一、 问题分析

1.1 问题描述

近年来，随着各类社情民意相关的文本数据的不断攀升，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的挑战。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。

问题给出了四个附件，附件一包括对留言进行分类的三级标签体系，附件二给出了已经完成一级分类的留言数据，包括留言编号、留言用户、留言主题等六项数据，附件三给出了包括留言时间、赞成数、反对数等七项指标的留言数据，附件四给出了包括答复意见、答复时间等七项内容的留言数据。

第一个问题要求按照附件一的划分体系对留言进行分类，建立关于留言内容的一级标签分类模型。第二个问题要求对附件三的留言内容进行归类，定义合理的热度评价指标，并给出评价结果。第三个问题要求参照附件四的答复意见，给出对答复意见的质量的评价方案。

1.2 论文的结构安排

本文共分为 n 章，各章内容安排如下：

第一章，对论文需要解决的问题进行描述，并简单介绍整篇论文的结构安排。

第二章，对数据进行预处理，方便后续的问题解决。

第三章，

二、 数据预处理

在数据挖掘过程中，数据预处理是第一步，同时也是很重要的一步，数据预

处理的好坏直接决定着之后特征提取、分类预测等步骤是否能顺利进行。在对数据进行分析之前和分析过程中，我们逐渐对题目所给的数据的认识的逐步加深，并最终得出了一套较为完整的数据预处理流程。

2.1 编码一级标签

附件二所给的一级标签为中文字符，不利于后续数据处理，因此我们使用 LabelEncoder 对一级标签进行编码，将七类一级标签分别标识为 0~6，并进行分类统计，便于后续构建模型。

	一级标签	type	count
0	城乡建设	4	2009
1	劳动和社会保障	1	1969
2	教育文体	5	1589
3	商贸旅游	3	1215
4	环境保护	6	938
5	卫生计生	2	877
6	交通运输	0	613

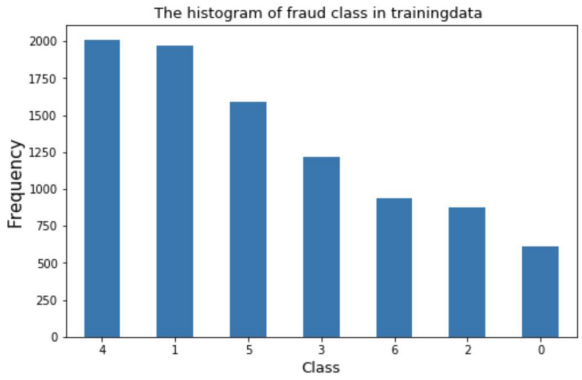


图 1 编码结果

2.2 留言文本预处理

在文本分类的流程中，文本预处理是对文本符号处理、分词、提取关键词、去除停用词等一系列操作的统称。经过预处理操作后，文本表示的质量得以提高，对于分类没有帮助的噪声得以减少。

2.2.1 数据清洗

首先使用 pandas 载入附件，并将数据以 data_frame 形式存储。我们发现留言中不仅包括了中文字符，还包括了数字、英文字符、标点等非常规字符，这些都是无意义的，并且会逃避分类器的识别，因此我们利用正则表达式定义了一个删除除字母、数字、汉字以外的所有符号的函数，从而避免干扰分类器性能，同时也可以减少分类负担。我们将数据清洗后的数据存储在 clean_review 列中。

2.2.2 jieba 分词

分词是中文文本分析的重要内容，正确的分词可以帮助我们更好的构建模型、运用算法分析。我们采用的是 jieba 库进行分词。

2.2.3 绘制词云

将分词后的数据按照七种不同一级标签进行词频统计，绘制词云。通过词云，我们发现一些词，例如：政府、老百姓、部门、希望等，对于后续区分毫无帮助且会造成偏差，我们将这些词筛选出来，设为自定义停用词表。

图 2 城乡建设类词云图

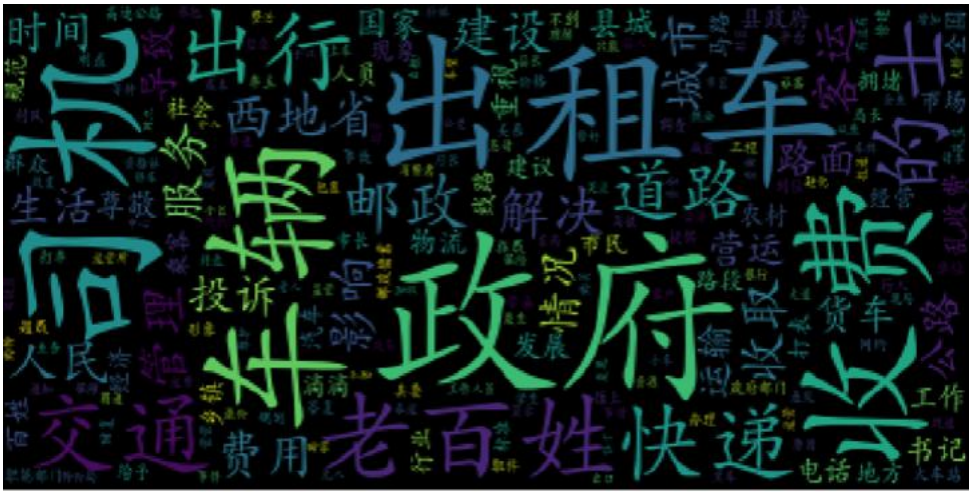


图 3 交通运输类词云图

2.2.4 去除停用词

无论是中文中，还是英文中，都有用来起连接作用的连词、虚词、语气词等无意义的词，这些词没有具体含义，只是起着衔接句子的作用。这些词对文本分析没有任何帮助，因此我们对分词后的数据进行了停用词处理。我们选用的停词表为中文停用词表、哈工大停用词表、四川大学机器智能实验室停用词库、百度停用词表以及自定义停用词表。自定义停用词表中为在数据挖掘过程中我们发现的一些无意义的词汇。

2.2.5 提取关键词

关键词提取是指从文本中确定一些能够描述文档含义的术语的过程，精炼分词结果，以帮助我们更好地构建分类模型。我们采用 jieba 库中的 textrank 方法和 tf-idf 方法来提取关键词，经过初步分词结果的对比，我们选择了效果更优的 textrank 方法。

$$WS(V_i) = (1-d) + d * \sum_{j \in \ln(V_i)} \frac{w_{jk}}{\sum V_k \in out(V_j) w_{jk}} WS(V_j)$$

此方法在 PageRank 的公式的基础上，引入了边的权值的概念，代表两个句子的相似度。从一段给定的文本中自动抽取出若干有意义的词语或词组，精炼分

词结果对已经分词的数据进行关键词提取，仅提取地名、名词、动名词、动词，并将提取到的词存储在 cut_review 列中。

	留言编号	留言用户	留言主题	留言时间	留言详情	一级标签	type	clean_review	cut_review
8811	137748	U0003355	A市四医院欺骗性回复让百姓心寒	2015/11/15 18:55:54	\n \n 再议A市四医院欺骗性回复到底为何！！！ 6月16日发贴，...	卫生计生	2	再议A市四医院欺骗性回复到底为何6月16日发贴认为A市四医院存在没有进行常规检查的情况下对儿...	检查 医院 耳鼻喉科 回复 情况 提出 医生 过度 患儿 关系 市四 孩子 就诊 公布 接受...
9196	352348	U0005416	咨询二女户计划生育补助金的政策问题	2019/12/19 9:21:35	\n \n 领导您好：我要咨询这个民生问题也携及到一部分群众的心声，就是农村...	卫生计生	2	领导您好我要咨询这个民生问题也携及到一部分群众的心声就是农村人群以前为了执行计划生育政策只生...	老家 农村 优惠 城市 咨询 执行 山林 购房 承包 土地 携及 取消 给于 合乎 政府 补...
1371	132255	U0004866	L3县大盒坪乡廖家坪村2017年贫困户危房改造补贴到底有没有?	2018/6/5 16:17:53	\n\t\t\t\t\t\n\t\t\t\t\t\t书记您好！我是大盒坪乡廖家坪村民张强致，再次...	城乡建设	4	书记您好我是大盒坪乡廖家坪村民张强致再次问政关于2017年贫困户危房改造补贴到底有没有今天下...	有没有 政策 贫困户 书记 补贴 老百姓 危房改造 解答 村子 村民 补田 没钱 乡镇 大盒...
898	94227	U0001693	J10 县五一路利必诚巷占道经营现象严重	2019/6/3 18:33:40	\n\t\t\t\t\t\n\t\t\t\t\t\t我是一名住在J10县五一路利必诚巷子的普通...	城乡建设	4	我是一名住在J10县五一路利必诚巷子的普通老百姓因为利必成巷子紧挨红太阳农贸市场每天都	巷子 小贩 状态 超市 管理 中央 图片 靠近 马路 沿街 叫卖 处于 行车 影响 出行

图 4 数据预处理后的附件一

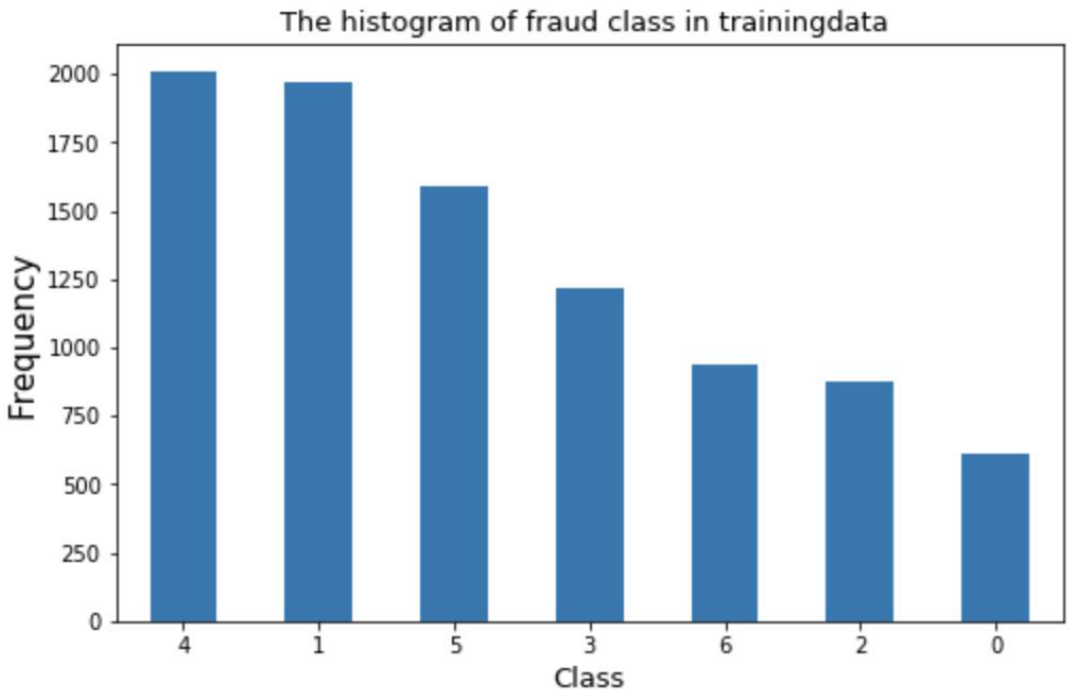
三、 问题一

3.1 LSTM 模型

3.1.1 向量化处理

我们将数据预处理过的数据进行向量化处理,我们要将每条 cut_review 转换成一个整数序列的向量，设置最频繁使用的 50000 个词，并且设置每条 cut_review 最大的词语数为 250 个(超过的将会被截去,不足的将会被补 0)。最终得到共有 20179 个不相同词语。

3.1.2 划分训练集和测试集



由于数据之间

四、问题二

4.1 问题归类

4.2 热度排行——Reddit 算法

4.2.1 数学原理

Reddit 的排名算法主要与以下内容有关：

1、文章的发表时间 t

$$t = \text{发表时间-2005年12月08日7:46:43}$$

用来标注留言新旧程度的单位为秒，其使用 Unix 时间戳（从 1970 年 1 月 1 日到当前时间的秒数）进行的计算，代码中的 1134028003 代表的日期为 2005 年 12 月 8 日 7:46:43，代表 Reddit 这个网站的上线时间。我们将其修改为 1496914280 代表的日期为 2017 年 6 月 8 日 17:31:20，为附件三中最早一条留言发表时间。通过上面的公式可以看到一旦留言发表，t 就是固定值，不会随时间改变，而且帖子越新，t 值越大。

修改后公式：

$$t = \text{发表时间} - \text{2017年6月8日17:31:20}$$

发表时间和热度排名的影响可以被概括如下:

（1）发表时间对排名有很大影响，该算法使得新的话题比旧的话题排名靠前。

（2）话题的得分不会因为时间的流失而减少，但是新的话题会比旧的话题得分高。

2、赞成票与反对票的差 x

$$x = \text{赞成票-反对票}$$

这是由于 Reddit 提供了投反对票的功能，所以可以使一些具有争议的话题会排的较后。而附件三中亦提供了反对数和点赞数的数据

修改后的公式：

$$x = \text{点赞数-反对数}$$

3、投票方向 y

y 是一个符号变量，表示对文章的总体看法。如果赞成票居多，y 就是 +1；如果反对票居多，y 就是–1；如果赞成票和反对票相等，y 就是 0。y 是文章评价的一种定性表达，0 表示没有倾向，大于 0 表示正面评价，小于 0 表示负面评价。

4、帖子的受肯定程度 z

z 表示赞成票超过反对票的数量。如果赞成票少于或等于反对票，那么 z 就等于 1。

结合以上几个变量，Reddit 的最终得分计算公式如下：

$$S_{\text{core}} = \log_{10} Z + \frac{yt}{45000}$$

这个公式可以分成两个部分来讨论：

1、$\log Z$

这个部分表示，赞成票超过反对票的数量越多，得分越高。 需要注意的是，这里用的是以 10 为底的对数，意味着 z=10 可以得到 1 分，z=100 可以得到 2 分。也就是说，前 10 个投票人与后 90 个投票人（乃至再后面 900 个投票人）的权重是一样的，即如果一个帖子特别受到欢迎，那么越到后面投赞成票，对得分越不会产生影响。而当反对票超过或等于赞成票，z=1，因此这个部分等于 0，也就是不产生得分。

通过观察发现，附件三中赞成数、反对数数值较小，因此我们将公式修改 logZ 为 Z。

2、$\frac{yt}{45000}$

这个部分表示，t 越大，得分越高，即新帖子的得分会高于老帖子。它起到自动将老帖子的排名往下拉的作用。 分母的 45000 秒，等于 12.5 个小时，也就是说，后一天的帖子会比前一天的帖子多得 2 分。结合前一部分，可以得到结论，如果前一天的帖子在第二天还想保持原先的排名，在这一天里面，它得到的净赞成票必须增加 100 倍。

y 的作用是用来产生正分和负分。当赞成票超过反对票时，得分为正；当赞成票少于反对票时，得分为负；当两者相等，得分为 0。这就保证了得到大量净赞成票的文章，会排在前列；得到大量净反对票的文章，会排在最后。投票对于总分的贡献不大，但是当投票的意见倾向发生变化时（由正面评价转向负面评价），投票对于总分的作用却是决定性（Y 的取值）。