“智慧政务”中的文本挖掘应用

摘要：近年来，各类社情民意相关的文本数据量不断攀升，这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，为此，本文利用自然语言处理和文本挖掘的方法解决群众留言分类、热点问题挖掘、答复意见评价这三个问题。

针对问题一：首先将附件 2 中的数据进行了数据抽取、去重、中文分词、以及停用词过滤等数据预处理。然后利用 TF-IDF 权重法提取特征词，构造词汇-文本矩阵。再通过多项式朴素贝叶斯模型和支持向量机模型对留言详情进行分类。将训练后的两个模型分别测试集进行预测，得出的准确率的 88.03%和 90.43%。所以选择支持向量机模型对留言内容做标签分类。

针对问题二：首先把某一时间段、某一特定地点、某一特定人群的相同留言问题归并，对附件 3 中的数据做预处理操作，得到文本的特征项向量，然后统计特征项词频并分配权重，通过使用向量间的夹角余弦值方法计算文本间的相似度。最后建立评价指标并量化指标，给出了排名前 5 的热点问题。

针对问题三：首先对群众留言反映的问题与相关部门的答复意见这二者的文本信息数据做关联分析、碰撞比对等处理，从而设立一套完整的评价体系来评价相关部门对群众留言的答复意见的质量。

关键词： TF-IDF 权重法 支持向量机 夹角余弦值 相关性分析

一：群众留言分类

1.1 问题一流程图

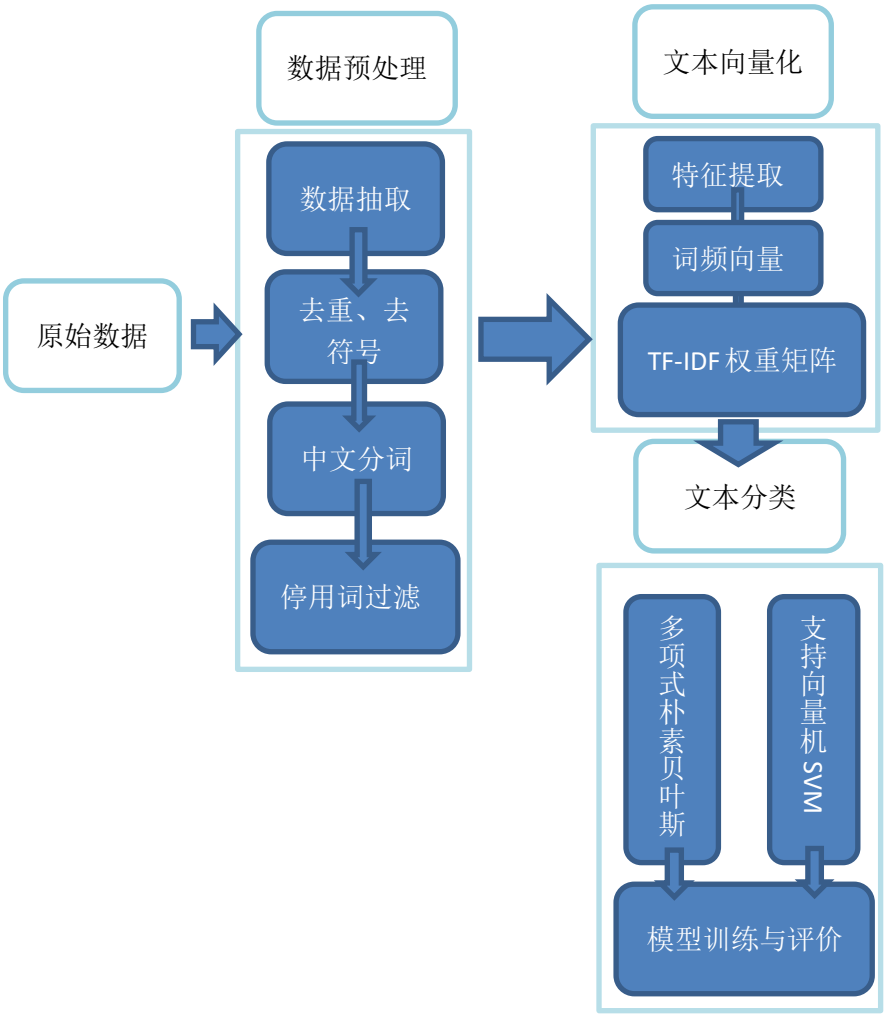

图 1-1 流程图示意

1.2 数据预处理

1.2.1 数据抽取

通过观察数据，发现一级标签的分布并不是均衡的。为了避免出现欠抽样和过抽样的问题，我们对一级标签分类各抽取了 300 条数据，抽取后数据共 2100 条。

一级标签数据分布情况如 1-2：

图 1-1 一级标签数据分布情况

同时我们将一级标签转换成 id，以便后续进行模型的训练。

一级标签转化成 id 对应情况如 1-3：

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

图 **错误!文档中没有指定样式的文字。**-3 一级标签转化成对应 id

1.2.2 对留言详情去重

人们在进行留言时可能会出现重复提交的情况，因此需对留言详情进行去重，保留其中一条即可。去重后数据剩余 2087 条。另外在留言详情中存在非文本字符，因此采用正则表达式将其去除，从而节省存储空间和数据处理时间。去除非中文字符后部分效果如 1-4：

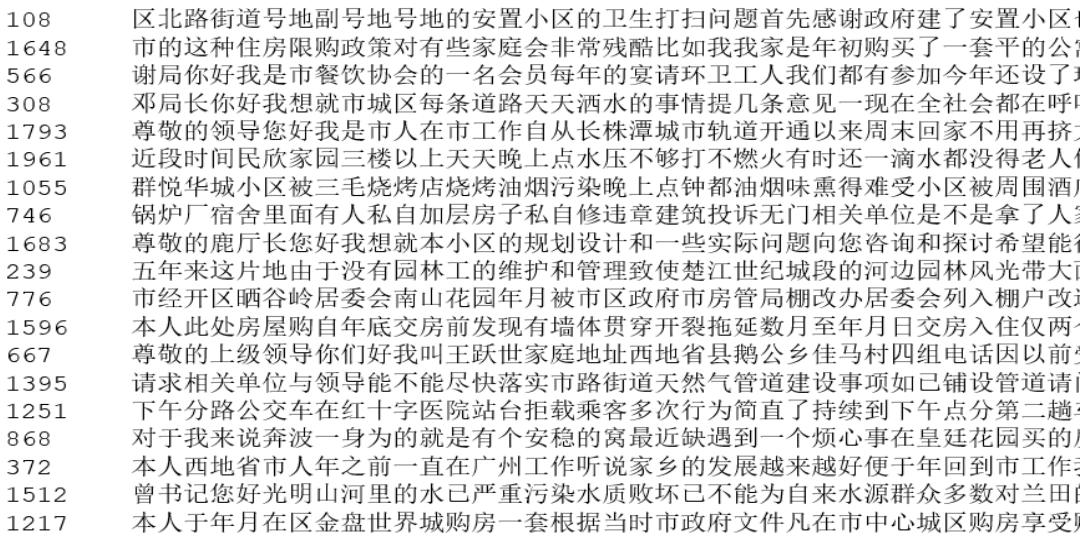

图 **错误!文档中没有指定样式的文字。**-4 去除非中文字符部分效果

1.2.3 对留言详情进行中文分词

为了对文本进行特征提取和向量化表示，我们需将处理后的文本数据进行中文分词。我们采用的是 python 中的 jieba 库进行中文分词。Jieba 采用了基于前缀词典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 (DAG)。同时采用了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。对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法，使中文分词能达到一个好的效果。分词后部分结果如图 1-5：

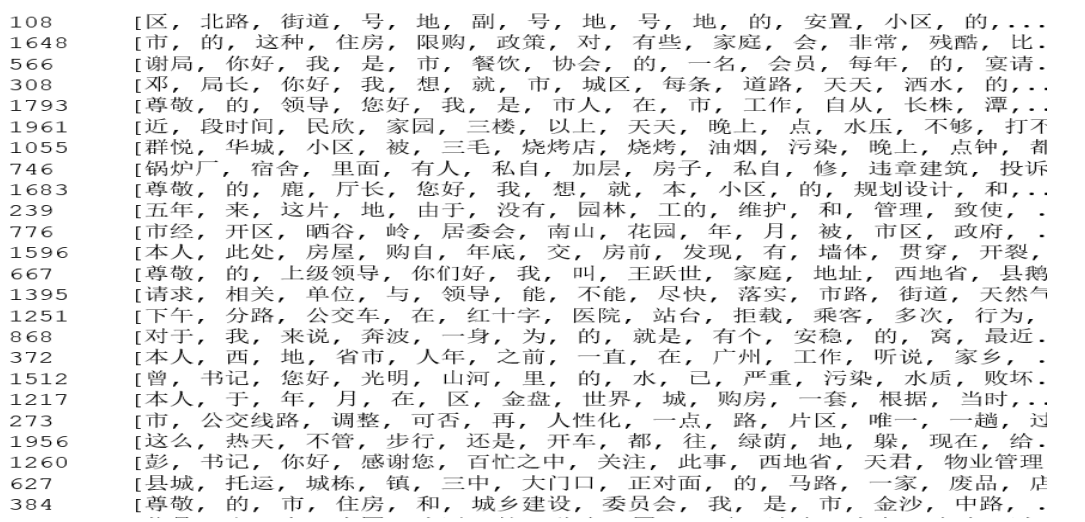

图 **错误!文档中没有指定样式的文字。**-5 部分分词效果

1.2.4 对留言详情去停用词

分词后的结果中，存在大量限定词。这些词语只是用来对文本的名词进行概念表达，而实际没有任何含义，且对文本分类无任何帮助。在文本中去掉这些停用词能够使模型更好地去拟合实际的语义特征，从而增加模型的泛化能力。

去除停用词部分效果如图 1-6

[区， 北路， 街道， 号， 副， 号， 号， 安置， 小区， 卫生， 打扫， 感谢， 政府...
[市， 住房， 限购， 政策， 家庭， 残酷， 我家， 年初， 购买， 一套， 平， 公寓楼...
[谢局， 市， 餐饮， 协会， 一名， 会员， 宴请， 环卫工人， 参加， 还设， 环卫工人.
[邓， 想， 市， 城区， 每条， 道路， 天天， 洒水， 事情， 提， 几条， 意见， 全...
[市人， 市， 工作， 长株， 潭， 城市轨道， 开通， 周末， 回家， 不用， 再挤， 大...
[段时间， 民欣， 家园， 三楼， 天天， 晚上， 点， 水压， 打不燃火， 一滴水， 没得.
[群悦， 华城， 小区， 三毛， 烧烤店， 烧烤， 油烟， 污染， 晚上， 点钟， 油烟味，..
[锅炉厂， 宿舍， 有人， 私自， 加层， 房子， 私自， 修， 违章建筑， 投诉无门， 相.
[鹿， 厅长， 想， 小区， 规划设计， 咨询， 探讨， 希望， 解答， 小区， 廉租房， ...
[五年， 这片， 园林， 工的， 维护， 管理， 致使， 楚江， 世纪， 城段， 河边， 园...
[市经， 开区， 晒谷， 岭， 居委会， 南山， 花园， 市区， 政府， 市， 房管局， 棚...
[房屋， 购自， 年底， 交， 房前， 发现， 墙体， 贯穿， 开裂， 拖延， 数月， 年月...
[上级领导， 你们好， 王跃世， 家庭， 地址， 西地省， 县鹅公乡， 佳马村， 四组， 电
[请求， 相关， 单位， 落实， 市路， 街道， 天然气， 管道， 建设， 事项， 铺设， ...
[下午， 分路， 公交车， 红十字， 医院， 站台， 拒载， 乘客， 持续， 下午， 点分，..
[奔波， 一身， 有个， 安稳， 窝， 缺， 烦心事， 皇廷， 花园， 买， 房子， 配套，...
[西， 省市， 人年， 广州， 工作， 听说， 家乡， 发展， 越来越， 回到， 市， 工作...
[光明， 山河， 里， 水， 污染， 水质， 败坏， 自来水， 源， 群众， 对兰田， 自来...
[区， 金盘， 世界， 城， 购房， 一套， 市政府， 文件， 市中心， 城区， 购房， 享...
[市， 公交线路， 调整， 可否， 人性化， 一点， 路， 片区， 唯一， 一趟， 过河， ...
[热天， 步行， 开车， 绿荫， 躲， 行人， 清凉， 建国， 栽， 梧桐树， 眼看， 县，...
[彭， 感谢您， 百忙之中， 关注， 此事， 西地省， 天君， 物业管理， 有限公司， 对县
[县城， 托运， 城栋， 镇， 三中， 大门口， 正对面， 马路， 一家， 废品， 店乱， ...
[市， 住房， 城乡建设， 委员会， 市， 金沙， 中路， 福临， 世家， 栋， 单元， 业...
[临县， 安， 家园， 小区， 栋， 下雪， 冰冻， 水表， 冻破， 周末， 融雪， 化， ...
[天然气， 沧水， 铺， 整整， 五年， 燃气公司， 只通， 小区， 临街， 散户， 问， ...

图 1-6 部分去除停用词后分词效果

1.3 留言详情向量化

1.3.1 对留言详情进行分类

为了使文本数据转化为计算机能够处理的数值型数据。我们需将文本特征进行表达。目前文本特征表达模型有向量空间模型、布尔模型和概率模型等。

向量空间模型将每一个文本表示为向量空间的一个向量，并以每一个不同的特征项（词条)对应为向量空间中的一个维度，而每一个维的值就是对应的特征项在文本中的权重，这里的权重可以由 TF-IDF 等算法得到。向量空间模型就是将文本表示成为一个特征向量：

$$V(d) = (t_1, w_1(d), t_2, w_2(d), \cdots, t_n, w_n(d))$$

其中，$t_i(i = 1,2,\cdots,n)$为文档d中的特征项，$w_i(i = 1,2,\cdots,n)$为特征项的权值，可由 TF-IDF 算法得出。

1.3.2 TF-IDF 算法

- 第一步，计算词频

$$\text{词频(TF)} = \text{某个词在文章中出现的次数}$$

不同的文章字数不一样，为了便于比较，通常进行了一个“标准化”。

$$\text{词频(TF)} = \frac{\textit{某个词在文本中出现的次数}}{\textit{文本的总词数}}$$

或

$$\text{词频(TF)} = \frac{\textit{某个词在文本中出现的次数}}{\textit{文本中出现次数最多的词的出现次数}}$$

- 第二步，逆文档频率

需要一个语料库（corpus），用来模拟语言的使用环境。

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

如果一个词越常见，那么分母就越大，逆文档频率就越小越接近 0。分母之所以要加 1，是为了避免分母为 0（即所有文档都不包含该

词）。log 表示对得到的值取对数。

- 第三步，计算 TF-IDF

TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。计算出文档的每个词的 TF-IDF 值，然后按降序排列，取排在最前面的几个词。

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

实际分析得出 TF-IDF 值与一个词在留言详情中出现的次数成正比，某个词对一级标签分类重要性越高，TF-IDF 值越大。计算留言详情中每个词的 TF-IDF 值，进行分类排序，次数最多的即为要提取的一类一级标签关键词。

1.3.3 生成 TF-IDF 词频矩阵

利用 TF-IDF 算法得出相应的词频矩阵如图 1-1：

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

图 **错误!文档中没有指定样式的文字。**-7 TF-IDF 转换后的词频矩阵

生成留言详情的 TF-IDF 权值矩阵后，根据每一个一级标签(id)的 TF-IDF 权值矩阵，对留言进行分类。我们采用了多项式朴素贝叶斯算法和支持向量机算法对留言详情进行分类。

1.4 模型建立

1.4.1 多项式朴素贝叶斯（Multinomial Naïve Bayes）

（1）基本思想

- 计算文本属于类别的概率
- 文本属于类别的概率等于文本中的每个词属于类别的概率的综合表达式。

（2）公式

$$\mathrm{P}(x_i|y_k)=\frac{N_{y_k i}+\alpha}{N_y+\alpha n}$$

其中$N_{y_k i}=\sum_{x\in T}x_i$ 表示在训练集 T 中类y_k具有特征i的样本的数量，$N_y=\sum_{i=1}^{|T|}N_{yi}$ 表示训练集 T 中类y_k 的特征总数。平滑系数$\alpha>0$ 防止零概率的出现，当$\alpha=1$ 称为拉普拉斯平滑，而$\alpha<1$称为 Lidstone 平滑。

1.4.2 支持向量机算法——SVM

SVM 多类分类方法的实现根据其指导思想大致有两种：

（1）将多类问题分解为一系列 SVM 可直接求解的两类问题，基于这一系列 SVM 求解结果得出最终的判别结果。

- 针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率。
- 它通过非线性变换，将输入向量映射到一个高维空间 H。
- 在 H 中构造最优分类超平面，从而达到最好的泛化能力。

（2）通过对前面所述支持向量机分类器中的原始优化问题的适当改变，使得它能同时计算出所有多类分类决策函数，从而“一次性”地实现多类分类。

原始问题可改写为：

$$min: 1/2 \sum_{m=1}^{k} ||w^m||^2 + C \sum_{i=1}^{n} \sum_{m \neq y_i} \varepsilon_i^m$$

$$S.T(w_i \cdot x_i) + b_i \geq (w_i \cdot x_i) + b_m + 2 - \varepsilon_i^m \quad , \varepsilon_i^m \geq 0$$

式中：$i = 1$，2，～，n，n为样本数量；$m = 1$，2，～，k，k为类别数量。这样就可得到决策函数：

$$f(x) = \max_i [(w_i + x) + b_i]$$

判别结果为第i类。虽然第（2）种思想看起来简单，但由于最优化问题求解过程复杂，计算量太大，因此未被广泛应用。

1.5 结果分析

将已经整理好地数据进行随机划分，其中 90%作为训练集，其余作为测试集。通过训练后，多项式朴素贝叶斯模型得分为 88.04%，支持向量机模型的得分为 90.43%

然而分类模型对测试集进行预测而得出的准确率并不能很好地反映模型的性能，为了有效判断一个预测模型的性能表现，需要结合真实值，计算出精确率（precision）、召回率(recall)、F1 值和 Cohen’s Kappa 系数等指标来衡量。

$$\text{精确度(Precision)} = \mathrm{P} = \frac{TN}{TN + FN}$$

精确度越高，模型某类的分类效果越好。

$$\text{召回率(recall)} = \mathrm{R} = \frac{TN}{TN + FP}$$

召回率越高，表示模型分类误分概率越低，模型效果越好

$$\mathrm{F}=\frac{(\alpha^2+1)*P*R}{\alpha^2*(P+R)}$$

当参数$\alpha=1$时，就是最常见的 F1 值，即：

$$\mathrm{F1}=\frac{2*P*R}{P+R}$$

综合考虑精确度与召回率。

	precision	recall	f1-score	support
0	0.72	0.85	0.78	27
1	0.85	1.00	0.92	34
2	0.96	0.81	0.88	32
3	0.85	0.96	0.90	23
4	0.90	0.85	0.88	33
5	0.92	0.72	0.81	32
6	0.89	0.89	0.89	28
avg / total	0.87	0.87	0.87	209

图 **错误!文档中没有指定样式的文字。**-8 多项式朴素贝叶斯模型

	precision	recall	f1-score	support
0	0.80	0.89	0.84	27
1	0.94	0.97	0.96	34
2	0.96	0.78	0.86	32
3	0.85	1.00	0.92	23
4	0.97	0.88	0.92	33
5	0.85	0.91	0.88	32
6	0.96	0.93	0.95	28
avg / total	0.91	0.90	0.90	209

图 **错误!文档中没有指定样式的文字。**-9 支持向量机模型

通过综合比较，我们发现用支持向量机算法训练的分类模型整体效果比多项式朴素贝叶斯要好。因此我们选择支持向量机算法对留言详情进行分类。

二、热点问题挖掘

2.1 留言归并处理

对于留言的归并，首先对文本进行分词预处理，根据地点名词提取作为特征项，建立文本的特征项向量，这样就可以通过这些向量来表示文本，特征项分配权重来体现关键词语的重要程度，对整个留言的具有识别和描述能力。

提取地点名词，我们采取的方法是：以地点实体的指示词、标识出来的地名以及事件词的出现位置来提示地点实体可能出现的位置，再综合运用地点实体内部组成特点、上下文边界规律来判定其是否要抽取的对象，并确定它的边界。

具体操作流程如下：

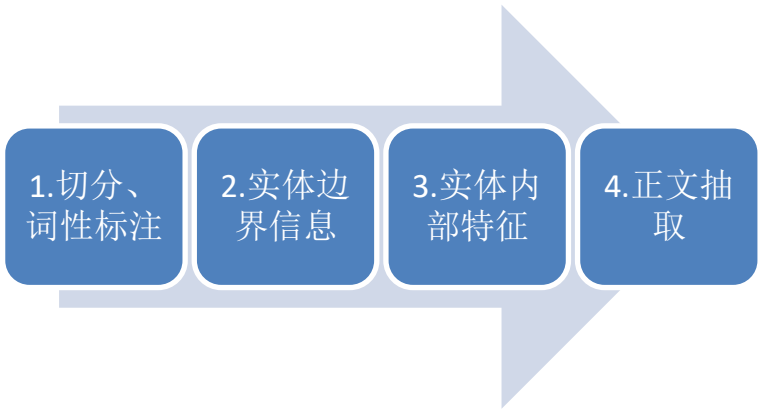

图 2-1 流程图

首先我们发现留言内容中的地点实体应该是满足以下三个条件的对象：

- 第一个条件：是实体名词或名词短语；
- 第二个条件：是对某个地理位置的最长表达；
- 第三个条件：是与事件关联的。

然后我们发现地点实体有以下两个界定的原则

- 第一个原则：并列结构的名词短语有多个中心词，根据各个最小的描述分成多个独立实体；
- 第二个原则：同位结构短语作为一个实体看待。

2.2 计算文本的相似度

（1）符号约定

1．特征项（Term）是指根据留言内容提取出的有代表意义的特征单元，我们记特征项为 T；

2．特征项权重（Weight）表示特征项表达能力的大小，通常权重越高，特征项权重越大，我们用 W 表示特征项权重；

3．相似度（Similarity）是两个文本之间相似程度的量化表示，记相似度为 S。

（2）模型表示

假设一共有n个留言，每个留言分别由 m 个特征项 T 组成，这些特征项互不相关。设$\mathrm{Di} = (\mathrm{T1}, \mathrm{T2}, \ldots, \mathrm{Dn})$对应文本的特征项Ti的权重，以 Ti为 m 维空间的坐标轴， Wi为相对应的坐标值， Di可以表示为$\mathrm{Di} = (\mathrm{W1}, \mathrm{W2}, \ldots, \mathrm{Wm})$。通过以上的分析，可以对文本集合表示成如下的空间向量模型：

$$\mathrm{D} = \begin{pmatrix} D1 \\ . \\ . \\ . \\ Dn \end{pmatrix} \begin{pmatrix} W11 & W12 & \ldots & W1m \\ W21 & W22 & \ldots & W2m \\ & \ldots & & \\ & \ldots & & \\ Wn1 & .Wn2 & \ldots. & Wnm \end{pmatrix}$$

（3）TF-IDF 权重计算法

在 TF-IDF 权重计算法中，最后阶段的权重系数往往是根据文本特征项的 TF 与 IDF 两个指标来确立的，最为常用的方法是采用两者的乘积的方式描述特征项的权重大小。TF-IDF 权重计算法是基于频率统计的原则，计算方式简单，同时具有高效的线性复杂度的特点，适合于系统文本库中文本数量较大的情况，是目前权重计算中最具有代表性的方法。

（4）VSM 算法

在经过特征项的选取、权重的计算过程之后，可以建立文本特征项向量，文本信息表达能力将以向量的形式表现出来，在此基础上我们用两个空间向量夹角的余弦进行相似度计算。

（5）相似度计算公式

假设在留言 D1 和 D2 中空间向量的表示分别为$(\mathrm{T11},\mathrm{T12},\ldots,\mathrm{T1m})$和$(T21,T22,\ldots,T2m)$,计算公式如下：

$$\mathrm{Sim}(D_1,D_2)=\cos\theta=\frac{\sum_{i=1}^{m}T_{1i}*T_{2i}}{\sqrt{\sum_{i=1}^{m}T_{1i}^2}+\sqrt{\sum_{i=1}^{m}T_{2i}^2}}$$

2.3 构建热度评价指标

根据非常规突发事件的发生，损害到了大多数民众的利益，导致民众情绪的高涨，引起大家强烈的控诉。

我们从附件 3 中利用数据透视表统计民众发言最多的时间，可以从中发现热点问题发生的大致时间段。如下图：

2020年	
1月	80
2019年	
1月	359
2月	256
3月	368
4月	397
5月	387
6月	338
7月	417
8月	391
9月	387
10月	296
11月	280
12月	366
2018年	
5月	1
10月	1
11月	1
2017年	
6月	1
总计	4326

图 2-2 热点问题发生的大致时间段

接着我们利用指标来评价：

$$\text{正指标：}\ y_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_{ij}) - \min(x_{ij})}$$

$$\text{逆指标：}\ y_{ij} = \frac{\max(x_{ij}) - x_{ij}}{\max(x_{ij}) - \min(x_{ij})}$$

式中 $\max(x_{ij})$、$\min$ （x_{ij}）分别为指标评价值的最小值和最大值

我们把附件中的数据代到相应的公式，通常存在三个问题：第一 ，指标体系具有很强的随意性和主观性 ，缺乏合理的论证过程；第二 ，指标体系缺乏普适性；所以为了验证指标体系是否适科学有效，本文对指标体系的合理性进行测试。

2.4 评价结果

2.4.1 排名前五的热点问题

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	五颗星	2019/11/15 至 2020/1/15	A2 区丽新城附近建搅拌站噪音扰民	A2 区丽发新城附近修建搅拌站，污染环境，影响生活
2	2	四颗星	2019/8/1 至 2019/8/21	A 市伊景园滨河苑	A 市伊景园滨河苑捆绑车位销售
3	3	三颗星	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
4	4	二课星	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
5	5	一颗星	2019/7/12 至 2020/2/21	A 市金星北片月亮岛	关于 A 市金星北片 110kv 及以上高压线的意见

表 2-1 排名前五的热点问题

2.4.2 相应热点问题对应的留言信息

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	188809	A909139	梅路丽发新城居民区附近	2019/11/19 18:07:54	区旁50米处建搅拌站，运渣车吵得	0	1
1	189950	A909204	区丽发新城附近建搅拌站吗	2019-11-13 11:20:21	米的地方建搅拌站。可想而知，一	0	0
1	190108	A909240	丽发新城小区旁边建搅拌站	2019-12-21 15:11:29	严重影响几千名学生的健康，很	0	1
1	190523	A00072847	违建搅拌站，彻夜施工扰	2019/12/26 13:55:15	染严重；3、搅拌站几百米外就	0	0
1	199379	A00092242	新城附近修建搅拌厂，严重	2019/11/25 10:17:56	居民因此还得了疾病住院，该地区	0	0
1	203393	A00053065	面建设混泥土搅拌站，粉	2019/11/19 14:51:53	了巨大的粉尘，严重影响居民健	0	2
1	208714	A00042015	附近修建搅拌站，污染环	2020-01-02 00:00:00	气质量和声环境质量急剧下降，	0	4
1	213464	A909233	新城小区附近违建搅拌站	2019-12-10 12:34:21	大型搅拌站。该搅拌站的设备太	0	0
1	213930	A909218	附近违规乱建混凝土搅拌	2019-12-27 23:34:32	居民强烈呼吁政府和有关职能部	0	0
1	214282	A909209	小区附近搅拌站噪音扰民	2020-01-25 09:07:21	吵天天吵，烦死了不仅吵还臭！	0	0
1	215563	A909231	新城小区旁边的搅拌厂是	2019-12-06 12:21:32	还产生了噪音和灰尘。这给小区居	0	0
1	215842	A909210	区丽发新城小区附近太吵	2020-01-26 19:47:11	一个搅拌厂是怎么回事！下班回	0	0
1	216824	A909214	砂石料噪音污水影响丽发	2019-12-25 12:15:57	解，这些严重扰民的噪音是从位	0	0
1	217700	A909239	城小区旁的搅拌站严重影	2019-12-21 02:33:21	范围内搬迁到丽发新城小区旁边	0	1
1	222831	A909228	染的A2区丽发新城附近环	2019-12-22 10:23:11	么城里能修改产生大量灰尘的搅	0	0
1	225217	A909223	新城附近修建搅拌厂严重	2019-11-15 09:17:36	附近建一搅拌站,每天尘土飞扬,噪	0	0
1	231136	A909204	区丽发新城附近建搅拌站吗	2019-12-02 11:20:21	拌站。距离上次投诉已经过去一个	0	0
1	233158	A909242	新城小区旁建搅拌厂严重	2019-12-05 08:46:20	上班不在家我还能忍忍，但周末	0	0
1	234327	A909212	噪声不断的丽发新城小区	2019-12-26 21:44:13	能休息，还产生大量粉尘和污染	0	0
1	235362	A909215	新城小区附近水泥搅拌站	2020-01-06 20:45:34	尘肆虐，严重危害居民身体健康！	0	0
1	238212	A909203	新城小区附近建搅拌站合	2019-12-12 10:23:11	小区作为居民区应是一个安静的好	0	0
1	239336	A909213	丽发新城小区遭搅拌站严	2019-12-11 11:44:11	厂成日运作，离居民区非常近！	0	0
1	239648	A909211	新城小区附近搅拌站明日	2020-01-06 22:41:31	有灰尘颗粒！都不敢开窗透气了，	0	0
1	243692	A909201	城小区附近的搅拌站噪音	2019-11-15 11:23:21	搅拌站的灰尘极大，都飘到小区里	0	2
1	244335	A909135	发新城社区搅拌站灰尘，	2019/12/2 12:11:23	把特大型搅拌站，水泥厂从绿心湖	0	0
1	244512	A00094706	发新城小区粉尘大的孩子	2019-12-05 20:57:50	尘很大,大人都无法正常呼吸,更	0	1
1	253040	A909202	区丽发新城附近建搅拌站吗	2019-12-04 12:10:21	孩在家里根本无法正常休息！搅拌	0	0

图 2-3 排名前五的热点问题明细示意图

三、答复意见的评价

评价体系指标主要是从答复的相关性、完整性、可解释性等角度入手展开设定。相关性是指答复意见的内容是否与群众留言中所问的问题相关，完整性是指是否满足某种规范，可解释性是指答复意见中内容的相关解释。采用指数评价体系设立一套评价方案。

3.1 评价体系指标量化处理

3.1.1 答复意见的相关性

（1）一方面我们将相关性描述作量化处理。

- 第一步：数据预处理

  我们对群众留言反映的问题与相关部门的答复意见这二者具备相关性的变量元素的文本数据进行预处理，主要是分词、去停用词、将词组向量化等常规操作。

- 第二步：神经网络

  对变量元素进行分析，将预处理得到的词向量交给 bi-LSTM 网络，用于挖掘语句关系，同时为了降噪，将群众留言的句意与答复意见的句意各自嵌入到一个向量，这里使用 Sentence2vec 方法。

- 第三步：相关性分析

  对两者作关联分析、碰撞比对，分析相关性的元素之间存在着的一定的联系或者概率情况，从而衡量两个变量因素的相关密切程度，模型可对每一个问答度给出一个评分，评估他们是否是相互对应的问题与回答。

- 第四步：阈值法

  我们多次实验后设置一个阈值，如果得到的结果大于阈值，相关性强，且超出越多相关性越强，相应的，得到的结果小于阈值时，相关性弱，且落后越多相关性越弱。

（2）另一方面我们认为相关性是占整个评价体系中分值较为高的一个方面。如果相关部门的答复意见跟所问的问题关联不大，说明答复意见的质量明显不高，这样的答复定是相关部门敷衍了事的结果且不为群众所满意。

3.1.2. 答复意见的完整性

- 匹配分析

  考虑相关部门答复的完整性，就是看这篇答复是否满足某种规范、某种模板，当我们自行设立好一个有关答复意见的规范或模板后，就把问题看作是一个匹配问题，将相关部门的答复意见文本与设立好的模板文本进行比对、匹配，同样时分析二者的相关密切程度，当然让这一切结果可靠的前提是设立的有关答复意见的模板是合理的有说服力的。

- 规范定义

  答复规范我们定义如下，大致包括三个步骤：一是开头用语是否规范，一般是：“XXX，您好！您的留言已收悉，经 XXX 研究，现答复如下。”写明答复意见的回应对象同时措辞简明得体；二是相关部门表明了对群众留言反映的问题进行了审查，是否作了认真调查、检索，确认群众留言内容反映的情

况如实的工作，简明说下审查结果；三是书写的答复文书是否有对群众留言问题解答的疑惑方面内容或是解决办法方面的内容。

3.1.3 答复意见的可解释性

可解释性是指答复意见中内容的相关解释。要将可解释性描述量化，可先分词处理答复意见中对内容的相关解释部分，去停用词后，计算相似度。

3.2 评价体系构建

3.2.1 流程图

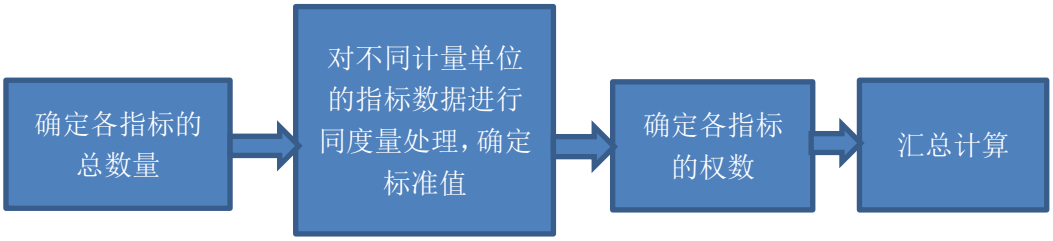

图 3-1 评价体系构建流程图

3.2.2 指数评价体系

评价的指标主要有三个：相关性、完整性、可解释性。运用这三个指标，对相关部门的答复意见进行评价，并根据各个指标的不同权重，进行综合评价。相关性是占整个评价体系中分值较为高的一个方面，完整性和可解释性占比稍低，我们根据不同指标的重要性进行加权处理。指标的完成值除以指标的标准值，乘以各自权数，加总后除以总权数即得到结果。

四、结论

对智慧政务中信息进行分析研究，了解社会和相关民众的需求特点与趋势，对广大的政务人员有重大意义，同时也是文本分析的一个课题、一个难题。传统的人工服务已经不能满足数据量庞大的数据信息。

本文采用朴素贝叶斯算法和支持向量机算法去构建了关于一级标签分类模型。分析了群众的热点问题，这有利于相关部门有针对性的处理问题。同时针对相关部门的答复意见做出了相应的评价，给出了一套完整的评价方案。

五、参考文献

[1]吴致晖，刘洪伟,陈丽.高效朴素贝叶斯 Web 新闻文本分类模型的简易实现.

[2]孙润志.基于语义理解的文本相似度计算研究与实现

[3]高燕.事件报道中地点实体的提取研究.2011 年.第 S1 期

[4]唐凌志.基于语义理解的论文相似度研究[D].

[5]金希茜.基于语义相似度的中文文本相似度算法研究[D]. 浙江工业大学,2009.