基于自然语言处理的智慧政务应用

摘 要

网络问政是我国电子政务系统中的重要组成部分，其目的是为了政府能够更容易的了解民生民情，重视群众在网络中表达的问题，且通过各种途径进行回应并有效解决。大部分网络问政平台在处理群众留言时依然采用人工分类的形式，它具有工作量大、效率低、差错率高等特点，随着时代发展，**智慧政务**行业相关部门需要利用政务服务网系统数据，对群众留言建立更加智能、科学的分类系统。不仅如此，还要进行热点问题的挖掘和建立相关部门答复意见的评价方案，针对性处理问题，从而进行更优质的便民服务。本文将对大量数据库进行分析，解决所面临的问题。

问题一，我们首先对所给数据进行机器化语言处理（采用 Python 的 xlrd 及 xlwt 库对留言进行读取和编写，运用 jieba 库分词从而构建词袋模型），其次采用基于 **NLP 系统**的 **Word2Vec** 的方法，进行降维处理。将词转换为词向量的形式建立分类模型，最后，再提高分类器性能并对分类模型进行测试评价。

问题二，我们运用 Python 进行文本挖掘，对文本分词后，过滤掉语气助词一类的词；再利用词频分析，根据留言分词内容在文档中出现的次数和点赞数进行统计分析，为了更方便浏览文本主旨，可以使用词云对文本中词频较高的分词给与视觉上的突出；最后运用文本挖掘中主题建模的方法，提取留言内容的主题信息，以及通过 **TF-IDF 算法**实现将相似问题进行归类，进行残差分析。通过模型及热度评价指标找出留言群众最关心的 5 个热点问题，进而找到所有相关热点问题的留言并制成表格。

问题三，我们基于神经网络的分析，选取逻辑回归、支持向量机和随机森林三种分类模型，对答复内容进行三层递进式训练和检验，将分别从答复意见的时间特征、初始结构特征和文本内容特征三个维度构建答复质量评价分类器。

关 键 词 ： **NLP 系统　Word2Vec 模型　降维处理　TF-IDF 算法　残差分析　神经网络　逻辑回归模型　支持向量机模型　随机森林模型**

Abstract

As an important part of China's e-government system, online political inquiry aims to make it easier for the government to understand the people's livelihood, pay attention to the problems expressed by the people on the Internet, and respond to and effectively solve the problems through various channels. Most of the network political platform still use the form of manual classification when dealing with the comments of the masses, it has the characteristics of large workload, low efficiency, high error rate, with the development of The Times, the relevant departments of the intelligent government industry need to use the government service network data to establish a more intelligent and scientific classification system for the comments of the masses. Not only that, but also to carry out the mining of hot issues and establish the relevant departments to reply to the comments of the evaluation program, targeted to deal with the problem, so as to carry out better quality services for the convenience of the people.This article will analyze a large number of databases to solve the problem.

In problem 1, we first process the given data in machine language (using Python's XLRD and XLWT libraries to read and write the message, using jieba library word segmentation to build the word bag model), and then use Word2Vec method based on NLP system to reduce the dimension.

In question 2, we use Python for text mining, and filter out words like modal particles after text segmentation.Word frequency analysis is used to conduct statistical analysis based on the frequency of word segmentation in the document and the number of thumb up. In order to browse the text theme more conveniently, word cloud can be used to give visual prominence to word segmentation with higher word frequency in the text.Finally, the method of topic modeling in text mining is used to extract the topic information of message content, and similar problems are classified by tf-idf algorithm to carry out residual analysis.Through the model and the heat evaluation index to find out the most concerned about the five hot issues, and then find all the relevant hot issues of the comments and make a table.

In question 3, we analysis based on neural network, logistic regression, support vector machines and random forests three kinds of classification model, the three layers of progressive training and the content of the reply inspection, respectively from the reply on time, the initial structure characteristics and text content reply three dimensions to build quality assessment classifier.

Key words: NLP system Word2Vec model dimension reduction process
Residual analysis The neural network Logistic regression model
Support vector machine model Random forest model

目录

一、问题重述

1.1 问题背景

近年来，互联网、微博等新兴媒体形式的普及给予了公共讨论的广阔空间，越来越多的市民群众可以利用网络发表自己的意见、诉说自己的需求。随着网民规模的不断扩大，网络媒体、政府门户网站、微博等纷纷推出问政平台，互联网不仅成为生活方式的补充，也成为人民问政的地带。网络问政逐渐成为一种新的发展趋势，这种趋势与互联网的发展是分不开的。

网络问政的发展，一方面有利于公众积极参与社会问题的讨论，承担社会责任；另一方面，政府能够更容易的了解民生民情，重视群众在网络中表达的问题，且通过各种途径进行回应并有效解决。数据显示，近些年移动端网民留言量逐渐上升，同时政府不断鼓励、促进基层单位参与留言处理，解决民生问题，从而加入“领导留言板”的市县不断增加。

1.2 本问题研究现状

电子政务是指公共管理部门全面运用现代信息技术、通信技术和管理理论，将管理和服务通过网络技术进行集成，实现组织结构和业务流程的优化重组，超越世间、空间和部门分割的限制，全方位地实施对社会的管理职能，是一种全新的管理方式，向社会提供优质、规范、透明的公共服务。

现阶段，我国电子政务整体尚处在初步发展阶段，尽管媒体功能已经基本成熟，但电子政务绩效评估仍存在不足，同时信息化政策尚待完善，发展呈现不平衡。例如，在处理网络问政平台的群众留言时，首先需要按照一定体系将留言分类，而这类环节还是通过大量人力完成，存在工作量大、效率较低的问题，随着信息化浪潮的来临，建立一个便捷高效、迅速可靠的信息管理系统已经成为提高效率问题的重要方案[1]。

总体而言，我国的电子政务建设虽然取得了重要进展，但整体水平较低，地区发展不平衡，尚处于电子政务的起步阶段。

1.3 结合现实情况，本次需要解决的问题

问题 1：根据所给数据及划分体系，建立关于留言内容的一级标签分类模型。

问题 2：参考问题 1 的分类模型，将留言内容为某时段内反应特定地点或特定人群问题的留言归类，并且制定科学的热度评价指标，建立热点问题评价模型。与此同时，根据数据制作“热点问题表.xls”和“热点问题留言明细表.xls”。

问题 3：针对数据中相关部门对留言用户的答复内容，从其相关性、完整性等多角度分析制定一套评价方案且将其实现。

二、问题分析

综述

大部分电子政务系统在处理群众留言时依然采用人工分类的形式，它具有工作量大、效率低、差错率高等特点，随着网络问政平台的不断发展，智慧政务行业相关部门需要利用政务服务网系统数据，对群众留言建立更加智能、科学的分类系统。不仅如此，从对留言内容分析进行挖掘热点问题和建立相关部门答复意见的评价方案知道，建立完善的电子政务系统有助于提高政务行业的工作效率，针对性处理问题，从而进行更优质的便民服务。

问题给出四个附件数据，附件一提供群众留言内容分类三级标签体系，将群众留言从三级标签开始分类，将 517 个数据划分了 15 个一级标签类别。附件二包括留言用户的留言内容数据，包括留言编号、留言用户、留言主题、留言时间、留言详情、一级分类 6 个板块。附件三在附件二划分的基础上，新增了点赞数和反对数两个指标。附件四给出相关部门对留言的答复意见数据，包括答复意见、答复时间等 7 个板块。

2.1 问题一的分析

问题一要求建立关于留言内容的一级标签分类模型。基于 NLP（自然语言处理）系统，建立智能分类模型首先要将留言内容录入，其中采用 Python 的 xlrd 及 xlwt 库对附件 2 中文本进行读取和编写；其次将留言内容进行重新组词，此时运用 jieba 库进行分词工作构建**词袋模型**；由于自然语言具有复杂性的特点，例如，自然语言处理中两个近义词或反义词，可能出现在编码上完全不同但在语义上高度相关的情况，所以我们决定利用**独热表示**和**分布式表示**来建立 gensim.word2vec **模型训练**，将词转换为词向量的形式建立分类模型；最后，提高分类器性能并对分类模型进行测试评价。在实际问题中，

要注意解决文本语义交叉的问题、将长文本转化为短文本的问题、将多分类问题转化为多个二分类问题的处理。

2.2 问题二的分析

问题二要求将某一段时间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，按照附件内容给出排名前 5 的热点问题及建立全面的热点问题明细表。首先在一级标签分类模型的基础上，运用 Python 进行**文本挖掘**，在对文本进行分词后，过滤掉语气助词一类的词；其次利用**词频分析**，根据留言分词内容在文档中出现的次数和点赞数进行统计分析，如果分词中包含部分停用词，可以通过 if 判断移除，为了更方便浏览者对文本主旨的查看，可以使用词云对文本中词频较高的分词给与视觉上的突出；最后运用文本挖掘中**主题建模**的方法，提取留言内容的主题信息，以及通过 **TF-IDF 算法**实现对关键词的提取。将相似问题进行归类，通过模型及热度评价指标找出留言群众最关心的 5 个热点问题，进而找到所有有关热点问题的留言并制成表格。

2.3 问题三的分析

问题三要求建立一个关于答复内容的评价模型，我们将分别从答复意见的时间特征、初始结构特征和文本内容特征三个维度构建答复意见评价特征体系。首先通过**无量纲化处理**将数据转化为统一规格，对照留言时间和答复时间给其制定指标来作为衡量答复的及时性的时间特征；其次，我们将答复内容的文本长度、平均句长、标点符号的数目等等看做初始结构特征，用(ans_length,）等代码和对其定义的指标实现答复意见的初始结构特征的评价；关于文本内容特征，其包括答复内容的相关性、可解释性、完整性等指标，我们选择**支持向量机模型**和**组合分类模型**作为待构建的模型，采用**皮尔森相关系数法、最大互信息系数法、L1 正则化、L2 正则化、递归特征消除法、随机森林平均不纯度减少法**对构造的特征进行评分，并对得分实施归一化处理，解决了答复内容的可解释性和完整性的评价问题。对于答复内容的相关性，要通过 word2vec 训练词向量，在词粒度上挖掘文本的语义信息，通过文本主词干及相关词集实现语义扩展，然后使用 **LDA 挖掘模型**实现建立主题相关性的特征。最后采用算法 SVM、LR 及**决策树 C4.5** 构建答复质量评价分类器。做到将问题量化，通过计算指标来建立评价模型。

2.4 数据分析及预处理

在本文模型中，我们需要采用机器化语言的方式进行分类，即将自然语言处理转化成机器语言处理。因此，我们引用 **UTF-8 模型**和 **jieba 分词模型**完成将自然语言录入机器并将其进行分词的过程，同时这两大步骤也作为我们的数据预处理部分。

我们在机器化语言的进程中发现，在进行语音处理时，是以数字计算为基础，根据获得的参数变化规律合成语音信号，进而实现语句的识别；在图像处理中，根据图像像素进行识别从而进行操作。因此处理自然语言也可以使用大致的方法，但由于自然语言的复杂性、多样性，无法将所有语言都进行数字化的处理。特别的，要想将汉字录入系统，往往存在内存空间和字节顺序的问题。

为了解决这些困难，我们引用了 **UTF-8**（Unicode Transformation Format）编码处理汉字来完成自然语言处理数字化表示，如下图所示。

位置	UTF-8 编码表	GB-2312 编码表	汉字与字符
1	00A4H	A1E8H	○
2	00A7H	A1ECB	§
…	…	…	…
586	4E01H	B6A1H	丁
587	4E03H	C6DFH	七
…	…	…	…
7444	FFE3H	A1EAH	￣
7445	FFE5H	A3A4H	￥

图 1. 语言数字化示意图

计算机处理文本时，需要先将文本转换为数字。最早只有 127 个字符被编码到计算机，包括大小写字母、数字和部分符号，这种编码称为 ASCII 码，例如 A 的编码是 65。通常情况下 8 个 bit 组成一个字节，同时一个 ASCII 码就是一个字节，但如果要表示中文，至少需要两个字节，由此引入 Unicode 编码解决将自然语言数字化处理。

Unicode 编码扩展自 ASCII 码，其规模可以容纳 100 多万个符号，是一个很大的集合，且通常用两个字节表示一个字符。但使用 Unicode 编码需要更多的存储空间，同时它只规定了符号的二进制代码，而没有规定代码如何存储，进而出现了更统一的、“可变长编码”的 UTF-8 编码。

Unicode 有多种实现方式，其中 UTF-8 是互联网上使用最广的一种实现方式，它可

以使用 1~6 个字节表示一个符号，根据不同符号变化字节长度。它的编码规则为：1) 对于单字节的符号，字节的第一位设为 0，后面 7 位为这个符号的 unicode 码。2) 对于 n 字节的符号(n>1)，第一个字节的前 n 位都设为 1，第 n+1 位设为 0，后面字节的前两位一律设为 10。剩下的没有提及的二进制位，全部为这个符号的 unicode 码。

在 UTF-8 编码规则中，如果是一个字节，那么最高的比特位为 0，这样可以兼容 ASCII 码，且 UTF-8 是包括 UTF-32 和 UTF-16 在内的三种方式中唯一兼容 ASCII 编码的。此编码使用 1-6 个字节为每个字符编码，其中 128 个 ASCII 字符只需要一个字节编码，大部分汉字使用三个字节编码。UTF-8 使用尽量少的字节存储一个字符，这样不仅节省存储空间，还能在传输时节省流量。另外，Unicode 和 UTF-8 之间的转换可以通过程序实现。下表总结了 UTF-8 编码规则，字母 x 表示可用编码的位。

```
//#txt---
   |  Unicode符号范围        |  UTF-8编码方式
 n |  (十六进制)             |  (二进制)
---+-----------------------+
 1 | 0000 0000 - 0000 007F |                                                0xxxxxxx
 2 | 0000 0080 - 0000 07FF |                                       110xxxxx 10xxxxxx
 3 | 0000 0800 - 0000 FFFF |                              1110xxxx 10xxxxxx 10xxxxxx
 4 | 0001 0000 - 0010 FFFF |                     11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
 5 | 0020 0000 - 03FF FFFF |            111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
 6 | 0400 0000 - 7FFF FFFF | 1111110x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
//#txt---end
```

图 2. UTF-8 编码规则

将语言机器化后，我们采用 **jieba 分词**将文本内容去符号重新组词，从而构建**词袋模型**。在汉语中，词是最小的能够独立活动的有意义的语言成分，汉语是以字为单位，与西方语言不同，词与词之间没有空格之类的标志指示词的边界。分词问题是中文文本处理的基础性工作，分词好坏对后面的中文信息处理起着关键作用。jieba 分词则运用算法实现分词的形式，通过一个中文词库，将待分词的内容与分词词库进行比对，生成有向无环图(DAG)，经过图结构和动态规划方法找到最大概率的词组。除了分词，jieba 库还提供了自定义中文词组的功能[2]。

jieba 分词是基于前缀词典、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况所构成的有向无环图(DAG)。例如，句子“抗日战争”生成了{0:[0,1,3]}

这样一个简单的 DAG，就是表示 0 位置开始，在 0，1，3 位置都是词，就是说 0~0，0~1，0~3 即“抗”，“抗日”，“抗日战争”这三个词，其在 dict.txt 中是词。

其次，jieba 分词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。DAG 的起点到终点会有很多路径，因此需要找到一条概率最大的路径，基本思路就是对句子从右往左反向计算较大概率，依次类推，最后得到较大概率路径及较大概率的切分组合。jieba 对于未登录词（没有在前缀词典里收录的词），采用了基于汉字成词能力的 HMM 模型，HMM 模型有 5 个基本组成：观测序列、状态序列、状态初始概率、状态转移概率和状态发射概率。出现未登录词，句子会作为观测序列，当有新句子进来时，具体做法为：先通过 Viterbi 算法求出概率最大的状态序列，然后基于状态序列输出分词结果。[1]

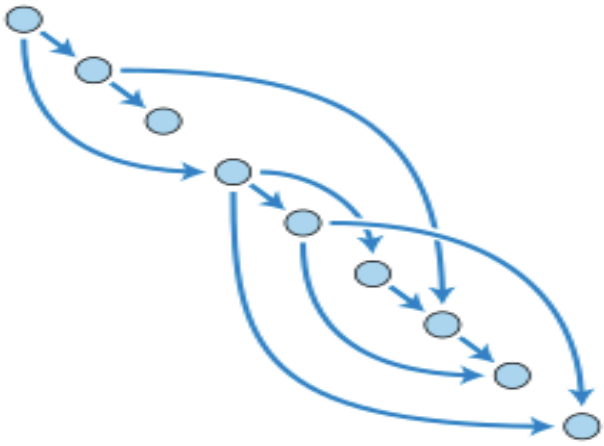


图 3. 常见的有向无环图（DAG）

jieba 分词支持三种模式：（1）全模式，把句子中所有可成词的词语都扫描出来，速度极快。（2）精确模式，它将句子精确切分，没有多余重复的内容，更适合文本分析。（3）搜索引擎模式，在精确模式的基础上，对长词再次切分，提高搜索的召回率。我们在分析问题的基础上，选择了**精确模式**。具体分法如下图所示：

```
1 [全模式]:  我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学
2 [精确模式]:  我/ 来到/ 北京/ 清华大学
3 [默认模式]:  我/ 来到/ 北京/ 清华大学
4 [搜索引擎模式]:  我/ 来到/ 北京/ 清华/ 华大/ 大学/ 清华大学
```

图 4. jieba 分词三种模式图解

综上而言，jieba 分词通过 DAG、动态规划，或是对于未登录词的 HMM 模型等方法，确保了我们将 Excel 中的文本机器化后，精准高效的将一个汉字序列切分成一个个单独的词，进而得到词袋模型。

三、模型假设

基于对问题的分析，我们做出如下假设：

1. 假设题目所给的数据真实可靠；
2. 假设政府部门对每条留言都进行回复；
3. 假设所给的文本数据中，无乱码和特殊字符；
4. 假设附件二中的一级分类标签种类为标签个数；
5. 假设群众在反馈给政府部门热点问题时，每位居民针对某一问题只反映一次，无就同一问题多次反馈的情况；
6. 假设部门在向用户回复问题时，行文无拼写错误及语法问题。

四、定义与符号说明

符号	符号说明
TF	特征词在文本中出现的频率
IDF	所有留言文本中出现某一特征词的留言数
X、Y	变量
u	样本均值
σ	样本标准差
$cov(X,Y)$	协方差
r	样本相关系数

(这里只列出论文各部分通用符号，个别模型单独使用的符号在首次使用时再进行说明。)

五、模型的建立与求解

综述

由于第一问需求分类模型，所以采用基于 NLP 的 **Word2Vec** 的方法，进行**降维算法**的运算。第二问采取 **TF-IDF 算法**，并根据合理的**热度评价指标**进行热点问题的筛选，并进行**残差分析**。第三问基于**神经网络**的分析，采用算法 SVM、LR 及**决策树** C4.5 构建答复质量评价分类器。

5.1 问题一的模型建立与求解

为了更直观反映数据，所以在大量的数据中提取出有效数据、减小数据误差，需对数据中的每一组数据合理性进行统计分析，缺失数据不予考虑，对特征数据进行提取建立模型。

首先，我们利用 Python 中的 xlrd 和 xlwt 库对 Excel 文件进行处理。其中 xlrd 主要是提供快捷的读取方式，通过读取表格读取单元格，获取表格的数据。在 xlwt 中正常输入中文会出现编码错误，说明 xlwt 默认支持 ASCII 码，可以使用 UTF-8 编码，即代码中的“encoding=’utf-8’”。针对附件 2 表格中“留言主题”和“留言详情”两列内容进行提取，读取过程的代码如下：

```
data = xlrd.open_workbook('fujian2.xlsx')
file = open('fj2.txt', 'w', encoding='utf-8')
#print(data.sheet_names())
#['Sheet1']
#table = data.sheet_by_name('Sheet1')
table = data.sheet_by_index(0)###根据索引获取工作表的内容
#print(table.name,table.nrows,table.ncols)###获取行列数
##col_3 = table.col_values(2)
for i in range(1,table.nrows):
    for j in [2,4,5]:
        col = table.cell_value(i,j)
        col1 = col.strip()
        file.write(col1)
file.close()
```

其次用 jieba 库对所提取内容进行分词处理，构建**词袋模型**。jieba 分词可支持精确模式、全模式和搜索引擎模式，其中，搜索引擎模式是基于精确模式，对长词进行二次切分[4]，在对比三种模式的分词结果后，本文选择利用第一种模式精确分词。jieba 分词采用基于 Trie 树结构的算法，可以有效地实现词图扫描，进而得到句子中所有成词的可能，将每个词的出现次数转换为词频，根据词频可以得到最合理的分词情况[5]。jieba 分

词的部分代码如下：

```
txt = open('fj2.txt', 'r', encoding='utf-8').read()
words = jieba.lcut(txt,cut_all=False)
def text_save(filename, data):#filename为写入CSV文件的路径，data为要写入数据列表.
    file = open(filename,'w',encoding='utf-8')
    neirong = []
    for i in range(len(data)):
        s = str(data[i]).replace('[','').replace(']','')#去除[],这两行按数据不同，可以选择
        s = s.replace(',','').replace("'",'') .replace('"','')+ ' '    #去除单引号，逗号，每行末
        file.write(s)
        neirong.append(s)
    print(neirong)
    file.close()
    print('保存文件成功')
text_save('fj2SetDone.txt',words)
```

再将文本中无实际意义的词语进行删除，部分代码如下：

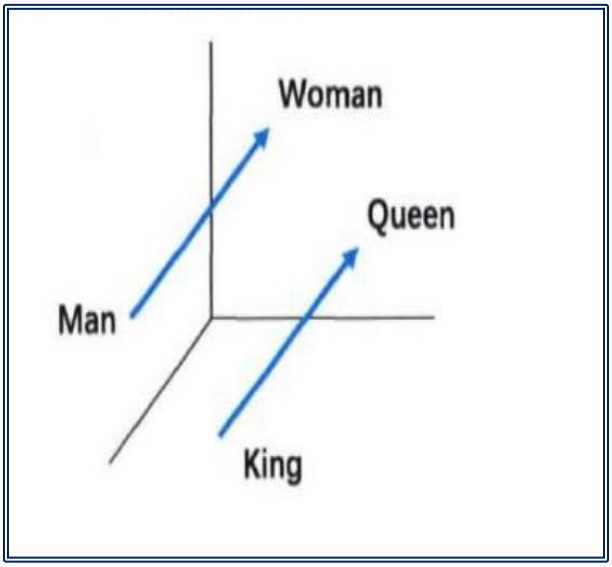
```
data = xlrd.open_workbook('fujian2.xlsx')
table = data.sheet_by_index(0)
f = open('fj2second.txt','w',encoding='utf-8')
for i in range(1,table.nrows):
    col = table.cell_value(i, 2)+ '\n'
    temp = re.sub('[a-zA-Z]','', col)
    tem = re.sub(r'\d', '', temp)
    te = re.sub(r'区','',tem)
    t = re.sub(r'市', '', tem)
    f.write(t)
f.close()
```

面对分词后庞大的数据库，我们需要根据语义将其分类。考虑到分类的准确性和科学性，这里引入**独热表示** (One-Hot Representation) 和**分布式表示** (Distributed Representation) 。其中独热表示又称独热编码，即将每个词用 0 和 1 构成的 N 维稀疏向量来进行表示，其中 N 为词汇表中单词的总数。在向量中，每个词都将与之对应的维度置 1，其余维度的值均为 0。此时任意两个词之间都是孤立的，光从这两个向量看不出两个词是否存在关系。而分布式表示[8]刚好解决了这个问题，它是一类将词的语义映射到向量空间中的自然语言处理技术，每一个词用特定的向量来表示，向量之间的距离一定程度上表征了词之间的语义关系，即两个词语义相近，在向量空间的位置也相近。分布式表示中典型的代表是词嵌入 (Word Embedding) ，通过神经网络或者矩阵分解等降

维技术，表征文本中单词的共现信息。我们采用这种分布式表示的方法来进行文本阅读模型中的单词嵌入问题，从而避免传统语言模型中的“维度灾难”和“词汇鸿沟”情况。

独热表示和分布式表示用图解如下：

Man	[1,0,0,0]
Woman	[0,1,0,0]
King	[0,0,1,0]
Queen	[0,0,0,1]



我们通过利用独热表示和分布式表示来建立 **gensim.word2vec 模型训练**，它是将词转换为词向量的形式建立的分类模型。

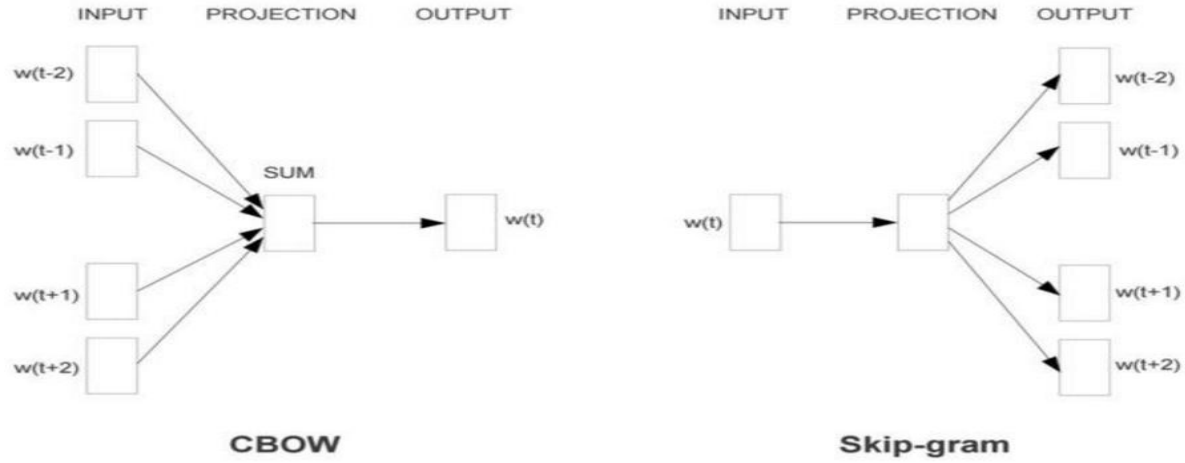
Word2Vec 实际是一种浅显的神经网络模型，它有两种网络结构，分别是 CBOW 和 Skip-gram。CBOW 的目标是根据上下文出现的词语来预测当前的生成概率，而 Skip-gram 的目标则是根据当前词来预测上下文中各词的生成概率。CBOW 和 Skip-gram 都可以表示成由输入层，隐含层和输出层组成的神经网络。

输入层中的每个词由独热编码方式表示，在隐含层中，K 个隐含单元的取值可以由 N 维输入向量以及连接输入和隐含单元之间的 NxK 维权重矩阵计算得到。在 CBOW 中，还需要将输入词所计算的隐含单元进行求和。输出层向量的值可以通过隐含层向量（K 维），以及连接隐含层与输出层之间的 KxN 维权重矩阵计算得到。输出层也是一个 N 维向量，每维与词汇表中的一个单词相对应。最后输出层向量应用 SoftMax 激活函数，可以计算出每个单词的生成概率。

接着训练神经网络的权重，使得语料库中所有单词的整体生成概率最大化。从输入层到隐含层需要一个维度为 NxK 的权重，从隐含层到输入层又需要一个维度为 KxN 的权重矩阵，学习权重可以用反向传播算法实现，每次迭代时将权重沿梯度更优的方向进行一小步的更新。但是由于 SoftMax 激活函数中存在归一化项的缘故，推导出来的迭代公式需要对词汇表中的所有单词进行遍历，使得每次迭代过程非常缓慢，由此产生了

Hierarchical Softmax 和 Negative Sampling 两种改进方法，训练得到的维度为 NxK 及 KxN 的两个权重矩阵之后，可以选择其中一个作为 N 个词的 K 维向量表示。

具体流程图如下：



在 gensim 中，word2vec 相关的参数都在包 gensim.models.word2vec 中。我们将附件 1 中一级标签分类的词语导入建立分类模型，检验句子和标签的复杂度，部分代码如下：

```
model = word2vec.Word2Vec.load("test_01.model")
zidian = {}
list2 = []
list1 = ['城乡建设',
    '党务政务',
    '国土资源',
    '环境保护',
    '纪检监察',
    '交通运输',
    '经济管理',
    '科技与信息产业',
    '民政',
    '农村农业',
    '商贸旅游',
    '卫生计生',
```

```
    '劳动和社会保障',]
words = xlrd.open_workbook('fujian2.xlsx')
table = words.sheet_by_index(0)
neirong = []
stopwords = set(open('stopword.txt', encoding='utf8').read().strip('\n').split('\n'))  # 读
file = open("fj2second.txt",'r',encoding='utf-8')
for line in file:
    if not line:
        break
    else:
        txt = jieba.lcut(line.strip())
        list2.append(txt)

for i in range(len(list2)):
    try:
```

为了提高分类器的性能，使模型更准确，我们采用了**降维处理**。由于文本数据的高维性和稀疏性，往往使得算法在训练和分类时间上开销很大，而过多的特征也会导致维数过剩，所以要采用特征降维的方法。其具体原理首先在本文分类的特征选择中，假设某个特征词 a 与一个类别 c 是相互独立的,再计算假设条件下的理论值，然后利用下式计算每个样本的观测值，和样本对应下的理论值的偏差。最后按偏差大小对特征词排序，选择 topn 作为特征。

具体公式采用：

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - E)^2}{E}$$

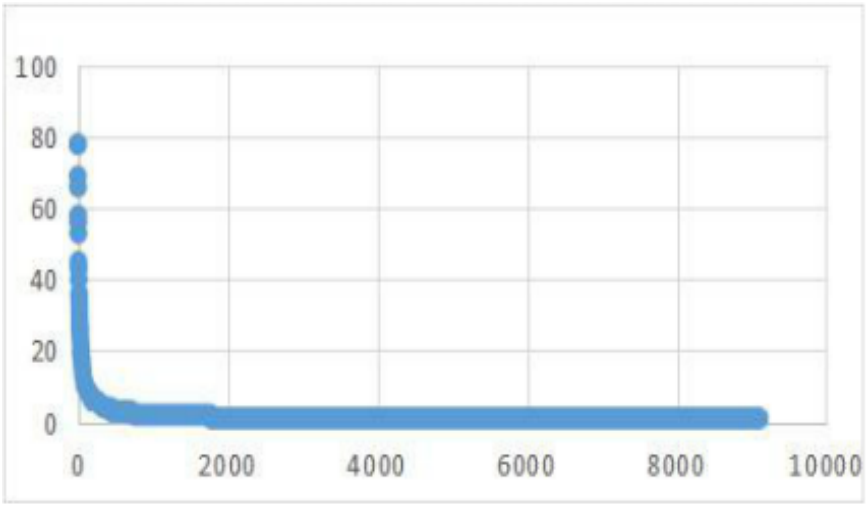
最终，我们完成了建立关于留言内容的一级标签分类模型，通过 F-Score 和准确率对模型进行检验，符合率为 0.93，符合理想状态。

5.2 问题二的模型建立与求解

文本数据挖掘是指从文本数据中抽取有价值的信息和知识的计算机处理技术，是抽取有用、有效、散布在文中的有价值信息，文本挖掘的方法有很多，其中文本分类和聚类是两种最基本的挖掘功能。文本挖掘是信息挖掘的一个研究分支，它涵盖了多种技术，包括数据挖掘技术、信息检索、机器学习、自然语言处理等。文本挖掘的关键在于词汇切分，但汉语语言的复杂性，使得中文词汇的切分成为一个很困难的问题[7]。

热度评价指标对筛选热点问题起决定性作用，那么如何制定热度评价指标呢？这里

采用热度加权的方式。我们利用**词频分析**，根据留言分词内容在文档中出现的次数和点赞数进行统计分析，赋予点赞数和分词重复数等不同的热度权重，构建了一个新的热度评价指标。文本部分词频数如下图所示：



在寻找热点问题的过程中，对文本进行聚类处理时，文本特征词的选取十分重要。我们希望获取的词语既能够保留原留言的信息，又能够反映词语的相对重要性，但如果保留所有词语，就可能会影响挖掘结果，因此对分词结果进行特征降维很重要。

进行**特征降维**有两种方法，一是特征选择（Feature Selection），即选取最合适的、具有代表性的维度；二是特征提取（Feature Extraction），主要是通过映射把原始特征变换为较少的新特征。本文使用特征提取的方法，根据词语在文本中的重要程度对其赋予一定的权重，提取部分权重较大的词语作为特征词，以此减少特征词的数量，同时达到降维的目的[8]。另外，我们引入 **TF-IDF 算法**，通过特征提取进行特征选择后，使用 TF-IDF 权重算法，进而完成文本表达。

具体过程如下：

要确定对“留言主题”和“留言详情”哪一列进行权重计算，需要对文本内容具体分析。以“建设”一词为例，“留言主题”中出现 163 次，其中标签为“城乡建设”的有 90 次，“留言详情”中出现 952 次，其中标签为“城乡建设”的有 440 次。从“留

言主题”中看，“建设”一词可以较好的代表“城乡建设”。

下图为其中两词出现次数：

分类		留言主题	留言详情
建设	所有标签	163	952
	城乡建设	90	440
设施	所有标签	35	386
	城乡建设	22	148

以留言 1783 为例，“留言主题”中出现“建设”1 次，“留言详情”中出现“建设”2 次。

TF-IDF（词频-逆文件频率）是一种用于资讯检索与资讯探勘的常用加权技术，它认为：字词的重要性随着他在文本中出现的次数成正比增加，但会随着它在所有文本中出现的频率成反比下降。词频（Term Frequency）指特征词在文本中出现的频率，

$$\mathrm{TF} = \frac{特征词出现的次数}{文本中所有词语数目} = \frac{m}{M}$$

但是一些通用的词语并不能表示主题内容，有时出现频率较少的词语更能表达留言的主题，因此不能只使用 TF。作为权重，指标需要具有强大的预测主题的能力，而逆文件频率（Inverse Document Frequency）表示所有留言文本中出现某一特征词的留言数，

$$\mathrm{IDF} = \log\left(\frac{文件总数}{包含特征词的文件数} + 0.01\right)$$

某一特征词在文件内的高频率出现，以及在整个文件集合中低文件频率，即 TF-IDF，可以过滤掉常见通用的词语，进而保留有代表性的的词语。

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF} * \mathrm{IDF}$$

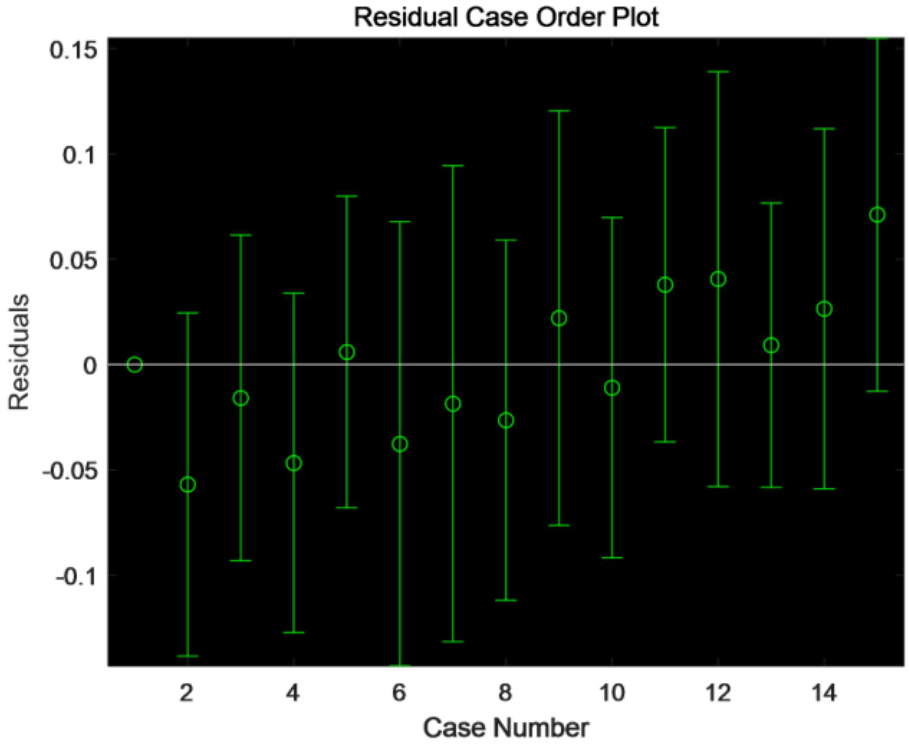
TF-IDF 是信息检索领域常用的方法，TF 反映文件的内部特征，IDF 反映文档间的特征[9]。

回归系数区间估计取值 bint							
-3.2726	-0.0013	-0.0027	-0.0060	-0.0132	-0.0004	-0.0005	-0.0004
3.7395	0.0026	0.0040	0.0008	0.0065	0.0003	0.0003	0.0004

通过计算回归系数区间的估计取值可知各变量系数 b=(i=1,2,3,4,5,6,7,8)

= 0.2335， 0.0006， 0.0007， -0.0026， -0.0033，-0.0000，-0.0001， 0

得到**残差图**如图所示：



此时用于检验回归模型的统计量 stats：

相关系数 r^2 = 0.9833 F= 0.8420，与 F 对应的概率 P=0.9523，

显著性水平 alpha= 0.01，

由此可知该模型可以有效筛选热点问题。

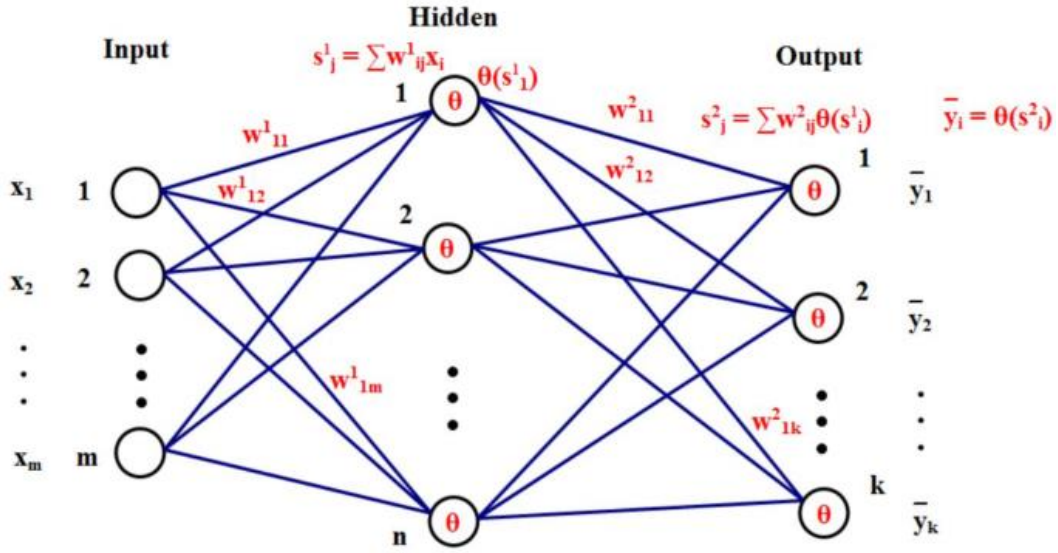
最终，整理出前五类热点问题，如下图所示：

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	2335	2019/1/17 至2019/3/1	A市58车贷案关注人群	承办A市58车贷案警官应跟进关注留言
2	2	2291	2019/9/11至2019/11/13	A市K9县居民	A市A5区汇金路五矿万境K9县存在一系列问题
3	3	1988	2019/8/6至2019/9/5	A4区居民	A4区绿地海外滩小区距长赣高铁距离过近
4	4	1854	2019/4/11至2019/5/24	梅溪湖金毛湾业主	反映A市金毛湾配套入学的问题
5	5	977	2019/3/26至2019/4/9	A市润和又一城、三润城、润和紫郡、润和长郡、润和美郡、润和星城、润和滨江府、明发国际城、星澜之悦的广大业主	关于A6区月亮岛路沿线架设110kv高压线杆的投诉

5.3 问题三的模型建立与求解

对于本题，我们将从答复意见的时间特征、初始结构特征、文本内容特征三个维度分析，就其答复内容的及时性、相关性、可解释性、完整性等指标建立评价模型。

在这里我们采用**神经网络**的分析方法，流程图如下：



该神经网络的原理为在输入层与输出层之间增加若干层(一层或多层)神经元，即隐单元，它们与外界没有直接的联系，但其状态的改变，则能影响输入与输出之间的关系，每一层设置了若干个节点。由此，我们得到了分析答复意见的方向。

无量纲化处理是综合评价步骤中的一个环节，通过无量纲化处理，单位不同的各特征之间才具有可比性，目前常见的无量纲化处理方法主要有极值化、标准化、均值化以及标准差化，其中最常用的方法是标准化。标准化方法值反映指标之间的相互影响，这种方法不适用于多指标的综合评价中，更适合数据样本量足够大的场景，因此本文使用z-score 标准化方法。

z-score 标准化方法的公式为

$$x' = \frac{x-u}{\sigma}$$

其中 x 为原始数据，u 为样本均值，σ为样本标准差，若$x\sim N(u,\sigma^2)$，

则，$y=\frac{x-u}{\sigma}\sim N(0,1)$

可以看出，z-score 标准化方法即将原始数据集标准化为接近标准正态分布的数据集，比较适合本题处理较大数据量[10]。

将所有数据转化为统一规格后，对照留言时间和答复时间给其制定指标来作为衡量

答复的及时性的时间特征；同时，我们将答复内容的文本长度、平均句长、标点符号的数目等等看做初始结构特征，结构化特征是指一个答案的直观体现或直接统计得出的特征。包括以下方面[11]：

1、文本长度（ans_length，数值型）。

用户对答案最直观的感受是文本长度，文本长度与信息量正相关。

2、平均句长（avg_stc_len，数值型）。

长句和短句在表达力方面存在显著差异。

3、标点符号占比（punc_ratio，数值型）。

如果标点符号占比过大，其所包含的有价值信息就相对不足。

4、格式是否良好（well_formed，二值变量）。

作为一种视觉语法，对在线阅读的体验有较大影响。

5、是否有参考文献（has_refer，二值变量）。

在线内容评估其可解释性。

关于文本内容特征，其包括答复内容的相关性、可解释性、完整性等指标，我们选择**支持向量机模型**和**组合分类模型**作为待构建的模型，采取**皮尔森相关系数法**、**最大互信息系数法**、**L1 正则化**、**L2 正则化**、**递归特征消除法**、**随机森林平均不纯度减少法**对构造的特征进行评分，计算其均值并作出分析。

皮尔森（pearson）相关系数是衡量线性关联性的程度，两个变量 X、Y 之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

由此可以得到总体相关系数，以此为例估算样本的协方差和标准差，便可以得到样本相关系数：

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

最大互信息系数（MIC）用于衡量两个变量的线性或非线性强度，MIC 度可以捕获变量间广泛的关系，不限于特定的函数关系。MIC 的计算首先对变量 X、Y 构成的散点图进行网格化，求出最大的互信息值，其次对最大的值进行归一化，最后选择不同度量下互信息的最大值作为 MIC 值：

$$\mathrm{MIC(D)} = \max_{XY<B(n)} M(D)_{X,Y} = \max_{XY<B(n)} \frac{I(D,X,Y)}{\log(\min(X,Y))}$$

L1 正则化和 L2 正则化可以看做是损失函数的“惩罚项”，即对损失函数中的某些参数的限制。在 Lasso 回归中，Python 中 Lasso 回归的损失函数为

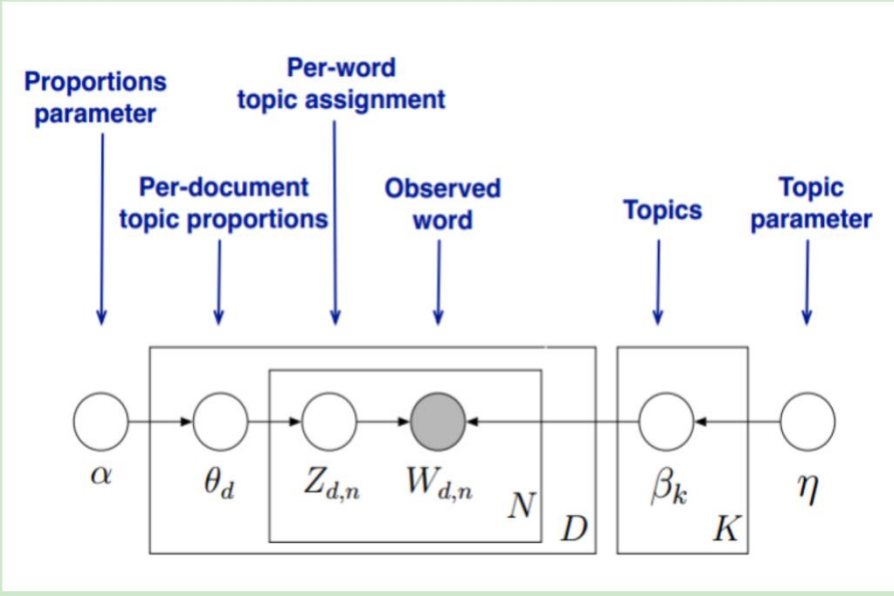
$$\min_{\omega} \frac{1}{2n_{samples}} \left|\left|X_\omega - y\right|\right|_2^2 + \alpha\left|\left|\omega\right|\right|_1$$

式中的 $\alpha\left|\left|\omega\right|\right|_1$即为 L1 正则化项，L1 正则化是指权值向量ω中各个元素的绝对值之和。L1 正则化有助于生成稀疏权值矩阵，可以用于特征选择。在 Ridge 回归中，Python 中 Ridge 回归的损失函数为

$$\min_{\omega} \left|\left|X_\omega - y\right|\right|_2^2 + \alpha\left|\left|\omega\right|\right|_1$$

式中的 $\alpha\left|\left|\omega\right|\right|_1$即为 L2 正则化项，L2 正则化是指权值向量ω中各个元素的平方和求平方根。

对于答复内容的相关性，要通过 **word2vec** 训练词向量，在词粒度上挖掘文本的语义信息，通过文本主词干及相关词集实现语义扩展，然后使用 **LDA 挖掘模型**[12]。实现建立主题相关性的特征。下图是 LDA 模型的具体实现流程图：



最后采用算法 **SVM**、**LR** 及**决策树 C4.5** 构建答复质量评价分类器。做到将问题量化，通过计算指标来建立评价模型。在基础模型中，逻辑回归（LR）、支持向量机（SVM）和随机森林（C4.5）的分类性能接近。随着特征体系的丰富，3 种模型的性能均有所提高，尤其是随机森林作为组合分类模型的优势越来越明显，与另外两种模型形成了较大差距。如果考虑我们前面所说的三个维度，则有了 7 个组合方式。

多种特征组合模型性能综合评价如下图所示：

特征组合	模型	准确率	灵敏度	特异性	精确率	召回率	F1 值	AUC
1	LR	0.831	0.798	0.86	0.837	0.798	0.817	0.831
	SVM	0.828	0.789	0.864	0.845	0.789	0.816	0.829
	RF	0.825	0.819	0.829	0.787	0.819	0.802	0.817
2	LR	0.767	0.705	0.838	0.833	0.705	0.764	0.773
	SVM	0.726	0.673	0.782	0.766	0.673	0.717	0.73
	RF	0.821	0.825	0.818	0.768	0.825	0.795	0.816
3	LR	0.752	0.952	0.694	0.475	0.952	0.634	0.727
	SVM	0.715	0.948	0.661	0.391	0.948	0.553	0.686
	RF	0.897	0.891	0.902	0.881	0.891	0.886	0.895
1 +2	LR	0.835	0.804	0.862	0.839	0.804	0.821	0.835
	SVM	0.829	0.814	0.842	0.807	0.814	0.811	0.828
	RF	0.853	0.853	0.853	0.815	0.853	0.834	0.854
1 +3	LR	0.856	0.823	0.887	0.87	0.823	0.846	0.857
	SVM	0.847	0.82	0.871	0.848	0.82	0.834	0.847
	RF	0.919	0.917	0.92	0.902	0.917	0.91	0.917
2 +3	LR	0.846	0.846	0.846	0.806	0.846	0.825	0.842
	SVM	0.797	0.859	0.763	0.658	0.859	0.745	0.784
	RF	0.916	0.913	0.918	0.899	0.913	0.906	0.914
1 +2 +3	LR	0.861	0.83	0.888	0.87	0.83	0.85	0.862
	SVM	0.847	0.82	0.87	0.847	0.82	0.833	0.847
	RF	0.921	0.92	0.922	0.904	0.92	0.912	0.921

最终得下方特征得分图：

维度	特征	Corr.	MIC	Lasso	Ridge	RFE	MDI	均值
结构化特征	*ans_length*	0.27	0.91	0	0	0.33	1	0.41
	avg_stc_len	0.1	0.32	0.01	0	0.53	0.03	0.17
	punc_ratio	0.1	0.42	0	0.37	1	0.05	0.32
	word_ratio	0	0.47	0	0.04	1	0.04	0.26
	re_edit	1	0.56	1	0.34	1	0.03	0.66
	well_formed	0.48	0.36	0	0.1	0.8	0	0.29
	has_refer	0.01	0	0	0.07	0.87	0	0.16
	has_img	0.42	0.3	0	0.16	1	0	0.31
	img_num	0.05	0.27	0.01	0	0.2	0.02	0.09
	has_href	0.29	0.22	0	0.08	0.93	0	0.25
	href_num	0.03	0.24	0	0	0.23	0	0.08
文本特征	*sentiment*	0.05	0.04	0	0	0.67	0	0.13
	entropy	0.57	0.9	0	1	1	0.04	0.59
	similarity	0.04	0.11	0	0.03	0.73	0.03	0.16
社交属性	*thanks_num*	0.14	0.98	0	0	0.2	0.04	0.23
	vote_num	0.12	1	0	0	0	0.42	0.26
	follower_num	0.1	0.86	0	0	0.07	0.03	0.18
	view_num	0.11	0.83	0	0	0.13	0.03	0.18
	ans_num	0	0.34	0	0	0.47	0.12	0.16
	edit_num	0.01	0.3	0	0	0.27	0.03	0.11

六、模型的误差分析

6.1 误差原因

1、由于数据是现实搜集数据，所得的某些数据必然会存在较大误差。

2、模型的建立基于假设，所筛选的数据不准确也会对结果造成影响。

3、若模型建立不合适，结果也会有较大误差。

6.2 误差分析

1、在问题一中，采用分布式表示的方法，但在汉语中有些词义不呈线性。

2、在一些计算过程中，忽略了小数，可能会造成误差。

3、测量数据本身可能存在误差。

七、模型评价与推广

7.1 模型评价方式

对于问题一、二：

本文主要是针对文本数据进行一个分类，我们选择**准确率**和 **F-Score** 两种指标对分类模型进行评价检验，下表说明了评价指标的相关参数。

真实情况	预测留言所属类别	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

其中，

1.TP（ True Positive）：实际为正例且被划分为正例的留言数；

2.FP（False Positive）：实际为反例但被划分为正例的留言数；

3.FN（False Negative）：实际为正例但被划分为反例的留言数；

4.TN（ True Negative）：实际为反例且被划分为反例的留言数。

准确率（Accuracy）是指对于给定的测试数据集，分类器正确分类的样本数与被分类总数之比，因此准确率是对全部数据的判断，包括正例和反例。根据准确率，我们的

确可以得到一个分类器是否有效，通常来说，准确率越高，分类器越好，但它不总是能有效的评价一个分类器，例如在样本不平衡的情况下，得到高准确率就没有意义。因此，我们不能只用准确率评价我们的分类模型。

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

精确率（Precision）和召回率（Recall）是信息检索领域两个最基本的指标。精确率也称为查准率，是针对预测结果而言的，即正确预测为正例的占全部预测为正例的比例。但精确率和准确率是完全不同的概念，精确率代表的是对正例结果中的预测准确程度。

$$\text{Precision} = \frac{TP}{TP + FP}$$

召回率也称为查全率，是针对原样本而言的，即正确预测为正例的占所有实际为正例的比例。

$$\text{Recall} = \frac{TP}{TP + FN}$$

精确率和召回率相互影响，理想状况下希望两者都高，但实际上两者是此消彼长的：当精确率高时，召回率就低；当召回率高时，精确率就低，因此需要在实际情况中合理判断。

很多时候选择用参数控制精确率和召回率，通过修改参数能够得出一条与两个变量相关的曲线（POC），曲线与 x、y 轴围成图形的面积成为 AUC，AUC 可以综合的衡量预测模型的好坏，但由于计算过程复杂，便提出 F-Score 作为综合指标评价分类。定义：

$$\frac{2}{F_1} = \frac{1}{Precision} + \frac{1}{Recall}$$

可得

$$F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}$$

对于问题三：

基础模型是指仅包含了结构化特征的模型，由于结构化特征是客观反映答案特点的特征，可以直接抓取得到或通过简单计算得到，将此模型作为基准。交叉验证是一种模型检验方法，用于评价模型的泛化能力。采用 **10 折交叉验证**对答案质量评价模型进行训练。使用这种方法，我们将数据集随机分成 10 份，使用其中 9 份进行训练而将另外

1 份用作测试。该过程可以重复 10 次，每次使用的测试数据不同，最终会得到混淆矩阵。

例如 SVM 基础模型的混淆矩阵：

频数值		实际值	
		正例	负例
预测值	正例	2 268	416
	负例	607	2 643

7.2 模型的优点

将所有合适的数据分别分析，全面考虑到了各种因素；基数大，误差可以尽量减小；建立模型采用公式和图像相结合的方式，使模型更加智能、科学。

7.3 模型的缺点

1、神经网络在求解的时候按照已知类别样本计算,但是对于未知类别样本应用判别函数时不做任何监督。

2、当达到此极限时，随训练能力的提高，预测能力反而下降，即出现所谓“过拟合”现象。

参考文献

[1] 陈国柱.浅谈我国电子政务存在的问题与对策[J].科技信息，2013:76-77.

[2] 徐博龙. 应用 Jieba 和 Wordcloud 库的词云设计与优化[J].福建电脑,2019:25-28.

[3] 何南思,杨霁.《Python 高级编程》课程信息化教学设计初探——以“Jieba 库应用”课程单元为例[J].信息与电脑(理论版),2019(09):239-240+243.

[4] 黎曦.基于网络爬虫的论坛数据分析系统的设计与实现[D].华中科技大学，2018.

[5] 邢彪.基于 jieba 分词搜索与 SSM 框架的电子商城购物系统[J].信息与电脑，2018:104-108.

[6] 孙飞,郭嘉丰,兰艳艳,徐君,程学旗.分布式单词表示综述[J].计算机学报,2019,42(07):1605-1625.

[7] 胥桂仙，许建潮，连远锋，李昱翠.文本挖掘中的特征表示及聚类方法[J].吉林工学院学报，2002，23(3):12-15.

[8] 叶雪梅，毛雪岷，夏锦春，王波.文本分类 TF-IDF 算法的改进研究[J].计算机工程与应用，2019，55(2):104-109.

[9] 覃世安，李法运.文本分类中 TF-IDF 方法的改进研究[J].现代图书情报技术，2013，238(10):27-30.

[10] 张志辉. 论文影响力的线性学科标准化方法研究[D].上海交通大学,2015.

[11] 王伟,冀宇强,王洪伟,郑丽娟.中文问答社区答案质量的评价研究:以知乎为例[J].图书情报工作,2017,61(22):36-44.

[12] 胡鹏辉. 基于多模型的问答社区答案质量评价研究[D].南京师范大学,2019.