“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。因此，运用网络文本分析和数据挖掘技术对群众留言与意见的研究具有重大的意义。

首先介绍语义处理背景和近况，分析此次文本挖掘实践的必要性和可行性，并对这次使用到的一些工具库进行介绍。然后我们对两种主流的提取关键词的算法进行了分析，介绍了它们的特点与实现方法，为后文解决实际问题做适当的铺垫。

对于问题 1，对文本分类的处理过程为将附件 2 作为训练文本集并进行预处理，然后对同一个一级标签下的留言主题利用 TF-IDF 算法，TextRank 算法提取关键词作为特征，建立词典，然后再对附件 3 中的每一条留言主题利用 TF-IDF 算法、TextRank 算法和中科院计算所 NLPIR 语义分析系统分别提取关键词，综合三种方法构造留言分类的模型，对于这些模型，我们介绍了它们的主要思想，通过实验对这些模型进行评价并加以改进，最终确定了我们的分类模型。最后，我们总结了对于问题 1 的实现过程，得到最终的分类结果。

对于问题 2，选择了合适的方法来对留言按照特定地名或特定人群归类，其中使用到了文本相似度的原理对特定关键词计算相似度。分析了热点问题的特征，并基于此构造了热点指数公式，然后经过实验发现了该公式的不足，于是在此基础上又提出了改进的热点指数公式，综合热点问题的点赞数，反对数，留言数从而全面的评价了一类问题的热度。最后根据热度指数实现了热度的排序，根据一类问题的所有留言确定了热点问题的时间范围。

对于问题 3，先对附件 4 进行数据预处理，通过分析群众留言与答复意见之间的联系，确定了以关键词作为核心的评价标准，并且结合实际，将答复意见的相关性与完整性结合起来考虑，提出了对答复意见的相关性与完整性评价方案。同时，我们还考虑了答复意见的及时性，提出针对答复意见的及时性评价方案，从而对答复意见的质量更加全面的考虑。

关键词 TF-IDF 算法；TextRank 算法；NLPIR 语义分析系统；自然语言处理

一、问题重述

本次数据挖掘挑战赛的“智慧政务”中的文本挖掘，是利用自然语言处理和文本挖掘的方法处理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。从而缓解行留言划分和热点整理的工作量，提升政府的管理水平和施政效率。需要完成以下三个问题：

1. 根据已给出的人工对留言分类的数据，建立关于留言内容的一级标签分类模型。
2. 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，按格式给出排名前 5 的热点问题，还有具体留言信息。
3. 针对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 语义处理背景

2.1.1 必要性分析

2.1.1.1 自然语言处理技术

NLP (Natural Langunge Possns,自然语言处理)是计算机科学领域以及人工智能领域的一个重要的研究方向，它用计算机来处理、理解以及运用人类语言(如中文、英文等)，达到人与计算机之间进行有效通讯。所谓“自然”乃是寓意自然进化形成，是为了区分一些人造语言，类似 C++、Java 等人为设计的语言。从 20 世纪 40 年代算起，自然语言处理的研究已经有 60 多年的历史了，随着网络时代的到来，它已经成为了现代语言学中一个颇为引人注目的学科[1]。

NLP 被应用到了许多领域，例如机器翻译，文本分类，情感分析等，常被应用在医疗、金融、教育、司法领域。在未来，自然语言处理的发展趋势是 NLP 与更多的领域深度结合，为各行各业创造更多价值。同时，在研究领域，自然语言处理的重点将逐渐从词法分析到句法分析再到语义分析完成转换，也会紧密的结合人工智能发展。

2.1.1.2 关键词提取技术

关键提取在文本挖掘领域是一个很重要的部分，在自然语言处理领域，对于大量的文本数据，核心就是要把最关键的问题提取出来，而一般对于一段文本，往往可以通过几个关键词反映整段文字的主题思想，同时关键词提取的准确程度也直接影响了对自然语言处理的最终结果。在这次“智慧政务”中的文本挖掘应用中，我们实现的方法也与关键词提取密不可分。

如今常用的关键词提取算法有 TF-IDF 算法，TextRank 算法，这次文本挖掘应用中，我们除了使用到了上述两个提取关键词的算法，我们还使用了 nlpir 汉语分词系统。

2.1.2 可行性性分析

我们主要使用基于 python 的 jieba 和 pynlpir 两种分词工具包（下称 jieba、

NLPIR）。它们准确性较高，安装使用非常方便，可扩展性较好，能够满足大数据下对复杂语义的分析场景。因此我们选择了这个两个分词包作为辅助，利用这两个分词工具包中提供的各种功能，帮助我们完成文本挖掘的任务。

2.1.2.1 jieba 分词工具

Jieba 分词工具是一款 python 中文分词组件，支持简、繁体中文，高级用户还可以加入自定义词典以提高分词的准确率。它拥有不同的分词模式来满足不同的任务要求。并且它自带有关键词提取算法 TF-IDF 和 TextRank 两个，文章下面会对它们进行详述。

2.1.2.2 NLPIR 汉语分词系统

另外一个对中文分词有非常好的效果的是中科院研究的分词系统，也称 NLPIR 汉语分词系统，其主要功能包括中文分词；英文分词；词性标注；命名实体识别；新词识别；关键词提取；支持用户专业词典与微博分析。NLPIR 系统支持多种编码、多种操作系统、多种开发语言与平台。我们在 python 中使用 PyNLPIR 包来调用 NLPIR 分词功能，其中提取关键词功能采用采用交叉信息熵的算法自动计算关键词。

2.1.3 主要算法详析

2.1.3.1 TextRank 算法

TextRank 算法将文本中的词视作了图中的节点，将节点之间的边看作为语法关系，即如果两个词存在一定语法关系，则这两个词在图中就会有一条边相互连接，通过一定的迭代次数，节点就会有越来越多的边与之相连，这些边的数目就可以作为节点的权重，最终不同的节点会有不同的权重，权重高的语法单元可以作为关键词[2]。

TextRank 迭代计算公式为：

$$PR(V_i) = (1-d) + d * \sum_{j \in \mathrm{In}(V_i)} \frac{1}{|Out(V_i)|} PR(V_i)$$

TextRank 关键词提取算法为：

（1）把给定的文本 T 进行分词和词性标注处理，并过滤掉停用词，只保留指定词性的单词,组成候选关键词集 S=[$t_1, t_2, t_3, \ldots, t_n$],其中$t_i$为候选关键词。

（2）由 S 构建候选关键词 G = (V,E)图出现在一个窗口中的词汇之间相互形成一条边。

（3）根据迭代计算公式计算各个节点的权重。

（4）对节点权重进行排序，从而得到权重最高的一些词，作为关键词。

2.1.3.2 TF-IDF 算法

TF-IDF 算法用来衡量一个关键词所能提供的信息[3]。是根据出现频率来判断是否为关键词的。TF-IDF 计算关键词的步骤为：

（1）计算词频(TF)，词频表示关键词在文档中出现的频率。表达式见下：

$$词频(TF)=\frac{关键词出现的次数}{文本总词数}$$

（2）计算逆文档频率(IDF) ,反映关键词的普遍程度,当一个词越普遍,有大量文档包含这个词时，其 IDF 值越低；反之，则 IDF 值越高。IDF 如式：

$$逆文档频率(IDF)=\log(\frac{语料库总文本数}{包含该词的文本数+1})$$

3.计算关键词在文档中的 TF-IDF 值：

$$IDF\text{-}IDF=词频(TF)*逆文档频率(IDF)$$

总体来说，TF 刻画了某词对某篇文档的重要性，而 IDF 刻画了某词对整个文档集的重要性。当一个词在文档频率越高但是在整个文档集中的普遍率低的时候，它的 TF-IDF 值就会越高，所以通过 TF-IDF 算法，可以过滤一些常见词，从而保留包含更多关键信息的重要词。

2.2 问题 1 分析过程

2.2.1 数据处理

由于留言主题基本上能够概括整条留言的留言详情，反映留言的主要信息，并且考虑到留言详情中无关信息较多，内容较大，分析起来比较复杂，所以对于一条留言，我们直接将留言主题作为语料库。

为提取留言主题关键字做准备工作，提高关键字提取正确率，我们可以自行添加在 jieba 中构建词典，这些词将同样来自语料库，我们使用中科院计算所 NLPIR 语义分析系统对留言主题提取关键词，并且取权重大于 2 的作为新词加入词典中，这些词带有留言主题的一些特征，从而在接下来使用 jieba 提取留言主题的关键词过程中可以分出原始词库中没有的词以及优先分出一些词，使得结果更加准确。

提取留言主题的关键词，将每一个一级标签下的留言主题的关键词整合到一起，作为机器学习的结果，之后在识别到了这些关键词的时候能够将留言分类到对应的一级标签上。首先，先要对这些留言主题进行中文分词，将留言主题中的文本信息转换为计算机能够识别的结构化信息。然后提取这些分词中重要的词汇作为关键词。这里使用了 TextRank 算法，和 TF-IDF 算法，采用了两种提取关键词方式相结合的办法来找到关键词。对每一个留言主题而言，由于其句子较短，所以设置最大关键词数目为 4 就能较好的概括一个留言主题，并且只取词性为名词，地名，动词，动名词作为关键词，因为这些词性的词反映的信息量更多，更

切合比赛要解决的留言问题提取条件。

由于留言主题杂乱、多样，其中有的语句含义不明确或者存在歧义，导致只使用一种提取关键词的方式得到的结果存在一些无意义的词汇，这些词在后面的过程中会成为干扰项，所以取这两种方法得到的结果的交集作为关键词结果，可以有效地剔除那些无效词，从而获得有效的，含义明确的结果。

2.2.2 分类模型

这种分类模型比较容易实现，在一段留言主题中提取了 3 个关键词后，分别对这些关键词在所有的标签关键词词典中查重。具体步骤为：

（1）统计关键词出现位置，对第 i 个关键词有

$$w_i = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_{14} \\ a_{15} \end{pmatrix} \text{其中} a_k = \begin{cases} 1, & \text{第 i 个标签的词典包含} w_i \\ 0, & \text{第 i 个标签的词典不包含} w_i \end{cases}$$

（2）取$w_1 + w_2 + w_3$最终的结果，这 15 个数就表示相应的标签的关键词与此 3 个关键词的重复个数。

（3）根据结果的具体情况分类：

①有一个唯一最大值，取该数的序号作为分类结果，该序号表明分类到第几个标签

②有重复的最大值，由于取关键词时，3 个关键词是按照权重排序的，所以第一个关键词的权重最大，将第一个关键词的结果作为最终分类结果，如果第一个关键词出现在了多个标签下，那么该留言主题在分类过程中存在争议，列入无法分类的情况。

③如果最终结果全为 0，即这三个关键词在整个关键词词典中都没有找到时，列入无法分类的情况。

这种单纯比较关键词在每个标签下的出现的数量作为分类依据的方法比较容易实现，并且思路简单清晰，但是经过我们的实验也发现了不少的问题。

首先，从最终结果的分类标准中，我们可以看到会出现关键词出现次数相同的情况，总共从一条留言主题中只提取 3 个关键词，数量较少，导致最终结果被局限在了一个很小的范围内，所以很容易就被列入了无法分类的情况。

其次，当所有的关键词没有被匹配到的时候，结果会被列入无法分类的情况，但是我们还是尽量希望找到一些因素来为它完成归类。

最后，这种方法没有考虑到词典中标签关键词数量，比如“城市建设”这个标签，由于我们的训练文本中“城市建设”标签下的留言较多，导致训练得到的该标签下的关键词也相应的比别的标签关键词更多，所以在分类时，就更容易在“城市建设”标签下匹配到关键词，所以这类标签无疑就会在这种分类模型下占到优

势，相应的留言就会更容易的被分在城市建设下，其中就会出现不少的错误分类结果，降低查准率。

为了上一个模型产生的问题，我们使用了另外一个文本分类模型，多重伯努利模型。这种模型使用最大似然估计来估计概率，对于一个留言主题R_i，它属于类c_j的概率为：

$$P = \sum_{k=0}^{n} P(t_k|c_j) = \frac{df(t_k, c_j)}{N(c_j)}$$

其中: t_k表示R_i中的关键词，n 为从一个R_i中提取到的关键词数量，$df(t_k, c_j)$表示类别c_j含有词组t_k的关键词数量，$N(c_j)$表示类c_j下总词数。

经过实验发现，这种模型能够在一定程度上改善分类结果，平衡标签关键字词典造成的标签之间分类不平衡的现象。

然而这种估计方法还是会存在零概率问题，所以需要采用平滑技术来克服零概率问题，多重伯努利模型的贝叶斯平滑估计式如下：

$$P(t_k|c_j) = \frac{df(t_k, c_j) + \alpha_k}{N(c_j) + \alpha_k + \beta_k}$$

其中：α_k与β_k的选择取决于t_k。

一般的取$\alpha_k = 1$，$\beta_k = 0$，但是经过我们的大量实验发现这样的选取方式会出现更多的错误分类情况，经过我们不断测试α_k和β_k的取值，都发现无法达到预期的效果。

出现这种情况的原因是因为一个留言主题R_i中提取到的关键词数目太少了，导致每一个关键词都对最终的分类结果有着强烈的影响，对于这次的实验内容，不适合采用这种平滑技术，所以最终我们选择的方案还是简单的多重伯努利模型，接受出现零概率的情况。

2.2.3 实现过程

2.2.3.1 关键词处理

同样按照训练文本集预处理的方法对待分类的文本集处理数据，提取关键词。然而经过反复的实验发现对于一段留言主题这样的短文本很难提取所有有效的关键词。然而留言主题中的信息一般都集中在几个比较重要的关键词中，如果漏掉了任何一个关键词，这会严重的影响之后对于文本分类的正确率。

单一的使用某个算法获得的关键词作为分类的依据会漏掉留言主题中另外一部分文本信息。经过我们的实验发现，对于同一段留言主题，采用不同的提取关键词算法往往会得到不同的结果，所以为了尽量将一段留言主题中的所有有效

信息都作为留言分类中的决策因素，我们提出了一种新的提取关键词方法：

同时采用 pynlpir 的提取关键词算法，TextRank 算法，TF-IDF 算法对同一段留言主题提取关键词，分别按照这三类提取关键词的结果进行留言分类，最后综合考量它们的结果得到正确概率最大的可靠结果。

目前，大部分电子政务系统依靠人工根据经验进行处理，存在工作量十分巨大，效率较低，差错率高等诸多问题，对此本项目提出了合理的解决方案，结合 TF-IDF 算法，TextRank 算法和 NLPIR 语义分析系统提取关键词的方法，依次对 15 个一级标题进行了关键词的提取和分类，其准确性高，可以达到预期的要求。

特别的，这三种算法中 TF-IDF 优点是实现简单，相对容易理解。但是 TF-IDF 算法提取关键词的缺点也很明显，它严重依赖语料库，需要选取质量较高且和所处理文本相符的语料库进行训练。依赖于语料环境，这给他带来了统计上的优势——它能够预先知道一个词的重要程度而 TextRank 只依赖文章本身,它认为一开始每个词的重要程度是一样的。

结合这三种算法优缺点，我们对这三个算法提取出的关键词进行处理，即取交集，认为三种方法同时获取的相同的关键词是有效的，如果出现了一个词汇存在于多个一级标题，认为这个词汇不具有有效性，即认为该词汇失效，对此我们建立了关键词字典库，来进行分类处理。分类情况见下表：

表 1 关键词表格

城乡建设:	更换 建档 受害 生意 改组 购房 群租房 龙潭 老峰 划拨 村民 项目 荣县 工地 属区 大道 不善 无烟 渔米 四区 华庭 塘镇 摆摊 铁线 秀峰 创卫 入户 验收 市松 茶场 转为 群租 公园 单厂 省道 颐和源 规门 城星 星镇 收益 政府 桥镇 种菜 泥巴路 通道 挡住 黄花 市岩 修噪 费用 停车 修剪 增加 市环 环卫局 要加 装防 冲棚 澧源镇 不合格 抵押 北路 提取 条例 私房 滨湖 市珍 改征 电梯 溪乡 办公楼 站点 市阳光 办事处 东段 游乐设施 无端 活动 没钱 经济房 上露 装空用地 故事 路展览馆 瘫痪 输送 农校 盗网 成麻 评标 属院 市康 水湾 事业局 洪道 篡改 公共场所 池子 冲村 索赔 权益 经典 吹倒 惠州 金城 市领域 消防车 双江 江东 体育场 门卫室 街棚 锦苑 养老推进 方塘 江北 环卫 符合 供应 帐子 铺设 安沙城北 领御 家属区烧开 市马家 火炬 厅长 爱心 方面 农村 沙坪 健身 解决 不用 兑现人群 山区 衙门 绿地 大员 渗水 路段 口镇 开标 东路 实事求是 等（未一一列举，详见 keyword_dic.txt）

环境保护	受害 经东 家园 村灌 资源 房燃煤 首镇 沥青 取缔 村石 县城 设备 行政 村民 噪音扰民 水生 青山 河长 请求 工地 属区 大道 区映 爱村 炼油厂 工业园 区岩 推广 偷放 砂石料 辐射 市磷 塘镇 企业 枫山 文桥 测试 中波 公园 镇大河 制衣厂 发射 新龙花园 电解厂 政府 桥镇 景湾 村养猪 紫金 毒灰尘 玻璃 洞养 区烟厂 大源 冷区 双井 市二 水泥路 市堤 沸石场 雄村 漏油 土局 村造纸 市环 村口 网网 脱有 关停 冠市 楼顶 业园 田地 污口 排废 营业 区朝 毒烟 命脉 大河洲 北路 厂污水 方包 铁艺 水泥 环绕 编织袋 大庆 老湾 青云 电梯 气味 扰民 喷漆 县苗 陶粒厂 香林村 罗塘村 喷油 市油 耕种 仙桥村 滑石 背村 木工厂 塑料厂 村长 公路 松林村 反映 变电 联造 柴油 冷头 镇友 搅拌场 衣厂 洗涤 信采 市络 嘉乐园 生存 压件 制止 青采 跪求 混凝 市力 化学剂 排乱 冲村 山界 大园 光药厂 厂重 等（未一一列举，详见 keyword_dic.txt）
交通运输	库堤 糊路 区茶 市经济 全国 县马 盲区 管管 寄送 中途 燃气 塔边 运输 超载 沥青 龙潭 抽扣 通水 卖手 圆通 县城 大力度 站候 路铺 汨湖 村民 罢工 邮政 污水 局网 呼吁 形式 请求 岐山 大道 市中心 流程 自用 预备 工业园 快递 天易 喊价 环境 整顿 软件 规范 攸快 向取 打印 资格 市普通 货车 不到 工行 大桥 任务 打不到 汇报 政府 桥镇 花路 缴费 农路 新塘 六省 出台 工板 排水 桥路面 人民 问题 通道 坝路 上涨 县滨 消耗 运营权 记为 业资格 认字 费用 水产局 市民 停车 扰乱 音屏 能取 增加 组乡 土运输车 领到 合伙 挂失 县原 驾驶员 侵占 重开 开通 打通 任意 县应 借母 硬化长 读表 加价 县人 标路 设计 资江 冲断 普利桥 县五 冲采 湄江村 等（其中未一一列举，详细信息请查看 keyword_dic.txt）
……	（未一一列举，详见 keyword_dic.txt）

根据上述关键词，项目比较的是所含关键词数量的多少进行判断的，如果考率的文字中含有该标签下词汇越多，证明和该标签紧密程度越高，反之，则认为紧密性越差。如果存在两个一级标题中词汇数量一样多，则认为本次判断失校，即认为是未查全的情况。

$$\begin{cases} max\{\alpha_1,\alpha_2,\ldots,\alpha_n\}\text{存在} & \text{有效} \\ max\{\alpha_1,\alpha_2,\ldots,\alpha_n\}\text{不存在} & \text{失效} \end{cases}$$

同时考虑到可能会出现查错的情况，即当某一留言主题和某一留言详情包含其他种类词汇过多或者关键词提取不够全面因而出现查错的现象。此次与准确率相关，是这次实验的重要的考察标准之一。

本次实验采用了两种方法进行实验，第一种方法是只考虑留言主题信息，认为留言主题一般是包含了主要的信息，是留言信息概括信息，将能够有效的把信息全面覆盖，认为可以达到实验的预计要求，并且大大减小了实验运行时间。第二种方法是综合考虑留言主题和留言信息，并且考虑包含关键词的方法，认为包含的关键词越多，则认为相关性越高，越有可能与那些主题相关，结合关键词的

个数大小，最终确定该主题的类别，这种方法有效的提高准确性，但是程序运行时间显著增加。

2.3.3.2 训练与评估

我们根据 F-score 对分类方法进行评价，计算 F1 值。

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

其中方法一准确率相对较低，经过改进的方法后，方法二在准确率有了显著的提高，能够达到实验的目的，准确率和召回率是互相影响的，理想情况下肯定是做到两者都高，但是一般情况下准确率高、召回率就低，召回率低、准确率高，两者基本上是出现此消彼长的情况。

方法一：

表 2 一级标签与测量结果信息（方法一）

一级标签数量	9210 non-null object
测试结果数量	8938 non-null object
一级标签主题	城乡建设 环境保护 交通运输 教育文体 劳动和社会保障 商贸旅游 卫生计生
测试结果主题	城乡建设 环境保护 交通运输 教育文体 劳动和社会保障 商贸旅游 卫生计生

表 3 一级标签与测量结果个数（方法一）

一级标签		测量结果	
城乡建设	2009	城乡建设	1839
环境保护	938	环境保护	478
交通运输	613	交通运输	392
教育文体	1589	教育文体	731
劳动和社会保障	1969	劳动和社会保障	340
商贸旅游	1215	商贸旅游	261
卫生计生	877	卫生计生	216
总体	9210	总体	8938

表 4 F-score 计算方法（方法一）

测试结果		
标签类型	准确率	召回率
城乡建设	0.9138	0.9701
环境保护	0.5095	0.9712
交通运输	0.6394	0.9787
教育文体	0.4600	0.9540
劳动和社会保障	0.1726	0.9162
商贸旅游	0.2148	0.9547
卫生计生	0.2462	0.9418

总体	0.4622	0.9704
F-score 计算结果	0.5827	

对于方法一来说，整体准确率偏低，召唤率较高，但是计算的 F-score 整体来说情况一般，主要是教育文体，劳动和社会保障，商贸旅游，卫生计生准确率过低，但是整体结果不理想，所以基于方法一，方法二有所改进，针对不再是留言主题信息，而是更加全面的留言主题和留言详细，其样本更大，准确性更高，更加符合真实情况。

方法二：

表 5 一级标签与测量结果信息（方法二）

一级标签数量	9210 non-null object
测试结果数量	9210 non-null object
一级标签主题	城乡建设 环境保护 交通运输 教育文体 劳动和社会保障 商贸旅游 卫生计生
测试结果主题	城乡建设 环境保护 交通运输 教育文体 劳动和社会保障 商贸旅游 卫生计生

表 6 一级标签与测量结果个数（方法二）

一级标签		测量结果	
城乡建设	2009	城乡建设	1856
环境保护	938	环境保护	558
交通运输	613	交通运输	145
教育文体	1589	教育文体	1328
劳动和社会保障	1969	劳动和社会保障	1772
商贸旅游	1215	商贸旅游	552
卫生计生	877	卫生计生	581
总体	9210	总体	9210

表 7 F-score 计算方法（方法二）

测试结果		
标签类型	准确率	召回率
城乡建设	0.9238	1
环境保护	0.5948	1
交通运输	0.2365	1
教育文体	0.8357	1
劳动和社会保障	0.8999	1
商贸旅游	0.4543	1
卫生计生	0.6625	1
总体	0.7375	1
F-score 计算结果	0.7375	

调整了计算方法，采用关键词包含的方法，有效的提高了准确率，召回率也

有显著的提高，即使运行的时间增加，但是能够达到预期的目标。

综合上述的两种方法，我们可以了解到商贸旅游的整体准确率较低，而城乡建设准确率较高，其主要原因在于关键词字典库的大小以及准确性，在接下来实验中将有效的改进字典库，以求进一步提高准确率。

由于我们采用了三种提取关键词算法，根据这三种结果分别按照分类模型对留言主题分类，最终得到了 3 中分类结果。将每一个结果看作一个分类建议，通过下面的逻辑确定最终的分类结果。

1)有相同的分类建议，该标签为最终结果。

2)没有相同的分类建议，列入了无法分类的情况。

这种分类方式的思想简单，简言之，即少数服从多数。同时经过我们的测试发现，这种方式也是得到分类正确率最高的方式。

2.3 问题 2 分析过程

2.3.1 问题 2 模型的建立

为了找到热点问题，简化分类过程，就需要通过对留言的特征进行分析，找出最为关键的特征。由于这是向政府进行的留言，因此文本中最具特征性的关键词就是地点和人物。所以对于问题识别，如何从众多的留言中识别出相似的留言，应首先确定该问题的针对的特点地点与人群，再按照这些关键词将所有留言进行分类。为了精确地将特定地点或特定人群的数据归并，需要对地点关键词做进一步的处理，定位到某市某地区，然后再在按照特定地点/人群分类，按照关键词识别出相似问题。

针对时间范围的提取，首先记录关键词重合度高的“词条”的行标，建立热点事件和多条留言的映射，再对一个热点所对应的多条留言日期进行比较，选择时间最远和最近的两个日期作为时间范围。

热度指数则是综合点赞、反对和问题出现频率做出综合研判，小组按照简单到复杂的逻辑对热度指数公式进行不断地优化。

2.3.2 数据处理

特点地点关键词的处理方法为：获得所有的地点、人群最小细分化关键词$(A_1, A_2, A_3, \dots, A_n)$,将所有留言与这些关键词匹配，得到留言$R$的关于特定地点、人群关键词的向量$M = (a_1, a_2, a_3, \dots, a_n)^T$,其中

$$a_i = \begin{cases} 1, A_i \in R \\ 0, A_i \notin R \end{cases} (i = 1,2, \dots n)$$

特定地点/人群的分类方法为：

①将留言R_1作为特定地点/人群组 1，该组的关键词向量为M_1。

②将留言R_2的M_{R_2}与M_1比较，如果$\left|M_{R_2}\right| = |M_1| = M_{R_2} * M_1$，即$M_1$与$M_{R_2}$完

全相同，那么说明这两个留言属于同一特定地点/人群，R_2也属于组 1。否则，R_2记作组 2。

③按照上述过程依次比较留言R_3，R_4，…，遍历所有的留言，得到分类后的特定地点/人群组 1，2，…。

对于那些有精确的特定地点的留言，可以初步确定其反映的问题一般都为同一问题，则可以直接确定它们的热点问题 ID，而有一些留言的地点描述没有那么精确，但是它所反映的问题可能已被确定为前面的某一组了。比如留言 1 中表示地点的“A 市中南大学楚雅医院”和留言 2 中“楚雅医院”，反映的是同一个地方，但仅通过特定地点无法将它们分到同一组，这个时候就需要借助其他的办法来将该留言准确的归类。

为了进一步提高分类的精确度，在上述的留言分类过程中添加步骤，如果$\left|M_{R_i}\right| > \left|M_j\right| = M_{R_i} * M_j$，即$R_i$的特定地点/人群的关键词包含有第 j 组的全部关键词，而且比第 j 组的全部关键词更多，那么将第 j 组的关键词向量替换为M_{R_i}，R_i添加到第 j 组中。而当$\left|M_j\right| < \left|M_{R_i}\right| = M_{R_i} * M_j$时，即$R_i$的全部的特定地点/人群的关键词都属于第 j 组的全部关键词，但不包括第 j 组的全部关键词，这时R_i就有可能属于第 j 组，我们需要对问题关键词进行分析，通过对问题的内容判断是否R_i属于第 j 组[4]。提取出已经分好组的问题关键词的向量$N_j = (b_{1j}, b_{2j}, b_{3j}, \dots, b_{nj})^T$。由于同一组中，留言之间会有重复的关键词，为了避免干扰，需要对这些关键词归一化。得到的向量N不含有重复的关键词。取当前留言的关键词的向量$N_{R_i} = (b_{1R_i}, b_{2R_i}, b_{3R_i}, \dots, b_{nR_i})^T$。然后利用余弦相似度计算关于留言具体内容的关键词相似度。

$$Sim(N_j, N_{R_i}) = \frac{\sum_{k=1}^{n}(b_{kj} * b_{kR_i})}{\sqrt{\sum_{k=1}^{n} {b_{kj}}^2} * \sqrt{\sum_{k=1}^{n} {b_{kR_i}}^2}}$$

利用两个向量的夹角的余弦值来衡量两个文本间的相似度，当该值达到一定大小后，可以推定R_i属于第 j 组，这些留言反映的是同一个问题[5]。

2.3.3 实现过程

热度指数评价指标：

一类留言问题的热度，最简单的评价热度指数的方式就是根据点赞数与反对数的多少确定。于是我们可以初步定义热度指数公式：

$$heat_i = \frac{\sum_{k=1}^{n}(agree_k + disagree_k)}{100}$$

其中$heat_i$表示第i类热点问题的热度指数，n表示第i类热点问题的留言总数，

$agree$和$dissagree$分别代表某一条留言的点赞数与反对数。

按照这种热度指数公式，经过实验发现许多点赞数和反对数为 0 的留言没有被考虑进来。某种社会问题可能被群众反复留言，但是网络上发表得到的反响并不一定很大。该公式却忽略掉了这类问题，因此需要对某类（潜在的热点问题）问题的出现频率进行描述从而优化指标计算公式。此外，考虑实际情况下，点赞和反对相对比较容易，而当一个问题真的达到一定严重程度的时候，群众才会去发表一个留言来反映问题，所以一条留言对于热度指数的提升是远远大于对留言的点赞或反对的。并且考虑热点问题往往与几个关键词紧密联系，在热点指数中需要加入对于留言本身内容的影响因素，而不是仅仅根据点赞反对数判断是否为热点。所以我们提出了改进的热度指数公式

$$HEAT_i = \frac{\max(heat_k)}{100}$$

其中i表示第i类热点问题，k表示该类热点问题的第 k 条留言

$$heat_k = agree_k + disagree_k + heat\_rate_{ik}$$

其中$heat_rate_{ik}$表示第i类热点问题中第 k 条留言的关键词匹配权值。关键词匹配权值$heat_rate_{ik}$的计算方法与该留言的关键词$words_k$和高频关键词hf_words有关。

得到留言的关键词$words_k$和高频关键词hf_words的过程为：

1) 分别使用三种提取关键词方法对留言提取关键词得到w_1，w_2，w_3。
2) 取$w_1 \cap w_2 \cap w_3$，得到$words_k$。
3) 按照步骤 1，2 遍历所有留言，取$words_1$，$words_2$，…，$words_n$的并集，得到hf_words。

通过取$words_k$和hf_words的交集，我们就可以直到在一条留言中，它提到的高频关键词有多少个，从而模拟对于留言内容的考察对热点指数的影响。

我们衡量$heat_rate_{ik}$的标准见表 1：

表 8　$heat_rate_{ik}$取值标准

$words_k \cap hf\_words$	$heat\_rate_{ik}$
8 个	50
7 个	45
6 个	40
5 个	35
4 个	30
3 个	25
2 个	20
1 个	10

经过多次实验发现，按照上述取值标准得到的热点指数最能符合我们的预期结果。细化关键词数目的原因，一个是考虑到关键词的作用是文本的核心意义，hf_words本身包含了所有高频词汇，但是这些词汇并不重复。如果某一组关键词的命中率极高，说明这组关键词所代表的问题反复以不同角度提出，侧面说明它是热点问题。二是由于赋值区间较大，为了防止出现大量留言问题热度指数相同的情况，对关键字数目进行了细化。

热度排名：

统计并且计算了热点问题的热度指数之后，我们对这些热度指数按照大小排列的顺序即可得到热点问题的热度排名。

时间范围：

解决时间范围，关键在于建立良好的映射关系，代码中巧妙引入了 page 和 hits 两张表，从而对列表下标，Excel 行标和热点问题 ID 形成双重映射。对映射表遍历同一热点问题的每一条留言，找到最早的留言与最近的留言，将读出的字符串进行格式转化，生成 datetime 对象，很方便的对各个留言时间比较，得到一类热点问题的时间范围。

2.4 问题 3 分析过程

2.4.1 问题 3 模型的建立

在问题 3 中同样需要使用关键词提取方法。此外，答复意见不仅仅要囊括留言主题中的地点、人群和事件关键词，还应该个性化、专业化地解决留言详情中的具体问题，落实到具体的人和事，从而避免答复的片面和敷衍情况。因此单凭留言主题所提取的关键字会遗漏大量细节，为了更好的比较留言与答复的相关性，必须要考虑留言详情的内容。

鉴于使用关键字高效地评价答复意见的优秀程度，小组采取以下评判策略：

①通过计算留言主题与留言详情的关键词和答复意见的关键词的重合程度来判断答复的相关性，决定是否有可解释性。

②通过关键字提取数目的多少以及切合程度反映的回复字数来判断完整性是否较好。

③通过判断留言时间和答复时间的距离来评判答复的及时性。

完整性是在相关性的基础上划分的，及时性占总分的 40%，相关性占 60%，而完整性决定着相关性分数档中的取值。

依照前面解决问题的思想，为了提高关键词的提取正确率，使提取到的关键词更加全面，我们需要为 jieba 添加自定义的词典。这个过程在数据预处理阶段进行。

考虑到留言详情和答复意见更加口语化，包含的信息分布相比留言主题更加

的分散，为了节省存储空间和提高提取关键词效率，引入停用词在这个阶段就显得十分重要，来自动忽略某些字或词。

2.4.2 数据预处理

首先对所给数据清洗，由于 Excel 表格中两列日期（留言时间、答复时间）格式非文本，因此在读取过程中会出现格式不统一的现象。先将两列日期单元格更改为“文本”，并且将不同格式的日期进行修正，方便读入程序。

将留言主题、留言详情、答复意见的文本经过分词处理，人为粗筛出一些高频词语和无实意的词语片段或单字，并将其分别添加入用户词典和停止词词典中，从而提高关键字提取效果。

2.4.3 实现过程

文本关键词提取：

按照前面的思想，我们使用中科院分词包 pynlpir 的提取关键词算法，TextRank 算法，TF-IDF 算法对同一段文字提取关键词来保证提取关键词的准确性和完整性。使用中科院分词 pynlpir 的提取关键词算法提取的关键词列表记为w_1，使用 TextRank 算法提取的关键词列表记为w_2，使用 TF-IDF 算法提取的关键词列表记为w_3，使用使用它们的交集作为最终提取关键词结果：

$$w = w_1 \cup w_2 \cup w_3$$

按照上述方法，我们分别对留言与答复的文本提取关键词，将留言主题与留言详情整合到一起，提取关键词得到w_q，对相应的答复意见提取关键词得到w_a。

相关性与完整性评价：

评价答复意见的相关性，我们给出的评价方案主要考虑w_q与w_a的重合程度判断标准，并将其分为了以下四个层面：

$$\begin{cases} w_a \cap w_q > 10, \text{相关性很强} \\ w_a \cap w_q \in (8,10], \text{相关性良好} \\ w_a \cap w_q \in (6,8], \text{相关性一般} \\ w_a \cap w_q \in (3,6], \text{相关性较差} \\ w_a \cap w_q < 3, \text{相关性很差} \end{cases}$$

而从完整度的角度出发，我们考虑主要元素的就是w_a的数目，需要注意的是完整度不能仅凭关键词w_a的数量$N(w_a)$决定，例如有的答复意见泛泛而谈，但是却与群众的留言毫无关系，这时$N(w_a)$虽然大，但是相关性却很差，所以我们需要结合相关性和完整性来统一的考虑一个答复意见的质量，我们给出的评分标准如下表：

表 9 相关性与完整性评分标准

$w_a \cap w_q$	$N(w_a) \leq 20$	$20 < N(w_a) \leq 30$	$30 < N(w_a)$
小于 3 个	10	10	20
3 个到 6 个	10	20	30
6 个到 8 个	20	30	40
8 个到 10 个	30	40	50
大于 10 个	40	50	60

从评价方案中可以看出，我们整体对答复意见的相关性与完整性进行了考虑，并且相关性的作用大于完整性，这也合情合理，更加灵活变通。当$w_a \cap w_q$过小时，我们就会怀疑答复意见的质量，有理由的认为答复没有针对留言进行回复，而在回避问题。当$w_a \cap w_q$较大时，说明留言与答复整体贴合较紧密。这时$N(w_a)$越大，即答复的篇幅越长，我们认为答复就更加的充分、完整、全面，所以给出更高的分数。

答复的及时性评价方案：

除了答复意见的相关性与完整性，答复意见的质量还与答复意见的回复时间有关。答复时间及时，群众反映的问题更好的被发现和解决，也更有利于增加群众对政府的办事效率信心。由于给出了群众留言的时间与答复意见的回复时间，所以我们可以考量答复的及时性对答复意见的质量做出，具体标准见表 10：

表 10 答复间隔时间评分标准

间隔时间	分数
小于 14 天	40
14 至 30 天内	20
大于 30 天	0

一般把 14 天作为两个工作周，如果两个工作周内完成答复，符合政府部门正常的工作周期。但如果拖欠到 14 天以后，就认定为没有及时恢复群众的需求，需要一定程度上扣分。如果一个月后仍不恢复，说明办事效率过低，没有及时回复群众的关切，因此在及时性这个环节不得分。

参考文献

[1]冯志伟.自然语言处理的历史与现状[J].中国外语,2008(01):14-22.

[2]张莉婧,李业丽,曾庆涛,雷嘉丽,杨鹏.基于改进 TextRank 的关键词抽取算法[J].北京印刷学院学报,2016,24(04):51-55.

[3]叶雪梅,毛雪岷,夏锦春,王波.文本分类 TF-IDF 算法的改进研究[J].计算机工程与应用,2019,55(02):104-109+161.

[4]程锦彬,钱钢.基于用户活跃程度的网络话题热度计算[J].江苏科技信息,2013(02):25-29.

[5]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.

[6] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Association for Computational Linguistics, 2004.

[7] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. 2004, 60(5):493-502.