“智慧政务”中的文本挖掘应用

摘要

本文旨在基于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，通过数据分析与挖掘、特征工程以及自然语言处理方法，对各类社情民意相关的文本数据进行留言划分和热点整理。以及通过对答复的相关性、完整性、可解释性进行量化，从而对答复意见的质量给出一套评价方案。

针对任务一，由于数据出现较多专有名词，对分词带来一定的影响，因此首先通过爬虫技术爬取得到专有名词。加载到结巴分词库，通过结巴，对每一条数据的留言详情进行分词，去停用词。建立特征词典、建立语料库、用 TF-IDF 模型处理语料库以及对一级标签进行量化，最后通过机器学习的朴素贝叶斯模型和浅层神经网络对其进行训练。使用 F-Score 对多项式朴素贝叶斯模型评价得分为 0.8566，伯努利朴素贝叶斯模型评价得分为 0.8192，浅层神经网络的测试集精度为 0.8643。因此选取多项式朴素贝叶斯模型或者浅层神经网络模型作为留言内容的一级标签分类模型都能够达到分类目的。

针对任务二，利用爬取的特定地点对每条数据的留言主题进行匹配得到该数据的特定地点，再通过正则表达式提取地级市完成的归类任务。把点赞数量和分类数目作为热度评价的指标。对每个分类里面的数据查找得出点赞数最高的数据进行计算，由于点赞数和分类数目的重要程度不一，需要通过一个权值进行权衡。经过权值计算之后，对热度评价得分前五的分类提取，然后提取出得分前五的每个分类点赞数最高的数据进行留言主题提取，通过该主题进行文本相似度的计算，得出与该主题相似的所有数据。最后汇总在热点问题表和热点问题留言明细表中。

针对任务三，通过满意度和重要度作为答复意见质量的评价指标。满意度由完整性和时效性两个方面进行权衡并进行量化。而对于重要度，则对群众留言内容进行情感分析并打分，该分数取绝对值后作为重要度的量化结果。最后的评价得分由满意度和重要度两者量化的结果通过公式计算得到。分数越高，则答复意见的质量越好。反之则越差。

关键词：爬虫；TF-IDF；朴素贝叶斯；神经网络；文本相似度；

目录

摘要

一、 问题分析 3

1.1 任务一的分析 3

1.2 任务二的分析 3

1.3 任务三的分析 4

二、 数据准备 4

2.1 爬取特定地点 4

2.2 附件数据预处理 5

2.2.1 数据去重 6

2.2.2 数据缺失值处理 6

三、 模型假设 7

四、 任务一 7

4.1 中文分词 7

4.1.1 基于词典的分词方法 7

4.1.2 基于统计的分词方法 8

4.1.3 隐马尔可夫模型和 Viterbi 算法实现中文分词 8

4.1.4 数据分词后的效果 8

4.2 去停用词 8

4.3 文本向量化 9

4.3.1 基于文本集建立词典 9

4.3.2 基于词典建立语料库 10

4.3.3 doc2bow 函数 10

4.3.4 TF-IDF 模型处理语料库 10

4.4 朴素贝叶斯模型 11

4.5 浅层神经网络 11

4.6 模型评价 12

五、 任务二 13
5.1 留言归类 13
5.1.1 基于正则表达式匹配地点 13
5.1.2 基于地点词典匹配小区名称 14
5.1.3 归类汇总 15
5.2 热点挖掘 15
5.2.1 热度评价指标定义 15
5.2.1 热点问题挖掘实现 16
六、 任务三 18
6.1 满意度建模 18
6.1.1 答复意见完整性 18
6.1.2 答复意见时效性 19
6.2 重要度建模 19
6.2.1 基于词典的留言文本情感分析 19
6.3 答复意见的质量评价得分计算 20
6.4 模型评价 21
七、 参考文献 22

一、 问题分析

1.1 任务一的分析

通过附件 2 给出的数据，建立一个关于留言内容的一级标签分类模型。

留言数据多数是对某些特定地点进行投诉，首先从网上爬取这些特定地点，然后对数据进行地点的提取。通过分词技术将每条数据的留言进行分词，然后构建语料库，用 TF-IDF 模型处理语料库，以及对一级标签进行量化。将处理量化后的训练数据集，通过机器学习的贝叶斯模型和浅层神经网络两种方式对其进行训练，得到一级标签分类模型。通常使用 F-Score 对贝叶斯模型和浅层神经网络进行评价。

1.2 任务二的分析

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，得到排名前 5 的热点问题以及给出相应热点问题对应的留言信息。

利用任务一爬取的特定地点对每条数据的留言主题进行匹配得到该数据的特定地点，然后再通过正则表达式提取每条数据的市、区和县完成归类任务。

点赞数量高可以反映群众所关注的焦点。若对于一个问题，众多群众对其进行留言，那么可以认为该问题是群众急于需要解决的问题。因此分类里每个数据的数目也可以反映出群众关注的焦点问题。把点赞数量和分类数目作为热度评价的指标。

对每个分类里面的数据查找得出点赞数最高的数据。进行计算，由于点赞数和分类数目的重要程度不一，需要通过一个权值进行权衡。经过权值计算之后，对热度评价得分前五的分类提取，然后提取出每个分类点赞数最高的数据进行留言主题提取，构成热点问题表的问题描述。将热点问题的描述作为目标主题，选取出与目标主题相似的文本内容，从而实现热点问题留言明细表的统计。最后的结果分别汇总在热点问题表和热点问题留言明细表中。

1.3 任务三的分析

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。

对答复意见的质量给出一套评价方案，也是对群众收到答复时的情感预测。首先要建立评价指标。从群众收到答复的通常的反应这个角度去分析。群众收到答复的时候通常会有两个情感表现。第一、群众在收到答复意见时，觉得意见是否达到了自己的预期会作出一个情感的评分。第二、群众在看到答复意见的内容时是否感到满意作出一个评分。最终建立两个指标，分别是满意度和重要度。

对于群众收到的答复意见是否能够接受以及有没有达到预期。其实就是相当于群众的满意度预测。于是评价方案就由满意度和重要度两个指标所决定。

通过对答复的内容从完整性和时效性去综合预测群众的满意度。而重要度则对群众的留言进行情感分析，从而得出群众对该留言的重要程度。

最后根据重要度和满意度去评价答复意见的质量。

二、 数据准备

2.1 爬取特定地点

数据里面有大量的特定地点如图 2-1，因此需要从网上爬取得到该特定地点。

留言编号	留言用户	留言主题
189294	A00083527	A市南站里面的12306客服虚设
189313	A000103797	A市磁悬浮列车只能现金支付很不方便
189333	A00093505	反映A2区卢富原著学区划分不合理的问题
189345	A00077163	A4区东风街道蚌塘社区自来水管网改造设计存在严重缺陷
189381	A000109815	A市万科魅力之城商铺无排烟管道，小区内到处油烟味
189456	A00028347	A3区云栖路云栖谷小区辅道好多乱停车的

图 2-1 特定地点示例

选取安居客作为爬取特定地点的网址，如图 2-2

图 2-2 爬取特定地点的网址

总共爬取了 477221 条数据，汇总在小区名称和小区地址的 Excel 表格中，部分数据如图 2-3

1	小区名称	小区地址
23667	金茂梅溪湖	[岳麓区-梅溪湖] 梅溪湖南路100号
23668	旭辉御府	[岳麓区-梅溪湖] 东方红路
23669	澳海澜庭	[望城区-金星北大道] 银星路999号
23670	万科魅力之城	[雨花区-体育新城] 劳动东路1299号
23671	山水湾	[星沙-万家丽北路] 万家丽北路88号
23672	保利国际广场	[天心区-书院路] 书院路9号
23673	中建江山壹号	[天心区-书院路] 湘江中路,近橘洲湾路
23674	中城丽景香山	[雨花区-劳动中路] 万家丽中路二段239号
23675	和泓梅溪四季(一期)	[岳麓区-梅溪湖] 映日路599号
23676	保利西海岸	[岳麓区-银盆岭] 银盆岭路228号
23677	碧桂园山湖城(别墅)	[宁乡市-金洲区] 金洲大道东288号
23678	保利香槟国际	[星沙-万家丽北路] 湘龙西路42号
23679	梅溪青秀	[岳麓区-梅溪湖] 沐风路19号
23680	保利花园	[天心区-省政府] 湘府西路229号
23681	中粮北纬28度	[望城区-雷锋大道北] 雷锋大道,近普瑞大道
23682	时代倾城(一期)	[望城区-金星北大道] 银星路
23683	盈峰翠邸	[望城区-金星北大道] 金星北路四段229号

图 2-3 爬虫部分数据

2.2 附件数据预处理

在数据分析当中，未清洗的数据通常会出现异常。一般有数据重复以及数据缺失的情况出现，如果我们不对这些情况进行处理，这会对我们后续的分析带来很多的影响。

2.2.1 数据去重

对于附件 2，附件 2 中的数据会发现有些群众会进行重复的留言如图 2-4，重复的数据在后续训练中会影响训练效果，因此需要去重处理。去除附件 2 留言详情重复数据 158 条。

1979年西地省政府为了鼓励干部、职工计划生育，而制定了“西地省城镇独生子女父母奖励制度”即独生子女父母退休时可享受增发5%的基本工
2001年华能F市电厂为57名身体健康、年龄33—45周岁的大集体职工办理正式退休。1979年国家为治理三废，投资五百万兴办，较现代化的粉煤
2003年，以腐败分子张恩照为首的建行决策者，采取错误的方式进行一次性减员11万，把一批为建行工作了几十年的员工推向社会失业大军中，
2005年期间，临G5县粮食局依据中共中央“关于国有企业改革和发展若干重大问题的决定”，西地省人民政府办公厅“关于进一步深化粮食购销
2006年10月13日，楚人发[2006]142号，关于印发《西地省公务员登记工作有关问题年处理意见》的通知　　关于机关违规进入人员的处
2008年2月，市人事局和市检察院对本人违法办理退职，2016年10月，法院判决撤销，判决生效已经7个月(具体见人社局收到的行政判决书)，请
2009年，以基药制度为杠杆的新医改首先在基层推开。基本药物制度的推出，原本目的是解决老百姓“看病难、看病贵”问题。不过，基药制度
200名社会弱势群体的泣血呐喊　　------关于长期苛扣巡逻队员工资、不购买社会保险的控告书　　控告人：J7县200名治安巡逻队员　　被控
200名社会弱势群体的泣血呐喊　　------关于长期苛扣巡逻队员工资、不购买社会保险的控告书　　控告人：J7县200名治安巡逻队员　　被控
2010年12月13日我在西地省C市中心医院妇产科经检查确认怀孕，预产期为2011年8月7日。我怀孕后一直在该院妇产科遵医体检。2011年4月11日
2010年下班途中发生交通事故，交警责任认定是同等责任，2011年申请工伤认定时，劳动部门以《工伤保险条例》第十六条第一款为由不予认定
2011年的高级职称还要等到什么时候评呢?听说你彭厅长要求所有参评者不管什么理由你都要求刷下百分之三十.你真不是个什么东西,只会向老
2011年内物价上涨，听说行政事业单位退休人员每人发了千多元物价补贴。企业退休人员养老金低得多，反而无分文物价补贴，这是为什么？对

图 2-4 附件 2 部分重复数据

对于附件 3，由于任务二需要提取出热点问题的每一条具体数据，因此并不进行去重处理。

2.2.2 数据缺失值处理

在附件 3 里出现了数据缺失的情况，我们需要对其进行补全或者去除的处理。因为在任务二当中我们尽量避免删除题目所提供的数据，所有我们对于附件 3 中出现的缺失情况进行了补全的处理。在附件 3 中出现了一处缺失情况，进行文本相似的补全处理，补全前后如图 2-5，图 2-6

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
189587	A00018292	家冲Ⅰ、Ⅱ线500kV线路杆迁	2019/1/13 18:44:38	行充分论证。市县约	0	0
189635	A00029819	A1区桐阴里小区一直夜间施工	2019/7/15 21:27:02	直夜间施工，对周	0	0
189663	A00083693	34层高楼用1.2的栏杆作为生	2019/5/11 19:03:58	商的某些工作人员	0	0
189733	A0009754	A市限卖房产政策一刀切	2019/10/10 17:09:47	致我无法启动我的创	1	0
189739	A00051608	区西湖街道茶场村五组是如何	2019/9/12 8:30:47	？？周边都拆迁了，	0	0
189856	A00073717	）	2019/7/3 11:53:35	阳一等就不知道是多	0	1

图 2-5 数据补全前

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
189587	A00018292	兴-艾家冲Ⅰ、Ⅱ线500kV线路杆迁工程临	2019/1/13 18:44:38	行充分论证。市县约	0	0
189635	A00029819	A1区桐阴里小区一直夜间施工	2019/7/15 21:27:02	直夜间施工，对周	0	0
189663	A00083693	复式楼34层高楼用1.2的栏杆作为生命的	2019/5/11 19:03:58	商的某些工作人员	0	0
189733	A0009754	A市限卖房产政策一刀切	2019/10/10 17:09:47	致我无法启动我的创	1	0
189739	A00051608	问A3区西湖街道茶场村五组是如何规划	2019/9/12 8:30:47	？？周边都拆迁了，	0	0
189856	A00073717	A3区保利麓谷林语小区外墙渗水日益严	2019/7/3 11:53:35	阳一等就不知道是多	0	1

图 2-6 数据补全后

三、 模型假设

1) 数据来源准确可靠

四、 任务一

4.1 中文分词

分词是汉语自然语言处理的第一项核心技术，词是中文表达语义的最小单位，分词的结果对中文信息处理至为关键。文本分词直接决定后续模型的效果。因此自然语言处理的核心步骤在于分词的效果。

4.1.1 基于词典的分词方法

基于词典的分词方法首先需要建立一个足够大的词典，然后依据一定的策略扫描句子，若句子中的某个字串与词典中的某个词匹配，则分词成功。该方法简单实用，但又过于依赖词典，分词效果缺少保障

常见的有：正向最大匹配法、逆向最大匹配法、双向最大匹配法和最少词数分词法。

1) 最大正向匹配法

   对输入的句子从左至右，以贪心的方式切分出当前位置上长度最大的词，组不了词的字单独划开。其分词原理是：词的颗粒度越大，所能表示的含义越精确。

2) 逆向最大匹配法

   原理与正向最大匹配相同，但顺序不是从首字开始，而是从末字开始，而且它使用的分词词典是逆序词典，其中每个词条都按逆序方式存放。在实际处理时，先将句子进行倒排处理，生成逆序句子，然后根据逆序词典，对逆序句子用正向最大匹配。

4.1.2 基于统计的分词方法

基于统计的分词方法是从大量已经分词的文本中，利用统计学习方法来学习词的切分规律，从而实现对未知文本的切分。

常用的统计学习方法有：隐马尔可夫模型（HMM）等

4.1.3 隐马尔可夫模型和 Viterbi 算法实现中文分词

隐马尔可夫模型是对序列进行标注，将分词问题转化为字的分类问题，每个字有 4 种词位（类别）：词首（B）、词中（M）、词尾（E）和单字成词（S）。由字构词的方法并不依赖于事先编制好的词典，只需对分好词的语料进行训练即可。当模型训练好后，就可对新句子进行预测，预测时会针对每个字生成不同的词位。

隐马尔科夫的预测问题是要求图中的一条路径，使得该路径对应的概率值最大。在计算该步骤上穷举法的计算量非常大。通过 Viterbi 算法利用动态规划的思想来求解概率最大路径。从而较好地解决了隐马尔科夫的预测问题，实现分词。

4.1.4 数据分词后的效果

由于文本当中存在大量的特定地点如 A 市,万科魅力之城等，通过分词后这类的词容易会被切分，无法保留完整性，因此在分词之前需要将事先在网上爬取到的地点以及自定义的词组成新的词库，再对附件 2 中的留言主题以及留言详细进行中文分词。分词示例如下表 4-1

原句子	A 市西湖建筑集团占道施工有安全隐患
中文分词后	['A 市', '西湖', '建筑', '集团', '占', '道', '施工', '有', '安全隐患']

表 4-1 分词示例

可以看到分词的效果能够达到后续的处理要求。

4.2 去停用词

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，如标点符号、语气词或

者该词应用广泛等等，这些字或词即被称为 Stop Words（停用词）。在后续模型当中这些词会影响训练的效果，因此需要进行去除。

这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。如今网上已经整理了许多停用词表，借用网上的停用词表进行去停用词后的部分数据如图 4-2，全部数据存放在附件二分词后的 Excel 文件中。

留言主题去停用词
A市 西湖 建筑 集团 占 道 施工 安全隐患
A市 在水一方 大厦 人为 烂尾 多年 安全隐患
投诉 A市 A1区 苑 物业 违规 收 停车费
A1区 蔡锷 南路 A2区 华庭 楼顶 水箱 长年 不洗
A1区 A2区 华庭 自来水 好大 一股 霉味
投诉 A市 盛世耀凯 小区 物业 无故 停水
咨询 A市 楼盘 供暖 一事
A3区 桐梓 坡 西路 可可 小城 长期 停水 得不到 解决
C4 市 收取 城市 垃圾处理 费 平等
A3区 魏家坡 小区 脏乱差
A市 魏家坡 小区 脏乱差

图 4-2 部分去停用词后的数据

4.3 文本向量化

文本表示是自然语言处理中的基础工作，文本表示的好坏直接影响到整个自然语言处理系统的性能。文本向量化就是将文本表示成一系列能够表达文本语义的向量，是文本表示的一种重要方式。目前对文本向量化大部分的研究都是通过词向量化实现的，也有一部分研究者将句子作为文本处理的基本单元，于是产生了 doc2vec 和 str2vec 技术。

4.3.1 基于文本集建立词典

将附件 2 分词后的文本集形成词典，以便于后面建立语料库使用，示例如表 4-3

文本集	[['坚果', '果实'], ['坚果', '实在', '好吃']]
词典	(['坚果', '果实', '好吃', '实在'])

表 4-3 建立词典示例

4.3.2 基于词典建立语料库

利用构建好的词典建立语料库，语料库即存放稀疏向量的列表，示例如表 4-4

文本集	[['坚果', '果实'], ['坚果', '实在', '好吃']]
词典	{'坚果': 0, '果实': 1, '好吃': 2, '实在': 3}
语料库	[[(0, 1), (1, 1)], [(0, 1), (2, 1), (3, 1)]]

表 4-4 建立语料库示例

通过 doc2bow 函数处理附件 2 文本集建立的词典，得到语料库。需要进一步进行权值的优化。

4.3.3 doc2bow 函数

1、将所有分词汇集成一个集合，并对每个词分配一个序号。

以['坚果', '实在', '好吃']为例

对每个分词分配序号：坚果→0；好吃→2；实在→3

转变成：[0, 3, 2]

2、转换成稀疏向量

0 有 1 个，即表示为(0, 1)， 2 有 1 个，即表示为(2, 1)，3 有 1 个，即表示为(3, 1)

最终结果：[(0, 1), (2, 1), (3, 1)]

4.3.4 TF-IDF 模型处理语料库

TF-IDF（term frequency－inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。

TF-IDF 是一种统计方法，用以评估一个词对于一个文件集或一个语料库中的其中一份文件的重要程度。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。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

计算公式为

$$TF-IDF=TF*IDF \tag{1.1}$$

计算示例如下表 4-5

文本集	[['坚果', '果实'], ['坚果', '实在', '好吃']]
词典	{'坚果': 0, '果实': 1, '好吃': 2, '实在': 3}
语料库	[[(0, 1), (1, 1)], [(0, 1), (2, 1), (3, 1)]]
TF-IDF	([[0.57973867, 0, 0, 0.81480247], [0.44943642, 0.6316672 , 0.6316672 , 0]])

表 4-5 TF-IDF 模型处理语料库示例

经过一系列的计算之后，得到附件 2 分词之后 TF-IDF 处理后的语料库，为接下来的贝叶斯模型训练做准备。

4.4 朴素贝叶斯模型

朴素贝叶斯（Naive Bayes）模型是一个非常常用的分类模型，它主要是应用了数理统计中非常重要的贝叶斯公式：

$$P(B_i \mid A)=\frac{P(B_i)P(A \mid B_i)}{\sum P(B_j)P(A \mid B_j)} \tag{1.2}$$

朴素贝叶斯模型常常用于文本分类，在这里选取了多项式朴素贝叶斯模型和伯努利朴素贝叶斯对比进行训练。

通过 sklearn 引入贝叶斯分类器，对处理好的附件 2 留言详情的 TF-IDF 语料库以及对应的一级标签进行训练。从而得到留言内容的一级标签分类模型。

朴素贝叶斯模型的优缺点总结如下：

优点：模型简单，训练速度快。有数学理论的支持，易解释。

缺点：容易出现欠拟合。在变量之间取值较为独立时，模型效果才较好。

4.5 浅层神经网络

神经网络被设计成与生物神经元和神经系统类似的数学模型，这些模型用于发现被标注数据中存在的复杂模式和关系。一个浅层神经网络主要包含三层神经

元-输入层、隐藏层、输出层。

将一级标签转化为 one-hot vectors 形式，一个 one-hot 向量除了某一位的数字是 1 以外其余各维度数字都是 0。如一级标签总共有 7 个类别，那么城乡建设的 one-hot 向量为[1, 0, 0, 0, 0, 0, 0]。

1) 导入处理后的留言详情 TF-IDF 语料库以及转化后的一级标签
2) 定义参数：

   输入：data_train[None,1000], 权重：w[1000, 7], 偏置项: b[7],

   输出: y[None,7]
3) softmax 函数对输出结果进行映射，sorfmax 的计算公式为

$$f(x_i)=\frac{e^{x_i}}{\sum_i e^{x_i}} \tag{1.3}$$

4) 定义损失函数（交叉熵），其公式为

$$H_{y'}(y)=-\sum_i y_i' \log(y_i) \tag{1.4}$$

5) 最小化损失函数

   权值优化算法:梯度下降法、设置 0.3 的学习速率
6) 通过 TensorFlow 进行网络的搭建和训练

   将数据集切分为训练集占 80%和测试集 20%，经过两万次训练后，训练集精度达到 0.8963，测试集精度达到 0.8643.

   效果能够达到预期。模型较好。

4.6 模型评价

通常使用 F-Score 对分类方法进行评价：

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{1.5}$$

利用公式进行计算多项式朴素贝叶斯模型的 F_1 值为 0.8566，而伯努利朴素贝叶斯模型的 F_1 值为 0.8192。多项式朴素贝叶斯模型对预测的效果更好，而且

也达到了预测的目的。

因此选用多项式朴素贝叶斯模型或者浅层神经网络为留言内容的一级标签分类模型都能达到一级标签分类的目的。

五、 任务二

5.1 留言归类

在附件 3 里留言存在着大量的特定地点如市、区、县、小区等，又或者特定人群如学生等如图 5-1 所示，因此需要对反映特定地点或特定人群问题的留言进行归类，归类后才能更好地进行热点问题的挖掘。

留言主题
A3区一米阳光婚纱艺术摄影是否合法纳税了？
咨询A6区道路命名规划初步成果公示和城乡门牌问题
反映A7县春华镇金鼎村水泥路、自来水到户的问题
A2区黄兴路步行街大古道巷住户卫生间粪便外排
A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民

图 5-1 部分留言主题的特定地点

5.1.1 基于正则表达式匹配地点

正则表达式，计算机科学的一个概念。正则表达式通常被用来检索、替换那些符合某个模式(规则)的文本。

在留言主题以及留言内容当中基本都会出现市、区或者县。这些名词都能满足“字母+数字+市区县”这样一个模式。

利用该模式组成一个“规则字符串”。借用正则表达式匹配出留言主题以及留言内容文本当中符合要求的字符串。从而实现地级市的提取，部分数据如图 5-2 。

留言编	留言用户	留言主题	留言详情	市	区、县
188006	A000102948	A3区一米阳光婚纱艺术摄影是否合法纳税了？	座落在A市A3区联丰路米兰春天G2栋320，一家	A市	A3区
188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌	A市A6区道路命名规划已经初步成果公示文件，	A市	A6区
188031	A00040066	反映A7县春华镇金鼎村水泥路、自来水到户的问	本人系春华镇金鼎村七里组村民，不知是否有	A市	A7县
188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	靠近黄兴路步行街，城南路街道、大古道巷、	A市	A2区
188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施	A市A3区中海国际社区三期四期中间，即蓝天璞	A市	A3区
188073	A909164	A3区麓泉社区单方面改变麓谷明珠小区6栋架空层	作为麓泉社区麓谷明珠小区6栋居民，我们近期	A市	A4
188074	A909092	A2区富绿新村房产的性质是什么？	“二高一部”发出关于针对非法集资的打击的通	A市	A2区
188119	A00035029	对A市地铁违规用工问题的质疑	我是一名在A市某地铁站上班的安检员，我是由	A市	nan
188170	A88011323	A市6路公交车随意变道通行	12月21日下午17时52分许，6路公交车（司机座	A市	A3区
188249	A00084085	A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁	保利麓谷林语桐梓坡路与麓松路交汇处地铁凌	A市	A3区
188251	A00013092	A7县特立路与东四路口晚高峰太堵，建议调整信	近来，下午晚高峰五点半左右，经过特立路与	A市	A7县

图 5-2 部分数据地级市的提取

5.1.2 基于地点词典匹配小区名称

利用事先爬取到的小区数据，抽取当中的小区名称数据，然后对其去除括号及括号里的内容，去除前后部分数据如图 5-3。然后进行去重处理，以便后续的查找工作。

小区名称	
锦绣花园(龙安区)	锦绣花园
华强城米兰达(六期)	华强城米兰达
华强城卡塞雷斯(三期)	华强城卡塞雷斯
华强城桑堤亚纳(一期)	华强城桑堤亚纳
瀚林苑(文峰区)	瀚林苑
双泰苑(东区)	双泰苑
水木清华(一期)	水木清华
康乐花园(东区)	康乐花园
华强城圣波拉(五期)	华强城圣波拉
华强城凯蒂斯(四期)	华强城凯蒂斯
华强城莱昂郡(二期)	华强城莱昂郡

图 5-3 小区名称处理后部分数据

经过清洗处理后，将地点数据构建成一个地点词典，然后对其进行遍历匹配查找，检索文本当中是否含有词典里的小区，若有，则提取出来，否则为空。从而达到从文本当中提取小区名称的目的，提取后的部分数据如图 5-4。

留言主题	留言详情	小区
A3区一米阳光婚纱艺术摄影是否合法纳税了？	座落在A市A3区联丰路米兰春天G2栋320，一家	一米阳光
咨询A6区道路命名规划初步成果公示和城乡门牌	A市A6区道路命名规划已经初步成果公示文件，	nan
反映A7县春华镇金鼎村水泥路、自来水到户的问	本人系春华镇金鼎村七里组村民，不知是否有	春华镇金鼎村
A2区黄兴路步行街大古道巷住户卫生间粪便外排	靠近黄兴路步行街，城南路街道、大古道巷、	步行街
A市A3区中海国际社区三期与四期中间空地夜间施	A市A3区中海国际社区三期四期中间，即蓝天璞	中海国际社区
A3区麓泉社区单方面改变麓谷明珠小区6栋架空层	作为麓泉社区麓谷明珠小区6栋居民，我们近期	明珠小区
A2区富绿新村房产的性质是什么？	“二高一部”发出关于针对非法集资的打击的通	富绿新村
对A市地铁违规用工问题的质疑	我是一名在A市某地铁站上班的安检员，我是由	nan
A市6路公交车随意变道通行	12月21日下午17时52分许，6路公交车（司机座	nan
A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁	保利麓谷林语桐梓坡路与麓松路交汇处地铁凌	保利麓谷林语
A7县特立路与东四路口晚高峰太堵，建议调整信	近来，下午晚高峰五点半左右，经过特立路与	特立路与东四路口
A3区青青家园小区乐果果零食炒货公共通道摆放	还我宁静我要复习迎考，大半年底商空调/冰柜	青青家园

图 5-4 部分数据小区名称提取

5.1.3 归类汇总

将数据提取到的地级市和小区名称汇总在一起，为后续分析做准备。如图 5-5

留言主题	留言详情	小区	市	区、县
A3区一米阳光婚纱艺术摄影是否合法纳税了？	座落在A市A3区联丰路米兰春天G2栋320，一家	一米阳光	A市	A3区
咨询A6区道路命名规划初步成果公示和城乡门牌	A市A6区道路命名规划已经初步成果公示文件，	nan	A市	A6区
反映A7县春华镇金鼎村水泥路、自来水到户的问	本人系春华镇金鼎村七里组村民，不知是否有	春华镇金鼎村	A市	A7县
A2区黄兴路步行街大古道巷住户卫生间粪便外排	靠近黄兴路步行街，城南路街道、大古道巷、	步行街	A市	A2区
A市A3区中海国际社区三期与四期中间空地夜间施	A市A3区中海国际社区三期四期中间，即蓝天璞	中海国际社区	A市	A3区
A3区麓泉社区单方面改变麓谷明珠小区6栋架空层	作为麓泉社区麓谷明珠小区6栋居民，我们近期	明珠小区	A市	A4
A2区富绿新村房产的性质是什么？	“二高一部”发出关于针对非法集资的打击的通	富绿新村	A市	A2区
对A市地铁违规用工问题的质疑	我是一名在A市某地铁站上班的安检员，我是由	nan	A市	nan
A市6路公交车随意变道通行	12月21日下午17时52分许，6路公交车（司机座	nan	A市	A3区
A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁	保利麓谷林语桐梓坡路与麓松路交汇处地铁凌	保利麓谷林语	A市	A3区
A7县特立路与东四路口晚高峰太堵，建议调整信	近来，下午晚高峰五点半左右，经过特立路与	特立路与东四路口	A市	A7县
A3区青青家园小区乐果果零食炒货公共通道摆放	还我宁静我要复习迎考，大半年底商空调/冰柜	青青家园	A市	A3区

图 5-5 部分数据地点提取

5.2 热点挖掘

5.2.1 热度评价指标定义

假如一条留言内容有较多人的点赞，那么视为该内容受到群众的关注。可以视为评价热点问题的指标之一。附件 3 中部分点赞数目靠前的留言文本如图 5-6。

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
208636	A00077171	A市A5区汇金路五矿万境K9县存在一	2019/8/19 11:34:04	咬人，请问有人对	0	2097
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，A3	5	1762
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
217032	A00056543	严惩A市58车贷特大集资诈骗案保护	2019/2/25 9:58:37	小股东、苏纳弟弟苏	0	790
194343	A000106161	承办A市58车贷案警官应跟进关注留	2019/3/1 22:12:30	侦并没有跟进市领	0	733
263672	A00041448	A4区绿地海外滩小区距长赣高铁最近	2019/9/5 13:06:55	复到我如下问题：1	0	669
193091	A00097965	A市富绿物业丽发新城强行断业主家水	2019/6/19 23:28:27	提供地摊上买的收	0	242
284571	A00074795	建议西地省尽快外迁京港澳高速城区	2019/1/10 15:01:26	、长浏高速出口，这	0	80
200667	A00079480	请问A市为什么要把和包支付作为任务	2019/1/16 17:01:25	层工作者也不理解，	0	78
262052	A00072424	关于A6区月亮岛路沿线架设110kv高压	2019/3/26 14:33:47	14号令《建设项目环	0	78
226723	A00040222	A市三一大道全线快速化改造何时启动	2019/9/15 15:31:19	改造，打通机场北通	0	66
272089	A00061602	关于A6区月亮岛路110kv高压线的建设	2019/4/9 17:10:01	地省体操学校、西塘	2	55
281898	A00096623	A市长房云时代多栋房子现裂缝，质量	2019/2/25 15:17:38	检站处理，陷入了与	5	55

图 5-6 点赞数量靠前的部分数据

点赞数量高固然可以认为是群众所关注的问题，但点赞数量不高则未必是没有受到群众们的关注。因此需要建立另一个评价指标。

对于一个问题，如果众多群众对其进行留言，那么可以认为该问题是群众急于需要解决的问题。在归类完成之后，我们就可以对其进行分类统计，得到每个分类里数据的数目，，把这个数目作为热度评价的指标之一。

综上，把点赞数和分类数目作为热度评价指标。

5.2.1 热点问题挖掘实现

对归类后的数据通过 pandas 以小区名称进行分类汇总，得到分类数目，如图 5-7

索引	数目
丽发新城小区	51
滨河苑	41
魅力之城	21
58车贷	13
深业睿城	11
万国城	10

图 5-7 部分分类数目

然后再对每个分类里面的数据查找得出点赞数最高的数据。进行计算，由于点赞数和分类数目的重要程度不一样，需要通过一个权值进行权衡。最后热度评价得分的计算公式为

$$score = 0.65m + 0.0035n \tag{1.6}$$

式（1.6）中，m 为分类数目，n 为最高点赞数。

计算示例，选取丽发新城小区，该分类数目为 51，最高点赞数为 24，则

$$score = 51 \times 0.65 + 24 \times 0.0035 = 33.234$$

对热度评价得分前五的分类提取，然后提取出每个分类点赞数最高的数据进行留言主题提取，构成热点问题表的问题描述。热度评价得分前五如图 5-8，前五分类中点赞数最高的数据如图 5-9。

小区	热点评价指标
丽发新城小区	33.234
滨河苑	26.692
魅力之城	13.671
五矿万境	13.1895
58车贷	11.3235

图 5-8 热度评价得分前五

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	小区
208285	A909205	投诉小区附近搅拌站噪音扰民	2019-12-15 12:32:11	尊敬的领导，我是A…	0	24	丽发新城小区
191001	A909171	A市伊景园滨河苑协商要求购房同时必须购买车位	2019-08-16 09:21:33	商品房伊景园滨河苑…	1	12	滨河苑
272122	A909113	A5区劳动东路魅力之城小区…	2019/08/01 16:20:02	局长：你好，A5区劳…	0	6	魅力之城
208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019/8/19 11:34:04	我是A市A5区汇金路…	0	2097	五矿万境
220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	尊敬的胡书记：您好…	0	821	58车贷

图 5-9 前五分类中点赞数最高的数据

同时该问题描述也作为后续文本相似度的搜索词。文本相似度的基本思路流程图如下图 5-10 所示

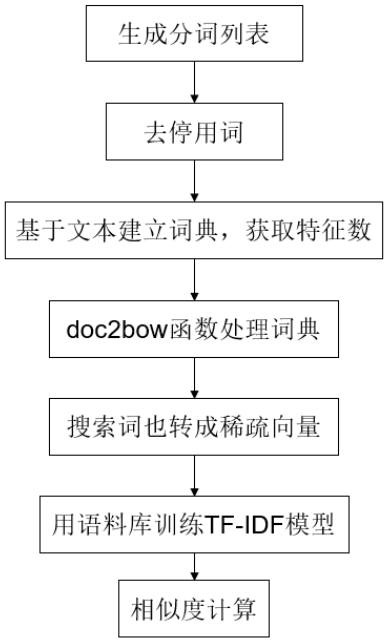

图 5-10 文本相似度流程图

通过一系列流程之后，最后进行相似度地计算，这里采取余弦相似度进行计算，公式如下

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{1.7}$$

式（1.6）中，A_i, B_i 分别代表向量 A 和 B 的各分量。

经过计算得出分数，通过一个阈值，选取出与目标主题相似的文本内容，从

而实现热点问题留言明细表的统计。最后的结果分别汇总在热点问题表和热点问题留言明细表中。部分热点问题留言明细如图 5-11。

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
1	208285	A909205	投诉小区附近搅拌站噪音扰民	2019-12-15 12:32:11	尊敬的领	0	24
1	255008	A909208	投诉小区附近搅拌站噪音扰民	2019-11-18 12:23:22	暮云街道	0	0
1	272224	A909224	丽发新城小区噪音大粉尘大,求撤走搅拌站	2020-01-09 19:46:10	我是暮云	0	1
1	239648	A909211	A市A2区丽发新城小区附近搅拌站明目张胆污染玉	2020-01-06 22:41:31	丽发新城	0	0
1	272447	A909206	投诉小区附近建设搅拌站	2019-11-20 19:12:22	投诉A市暮	0	0
1	225217	A909223	A2区丽发新城附近修建搅拌厂严重影响睡眠	2019-11-15 09:17:36	我已经好	0	0
1	281546	A0005147(	丽发新城小区附近搅拌站粉尘大,无法呼吸	2019-11-29 14:19:27	我是暮云	0	1
1	267050	A909227	噪音、灰尘污染的A2区丽发新城附近已扰乱居民	2019-11-02 10:18:00	A2区丽发	0	0
1	233158	A909242	丽发新城小区旁建搅拌厂严重扰民!	2019-12-05 08:46:20	本人是丽	0	0

图 5-11 部分热点问题留言明细

六、 任务三

对答复意见的质量给出一套评价方案，设立了满意度和重要度两个指标。两个指标进行建模，从而量化两个指标，再对答复意见的质量作出评价。评价方案的流程图如下图 6-1

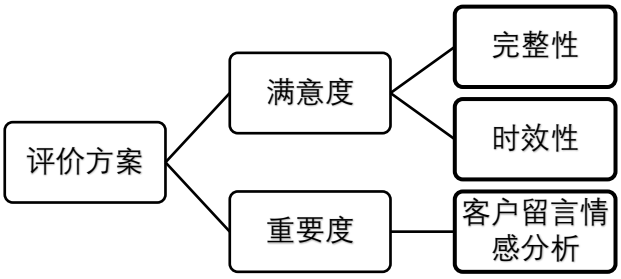

图 6-1 评价方案流程图

6.1 满意度建模

满意度的量化从完整性和时效性两个角度进行权衡。

6.1.1 答复意见完整性

一般地，客服在进行答复的时候，会按照的一定模板去进行答复。通常的模板包括、对对方的称呼或者问候，留言的详细内容以及最后的致谢三个部分。如果缺少一个部分那么则可以说该答复意见不完整。

因此对答复意见是否完整就可以将其分割成三个部分进行量化，检索每个部分是否满足模板，若满足则为 1，不满足则为 0。示例如表 6-2

答复意见	网友“UU00877”您好！您的留言已收悉。现将有关情况回复如下：经查，A4 区政府已责成区市政局牵头，区城乡建设局、区规划分局配合进行具体选址，招标（邀标）进行方案设计等，尽快启动万国城小区人行天桥建设并投入使用，方便出行。感谢您对我们工作的支持、理解与监督！2019 年 1 月 8 日		
	问候	答复内容	致谢
1：有，0：无	1	1	1

表 6-2 完整性量化

最后通过权值进行最后的模板得分计算，公式如下

$$T-score=0.2A+0.6B+0.2C \tag{1.8}$$

式（1.8）中，A 为问候，B 为答复内容，C 为致谢

计算示例为$T-score=0.2\times1+0.6\times1+0.2\times1=1$

从模板匹配的角度，实现对完整性的得分计算。

6.1.2 答复意见时效性

答复意见时效性是一个重要的指标，通常群众收到答复的时间过久，将会产生不满。

因此通过计算群众留言的时间到答复意见的时间差，可以评价群众的满意度。

答复意见的时间越早，那么群众的满意度则会越高。反之则越差。

6.2 重要度建模

6.2.1 基于词典的留言文本情感分析

情感分析（SA）又称为倾向性分析和意见挖掘，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程，其中情感分析还可以细分为情感极性（倾向）分析，情感程度分析，主客观分析等。

通过制定一系列的情感词典和规则，对文本进行段落拆借、句法分析，计算情感值，最后通过情感值来作为文本的情感倾向依据。

在自然语言文本挖掘中情感分析的基本思路如图 6-3。

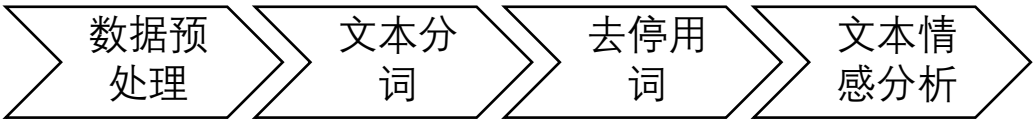

图 6-3 情感分析基本思路流程图

计算情感分析得分之后将其进行取绝对值作为群众的留言的重要度，部分重要度得分如图 6-4。

留言主题	留言时间	答复时间	重要度得分
A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	2019/5/10	0
A3区潇楚南路洋湖段怎么还没修好？	2019/4/24 16:03:40	2019/5/9	0.196399447
请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	2019/5/9	4.000751585
在A市买公寓能享受人才新政购房补贴吗？	2019/4/24 15:07:30	2019/5/9	5.123397098
关于A市公交站点名称变更的建议	2019/4/23 17:03:19	2019/5/9	1.109459973
A3区含浦镇马路卫生很差	2019-04-08 08:37:20	2019/5/9	0.451423893
A3区教师村小区盼望早日安装电梯	2019/3/29 11:53:23	2019/5/9	2.124614687
反映A5区东澜湾社区居民的集体民生诉求	2018/12/31 22:21:59	2019/1/29	2.436357742
反映A市美麓阳光住宅楼无故停工以及质量问题	2018/12/31 9:55:00	2019/1/16	0.278326756
反映A市洋湖新城和顺路洋湖壹号小区路段公共绿化	2018/12/31 9:45:59	2019/1/16	1.163057425
反映A2区大托街道大托新村违建问题	2018/12/30 22:30:30	2019/3/11	0.584716793
A5区鄱阳村D区安置房人防工程的咨询	2018/12/29 23:27:51	2019/1/29	4.247770862
A4区万国城小区段请求修建一座人行天桥或者地下	2018/12/29 11:55:34	2019/1/14	0.741969043
举报A市芒果金融平台涉嫌诈骗	2018/12/28 17:18:45	2019/1/3	3.463082883
建议增开A市261路公交车	2018/12/28 7:53:25	2019/1/14	0.600581451
关于A市新开铺路与披塘路交叉路口通行安全问题的	2018/12/27 15:18:07	2019/3/6	0.810677019

图 6-4 部分重要度得分

6.3 答复意见的质量评价得分计算

评价由满意度和重要度共同决定，得分与答复相隔的天数成反比，利用重要度去权衡完整性和时效性的总分得分。最后计算出评价得分。

计算公式为

$$lastsroce = (\frac{1}{day} + T - sroce) * \frac{1}{f} \tag{1.9}$$

式（1.9）中，day 为答复相隔的天数，$T - sroce$ 为 完整性得分，f 为重要度得分。

由式中可以得出，对于重要程度较低的时候，答复的相隔天数和完整性对于

群众的满意程度并不会有太大的影响。而当重要程度较高的时候，答复的相隔天数和完整性就会很大程度地决定群众的满意程度。

最后通过 Python 计算得出所有答复意见的评价得分，部分得分如图 6-5

答复意见	答复时间	最终得分
现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑	2019/5/1(	10.65679
网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的	2019/5/9	3.602716
市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已	2019/5/9	0.260383
网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市	2019/5/9	0.166111
网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于	2019/5/9	0.383389
网友“A00077538”：您好！针对您反映A3区含浦镇马路卫生很差的问题,A3	2019/5/9	2.937185
网友“A000100804”：您好！针对您反映A3区教师村小区盼望早日安装电梯	2019/5/9	0.460497
网友“UU00812”您好！您的留言已收悉。现将有关情况回复如下：一、关于	2019/1/29	0.40809
网友“UU008792”您好！您的留言已收悉。现将有关情况回复如下：据查，	2019/1/16	2.806053
网友“UU008687”您好！您的留言已收悉。现将有关情况回复如下：您所反	2019/1/16	0.998608
网友“UU0082204”您好！您的留言已收悉。现将有关情况回复如下：经查，	2019/3/11	1.481105

图 6-5 部分评价得分

6.4 模型评价

优点，模型综合了满意度和重要度两个方面去分析，模型较为全面

缺点，模型在量化满意度的完整性时会存在一定误差，以及基于词典的方法分析情感，过于依赖词典，导致模型效果较为一般。可以从这两方面进行优化。

七、 参考文献

[1]奉国和,郑伟.国内中文自动分词技术研究综述[J].图书情报工作,2011,55(02):41-45.

[2] 邹佳伦, 文汉云, 王同喜. 基于统计的中文分词算法研究[J]. 电脑知识与技术, 2019, 015(004):149-150,153.

[3]施聪莺,徐朝军,杨晓江.TFIDF 算法研究综述[J].计算机应用,2009,29(S1):167-170+180.

[4]徐文海,温有奎.一种基于 TFIDF 方法的中文关键词抽取算法[J].情报理论与实践,2008(02):298-302.

[5]王国才. 朴素贝叶斯分类器的研究与应用[D]. 重庆交通大学.

[6]杨立公,朱俭,汤世平.文本情感分析综述[J].计算机应用,2013,33(06):1574-1578+1607.

[7] 夏明军. 基于数据挖掘的移动用户满意度分析[D]. 2006.