“智慧政务”中的文本挖掘应用

摘要

本文根据附件给出的信息进行数据探索、文本预处理、文本向量化、构建留言划分模型、构建留言热点主题模型等工作。旨在为自然语言处理技术以及文本挖掘技术在“智慧政务”中的作用提供新思路与新方向。

针对任务一，首先通过爬取新数据进行数据增强；为了避免数据不平衡造成的影响，对每个标签随机选取 1200 条数据作为实验数据；接着利用正则表达式对留言详情内容进行清洗，利用 jieba 分词进行分词操作，对分词后的数据进行去停用词，得到可用数据。选择 TF-IDF 模型进行特征提取，将处理后的数据按照 8：2 的比例划分为训练集和测试集，选取线性支持向量机、朴素贝叶斯、逻辑回归三个模型分别进行查准率（P)，查全率（R）和 F 值验证，进行模型评价与分析。

针对任务二，首先确定热度评价指标——热度指数=点赞数-反对数，通过文本筛选，筛选出热度指数大于 0 的数据。经过文本内容清洗，添加用户自定义词典，分词，词性标注，去停用词操作，得到包含地名（ns），机构团体（nt），其他专有名词（nz）以及未知词性（x）的词语，构建 LDA 主题模型；根据 LDA 主题模型结果，在潜在主题中寻找可能的地名，对附件三进行文本筛选，筛选出 22 个地点留言记录；根据热度指数高低，挖掘出热点问题。

针对任务三，用 TF-IDF+余弦相似度的方法去量化答复意见的相关性和完整性；而答复意见的可解释性，可以结合满意度评分来量化。

关键词： 自然语言处理　TF-IDF 算法　文本分类　文本挖掘　LDA 主题模型

目录

一 问题分析

近些年来，随着大数据、云计算、人工智能等技术等概念的出现与发展，创新技术已经成为全球经济发展的最重要推动力之一，人工智能快速赋能传统产业。互联网创新技术的推陈出新，不仅促进了各行各业的蓬勃发展，对于地方政府的政务管理运行，如果运用得当，也是能起到事半功倍的作用。4 月 28 日，中国互联网络信息中心(CNNIC)发布第 45 次《中国互联网络发展状况统计报告》。报告显示，截至 2020 年 3 月，我国网民规模为 9.04 亿，互联网普及率达 64.5%，较 2018 年底提升 4.9 个百分点；我国手机网民规模达 8.97 亿，较 2018 年底增长 7992 万，我国网民使用手机上网的比例达 99.3%，较 2018 年底提升 0.7 个百分点。网络的普及不仅丰富了人们的吃喝玩乐、衣食住行，同样的也为政府了解民意、 汇聚民智、凝聚民气的提供了新的重要渠道，随着大家对微信、微博、市长信箱、阳光热线等网络问政平台的熟悉，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，面对着日益庞大的数据量，传统的数据分析方式愈感力不从心，认知技术包括数据挖掘、机器学习和自然语言处理正在取代传统分析方法。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。

针对题目中提出的留言划分以及热点整理的相关问题，我们根据附件给出的信息进行了数据准备、文本预处理、文本向量化、构建留言划分模型、构建留言热点主题模型等工作，旨在为自然语言处理技术以及文本挖掘技术在“智慧政务”中的作用提供新思路与新方向。

二 数据探索

2.1 数据分布

将附件二中的数据导入程序中，运行发现数据中包含了 7 种一级标签，它们的分类及分布为：

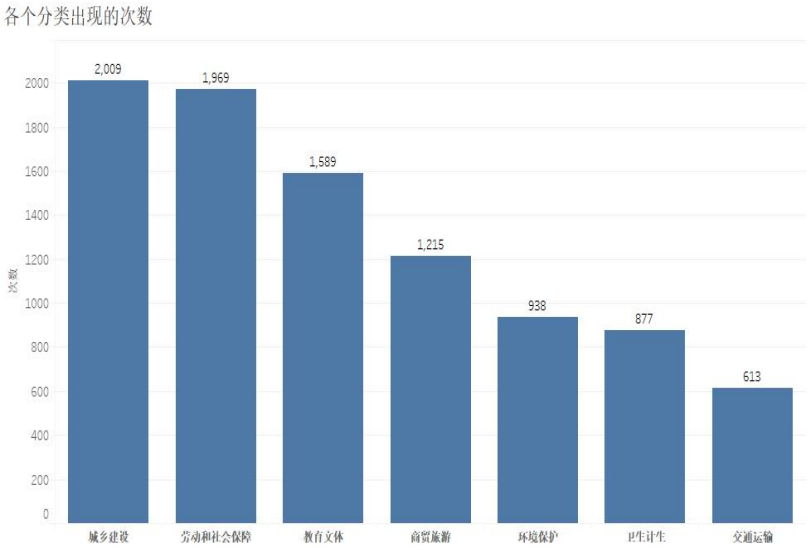


图 1　附件二数据分布情况

通过图 1 可直观的看出，数据分布存在不均匀现象，城乡建设类有 2009 条数据，而卫生计生类仅有 613 条数据，数据量相差较大。

2.2 数据增强

为避免数据量相差过大对分类器性能产生影响，本文先进行数据增强。

数据增强方法大致分为两种：

①　获取新的数据（从网页中爬取新的数据）；

②　对数据进行增强（利用已有的数据创造更多的数据）

本文采用第一种方法，新的数据来源于网页湖南省长沙市市长信箱，具有一定的权威性。由于各分类数据量有所不同，本文取中位数 1215≈1200 为标准，分别对环境保护类，卫生计生类，交通运输类进行数据增强。

将新增数据经过整理，分别命名为“环境保护新增数据”，“卫生计生新增数据”，“交通运输新增数据”。

将附件 2 的数据与新增后数据进行数据合并，整理后命名为“附件 2（完整数据）”。

2.3 数据抽样

在进行了数据增强后，数据分布为：

表 1　数据增强后数据分布情况

一级标签	次数
城乡建设	2009
劳动和社会保障	1969
教育文体	1589
商贸旅游	1215
环境保护	1200
卫生计生	1200
交通运输	1200

从表中可看出，各个分类的数据量已经相差不大，但数据依然不平衡，于是本文通过数据抽样对数据进行平衡处理。

数据抽样方式：

①　欠抽样：针对总体数据当中的多数类别，通过对多数类当中进行随机抽样的方法，减少多数的样本量，以此来降低数据集的不平衡程度；

②　过抽样：针对总体数据当中的少数类别，通过对少数类别当中进行复制，为少数类别增加额外的样本量，以此来降低数据集的不平衡程度。

显然，我们只需要对城乡建设类，劳动和社会保障类，教育文体类，商贸旅游类进行欠抽样，样本量为 1200。

三 文本预处理

3.1 结构化数据和非结构化数据

结构化数据：可以表示成多行多列的形式，并且每行（列）都有具体的含义。

非结构化数据：无法合理的表示为多行多列的形式，即使能表示，每行每列也没有实际的含义。

而文本数据，正是一种非结构化数据。因此预处理步骤大致为：

3.2 数据清洗

一般来说，数据清洗是指在数据集中发现不准确、不完整或不合理的数据，并且对这些数据进行修补或移除以提高质量的过程。包括处理缺失值，重复值，文本内容清洗等。

	留言详情	一级标签
9053	\n \n 我是一名4岁儿童的家长，住在A1区，我和老公都是单位上班的，工作...	卫生计生
10058	请问目前疫情期间自驾往返湖南省内城市之间，中间停留1日，是否需要回家后在家隔离14天	卫生计生
9914	尊敬的市长您好：　我家住芙蓉区德政园社区，近期因肺结核病在长沙市中心医院住院，听闻国家...	卫生计生
4967	\n\n西地省D7县栗江镇实行期末考试教师排名制度已经多年，并且排名靠前的就拿那些靠后的老师...	教育文体
5090	\n\n教师本是一门高尚的职业，但是如今教师却成了一门养家糊口都难的职业。尤其处在外地的教师...	教育文体
9834	我是2008参加的护士资格考试，2010年注册护士执业证，当时没有发护士资格证，解释是200...	卫生计生
9602	泉塘地区无社区公交到达地铁站，现有809线路过长，车辆配置多，但市区拥堵，车辆回场慢，造...	交通运输
7604	\n\t\t\t\t\t\n\t\t\t\t\tK市发展和改革委员会于2015年12月16日...	商贸旅游
10272	尊敬的胡市长：您好！长沙的街道一夜之间突然多了很多高高耸立的柱子，都是夜晚施工完成的。我通过...	环境保护
3883	\n \n　我是一名小县城一中的一名高三学生，高三的学习进度被学业水平考...	教育文体
3304	\n\t\t\t\t\t\n\t\t\t\t\tD11市大义镇设有一个快递代收发点，所有快递...	交通运输
9722	区、街道各位相关领导，您好！我是开福区捞刀河镇大星村的村民。今年20岁，作为一名学生，住在这...	交通运输
1551	\n\t\t\t\t\t\n\t\t\t\t\t　M6州M7市乾州新城，被M7市委、市政府...	城乡建设
6556	\n \n　西地省傲派自动化设备有限公司长期拖欠工资长达5个月之久，给企...	劳动和社会保障
3061	\n\t\t\t\t\t\n\t\t\t\t\t交通部2016年下文，要在全国推行城乡客运一...	交通运输

图 2　未进行清洗前的部分数据

	留言详情	一级标签
6857	尊敬的领导您好我是一名刚毕业工作了半年的大学生2018年9月办理了人才新政中的毕业生生活及住...	劳动和社会保障
4202	尊敬的杨书记我是一名初三毕业生的家长在万般无奈的情况下向您反映一个问题今年示范性高中A7县一...	教育文体
5398	尊敬的彭厅长2011年西地省卫生高级职称评审大部分的县市都公示了但仍有部分县市没有公示我们就...	劳动和社会保障
6050	省市社保局我们是一群流浪者———早一二十年前响应政府号召从单位辞职下海到民营科技企业和外资企...	劳动和社会保障
471	尊重市领导您好C市精美五大工程和美丽乡村建设作为金薮镇政府会纳入C市精美五大工程和美丽乡村建...	城乡建设
7948	城南社区禾塘村铺天垅罗子坡水库食品黑作坊无牌无证每天生产上吨豆腐供入市场对周边环境污染性很大...	商贸旅游
5010	作为一个家长一个普通公民我想投诉西地省G7县教育局G7县一中在2019年的中考招生中花样百出...	教育文体
8117	尊敬的李局长你好位于株州的西地省爱唯新贸易有限公司在销售千金养生产品利用微信微商网络进行层级...	商贸旅游
43	建议长株潭城轨最早一班城铁提早时间现在最早一趟A市西到C市的城铁是7点出发要8点半左右才能到...	城乡建设
8533	山门镇人民医院挂号费这几年来一直都是高额收取的也不知道是从哪一年开始的2013年春节我回家过...	卫生计生
8748	尊敬的张主任您好自2012年9月1日起正式推行农村孕产妇县乡住院分娩基本医疗全免费工作1明确...	卫生计生
7843	领导你好我是一个规矩做生意的良民百姓想跟您反应一个事情我是于今年进入位于解放西路司门口老街的...	商贸旅游
9512	长沙市民公共交通出行太不方便了公交扫码一个APP线下公交卡nfc充值一个APP公交查路线时间...	交通运输
2077	B9市楚联纸厂是整治落后产能企业关停对象国家按政策已补助二百多万元且该厂整合前是B9市环保审...	环境保护
3718	西地省G市G9市教育局没有做好免除城市义务教育阶段学生学杂费的通知国发200825号文件中间...	教育文体

图 3　数据清洗后的部分数据

3.3 文本分词

3.3.1 结巴分词

结巴分词是一种使用 Python 语言开发的中文分词工具[1]。它有三个主要特点：

a）支持三种分词模式：精确模式、全模式、搜索引擎模式；

b）支持繁体分词；

c）支持自定义词典。

结巴分词的实现基于以下三个原理：

a）基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（Directed Acyclic Graph DAG)；

b）采用动态规划查找最大概率路径，找出基于词频的最大切分组合；

c）对于未登录词，采用了 Viterbi 算法和基于汉字成词能力的 HMM 模型。

本文采用结巴分词中的精确模式。该模式是结巴分词中最基础和自然的模式，适合文本分析。

	留言详情	一级标签
7634	<generator object Tokenizer.cut at 0x0000026E4...	商贸旅游
9662	<generator object Tokenizer.cut at 0x0000026E4...	交通运输
7509	<generator object Tokenizer.cut at 0x0000026E4...	商贸旅游
7878	<generator object Tokenizer.cut at 0x0000026E4...	商贸旅游
1666	<generator object Tokenizer.cut at 0x0000026E4...	城乡建设
4492	<generator object Tokenizer.cut at 0x0000026E4...	教育文体
7858	<generator object Tokenizer.cut at 0x0000026E4...	商贸旅游
265	<generator object Tokenizer.cut at 0x0000026E4...	城乡建设
3382	<generator object Tokenizer.cut at 0x0000026E4...	交通运输
4949	<generator object Tokenizer.cut at 0x0000026E4...	教育文体
3533	<generator object Tokenizer.cut at 0x0000026E4...	交通运输
4303	<generator object Tokenizer.cut at 0x0000026E4...	教育文体
8423	<generator object Tokenizer.cut at 0x0000026E4...	卫生计生
8389	<generator object Tokenizer.cut at 0x0000026E4...	卫生计生
8510	<generator object Tokenizer.cut at 0x0000026E4...	卫生计生

图 4　分词后的部分数据

3.4 停用词处理

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，这些字或词被称之为停用词。

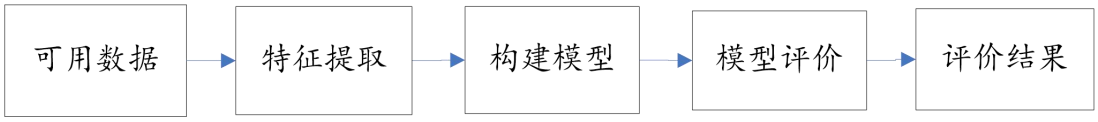
停用词在句中大量出现，却对语义分析没有帮助。对于这种词，我们一般会删除。同时经过停用词处理可以降低存储空间的消耗以及减少计算时间。

	留言详情	一级标签
1438	上半年 修路 路灯 半年 全停 月份 一段时间 10 月份 停 每天晚上 小孩 放学 回家 ...	城乡建设
4058	尊敬 蔡 局长 您好 I I3 一位 初中生 哥哥 妹妹 休 病假 一期 孩子 下期 开学 ...	教育文体
8788	2015 九月 二十五日 办理 二胎 准生证 资料 交给 K1 七里 办事处 计生办 工作人...	卫生计生
303	邓 局长 你好 一位 市民 2015 31 号 A1 K4 城管 没收 一辆 电动 三轮车 ...	城乡建设
3479	您好 请求 老百姓 解决 I 交通 禁止令 该令 闹 百姓 怨声载道 沸沸扬扬 究其原因 盲...	交通运输
3222	L5 县低 庄到 L3 公路 有没有 规划 修通 大渭溪 乡 L5 县低 庄到 L3 公路 ...	交通运输
7676	12 乘坐 L5 县至 溪口 线路 客车 岁 儿子 从后 溪口 办事 售票 儿子 买 全票 ...	商贸旅游
8392	厅长 您好 希望 百忙之中 制止 这件 发生 良知 正义感 公民 希望 卫生事业 阳光 希望...	卫生计生
3514	最为 拥堵 道路 规划 落后 路网 完善 高架 立交桥 规划 增加 十字路口 人行天桥	交通运输
6169	企业 职工 改革 红利 2005 国企 改制 一刀切 退休 政策 退休 国企 改制 分配 一...	劳动和社会保障
10323	近日 环保局 交通运输 局市 质监局 物价局 交警支队 部门 联合 下文 2013 长沙 正...	环境保护
6051	父亲 2012 11 16 下班 途中 因车祸 去世 D9 交警大队 认定 交通 意外 父亲...	劳动和社会保障
9096	位于 K3 县乐 泰州 超 市场 商 生鲜 贩卖 牛腩 牛肉 拿母 猪肉 加工 牛腩 卖 超...	卫生计生
4722	张丽系 A8 乡镇 在职 在编 高中化学 老师 强制 合同期 强制 合同 参加 2018A6...	教育文体
1042	K 帝王 广场 栋 开发商 省佳煌 房地产 开发 有限公司 栋 结构 体系 框剪 结构 装修...	城乡建设

图 5　去停用词后部分数据

四 群众留言分类

基本流程：



4.1 特征提取

4.1.1 概述

文本的表示及其特征项的提取是文本挖掘的一个基本问题，它把从文本中抽

取出的特征词进行量化来表示文本信息。将它们从一个无结构的原始文本转化为结构化的计算机可以识别处理的信息。（即对文本进行科学的抽象，建立它的数学模型，用以描述和代替文本，使计算机能够通过对这种模型的计算和操作来实现对文本的识别）

由于文本是非结构化的数据,要想从大量的文本中挖掘有用的信息就必须首先将文本转化为可处理的结构化形式。

特征提取的过程[2]：

(1) 对原数据集进行分词、去停用词等预处理，得到一个初始特征集 T；

(2) 特征集合 T 进行权重分配，并按权重值降序排列得到特征集 T1；

(3) 根据对应评估函数，选取一个得到最能代表文本类别信息的最优特征子集 T2。

4.1.2 特征提取模型

(1) 词袋模型[3]

Bag of Words，也称作“词袋”。它用于描述文本的一个简单数学模型，也是常用的一种文本特征提取方式。在信息检索中，词袋模型假定对于一个文本，忽略其次序和语法，仅仅当作是该文本中若干个词汇的集合。该文本中，每个词汇都是互不相关的，每个词汇的出现都不依赖于其他词汇。也就是说，文本中任意一个单词不管出现在任意哪个位置，都不会受到其他因素的影响。

模型步骤及例子

句子 A：Tom likes to play basketball.Mike likes too.

句子 B：Mike also likes to play tennis.

1、文本分词：把每个文档中的文本进行分词

2、构建词汇表：把文本分词得到的单词构建为一个词汇表，包含文本语料库中的所有单词，并对单词进行编号，假设词汇表有 n 个单词，单词编号从 0 开始，到 n-1 结束，可以把单词编号看作是单词的索引，通过单词编号可以唯一定位到该单词。即：[Tom,likes,to,play,basketball,Mike,too,also,tennis]

3、词向量表示：每个单词都表示为一个 n 列的向量，在单词编号（词汇索引）位置上的列值为 1，其他列的值为 0。即：

句子 A： [1,1,1,1,1,1,1,0,0]

句子 B： [0,1,1,1,0,1,0,1,1]

4、统计频次：统计每个文档中每个单词出现的频次。即：

句子 A：[1,2,1,1,1,1,1,0,0]

句子 B：[0,1,1,1,0,1,0,1,1]

这两个词频向量就是词袋模型，可以很明显的看到语序关系已经完全丢失。

（2）TF-IDF 模型[3]

TF-IDF（词频-逆文档频率法）作为一种加权方法，在词袋模型的基础上对词出现的频次赋予 TF-IDF 权值，对词袋模型进行修正，进而表示该词在文档集合中的重要程度。

模型步骤：

1、统计词频 TF：统计每个词在文本中出现的次数，出现的越频繁，那么就越可能是这个文章的关键词。

$$词频(TF)=\frac{某个词在文档中出现的次数}{文档中所有词汇出现的次数总和}$$

2、计算逆文档频率 IDF：指逆向文本频率，是用于衡量关键词权重的指数。一个词在语料库中越少见，它的权重就越大；反之，一个词在语料库中越常见，它的权重就越小。

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

3、计算 TF-IDF：用于衡量某个词在文章中的重要性。TF-IDF 与该词在文章中的出现次数成正比，与该词在整个语料库中的出现次数成反比。

$$TF-IDF=TF*IDF$$

词袋模型只统计词语是否出现或者词频，会被无意义的词汇所影响。因此本文选择 TF-IDF 模型对留言详情进行特征提取。

4.2 构建模型

在机器学习中，常用的分类模型有决策树，神经网络，支持向量机，朴素贝叶斯，逻辑回归等。在研究过程中，对于文本分类而言，我们根据数据量的大小、

特征值的差异选取了支持向量机、逻辑回归以及朴素贝叶斯模型进行了交叉验证并确定模型。

4.2.1 模型简介

支持向量机（Support Vector Machine，SVM）是一个二分类的模型，其决策边界是对学习样本求解的最大边距超平面；SVM 使用铰链损失函数计算经验风险并在求解系统中加入了正则化项以优化结构风险，是一个具有稀疏性和稳健性的分类器。SVM 还可以通过核方法进行非线性分类；在文字识别、文本分类、图像分类等模式识别问题中有得到应用。

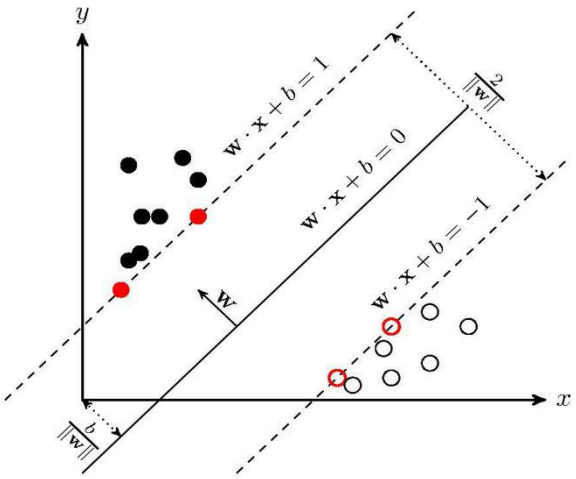


图 6　支持向量机图示

逻辑回归是一种简单，常见的二分类模型，通过输入未知类别对象的属性特征序列得到对象所处的类别。虽然带着回归的字样，但是逻辑回归属于分类算法，可以用来进行多分类操作；

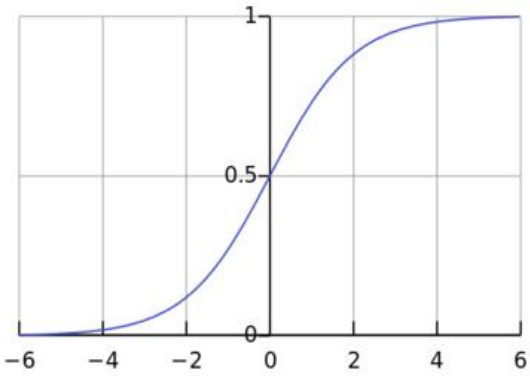


图 7　逻辑函数图示

$$h_\theta(x;\theta)=\frac{1}{1+e^{-\theta^T x}} \tag{1}$$

朴素贝叶斯算法[4]是一种以贝叶斯相关理论为基础的最常用的方法，也是最简单的贝叶斯分类器，具有很好的可解释性；在概率论与统计学中，贝叶斯定理表达了一个事件发生的概率，而确定这一概率的方法是基于与该事件相关的条件先验知识，而利用相应先验知识进行概率推断的过程为贝叶斯推断。朴素贝叶斯算法的健壮性比较好，对于不同类型的数据集不会呈现出太大的差异性。当数据集属性之间的关系相对比较独立时，朴素贝叶斯分类算法会有较好的效果。

$$P(h\mid D)=\frac{P(D\mid h)P(h)}{P(D)} \tag{2}$$

4.2.2 交叉验证

在上述三个模型的基础上，使用 F-Score 对分类结果进行评价。

$$F_i=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{3}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

4.3 模型评价

4.3.1 数据增强前

将处理后的数据按照 8:2 的比例划分为训练集和测试集进行验证，在对模型经过参数调整和分析后，得到各个模型的数据如下：

表 2　数据增强前各模型评估结果

模型	查准率 P	查全率 R	F 值
线性支持向量机	0.89	0.89	0.90
逻辑回归	0.89	0.89	0.89
朴素贝叶斯	0.87	0.74	0.81

4.3.2 数据增强后

得到各个模型的数据如下：

表 3　数据增强后各模型评估结果

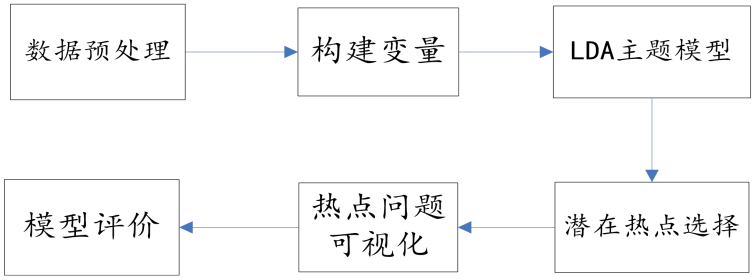
模型	查准率 P	查全率 R	F 值
线性支持向量机	0.91	0.91	0.91
逻辑回归	0.89	0.89	0.89
朴素贝叶斯	0.88	0.87	0.88

4.3.3 模型结果评价

通过对数据增强前后的线性支持向量机、逻辑回归和朴素贝叶斯三个模型对文本数据的查准率（P）、查全率（R）和 F 值的观察与比较可以看出：数据增强对线性支持向量机和朴素贝叶斯这两个模型的效果有不同程度的提升，其中对朴素贝叶斯的效果最为显著，查全率提高了 13%，F 值也有了 7%的提高。无论是否数据增强，逻辑回归的变化随数据增强变化并不明显，F 值保持在 0.89 的水平，而线性支持向量机的效果都是最佳，分类准确率达到了 91.27%，F 值达到了 91.19%。有着比传统的实验模型具有更好的效果，所以本文最终选择线性支持向量机模型作为分类模型。

五 热点问题挖掘

基本流程：



5.1 热点问题

随着网络技术的快速发展，互联网成为了信息的集散地，对这些信息进行分析，可以及时了解到当前社会人们所关心的热点。热点问题的识别一方面可以拉近政府与群众之间的距离，帮助政府及时了解群众的迫切诉求；另一方面可以及时了解社会动向，为政府各项决策提供有力的数据支持。在这样的需求下，热点话题挖掘（Topic Detection，TD）成为了一个极其重要的研究问题。

所谓热点问题，就是对一定来源的网络数据进行分析，综合利用统计、聚类

等方法从大量的数据中识别出被市民广泛讨论的热点。

在本文中，热点问题指某一时段内群众集中反映的某一问题。

5.2 基于 LDA 主题模型的热点问题挖掘

5.2.1 LDA 模型[5]

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。该模型是假设文档是由一系列潜在的主题混合而成，而主题是由词汇表中所有的词汇混合而成，不同的文档主要原因在于它们是由不同的主题按照不同的比例混合构建而成。

假如给定一个问题数据集 R 包含有 m 个问题数据 $\{r_1, r_2, \cdots r_m\}$，在这 m 个问题里面分布着 k 个主题 $\{t_1, t_2, \cdots t_k\}$，问题里面所有的特征词汇聚成一个特征词汇表 V 记为 $\{w_1, w_2, \cdots w_n\}$，词汇记号与该词所在的问题在整个数据集中的位置有关，那么词汇 w_i 在问题 r_m 这个位置的概率就可以表示为：

$$p(w_i \mid r_m) = \sum_{j=1}^{T} p(w_i \mid z_i = j) p(z_i = j \mid r_m) \tag{4}$$

其中 z_i 是潜在变量，表示词汇 w_i 的主题序号，$p(w_i \mid z_i = j)$ 表示词汇 w_i 被分配到第 j 个主题的概率，$p(z_i = j \mid r_m)$ 表示第 j 个主题在问题 r_m 中的概率。

这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了简单的数字信息，这简化了问题的复杂性，同时也为改进模型提供了可能。

5.2.2 基于 LDA 的热点问题识别

在热点问题挖掘的数据中，留言中的特征词是整个模型中可观察的唯一变量。LDA 模型在已知主题数目的情况下，通过调节特征词语在潜在主题上的概率分布完成每条留言的生成过程。在此过程中，可以获得每个特征词语在各个潜在主题上的概率分布情况以及每条留言在这些潜在主题上的概率分布情况；可以知道，如果一条留言在某个潜在类别上的概率分布值越高，那么这条留言成为该留言类别的可能性也就越大。如果某个特征词在某个类别上的概率分布值越高，说明此特征词对该类别的贡献也就越大，也就越有可能成为该主题的热点特征词语。

通过观察附件三数据发现，留言信息里包含着不同的地点，不同的人群，反映不同的问题。本文基于 LDA 主题模型，旨在提取大概率出现的“地名”词汇。

5.3 模型建立过程

5.3.1 数据预处理

本文将热度评价指标定义为：热度指数=点赞数-反对数。数据预处理的基本步骤为：

（1）数据筛选：针对附件三，筛选出热度指数大于 0 的数据，共计 1226 条留言信息，保存为“热度指数大于 0.xlsx”

（2）分词：通过 jieba 库对留言主题进行分词

（3）词性标注[6]**：**所谓词性标注就是为每个词的词性加上标注，也就是确定该词属于名词、动词、形容词还是其他词性的过程，它也是自然语言处理领域的基础，也是像机器翻译、信息检索等应用中一个不可缺少的环节。

考虑到需提取大概率出现的“地名”词汇，因此在词性标注后，只留下词性为地名(ns)，机构团体(nt),其它专有名词（nz)以及未知词性(x)的词语。而对于有些词语，jieba 不能准确地进行分词处理；例如：“五矿万境”会分成“五矿”、“万境”。于是构造用户自定义词典，将一些不应该分开的词汇添加进去，命名为“addword”。

（4）**去停用词：**删除掉在句中大量出现，却对语义分析没有帮助的词。

5.3.2 输入向量的构造

将预处理后的数据，先建立用户词典；根据用户词典，构造出一个形如[(0, 1), (1, 1)]的结构数据，一个中括号代表一条留言，一个括号代表一个特征词，括号里的第一个数字表示这个词在用户字典中的位置，每一个词在用户词典里都有唯一一个属于自己的位置，也就是说这个数字决定了这个词；括号里的第二个数字代表了这个词在这句话里面出现的次数。这样，我们就对 LDA 模型所需要的输入向量构造完成。

5.3.3 模型构建与结果

本文选取 gensim 库中的 LDA 方法对数据进行模型构建。

（1）主题数的确定

通过查阅资料，本文选择了一个数值定量评估的方法“主题相干性”[7](topic coherence）来对确定主题数。

人们对于主题模型的理解更倾向于属于同一主题的单词在语料库中共同出现的频率。“c_v topic coherence”做的就是这样的工作。gensim 提供了几种不同的主题相干性测量方法，其主要的不同在于“共现”的定义不同。Palmetto Online Demo 这里定义了几种不同的共现定义，其中 C_v,C_UCI,C_NPMI 为 gensim 所采取的可选方法。通常情况下，主题相干性数值越高，说明模型产生文档的能力越好，模型的推广性也就越好。

实验中，测试了主题数为 10，20，30，40，50 的情况，如图所示：

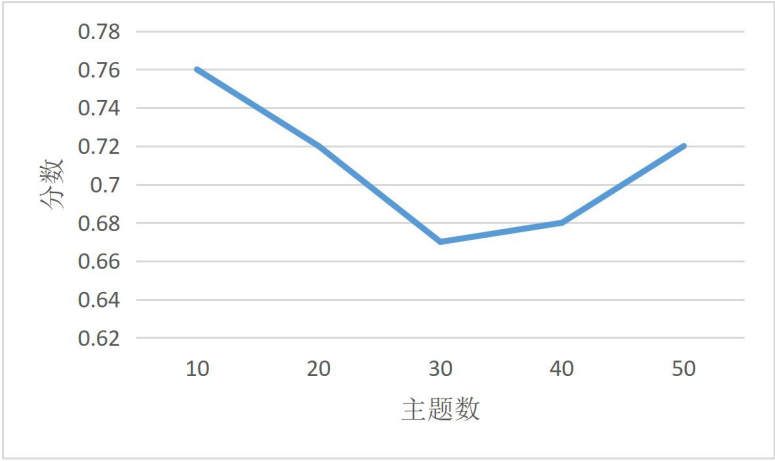


图 8　主题相干性曲线

由图可知，主题数为 10 时，主题相干性数值最高，此时模型的性能最佳。因此，确定主题数最优值为 10。

（2）模型结果

确定主题数后，剩下的 alpha 和 beta 参数我们根据经验值设定为 alpha = K/50 和 beta = 0.1；将向量导入模型中运行，得到了词汇表中的特征词语在 10 个潜在主题上的概率分布。通过观察输出结果，在潜在主题中寻找到可能的地名；根据潜在主题中出现的地名对附件三进行文本筛选，筛选出分别属于 22 个地点

的留言记录，最后根据热度指数的高低，选取了前五类问题作为热点问题，保存为“热点问题留言明细表.xls”。

5.4 热点问题可视化

针对热点问题明细表的五类数据的留言内容，绘制词云图。

（1）热点问题 1：

图 9　热点问题 1 词云图

（2）热点问题 2：

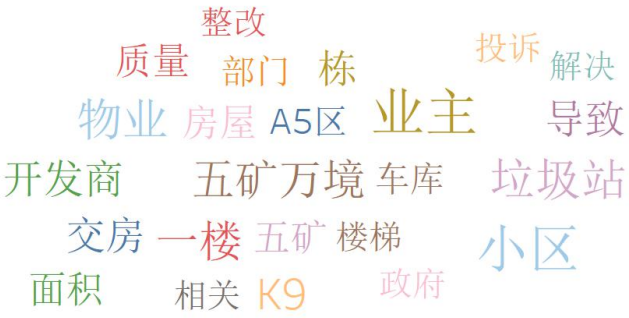


图 10　热点问题 2 词云图

（3）热点问题 3：

图 11　热度问题 3 词云图

（4）热点问题 4：

图 12　热点问题 4 词云图

（5）热点问题 5：

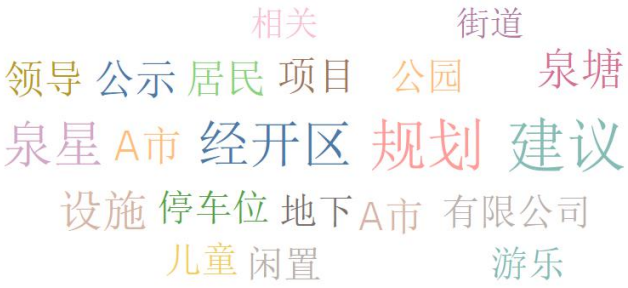


图 13　热点问题 5 词云图

根据五张词云图，归纳总结出地点以及问题，保存为“热点问题表.xls”

5.5 模型评价与建议

5.5.1 模型评价

LDA 主题模型，从效果上来看与聚类模型有不少相似的地方，主题模型输出的潜在主题是聚类中心，就像留言和多个类簇（类别）关联，对总结归纳留言集合比较有帮助。

与聚类算法相比 LDA 的不同之处在于后者对每条留言在主题上的分布提供了一个留言的简洁总结，在这个降维了的特征空间中进行留言比较，比在原始的词汇的特征空间中更有意义。LDA 模型还有一个好处，就是可以推断数据集的潜在主题数量，K 个主题即是 K 个数值特征，这些特征还可以被用在像逻辑回归或者决策树这样的算法中用于预测任务。

5.5.2 建议

本文使用 LDA 模型也有不足之处。

（1）有的留言文档太短，不利于 LDA 模型的训练；

（2）在对数据进行分词和词性筛选的时候，发现因为分词语料库的限制性，对文本数据的分词结果不太理想；例如“北山镇”被分成了“县北”，而“人民西路”、“楚龙西路”、“福元西路”都被分成了“西路”；这就导致了词性筛选的时候难免把部分“精华”丢弃，而又留下了部分的“糟粕”，由此影响了潜在主题的结果输出。

对于以上问题，我们建议可以通过地方政府对当地的特征词语收集与整理，例如各地区名，村落名，小区名等录入自己的数据库，形成一套属于当地，适用于当地的地方语料库，再结合分词语料库进行分词，相信这样无论是对分词结果还是模型的输出结果，效果都会有很大的提升。

六 答复意见的评价

通过查阅资料，我们了解到人们对数据质量评价体系的描述基本上是从准确性、时效性、相关性、客观性、可衔接性、完整性、可理解性、透明性、可操作性、可取的性、可解释性、效益性、安全性等方面展开的。

本文选取相关性，完整性，可解释性三个指标对答复意见进行评价。

1、相关性和完整性

本文采取 TF-IDF+余弦相似度的方法，去衡量相关性和完整性。

余弦相似度[8]：通过两个向量之间的夹角来衡量向量相似性。余弦值接近 1，夹角趋于 0，表明两个向量越相似，余弦值接近于 0，夹角趋于 90 度，表明两个向量越不相似。

$$\cos(\theta)=\frac{a-b}{\|a\|\bullet\|b\|}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}} \tag{5}$$

通过自定义余弦相似度函数，求出对应的答复意见与留言详情的相似度。

留言详情	答复意见	答复时间	相似度
2019年4月以来位于A市A2区桂花坪街道的A2区公安分局宿舍区景春华苑出现了一番乱象该小区...	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区景春花苑物业管理有问题"的调查核...	2019/5/10 14:56:53	0.372128
潇楚南路从2018年开始修到现在都快一年了路挖得稀烂用围栏围起一直不怎么动工有时候今天来台挖...	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问题,A3区洋...	2019/5/9 9:49:10	0.031951
地处省会A市民营幼儿园众多小孩是祖国的未来但民营幼儿园教师一直都是超负荷工作且收入又是所有行...	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"的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	0.409261
尊敬的书记您好我研究生毕业后根据人才新政落户A市想买套公寓请问购买公寓能否享受研究生3万元的...	网友"A000110735"：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	0.334450
建议将白竹坡路口更名为马坡岭小学原马坡岭小学取消保留马坡岭	网友"A0009233"，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议"白竹坡...	2019/5/9 9:51:30	0.553283

图 14：部分数据相似度

2、可解释性

关于可解释性这一部分，我们可以先收集准备数据，包括市民问题留言，地方政府相关部门回复，以及对应的满意度评分，具体实现如下：

（1）我们可以在用户的留言得到地方政府相关部门的解决回复之后，给用户提供一个评分选项，让用户按照对相关部门回复与处理的结果的满意度评分，分数跨度设置一到十分，满分十分为十分满意；

（2）在用户评分后如果不是十分满意，就可以再给用户提供一个多条选择题的政府处理问题反馈调查问卷，用来获取群众不够满意的原因，比如说是因为处理的时间太久还是处理的效果太差；

（3）在获取足够的数据之后，我们就可以对用户的问题留言，和相关部门的回复，以及满意度评分与反馈问卷的答案一起，先提取留言和留言回复的相关文本特征，对不同的留言特征以及回复特征建立基于评分的联系，以及对不同评分之下的反馈运用主成分分析的方法，寻找评分背后的可解释性规律。

3、评价方案

最后，我们根据以上所提出的相关性完整性和可解释性，建立适用于当地政府的留言回复评分系统，在系统对留言回复自动评分的同时，可以继续收集数据，来对评分系统提供更丰富的数据，以提高评价的准确率。

七 参考文献

[1] 王杨,许闪闪,李昌,艾世成,张卫东,甄磊,孟丹.基于支持向量机的中文极短文本分类模型[J].计算机应用研究,2020,37(02):347-350.

[2] 徐冠华,赵景秀,杨红亚,刘爽.文本特征提取方法研究综述[J].软件导刊,2018,17(05):13-18.

[3] 黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[J].软件工程,2020,23(03):1-3.

[4] 邸鹏,段利国.一种新型朴素贝叶斯文本分类算法[J].数据采集与处理,2014,29(01):71-75.

[5] 余传明,张小青,陈雷.基于 LDA 模型的评论热点挖掘:原理与实现[J].情报理论与实践,2010,33(05):103-106.

[6] 梁喜涛,顾磊.中文分词与词性标注研究[J].计算机技术与发展,2015,25(02):175-180.

[7] Michael Röder,Andreas Both,Alexander Hinneburg.Exploring the Space of Topic Coherence Measures.Pages 399–408,2015

[8] 武永亮,赵书良,李长镜,魏娜娣,王子晏.基于 TF-IDF 和余弦相似度的文本分类方法[J].中文信息学报,2017,31(05):138-145.