“智慧政务”中的文本挖掘应用

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.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。

针对问题一，建立关于留言内容的一级标签分类模型，并且使用 F-Score 对分类方法进行评价。因此我们首先通过对数据进行文本去重，短句删除，去除特殊符号和停用词，去除异常值等操作对数据进行清洗，将 9210 条数据清洗至 8942 条数据。我们在预处理的基础上利用 jieba.analyse.extract_tags 方法基于 tfidf 算法进行留言主题和留言详情的一些关键词的提取，并构建词袋，将得到的矩阵作为 RidgeClassifier 模型的输入，利用 8 折交叉验证法构造测试集和验证集，最终得到验证集岭分类的准确率达到 91.24%，同时将留言主题和留言详情提取出关键词作为 Bert 模型的输入进行分类，分类准确率达到 98.21%。

针对问题二，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。因此我们首先按照同问题一同样的方法对附件 3 中的数据进行处理，个别同问题一进行区别处理。其次我们利用语义层面的多维度聚合融合机理进行文本聚合，其中包括主题词聚合、本体聚合、元数据聚合、摘要聚合、地区聚合五个维度分别进行相似度的计算，剔除一些相近似的文本，数据由原始的 4326 条数据，聚合后剩余 3273 条，接着利用层次聚类分析对留言归成 15 类，并将之前剔除的相似数据放入对应类别。接着我们构建热度评价体系，这部分可以根据问题一中的特征提取算法，从主题热度、内容特征热度、群众认可度这三方面作为一级指标进行构建，后用因子分析法对指标体系进行界定和评价，并得出排名前五的热点问题。

针对问题三，我们先对所给数据进行了简单的描述性统计。通过描述性统计与实际数据分析，对该问有了一个初步的了解。其次，我们考虑从三个方面出发，提出了一个评价指标方法。具体而言，我们从一个答复的相关性,完整性和时效性三个方面出发，分别给出了各方法的得分函数，并利用实际数据验证分析通过不同分割力度上的非监督文本相似性加权，我们给出量化答复与提问之间的相似性的方法。通过答复字数与文本匹配，我们给出了量化答复完整性的方法。在考虑答复时间与留言时间之间的时间差后，通过对其拟合分布，利用对应分布的分位数信息来作为时效性的得分。实际数据效果显示，我们所提出的得分方法，不但在各个方面都具有对异常数据的敏感性，而且其对大多数常见情况也表现良好。

关键词：岭回归分类；Bert 模型分类；多维度融合聚合；层次聚类法；因子分析法；

一、问题重述

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。

利用自然语言处理和文本挖掘的方法解决以下的问题。

1、群众留言分类

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，差错率高等问题。根据给出的数据，建立关于留言内容的一级标签分类模型。

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。请根据附件 3 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。

3、答复意见的评价

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。

二、问题分析

2.1 问题一的分析

问题一主要是让我们建立关于留言内容的一级标签分类模型，使用 F-Score 对分类方法进行评价。

首先我们通过对数据进行文本去重，短句删除，去除特殊符号和停用词，去除异常值等操作对数据进行清洗，将 9210 条数据清洗至 8942 条数据。

其次对于群众留言数据的特征提取，我们在预处理的基础上利用 jieba.analyse.extract_tags 方法基于 tfidf 算法进行留言主题和留言详情的一些具有代表性的动词和名词关键词的提取。

最后我们分别采用了 RidgeClassifier 模型和 Bert 模型进行分类。

2.2 问题二的分析

问题二需要我们附件 3 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。

首先按照同问题一同样的方法对附件 3 中的数据进行处理，个别不同进行区别对待。

其次我们利用语义层面的多维度聚合融合机理进行文本聚合，其中包括主题词聚合、本体聚合、元数据聚合、摘要聚合、地区聚合五个维度分别进行相似度的计算，剔除一些相近似的文本，数据由原始的 4326 条数据，聚合后剩余 3273 条，接着进行层次聚类分析对留言进行归类。

接着我们构建热度评价体系，这部分可以根据问题一中的特征提取算法，从主题热度、内容特征热度、群众认可度这三方面作为一级指标进行构建，后用因子分析法对指标体系进行界定和评价。

2.3 问题三的分析

针对相关部门对留言的答复意见，本文从答复的相关性、完整性和时效性三个角度出发，针对答复意见的质量，提出一套合理可行的评价方案。其中，相关性是指有关部门答复意见与群众留言内容之间的文本相似性，完整性是指有关部门答复内容具体并且能实际解决问题，时效性是指有关部门是否能在合理的时间期限内对群众留言内容给出答复。针对这三方面，我们计划提出了一套合理的答复意见评价准则。相关性方面，,我们通过不同分割粒度下的非监督相似性算法加权给出了一个合适的相似性得分模型；完整性方面，我们同时考虑答复字数与额外信息，给出了一个适当的完整性得分模型；.时效性方面，通过拟合答复时间与留言时间之间的差异，利用有关分布的分位数信息作为时效性得分模型。

三.基本假设

考虑到数据的缺失、数据异常值、设备异常等问题，以及为了可以顺利的完成本研究，本文所研究的内容都基于以下假设：

(1) 假设数据采集期间，留言系统运行正常；

(2) 留言时间等各项都记录准确；

(3) 数据标签标注正确；

四、数据预处理

首先要进行文本数据的预处理，文本数据里面存在大量价值含量低甚至没有价值含量的内容，如果将这些文本内容也引入进行分词、词频统计乃至情感分析等，必然会对分析造成很大的影响，得到的结果的质量也必然是存在问题的。那么，在利用这些文本数据之前就必须先进行文本预处理，把大量的此类无价值含量的评论去除。

4.1 文本去重

文本去重，顾名思义就是去除文本数据中重复的部分。我们对以下两种情况进行去重处理

1、同一个人可能会出现重复的留言，当一个问题在一定时间内没有得到解决或回复，同一个人可能会出现相同或相似的重复留言。

2、在大多数情况下，不同人之间的有价值的留言不会出现完全重复，如果出现了不同人留言的完全重复，这些留言的意义不是很大，一般是直接复制、粘贴其他人的留言，显然第一条最有价值。

相近的预料表达的含义相差不多，选择相似度高达 90%以上的预料进行删除，对于完全重复的语料直接采用最简单的比较删除法就好。

4.2 短句删除

由语言的特点知道，字数越少所能表达的意思就越少，想要表达一些相关的

意思就一定要有相应量的字数，过少的字数的评论必然是没有任何意义的留言数据，为此，就要删除掉过短的留言文本数据，以去除掉没有意义的评论，例如，

1、原本就果断的评论文本，如“很不好”。

2、经去词处理后过短的文本。

对于保留的评论的字数下限的确定，我们在此处设定下限为 5 个国际字符，即经过预处理后的语料若小于等于 5 个国际字符，则将该语料删去。

4.3 文本评论分词

本文采用 Python 的中文分词包“jieba”(结巴分词)，对附件中的留言数据进行中文分词。“结巴分词”提供分词、词性标注、未登录词识别，支持用户词典等功能。经过相关测试，此系统的分词精度高达 97%以上。为进一步进行词频统计，分词过程将词性标注作用去掉。

五、模型的建立与求解

5.1 群众留言分类问题的建立与求解

本次建模针对来自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众留言的答复意见数据，对文本进行基本的机器预处理、中文分词、停用词过滤后，通过建立包括栈式自编码深度学习、语义网络、RidgeClassifier 与 textrank 主题模型等多种数据挖掘模型，实现对文本评论数据的倾向性判断以及所隐藏的信息的挖掘并分析，以期望得到有价值的内在内容。

5.1.1 问题一数据的预处理

于附件 2 中的数据，我们做了如下预处理：

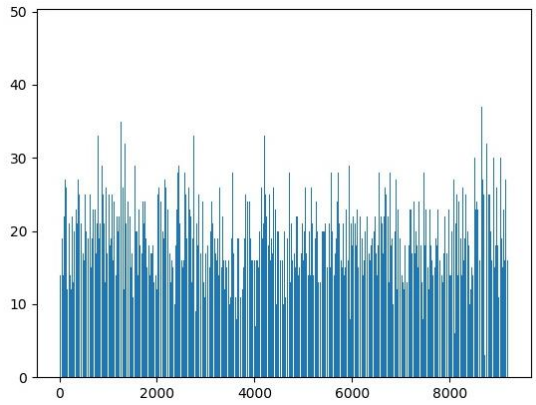
1、采用了数据清洗；去除“留言主题”、“留言详情”完全重复的语料，对于完全重复的数据只保留第一条。部分重复数据显示如下（全部数据见附件-第一问运行结果）：

7289	42091	U0005261	C1区农产品市场有无检验检疫证明啊？	2018/12/7	投诉了沙子岭往C市方	商贸旅游
7291	42395	U0003769	C5市润泽东方实景演出开演前有人卖假字	2018/6/8	C5市演出剧场开演前	商贸旅游
7297	44244	U0008135	C1区楠竹山镇私屠滥宰现象严重	2019/11/19	C市C1区楠竹山镇私屠	商贸旅游
7298	44331	U000467	C2区中瀚财富广场老旧电梯改造为什么那	2019/8/9	C2区中瀚财富广场老	商贸旅游
7309	48188	U0003789	餐馆饭馆的餐位费变茶水费，性质难道不	2018/7/11	现在不是取消了餐馆	商贸旅游
7316	49662	U0005261	C1区农产品市场有无检验检疫证明啊？	2018/12/7	投诉了沙子岭往C市方	商贸旅游
7317	49757	U0003769	C5市润泽东方实景演出开演前有人卖假字	2018/6/8	C5市演出剧场开演前	商贸旅游
7320	50044	U0008135	C1区楠竹山镇私屠滥宰现象严重	2019/11/19	C市C1区楠竹山镇私屠	商贸旅游
7321	50235	U000467	C2区中瀚财富广场老旧电梯改造为什么那	2019/8/9	C2区中瀚财富广场老	商贸旅游

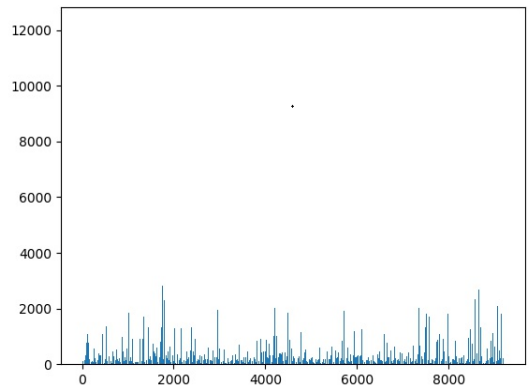
2、进行文本去重；去重前后按照一级标签数据分布如表 5-1 所示；去重前后留言主题和留言详情长度对比如图 5-2 所示；

表 5-1 去重前后一级标签分布表

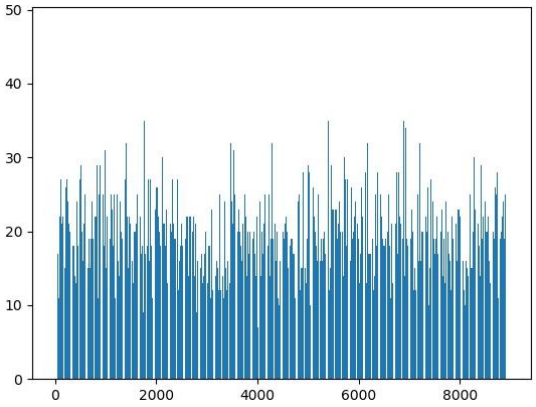
一级标签	城乡建设	劳动和社保	教育文体	商贸旅游	环境保护	卫生计生	交通运输
去重前	2009	1969	1589	1215	938	877	613
去重后	1981	1922	1544	1150	918	860	591



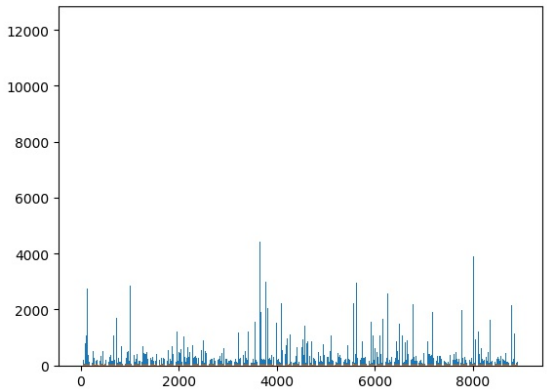
去重前留言主题长度



去重前留言详情长度



去重后留言主题长度



去重后留言详情长度

图 5-2 去重前后留言主题和留言详情长度对比图

我们可以看出原始数据的留言详情长度最大为 12226，最小为 12； 留言主题长度最大为 48，最小为 2；共 9210 条数据，去重后剩余 8966 条，不存在缺失值；且去重后,长度的最值未发生变化。

3、短句删除；其中设定删除下限为 20，删除详情长度小于等于 20 的样本，剩余 8942 条样本。接下来删除每条留言详情中长度小于等于 6 的句子。例如：不甚感谢,大家都知道,为老百姓做主等短句,此时留言详情长度最大值为 11562，最小值为 7。（短句删除后的全部数据见附件-第一问运行结果）

4、去除空格等特殊符号和去停用词等操作；经对数据探索性分析，发现原始数据中存在大量的空格、回车等符号，比如\n\t\u3000\xa0。对其进行相应的数据清洗。

5、去除异常值；由于原始数据量大，这类数据所占比例较小，对于问题影响不大，因此对其进行丢弃处理。例如以下图 中人为标注的标签错误

8064	154109	U0002489	B市发K市的物流单位联合经营，…	2017/10/15 12:47:13	目前B市发K市物流由三家物流单…	商贸旅游
2995	34376	U000938	B市发K市的物流单位联合经营，…	2017/10/18 21:46:46	目前B市发K市物流由三家物流单…	交通运输
2997	34380	U0002489	B市发K市的物流单位联合经营，…	2017/10/15 12:47:13	目前B市发K市物流由三家物流单…	交通运输

5.1.2 基于 tf-idf 算法的特征提取

tf-idf 算法用一种统计学的方法来衡量一个词语在文本中的重要程度，常被

用于信息提取、文本挖掘等场景之中。该算法的核心便是计算一个文本中某个词语的 tf 值与 idf 值。

tf 是指文本中的词频。衡量一个词语在文档中的出现频率有很多方法，最简单也足够有效的，便是直接计算这个词出现的次数，来作为这个词的 tf 值。

idf 是指“逆文档频率”，是一个用来衡量一个词常见程度的值。这个值的计算不应该基于单个文档，而应该考虑所有要进行分析的文档，来得出结果。

$$idf = \log \frac{\text{语料库文档总数}}{\text{包含该词的文档数}+1}$$

其中，分母处加一是为了防止某个词从未出现过而导致分母为 0。

Tf 与 idf 显然都与一个词的重要程度正相关，所以将其直接相乘，以乘积比较不同词之间的重要程度差异。

对留言主题和留言详情分别提取前 2 和前 5 个关键词，得到的数据如下表：

留言编号	主题_关键词	详情_关键词
24	安全隐患西湖施工建筑集团	路段未管杆圈文明城市车流安全隐患西行加油站上下班人行道路灯整改人流围墙燕子西湖便道路口安置大道
37	在水一方烂尾安全隐患大厦人为	在水一方过往行人锈迹斑斑四楼主干道护栏一楼人行道牵头倒塌拆除占用书院水电大厦请求人为设施车辆危机
83	停车费投诉违规物业市区	业主物业车位收费小区业主小区停车停车费第条程明西地省相关社区征得物业管理业主大会服务地下区苑公摊
303	不洗区华庭蔡锷水箱楼顶长年南路	不洗华庭水是健康霉味致癌物水箱环保部门楼顶长年自来水区区必不可少供水用品日常生活小区高层龙头二次
379	耀凯停水无故盛世投诉小区物业	小区业主停水停电物业公司耀凯楼楼水电费业委会物业费为所欲为问一问足额职能部门入住平方供水供电两层盛
382	供暖楼盘一事咨询	供暖近年西地省楚江规划气候具体位置片区阴冷楼盘常年潮湿恶劣月亮地处不知地区
445	桐梓停水小城可可西路得不到解决长期	停水自来水厂胡书记居民小区业主桐梓单位解决业委会苦不堪言感激不尽居委会恳请给与小城您好可可西路家住
476	垃圾处理收取平等城市	垃圾处理小区城市物业公司小区业主代收垃圾环卫局确实业主您提给予力度拒交梅家田依规费是雅优翠园费二希
530	脏乱差魏家坡小区	小区小区业主绿化带社区商业街破烂不堪拨款停车业主每次奥克斯无人管您们好魏家坡巷脏差搞人管文明小区魏
673	泰华小区业主业委会一村第四届侵占涉嫌小区非法公共资金	居委会业委会业主非法滨江小区业主泰华小区社区一村公章第四届违章建筑委员会谋利资料选举第二届委员年月
994	壹号溪湖区梅御湾用水业主	自来水溪湖停水用水恳请电话壹号住梅御湾楼自年就会水转凉洗个打不亮用据说楼停到热水澡百姓生活宜居
1005	限电区鸿涛关水入住翡翠强行业主	余晖业主交房物业限电延迟通知停水第二次同意协议房地产情况解决社区小区合同一栋栋一忍再忍拒绝

关键词提取的 Python 代码如代码清单清单 5-3 所示：

代码清单 5-3　关键词提取

```
def keywords_tfidf(data):#data 为预处理后的数据
    keyword = []
    for i in range(len(data)):
        corpus = data[i]
        keyword_tfidf = jieba.analyse.extract_tags(corpus)
        keyword.append(''.join(keyword_tfidf))
    return keyword
```

5.1.3 模型构建

1）**基于岭回归（RidgeClassifier）进行分类**

岭回归(Ridge Regression), 在最小二乘估计问题的基础上，向离差平方和增加了一个 L2 范数的惩罚项，Ridge 类有一个分类器的形式：RidgeClassifier. 该分类器首先把二值 targets 转换到 {−1,1}\{-1, 1\}{−1,1}, 然后把它作为一个回归任务，优化相同的目标函数。预测类对应回归预测的符号函数。对于多水平分类问题，则视为一个多输出的回归，预测类对应回归的最高值输出。

当前我们通过 tfidf 算法，将对留言主题和留言详情提取出关键词，构建词袋，得到的矩阵作为输入，然后将 8942 条输入数据分割成为 8 个子样本，1 个单独的子样本被保留作为验证集的数据，其他 7 个样本用来训练。交叉验证重复 8 次，每个子样本验证一次，平均 8 次的结果或者使用其它结合方式，最终得到一个单一估测。也就是 8 折交叉验证法，接下来进行岭回归分类，岭分类的准确率结果达到 91.24%。

2）**基于 Bert 模型的留言分类**

BERT 是一种新的语言表征模型，它用 Transformer 的双向编码器表示，意思它在处理一个词的时候，能考虑到该词前面和后面单词的信息，从而获取上下文的语义。与最近的其他语言表示模型不同，BERT 旨在通过联合调节所有层中的上下文来预先训练深度双向表示。因此，预训练的 BERT 表示可以通过一个额外的输出层进行微调，适用于广泛任务的最先进模型的构建，比如问答任务和语言推理，无需针对具体任务做大幅架构修改。

我们把利用 tfidf 算法将附件中留言主题和留言详情提取出关键词作为输入。利用 8 折交叉验证法构造测试集和验证集，通过把给定标记对应的标记嵌入、句子嵌入和位置嵌入求和来构造其输入表示，利用 bert 模型分类准确率达到 98.21%。

3）**模型融合**

模型融合是机器学习中经常使用到的一个利器，它通常可以在各种不同的机器学习任务中使结果获得提升。顾名思义，模型融合就是综合考虑不同模型的情况，并将它们的结果融合到一起。在这里，我们选择从提交结果文件中融合，因为这样做并不需要重新训练模型，只需要把不同模型的测试结果找出来，然后采取某种措施得出一个最终结果就可以。最终测试集的 f1 分数达到了 98.21%。（融合结果见附件-第一问运行结果）

5.1.4 结果分析

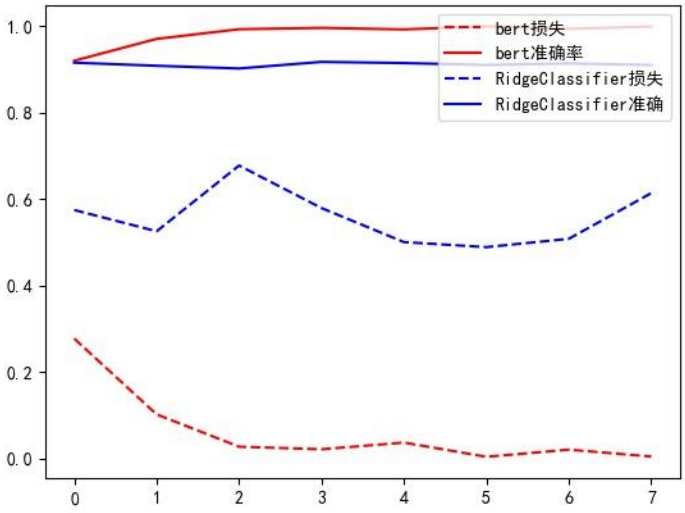


图 5-4 岭回归分类和 Bert 模型分类结果对比

通过图 5-4 我们可以看出，通过 8 次交叉验证选取不同的验证集的分类结果中 BERT 模型的留言分类结果优于使用岭回归的分类结果。同时采用平方损失作为损失函数，bert 模型分类损失也是远低于岭回归分类结果损失，综上我们选 BERT 模型进行分类，分类准确率达到 98.21%。

5.2 问题二模型建立与结果分析

5.2.1 问题二数据的预处理

附件 3 中数据原始留言详情长度最大为 6999，最小为 9；留言主题长度最大为 37，最小值为 0（也就是缺失值），对于缺失值我们直接删除；

对于文本去重、短句删除、去除空格等特殊符号和去停用词等操作，同问题

一处理不同的是对留言详情处理：删除每个样本中长度小于 5 的句子，剩余留言详情最长 6886，最短 9。（具体结果见附件-第二问运行结果）

值得注意的是由于一些留言主题较短，去掉停用词后，可能不存在关键词，因此我们进行如下缺失值处理。

缺失值处理：对于留言主题缺失值，选择留言详情关键词的前 5 个关键词进行填充；对于留言主题关键词缺失值，选择留言详情关键词的前 2 个关键词进行填充；

同时对于数据中出现的异常数据（64,65 行数据），如下图所示，我们采取的处理方式为：对其进行预处理（去除括号）后，按留言主题缺失的方式处理。

64	189739	A00051608	请问A3区西湖街道茶场村五组是如何规划的	2019/9/12 8:30:47	…	0	0
65	189856	A00073717	）	2019/7/3 11:53:35	…	0	1

5.2.2 留言分类模型构建及结果分析

1. 利用语义层面的多维度聚合融合机理进行文本聚合

我们在数据预处理的基础上，再利用 jieba.analyse.extract_tags 方法基于 tfidf 算法进行留言主题和留言详情的关键词的提取。然后针对如下五个维度进行聚合，即提取出关键词后，构建词袋，得到 tfidf 矩阵，利用余弦相似度进行相似度的计算。

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。相比距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上。

其中五个部分包括：

主题词聚合：对留言主题的关键词和详情的关键词分开进行相似度计算，选择的相似度阈值分别定为 0.9 和 0.8，将相似度高于阈值的几个数据提取出来，作为聚合结果 ；

本体聚合：留言主题和留言详情去除特殊符号和停用词后本身进行切词，分开进行相似度计算，相似度阈值分别定为 0.65 和 0.66，将相似度高于阈值的几个数据提取出来，作为聚合结果 ；

元数据聚合：由于同一人在较短时间内进行的留言，大概率上是重复的，所以在此我们主要指留言用户和留言时间进行相似度计算，相似度阈值定为 0.9，将相似度高于阈值的几个数据提取出来，作为聚合结果 ；

摘要聚合：留言主题和详情的关键词作为总的摘要部分进行相似度计算，相似度阈值定为 0.7，将相似度高于阈值的几个数据提取出来，作为聚合结果；

地区聚合：对原始留言主题中的位置信息进行相似度计算，相似度阈值定为 0.68，将相似度高于阈值的几个数据提取出来，作为聚合结果；

最终将五个部分的聚合结果进行融合后，保留相似样本的第一个，其他相似的样本进行删除。经过这一系列操作后，数据由原始的 4326 条数据，多维聚合后剩余 3273 条。

具体操作过程由图 5-5 如示：

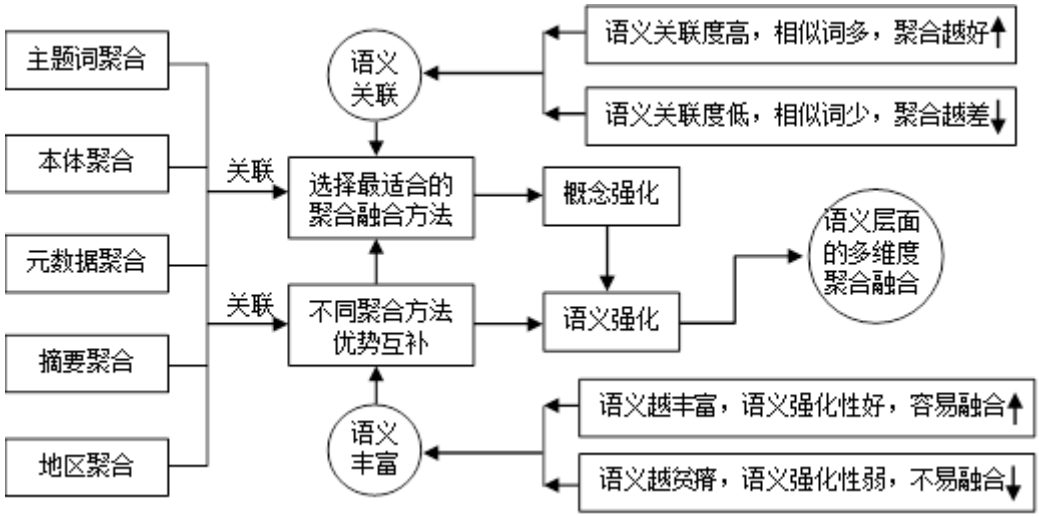


图 5-5 多维度聚合融合机理进行文本聚合流程

2.利用层次聚类进行留言的归类

层次聚类法（Hierarchicalmethods）先计算样本之间的距离。每次将距离最近的点合并到同一个类。然后，再计算类与类之间的距离，将距离最近的类合并为一个大类。不停的合并，直到合成了一个类。其中我们采用的类与类见的距离的计算方法为最短距离法，类内距离采用最常用的欧式距离。

我们根据第一步的多维度聚合融合机理方法更新文本后，重新构建词袋空间，得到 tfidf 矩阵，利用层次聚类对文本进行分类，得到的谱系聚类图如下图所示：

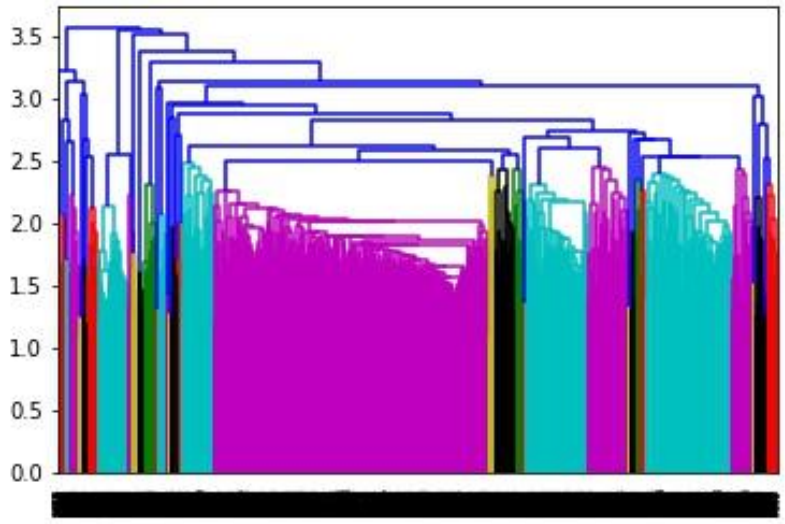


图 5-5 谱系聚类图

由谱系聚类图中我们可以看出，如果从高度 2.0 截取，会分成 150 类，从高度 2.5 截取。可以分为 34 类，为了和第一文中根据一级标签分类数一致，我们选择从高度 2.9 截取，将文本数据总共分成 15 类

最后，将上述利用多维度聚合后剔除的相似数据归并到 15 类数据中得到最终的分类结果。（详细的分类结果在附件-第二问运行结果中分类结果显示）

5.2.3 热度评价体系的构建及结果分析

1、热度评价体系的构建

该题需要我们对第二问的分类结果进行热度评价并进行排序，首先构建一个热度评价体系，我们拟从主题热度，内容特征热度，传受众特征热度三个维度选取关键词，留言字数，留言数目，出现及时性，点赞数以及反对数这六个要素对热点问题进行热度定量评价。具体指标构建情况，我们由图 5-6 展示：

图 5-6 热度评价体系构建模型

1、主题热度；我们选取每一类中关键词数目占整个关键词数的比重作为该主题的热度。

2、内容特征热度影响力；我们通过留言字数，留言数目，出现的及时性三个二级指标进行反映。首先字数越多其表达出的信息也越充分。该类留言数目越多也体现热度高；出现的及时性，我们根据图 5-7 可以看出，

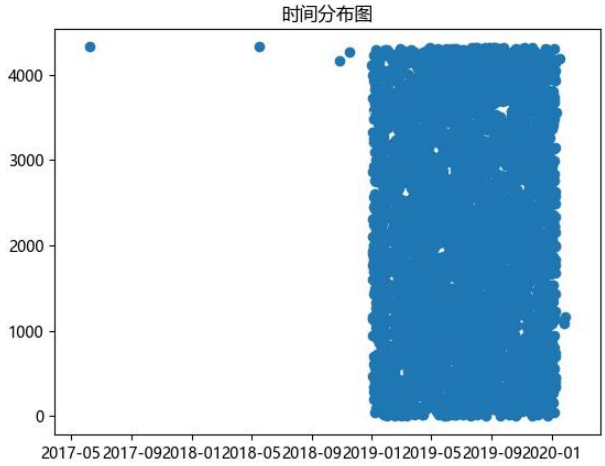


图 5-7 时间分布图

由图 5-7 可以看出留言时间主要集中在 2019 年 1 月到 2020 年 1 月，因为留言出现的及时性可以反映出留言内容如今的热度，所以我们选择了留言时间在 2019 年 1 月 1 日到 2020 年 1 月 1 日之间的数据进行热度评价，时间范围之外的进行剔除。

3、受众特征热度影响力。受众特征是指受众看到留言信息后所产生的的态度和看法，受众的活跃度对留言热度产生了极大的影响，所以我们选择将每类留

言的点赞数减反对数作为受众影响力。

2、运用因子分析方法对指标体系进行界定和评价

首先我们对搜集的数据进行因子分析可行性检验，通过样本 KMO 检验与 Bartlett 球形检验，表明统计数据适合做因子分析。

1、首先利用碎石图观察所需提取公因子的个数；如图 5-8 所示

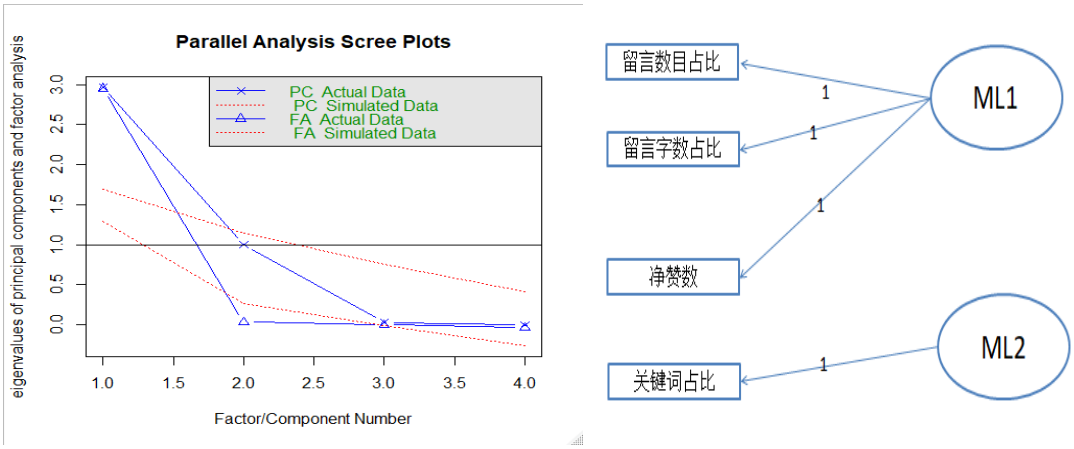


图 5-8 因子分析碎石图以及载荷矩阵

图 5-8 中左图显示的是主成分分析和因子分析两种方法的碎石图，我们可以看出 PCA 算法在第二个因子处出现拐点，同时前两个因子特征值大于 1，因子分析法是在第一个因子处就出现拐点且因子特征值大于 1，综合实际情况，我们选择两个公因子进行分析。

图 5-8 中右图绘制的是因子分析的载荷矩阵，我们可以看到第一公因子主要是由留言数目占比、留言字数占比、净赞数三部分组成，第二公因子主要是由关键词占比构成。

2、综合因子计算过程

表 5-9 主因子的一些数字特征

公因子	特征值	方差贡献率 %	累计方差贡献率 %
F1	2.948	73.7	73.7
F2	0.996	24.9	98.6

由表 5-9 我们可以看出第一个公因子的贡献率为 73.7%，第二个因子的贡献率为 24.9%，两个因子的贡献率和达到 98.6%

根据旋转因子载荷矩阵中各项指标与公共因子的相关系数值，我们可以列出各公共因子的计算方程，分别为

$$F_1=0.977X_1+0.998X_3+0.999X_4 \quad (1)$$

$$F_2=0.997X_2 \quad (2)$$

其中 X_1 代表净赞数， X_2 代表关键词占比， X_3 代表留言字数占比， X_4 代表留言数目占比。

在根据表 XX 中三个公共因子的方差贡献率对留言热度进行加权汇总，可以得到最终综合因子 F 的计算公式见公式（3）

$$F=0.747F_1+0.252F_2 \quad (3)$$

3、热度排名

根据因子分析得出的留言热度综合得分计算公式对本次分类结果进行排名，排名结果如下图所示。（具体结果见附件-热点问题表，热点问题留言明细表）

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	2.634163	2017/06/08至2020/01/26	A市教育、民政	经济学院强制学生实习、噪音扰民及其他民生政务
2	2	0.408029	2019/01/01至2020/01/06	A市城乡建设	建筑的质量、施工、利润等问题
3	3	0.234147	2019/01/03至2020/01/07	西地省劳动和社会保障、教育	工资拖欠、公司诈骗、学校建设、医疗门诊等问题
4	4	0.076103	2019/01/01至2020/01/06	A市环境和民生	小区物业乱收费、停水、停电、环境差
5	5	0.067419	2019/01/03至2020/01/07	A市线路规划	地铁线路施工存在的问题与建议

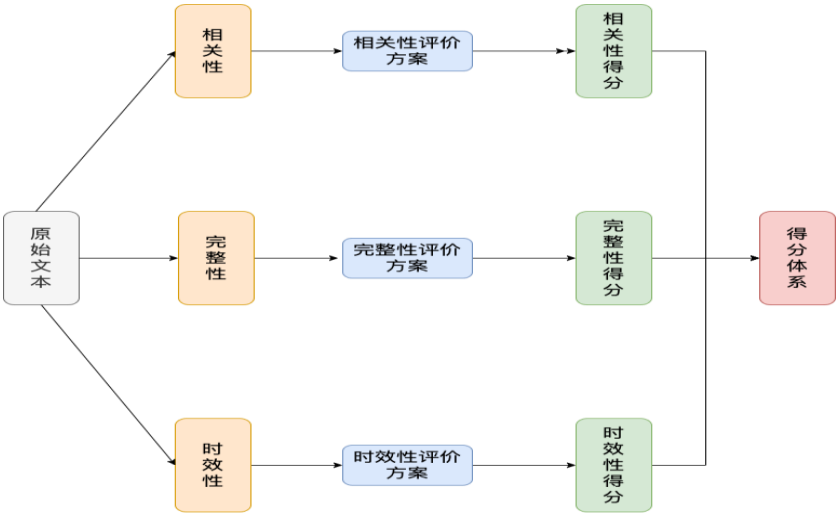
5.3 问题三模型建立与结果分析

针对相关部门对留言的答复意见，本文从答复的相关性、完整性和时效性三个角度出发，针对答复意见的质量，提出一套合理可行的评价方案。

相关性是指有关部门答复意见与群众留言内容之间的文本相似性；

完整性是指有关部门答复内容具体并且能实际解决问题；

时效性是指有关部门是否能在合理的时间期限内对群众留言内容给出答复。具体评价方案流程我们由图 5-10 显示



5-10 评价方案流程

5.3.1 问题三数据的预处理

1、在未对附件 4 中的数据进行任何处理前,先对其留言主题字数，留言详情字数,答复详情字数和答复时间与留言时间之间的时间差进行一个简单的描述性统计。具体情况如下表所示：

表 5-11 数据描述性统计表

	留言主题字数	留言详情字数	答复详情字数	时间差(day)
mean	18.547940	318.482599	360.256037	20.373224
std	4.672109	383.181847	429.929393	40.648486
min	8.000000	12.000000	3.000000	0.000000
25%	15.000000	102.000000	145.000000	5.000000
50%	18.000000	192.000000	276.000000	11.000000
75%	22.000000	369.000000	444.000000	22.000000
max	36.000000	3055.000000	7883.000000	1161.000000

其中,先对留言时间和答复时间进行截断处理，只保留其年月日信息,再定义答复时间与留言时间之间的差值为时间差，其单位为天。

从表 5-11 可以看到，留言主题的字数分布情况良好，但留言详情字数与答复详情字数跨度很大。值得注意的是，答复详情中最短的字符串长度只有 3，这说明一部分答复中存在敷衍了事的情况。大部分的答复字数都在 145-369 范围内，并且出现了 7883 字数这样的极端情况。对答复详情字数的一种普遍且合理的看

法是，答复字数越多，答复详情内容充实和对应留言被合理解决的概率也越大。再考虑时间差因素，不难看出，存在群众留言当天即可得到反馈的情况，将近一半的群众留言能在 11 天之内得到有关部门的答复，而大部分情况都能在 1 个月内得到有关部门的反馈。同时这里也出现了 1161 天这样的异常时间间隔。

2、通过上述描述性统计，进一步探索其中出现的异常情况。例如在答复意见中，出现如下异常情况：

表 5-12 答复意见异常情况表

留言标号	答复详情
25431	2016 年 6 月 12 日
37459	2019 年 1 月 14 日
37482	2018 年 12 月 12 日
101636	0000-00000000 ...
10924	网友：您好！留言已收悉
25918	UU0081182

由上表可以看出,存在一部分答复仅包含了年月日等时间信息，一部分答复为形如“网友：您好！留言已收”简单无意义句子，一部分答复仅为留言群众的用户编码。

针对留言详情和答复字数都出现极端情况下，为了保证留言详情和答复详情之间的文本相似性效果,对其字数之间的差值也进行了一个简单的描述性统计。

表 5-13 字数差值的描述性统计

	Q&A 字数差异
mean	-41.773438
std	516.241753
min	-7844.000000
25%	-215.000000
50%	-60.000000
75%	100.500000
max	2730.000000

虽然大部分情况下，留言和答复之间的差异都尚可，但也出现了-7844 和 2730 这样的异常情况。而在文本相似度方面，句子越长其越容易得到更高的相似度评分，但这并不能反映答复详情与留言详情之间的真实相似情况。

针对上面简单的描述性统计结果，针对问题三的原始数据，我们考虑对其进行如下预处理过程：

1、对文本进行常规的文本去重，去除空格等标点符号，去除例如“https://baidu,com”这样异常字符串和去除停用词。

2、针对某些答复详情在处理后出现空字符串情况，将其特殊赋值以保证有关算法的正常运行和效果的合理性。

3、针对留言字数与答复详情字数都出现了极端情况，可以考虑使用自然语言处理中的有关算法对其提取摘要,降低留言详情与答复详情字数差异，保证有关算法结果的合理性。

5.3.2 模型的建立与求解

一、相关性

在自然语言信息处理过程中，文本相似度起着至关重要的作用,其广泛应用于信息检索，文本挖掘，机器翻译等细分领域。在许多 NLP 实际应用中，我们

经常需要判断两篇文章是否相似，并计算两篇文档相似程度，并以此服务为下游任务。例如在对语料进行预处理时基于文本相似度，把重复度高的样本挑选并删除；在现有的聚类算法中，改写距离函数，从而实现热点话题的探索；在问答系统中，对用户问题进行反馈,在现有的回复库中提供适当的选择。文本相似度计算方法一般由两个关键部分构成，即文本表示模型和形似度度量方法。前者负责将文本表示为数值向量，即提供特征；后者在前者的基础上，基于数值结果来反馈文本之间的相似程度。

我们认为一个合理的答复，应该在能充分体现留言中的关键内容。具体而言，即使在不同的文本切分粒度和特征构建方法下，一个合理且完善的答复都应与其对应的留言内容体现出较高的文本相似度。在该问中，基于所给定数据我们考虑使用一系列无监督性的文本相似度的加权形式来作为相关性的得分评价准则。我们认为一个的文本相似度算法，其方差越大越能反映其划分能力。因此，在权重的选取方面，考虑采用各算法在所给数据上的方差与总方差的比值作为其权重。

针对原始数据字符之间长短不一的情况，我们优先考虑直接使用 Simhash 算法对留言内容与答复内容之间的相似度进行初步的估计。基于语义词典计算文本相似度在近年一直被广泛使用，常用的中文语义词典为《知网》和《同义词词林》。因此我们考虑通过自然语言处理相关算法提取留言与答复的短摘要,降低其文本字数之间的差异，再利用基于语义词典计算文本相似度算法来对其相关性进行进一步的估计。

1．所涉及到的文本相似度方法简要说明

Simhash 算法是敏感哈希算法在文本特征提取任务中的应用。它会把一篇文档映射为一个长度为 64、元素值为 0 或 1 的一维向量，这样我们就可以使用某种距离计算方式，计算两篇文本的距离和相似度了。一般来说，我们考虑海明距离。Simhash 算法对文本的“相同”与否特别敏感，因此非常适合用来判断两篇文档内容是否相同。另外 Simhash 算法计算简单，速度上有一定优势,可以用来对海量文档进行去重。基于语义词典的相似性算法通常采用层级体系，不同级别的分类结果可以为自然语言处理提供不同的服务。该系列算法从词语的语义出发，根据词语的同义项在词林中的位置和编码，计算出其词语的相似度。过去的诸多研究分析均表明，对词义进行有效的扩展或者对关键词做同义词替换可以明显的改善信息检索、文本分类和自动问答系统的性能。

2．实际数据分析结果

表 5-14 三种相似度算法结果比较

	Hownet	Cilin	simhash
mean	0.500570	0.581847	0.568054
std	0.113661	0.089690	0.079027
min	0.000000	0.000000	0.328125
25%	0.432844	0.526316	0.515625
50%	0.504493	0.585000	0.562500
75%	0.574374	0.640000	0.625000
max	0.922222	0.920000	0.859375

通过对附件 4 数据进行上述预处理后，分别使用上述三种相似度算法对其相似程度进行估计。可以看出提取摘要后，基于语义词典的无监督相似性算法的标准差均更大，而 simhash 算法效果次之。可以肯定的是，上述三个算法虽然性能之间存在差异，但均能良好的反映留言详情与答复详情之间的相似程度。

3. 相似性得分

通过实际数据分析结果，我们采用各算法的方差与总方差的比值作为加权和的各权重。具体值如下表所示：

	Hownet	Cilin	simhash
权重	0.48	0.30	0.22

即对于一个新样本 x，其相似性得分(similarity score)为：

$$Score_{similarity}(x) = (0.48 * Score_{howner}(x) + 0.3 * Score_{cilin}(x) + 0.22 * Score_{simhash}(x)) * 100)$$

4. 实际数据效果

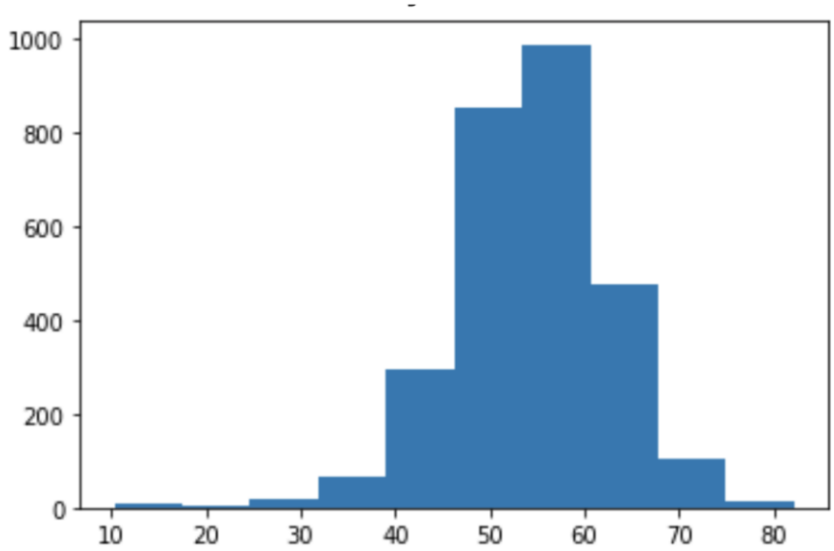


图 5-15 相似性得分直方图

上图横坐标为相似性得分，纵坐标为区间计数。从上图可以看出,大部分样本的相似性得分都集中在 50-60 范围内.回顾实际数据，大部分的答复内容也集中在转交其他部门待处理的形式上,并没有正面地对用户提问进行答复，其相似性得分也不会很高。而对于答复合理且详细的情况,其得分也能体现出来。

二、完整性

若一个答复是完整的，意味着该答复意见不但能有效针对用户留言，同时也提供了一定的解决方案。首先，回顾实际数据,答复意见的一般格式均为“对群众留言的回述+解决方案+回复时间”。当然表 5-12 中也列出了一些现存的异常回复情况，通常为无意义的时间信息或者重复了用户编码。这提示我们可以从答复字数的角度出发。当一个答复字数小于一个阈值时，其提供可行解的概率也会越小,也不可能为一个完整答复。

另一方面，回顾实际数据，其已对一些敏感信息做了脱敏处理，例如地点信息，联系电话，有关行政法律文件等。这导致现有的方法很难精细地对其提取对应的有效内容。因此，我们只能通过匹配关键词来初步判断该回复中是否存在存在移交相关部门，是否存在有关行政公文等信息。

1. 完整性得分

完整性得分为三个部分得分之和，具体而言，先针对答复字数我们给出一个得分方法。再识别答复内容是否存在移交其他部门的信息，若存在，给出一个相应的额外得分。最后识别其是否有政府公文信息，若存在,给出一个相应的额外得分。

从表 5-11 出发，下表为一个简单可行的字数得分：

答复字数区间	得分
[0,11]	0
[12,145]	20
[146,276]	40
[277,444]	60
[445,∞)	80

除此之外，也包含了两个额外得分，具体如下表所示

额外项	得分
移交其他部门信息	10
政府公文信息	10

2. 实际数据效果

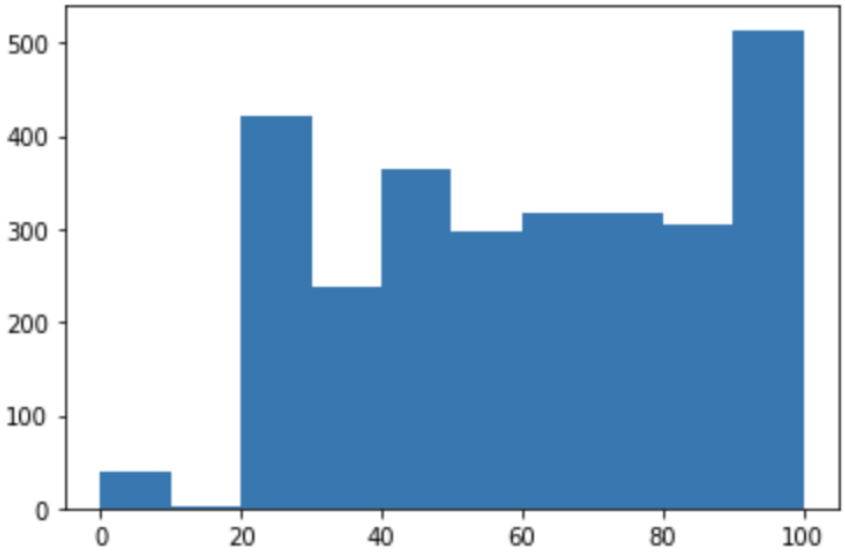


图 5-16 完整性得分直方图

上图横坐标为完整性得分，纵坐标为区间计数。该得分方法既保证了一部分高质量回复分数高，也区别开了前文所述的低质量无意义回复。针对普通回复，也保证了一定意义上的差别性。

三、时效性

在数据预处理过程中，对时间差特征进行了简单的描述性统计。为了进一步对其建模，探索更多的信息，我们考虑对其拟合分布。对大于 200 的数据，我们进行通过有关算法，得到拟合结果如下表 5-17 所示：

表 5-17 分布拟合结果

distribution	Sum square error	aic	bic
wald	0.000100	1550.795617	-47951.020255
gilbrat	0.000136	1486.265003	-47104.005170
lognorm	0.000197	1467.272098	-46061.189514
johnsonsu	0.000198	1468.757644	-46027.594028

表 5-17 列出了前 4 个均方误差最小的分布拟合结果。其中均方误差(sum square error)为实际数据与给定参数后所产生对应分布随机数的误差的 2 范数，其主要反映模型拟合的好坏程度；AIC，BIC 信息量广泛用于统计领域,主要通过这两个信息量来作为模型筛选的准则。一般来说，AIC 和 BIC 越大,对应的模型越符合实际情况。

更一般的，可以得到其直方图，具体如下图所示：

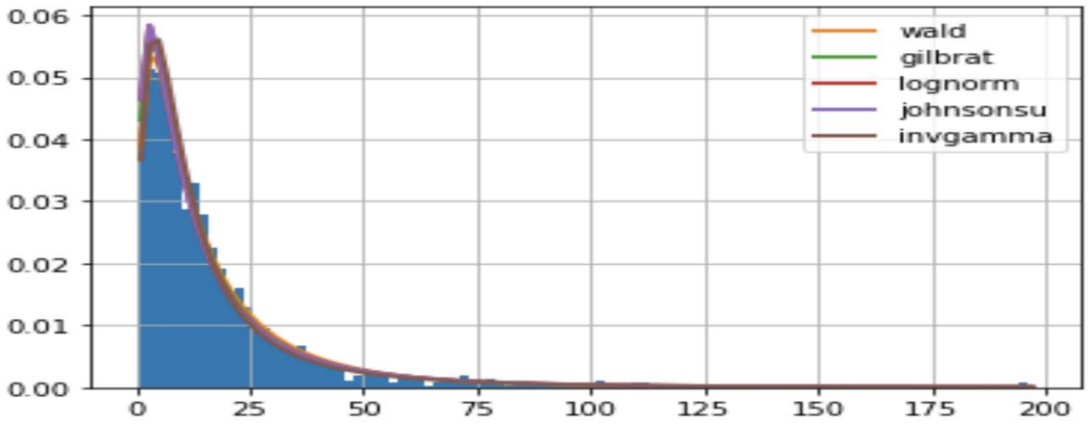


图 5-18 拟合效果直方图

可以看出时间差特征在这个分布上拟合情况均良好，其均方误差很小，AIC 和 BIC 准则尚可。为了简单，我们采用三参数对数正态分布作为时间差特征的先验信息。并且从上图也直接体现了拟合情况，这说明我们对时间差特征拟合分布是合理的。

1. 时效性得分

针对时效性得分，一个合理且可行的方案应该满足以下两方面：

1. 得分应该随时间差的增长而递减。即有关部门回复的越快，其时效性得分就应该越高。与之相反，有关部门对群众留言越不关心，其时效性得分就应该越低。
2. 大部分情况下，即 25%分位点至 75%分位点这部分数据，其得分情况应该起伏不大。换句话说，时效性得分标准一方面应该对密集型数据迟钝，而对异常型数据敏感。

在上一节中，我们考虑采用三参数对数正态分布作为时间差特征的先验信息。从这个角度出发，一种简单且有效的定义得分函数的方法是：

$$Score_{time}(x) = (1 - F(x)) * 100$$

其中 F 为 x 对应累积分布函数。这个得分函数既保证了其对时间差特征异常情况敏感，也能保证对于大多数情况其分数合理。换句话说，我们采用其在特定

分布下的分位数信息来作为时效性得分，相对于定义线形得分函数而言，其更合理也更能反映对应的统计学意义。

2. 实际数据效果

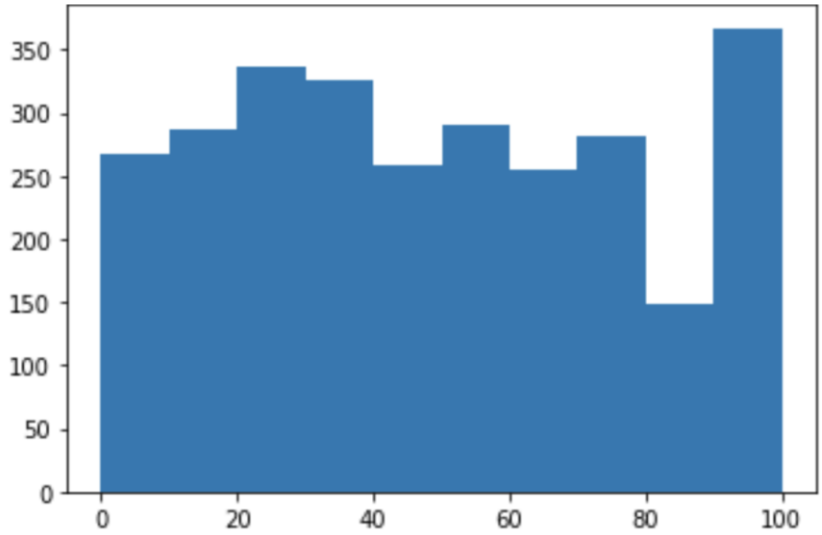


图 5-19 时效性得分直方图

上图横坐标为时效性得分，纵坐标为区间计数.从上图可以看出， 利用分位数信息来作为时效性得分，可以保证各分数区间内的样本数差异不大，有效地对时间差特征进行了差异化描述。结合图 5-17 来看,该方法是合理的。当时间差特征过小亦或者过大时，其时效性得分敏感；而当时间差特征集中在其均值 20 附近，其时效性得分迟钝。

六、模型优缺点分析

6.1 模型的优点

1、本文采用多种启发式算法对问题的不同角度进行求解，而不是盲目穷举。

2、bert 模型同时利用左侧和右侧的词语，可以对语义进行很好的学习，相对于 RNN 而言更加高效、能捕捉更长距离的依赖。

3、模型容易实现，如果想解决一个新的问题，对于岭分类算法只需针对新的问题重新进行编码就行；对于 bert 模型，只需要把处理后的文本直接塞入模型进行学习即可。

6.2 模型的缺点

1、岭分类算法对语义的学习较弱，对预处理要求较高。

2、bert 模型对硬件要求较高，需要大量的时间去训练大量的参数

参考文献

[1]杨开平. 基于语义相似度的中文文本聚类算法研究[D].电子科技大学,2018.

[2]李春婷. 基于语义相似度的文本聚类算法研究[D].重庆邮电大学,2017.

[3]梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.

[4]赵舒贞. 网络舆情预警指标体系的建构[D].云南师范大学,2013.

sun:

[5]李春林,冯志骥.基于文本挖掘的新能源汽车用户评论研究[J].特区经济,2020(04):148-151.

[6]李舟军,范宇,吴贤杰.面向自然语言处理的预训练技术研究综述[J].计算机科学,2020,47(03):162-173.

[7]刘思琴,冯胥睿瑞.基于 BERT 的文本情感分析[J].信息安全研究,2020,6(03):220-227.

[8]杨萌,张云中,徐宝祥.社会化标注系统资源多维度聚合机理研究[J].图书情报工作,2013,57(15):126-131.