基于文本挖掘和分类的智慧政务处理与评价体系

摘要

近年来，随着网络问政平台逐步发展，依靠人工来进行留言划分和热点整理的相关部门的工作迎来了极大挑战。但随着大数据、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。

问题一对于群众留言分类，在经过对数据的预处理之后，首先将要分类的七个一级标签数字化，通过剔除停用词，分词等一系列方式将文本数据量化，通过计算其 TF-IDF 值进而使用多种分类器模型进行 F-score 的比较，以及通过七维的混淆矩阵，以热力图的形式直观的比较各分类器的效果，获得更佳的文本分类器模型。

问题二对于热点问题挖掘，其难点在于如何准确的获得地点人群信息。在尝试 k-mean 聚类效果不佳的情况下，通过分析数据类型本文将待挖掘问题分区细化。在更小的样本下词频的统计，从而获得各个分区较有可能成为热点新闻的地点以及人群。通过反馈维度和时间维度两个维度来建立起热度评价体系模型。

问题三对留言答复质量进行评价，本文首先对答复质量评价进行指标的提取，得到信息量、相关性、完整性和可解释性这四个指标，接着对每个指标进行量化处理并计算出具体数值，然后通过专家打分法得到指标的权重，并用灰色关联度修正权重得到最终的质量评价模型。

关键词：文本向量化，TF-IDF 值，F-score，文本相似度，灰色关联法

目录

一、问题重述

1.1 问题背景

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。

1.2 明确目标

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，且差错率高等问题。本文将根据附件 2 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，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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。本文根据附件 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、数据预处理

2.1 数据存储

将附件二中共有 9210 条数据以 dataframe 导入 python 储存。首先需要对数据进行空值以及重复数据的筛查。经过筛查，9210 组评论之中没有存在重复的样本以及缺失样本的情况。

```
在留言详情列中总共有 0 个空值.
在一级标签列中总共有 0 个空值.
```

2.2 删去空白符，换行符

由于评论之中存在大量的空格，换行符等不具备词意的“分割点”。需要在数据预处理阶段将这些删去。部分处理好的数据如下图所示。此时数据中不再含有\n，\t 等符号。

	留言详情	一级标签
1011	尊敬的张书记：您好!楚江纸业春江路安置小区（市房产局对面）建成已经两年有余，至今小区内各条道...	城乡建设
6471	领导：您好！1、现就小孩的社保进行咨询，我县小孩从入读幼稚园开始，学校就办理了社保卡，那我想...	劳动和社会保障
8110	遵敬的李局长：我叫王坚良，是福建泉州人。在今年十月份我上A市与西地省瑞红商贸发展有限公司签订...	商贸旅游
4962	想咨询下小孩子入学的一些细则政策，把教育局报名系统上显示的A市基础教育科所有电话：市、区全部...	教育文体
819	尊敬的易书记：市政府后面的五岭阁公园标致性建筑五岭阁，从建成到现在，就从未对普通市民开放过。...	城乡建设
1518	M市房产信息网以前能查到个人购房合同备案信息等情况，近段时间相关页面打开后一片空白，没有显示...	城乡建设
4064	在国家鼓励民间资本办学的政策下G5县却催生出了一种异胎“民办公助”的办学模式，即私人办学，政...	教育文体
430	本人于2016年10月申请住房公积金贷款，当时所有资料齐全，并且按照住房公积金要求，多次来签...	城乡建设
3810	A7县当年的代课教师都摸底调查造册了，补贴费用为什么还没发下来?	教育文体
3027	村里的路是稀烂的，是越野车试车的地方，两驱进不来，要四驱。村里的石头被乱开采，路上跑的全是后...	交通运输

图 1 删去空白符和换行符的部分数据

2.3 删除停用词

该步骤与第 2 步有些类似。这里主要对中文评论进行标点符号，特殊符号，以及一些无意义的常用词如“吗，呢，啊”带有感叹色彩的词，“然而，然后”连接词以及“你，我，他”代词等[1]。在网上下载常用停用词表（https://blog.csdn.net/qq_34696236/article/details/80536783）删除后如下所示。

	留言详情	一级标签	code	clean_review
8860	K6县中医院陈彬彬，身为医务人员，计生专干，知法犯法，自从管理开具出生证以来，利用手中权力，...	卫生计生	6	K6县中医院陈彬彬身为医务人员计生专干知法犯法自从管理开具出生证以来利用手中权力多次收受他人...
4813	对于教职员工所关切的进岗问题，一所学校领导工作分数高达整个教育生涯的百分之三十以上，而另一所...	教育文体	3	对于教职员工所关切的进岗问题一所学校领导工作分数高达整个教育生涯的百分之三十以上而另一所没有...
3371	省局领导：我叫刘平剑男工作单位是J3县文化广电新闻出版局。是一名老邮政特邀监督员。近日，我收...	交通运输	2	省局领导我叫刘平剑男工作单位是J3县文化广电新闻出版局是一名老邮政特邀监督员近日我收到省邮政...
6463	2017年7月14日，市直事业单位公开招聘简章发布之后，本人意向职位是K市一中档案员之职位，...	劳动和社会保障	4	2017年7月14日市直事业单位公开招聘简章发布之后本人意向职位是K市一中档案员之职位于是本...
7499	我是J8县人，后搬到广东韶关。心系家乡情。高速公路甚至高速铁路的规划建设，就好比打通了J8县...	商贸旅游	5	我是J8县人后搬到广东韶关心系家乡情高速公路甚至高速铁路的规划建设就好比打通了J8县的任督二...
3928	现在有好些老师好多年了都在停薪留职，请问这种情况教育部门允许吗？停薪留职老师的工资都被谁领用...	教育文体	3	现在有好些老师好多年了都在停薪留职请问这种情况教育部门允许吗停薪留职老师的工资都被谁领用了谈...
4850	现D6县洪市镇洪市中学领导换届，学校管理混乱希望上级领导干部亲临视察为学生主持公道。1.食堂...	教育文体	3	现D6县洪市镇洪市中学领导换届学校管理混乱希望上级领导干部亲临视察为学生主持公道1食堂承包私...
7347	2018年9月20日，中午正是学生下课回家吃饭时间，一个学生骑自行车金桂园附近马路骑行道上，...	商贸旅游	5	2018年9月20日中午正是学生下课回家吃饭时间一个学生骑自行车金桂园附近马路骑行道上一辆无...
349	在2019年我们小区部分业主查出A市A1区星典时代小区维修基金使用情况，并发至小区业主群里，...	城乡建设	0	在2019年我们小区部分业主查出A市A1区星典时代小区维修基金使用情况并发至小区业主群里发现...
771	柳叶大道与G1区大道交汇处的邮政大楼门前的人行道，坑坑洼洼，晴天一身灰，雨天一路泥，与G市的...	城乡建设	0	柳叶大道与G1区大道交汇处的邮政大楼门前的人行道坑坑洼洼晴天一身灰雨天一路泥与G市的卫生城市...

图 2 删去停用词的部分数据

三、群众留言分类问题的分析与解决

由附件一可以得知一级标签的所有类别。附件二给出了留言编号，留言用户，留言主题，留言时间，留言详情以及一级标签六列数据。问题一重点在于使用恰当的分类器将留言分类，以克服人工时间长，效率低以及差错率高的缺点。

作为一种有监督的机器学习方法，文本分类可以划分成两个环节，以此为文本的表现形式和分类算法。文本的表现形式对文本分类结果产生重要的影响，本文主要采取切分词的方法将文本数据向量化，并通过三种模型之间的比对得到更佳的分类器模型。

3.1 将标签数字化

问题一由于是分类问题，为了便于后面分类器的训练。在数据预处理中需要增加“将标签数字化”这一过程。增加一列后数据如下所示。

	留言详情	一级标签	code
5794	尊敬的陈书记：您好！我于是2012年11月在C市银行业协会开车，而银行业协会一直未给我签订劳...	劳动和社会保障	4
6996	本人因在A市工作期间，不慎受伤被路人送往A市区三甲医院西地省旺旺医院进行治疗，因医院反映本人...	劳动和社会保障	4
4076	B市幼儿园601钻石分园国庆前突然开了个家长会，说要办理入园磁卡和土星网的幼儿教育的上网资格...	教育文体	3
2875	蒋厅长：您好！目前K市K1区珠山镇内小冶炼厂到处可见，特别是原于家乡中学内的碳酸锰厂严重违反...	环境保护	1
6937	本人有晚辈在F市社保局的大厅窗口做临时工，经常会和我说一些工作的情况，我整理了一下，因为去年...	劳动和社会保障	4
4166	我是1983年7月技校毕业参加工作.井下工作了9年，1991年元月调入矿子弟学校工作.中级职...	教育文体	3
2147	我们茅塘村粉必组几十户村民联名投诉位于C3县107国道边中路铺镇荷塘开发区内“C市三兴工业科...	环境保护	1
5109	快到年底了，D6县教育局很多股室开始大量催交论文：教研、基教、装备、国培、督导等等，每个类别...	教育文体	3
4339	尊敬的局长：您好！看到西地省积极申办全运会的消息很高兴，作为西地省儿女，真心希望能把全运会请...	教育文体	3
2816	A市A9市河，远大路起至营盘路止，可以各自向两端延长200米，A9市河东岸常常黑烟滚滚，乌烟...	环境保护	1

图 3 标签数字化后的部分数据

数字对应的一级标签如下图

	一级标签	code
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

图 4 数字对应的一级标签

3.2 数据可视化分析

通过 dataframe 的筛选计数，可以得到附件二中需要分类的一级标签共有 7 类。其中 7 类的具体评论数分布如下图所示。

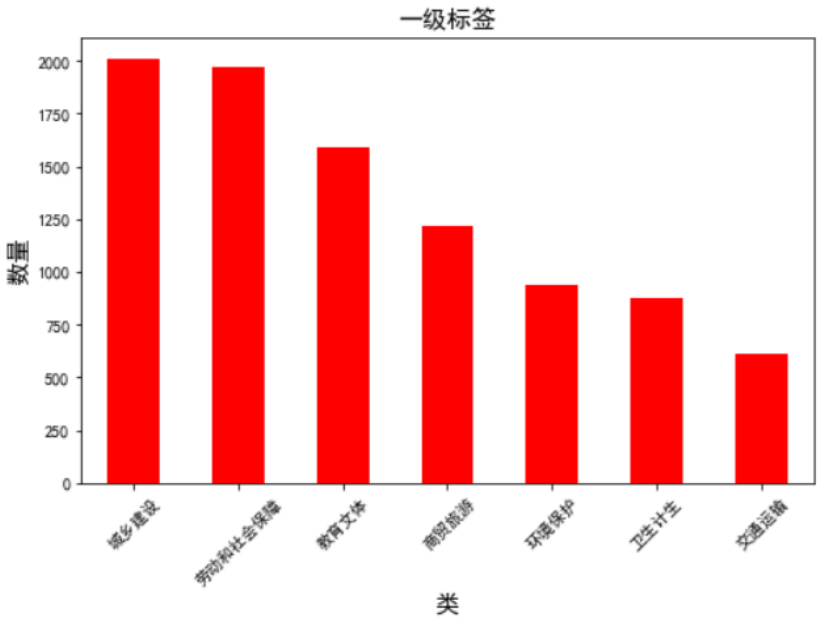

图 5 7 类具体评论数分布

从 training data 来看，没有哪一类数据过少。如果分类合理，算法正确，应该可以很好的训练模型得到正确的分类。

3.3 分类模型构建

3.3.1 分词

上述预处理之后得到的数据后，首先通过 jieba 模块对语句进行分词。

	clean_review	cut_review
	A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工...	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑...
	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏...
	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公...

图 6 分词后的部分数据

3.3.2 计算 TF-IDF 特征值

（1）原理及方法

TF-IDF（Term Frequency-InversDocument Frequency）是一种常用于信息处理和数据挖掘的加权技术。该技术采用一种统计方法，根据字词的在文本中出现的次数和在整个语料中出现的文档频率来计算一个字词在整个语料中的重要程度。它的优点是能过滤掉一些常见的却无关紧要本的词语，同时保留影响整个文本的重要字词。

TF（Term Frequency）表示某个关键词在整篇文章中出现的频率。

IDF（InversDocument Frequency）表示计算倒文本频率。

文本频率是指某个关键词在整个语料所有文章中出现的次数。倒文档频率又称为逆文档频率，它是文档频率的倒数，主要用于降低所有文档中一些常见却对文档影响不大的词语的作用。

计算方法：通过将局部分量（词频）与全局分量（逆文档频率）相乘来计算 tf-idf，并将所得文档标准化为单位长度。文件中的文档中的非标准权重的公式，如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_{i,j}$$

计算步骤：

①计算词频

$$词频(TF) = \frac{某个词出现的次数}{总词数}$$

②计算逆文档频率

$$逆文档频率(IDF) = \log\left(\frac{语料库的文档总数}{包含该词的文档数 + 1}\right)$$

③计算 TF-IDF 特征值

$$TF - IDF = 词频(TF) \times 逆文档频率(IDF)$$

（2）实现与结果

除了抽取评论中的每个词语外，由于两个相近之间可能存在一定的信息，因此需要将这些相邻的“连词”考虑进来。经过对公式的参数的刻画，得到了所需的分词后评论的 TF-IDF 特征值。第一个评论的部分特征值如下所示。在这个步骤中总共读取了 9210 条评论，711874 个词语以及连词。

```
(0, 24440)    0.08944097925743864
(0, 262094)   0.08839433053706158
(0, 614931)   0.1544227435821787
(0, 122638)   0.1544227435821787
(0, 446463)   0.1544227435821787
(0, 648659)   0.09384353736424383
(0, 175642)   0.11122929454454654
(0, 648914)   0.1906457064406779
(0, 103851)   0.0967068039172637
(0, 179272)   0.07192989115457359
(0, 649013)   0.09554553691580105
(0, 614802)   0.11846077082001787
(0, 332673)   0.07348453687864784
(0, 693230)   0.08343430976787648
(0, 505738)   0.13644175720109827
(0, 281202)   0.07832950279440827
(0, 697403)   0.06353745884518128
(0, 423067)   0.07325319735036723
(0, 237726)   0.10141529614895689
(0, 51576)    0.1080515432236505
(0, 445101)   0.06548902023657903
(0, 648476)   0.08674247611454806
(0, 103365)   0.11444885176080954
(0, 651659)   0.11673633535111942
(0, 280013)   0.07419884656404771
```

图 7 部分特征值计算结果

3.3.3 分类器模型

主要的文本分类有三种方法：一是基于统计的分类设计，主要包括朴素贝叶斯，k 最近邻，支持向量机；二是基于规则的分类设计，主要包括决策树，关联规则；三是基于连接的分类设计，主要指人工神经网络。本文主要从三个角度：logistic 回归，多项式朴素贝叶斯以及线性支持向量机进行文本分类器的设计。

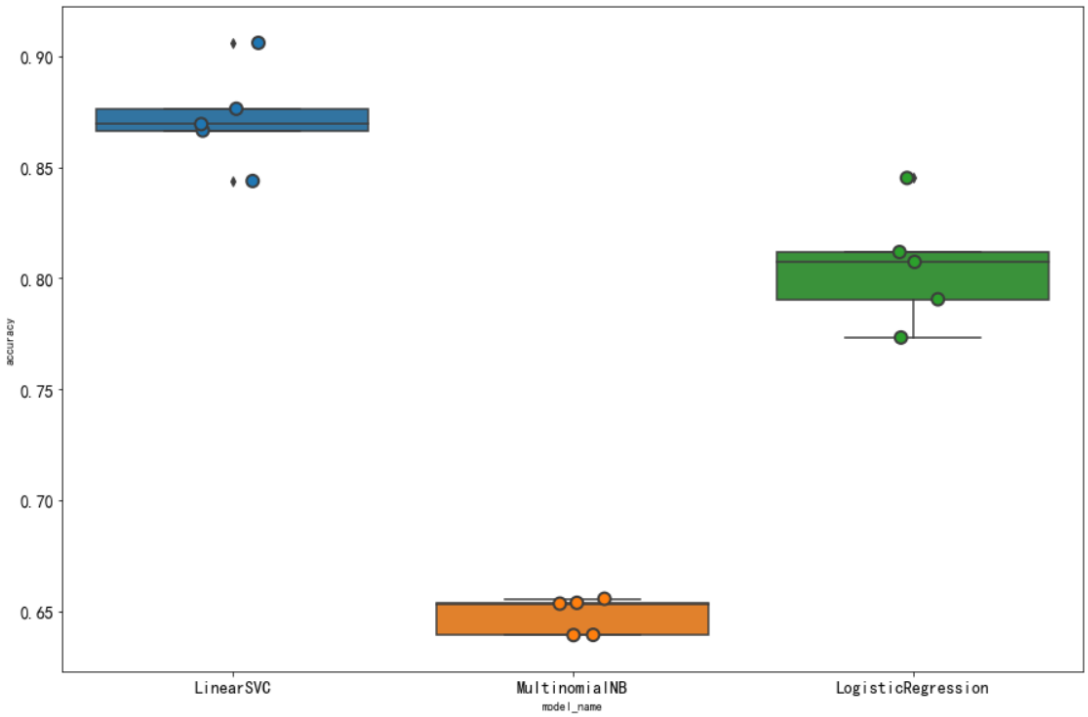

图 8 五折交叉验证

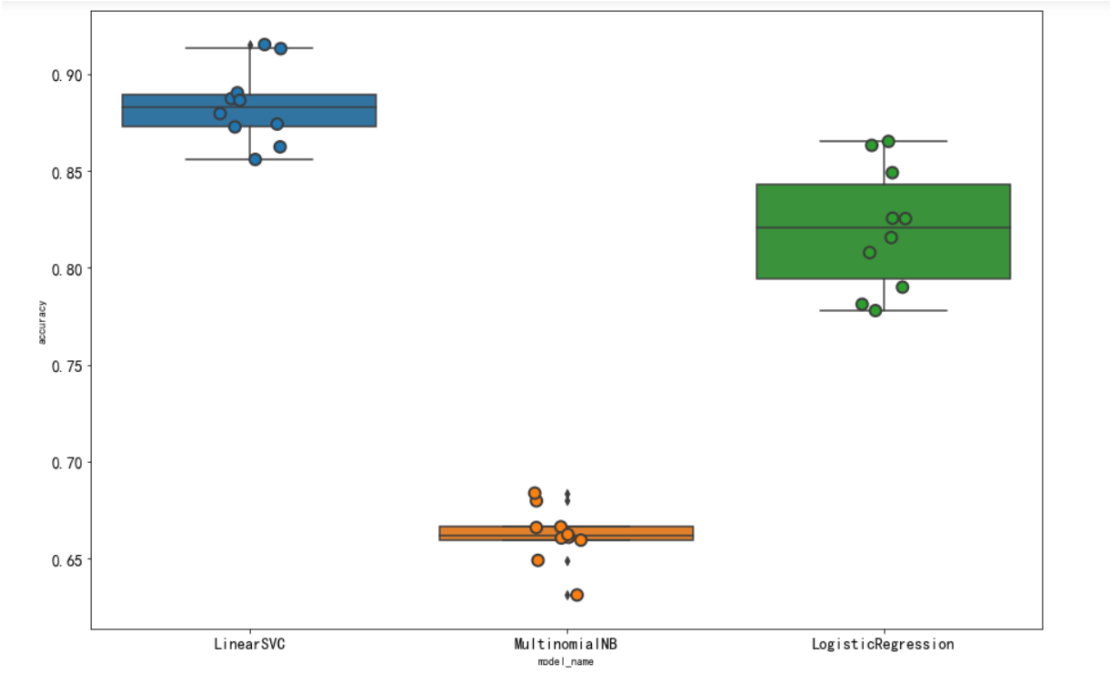

图 9 十折交叉验证

由上箱型图可知，在十折交叉验证下，支持向量机的准确度最高，均值已经超过了 85%。朴素贝叶斯的准确率的均值最低，logistic 回归模型虽然也有将近 85%的准确度，但方差过大效果并不好。因此这三种模型中，直观上看效果最佳的为支持向量机模型，下面以热力图的形式给出其混淆矩阵。

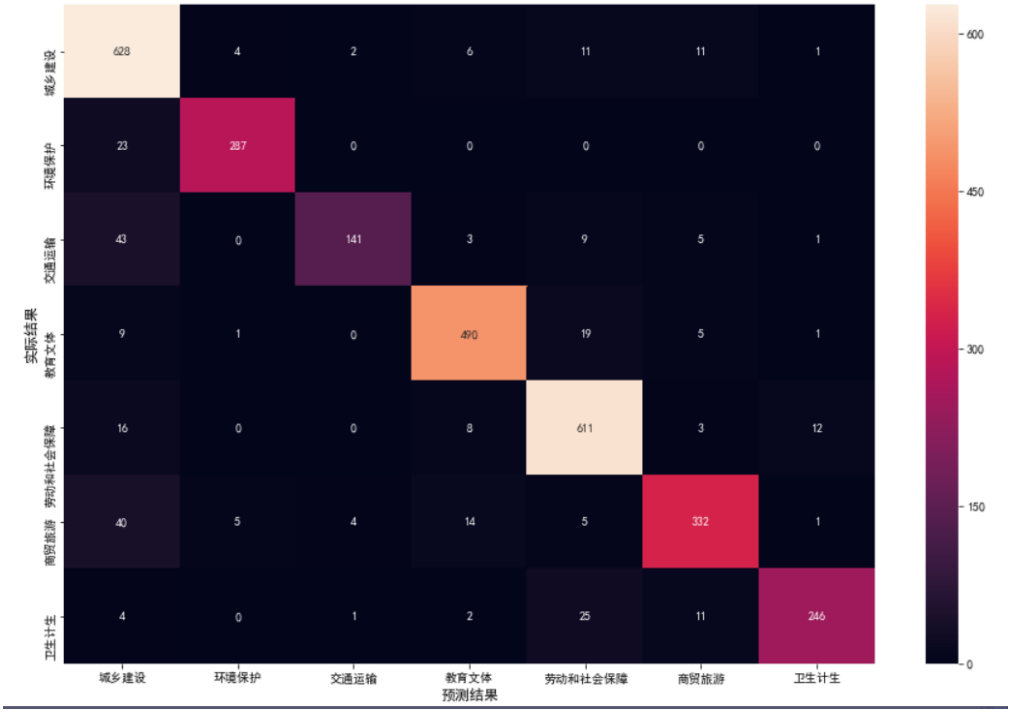

图 10 支持向量机混淆矩阵

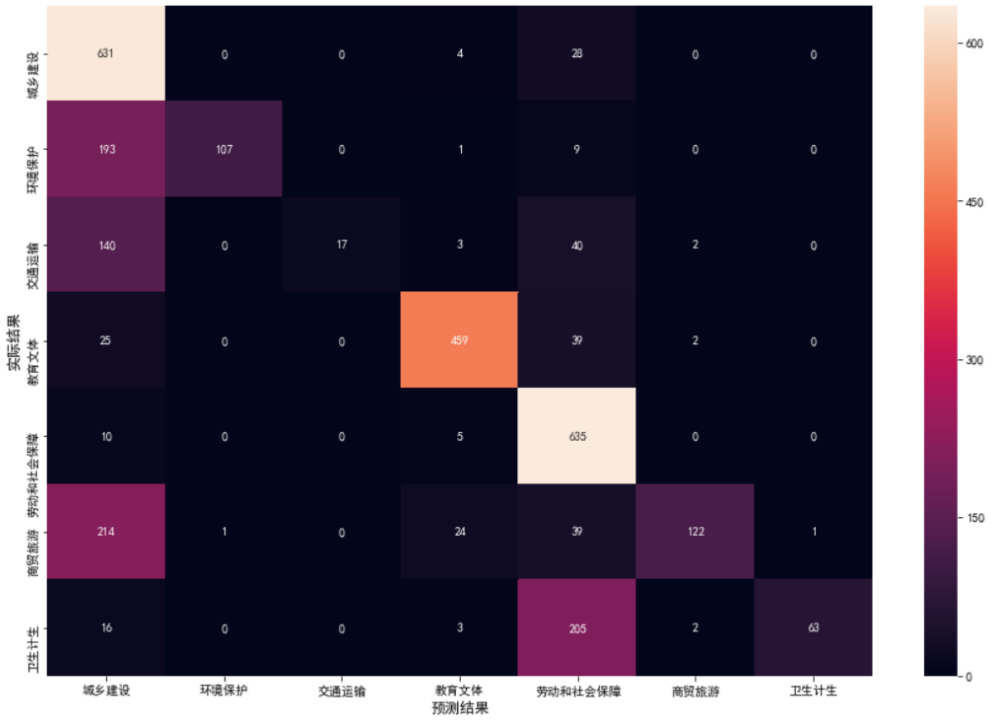

图 11 朴素贝叶斯混淆矩阵

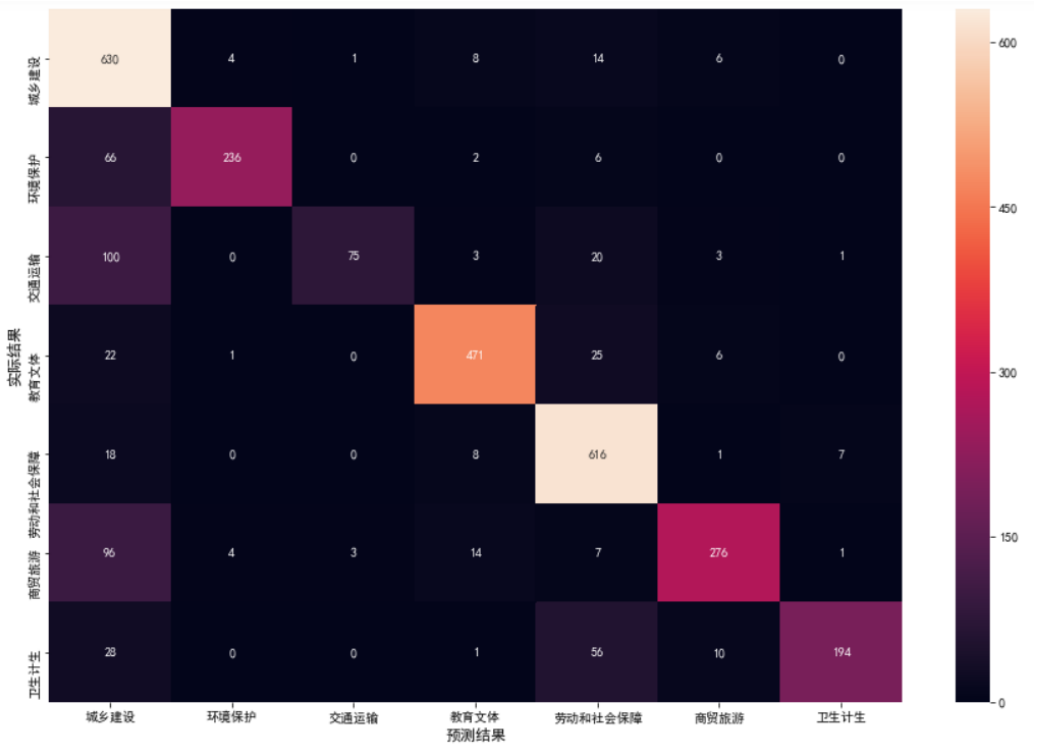

图 12 Logistic 混淆矩阵

在上示三张矩阵热力图中，主对角线为对类别的判断正确的情况，而三种方法比较之下，支持向量机的预测效果较为良好。

2.3.4 F-score

精确率(Precision)和召回率(Recall)是常用的评价模型性能的指标，精确率，即正确预测为正的占全部预测为正的比例；召回率，即正确预测为正的占全部实际为正的比例。一般情况下，Precision 高，Recall 就低，Recall 高，Precision 就低。因此我们需要综合权衡这 2 个指标，这就引出了一个新的指标 F-Score，这是综合考虑 Precision 和 Recall 的调和值，使用 F-Score 对分类方法进行评价。计算公式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2P_iR_i}{P_i + R_i}\right)$$

支持向量机，朴素贝叶斯，logistic 回归三种方法下各类 F-score 如下图所示。

	precision	recall	f1-score	support
城乡建设	0.82	0.95	0.88	663
环境保护	0.97	0.93	0.95	310
交通运输	0.95	0.70	0.81	202
教育文体	0.94	0.93	0.94	525
劳动和社会保障	0.90	0.94	0.92	650
商贸旅游	0.90	0.83	0.86	401
卫生计生	0.94	0.85	0.89	289

图 13 支持向量机 F-score

	precision	recall	f1-score	support
城乡建设	0.51	0.95	0.67	663
环境保护	0.99	0.35	0.51	310
交通运输	1.00	0.08	0.16	202
教育文体	0.92	0.87	0.90	525
劳动和社会保障	0.64	0.98	0.77	650
商贸旅游	0.95	0.30	0.46	401
卫生计生	0.98	0.22	0.36	289

图 14 朴素贝叶斯 F-score

	precision	recall	f1-score	support
城乡建设	0.66	0.95	0.78	663
环境保护	0.96	0.76	0.85	310
交通运输	0.95	0.37	0.53	202
教育文体	0.93	0.90	0.91	525
劳动和社会保障	0.83	0.95	0.88	650
商贸旅游	0.91	0.69	0.79	401
卫生计生	0.96	0.67	0.79	289

图 15 Logistic F-score

类型	F-score
支持向量机	0.893
朴素贝叶斯	0.547
logistic 回归	0.790

表 1 三种方法的 F-score 对比

由上述可知，三种方法比较之下，支持向量机的方法模型拥有最高的 F-score。

四、热点问题挖掘问题的分析与解决

4.1 问题二数据分析

在第一个问题中，我们主要计算文本的 TF-IDF 值，使文本数据得以量化，并通过分类器将标签与内容相对应的一种机器有监督的学习方法。而在第二个问题中，从原本的有监督分类问题转变成无监督的机器学习过程。第二问的关键在于找出热点问题。其中题干给与的信息只有点赞数和反对数。因此热度的定义主要与一段时间内留言的条数，“点赞”与“点灭”的个数相关。因此问题的关键转变成对相关主题的评论进行归类。我们尝试了多种方法对文本数据进行分析，包括 k-means 聚类多种聚类模型，但在计算模型的残差平方和以及轮廓系数，分类的效果不佳[2]。因此下文重新对数据结构进行分析。

4.1.1 残差平方和

$$SSE = \sum_{i=1}^{k} \sum_{p \epsilon C_i} |p - m_i|^2$$

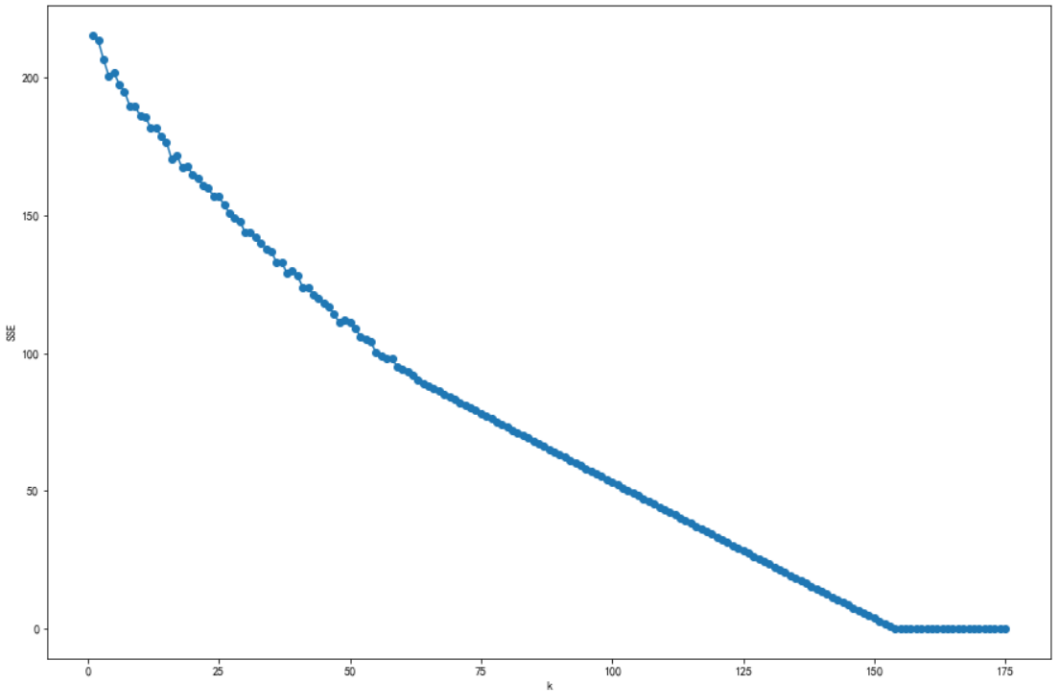

图 16 sse 随聚类个数的变化图（横轴为聚类簇数）

4.1.2 轮廓系数

轮廓系数（Sihouette Coefficient），某个样本点X_i的轮廓系数定义如下：

$$S = \frac{b - a}{\max(a, b)}$$

其中，a 是X_i与同簇的其他样本的平均距离，称为凝聚度；b 是X_i与最近簇中所有样本的平均距离，称为分离度。最近簇的定义为：

$$C_j = \arg\min_{C_k} \frac{1}{n} \sum_{p \epsilon C_i} |p - X_i|^2$$

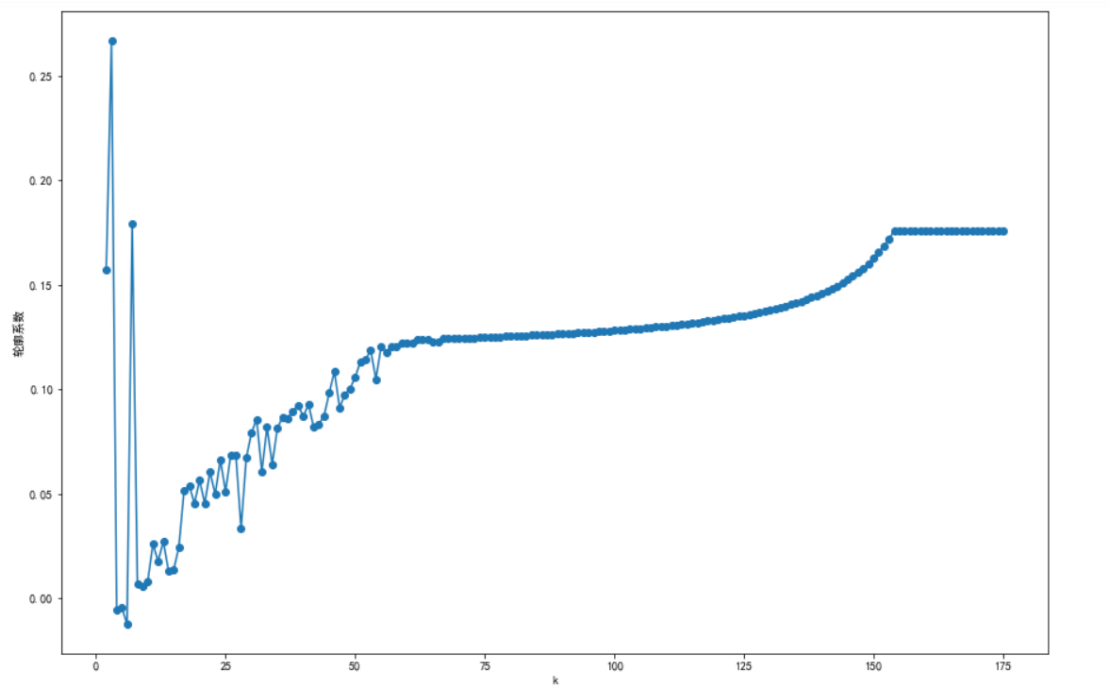

图 17 轮廓系数随聚类个数的变化图（横轴为聚类簇数）

4.2 数据类型

第二问的关键在于如何正确的将各留言主题分好类别，而分类的关键是在地点（人群）。考虑到评论中包含了地点和人群信息，我们尝试从数据本身的角度先对四千多条评论进行划分。通过对划分后的数据进行词频的统计，可以达到较为精准的关键词预测。

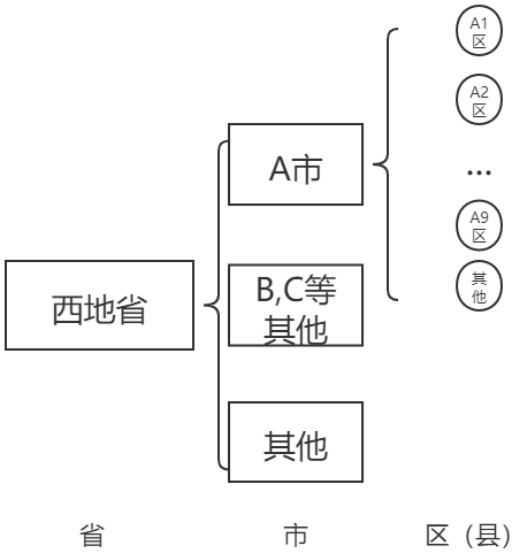

图 18 数据分类指标

4.3 热度评价体系的建立

通过聚类我们可以得到四类指标：同类问题留言数、同类问题留言点赞总数、同类问题留言反对总数、同类问题留言时间跨度。要及时发现热点问题，不仅需要考虑问题的反馈程度，即留言数与点赞反对数，同时也要考虑时间的跨度，否则长期累积下来的问题热度容易埋没短期内发生的热点问题，无法及时发现当下热点问题，因此四类指标又可分为反馈维度和时间维度两个维度[3]。

4.3.1 反馈维度

反馈维度包括同类问题留言数(N_1)、同类问题留言点赞总数(N_2)、同类问题留言反对总数(N_3)，是表征内容热度的最直接指标，一般采用对各项数据加权的方式来计算。考虑到点赞数与反对数是两项相反指标，在计算原始热度值时将两项指标之差，记为净点赞数作为指标进行加权。

对内容的原始热度值 H_a 根据上述指标采用线性加权的方式计算，计算方法如下 ：

$$Ha = \alpha \cdot N1 + \beta \cdot (N2 - N3)$$

α 为同类问题留言数对原始热度值影响权重，β 为净点赞数对原始热度值影响权重。考虑到短期内热点问题留言数往往高于净点赞数且更能详细地反映问题，因此留言数权重 α 应该大于净点赞数 β ，设置 α =4， β =2，代入公式如下：

$$Ha = 4N1 + 2(N2 - N3)$$

4.3.2 时间维度

时间维度用来控制热度随时间的衰减，距离当下比较久远的内容应该具有更低的热度值[4]。引入牛顿冷却定律作为时间衰减函数，即

$$H(t) = Ha \cdot exp[-\gamma \cdot \Delta t]$$

Δt 为该问题时间跨度，即第一条留言与最后一条留言的时间跨度，单位为天，即 24h。

γ 为“重力因子”，即该值越大，问题热度随着时间跨度增加下落越快。将 γ 设置为 0.005，代入公式如下：

$$H(t) = Ha \cdot exp[-0.005 \cdot \Delta t]$$

综上可得热度评价指标计算公式如下：

$$H = [4N1 + 2(N2 - N3)] \cdot exp[-0.005 \cdot \Delta t]$$

4.4 结果分析

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	37.2	2019/1/3 至 2019/1/7	西地省山千制药机械股份有限公司	西地省山千制药机械股份有限公司拖欠工资近半年
2	2	21	2019/4/17 至 2019/10/30	A 市 A5 区魅力之都	魅力之城小区一楼的夜宵摊严重污染附近的空气
3	3	14.9	2019/7/3 至 2020/1/26	A 市 A2 区丽发新城小区	A 市 A2 区丽发新城小区附近搅拌站明目张胆污染环境
4	4	11.6	2019/1/23 至 2019/7/18	A 市 A4 区凯乐国际城小区	A4 区凯乐国际城周边交通乱象
5	5	11.5	2019/3/26	A 市 A1 区东成大厦	请公开 A1 区东成大厦的相关规

			至 2019/7/15		划文件
6	6	11	2019/3/16 至 2019/10/28	A 市 A3 区梅溪湖街道	A3 区梅溪湖街道区牛聋子粉馆非法营业，噪音震人
7	7	7.3	2019/3/17 至 2019/8/22	A 市 A6 区二中	A6 区第二中学有偿补课
8	8	6.8	2019/1/14 至 2019/4/2	A 市 A7 区松雅小区	A7 县松雅小区 D 区马路市场问题何时得以解决
9	9	1.7	2017/6/8 至 2019/11/22	A 市经济学院	A 市涉外经济学院寒假过年期间组织学生去工厂工作

表 2 热点排名前 9 的问题

五、答复意见的评价问题的分析与解决

5.1 评价指标提取

指标提取是对评论数据进行分析的另一项重要任务,利用关联规则产生一系列名词和名词短语，然后对它们进行剪切，将最终得到的集合作为特征列表，本问题中根据研究目标，需提取的主要特征指标有相关性、可解释性、完整性和信息量等[5]。对以上特征指标可以通过如下方式获得：

（1）通过简单计算可以获得信息量、目标值(点赞数)和完整性指标，其中信息量就是下文的评论长度，一般以评论内容的字或词数为计;目标值(点赞数)作为在线评论质量的衡量，可以从网页上直接获取;完整性指标衡量在线评论的内容是否完整，即是否包含文本型和数值型两类评论，若包含两类数据则将该评论的完整性标注为 1，否则标为 0;

（2）改进公式可以获得可解释性指标，关于中文留言可解释性的研究，至今没有统一的定论和成型的算法，而英文可读性研究相当成熟，本文则借鉴英文可读性的衡量公式自动化可读性指数(Automated Readability Index，ARI)的值表示，即

ARI = 4.71*(总字符数/总字数)+ 0.5*(，急字数/总句数)一 21.43

其中字符数可按照中文一个汉字相当于两个字符来计算，暂时获得评论的可读性指数。

（3）利用编程处理可得到相关性(主题相关度)

5.2 信息量的判定与结果

从信息质量评价角度出发，回复信息质量评价指标分为内在特质、内容特质和描述特质三个方面被广泛采纳。针对研究对象具有情感丰富、内容随意与形式多变的特点，进行重新整合相应指标，抽取文本特征并分别归属到主客观、信息量和语义特征(一致性)三个类别，建立回复质量评价指标体系。本研究则进一步参考信息质量评价指标并结合本文数据特征，把答复内容特征归属到相关度和信息量中。

其中，信息量可以理解为评论信息的含量，如句量和词量，也可以表示为平均句长或单句长度。通常认为，评论的信息含量越大，评论的质量就越高，本研究中取评论长度代表信息含量，模型中用 Words 表示。

留言详情	答复意见	答复时间	Words
业公司却以交20万保证金，不	收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	83.4
后面的生意带来很大影响，里	需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	55.4
同时更是加大了教师的工作压	办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	81.25
落户A市，想买套公寓，请问	年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	45
“马坡岭小学”，原“马坡岭	保留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	46.33333
再把泥巴冲到右边，越是上下	于您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	105.5
为老社区惠民装电梯的规范	A市A3区人民政府办公室下发了《关于A市A3	2019/5/9 10:18:58	56
跑好远，天寒地冻的跑好远，	修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	48
也没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	58.5
立交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	67.66667
规划局审批通过《温室养殖	司支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	56.125
区安置房地下室近两万平方米	续，按长人防发[2014]7号文件要求，鄱阳	2019/1/29 10:52:01	74.8
峰，大量从小区开车出去的业	分局配合进行具体选址，招标（邀标）进行	2019/1/14 14:34:58	61.5
贵省相关政府部门的大力支持	映的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	45
小时以上！天寒地冻，其他公	正常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	62.33333

图 19 信息量判定的部分结果

5.3 相关性的判定与结果

留言答复的好坏首先需要通过文本相关性的检验，在分析文本相关性时，与前面两个问题的步骤相同。仍需要对原先的数据结构进行剔除停用词，分词两个步骤。通过词袋空间将切分好的词量化，将文本转化成二元组的向量，通过计算 TF-IDF 值对语料库建模，从而计算出测试文本和训练文本的相似程度。

在第三问的相关性问题中，“答复意见”与“评论主题”以及“评论详情”的相关程度是我们关心的。因此我们将“评论主题”和“评论详情”作为 Training_data，而我们所关心的“答复意见”为 Test_data。将相关程度的阈值设置为 0.9，就可以得到相关性较差的评论回复数量以及具体内容，其值记为 Relevancy。通过计算，可以得到 2817 组回复中，

共有 121 条相关性较差的回复。

['J市二手房收税问题咨询',
 '麻烦相关部门告知二手房具体税收规定家人有一套168平米的房子出售满两年非唯一住房房龄满了十年请问买方和卖方大概要交多少税麻烦告知具体计算标准比如多少平方以上怎么收百分之几二手房交易需要收哪些税怎么收',
 '您的留言已收悉我们已将您反映的问题转相关部门进行处理敬请关注后续回复谢谢']

图 20 留言编号为 90210 的答复意见被视为相关性较差回复

5.4 完整性判定与结果

留言答复的完整性主要体现在答复整体信息的结构上，简单的来说，完整性(Complete)可以理解为一条评论即包含数值型评论也包括文本型评论，完整的数据可以标记为 1，不完整则记为 0;

通过计算，可以得到 2817 组回复中，共有 263 条是属于不完整的答复。

12374	UU008582	省长株潭两型社会建设的	2014/4/15 11:22:02	任务定时间和质量	网友：您好！留言已收悉	2014/5/9 17:31:55	0
12383	UU008795	建议取消绕城高速收费	2014/4/13 17:26:33	上，每天来往均要	网友：您好！留言已收悉	2014/5/16 15:57:26	0
12387	UU0081236	县金洲新区高新安置房分	2014/4/10 13:38:43	经有三年多了，未	网友：您好！留言已收悉	2014/5/16 15:55:37	0
12415	UU0081509	A3区含浦镇白鹤社区的用地	2014/3/27 17:16:36	寻这些难道都不是	网友：您好！留言已收悉	2014/5/9 17:28:09	0
12451	UU0081019	书记增加A市交通辅警工资	2014/3/11 17:02:28	很低，住房公积金	网友：您好！留言已收悉	2014/4/28 16:06:58	0
12452	UU0081171	于A市旅游发展的看法和建	2014/3/11 11:56:22	地现状也亟需集	网友：您好！留言已收悉	2014/4/28 16:05:42	0
12458	UU00854	市镇头镇连山村交通不便	2014/3/7 22:39:35	不断的扩宽提质，	网友：您好！留言已收悉	2014/4/14 12:13:50	0
12614	UU0081308	黄兴小学问题塑胶跑道铲除	2013/11/3 22:34:11	毒中毒，发生皮炎	容易对草坪造成根本性的破坏。如果使用沙质或土质的	2013/12/27 11:32:10	1
12649	UU008205	决北京御园小区电梯安全	2013/10/18 16:49:10	联系，多次反映问	分局联系，组织技术人员对电梯进行了现场检查和故障	2013/11/22 16:50:59	1
12663	UU0081891	请求解决异地网上认证	2013/10/13 15:19:48	认证是一件解决	人员，您在网上认证不成功，您可前往西地省社会保险	2013/12/11 15:42:44	1
12669	UU0081548	采取有力措施治理好A市的	2013/10/9 17:40:51	污染了。 1，	加快淘汰黄标车；全面启动储油库、加油站和油罐车的	2013/12/27 11:23:51	1
12701	UU0082056	在泉塘物流园中建一所公立	2013/9/6 17:47:22	海德公园、A1区	宅小区以南，总投资1225.52万元，占地面积2714.7平	2013/10/17 17:23:25	1
12708	UU0081902	A5区华盛世纪新城业主无	2013/9/3 17:07:48	场、山水熙园、	步行300米至洞井路高架桥下137路（劳动广场—洞井路	2013/11/22 16:39:16	1
12712	UU0081942	决清水塘三小学生出行安	2013/9/2 20:32:40	路车多速度又快	核后认为，由于该路口南口为陈路渔规划路，为避免重	2013/10/31 16:14:42	1
12724	UU008826	星沙镇板桥小区一条街扰	2013/8/28 14:49:19	时候 就必须赶紧	前三包责任书》立即进行整改，杜绝扰民事件的再次	2013/9/27 18:14:35	1
12745	UU008197	都小学学生们一个安全出	2013/8/20 14:34:36	的汽车车流量实	后，区市政局决定在9月上旬，对原有的热溶标线减速	2013/10/17 17:21:32	1
12755	UU0081083	镇玉江村农户希望热天能	2013/8/14 12:04:22	开，真的是折磨	农网改造时设计的标准跟不上发展需要，需要进行再次	2013/10/31 16:06:55	1

图 21 完整性判定的部分结果

5.5 可解释性判定与结果

可解释性(accountable)本来指的是可以追踪到数据的来源，广义上的可解释性指在我们需要了解或解决一些事情的时候，我们可以获得所需要的足够的可以理解的信息，如果在一些情境中我们无法得到相应的足够的信息，那么这些事情都我们来说是不可解释的。在留言答复中，答复中能够找到关键的可以理解的信息去支撑此回复，则认为此回复是具有可解释性的。

对于留言回复而言，可解释性对于中文来说没有多大意义，我们可以理解为可读性，留言回复的可读性或读者的理解能力可以用自动化可读性指数 ARI （ Automated Readability Index ）来表示，在过去的信息科学研究中用可读性来定性地检验一些(英文)文本的特征，表示某文本内容吸引人的程度，而 ARI 依赖于文本的字符数，又比其他的可读性测试方法准确率高，因而经常被用于类似的研究中。计算可以直接采用公式:

$$ARI = 4.71*(总字符数/总字数) + 0.5*(总字数/总句数) - 21.43$$

其数值近似等于我们可能理解一段文字的最低程度。国内关于中文可读性的研究尚不成熟，没有现成的公式可用，因此本文仅借鉴英文可读性计算公式作为参考。

答复意见	答复时间	ARI
收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	25.39791
需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	11.4391
办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	24.35426
年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	6.460333
保留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	7.192134
于您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	36.47645
A市A3区人民政府办公室下发了《关于A市A3	2019/5/9 10:18:58	11.70054
修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	7.664323
单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	12.89231
部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	17.57737
司支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	11.75161
续，按长人防发[2014]7号文件要求，鄱阳	2019/1/29 10:52:01	21.33487
分局配合进行具体选址，招标（邀标）进行	2019/1/14 14:34:58	14.60439
映的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	6.303333
正常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	15.00078

图 22 可解释性指数 ARI 的部分结果

5.6 构建留言答复评价指标体系

本文运用的留言回复质量评价指标体系——1W2R3C 指标，包括信息量（Words）、相关度（Relevancy）、可解释性（Readability），完整性（Complete），能较全面地衡量留言回复的质量。

指标	说明
信息量（Words）	从内容上确保回复质量，以回复长度衡量(词/字数统计)
相关度（Relevancy）	回复内容与留言主题的相关性
可解释性（Readability）	用于检测回复的被理解程度，由 ARI 表示
完整性（Complete）	衡量回复是否包含数值型和文本型数据

表 3 留言答复评价指标及说明

5.7.确定评价指标权重

关于留言回复质量评价模型的研究，不同的指标权重将会得到完全不同的实验结论，因此权重的选择直接影响到最终模型的成败。大多研究中用到的简单方法有问卷调查法和专家打分法。为全面验证最终模型的性能，本文将模型回归分析得出的权重与专家打分法得出的

权重进行比较分析，且两者之间进行建模对比分析。

5.7.1 专家打分法确定权重 W1

专家打分法的原理同专家调查法类似，也称为德尔菲法专家打分法，指通过匿名的方式去征询相关专家的意见,然后对收集到意见进行统计、分析与归纳,并经过多轮的意见征询、反馈和调整后，最终确定预测结果。该方法因具有易于确定评价等级和标准、直观性强、计算方法简单、可以将能够进行定量与无法定量计算的评价项目都加以考虑等的特点，而备受大多研究者的欢迎。

由翻阅文献得到各种指标的专家打分法的权重如下表所示

指标	权重
信息量（Words）	0.22
相关度（Relevancy）	0.39
可解释性（Readability）	0.18
完整性（Complete）	0.21

表 4 留言答复的指标权重

5.7.2 灰色关联分析法修正的权重 W2

为了尽可能使指标权重达到科学、合理的标准，不产生过大偏差，我们认为通常研究中采用专家打分法，由于在专家选择的时候主观性很大，专家的意见也带有很强的主观性，因而得到的结果可能产生较大的偏差，另外在多因素研究中，各因素并不是独立存在的，它们之间的相互影响也不可忽略，于是本文运用灰色关联度来进一步修正各指标权重。

灰色关联分析可从多角度，多视角对物品的质量进行评价，由以往的医学领域中的广泛应用可知，该方法具有操作性强，效果好等优点。本研究考虑了各指标间可能存在某种相关性，引入灰色关联度，来修正专家打分的结果，大大降低权重的主观性[6]。

对于灰色关联分析中参考列的选择，在实际应用中会选择数据行中最大值或最小值，来求的最优答案;如果要确定指标的权重，选择最重要的指标即可，此时多为主观决断。在本文构建的指标体系中，选择相关性作为参考列。

引入关联度r_i后，修正专家打分方法得到的权重 W2，此时的在线评质量评价指标的权重变为:

$$W2 = W1 * r_i$$

其中，以相关性为参考列得出的关联度系数如下所示

Relevancy	Words	Complete	Readability
	0.499342751	0.485421885	0.491719958
	0.496749967	0.479276909	0.481441377
	0.499499455	0.491787956	0.496202716
	0.495665413	0.484859913	0.483810238
	0.469864404	0.497043199	0.451063812
	0.471462123	0.454967206	0.482314828
	0.498255249	0.486068597	0.488335985
	0.499633458	0.467307042	0.477779512
	0.468509397	0.49646034	0.448514223
	0.443234583	0.474027448	0.42564994

图 23 以相关性为参考的部分关联度

一般研究中为了方便计算，本文计算关联系数的均值作为关联度，因此，表中各个指标与相关性的关联度取值分别为：0.47906365、0.477893742、0.468160546。则最终修正后的指标权重为 W2=（0.105394003，0.39，0.086020873，0.098313715）。

5.7.3 留言回复质量评价模型

根据上文提取的指标特征，用上述方法构建本文的留言答复质量评价模型。在此用 Quality 表示留言答复的质量，得分记为 Q，可建立如下模型：

$$Q = 0.1054 * Words + 0.39 * Relevancy + 0.086 * Readability + 0.0983 * Complete$$

其中， Words 表示信息量，Relevancy 表示相关度， Readability 为可读性，Complete 为完整性。

留言编号	留言详情	答复意见	答复时间	质量得分
2549	业公司却以交20万保证金，不	收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	11.6212049
2554	面的生意带来很大影响，里	需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	7.29536248
2555	同时更是加大了教师的工作	办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	11.2102931
2557	落户A市，想买套公寓，请问	年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	5.6732929
2574	“马坡岭小学”，原“马坡岭	保留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	5.90452084
2759	再把泥巴冲到右边，越是上下	于您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	14.9828933
2849	为老社区惠民装电梯的规范	A市A3区人民政府办公室下发了《关于A市A3	2019/5/9 10:18:58	7.37318017
3681	好远，天寒地冻的跑好远，	修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	6.03036733
3683	没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	7.63463104
3684	立交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	9.14513807
3685	规划局审批通过《温室养殖	同支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	7.33441429
3692	区安置房地下室近两万平方米	续，按长人防发[2014]7号文件要求，鄱阳	2019/1/29 10:52:01	10.2881175
3700	峰，大量从小区开车出去的业	分局配合进行具体选址，招标（邀标）进行	2019/1/14 14:34:58	8.20097475
3704	省相关政府部门的大力支持	映的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	5.68710486
3713	小时以上！天寒地冻，其他公	正常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	8.21722735
3720	址：https://baidu.com/。	电的“披塘路路口两端各拆除20米中间花坛，	2019/3/6 10:26:14	8.8551901
3727	便以各种理由拒绝退货，并	根据您提供的信息进行投诉信息的登记分送	2019/1/3 14:02:47	5.51425376

图 24 留言答复质量评价的部分结果

六、总结与改进

6.1 总结

本文主要运用了文本数据处理和分类方法完成群众留言分类和热点问题挖掘的问题，运用专家打分法和灰色关联度修正法构建了答复质量的评价模型。

首先，我们对原始数据进行属性特点的分析，为留言数据制定多种的预处理方式，尽可以全面地完成文本数据的量化处理，进而有利于我们对数据进行针对性的操作。

其次，在完成数据预处理之后，我们将要分类的七个一级标签数字化，通过剔除停用词，分词等一系列方式将文本数据量化，通过计算其 TF-IDF 值进而使用多种分类器模型进行 F-score 的比较，并且以热力图的形式更直观地表现出模型间的对比效果。

最后，我们对答复质量评价进行指标的提取，得到信息量、相关性、完整性和可解释性这四个指标，接着对每个指标进行量化处理并计算出具体数值，然后通过专家打分法得到指标的权重，并用灰色关联度修正权重得到最终的质量评价模型。

6.2 改进

首先，在文本分类器的设计上，本文主要基于统计的分类思想。其优势在于使用特征表示文本，但同时也忽略了文本语言结构。当然，在查阅相关文献后，有许多机器学习方法如基于规则分类的决策树，以及基于连接分类的人工神经网络等等。相关的有监督的机器学习方法值得我们在以后的研究中进一步深讨。

其次，在留言答复质量评价模型中，本文基本实现对答复质量评价模型的研究，完成指标提取，构建模型等过程，但由于技术水平有限，处理过程中仍处在一些不足。例如在指标提取中，我们只选取了信息量，相关度，可解释性以及完整性四个指标，也导致在最终计算灰色关联度的时候，指标之间缺少足够的解释力。将文本有效性和一致性加入模型，应该可以增强关联系数获得更加精确有效的模型。但由于水平和时间受限，无法将这两个指标量化加入模型中，这在以后的研究中还需要进一步的完善。

再者，评价模型的权值的设定只用了原始数据作为实验对象，针对问题下手。模型在经过反复训练如 Bootstrap 重抽样进行重复实验，来充分验证模型的适用性。。

七、参考文献

[1]孙刚. 基于线性回归的中文文本可读性预测方法研究[D].南京大学,2015.

[2]李玮瑶,赵凯.基于特征提取的网络热点事件挖掘算法[J].计算机与现代化,2015(05):17-20.

[3]詹士昌. 牛顿冷却定律适用范围的探讨[J]. 大学物理, 2000, 019(005):36-37.

[4]littleMagic. 设计一个属于自己的内容热度值算法[EB/OL]. https://www.jianshu.com/p/fb454a4b383d ,2019-4-15.

[5]郭银灵. 基于文本分析的在线评论质量评价模型研究[D].内蒙古大学,2017.

[6]郑军伟. 基于灰色系统理论的数据关联度建模及其应用[D].杭州电子科技大学,2011.

八、附录

8.1 留言分类问题代码实现

```
import pandas as pd
import matplotlib
import os
import numpy as np
import matplotlib.pyplot as plt
import jieba as jb
import re
from sklearn.feature_extraction.text import TfidfVectorizer
#读文件
os.chdir(r'C:\Users\Administrator\Desktop')
data1=pd.read_excel('附件 2.xlsx')
df1=pd.DataFrame(data1)
data=df1['留言详情'].tolist()
price1 = [x.replace('\n', '').replace('\t', '').replace('\r', '').replace('\u3000', '').replace('\xa0','').replace(' ','') for x in data ]
df = pd.DataFrame(list(zip(price1,df1['一级标签'])),columns=['留言详情','cat'])
```

```
#部分可视化分析
#类别
d = {'cat':df['cat'].value_counts().index, 'count': df['cat'].value_counts()}
df_cat = pd.DataFrame(data=d).reset_index(drop=True)
#图示
plt.rcParams['font.sans-serif']=['SimHei'] #用来正常显示中文标签
df_cat.plot(x='cat',  y='count',  kind='bar',  legend=False,  figsize=(8,
5),color='red')
plt.xticks(rotation=45)
plt.title("一级标签",fontsize=15)
plt.ylabel('数量', fontsize=15)
plt.xlabel('类', fontsize=15)
#将标签数字化
df['code'] = df['cat'].factorize()[0]
cat_id_df                                =                                df[['cat',
'code']].drop_duplicates().sort_values('code').reset_index(drop=True)
cat_to_id = dict(cat_id_df.values)
id_to_cat = dict(cat_id_df[['code', 'cat']].values)
df.sample(10)
d2=df
d2.columns = ['留言详情','cat','code']
#加载停用词并剔除
def remove_punctuation(line):
    line = str(line)
    if line.strip()=='':
        return ''
    rule = re.compile(u"[^a-zA-Z0-9\u4E00-\u9FA5]")
    line = rule.sub('',line)
    return line

def stopwordslist(filepath):
    stopwords = [line.strip() for line in open(filepath, 'r', encoding='utf-
8').readlines()]
    return stopwords
stopwords = stopwordslist(r"C:\Users\Administrator\Desktop\StopWords.txt")
df['clean_review'] = df['留言详情'].apply(remove_punctuation)
#最终切词完毕的 dataframe
df['cut_review'] = df['clean_review'].apply(lambda x: " ".join([w for w in
list(jb.cut(x)) if w not in stopwords]))
#计算 TF-IDF
tfidf = TfidfVectorizer(norm='l2', ngram_range=(1, 2))
features = tfidf.fit_transform(df.cut_review)
labels = df.code
print(features.shape)
```

```
print('----------------------------')
print(features)

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC

from sklearn.model_selection import cross_val_score

%线性向量机，朴素贝叶斯，logistic 回归三种模型
models = [
    LinearSVC(),
    MultinomialNB(),
    LogisticRegression(random_state=0),
]
CV = 10
cv_df = pd.DataFrame(index=range(CV * len(models)))
entries = []
for model in models:
    model_name = model.__class__.__name__
    accuracies = cross_val_score(model, features, labels, scoring='accuracy',
cv=CV)
    for fold_idx, accuracy in enumerate(accuracies):
        entries.append((model_name, fold_idx, accuracy))
cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])
#绘制准确率的比较图
import seaborn as sns
plt.figure(figsize=(15, 10))
sns.boxplot(x='model_name', y='accuracy', data=cv_df)
sns.stripplot(x='model_name', y='accuracy', data=cv_df,
              size=10, jitter=True, edgecolor="gray", linewidth=2)
plt.yticks(size = 14)
plt.xticks(size = 14)
plt.show()
#以热力图的形式绘制混淆矩阵（以线性向量机为例）
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
model = LinearSVC()
X_train,    X_test,    y_train,    y_test,    indices_train,    indices_test    =
train_test_split(features, labels, df.index,

test_size=0.33, stratify=labels, random_state=0)
model.fit(X_train, y_train)
```

```
y_pred = model.predict(X_test)
conf_mat = confusion_matrix(y_test, y_pred)
fig, ax = plt.subplots(figsize=(15,10))
sns.heatmap(conf_mat, annot=True, fmt='d',
            xticklabels=cat_id_df.cat.values, yticklabels=cat_id_df.cat.values)
plt.ylabel('实际结果',fontsize=12)
plt.xlabel('预测结果',fontsize=12)
plt.show()
#计算各类的 F-1score（以线性向量机为例）
from sklearn.metrics import classification_report
print('accuracy %s' % accuracy_score(y_pred, y_test))
print(classification_report(y_test,
y_pred,target_names=cat_id_df['cat'].values))
```

8.2 热点问题挖掘代码实现

```
import jieba
from collections import Counter
#读文件
def GetTxtDataFromFile( filename ):
    with open( filename, encoding='UTF-8' ) as fp:
        txtConts = fp.read()
        return txtConts
#剔除停用词，分词
def CutWithStopWord( txtConts, stopWord ):
    cutList = []
    strList = jieba.cut( txtConts )

    for word in strList:
        if not( word in stopWord ) and len( word ) > 1:
            cutList.append( word )
    return cutList
#统计词频
def StatTopNWords( words, N ):
    wordList = []
    wordList = Counter( words )
    wordListNTop = wordList.most_common( N )
    return wordListNTop

def main():
    txtConts = GetTxtDataFromFile( r'C:\Users\Administrator\Desktop\text.txt' )
    stopWords = GetTxtDataFromFile( 'Stopwords.txt' )
    cutWordsResults = CutWithStopWord( txtConts, stopWords )
```

```
    wordListNTop = StatTopNWords( cutWordsResults, 50)

    print( 'Initial txtConts:\n', txtConts )
    print( 'cutWordsResults:\n', '/'.join( cutWordsResults ) )
    print( 'wordListNTop:\n', wordListNTop )


if __name__ == '__main__':
    main()
```

8.3 留言答复评价代码实现

8.3.1 相关性计算

```
import pandas as pd
import matplotlib
import os
import numpy as np
import matplotlib.pyplot as plt
import jieba as jb
import re
os.chdir(r'C:\Users\Administrator\Desktop')
data1=pd.read_excel('附件 4.xlsx')
df1=pd.DataFrame(data1)
data=df1['留言主题'].tolist()
price1     =     [x.replace('\n',     '').replace('\t',     '').replace('\r',
'').replace('\u3000', '').replace('\xa0','').replace(' ','') for x in data ]
data2=df1['留言详情'].tolist()
price2     =     [x.replace('\n',     '').replace('\t',     '').replace('\r',
'').replace('\u3000', '').replace('\xa0','').replace(' ','') for x in data2 ]
data3=df1['答复意见'].tolist()
price3     =     [x.replace('\n',     '').replace('\t',     '').replace('\r',
'').replace('\u3000', '').replace('\xa0','').replace(' ','') for x in data3 ]
df = pd.DataFrame(list(zip(price1,price2,price3)),columns=['留言主题','留言详情
','答复意见'])

def remove_punctuation(line):
    line = str(line)
    if line.strip()=='':
        return ''
    rule = re.compile(u"[^a-zA-Z0-9\u4E00-\u9FA5]")
    line = rule.sub('',line)
    return line
```

```
def stopwordslist(filepath):
    stopwords = [line.strip() for line in open(filepath, 'r', encoding='utf-8').readlines()]
    return stopwords

#加载停用词
stopwords = stopwordslist(r"C:\Users\Administrator\Desktop\StopWords.txt")
df['clean_review'] = df['留言主题'].apply(remove_punctuation)
df['clean_review2'] = df['留言详情'].apply(remove_punctuation)
df['clean_review3'] = df['答复意见'].apply(remove_punctuation)
count=0
cor=[]
for i in range(len(reviews)):
    all_doc=[]
    all_doc.append(reviews[i][0])
    all_doc.append(reviews[i][1])
    all_doc_list = []
    for doc in all_doc:
        doc_list = [word for word in jieba.cut(doc)]
        all_doc_list.append(doc_list)
    doc_test_list = [word for word in jieba.cut(reviews[i][2])]
    dictionary = corpora.Dictionary(all_doc_list)
    corpus = [dictionary.doc2bow(doc) for doc in all_doc_list]
    doc_test_vec = dictionary.doc2bow(doc_test_list)
    tfidf = models.TfidfModel(corpus)
    tfidf[doc_test_vec]
    index = similarities.SparseMatrixSimilarity(tfidf[corpus], num_features=len(dictionary.keys()))
    sim = index[tfidf[doc_test_vec]]
    if max(sim)<0.2:
        count+=1
        cor.append(i)
f3 =open(r"C:\Users\Administrator\Desktop\textcorrelation.txt",'w',encoding='utf-8',errors='ignore')
for i in c:
    f3.write(str(i))
    f3.write("\n")
f3.close()
```

8.3.2 完整性判定

```
import pandas as pd
```

```
import matplotlib
import os
import numpy as np
import matplotlib.pyplot as plt
import jieba as jb
import re
os.chdir(r'C:\Users\Administrator\Desktop')
data1=pd.read_excel('附件 4.xlsx')
df1=pd.DataFrame(data1)
data=df1['留言主题'].tolist()
price1     =     [x.replace('\n',      '').replace('\t',      '').replace('\r',
'').replace('\u3000', '').replace('\xa0','').replace(' ','') for x in data ]
data2=df1['留言详情'].tolist()
price2     =     [x.replace('\n',      '').replace('\t',      '').replace('\r',
'').replace('\u3000', '').replace('\xa0','').replace(' ','') for x in data2 ]
data3=df1['答复意见'].tolist()
price3     =     [x.replace('\n',      '').replace('\t',      '').replace('\r',
'').replace('\u3000', '').replace('\xa0','').replace(' ','') for x in data3 ]
df = pd.DataFrame(list(zip(price1,price2,price3)),columns=['留言主题','留言详情
','答复意见'])

def remove_punctuation(line):
    line = str(line)
    if line.strip()=='':
        return ''
    rule = re.compile(u"[^a-zA-Z0-9\u4E00-\u9FA5]")
    line = rule.sub('',line)
    return line

def stopwordslist(filepath):
    stopwords = [line.strip() for line in open(filepath, 'r', encoding='utf-
8').readlines()]
    return stopwords

#加载停用词
stopwords = stopwordslist(r"C:\Users\Administrator\Desktop\StopWords.txt")
df['clean_review'] = df['留言主题'].apply(remove_punctuation)
df['clean_review2'] = df['留言详情'].apply(remove_punctuation)
df['clean_review3'] = df['答复意见'].apply(remove_punctuation)
count=0
cor=[]
for i in range(len(reviews)):
    all_doc=[]
    all_doc.append(reviews[i][0])
```

```
    all_doc.append(reviews[i][1])
    all_doc_list = []
    for doc in all_doc:
        doc_list = [word for word in jieba.cut(doc)]
        all_doc_list.append(doc_list)
    doc_test_list = [word for word in jieba.cut(reviews[i][2])]
    dictionary = corpora.Dictionary(all_doc_list)
    corpus = [dictionary.doc2bow(doc) for doc in all_doc_list]
    doc_test_vec = dictionary.doc2bow(doc_test_list)
    tfidf = models.TfidfModel(corpus)
    tfidf[doc_test_vec]
    index          =          similarities.SparseMatrixSimilarity(tfidf[corpus],
num_features=len(dictionary.keys()))
    sim = index[tfidf[doc_test_vec]]
    if max(sim)<0.2:
        count+=1
        cor.append(i)
f3                  =open(r"C:\Users\Administrator\Desktop\textcorrelation.txt",
'w',encoding='utf-8',errors='ignore')
for i in c:
    f3.write(str(i))
    f3.write("\n")
f3.close()
```

8.3.3 可解释性指数计算

```
#句子数量
N2=[]
for i in range(len(df)):
    a=df['答复意见'][i].count('。')
    a=a+df['答复意见'][i].count('?')+df['答复意见'][i].count('!')
    if a==0:
        N2.append(1)
    else:
        N2.append(a)
#总字符数
N=[]
for i in range(len(df)):
    N.append(len(df['答复意见'][i]))
#总字数
N1=[]
for i in range(len(df)):
    N1.append(len(df['clean_review3'][i]))
```

```
ARI=[]
for i in range(len(N)):
    ARI.append(4.71*N[i]/N1[i]+0.5*N1[i]/N2[i]-21.43)
ARI
#读入 txt
f3      =open(r"C:\Users\Administrator\Desktop\ARI.txt",      'w',encoding='utf-
8',errors='ignore')
for i in ARI:
    f3.write(str(i))
    f3.write("\n")
f3.close()
```