“智慧政务”中的文本挖掘应用

摘要

近年来，由于各类网络问政平台逐步成为政府了解民意、集中民智等的重要渠道，各类与社情民意相关的文本数据不断增多，对于以往主要依靠人工进行文本数据处理的相关部门带来极大挑战。但是随着大数据、人工智能等技术的发展，建立基于自然语言处理的智慧政务系统对于政府的管理水平和施政效率都具有极大的推动作用。

自然语言处理是计算机科学领域和人工智能领域中的一个重要方向。可以解决文本分类、计算文本相似度、信息过滤、信息抽取等文本处理问题。本文根据已给出的群众问政留言记录等数据，利用自然语言处理中和和文本挖掘等方法解决三个问题。

针对问题一，本文主要应用自然语言处理中的文本分类方法建立了朴素贝叶斯文本分类模型。首先对文本进行基本去重、去空、分词和过滤停用词的预处理后，将文本内容进行 TF-IDF 特征向量化，将文字内容转化为数字向量的形式，然后建立相应的朴素贝叶斯文本分类模型。最后利用 F-Score 值对模型进行评价。得出各类别 F-Score 的平均值约为 0.90，可证明这个模型的效果还是不错的。

针对问题二，本文从相似留言归类、确定留言热度评价指标和得出评价结果三个方面进行分析。首先是相似留言归类，通过采用余弦相似度原理筛选出相似的留言主题，并归为一类，然后将同一留言内容下的特定的地点或人群与问题描述分开。其次确定留言数、互动量维度、时间维度和情感维度四个指标，将其应用变异系数的方法分别设置权重，根据这些权重计算出最后的热度评价结果。

针对问题三，通过分析问题可以得出，答复意见的质量和留言内容相关，通过求出二者的相关系数得出留言内容和答复意见存在一定的相关性。完整性则从结合答复意见是否满足固定的形式和字数得出，然后从答复意见的结构和要点分析其相关解释。最后结合相关性、完整性和可解释性三者对答复意见的质量给出一套评价方案。

关键词　朴素贝叶斯　变异系数　相似度　TF-IDF　自然语言处理

Abstract

In recent years, as various online political platforms have gradually become an important channel for the government to understand public opinion, the number of text data related to social conditions and public opinions has been increasing, which has brought great challenges to relevant departments that used to mainly rely on manual text data processing.However, with the development of big data, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing plays a great role in promoting the management level and efficiency of the government.

Natural language processing is an important direction in computer science and artificial intelligence.It can solve the problems of text classification, computing text similarity, information filtering and information extraction.In this paper, three problems are solved by means of neutralization of natural language processing and text mining.

Aiming at problem one, this paper mainly USES the text classification method in natural language processing to establish the naive bayesian text classification model.After the text is preprocessed with basic de-weighting, de-nullification, word segmentation and filtering stop words, the text content is transformed into tf-idf feature quantization, the text content is converted into the form of digital vector, and then the corresponding naive bayesian swinburne classification model is established.Finally, f-score is used to evaluate the model.The results show that the average value of various f-scores is about 0.90, which proves that the effect of this model is good.

In view of question two, this paper analyzes three aspects: classification of similar comments, determination of heat evaluation index of comments and evaluation result.Firstly, similar comments are classified. Similar message subjects are selected by using the principle of yu xuan similarity degree and classified into one category. Then, specific places or groups of people under the same message are

separated from the description of the problem.Secondly, the four indexes of the number of comments, interaction volume dimension, time dimension and emotion dimension were determined, and the weights were set by using the method of variation coefficient, and the heat evaluation results were calculated according to these weights.

Aiming at question three, it can be concluded by analyzing the question that the quality of the reply is related to the message content, and the correlation coefficient of the two can be concluded that there is a certain correlation between the message content and the reply content.The completeness comes from the combination of the form and the number of words of the reply, and then from the structure and the main points of the reply, the relevant explanation is analyzed.Finally, a set of evaluation scheme is put forward to evaluate the quality of the replies by the combination of relevance, completeness and interpretability.

Keywords Naive bayes similarity TF-IDF Natural language processing

目录

一、挖掘的背景与目标

1.1 挖掘背景

在当下时代，随着科学技术的进步，微信、微博、市长信箱、阳光热线等网络问政平台逐渐成为政府了解民意、汇聚明智、凝聚名气的重要渠道。各类社情民意相关的文本数据不断攀升，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。

1.2 挖掘目标

本次数据挖掘的的目标是利用给出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正留言的答复意见，利用自然语言处理和文本挖掘的方法解决以下三个问题：

（1）根据给出的数据处理网络问政平台的群众留言，按照一定的划分体系对留言进行分类，以便后续将群众分派至相应的职能部门处理。建立关于留言内容的一级标签分类模型并利用 F-Score 对分类方法进行评价。

（2）根据给出的附件 3 文本数据内容将某一时间段内特定的地点或特定人群问题的留言进行分类，定义合理的热度评价指标，最后给出评价结果。按照表 1 和表 2 给定的格式给出排名前五的热点问题和相应热点问题对应的留言信息，并保存这两个文件。

（3）根据附件 4 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，并尝试实现这一套方案。

二、问题的分析

2.1 问题一的分析

在这一问题中，需要根据给出的附件 2 的文本内容，建立有关于留言内容的一级标签分类模型。这属于一个文本多分类问题，通过分析文本数据集的特点，首先对文本数据进行了去重复值，去除空值，去除数字、字母，jieba 分词以及过滤停用词等操作对数据进行了预处理，接着运用 TF-IDF 算法对文本进行向量化，最后建立一个朴素贝叶斯分类模型对文本进行分类，通过对模型进行训练和优化，最后得出 F-Score 值对模型进行评价。

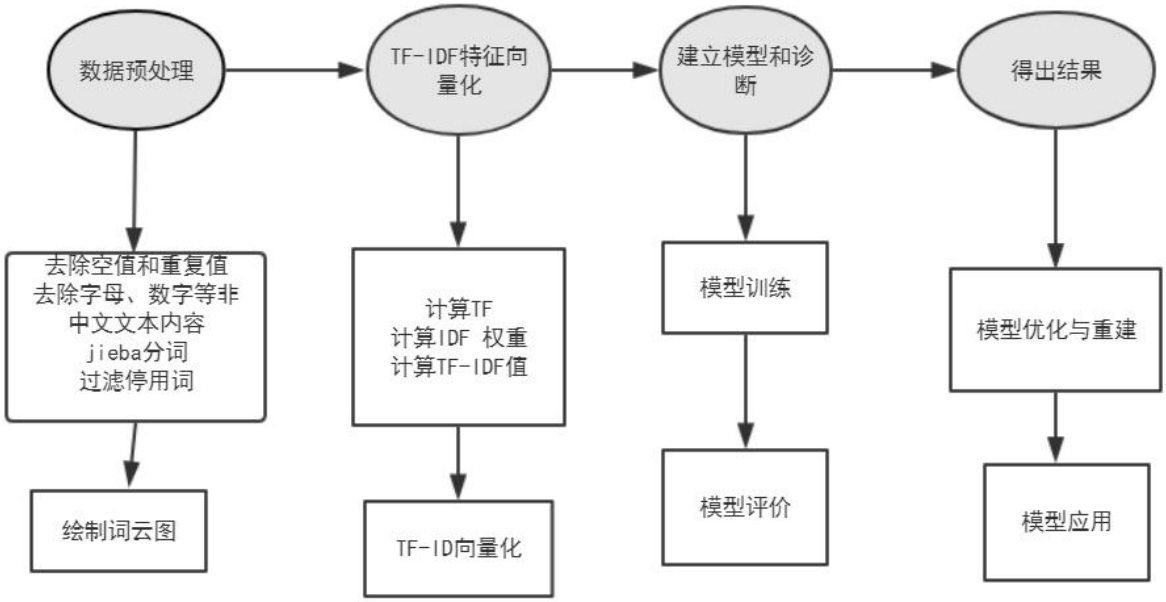

图 1：问题 1 流程图

2.2 问题二的分析

在这一问题中，主要是对留言内容按照特定的地点或人群进行归类，然后定义合理的热度评价指标，得出一个评价结果。简单可以分为留言归类、确定指标和得出结果这三个步骤去解决问题。首先是利用正则表达式的方法提取出特定的地点或人群，接着进行留言归类，本文利用了余玄相似度原理对留言内容进行归类，并确定对应的指标，通过变异系数的方法给每个指标分别设置权重，根据个指标的权重计算最后的得分，并给出评价结果。

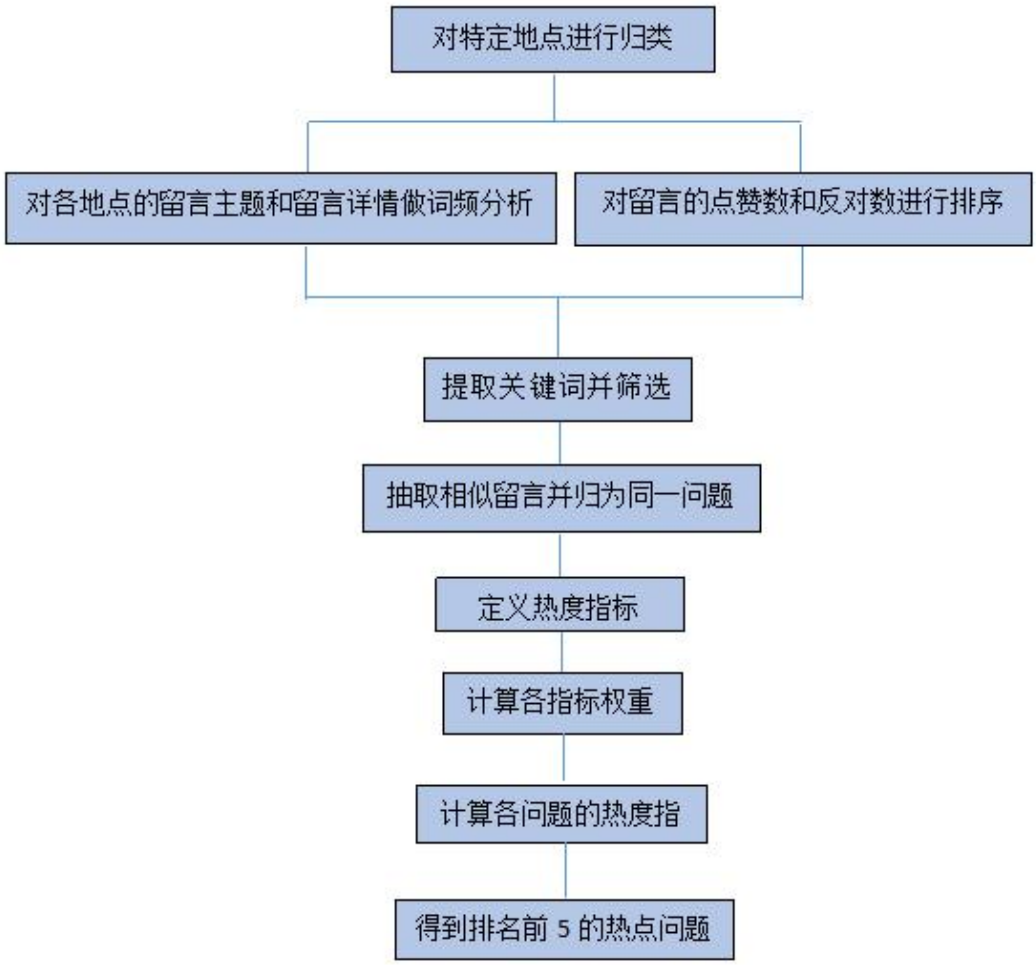

图 2：问题 2 流程图

2.3 问题三的分析

在问题三中，通过分析答复意见的相关性、完整性和可解释性三个角度，每个角度设定一些指标。相关性用留言详情和答复意见的相关系数来表示，通过对文本进行特征向量化处理得出相关系数。完整性用答复意见的时间间隔，答复的形式和字数来进行表示。可解释性通过分析意见叙述的结果来进行评价。最后集合三者给出评价方案，并通过确定的方案去实现。

三、问题一的方法及过程

3.1 数据预处理

3.1.1 去重、去空

根据分析提取附件 2 中的留言详情和一级标签这两类数据，并进行预处理。考虑到数据可能存在重复或者是空值的情况，因此在读取数据的时候，首先查看留言详情这一列数据的空值和重复值，并进行去重和去除空值操作。

3.1.2 去除非中文文本内容

由于留言主题内容都是中文，因此要对中文进行一些预处理工作。首先利用 Python 中的正则表达式（re）方法删除留言详情这一列的文本中的字母、数字、标点符号、字符串等汉字以外的所有字符。

3.1.3 中文分词

中文分词（Chinese Word Segmentation）指的是将一个中文文本按照某种序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

分词是文本挖掘预处理的重要一步，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 中，以中文文本的方式给出数据，为便于转换，先要对留言详情这些内容进行分词。这里采用的是 Python 的中文分词包 jieba 进行分词。

3.1.4 停用词过滤

在解析文本中有很多无效的词和一些标点符号，这些词也被称为停用词。在文本分词的处理结果上，过滤掉一些与本文主题内容无关的词，即称为过滤停用词。

停用词一般包括助词（如：“啊”，“哦”，“吧”等）、形容词、副词、大量重复出现的人称代词（如：“你”，“你们”，“我们”，“它”，“他”，“他们”等类似词语）、各种特殊符号（回车，空行等）以及标点符号等。这些词通常对文本主题和分类基本不起什么作用的词，删除这些词不仅可以节省存储空间，提高代码运行速度，还可以降低特征选取的维度，提高分类效率和准确度。

3.1.5 绘制词云图

词云图又叫文字云，是对文本数据中出现频率较高的关键词予以视觉上的突出，形成“关键词的渲染”，从而过滤掉大量的文本信息。

在本文中，通过绘制词云图来查看文本数据去重、去空、去除字母、数字以及过滤停用词之后的数据预处理效果。

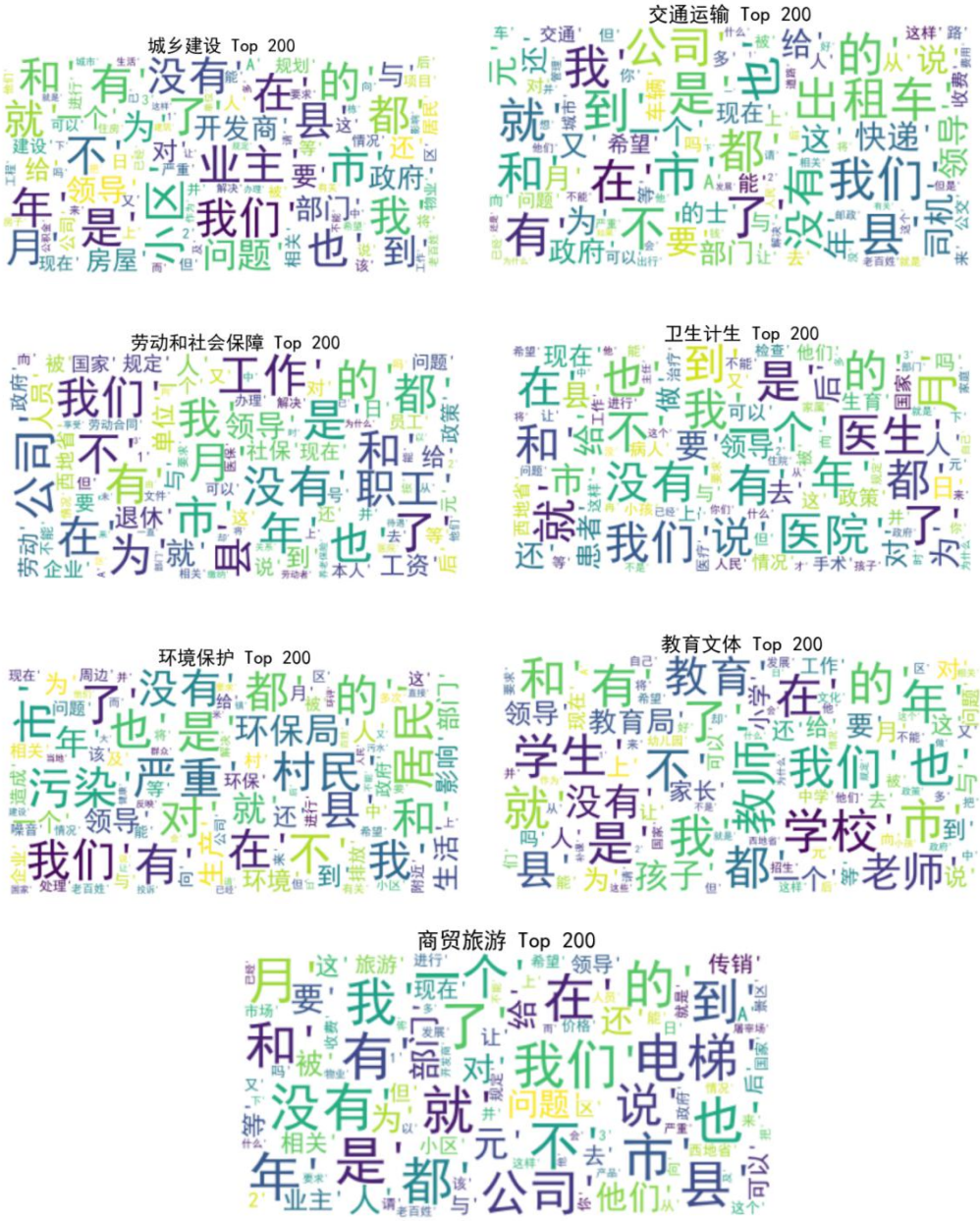

图 3：各类别词云图

3.2 TF-IDF 特征权重计算

在对留言详情进行数据预处理以及分词等操作后，需要将这些词语转换为向量，以供文本挖掘分析使用。这里采用 TF-IDF 算法，把留言详情转换为权重向量。

3.2.1 TF-IDF 简介

TF-IDF（Term Frequency–Inverse Document Frequency）是一种用于信息检索与数据挖掘的常用加权技术。它是一种统计方法，用以估计一个字词对于文本的重要程度。字数的重要性随着它在文本中出现的次数成正比增加。词频

（Term Frequency，TF）指的是某一个给定的词语再该文本中出现的频率。逆向文档频率（Inverse Document Frequency，IDF）是一个词语普遍重要性的度量。

某一特定文本内的高词语频率，以及该词在整个文档集合中的低文档频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 常用于过滤掉常见的词语，保留重要的词语。

3.2.2 基本原理

第一步：计算词频（Term Frequency，TF）：

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{ij}} \tag{1}$$

其中，TF_{ij} 表示词频，n_{ij} 表示某个词在文本中出现的次数，$\sum_k n_{ij}$ 表示该文本的总次数或该文本出现次数最高的词的出现次数。

第二步：计算 IDF 权重，即逆文本频率（Inverse Document Frequency），IDF 越大，说明这个特征在文本中的分布越集中，说明该分词在区分该文本内的内容属性能力越强。

$$idf_i = \log \frac{|D|}{\left|\left\{j : t_i \in d_j\right\}\right|} \tag{2}$$

其中，$|D|$ 是语料库中的文本总数，$\left|\left\{j : t_i \in d_j\right\}\right|$ 表示包含词语 t_i 的文本数目（即 $n_{i,j} \neq 0$ 的文本数目）。如果该词语不在语料库中，就会导致分母为 0，因此一般情况下，使用 $\left|\left\{j : t_i \in d_j\right\}\right| + 1$ 来表示包含词语 t_i 的文本数目。即：

$$idf_i = \log \frac{|D|}{\left|\left\{j : t_i \in d_j\right\} + 1\right|} \tag{3}$$

第三步：计算 TF-IDF 值（Term Frequency–Inverse Document Frequency）。

$$TF - IDF = TF * IDF \tag{4}$$

从 TF-IDF 定义可知，特征词在某个文本中的权重和它在当前文本中出现的频率成正比，与整个文本集中包含该特征词的文档数量成反比。TF-IDF 越大，表明包含该特征词的文本数量比较少并且该特征词在某个文本内容中属于高频词。

第四步：生成 TF-IDF 向量，具体步骤如下：

（1）使用 TF-IDF 算法，得出每个词语的权重；

（2）生成各个标签描述的 TF-IDF 权重向量。

```
(9054, 819875)
------------------------------
  (0, 484455)   0.14489802546436545
  (0, 754756)   0.14489802546436545
  (0, 481955)   0.14489802546436545
  (0, 352930)   0.13423371715911828
  (0, 484501)   0.14489802546436545
  (0, 723674)   0.14489802546436545
  (0, 397049)   0.10543539544512694
  (0, 803158)   0.14489802546436545
  (0, 329414)   0.13865980500966668
  (0, 743766)   0.14489802546436545
  (0, 119429)   0.13865980500966668
  (0, 740637)   0.14489802546436545
  (0, 515680)   0.14489802546436545
  (0, 55434)    0.14489802546436545
  (0, 350061)   0.14489802546436545
  (0, 545751)   0.14489802546436545
  (0, 281244)   0.14489802546436545
  (0, 488719)   0.13865980500966668
  (0, 805919)   0.1279954967044195
  (0, 330610)   0.14489802546436545
  (0, 581654)   0.14489802546436545
  (0, 800349)   0.14489802546436545
  (0, 388208)   0.13080057670311857
  (0, 704009)   0.13423371715911828
  (0, 741221)   0.14489802546436545
  :     :
```

图 4：文本 TF-IDF 特征词的表示

从上图可以看出，这里的维度为（9054，819875），数字 9054 表示这个数据集经过预处理后总共有 9054 条留言详情数据，819875 表示特征数量，包括全部留言详情数据中的所有词语数+词语对（即相邻两个词语的组合）的总数。

3.3 分类器的构造

构造分类模型是文本分类的核心。本文采用监督学习中的朴素贝叶斯分类模型，解决如何根据训练文本集学习构造一个分类器也称为分类函数，接着将其应用在测试文本数据集中华，通过判断其分类的 F-Score 指标进行相关参数的调整。本文使用的是 Python 中的 sklearn 朴素贝叶斯分类器 MultinomialNB。

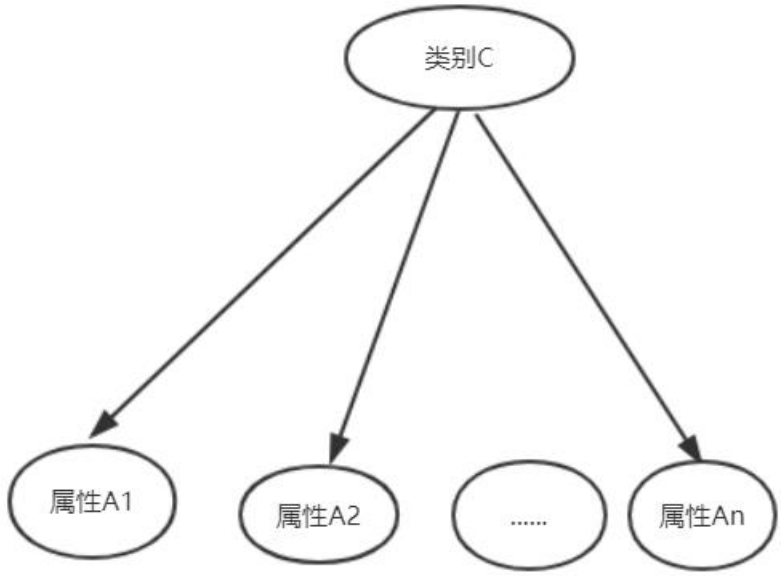

图 5：朴素贝叶斯分类模型的结构图

3.4 建立朴素贝叶斯分类模型

3.4.1 原理

朴素贝叶斯分类模型是一种贝叶斯分类器中应用最为广泛的模型之一，也是一种较为简单的构造分类器的方法，这一模型是将问题分为特征向量和决策向量，并假设问题的特征向量和决策向量之家都是互不相关的。

3.4.2 建模过程

第一步：求基于互相独立的特征项t_i如果出现在这个文档x中，它属于c类别的条件概论。即

$$P(C \mid X) = \frac{P[c]\prod_{i=1}^{n} P(t_i \mid c)}{\prod_{i=1}^{n} P(t_i)} \qquad (5)$$

其中，类别c的先验概论$P(c)$是根据已有的文本数据集所的出来的，$P(t_i \mid c)$是t_i出现在c类别文本中的条件概论；文档x为所包含的次数n。文本分类的目的是找出文本数据集中最有可能属于的哪个类别，即为怎样求得后验概率最大。此时，使用最大似然估计（MLE）的方法来求出$P(t_i \mid c)$和$P(c)$。

第二步：在最大似然估计下，类别c的先验概论为

$$P(c)=\frac{N_c}{N} \tag{6}$$

其中，N_c为训练数据集中类别c所包含的文本总数量，N为训练集中全部类别下的文本总数量，则条件概论为：

$$P(t_i \mid c)=\frac{T_{ct}}{\sum T_{ct}} \tag{7}$$

其中，T_{ct}是表示在c类别下包含特征性t的文本个数，$\sum T_{ct}$为c类别下所包含文本的总数目。

3.5 模型的评估

3.5.1 训练模型

在进行模型的测试时，需要将数据集进行划分，分为测试集合训练集。在给定的样本数据中，拿出大部分的样本数据作为训练集来训练模型，剩余的小部分样本数据使用刚建好的模型来进行预测。

接下来对训练好的模型进行查看混淆矩阵，并显示预测标签和实际标签之间的差异。混淆矩阵的主对角线表示预测正确的数量，除了主对角线外的其余都是为预测错误的。

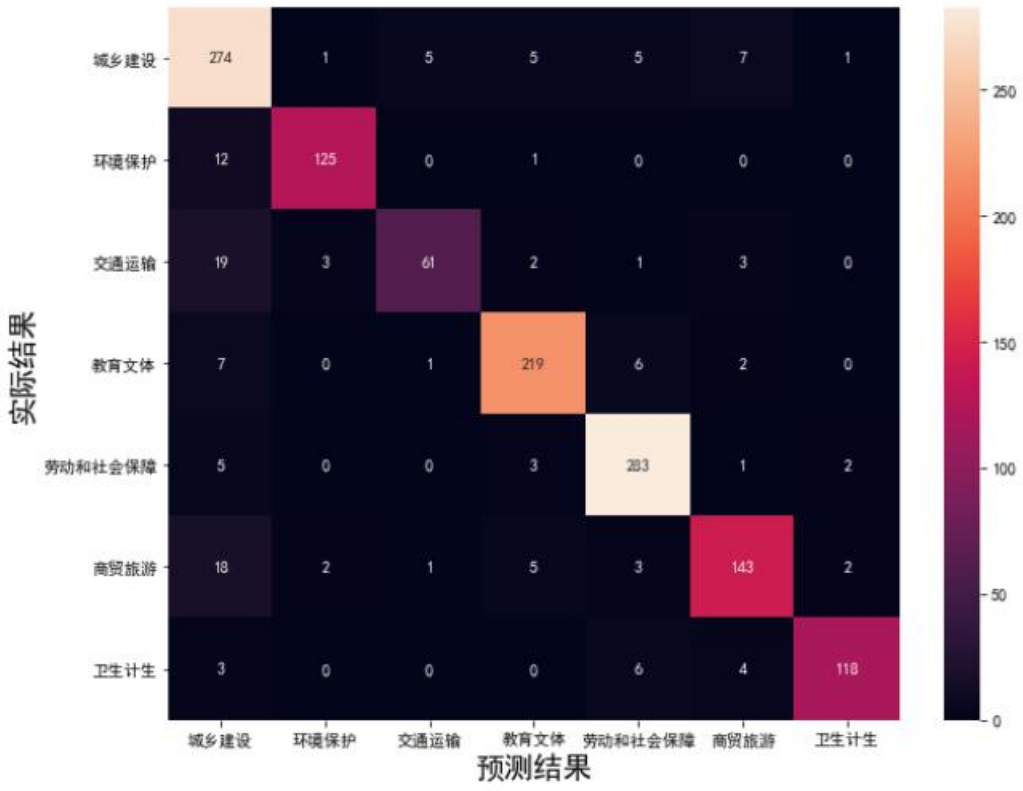

图 6：混淆矩阵图

从图 6 的混淆矩阵可以看出，“卫生计生”类的预测效果做准确，只有 3 例是预测失误的，而“交通运输”类和“商贸旅游”类的预测效果较差，接近 20 例为预测错误。

3.5.2 F-Score 分数

F-Score 也经常表示为 F1 值，F1 值是以每个类别为基础进行定义的，其中包括查准率（precision）和查全率（recall）。查准率也可以称为精确率，是指预测结果属于某一类的个体，实际属于该类的一个比例；查全率也经常称为召回率，是指被正确预测为某个类别的个体数量和数据集中该类别的个体总量的比例。

在分类模型中，经常使用 F-Score 对分类方法进行评价，即所得出来的 F1 值：

$$F1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{8}$$

其中，P_i 表示第 i 类的查准率，R_i 表示第 i 类的查全率。

accuracy 0.8999264164827079

	precision	recall	f1-score	support
城乡建设	0.81	0.92	0.86	298
环境保护	0.95	0.91	0.93	138
交通运输	0.90	0.69	0.78	89
教育文体	0.93	0.93	0.93	235
劳动和社会保障	0.93	0.96	0.95	294
商贸旅游	0.89	0.82	0.86	174
卫生计生	0.96	0.90	0.93	131
accuracy			0.90	1359
macro avg	0.91	0.88	0.89	1359
weighted avg	0.90	0.90	0.90	1359

图 7：基于 TF-IDF 特征处理实现结果

通过选取样本数据集的 85%作为训练数据集，其余的 15%作为测试数据集。图 7 显示测试集上 7 个类别的平均查准率约为 0.90，平均查全率也是约为 0.90，最终平均 F1 值约为 0.90。根据最终的平均 F1 值可知，运用朴素贝叶斯分类模

型的效果较好。

四、问题二的方法及过程

4.1 相似问题归类

4.1.1 提取特定地点或人群

在进行热点问题归类之前，首先通过提取关键词的方法先对地点词进行筛选，采用正则表达式的词语识别、和 Excel 表格的筛选功能先将同一地点的留言归为一类。

正则表达式，又称规则表达式。（英语：Regular Expression，在代码中常简写为 regex、regexp 或 RE)，计算机科学的一个概念。正则表达式通常被用来检索、替换那些符合某个模式(规则)的文本。

正则表达式是对字符串（包括普通字符（例如，a 到 z 之间的字母）和特殊字符（称为“元字符”））操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。正则表达式是一种文本模式，该模式描述在搜索文本时要匹配的一个或多个字符串。

```
for i in theme:
    s=re.findall('[A-Z][0-9]*[^a][\u4e00-\u9fa5]{1,9}',i)#获取地点关键词
    b= [ ''.join(s)][0]#列表转字符
    x.append(b)
res={} #创建字典
for i in x: #对于列表中的字符,
    if i not in res.keys():# 如果字符不是字典里的键，就创建一个对应的键值对
         res[i]=1
    else:#如果已经有了就叠加1个频次
        res[i]+=1
g=sorted(res.items(),key=lambda x:x[1],reverse=True)#排序对象是键值对，排序依据，升序排序 ，把各地点的频次排序
pd.DataFrame({'地点':b,'频次':g})#构建二维数据表
```

图 8：提取地点的代码图

通过使用正则表达式，对“附件 3”的留言主题进行特定地点（例如：A 市 A1 区、A9 市 A1 区、西地省和 A7 县等）筛选。一共可以分为 A 市、A 市 A1 区、A 市 A2 区、A 市 A3 区、A 市 A4 区、A 市 A5 区、A 市 A6 区、A9 市、A9 市 A1 区、A9 市 A4 区、A7 县和西地省共 12 个特定地点，分别放在不同的工作表里。

4.2 留言归类

将相同的地点进行归类之后，接下来需要按照特定的地点或者人群对相似的留言进行归类。

4.2.1 提取关键词

对附件三里面的留言主题，筛选出上述的 12 个特定地点。通过对每个工作表的留言主题和留言详情做数据预处理，和词频分析。得出频数较多的关键词。同时从另一方面对附件三每条留言的点赞数和反对数进行排序，从而可以分别得到点赞数和反对数较多的留言，从点赞数和反对数较多的留言进行关键词和关键地点提取。

4.2.2 相似留言归类

本文是通过对附件三表格里面的留言主题通过利用余弦相似度原理，将文本信息转化为向量进行筛选。得出特定的关键词和特定地点后，在附件三表格里面对留言主题进行筛选，在各特定地点表格的留言主题里面进行关键词的筛选，从而得到相似留言，并归为一个问题。将选取留言数量较多 或者点赞数和反对数量较多的问题另存为热点问题统计表。

余弦相似度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，余弦值越接近 1 就表明夹角越接近 0 度，也就是两个向量越相似，这就叫余弦相似性。在图 9 中，a，b 分别表示两个向量，(x,y)表示向量的坐标，θ为两个向量的夹角。

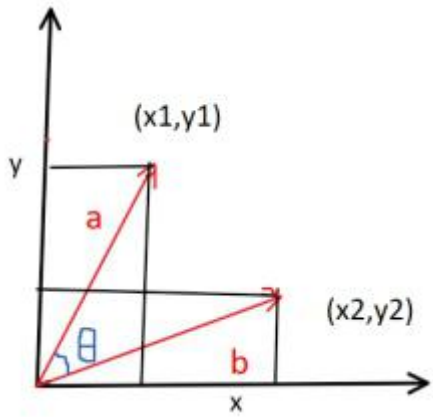

图 9：余弦示图

这样，问题就变成了如何计算，这两个向量的相似程度，我们可以把他们想象成空间中的两条线段都是从原点出发，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同，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，因为这方向正好相反，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。

4.2.3 计算思路

对于两个向量，如果他们之间的夹角越小，那么这两个向量是越相似的，明显相似性就是利用了这个理论思想，它通过计算两个向量的夹角的余弦值来衡量向量之间的相似度值，余弦相似性推导公式如下，余弦函数在三角形中的算公式为：

$$\cos(\theta)=\frac{a^2+b^2-c^2}{2ab} \tag{9}$$

向量a和向量b在直角坐标中的长度为$a=\sqrt{x_1^2+y_1^2},b=\sqrt{x_2^2+y_2^2}$，向量$a$和向量$b$之间的距离我们用向量$c$表示，就是上图中的黄色直线，那么向量$c$在直角坐标系中的长度为$c=\sqrt{(x_2-x_1)+(y_2-y_1)}$，将$a$，$b$，$c$带入三角函数的公式中得到如下的公式：

$$\begin{aligned}\cos(\theta)&=\frac{a^2+b^2-c^2}{2ab}\\&=\frac{x_1^2+y_1^2+x_2^2+y_2^2-(x_2-x_1)-(y_2-y_1)}{2\sqrt{x_1^2+y_1^2}*\sqrt{x_2^2+y_2^2}}\\&=\frac{x_1*x_2+y_1*y_2}{\sqrt{x_1^2+y_1^2}*\sqrt{x_2^2+y_2^2}}\end{aligned} \tag{10}$$

这是 2 维空间中余弦函数的公式，那么多维空间余弦函数的公式就是：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{11}$$

4.3 确定热度指标

通过分析问题，确定出五个评价指标，分别为：同一个地点同一个人群的留言总数，问题的总点赞数和反对数，留言问题的时间跨度和情感分析得分的平均数。

4.3.1 留言总数

在对相似问题进行归类后，每一个相似问题的留言总数也都不相同，留言总

数相对较多的说明这一问题被比较多人反映。通过计算同一留言问题的总数可以反映留言热度。

4.3.2 互动量维度

互动量维度即为留言的总点赞数和总反对数，在众多问题中，如果某一个问题被较多人点赞或者反对，说明这一问题相对受人关注，也可以定义为一个热度指标。

4.3.3 时间维度

时间维度可以简单理解为同一个问题的时间跨度，时间跨度较长的说明这一问题被大家持续关注，而时间较短的说明不那么引人关注。

4.3.4 情感维度

情感维度即为某一个留言的情感分析得分，通过对每条留言进行情感分析，得出一个分数，然后算出某一问题的所有留言的情感分析得数，求平均值，用平均值来表示这一问题的情感分析得分。

文本情感分析：又称意见挖掘、倾向性分析等。简单而言，是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。互联网(如博客和论坛以及社会服务网络如大众点评)上产生了大量的用户参与的、对于诸如人物、事件、产品等有价值的评论信息。这些评论信息表达了人们的各种情感色彩和情感倾向性,如喜、怒、哀、乐和批评、赞扬等。基于此，潜在的用户就可以通过浏览这些主观色彩的评论来了解大众舆论对于某一事件或产品的看法。

4.4 确定热度指标权重

由于评价指标体系中的各项指标的量纲不同，不宜直接比较其差别程度。为了消除各项评价指标的量纲不同的影响，需要用各项指标的变异系数来衡量各项指标取值的差异程度。因此需要对所确定的所有指标分别确定一个权重，从而进行评价。

本文采用的是变异系数法对各个指标来确定一个合理的权重。变异系数法是直接利用各项指标所包含的信息，通过计算得到指标的权重。是一种客观赋权的方法。此方法的基本做法是：在评价指标体系中，指标取值差异越大的指标，也就是越难以实现的指标，这样的指标更难反映被评价单位的差距。各项指标的变异系数公式如下：

$$v_i = \frac{\sigma}{x_i} (i = 1,2,3,4....n) \tag{12}$$

其中，v_i表示第i项指标的变异系数，σ_i表示第i项指标的标准差，$\overline{x_i}$为第i项指标的平均值。

则可得出各项指标的权重为：

$$w_i = \frac{v_i}{\sum_{i=1}^{n} v} \tag{13}$$

最终确定的权重：留言数占比 18.4%，反对数占比 28.3%，点赞数占比 31.9%，时间跨度占比 14.2%，情感分析平均数占比 0.72%。

4.5 评价结果

在热度评价分析的时候，需要量化评价指标，比较留言问题中哪个热度比较高以及如何对它们进行合理的排位，其中主要过程是先选出热度评价指标、计算指标权重和问题热度指数，然后根据热度指数对问题进行热度排名，进而得出评价排名结果。

图 10：评价分析过程的流程图

评价结果根据得出的各项指标权重，应用公式：

$$f(x) = 0.184x_1 + 0.283x_2 + 0.319x_3 + 0.142x_4 + 0.072x_5 \tag{14}$$

其中，$x_i(i = 1,2,3,4,5)$表示各项指标。

最终得出前五的热点问题分别为：

① 西地省 58 车贷立案无进展

②A 市 A5 区汇金路五矿万境 K9 县存在一系列问题

③A 市绿地海外滩二期和高铁规划建设存在问题

④A 市经济学院强制实习

⑤A 市地铁线路开通以及建设工程问题

五、问题三的分析过程及方法

问题三主要从三个方面评价答复意见的质量，分别为：相关性，完整性和可解释性。

5.1 相关性：

答复意见的内容是否与问题相关，是否具有针对性，能否通过要点论述建议和做出合理的答复。相关性分析一般是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密切程度。相关性的元素之间需要存在一定的联系或者概率才可以进行相关性分析。相关性不等于因果性，也不是简单的个性化，相关性所涵盖的范围和领域几乎覆盖了我们所见到的方方面面，相关性在不同的学科里面的定义也有很大的差异，通常可以根据相关系数来衡量这一个特性。

要计算文本之间的相关系数，通常需要将文本转化为向量来表示，进而求得相关系数。在统计学界中，通常认为相关系数 0.00 - ±0.30 是具有微相关，±0.30 - ±0.50 是具有实相关性，±0.50 - ±0.80 具有显著相关，±0.80 - ±1.00 是具有强相关性。

最后通过比较附件 4 中的留言详情和答复意见，计算两者之间的相关系数来描述两者之间的相关性和相关程度。可以得出留言详情和答复意见具有一定的相关关系。

5.2 完整性

留言答复意见的完整性是指答复的意见是否满足某种规范，一般情况来看，所遵循的规范中，留言回复的时效性、效率和答复的质量是较为重要的几项指标，针对留言的时间在规范的时间内做出一个合理有质量的建议措施是很有必要的，也符合我们实际生活的需要，同时，回复内容的长短也有一定的规范性，如果内容没有侧重点和简略部分，回复的意见可实施效率也比较低，为此，主

要按照答复意见的时间间隔，答复意见的形式和答复意见的字数这三个指标进行衡量留言答复的完整性特点。

5.2.1 留言回复的时间间隔

根据留言答复时间和留言时间，使用 Excel 函数的相关计算得出相应的时间间隔得出下面中的表 1：

表 1： 答复时间间隔数据表

留言时间	答复时间	答复时间间隔
2019/4/25 9:32	2019/5/10 14:56	15
2019/4/24 16:03:40	2019/5/9 9:49:10	15
2019/4/24 15:07:30	2019/5/9 9:49:42	15
2019/4/23 17:03:19	2019/5/9 9:51:30	16
2019/3/29 11:53:23	2019/5/9 10:18:58	41
2018/12/31 22:21:59	2019/1/29 10:53:00	29
2018/12/31 9:55:00	2019/1/16 15:29:43	16
2018/12/31 9:45:59	2019/1/16 15:31:05	16

找出最大和最小的时间间隔，按照合理的划分体系对其进行时间的划分，统计每个时段的留言数量，并绘制相应的时间间隔分布表，如表 2：

表 2： 时间间隔分布表

时间间隔	留言数量
0-240	2803
240-480	11
480-720	0
720-960	1
960-1200	1

集中分析答复时间，如图 11 的答复时间分布散点图上可以看出留言回复时间大致集中在留言之后的 3 个月内，而且时间点的分布较为均匀，少数留言答复比较延迟，要避免忽略关键性的问题，在时效方面还应该加强，但是总体上留言回复工作效率比较高，能够根据留言的具体时间及时对留言进行分析，给出合理的回复意见。

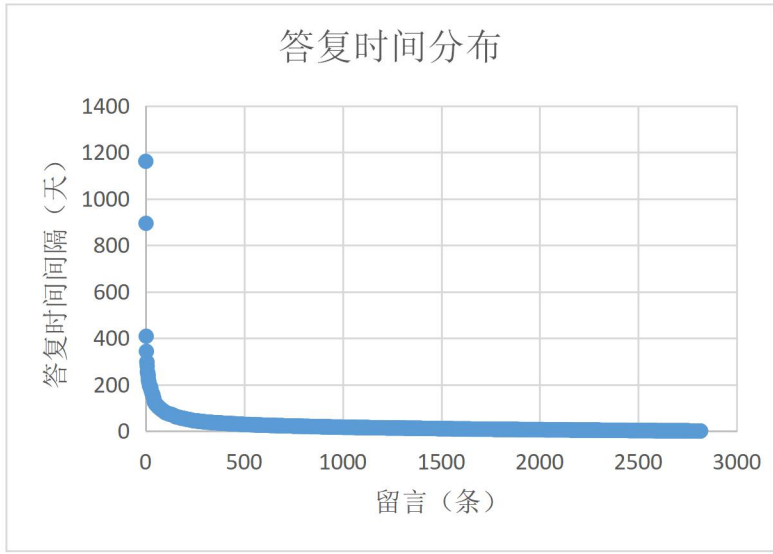

图 11：答复时间分布散点图

5.2.2 答复的留言的形式

通过筛选答复内容，筛选结果显示，其中留言的格式较为规范，开头、结尾、称呼和时间点等形式表达合理，较为规范；另外针对重要问题还能分点回答，从多个方面和角度进行较为全面的分析，层次分明，条例清晰，避免数据的交杂和重复，使人浏览的时候一目了然。

5.2.3 留言答复内容的字数

留言答复意见在较为全面答复留言的内容和问题的基础上，简要地表现出了主要的观点和看法，避免了文本内容过多累赘和重复，有条理和突出要点的中心，明确主要的内容，有针对性地回答留言内容的问题。

5.3 可解释性

答复意见的可解释性是关于答复意见中内容的相关解释，其合理性的判定标准是留言意见中的层次结构关系和要点分布，在留言中能有效地按一定的层次进行分点答复是比较直观的，通过分层作答的解释性和结构性相对比较合理优化，所得出的结论和给出的建议也是较为具体的评价指标。

5.3.1 意见叙述的结构

答复的内容中针对具体的留言问题做出合理的建议和相应的回答，能够集中问题的要点进行回答并且解释较为详细，地区性问题有一定的说服力，大部分问题的说明都有一定的理论基础和依据，比较有说服力，引用留言中的关键点开展开意见答复，避免了重要内容的缺失，整体结构较为完整和全面，抓住

了留言问题中的时间和地点这些重要的引导词，提出相应的答复意见，对建议部分给出了合理的分析，较为精准和明确。但是有一小部分没有具体答复和针对性性建议，如有些答复意见只是简单的回复一些“网友：您好！留言已收悉”这寥寥数语的内容，还需要注意数据过多带来的数据遗漏或是缺失，这一问题还有待改进。

六、参考文献

[1]张航. 基于朴素贝叶斯的中文文本分类及 Python 实现[D].山东师范大学,2018.

[2]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,2016,34(07):147-154.

[3]喻凯西. 朴素贝叶斯分类算法的改进及其应用[D].北京林业大学,2016.

[4]阿曼. 朴素贝叶斯分类算法的研究与应用[D].大连理工大学,2014.

[5]朱晓丹. 朴素贝叶斯分类模型的改进研究[D].厦门大学,2014.

[6]赵玎,陈贵梧.从电子政务到智慧政务:范式转变、关键问题及政府应对策略[J].情报杂志,2013,32(04-207+197.

[7]https://www.jianshu.com/p/fb454a4b383d

[8]https://blog.csdn.net/u013421629/article/details/81171361

[9]https://blog.csdn.net/jediael_lu/article/details/77863419