

## 第八届“泰迪杯”全国数据挖掘挑战赛论文报告

所选题目：C 题“智慧政务”中的文本挖掘应用

综合评定成绩：\_\_\_\_\_

评委评语：

评委签名：

# “智慧政务”中的文本挖掘应用

## 摘 要

在信息时代，对于大量的数据仅靠人工处理是机器繁琐与低效的，所以解决相应问题，我们首要考虑的是利用程序和机器学习来进行深层次的数据挖掘。利用相关程序和机器学习我们能高效的完成筛选关键信息、进行关键信息的识别、和根据关键信息的自动分类等建立目标体系所需的必要步骤，节省大量的时间与人力成本。

**针对问题一：**将附件 2 的数据进行预处理，用 jieba 库对中文进行分词，使用 TF-IDF 算法计算每个关键词的权重向量，将数据集的 70%作为训练数据集用于构建词袋，即训练模型，使用朴素贝叶斯分类方法，将另外数据集的 30%作为测试数据集代入模型，得出结果并用 F-Score 评价方法对分类的结果进行评价。

**针对问题二：**我们的热度评价指标由两部分组成，一部分是对于一个问题引起的关注度（即点赞与反对总数）的高低，另一部分是同类型的留言归类后，类型中留言数量的多少。首先每行数据列为单独的 txt 文本，然后运用 c++程序提取分好的时间段的所有数据集合里出现的高频词语整理出高频话题，再结合关注度选出最热话题。按照问题时间要求分出 5 个表格（问题），而后用 excel 将数据划分到对应的最热话题。

**针对问题三：**根据附件四中相关部门对留言的答复意见，提取答复意见的质量优劣性的特征，研究通过五个角度对回复信息进行量化描述，分别是回复内容是否具有相关性，回复内容是否具有可解释性，回复内容是否充实，回复内容是否具有时效性，回复内容套用模板，给出答复意见的一套评价方案。

**关键词：**朴素贝叶斯分类   TF-IDF 算法   机器学习   C++   回复质量评价

# Text mining application in "smart government"

## Summary :

In the information age, it is tedious and inefficient for a large number of data to be processed only manually. Therefore, to solve the corresponding problems, we should first consider using program and machine learning to carry out deep data mining. By using related programs and machine learning, we can efficiently complete the necessary steps of screening key information, identifying key information, and automatically classifying key information to establish the target system, saving a lot of time and labor costs.

For question 1: preprocess the data in Annex 2, use the Jieba database to segment Chinese words, use TF-IDF algorithm to calculate the weight vector of each keyword, use 70% of the data set as the training data set to build the word bag, that is, use naive Bayesian classification method, and substitute 30% of the other data set as the test data set into the model, get the results and conduct F-score Evaluate the result of classification.

For problem 2: our heat evaluation index consists of two parts: one is the level of attention (i.e. the total number of likes and objections) caused by one problem, and the other is the number of messages in the same type after classification. First, each row of data is listed as a separate TXT text. Then, C + + program is used to extract the high-frequency words in all the data sets of the divided time period to sort out the high-frequency topics, and then combined with the degree of attention to select the hottest topics. According to the problem time requirements, five tables (problems) are divided, and then the data are divided into the corresponding hottest topics by Excel.

For problem 3: ccording to the reply opinions of the relevant departments in Annex 4 to the message, extract the characteristics of the quality of the reply opinions, and study the quantitative description of the reply information from five perspectives, namely, whether the reply content is relevant, whether the reply content is interpretable, whether the reply content is substantial, whether the reply content is timely, and whether the reply content is applied Template, give a set of evaluation scheme of reply.

**Key words:** Naive Bayes classification   TF-IDF algorithm   machine learning

C + +   reply quality evaluation

# 目录

1.挖掘目的.....	5
2.分析方法与过程.....	5
2.1 总流程图.....	5
2.2 问题细明分析.....	6
2.2.1 问题一流程图与分析.....	6
2.2.2 问题二流程图与分析.....	6
2.2.3 问题三流程图与分析.....	7
3.数据预处理.....	7
3.1 机器学习数据处理.....	7
3.2 数据筛选.....	8
3.3 数据统计.....	8
4.模型建设与优化.....	10
4.1 一级分类模型建立.....	10
4.1.1 一级分类标签选取.....	10
4.1.2 TF-IDF 算法计算权值 .....	10
4.1.3 朴素贝叶斯分类.....	11
4.1.4 模型求解.....	12
4.2 热点问题 .....	14
4.2.1 高频词汇统计.....	14
4.2.2 热点问题与高频词选取.....	15
4.2.3 热门问题数量占比统计.....	16
4.2.4 热门问题关注度占比统计.....	17
4.2.5 综合热点问题统计.....	17
4.2.6 留言归类.....	18
4.3 答复意见评价.....	19
4.3.1 答复意见质量的优劣性特征.....	19
4.3.2 回复质量的描述方法.....	19
4.3.2.1 无效问题处理.....	20
4.3.2.2 回复信息质量的量化方法.....	20
4.3.3 回复质量评价模型.....	21
4.3.4 实验结果分析.....	21
4.3.4.1 回复质量评价模型计算步骤.....	21
5.结果分析.....	22
5.1.一级分类标签模型评估.....	22
5.2 答复意见的评价分析.....	23
6.模型评价.....	24
6.1 模型的优点.....	24
6.2 模型的缺点.....	24
6.3 模型的优化.....	24
7.参考文献.....	24

# 1.挖掘目的

1. 在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。所以我们建立关于留言内容的一级标签分类模型。用 F-Score 对分类方法进行评价
2. 热点问题挖掘，某一时段内群众集中反映的某一问题可称为热点问题。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。我们将某一时段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，并给出评价结果。按要求的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表。
3. 给出答复意见评价，针对附件 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。

# 2.分析方法与过程

## 2.1 总流程图

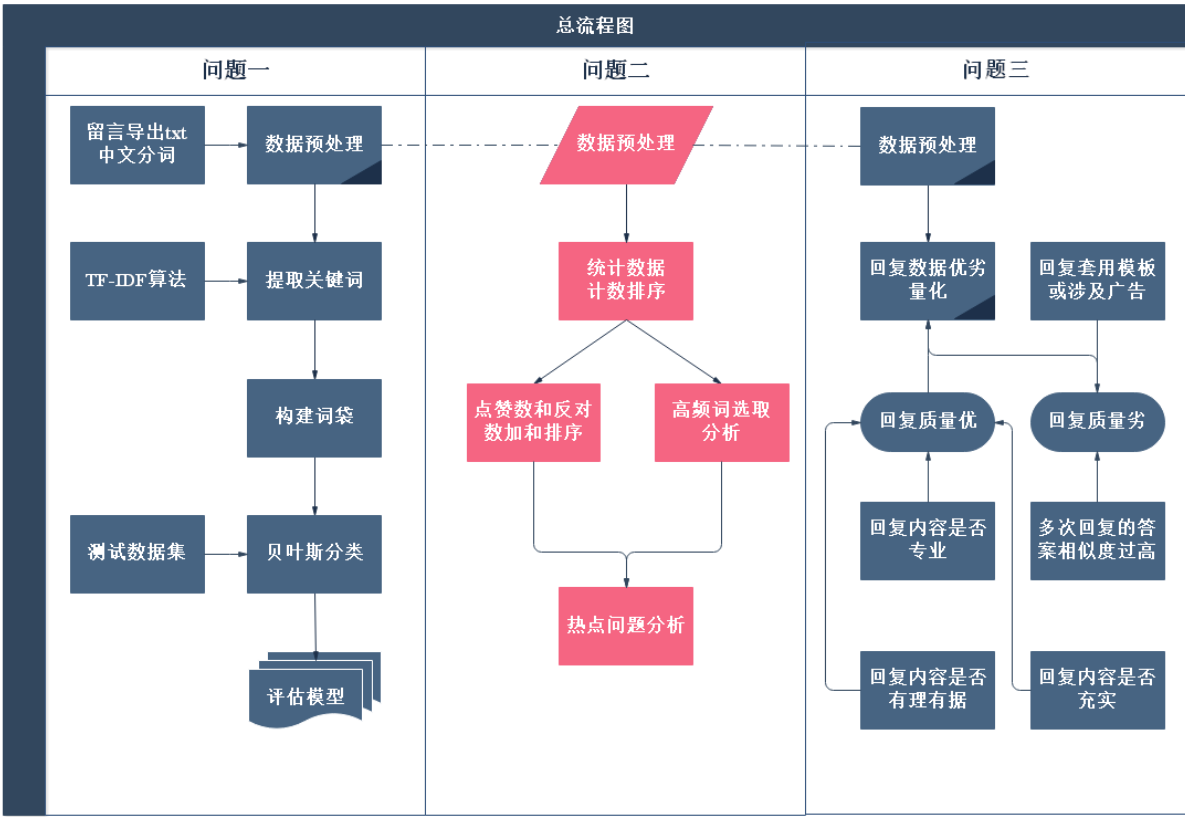


图 1 总流程图

2.2 问题明细分析

2.2.1 问题一流程图与分析

我们根据大量数据与人民网标准建立一级标签分类模型，首先要考虑一级标签下有多少种分类结果，通过“留言”进行一级标签分类，我们需要先识别其中的关键词，并评估留言中字词对于一个文件集或语料库中的一份文件的重要程度，即为频数多的字词分配权重，字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。总结便是一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能代表这篇文章。

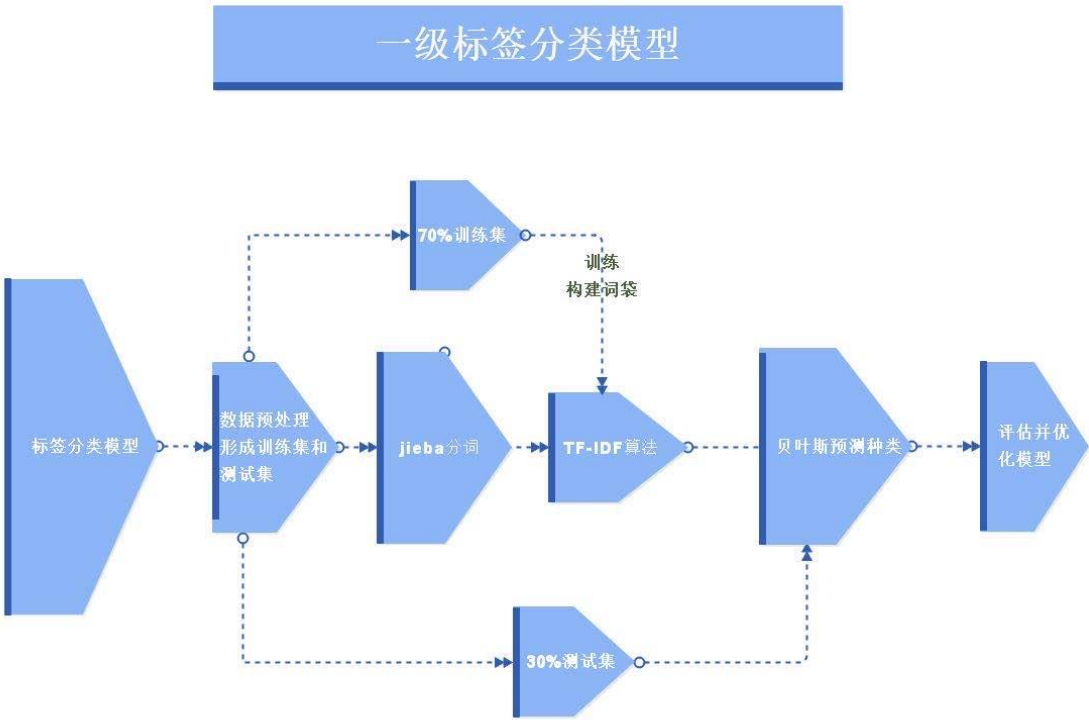


图 2 问题一流程图

2.2.2 问题二流程图与分析

问题二的关键在于精细筛选与处理每条留言，然后运用编程筛选出高频词再结合关注度综合评选出热点问题，再将所有留言数据依次归类。

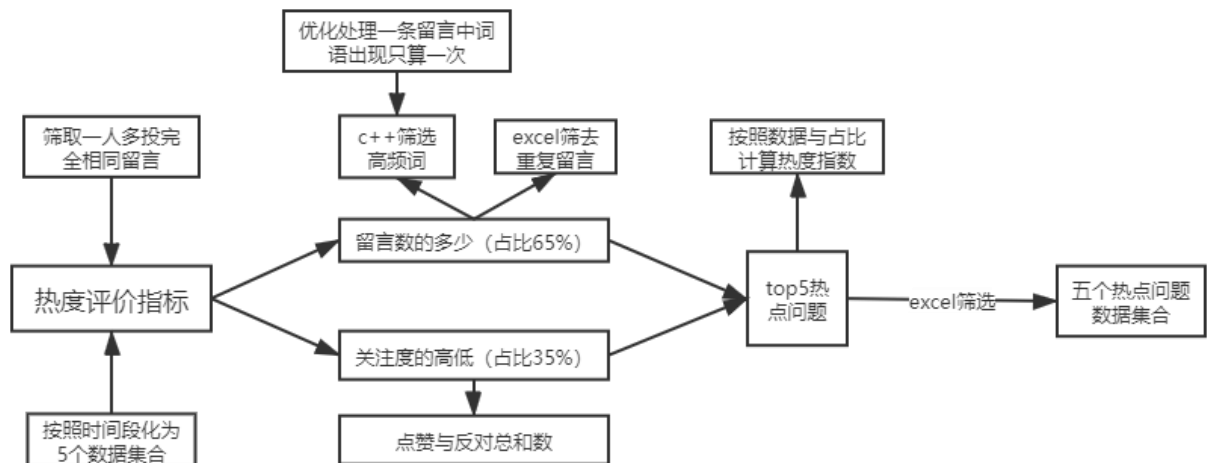


图 3 问题二流程图

### 2.2.3 问题三流程图与分析

通过研究五个角度对回复信息进行量化描述,分别是回复内容是否具有相关性,回复内容是否具有可解释性,回复内容是否充实,回复内容是否具有时效性,回复内容套用模板,给出答复意见的一套评价方案。

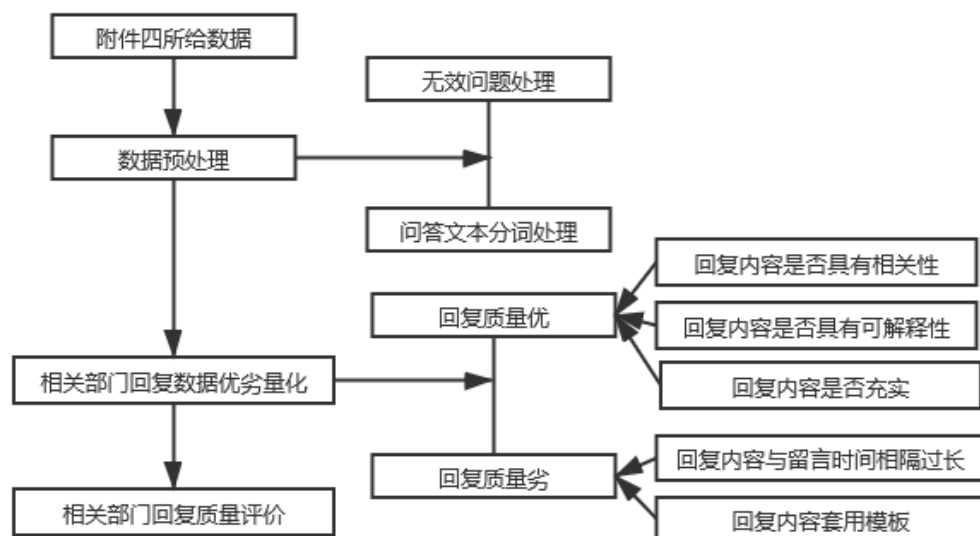


图 4 问题三流程图

### 3.数据预处理

### 3.1 机器学习数据处理

在建立一级标签分类模型之前，我们需要先对附件 2 的数据进行预处理，只保留留言主题和留言内容，使用自然语言 Python 中的 jieba 库和停顿词对留言主题和留言内容中的文本进行切割词语，从而可以为计算各词语权值的计算做准

备。

本题采取机器学习的思想，为了使模型达到最好的效果，同时为了避免出现过拟合和欠拟合的现象。通常，我们可通过测试数据来对模型的泛化误差进行评估并进而做出选择，为此需使用一个测试集来测试模型对新数据的判别能力，同时需要保证测试集与训练集之间尽可能的互斥，测试样本尽量不在训练集中出现过，未在训练过程中使用过。

因此，采取等比例随机抽取的方式，保证各类训练集和测试集所占总体的比例一致，将附件 2 所给的数据集，一分为二，将 70%的数据集作为训练集（原始数据），将另外 30%的数据集作为测试集用来对模型的泛化误差进行评估并进而做出选择。

### 3.2 数据筛选

首先直接用 excel 自带的“数据”“删去重复项”删去在留言详情里面完全相同的内容（可能为一人多次重复留言），并且在提取高频词时，为避免一个人留言中多次重复提及到一个词，我们运用 c++ 程序处理每项留言：使用 `map<string,bool>` 关联容器，将字符串和布尔值关联在一起，出现一次字符串时，使它的布尔值（标记值）设为 1，当再次遇到该字符串的时候，先判断它的布尔值是否是默认值（0），否说明已经出现，那么就不进行计数

由于留言具有时效性，所以结合文本量情况与实际情况我们将附录 3 中的数据分为六个时间段数据，2017 年与 2018 年数据过少合并为一，2019 有全年大量数据于是将每三个月数据依次按照时间顺序划为 1 份数据，共 4 份，2020 年数据适中，独立成为一份数据。

对于程序语言在进行留言处理时只需要“留言主题”与“留言内容”，所以在将 excel 每一行生成一份独立的 txt 文本时只需保留所需的两行与留言编号。

### 3.3 数据统计

在根据附录二设计一级标签分类模型时，我们先统计附录二中相关一级标签的种类：

表 1 一级标签数量统计表

一级标签	数量值
城乡建设	2009
环境保护	938
交通运输	613
教育文体	1589
劳动和社会保障	1969
商贸旅游	1215
卫生计生	877
(空白)	
总计	9210



所以根据附录 2 所建立的一级分类标签则是在原来一级标签分类中进行改善与调整。

在根据附录 3 划分热点问题时，在筛选完数据按时间划分成六份数据后，先根据 excel 升序排列统计出每个时间段的关注度前五的话题。

表 2 2017-2018 年高关注度留言统计

留言编号	点赞反对总和
360114	9
360113	3
286572	3

表 3 2020 年高关注度留言统计

留言编号	点赞反对总和
244098	7
281722	3
275804	2
214944	2
192140	1
262339	1
283973	1

表 4 2019 年 1-3 月份高关注度留言统计

留言编号	点赞反对总和
220711	821
217032	790
194343	733
284571	80
200667	78
262052	78
281898	60

表 5 2019 年 4-6 月份高度关注度留言统计

留言编号	点赞与反对总和
223297	1767
193091	242
272089	57
267630	42
257376	39
193286	32
248495	23

表 6 2019 年 7-9 月份高关注度留言统计

留言编号	点赞反对总和
208636	2097
263672	669
226723	66
203187	63
231690	27
232063	22
252226	21

表 7 2019 年 10-12 月份高关注度留言统计

留言编号	点赞反对总和
239595	44
205217	33
283631	31
253369	29
280425	29
285552	23
262258	22

## 4. 模型建设与优化

### 4.1 一级分类模型建立

#### 4.1.1 一级分类标签选取

首先根据网站提示，我们选取中国人民网的留言分类进行一级标签的确定。其次由于我们最终将附件二的数据进行归类，所以与附件二预处理的数据进行比较，我们删去附件二中没有的金融、文娱、政务选项，将劳动与社会保障划分为企业类与就业类，得到我们最终的一级分类标签。

表 8 一级分类标签表

一级分类标签							
城规	教育	医疗	环保	企业	就业	旅游	交通

#### 4.1.2 TF-IDF 算法计算权值

通过初步分析之后，基于 TF-IDF 权重策略（权重策略文档中的高频词应具有表征此文档较高的权重），建立 TF-IDF 模型并进行单词的权值计算，这里涉

及到以下两个公式：

关键词词频（TF）：指一篇文档中关键词出现的频率：

$$TF_w = \frac{\text{某个词}w\text{在文章中的出现次数}}{\text{文章的总词数}}$$

逆向文本概率（IDF）：用于权衡关键词权重的指数：

$$IDF = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条}w\text{的文档数} + 1}\right), \text{分母} + 1 \text{ 是为了避免分母为 } 0$$

得到词语上述的两个值之后求 TF-IDF 值

$$TF - IDF = TF * IDF$$

TF-IDF 值越大，则这个词成为一个关键词的概率就越大。也就更容易通过这个词进行区别分类。

#### 4.1.3 朴素贝叶斯分类

通过 TF-IDF 算法计算关键词的权值向量，在进行贝叶斯分类之前需要用到以下的公式计算概率。

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y|X_1, X_2, \dots, X_N) = \frac{P((X_1, X_2, \dots, X_N)|Y)P(Y)}{P((X_1, X_2, \dots, X_N))}$$

$$P((X_1, X_2, \dots, X_N)|Y) = P(X_1|Y)P(X_2|Y) \dots P(X_N|Y)$$

$$p(\text{类别}|\text{特征}) = \frac{p(\text{类别})p(\text{特征}|\text{类别})}{p(\text{特征})}$$

将朴素贝叶斯算法应用于本题一级标签分类问题，一名其中一条留言为例：

X = 内容，如“一、二年级每班只有一位代课老师兼班主任全日制，负责教孩子们的语文，数学”

Y = 类别，如“教育”、“教育以外的某个类别”

那么给定一个留言内容“一、二年级每班只有一位代课老师兼班主任全日制，负责教孩子们的语文，数学”，就可以根据贝叶斯公式计算出这个留言属于教育类的概率。

$P(\text{教育}|\text{一、二年级每班只有一位代课老师兼班主任全日制, 负责教孩子们的语文, 数学})$

$$= \frac{P(\text{"一、二年级每班只有一位代课老师兼班主任全日制, 负责教孩子们的语文, 数学"}|\text{"教育"})P(\text{"教育"})}{P(\text{"一、二年级每班只有一位代课老师兼班主任全日制, 负责教孩子们的语文, 数学"}|\text{"教育"})}$$

根据 jieba 库分割成词,

$$P(\text{"一、二年级每班只有一位代课老师兼班主任全日制, 负责教孩子们的语文, 数学"}|\text{"教育"}) \\ = P(\text{"一"、"二"、"年级"、"每班"、"只有"、"一位"、"代课老师"、"兼"、"班主任"、"全日制"、$$

$$\text{"负责"、"教"、"孩子"、"们"、"的"、"语文"、"数学"}|\text{"教育"})$$

根据条件独立性假设, 只要分别求出各词语在教育标签类下出现的概率, 即可判断这个留言是否属于教育类。

将思想衍生至本题, 同理当我们有样本 (包含特征和类别) 的时候, 我们非常容易通过  $p(x)p(y|x) = p(y)p(x|y)$  统计得到  $p(\text{特征}|\text{类别})$ , 即  $p(\text{特征})p(\text{类别}|\text{特征}) = p(\text{类别})p(\text{特征}|\text{类别})$ , 有

$$p(\text{类别}|\text{特征}) = \frac{p(\text{类别})p(\text{特征}|\text{类别})}{p(\text{特征})}。$$

通过独立假设, 对每个类别计算一个概率  $p(c_i)$ , 然后再计算所有特征的条件概率  $p(f_j|c_i)$ , 那么分类的时候, 我们就是依据贝叶斯找一个最可能的类别:

$$p(class_i|f_0, f_1, \dots, f_n) = \frac{p(class_i)}{p(f_0, f_1, \dots, f_n)} \prod_j^n p(f_j|c_i)$$

#### 4.1.4 模型求解

划分训练集和测试集, 使用附录中的 (分割数据集.py) 将附件 2 的数据集分割, 保证每个类都运行一次程序, 可以保证无论是训练集还是测试集中的各类数据的占比一致。数据集 (9210 个) 分后训练集 (6452 个) 和测试集 (2758 个)。运行附录中的 (贝叶斯算法.py) 得到以下截图的结果, 首先先将训练集代入到模型之中, 计算词语的权值向量, 后使用测试集进行一个结果预测, 而后 2758 个样本之中, 预测失败的有 565 个, 失败率约为 20.5%。将编号为 143145 的留言分割成各个关键词后:

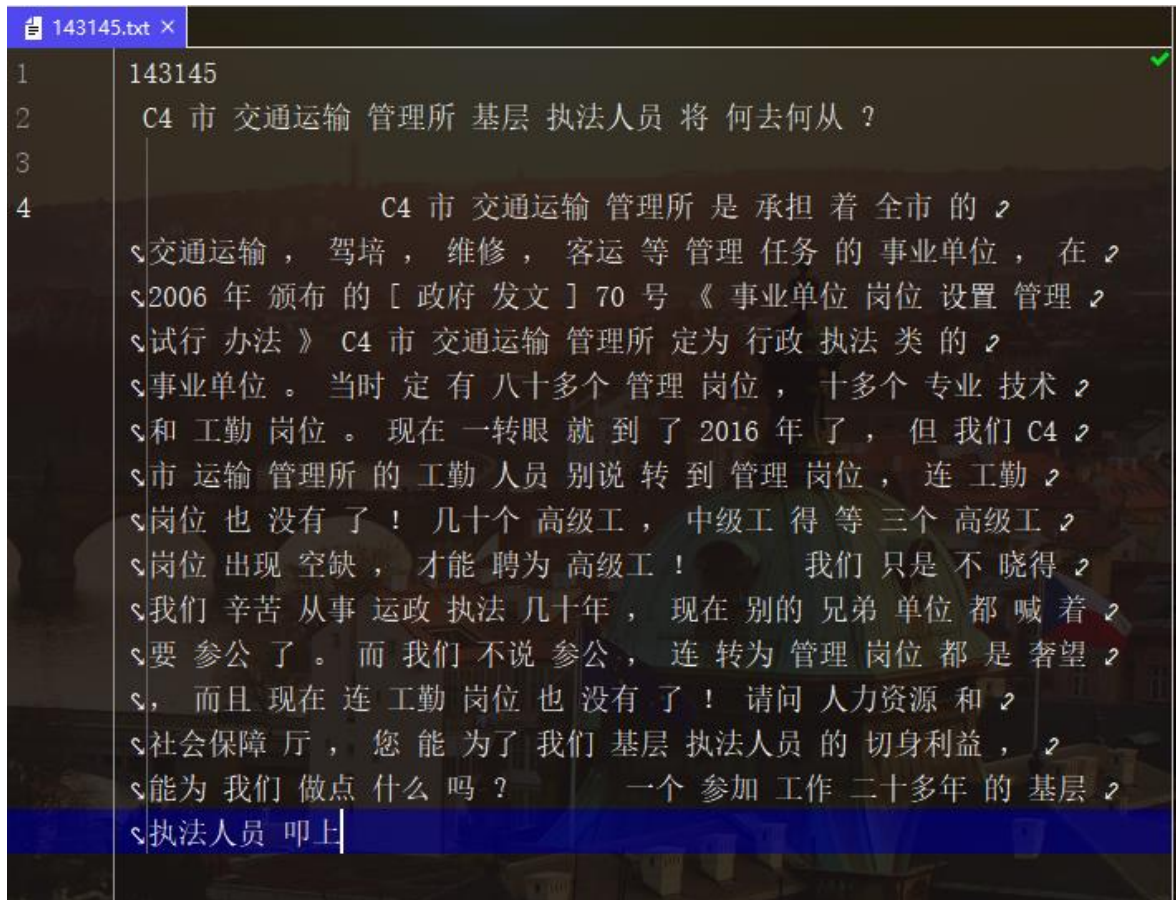


图 5txt 中文分词结果



图 6 测试数据集运行结果

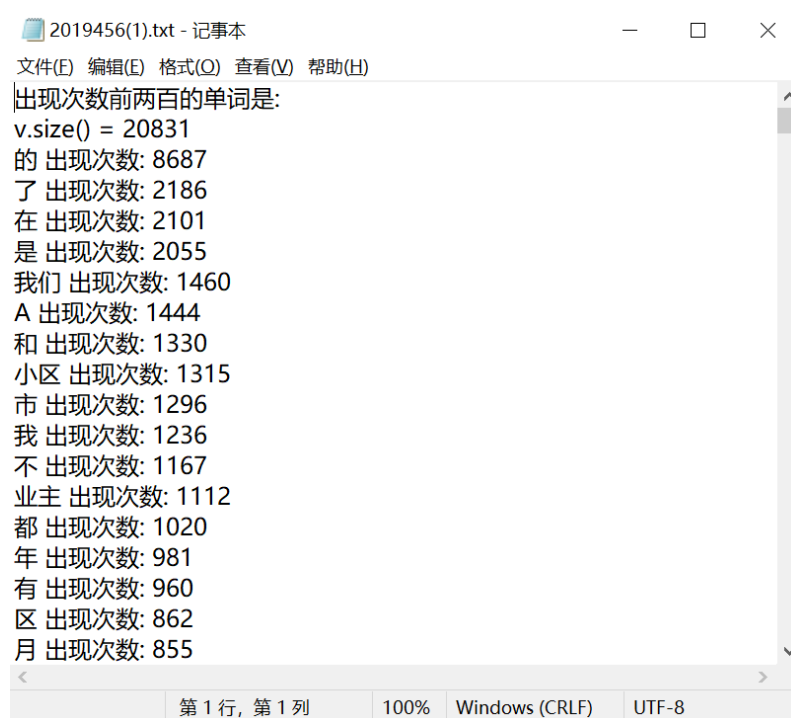
现实生活中，在市民们现根据大类自行选择了主题之后，根据以上机器学习的程序运用，我们将其中错误的分类进行整理筛选然后重新归类，确保留言与一级标签完全高精度吻合。（所有文件分类见附录）

## 4.2 热点问题

### 4.2.1 高频词汇统计

在数据处理的时候我们已经将附件 3 所有的数据按照时间序列分为了 6 份数据，每一份数据我们都运用 c++ 程序提取留言内容里统计出的高频词语。其中，程序实现是使用 c++ 语言自带的 stl 容器，循环遍历每个已只提取内容主题与内容详情的 txt 文件并读取里面的词语（字符串），通过 map 关联容器进行计数，最后对容器按词语出现次数从大到小排序（c++ 程序与所有结果见附件）。

以 2019 年 4-6 月运行程序结果为例：



```
2019456(1).txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
出现次数前两百的单词是:
v.size() = 20831
的 出现次数: 8687
了 出现次数: 2186
在 出现次数: 2101
是 出现次数: 2055
我们 出现次数: 1460
A 出现次数: 1444
和 出现次数: 1330
小区 出现次数: 1315
市 出现次数: 1296
我 出现次数: 1236
不 出现次数: 1167
业主 出现次数: 1112
都 出现次数: 1020
年 出现次数: 981
有 出现次数: 960
区 出现次数: 862
月 出现次数: 855
第 1 行, 第 1 列 100% Windows (CRLF) UTF-8
```

图 7 提取 2019 年 4-6 月高频词结果图

其中我们筛去如“的”“A”等此类无用词语，提取有用信息如“小区”“业主”等词语，将六份数据中所有高频词进行整合。

年份	高频词与出现次数							
2017、2018年	高频词	学院	溪湖	实习	公司	实习	强制	学校
	出现次数	7	7	6	4	4	4	4
	高频词	CBD	购房	学生	补贴/企业	寝室/老师		
2019年1-3月	出现次数	4	4	3	3	2		
	高频词	业主	小区	开发商	领导	政府	物业	公司
	出现次数	1062	940	515	472	444	404	392
2019年4-6月	高频词	西地省	社区	影响	建设	装修	购房	商品房
	出现次数	290	280	276	264	256	215	189
	高频词	小区	业主	居民	严重	公司	开发商	物业
2019年7-9月	出现次数	1315	1112	491	489	460	428	357
	高频词	西地省	学校	施工	小学/幼儿园	医院	周边	噪音
	出现次数	314	279	261	255	200	180	165
2019年10-12月	高频词	小区	业主	开发商	公司	建设	物业	学校
	出现次数	1146	1047	414	408	351	336	266
	高频词	施工	医院	西地省	幼儿园	道路	车辆	拆迁
2020年	出现次数	262	249	242	233	225	174	169
	高频词	业主	小区	开发商	公司	物业	学校	居民
	出现次数	908	800	422	340	335	326	320
2020年	高频词	社区	施工	噪音	中学	车辆	公交	小学
	出现次数	306	303	170	168	155	152	143
	高频词	小区	号线	公司	地铁	物业	施工	投诉
	出现次数	61	42	29	26	22	20	19
2020年	高频词	噪音	车位	南站	办证	社保		
	出现次数	18	17	15	13	11		

图 8 六时间段高频词汇汇总

#### 4.2.2 热点问题与高频词选取

在上统计的高频词中我们选取词性相近的作为一个独立的热点问题，而组成他的高频词则作为其关键词。根据高频词出现次数高低我们选取“社区小区物业与生活问题”、“开发商与相应公司问题”、“学生学校问题”、“施工建设噪音影响问题”、“医院与就医问题”五个主题词依次编号 1-5。

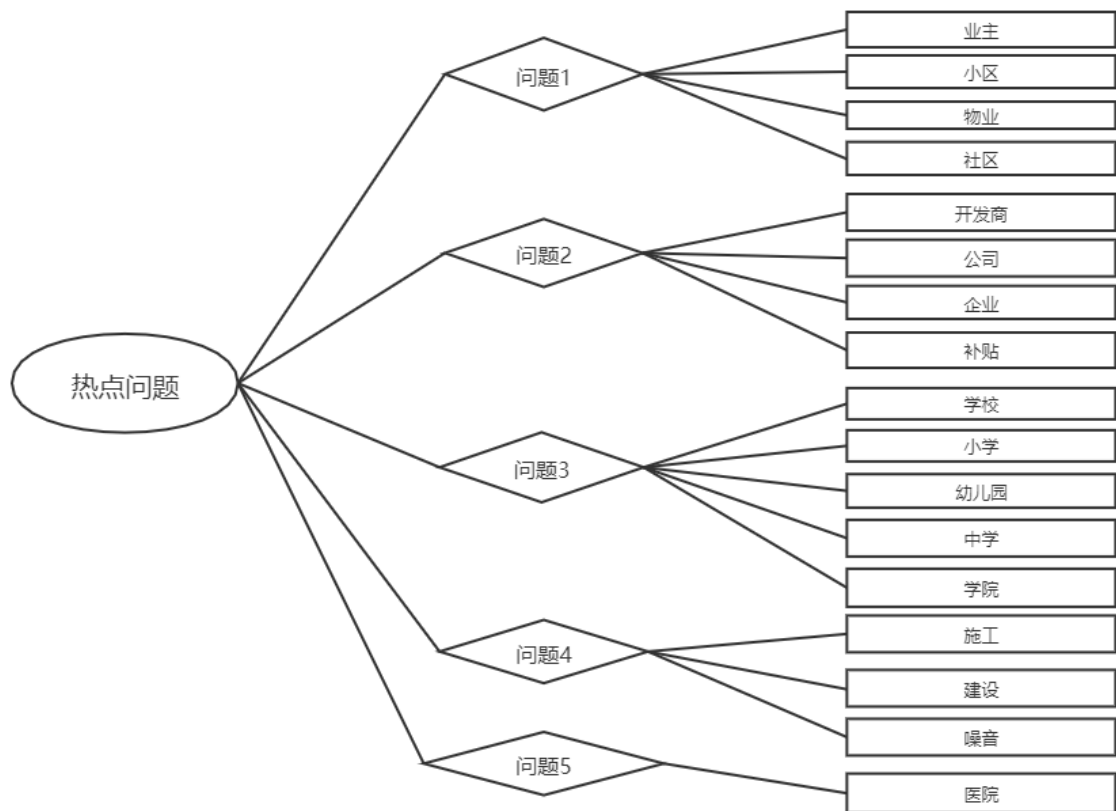


图 9 每类热点问题关键字划分图

#### 4.2.3 热门问题数量占比统计

首先，我们计算热门问题数量占比，用每个主题词统计出来的关键词数量总和来代替其热门问题出现的次数进行占比计算。

$$P_i = \frac{n_i}{n}$$

其中  $p_i$  代表主题词占比， $n_i$  对应编号主题中所属关键词出现的最高次数：

$$\left\{ \begin{array}{l} n_1(\text{社区小区物业与生活问题}) = \text{业主} + \text{小区} + \text{物业} + \text{社区} = 10431 \\ n_2(\text{开发商与相应公司问题}) = \text{开发商} + \text{公司} + \text{企业} + \text{补贴} = 3415 \\ n_3(\text{学生学校问题}) = \text{学校} + \text{幼儿园} + \text{小学} + \text{中学} + \text{学生} + \text{学院} = 1936 \\ n_4(\text{施工建设噪音影响问题}) = \text{施工} + \text{建设} + \text{噪音} = 1814 \\ n_5(\text{医院与就医问题}) = \text{医院} = 449 \end{array} \right.$$

$$n = n_1 + n_2 + n_3 + n_4 + n_5 = 18045$$

所以计算  $p_i$  可得：

$$p_{n1} = 0.5781 \quad p_{n2} = 0.1892 \quad p_{n3} = 0.1073 \quad p_{n4} = 0.1005 \quad p_{n5} = 0.0249$$



4.2.4 热门问题关注度占比统计

在计算关注度时，我们将附件三所有点赞数与反对数相加在进行数据排序得到关注度最高的留言：

	A	B	C	D	E	F	G	H
1	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	总和
2	208636	A00077171	K9县存在一系列问题	2019/8/19 11:34:04	狗咬人，请问有人对	0	2097	2097
3	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	纳入配套入学，A3I	5	1762	1767
4	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	案情消息总是失望，	0	821	821
5	217032	A00056543	A市58车贷特大集资诈骗案保	2019/2/25 9:58:37	小股东、苏纳弟弟苏	0	790	790
6	194343	A000106161	A市58车贷案警官应跟进关注	2019/3/1 22:12:30	圣侦并没有跟进市领	0	733	733
7	263672	A00041448	小区距长赣高铁最近只有30米	2019/9/5 13:06:55	复到我如下问题：1、	0	669	669
8	193091	A00097965	富绿物业丽发新城强行断业主	2019/6/19 23:28:27	提供地摊上买的收据	0	242	242
9	284571	A00074795	省尽快外迁京港澳高速城区	2019/1/10 15:01:26	、长浏高速出口，过	0	80	80

图 10 附件三留言关注度统计图

其中我们发现，留言 263672 与留言 193091、留言 19309 与留言 284571 都出现了明显的断层，所以基于数据数量考虑我们选取前六条留言进行统计。

留言 208636 ( $m_1$ )：留言主题“A 市 A5 区汇金路五矿万境 K9 县存在一系列问题”为我们找寻到了热点事件发生的地点，找到相对应时间段 2019 年 8 月，和前五个热点话题相似度低，于是我们将  $m_1$  新划为新的  $n_6$  类。

留言 223297 ( $m_2$ )：关于“A 市区金毛湾入学问题”无论是主题内容还是时间段都与与高频词中的“学生学校问题”主题相匹配，所以我们将其归为  $n_3$  类。

留言 220711 ( $m_3$ )：“请书记关注 A 市 A4 区 58 车贷案”的主题与时间线  $n_2$  相符合，所以将  $m_2$  归为  $n_2$  类。

留言 217032 ( $m_4$ )：“严惩 A 市 58 车贷特大集资诈骗案保护伞”与  $m_3$  同一事件。

留言 193091 ( $m_5$ )：“A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？”将其归为  $n_4$ ，时间为 2019 年 4 月-2019 年 9 月和 2020 年。

留言 194343 ( $m_6$ )：“A 市富绿物业丽发新城强行断业主家水”并找到相对应主题，归为  $n_1$  类。时间为 2019 年与 2020 年。

与  $Pn_i$  相同公式计算：

$$p_{m1}=0.3050 \quad p_{m2}=0.2569 \quad p_{m3}=0.2344 \quad p_{m5}=0.1066 \quad p_{m6}=0.0971$$

4.2.5 综合热点问题统计

我们设计的热度评价指标由两个部分组成，一部分是对于一个问题引起的关注度（即点赞与反对总数数）的高低，另一部分是同类型的留言归类后，类型中留言数量的多少。在现实生活中一般人们有急需解决的问题会主动自己留言，再去查看已存在的热点留言，所以两部分中我们决定“留言数量”占比 65%，“关注度”占比 35%。

$$\begin{aligned} p_1 &= 65\%p_{n1} + 35\%p_{m6} = 0.4097 \\ p_2 &= 65\%p_{n2} + 35\%p_{m3} = 0.2051 \\ p_3 &= 65\%p_{n3} + 35\%p_{m2} = 0.1597 \\ p_4 &= 65\%p_{n4} + 35\%p_{m5} = 0.1027 \\ p_5 &= 65\%p_{n5} + 35\%0 = 0.0162 \\ p_6 &= 65\%0 + 35\%p_{m1} = 0.1067 \end{aligned}$$

于是我们按照  $p_i$  大小进行归类得到我们数据“热点问题表.xls”

表 9 热点问题表

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	0.4097	2019 与 2020 年	A 市/小区业主	社区小区业主与物业有关的生活问题
2	2	0.2051	2017、2018、2019 与 2020 年	A 市/就业人员与企业公司	开发商与相应公司问题
3	3	0.1597	2017、2018、2019 年 4 月-2019 年 12 月	A 市金毛湾区/学生	学生关于入学和在学校的诸多相关问题
4	4	0.1067	2019 年 8 月	A 市 A5 区	是 A 市 A5 区汇金路五矿万境 K9 县...
5	5	0.1027	2019 年 4 月-12 月、2020 年	A4 区/居民	施工建设有关噪影响、环境污染等问题

#### 4.2.6 留言归类

我们将图 8 筛选对应的高频词作为关键词语，用数据预处理时处理好的无重复数据的 excel 进行筛选；第一轮筛选，先在附录 3 中依据时间段筛选出五个 excel 表格。第二次筛选在 excel 中新建一行关键词列加上我们的关键字，筛选时在每个表格里依次按照关键字进行交集化筛选（有关键字即保留），成功筛选完成后再将每个 excel 添加上其所属的问题 ID 最后合并即可。（筛选完成的全部 excel 表格见附录）

	A	B	C	D	E	F	G	H	I	J	K	L
1	问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数				
2	1	304503	A011244	什么时候能实行独生子女护理	2019-02-26 15:22:05	多，照顾四个老人真的	0	0				
3	1	313964	A108906	驾驶证期满换证，一个星期	2019-04-26 15:28:42	都快一个星期了都没	0	0				
4	1	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	必须去学校安排的几	0	0				
5	1	316619	A235259	问A市什么时候能普及5G网络	2019-05-14 11:22:13	为城区建成。看消息	0	0				
6	1	319659	A023956	市江山帝景新房有严重安全隐	2019-05-30 17:34:02	天气后过道全部是水	0	0				
7	1	321736	A9992521	市能不能提高医疗门诊报销范	2019-06-12 08:23:01	小孩体弱多病各种	1	0				
8	1	323034	A012414	务收费标准应考虑居民的经济	2019-06-19 17:46:24	费清运费按不超过30	0	0				
9	1	323149	A1241141	K3县乡村医生发卫生室执业许	2019-06-20 20:38:47	是证件下来啊。有些	0	0				
10	1	212323	A00020702	团要求员工购房时必须同时赠	2019-07-11 00:00:00	时也必须一对一购买	0	0				
11	1	360107	A0283523	力之城小区一楼的夜宵摊严重	2019-07-21 10:29:36	维护社会和谐稳定，	3	0				
12	1	360101	A324156	劳动东路魅力之城小区油烟	2019-07-28 12:49:18	没有。每天油烟直排	4	0				
13	1	213584	A909172	伊景园滨河苑定向限价商品房	2019-07-28 13:09:08	，无视法律法规，	0	0				
14	1	224767	A909176	苑车位捆绑销售！广铁集团	2019-07-30 14:20:08	，说什么预购不用！	0	0				
15	1	188801	A909180	河苑针对广铁职工购房的霸	2019-08-01 00:00:00	接着一个，首先未	0	0				
16	1	360108	A0283523	小区一楼的夜宵摊严重污染附	2019-08-01 16:20:02	维护社会和谐稳定，	6	0				
17	1	285897	A909191	伊景园滨河苑违法捆绑销售车	2019-08-01 20:06:52	团就要求捆绑购买车	0	0				
18	1	251601	A909187	A市伊景园滨河苑诈骗钱财	2019-08-01 22:42:21	的认购金后不与购房	0	0				
19	1	209506	A909179	新城坑客户购房金额并且捆绑	2019-08-02 16:36:23	同的那种，工作人员	0	0				

图 11 热点问题 1 留言明细表示例图

4.3 答复意见评价

根据附件四中相关部门对留言的答复意见，提取答复意见的质量优劣性的特征，研究如何通过不同角度对回复信息进行量化描述，给出答复意见的一套评价方案。

4.3.1 答复意见质量的优劣性特征

如何识别出最佳回复,如何衡量相关部门回复意见的质量，能切实解决公民问题是网络问政平台发展需要解决的重要问题，我们将相关部门在线回复质量优劣的特征进行归纳，得到以下优劣评判特征。

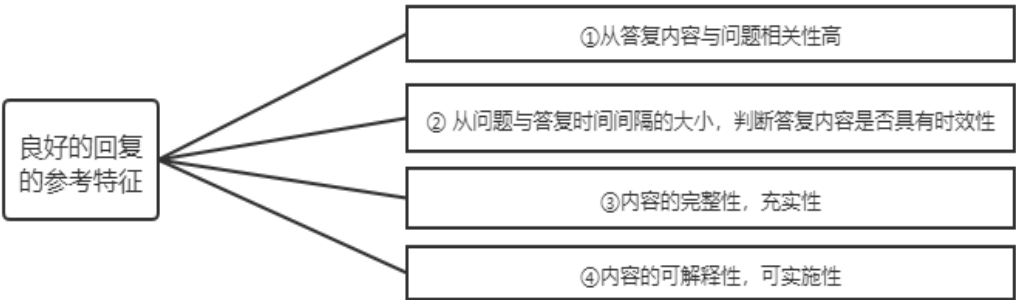


图 12 良好的回复的参考特征图

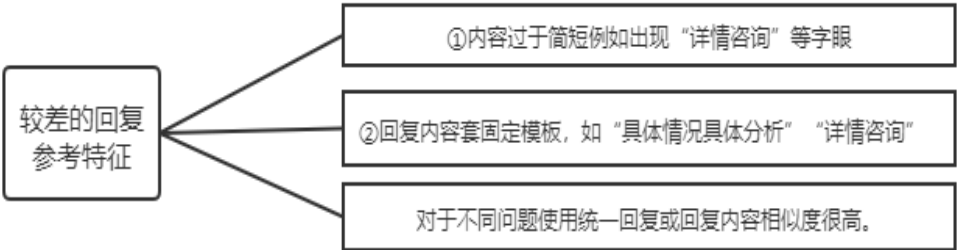


图 13 较差的回复参考特征图

在附件 4 给予的海量数据之下，如何基于这些特征利用人工智能技术对相关部门留言的答复意见进行评价，仍然是一个极具有研究价值的问题，下面从不同角度对相关部门对留言的答复意见的质量进行量化描述，从而建立评价模型。

4.3.2 回复质量的描述方法

相关部门对民众留言的答复意见的质量有很多种因素决定，必须从不同角度对其进行描述，下面，首先预处理附件 4 中的回复数据，将有问题的回答数据进行剔除，然后从五个方面对回复质量进行量化描述。

#### 4.3.2.1 无效问题处理

在回答系统的数据库的问题和回复文本中，有些留言描述不清，造成相关部门无法做出有效回复，这类数据称为无效数据，在数据处理中，需要将无效数据剔除评价系统，多数情况下，无效文本长度较短，且与一般问题的文本长度有明显差距。

在数据中，问题与回复是成对出现的，问题记为 Q，答案记为 A，|Q|为问题文本长度，取正整数  $k_0$  作为阈值，当  $|Q| \leq k_0$  时，我们就认为该条问题的数据为坏数据，不计入回复质量的评价范围内。数据处理后，对回复信息满足良好的记正分，反之，则记负分，得到评价得分记为 F。

#### 4.3.2.2 回复信息质量的量化方法

##### (1) 回复内容是否具有相关性

根据回复内容和问题中词汇的相似度来评定回复内容是否与问题具有相关性。利用 word2vec 可以计算出回复内容和问题中词汇的相似度：

$$\text{WORDSIM}(w_i, w_{pi}) = \frac{\sum_{i=1}^n (x_{i1} * x_{i2})}{\sqrt{\sum_{i=1}^n x_{i1}^2} * \sqrt{\sum_{i=1}^n x_{i2}^2}}$$

其中，两个词语  $w_i$  和  $w_{pi}$  的词向量表示为：  $w_i = (x_{11}, x_{21}, x_{31}, \dots, x_{i1} \dots x_{n1})$ ，  $w_{pi} =$

$(x_{11}, x_{21}, x_{31}, \dots, x_{i1} \dots x_{n1})$ ；  $n$  表示用 word2vec 训练词向量时设定的词向量的维度。

当  $w_i = w_{pi}$  时，可以通过系数  $1 + \lambda'$  增加相似度。由此，我们建立如下回复内容的专业程度的评价项  $F_1$ ：

$$F_1 = \begin{cases} \text{WORDSIM}(w_i, w_{pi}) & w_i \neq w_{pi} \\ (1 + \lambda') \text{WORDSIM}(w_i, w_{pi}) & w_i = w_{pi} \end{cases}$$

##### (2) 回复内容是否具有可解释性

相关部门的回复内容需要有可解释性与一定的可实施性，假设剔除部分无效数据后剩余文本数有  $N_0$  条比较规范的回答数据，统计回复信息数目  $N_1$ ，计算出出现频率，即  $N_1$  与  $N_0$  的比值，记为  $F_2$ 。

$$F_2 = \frac{N_1}{N_0}$$

##### (3) 回复内容是否充实

信息回复内容的详细程度与文本长度有直接关系，简短内容信息回复量一般不够，评分应该比较低，同时，过长的文本回复的评分不应该过高，因此使用对数函数来量化回复文本长度与评分分析，建立“回复内容是否充实”的评价项  $F_3$ 。

$$F_3 = \frac{1}{N_0} \sum_{i=1}^{N_0} \log_m L_i$$

其中  $L_i$  为针对第  $i$  问题回复的文本长度， $m$  为常数。

##### (4) 回复内容是否具有时效性

根据问题的答复时间  $T_d$  与留言时间  $T_l$  的差值，来对回复内容的时效性进行评分，差值小的评分高，差值大的评分低。建立“回复内容是否具有时效性”的

评价项  $F_4$

$$F_4 = \frac{k}{N_0} \sum_{i=1}^{N_0} (T_d - T_L)$$

其中  $k$  为常数。

#### (5) 回复内容套用模板

在回复内容中，有些则采用固定词语，如“请来电咨询”“网友：您好！留言已收悉”等，建立一个较差的回复的关键词组成集合  $T_{key}$ ，以出现频率作为评价项  $F_5$ 。

$$F_5 = \frac{N_{T_{key}}}{N_0}$$

### 4.3.3 回复质量评价模型

对上面的五项指标进行整合，计算出回复信息的得分情况，建立评价函数  $F$ ，

$$\begin{cases} F = M \cdot \lambda^T \\ \lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5), M = (F_1, F_2, F_3, F_4, F_5) \end{cases}$$

其中， $\lambda^T$  为  $\lambda$  向量的转置向量，向量  $\lambda$  和  $M$  为反应回复信息不同侧面的权重向量和得分向量。

### 4.3.4 实验结果分析

#### 4.3.4.1 回复质量评价模型计算步骤

(1) 剔除无效数据，由于考虑到在实际生活中，大多数问题很难用很简短的语句描述清楚，导致工作人员在回复的时候不能很好的理解问题所在，导致问题回复质量差。设  $Q = \{\langle aA \rangle\}$  为全部问答的集合，剔除以损害数据后的答集， $Q' = \{\langle Q, A \rangle | \langle Q, A \rangle \in Q, |Q| \geq 20\}$ ，即问题中的字的个数要多于 20。筛选后还剩 2805 条数据。

(2) 使用 word2vec 计算可得，留言与回复中的相关度  $F_1=0.8679$ 。

(3) 判断回复内容是否具有可解释性，一般会出现“根据.....相关规定”“法律法规”等专业词语，通过 Excel 筛选，在 2805 条有效数据中，有 1096 条出现了相关词汇。 $F_2=0.3907$

(4) 判断回复内容是否充实，通过 Excel 中的 LENB-LEN 对回复的单元格的字数进行计算并升序排序，发现大约字数在小于 60 的情况下，问题回复较差，没有实用性，使用公式  $F_3 = \frac{1}{N_0} \sum_{i=1}^{N_0} \log_m L_i$ ，计算可得  $F_3 = 0.3557$ 。

(5) 判断回复时间是否及时，具有时效性，我们通过 Excel 将回复时间与留言时间做差，求和等计算得到  $F_4=0.9126$ 。

(6) 在回复内容中，有些则采用固定词语，如“请来电咨询”“网友：您好！留言已收悉”等，用 Excel 筛选出 353 项此类回复， $F_5=0.9126$ 。

(7) 将各项得分向量表示为  $M = (0.8679, 0.3907, 0.3557, 0.9126, 0.1258)$ ，根据实际情况分配权重向量  $(0.30+0.15+0.15+0.25+0.15=1)$  可得回复质量评价函数值  $F=0.58161$ 。

所以综上我们可得评分为：

表 10 回复评价结果表

F <sub>1</sub>	F <sub>2</sub>	F <sub>3</sub>	F <sub>4</sub>	F <sub>5</sub>	F
0.8679	0.3907	0.3557	0.9126	0.9126	0.58161

5.结果分析

5.1.一级分类标签模型评估

根据测试集代入模型所得的结果，可得出下列表格，分别针对每个类别来说，将预测类别的组合划分为真正例（true positive）、假正例（false positive）、真反例（true negative）、假反例（false negative）四种情形。即下表 X.X 分类结果混淆矩阵

表 11 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假正例）
反例	FP（假正例）	TN（真正例）

因为本题的类别大于 3，因此将分类结果混淆矩阵扩大成八阶矩阵（见下表 X.X）。

查准率 P 与查全率 R 的定义：

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN}$$

表 12 分类结果混淆矩阵

真实情况	预测结果							
	城规	环保	交通	教育	就业	旅游	企业	医疗
城规	536	14	6	10	16	15	4	1
环保	14	266	0	0	0	1	0	0
交通	49	1	109	3	5	14	1	1

教育	13	2	1	438	9	5	7	1
就业	8	4	2	18	225	2	68	9
旅游	41	4	12	16	1	286	2	2
企业	5	0	0	23	111	1	105	9
医疗	7	7	1	4	13	12	4	215

$$\begin{aligned}
P_{\text{城规}} &= \frac{536}{673} \times 100\% \approx 79.6\%, R_{\text{城规}} = \frac{536}{602} \times 100\% \approx 89.0\%, F_{\text{城规}} \approx 84.1\% \\
P_{\text{环保}} &= \frac{266}{298} \times 100\% \approx 89.3\%, R_{\text{环保}} = \frac{266}{281} \times 100\% \approx 94.7\%, F_{\text{环保}} \approx 91.9\% \\
P_{\text{交通}} &= \frac{109}{131} \times 100\% \approx 83.2\%, R_{\text{交通}} = \frac{109}{183} \times 100\% \approx 59.6\%, F_{\text{交通}} \approx 69.4\% \\
P_{\text{教育}} &= \frac{438}{512} \times 100\% \approx 85.5\%, R_{\text{教育}} = \frac{438}{476} \times 100\% \approx 92.0\%, F_{\text{教育}} \approx 88.7\% \\
P_{\text{就业}} &= \frac{225}{380} \times 100\% \approx 59.2\%, R_{\text{就业}} = \frac{225}{336} \times 100\% \approx 67.0\%, F_{\text{就业}} \approx 62.8\% \\
P_{\text{旅游}} &= \frac{286}{336} \times 100\% \approx 85.1\%, R_{\text{旅游}} = \frac{286}{364} \times 100\% \approx 78.0\%, F_{\text{旅游}} \approx 81.4\% \\
P_{\text{企业}} &= \frac{105}{191} \times 100\% \approx 55.0\%, R_{\text{企业}} = \frac{105}{254} \times 100\% \approx 41.3\%, F_{\text{企业}} \approx 47.1\% \\
P_{\text{医疗}} &= \frac{215}{238} \times 100\% \approx 90.3\%, R_{\text{医疗}} = \frac{215}{263} \times 100\% \approx 81.7\%, F_{\text{医疗}} \approx 85.8\%
\end{aligned}$$

根据矩阵可计算出各类的查准率 P 和查全率 R，最后再根据 F-Score 评价方法的公式对模型进行评价

$$F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_i R_i}{P_i + R_i} = 0.764543 \times 100\% \approx 76.5\%$$

F1 的取值范围在 [0, 1] 之间，越高越好，本题模型评价所得结果为 0.764543，估测结果错误率为 20.5% 左右，由于企业和就业标签是劳动与社会保障分出来的，故两者会使模型的错误率增大，而交通和城规两者之间也具有一定的相似性，在样本数量足够大的情况下，可认为该模型可以很好地解决一级标签分类的问题。

## 5.2 答复意见的评价分析

算数据可以看出，留言与回复的相关度够好，除去个别情况，回复时间相对较及时，在回复内容上，很多都没有解决留言者的实际问题，只是套用模板，回复内容的专业度较低。



## 6.模型评价

### 6.1 模型的优点

- ①朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。
- ②对小规模的数据表现很好，能够处理多分类问题，对缺失数据不敏感，算法比较简单。

### 6.2 模型的缺点

- ①朴素贝叶斯分类，具有一定的狭隘性，由于贝叶斯算法是统计各个词语的词频，所以朴素贝叶斯失去了词语之间的顺序信息，在某些时候，词语顺序的不同，可能会导致分类结果出现错误。
- ②当样本数量太少或从未出现某个特征性词语时，不满足大数定律，计算的概率会失真。

### 6.3 模型的优化

当数据过小时，使用平滑技术处理太小的数据集，某些词汇在训练集中并未出现，会导致 $P(\text{"词语"}w|S) = 0$ ，于是在计算独立条件假设概率时，不同词语的概率相乘会使得整个乘积的结果为0，贝叶斯分类模型就达不到好的效果，于是采取平滑技术处理此类情况。

$$P(\text{"词语"}w|S)$$

$$= \frac{\text{某类中出现"词语"}w\text{的次数的总和}+1}{\text{某类中所有词出现次数（计算重复次数）的总和} + \text{被统计的词表的词语数量}}$$

给未出现在训练集中的词语一个估计的概率，相应地调低其他已经出现的词语的概率。平滑技术是因为数据集太小而产生的现实需求，如果数据集够大，平滑技术对结果的影响将会变小。

## 7.参考文献

- [1] 杨开平, 李明奇, 覃思义. 基于网络回复的律师评价方法[J]. 计算机科学, 2018, 45(09): 237-242.
- [2] 张帆. 贝叶斯算法在校园留言板垃圾过滤中的应用研究[D]. 郑州大学, 2016.
- [3] 马小龙. 网络留言分类中贝叶斯复合算法的应用研究[J]. 佛山科学技术学院学报(自然科学版), 2013, 31(02): 43-47+68.
- [4] <http://qzlx.people.com.cn/GB/382736/index.html>



