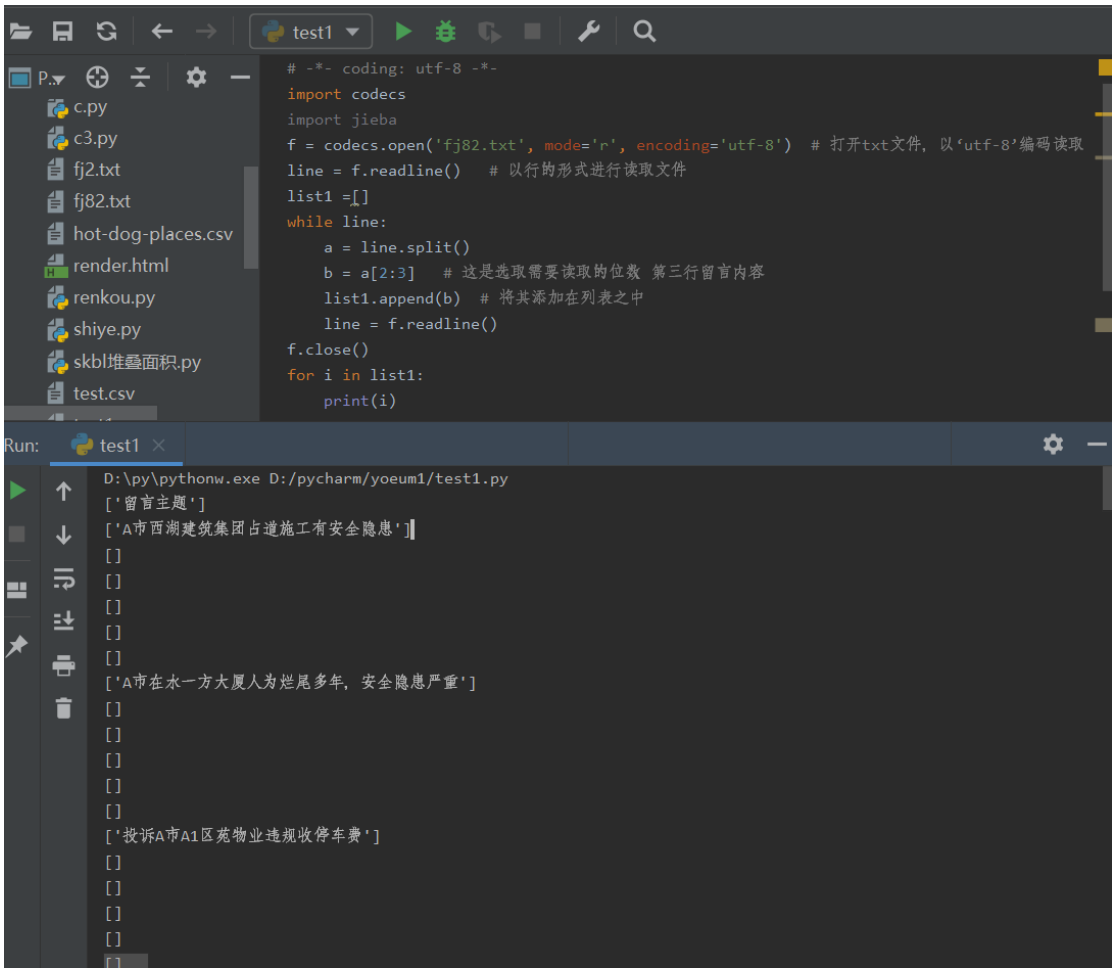

图 5 读取数据

从非结构化数据到结构化数据转化

将每个词表示为长长的向量，词袋：所有词的不重复构成

选取尽可能多的已标记的政务分类文本，进行 Word2vec 的词向量库建设

2） 借助权重 TF-IDF 算法，计算网络留言每个词在不同类别所占的比例，量化分类的权重，生成改进型的文本向量。增加词频信息（文本特征）并统计词频信息数量，词频越高，比较热点。具体介绍如下：

TF-IDF 权重策略指权重策略文档中的高度频词应具有表征此文档较高的权重，除非该词也是高文档频率词；

TF:term frequency 即关键词频，是指一篇文章中关键词出现在所有文档的频率

TF = N/M(N 出现的文频率/总的文本数 M)

IDF:Inverse document frequency 指逆向文本频率，是用于衡量关键词权重的指数，公式

$$\mathrm{IDF} = \log\left(\frac{D}{D_W}\right) \qquad \text{——式 1}$$

D_W 分母:表示在文本中出现的次数，越小则对应文本词的维系越强，D 总的文本数

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF} \times \mathrm{IDL} \qquad \text{——式 2}$$

该值越大，越说明对应词汇在文本出现的分类特性就越强，越大越有表征性。

3） 借助卷积神经网络 CNN 根据改进型的词向量，生成政务文本分类模型。输入神经网络模型，获取神经网络的层数、激活函数、分类方法、调优方法。

```
# cnn
conv_0=Conv2D(num_filters,kernel_size=(filter_sizes[0],embedding_dim),padding='valid',kernel_initializer='normal',activation='relu')(reshape)
conv_1=Conv2D(num_filters,kernel_size=(filter_sizes[1],embedding_dim),padding='valid',kernel_initializer='normal',activation='relu')(reshape)
conv_2=Conv2D(num_filters,kernel_size=(filter_sizes[2],embedding_dim),padding='valid',kernel_initializer='normal',activation='relu')(reshape)

maxpool_0=MaxPool2D(pool_size=(sequence_length-filter_sizes[0]+1,1),strides=(1,1),padding='valid')(conv_0)
maxpool_1=MaxPool2D(pool_size=(sequence_length-filter_sizes[1]+1,1),strides=(1,1),padding='valid')(conv_1)
maxpool_2=MaxPool2D(pool_size=(sequence_length-filter_sizes[2]+1,1),strides=(1,1),padding='valid')(conv_2)


concatenated_tensor = Concatenate(axis=1)([maxpool_0, maxpool_1, maxpool_2])
flatten = Flatten()(concatenated_tensor)
dropout = Dropout(drop)(flatten)
output = Dense(units=2, activation='softmax')(dropout)
model=Model(inputs=inputs,outputs=output)
```

图 6 cnn 建模

步骤二：模型准备：对原始数据进行预处理。

对于网络数据而言，其一个最鲜明的特点便是数据的随意性大，参差不齐、概念不清、数量级不同等诸多差异给后续的数据分析和挖掘带来诸多麻烦甚至错误。如果要对网络评论数据进行情感分析，就必须先将文本数据进行预处理，转化为

结构化的数据。故对原始数据进行按如下步骤处理以提高程序效率及模型的准确性。

1） 近义词和错词替换：在检查数据过程中避免部分词语出现拼写错误的情况而漏了数据，把网络留言错词语做纠正处理。

2） 去空、去重

对于存储的附件 3 的网络留言文本。每行代表了相关留言的但是难免出现两个完全一样的文本和一些空行。所以本文首先进行了“去重”、“去空”的预处理工作。在导入评论文本时，同时进行了是否为空的判断，若为空，则进行数据删除。若非空则将留言文本导进列表再进行去除重复处理，对于完全相同的记录，可以把一条作为正确的，删除其它重复的记录；对于同一对象实体的不同表现形式而形成重复记录的情况，可以选择特征属性进行判定，如果某些特征属性相同，就可以认为是重复记录，也可以先修正表示形式不同的字段的内容，归结为完全重复的记录，再进行处理，对于这种情况，两条记录是否重复需要通过特征字段值的匹配来决定。

```
StreamReader sr = new StreamReader("C:/Users/YoEum/Desktop/泰迪杯/data/fj2.txt",Encoding.UTF8);
String line;
while((line=sr.ReadLine())!=null)
{
   if(line.ToString()!="")//去掉空文本
   {
      CommentsList.Add(line.ToString());
   }
}//去空程序段
CommentsList2.Add(CommentsList[0]);
for(int i=1;i<CommentsList.COunt;i++)
{
   IsRepeated=false;
   for (int j=0;j<i;j++)
   {
      if(CommentsList[i].Equals(CommentsList[j]))
      {
         IsReapeated=true;
         break;
      }
   }
   if(!IsRepeated)
   {
      CommengtsList2.Add(CommentsList[i]);
   }
}//去重程序段
```

图 7 “去空”“去重”程序段

3） 中文分词

中文分词是中文文本处理的一个基础步骤，也是中文人机自然语言交互的

基础模块，指的是通过某种特定的规则，将中文文本切分成一个一个单独的词。不同于英文的是，中文句子中没有词的界限，因此在进行中文自然语言处理时，通常需要先进行分词，分词效果将直接影响词性、句法树等模块的效果。根据其特点，可以把分词算法分为四大类：基于规则的分词方法、基于统计的分词方法、基于语义的分词方法、基于理解的分词方法

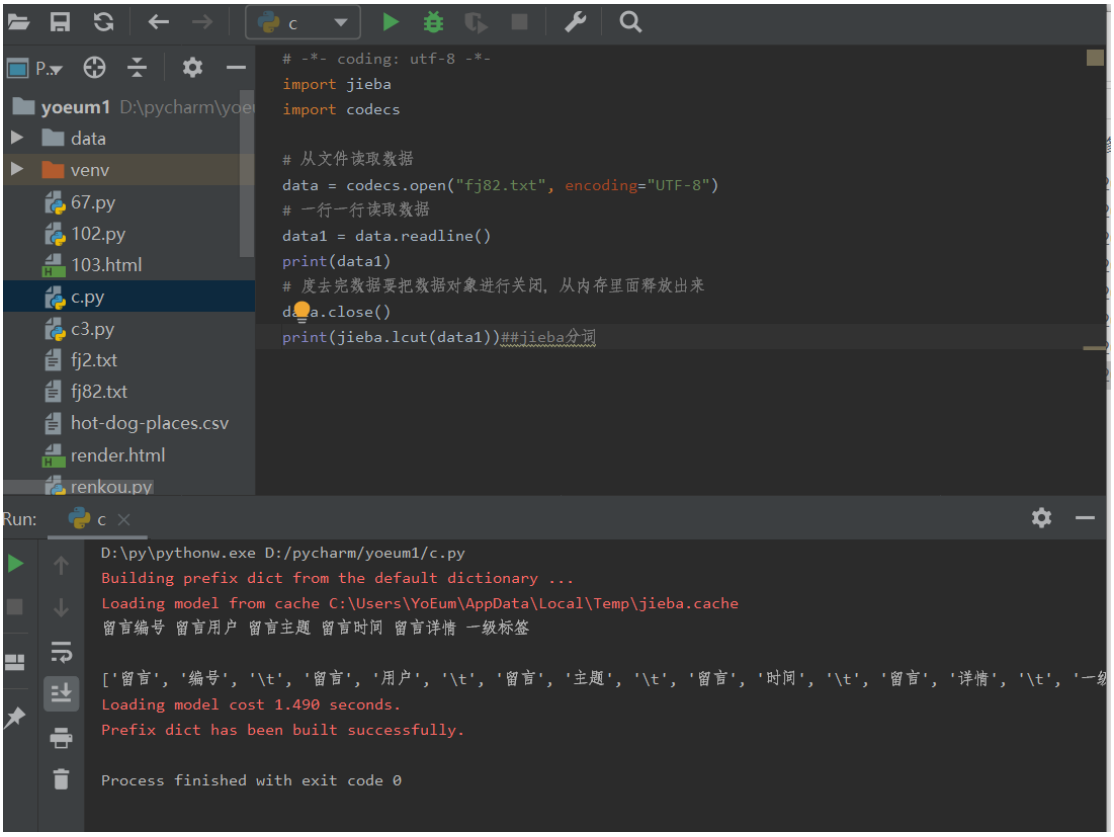

图 8 中文分词

4） 停用词过滤

留言文本在经过近义词与错词替换、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，需要将它们过滤掉，否则将会影响下文的分析的正确率。

5） 文本特征选择

特征选择(feature selection)，从原始的 d 维空间中，选择为我们提供信息最多的 k 个维(这 k 个维属于原始空间的子集)

特征提取(feature extraction)，将原始的 d 维空间映射到 k 维空间中(新的 k 维空间不输入原始空间的子集)

在文本特征属性选择阶段，一般用“词 t 与类别 c 不相关”作出假设，计

算出的卡方值越大，说明假设偏离就越大，假设越不正确。文本特征属性选择过程为：计算每个词与类别 c 的卡方值，然后排序取前 K 大的即可。接下来，有关计算卡方值的方法。

假设 n 个样本的观测值分别为 $x_1, x_2,\ldots, x_n$ ，它们的均值（期望）为 E，那么卡方值计算如下

$$x^2 = \sum_{i=1}^{n} \frac{(x_i - E)^2}{E} \qquad \text{——式 3}$$

如果计算出的 x^2 值与事先设定的阀值进行比较，如果 x^2 小于阀值，则原假设成立，否则原假设不成立。

6） 文本表示

文本表示既是通过某种形式将文本字符串表示成计算机所能处理的数值向量。在对于附件 3 网络留言进行文本表示根本原因是计算机不能直接对文本字符串进行处理，因此需要进行数值化或者向量化。本文本主要使用 One-hot Representation ,直译是独热编码，这种编码格式首先建立一个全局的完备的词典，该词典包含所有文本中的词，因此该方式表示后的形式为一个很大的 vector，vector 中只在该词出现的位置设置为 1，表示该词出现，其余全部为 0.这种形式的弊端。

步骤三：文本矩阵转化

1） 建立文本词集和文本词表：先按行读入数据集，由于一行中，两个 tab 之间的数据是无用的，因此舍弃掉前面的数据后，按照空格分隔字符串，得到一个个的词语，每分隔出一个 单词，就将其加入到文本词集 word_set 和文本词表 word_list 中（word_list 中的词语按出现顺序存储，且不重复；word_set 中的词语即按照词语出现的顺序存储，可重复，且不同行的单词之间用字符“#”隔开）

2） 建立矩阵 matrix 和 onehot（matrix[i][j]表示第 i 句留言中，在文本词表上位置为 j 的 词语出现的次数，其目的是方便后面写 TF 矩阵）：方法是通过简单的遍历，将 word_set 中的每一个词语与 word_list 中的词语作对照，找到单词在单词表中的位置 n，同时根据遍历过 的“#”字符的数目确定行数 m，最后另 onehot[m][n] 位置为 1，另 matrix[m][n]的位置的数值递增 1；

3） 建立矩阵 TF： TF 矩阵表示的是每一行文本中的每一个值标志对应的词语出现的次数 归一化，根据 matrix 矩阵很容易转化得到 TF 矩阵，只需要先将 matrix 矩阵同一行上的值全部相加，再用该词语出现的次数除以这个求和的结果即得到 TF 矩阵对应位置的值。

4） 建立矩阵 TF-IDF:TF 矩阵代表的是某个词语在一个文本上的出现频率，一般来说，出现的频率越高，该词语在该文本的重要性就越大，但是有些不怎么重要的词语（例如冠词，语气助词等）往往在多数文本中出现的频率都会很大，为了排除这一干扰，引入了 TF-IDF 矩阵

5） 将一个文本或者一句评论映射成一个词向量矩阵后，即将中文文本数据转化成计算机可以识别的信息格式，继而利用基于递归自编码的深度学习方法进行情感分析。文本矩阵转化过程通过编写程序产生随机的向量词表，每个词对应一个唯一的词标识号和词向量，词向量表生成后，通过扫描，将每句评论转化成一个词向量矩阵，将中文文本数据转化成数字数据——计算机可以识别的数据信息，进而进行文本情感分析。

图 9 词表

步骤四：情感文本分析

1） 构建词典。将评论文本归类为 4 类：通用情感词、程度副词、否定词、领域词。利用语义相似度计算方法计算词语与基准情感词集的语义相似度，以此推断该词语的情感倾向。

2） 构建倾向性计算算法。主要是利用情感词典及句式词库来分析文本语句的特殊结构和情感倾向词，采用权值算法代替传统人工判别或仅利用简单统

计的方法进行情感分类。然后给不同的情感词赋予不同的权值，然后进行加权求和。利用加权平均算法式

$$\overline{E}=\frac{\sum_{i=1}^{N_p} wp_i+\sum_{j=1}^{N_n} wp_j}{N_p+N_n} \qquad \text{——式 4}$$

计算（其中，N_p，N_n 分别代表表达正面情感和负面情感的词汇数目；wp_i，wp_j 分别代表正面情感词汇和负面情感词汇的权值），可以有效提高通用领域情感分类的效率和准确率。然后通过加权计算得到的结果确定阈值来判断文本倾向性。得到的结果评价采用正确率、召回率和 F 值来评判算法效果。

步骤五：属性提取并统计

1） 本步骤主要是结合步骤三得到词表和步骤四得到的情感分析结果，进行统计，得到网络留言级别、和热点问题、热点留言的排名和包含某属性的留言数据、消极留言评论、留言回复的百分比。

2） 利用程序遍历，统计分析得出包含某个属性相关词的留言数据中的积极评论与消极留言的数量，和各自占该留言的与该属性相关的所有评论数量的比重。并本文对评论信息分别进行词频统计，计算出重复的关键词出现次数，然后直接按照由大到小进行词频统计排序。部分统计结果如下：

word freq
A市 3857
小区 1856
问题 1728
政府 1402
严重 1327
业主 1106
什么 1051
公司 1002
影响 936
可以 903
工作 831
时间 830
情况 826
工程 779
安全 786
要求 783
经济 762
投诉 756
人员 711
建设 683
处理 648
环境 581
规划 576
管理 571
开发商 563
施工 543
施工 534
建议 534
导致 506
需要 496
扰民 492
噪音 452
周边 432
学校 431
项目 416
请求 415
怎么 411

图 10 词频

针对问题三：答复意见的评价

需要针对相关部门对留言的答复意见，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评价方案，并尝试实现。

相关性：答复意见的内容时是否与问题相关

完整性：是否满足某种规范

可解释性：答复意见中内容的相关解释（答复意见的回答、构建答复指标、怎么反映答复问题的质量、答复问题质量怎么样）

针对答复的信息意见，构建答复指标，衡量答复意见的答复质量。

具体步骤如下：

步骤一：计算留言与答复两两特征之间的相关性

1）使用 Python 的 sklearn，构建 DataFrame 格式数据，并将数据数组化。

2）对数据进行分词和停用词去除，然后使用‘’.join 连接列表

3）构建 np.vectorizer 向量化函数和调用函数进行分析和去除停用词

4）使用 from sklearn.metrics.pairwise import cosine_similarity，进行 TF-idf 词袋模型的构建，并对每一个留言内容做词统计

5）对两两样本之间做相关性矩阵，使用余弦相似度计算公式

$$\frac{\mathrm{a}.dot(b)}{sqrt(a^2+b^2)} \qquad \text{——式 1}$$

对数字映射后的特征做一个余弦相似度的匹配

```
import ...

f = codecs.open('fj82.txt', mode='r', encoding='utf-8')  # 打开txt文件，以‘utf-8’编码读取
line = f.readline()   # 以行的形式进行读取文件
list1 =[]
while line:
    a = line.split()
    b = a[2:3,5:6]   # 这是选取需要读取的位数 第三行留言内容
    list1.append(b)  # 将其添加在列表之中
    line = f.readline()
f.close()
for i in list1:
corpus = i

labels = ['留言主题', '留言主题', '一级标签', '一级标签', '留言主题', '一级标签']

# 第一步：构建DataFrame格式数据
corpus = np.array(corpus)
corpus_df = pd.DataFrame({'Document': corpus, 'categoray': labels})

# 第二步：构建函数进行分词和停用词的去除
# 载入英文的停用词表
stopwords = nltk.corpus.stopwords.words('english')
# 建立词分割模型
cut_model = nltk.WordPunctTokenizer()
```

图 11 相关性的计算

步骤二

数据库中的数据是从外界输入的，而数据的输入由于种种原因，会发生输入无效或错误信息。保证输入的数据符合规定，把附件 4 导入数据库，利用关系数据库管理系统将检查数据内容属性上的完整性。数据完整性（Data Integrity）是指数据的精确性（Accuracy）和可靠性（Reliability）。数据库采用多种方法来保证数据完整性，包括外键、约束、规则和触发器。系统很好地处理了这四者的关系，并针对不同的具体情况用不同的方法进行，相互交叉使用，相补缺点。

数据完整性由以下三个方面构成：

1） 域完整性：

是指一个列的输入有效性，是否允许为空值。强制域完整性的方法有：限制类型、格式或可能值的范围。

2） 实体完整性：

是指保证表中所有的行唯一。实体完整性要求表中的所有行都有一个唯一标识符。这个唯一标识符可能是一列，也可能是几列的组合，称为主键。也就是说，表中的主键在所有行上必须取唯一值。用 PRIMARY KEY 短语定义了关系的主键后,每当用户对基本表插入一条数据或者记录对主键进行更新操作时,关系型数据库都会对实体完整性进行检查,

3） 参照完整性：

是指保证主关键字和外部关键字之间的参照关系。它涉及两个或两个以上表数据的一致性维护。外键值将引用表中包含此外键的记录和被引用表中主键与外键相匹配的记录关联起来。在输入、更改或删除记录时，参照完整性保持表之间已定义的关系，确保键值在所有表中一致。关系模型额参照完整性在 CREATE TABLE 中用 FOREIGN KEY 外键定义,用 REFERENCES 短语指明这些外键参照哪些表的主键.

因此根据数据完整性用数据库检查附件 4 留言主题、留言时间、留言答复的相关完整性是否满足某种规范。用 CHECK 短语指定的列值应该满足的条件。

```
Create table Message(
Muser char primary key//主码，实体完整性，主码不为空
Mid int foreign key references id//参照完整性
Manser nvarchar(10000) not null,--留言答复定义完整性，其内容不为空，若为空则没解决相应留言问题
sname nvarchar(10000) unique,--用户定义完整性，主码唯一
)
```

图 12 数据完整性的检查

步骤三

广义上的可解释性指在我们需要了解或解决一件事情的时候，我们可以获得我们所需要的足够的可以理解的信息。在数据挖掘和机器学习场景中,可解释性被定义为向人类解释或以呈现可理解的术语的能力,从本质上讲,可解释性是类与决策模型之间的接口,它既是决策模型的准确代理,又是人类所可以理解的在自上而下的机器学习任务中,模型通常建立在一组统计规则和假设之上,因而可解释性至关重要,因为它是所定义的规则和假设的基石。此外,模型可解释性是验证假设是否稳健,以及所定义的规则是否完全适合任务的重要手段。与自上而下的任务不同,自下而上的机器学习通常对应于手动和繁重任务的自动化,即给定一批训练数据,通过最小化学习误差,让模型自动地学习输入数据与输出类别之

间的映射关系。在自下而上的学习任务中,由于模型是自动构建的,我们不清楚其学习过程,也不清楚其工作机制,因此,可解释性旨在帮助人们理解机器学习模型是如何学习的,它从数据中学到了什么,针对每一个输入它为什么会做出如此决策以及它所做的决策是否可靠。具体方法如下：

1） 在建模之前的可解释性方法

2） 建立本身具备可解释性的模型

3） 在建模之后使用可解释性方法对模型作解释

基于实例的方法，主要通过一些代表性的样本来解释分类结果的方法。将样本分成三个组团，分别找出每个组团中最具有代表性样例和重要的子空间，使用分类及其对应的代表性样本和代表性特征的子空间。

2.3 结果分析

问题一：

在众多留言中识别出相似的留言，把相似留言打上同一个标签放在一起，归为一类问题，进行不同部门的工作比较，进而可以提升政府的管理水平和施政效率。

问题二：

把特定地点或人群的数据归并，把相似的留言归为同一个问题。如下表 1、表 2 分别所示是附件 3 的热点问题留言明及热点问题排名，以便于提高和优化相关部门的工作，对于各类问题更有针对性，提高解决问题效率。

热度排名	问题 ID	热点指数	时间范围	地点/人群	问题描述
1	1	292	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	253	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
3	3	198	2019/02/26 至 2019/09/13	A2 区天悦幼儿园	幼儿园乱收费
4	4	162	2019/2/21 至 2019/07/09	A3 市居民	A3 市多处未拆迁
5	5	75	2019/07/23 至 2019/8/29	A 市发建开发公司	搞形式主义

表 1 热点问题表

问	留言	留言	留言主题	留言时间	留言详情	点	反

题ID	编号	用户				赞数	对数
1	360104	A012417	A市魅力之城商铺无排烟管道，小区内到处油烟味	2019/08/18 14:44:00	A市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天到凌晨还在营业……	0	0
1	3660105	A120356	A5区魅力之城小区一楼被 搞成商业门面，噪音扰民严重	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，噪声严重扰民，有很大的油烟味往楼上窜，没办法居住……	1	0
1	360106	A235367	A市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019年5月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
1	360109	A0080252	魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	您好：我是魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业，每到晚上睡觉耳边都充斥着吆喝……	0	0
2	360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	西地省A市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面，可是这样搞……	0	0
2	360111	A120445	A市经济学院组织学生外出打工合理吗？	2019/11/5 10:31:38	学校组织我们学生在外边打工,在东莞做流水线工作，还要倒白夜班。本来都在学校好好上课，十月底突然说组织到外省打工……	1	0
2	360114	A0182491	A 市经济学院变相强制实习	2019-04-28 17:32:51	系里要求我们在实习前分别去指的不同公司实训，我这的公司的工作内容和老师之前介绍以及我们专业乎不对口，不做满6个月不给实训分，不能毕业……	9	0
3	195453	A000109415	A2区天悦幼儿园的社团收费太贵了	2019/9/13 9:41:52	我有一个疑惑，天悦幼儿园到底是公办园还是私办园，为什么这么明目张胆的引进各种收取费用的社团？今天下午参加家长会，班主任直接在班级家长会上介绍有戏剧、舞蹈、演讲和口才、机器人、书法、跆拳道，费用从860-1300元一个学期…	1	0
3	283751	A00072134	A2区天悦幼儿园老师推荐孩子去金学教育	2019/2/26 23:03	我是天悦幼儿园的一位家长，在班主任口头推荐下将孩子送至天悦嘉园9栋101金学教育上兴趣班，通过孩	0	0

			上兴趣班		子上了兴趣班以后我存在以下几个疑点…		
4	286165	A0001 5615	A3 区银盆岭村这里为何不拆迁	2019/4/29 15:55:47	市府领导：A3 区银盆岭 岭村位于 A 市政府周边，岳华路以东，彩虹路以西，A3 区大道以北，是一个典型的城中村，但是该村自 2004 年起就被 A 市政府控规，所有土地除了村民安置地外全部划拨或者转让到各个用地单位…	0	1
4	192027	A0002 0229	反映 A3 区 A3 区街道安置小区惟盛园业主的房子问题	2019/2/20 18:04:32	A3 区 A3 区街道 148 亩安置小区惟盛园业主房子问题反映 1.公共面积分摊问题，实际每一层的公共面积 61.9 平方，而扣除业主 89.94 平方，多出的面积怎么来的？…	0	0
4	214709	A0005 1608	A3 区西湖街道茶场村五组何时启动拆迁？	2019/6/17 20:38:45	您好胡书记，据悉，A3 区西湖街道茶场村五组早在三年前已被 A 市 A3 区山国家大学科技城报批了用地红线手续，已经过去三年了，请问政府何时能启动拆迁呢？？看着旁边的六组已经被拆迁了，剩下我们这一小块块的村民也盼着能早日本小康!!谢谢政府。	0	0
5	197455	A0003 7975	举报 A 市发建集团大搞形式主义	2019/8/29 17:48:50	目前，A 市城市建设开发公司由发建集团代管。今年 6 月份，城建开发公司于二十年前建设的某小区的某处出现护坡垮塌，影响到小区外的几栋居民住宅。由于城建开发公司自身经济困难，资不抵债，几年前就已发不出工资，因此没有能力解决此问题，并已向政府相关部门递交了情况说明和报告。此事涉及人民生命财产安全…	1	0

表 2 热点问题留言明细表

问题三：

对于评论留言的答复，得出如下提到答复质量的结论：

1） 区分好评论提出的相关意见，是否具备准确性和科学性。

2） 根据问题所在，实际要解决的问题要区分好具体体现在哪些方面，并确定具体答复策略。

3） 根据附件 4 的留言答复通过检查留言文本数据的相关性、完整性和可解释性，理清问题关键所在，对于提高解决问题的质量很重要。

三、结论

总结本次比赛，我们根据“智慧政务”评论数据的特点，利用构建模型方法对热点问题分类处理以及统计分析出评论数据的情感倾向性，实现了本次的挖掘目标。

本次评论数据挖掘分析的过程中，每一步都通过程序实现，进行了大量的数据挖掘分析工作，实验中的每一步都有理有据，各个步骤之间联系密切，条理清晰且系统地完成了本次数据挖掘分析工作。但是在实验过程中依旧遇到了很多瓶颈问题，例如关于留言的标签分类的分析问题。

将上述得到的结果进行以下几个方面的详细分析：

1） 对于标签分类，方便后续分派到相应职能部门处理，落实到相关部门处理问题的分配，可以减少工作量、提高效率。
2） 对于热点关键词的提取，可以了解到当下热点问题，在相关职能部门处理问题上更加关注相关热点问题。
3） 对于不同的留言信息和主题进行不同属性分析，提炼出该数据的反映情况，并答复信息意见，使群众的问题得以解决。

四、参考文献

[1] 昌攀、曹扬、胥月、张鹏翔,《一种政务文本分类模型的构建方法与流程》，http://www.xjishu.com/zhuanli/55/201911123141_3.html,2020/04/03，2020/04/03
[2] 朱嫣岚，闵锦，周雅倩，等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报，2006，20(1):14 — 20.
[3] 张昊旻，石博莹，刘栩宏. 基于权值算法的中文情感分析系统研究与实现[J]. 计算机应用研究，201229 (12):4571 — 4573
[4] zsffuture,《NLP---文本情感分析》，https://blog.csdn.net/weixin_42398658/article/details/85222547，2019-01-03
[5] 致 Great 文本分类(下)-卷积神经网络(CNN)在文本分类上的应用。链接：https://www.jianshu.com/p/e9108a0f5ac8
[6] obvious 文本数据集的简单处 https://blog.csdn.net/obvious_/java/article/details/80649730
[7] 《机器学习-文本数据-文本的相关矩阵》,https://www.cnblogs.com/my-love-is-python/p/10325005.html，2019-01-26
[8]数据完整性 https://zhidao.baidu.com/question/1302199679650641979.html
[9] SQL 中，什么是数据完整性？数据完整性分为几种？https://zhidao.baidu.com/question/573722134.html?qbl=relate_question_0&word=%CA%B2%C3%B4%CA%C7%CD%EA%D5%FB%D0%D4

[10] 数据库的完整性包括哪三种？通过例子，如果需要也可以使用 SQL 语句，图标等，来说明数据库完整性的作用！
https://zhidao.baidu.com/question/164726796.html?qbl=relate_question_6&word=%CA%B2%C3%B4%CA%C7%CD%EA%D5%FB%D0%D4
[11]王小贱，《深度学习的可解释性研究（一）—— 让模型具备说人话的能力》，
https://zhuanlan.zhihu.com/p/37223341，2018-05-24
[12] 李进锋 杜天宇 李博 机器学习模型可解释性方法、应用与安全研究综述
纪守领 (浙江大学计算机科学与技术学院网络空间安全研究中心)
[13] 赵晓荣 叶呈成 黄佳锋 薛云 《基于电商平台家电设备的消费者评论数据挖掘分析》