# 群众问政留言信息的数据挖掘与综合分析

## 摘要

群众利益无小事，民生问题大过天。自古以来，处理解决好民生问题一直是政府的头等大事。习总书记曾说：“老百姓上了网，民意也就上了网。”近些年来网络问政平台逐渐成为了政府与民众沟通交流的重要方式，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。本文工作主要包括以下几个方面：

（1）数据的预处理。对所给数据中存在的数据缺失情况、数据异常情况和无意义数据进行分词和去停用词的处理，将处理后的数据作为研究数据使用。

（2）针对问题一，首先将原始数据提取并分析出一级标签下的内容，为了将非结构化的文本信息转化为计算机能够识别的结构化信息，利用文本分词和去停用词对文本进行格式转换，并用 FastText 模型对转换结果进行无监督训练，自动识别留言内容的一级标签分类。

（3）针对问题二，对留言主题进行数据挖掘，建立合理的一级、二级热点评论指标，结合 jieba 分词工具对主题进行分词和词频统计，最后用 LDA 算法归类出排名靠前的特征词，通过 Excel 特征词检索得出热度排名前五的热点问题，并建立留言主题热度综合评价指标体系计算热点问题的热度指数。

（4）针对问题三，随机抽取五条留言做评价方法样本，针对不同的答复意见，从相关性、完整性、可解释性以及时效性的角度作为答复意见的评价指标，构造加权标准化矩阵运用对数最小二乘法对指标赋予权重，再使用 TOPSIS 综合评价法评价抽取的五行样本内容。

关键词：FastText 模型、 jieba 分词工具、 LDA 算法 、TOPSIS 综合评价法

# 目录

# 第一章 绪论

## 1.1 研究背景

随着信息技术的飞速发展，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，而且各类社情民意相关的文本数据量不断攀升，大数据、 云计算、人工智能等技术也在其中得到了广泛的运用。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。对于这些数据分析从而得到更有价值的信息，已经成为政府部门着重研究的方向。

据中国统计年鉴数据显示，到 2019 年年底，连接网络用户的人员为 44928 万人，相对于 2018 年均稳步上升。网络问政越来越便利，群众问政留言记录的文本数量也在不停地攀升，这导致了相关部门利用人工对留言划分和热点整理的工作变得极其困难，使得政府不能快速的了解到人民的诉求，从而人民的问题不能得到快速的解决，这些都严重阻碍了人们生活水平的提高以及社会经济的发展。所以这迫使人们采用高新技术来解决处理留言记录的问题，而智慧政务系统便应运而生。

## 1.2 研究内容

（1）利用附件二中的数据，提取并分析出各个一级标签下的内容，利用文本分词和去停用词对文本进行格式转换，运用 FastText 模型对转换结果进行无监督训练，得以实现自动识别留言内容的一级标签分类。

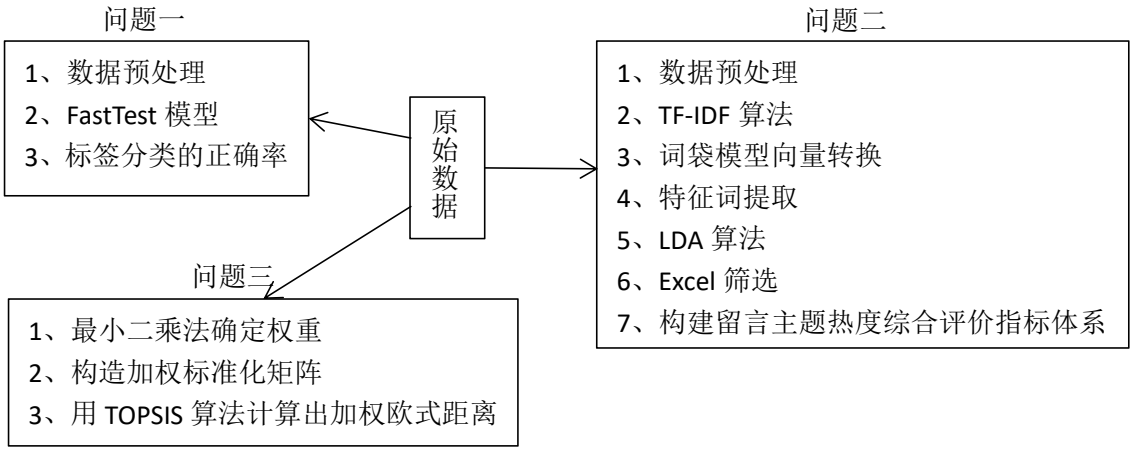
（2）利用附件三中的数据，根据对留言主题的挖掘，建立合理的一级、二级及三级热点评论指标，结合 jieba 分词工具对主题进行分词，对分词结果进行词频统计，最后用 LDA 主题聚类的方法归类出排名靠前的特征词，通过 Excel 特征词检索得出热度排名前五的热点问题。

（3）利用附件四中的数据，随机抽取五行内容做评价方法样本，针对不同的答复意见，从相关性、完整性、可解释性及时效性等角度作为答复意见的评价指标，构造加权标准化矩阵运用对数最小二乘法对指标赋予权重，再使用 TOPSIS

综合评价法评价抽取的五行样本内容。

## 1.3 研究路线

针对要解决的群众留言分类、热点问题挖掘及答复意见的评价问题，本文的研究路线如下所示：




**图一 研究路线图**

# 第二章 研究方法与过程

## 2.1 问题一研究方法与过程

### 2.1.1 流程图

针对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，我们在群众留言分类问题的解决过程如下：

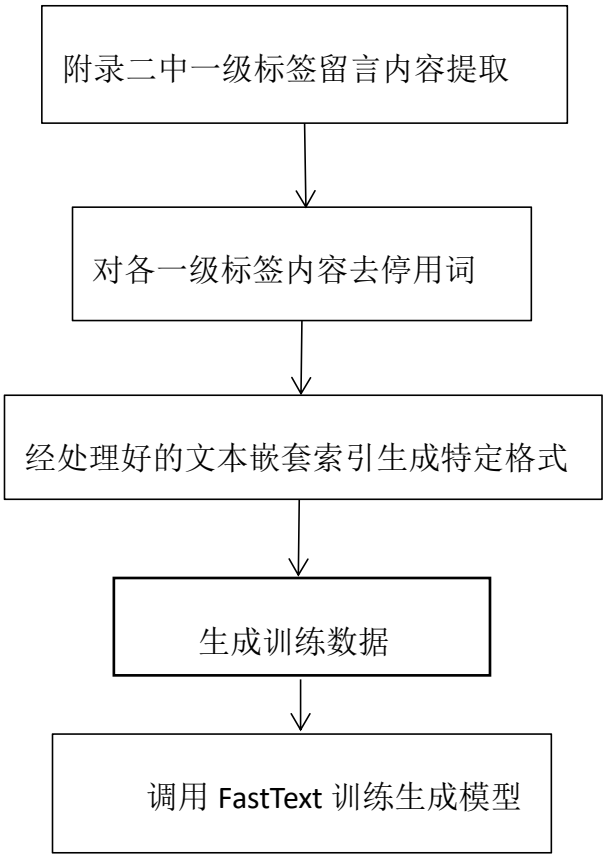



**图 2 群众留言问题流程图**

### 2.1.2 数据预处理

**群众留言的去重**

在题目给的数据中，可能会出现很多重复的数据。例如在一定的时间段，居民的需求相似，就会出现很多重复的信息。考虑到相关的政府部门需要处理上百万条的数据，因此去掉相同的留言。

**提取一级标签、去空值**

利用 Python 将附件二中的一级标签提取，考虑到 Python 中的字典在保存数据的时候，key 为相同的内容，value 取值为最后更新的值。因此在读取数据的时候，把一级标签作为 value 保存在 value 中。最后将一级标签为空的记录，干扰了问题的分析，采取直接过滤的方法，从文本中删除。提取出来的一级标签分别为“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”。

**群众留言去停用词，套索引、生成训练集**

在进行挖掘分析之前，先要把非结构化的文本信息转化为计算机可以识别的结构化的信息。在附件二的 Excel 表中，以中文文本的方式给出了数据。为了便于转换，现对于这些群众留言信息进行去停用词。自己建立了停用词表，数据保存在 stopwords.txt，利用 Python 删除了停用词。为了方便政府部门的的查找，用 Python 建立了 label 索引，也对于以后的研究提供了便利。生成七个训练集，来拟合数据模型，以便于索引的运用。

对于数据预处理的 Python 程序见附件 Untitled3.html，所得到的数据存入了 mytrain_data.txt。

### 2.1.3 FastText 模型

将训练好的词向量，进行文本分类，以便挖掘使用。这里采用 FastText 模型，进行文本分类。FastText 模型的具体原理如下：

第一：模型架构

FastText 模型架构如下图所示。FastText 模型输入一段文本，输出这个词序列属于不同类别的概率。序列中的词和词组组成的特征向量，特征向量通过线性变换映射到中间层，中间层在映射到标签。FastText 在预测标签时使用了非线性激活函数，但在中间层不使用非线性激活函数。

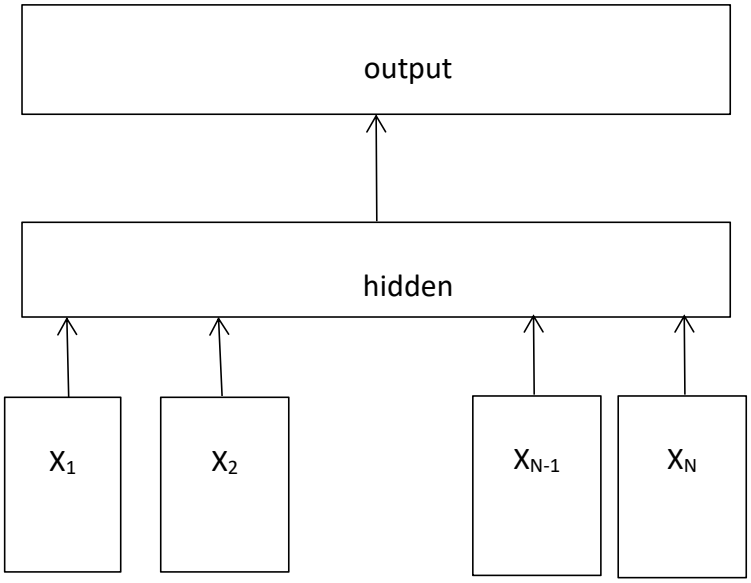



**图 3　FastText 模型架构**

第二：N-gram 特征

FastText 可以用于文本分类和句子分类。我们经常使用的特征是词袋模型，但是词袋模型不可以考虑词之间的顺序，所以加入了 N-gram 特征，而且为了提高效率，则需要过滤掉低频的 N-gram。Word2vec 把语料库中的每个单词当成原子，每个单词生成向量，忽略单词的内部的形态特征。为了克服这样的问题，使用了字符级别的 N-grams 来表示单词，用向量叠加表示词向量。

第三：层次 Softmax

为了改善改善运行时间，FastTest 模型使用了层次 Softmax 技巧。它建立在哈弗曼编码的基础上，对标签进行编码，能够极大地缩小模型预测目标的数量。分层 Softmax 的基本思想是使用树的层级结构代替扁平化的标准 Softmax，使得在计算 $P(y=j)$时，只需计算一条路径上的所有节点的概率值，无须在意其它结点。

下图是一个分层 Softmax 示例：

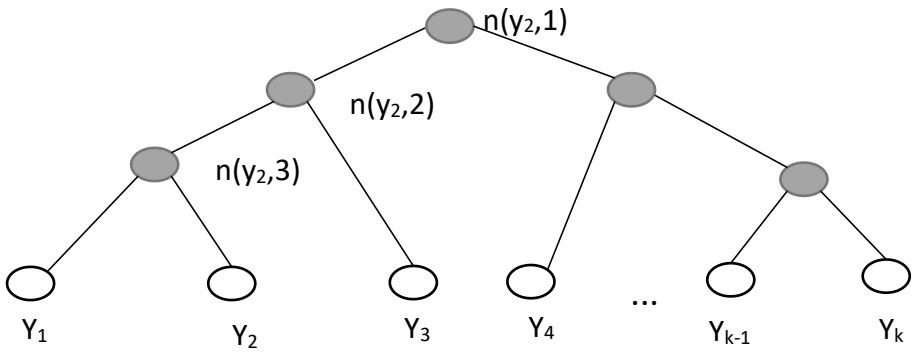



**图 4　分层 Softmax**

树的结构是根据类标的频数构造的哈弗曼树。$K$ 个不同的类标组成的所有叶子结点，$K$-$1$ 个内部节点作为内部参数，从根节点到某个叶子结点经过的节点和边形成一条路径，路径长度表示为 $L(y_i)$。于是 $P(y_i)$就可以被写成：

$$P(y_i) = \prod_{i-1}^{L(y_i)-1} \sigma([n(y_i,l+1) = LC(n(y_{i,}l)) \cdot \theta_{n_{(y_i,l)}}{}^{T} X])$$

其中，$\theta(\cdot)$表示 sigmoid 函数；$LC(n)$表示 n 节点的左孩子；$[x]$是一个特殊函数，被定义为：

$$[x]=\begin{cases} 1 & \text{If x==true} \\ -1 & \text{otherwise} \end{cases}$$

$\theta_{n_{(y_i,l)}}$是中间节点，$n(y_i,l)$的参数；$X$ 是 Softmax 层的输入。上图中高亮的节点和边是根节点到$y_2$的路径，路径长度 $L(y_2) = 4$，$P(y_2)$可以被表示为：

$$\begin{aligned} P(y_2) &= P(n(y_2,1),left \cdot P(n(y_2,2),left) \cdot P(n(y_2,3),right) \\ &= \sigma(\theta_{n_{(y_2,1)}}{}^{T} X) \cdot \sigma(\theta_{n_{(y_2,2)}}{}^{T} X) \cdot \sigma(-\theta_{n_{(y_2,3)}}{}^{T} X) \end{aligned}$$

于是从根节点走道叶子结点$y_2$，实际上是在做了三次二分类的逻辑回归。

第四：CBOW

输入层：个节点，上下文共 N 个词的词向量的平均值。输入层到输出层的连接边：输出词矩阵。输出层：个节点。第 n 个节点代表中心词的概率。

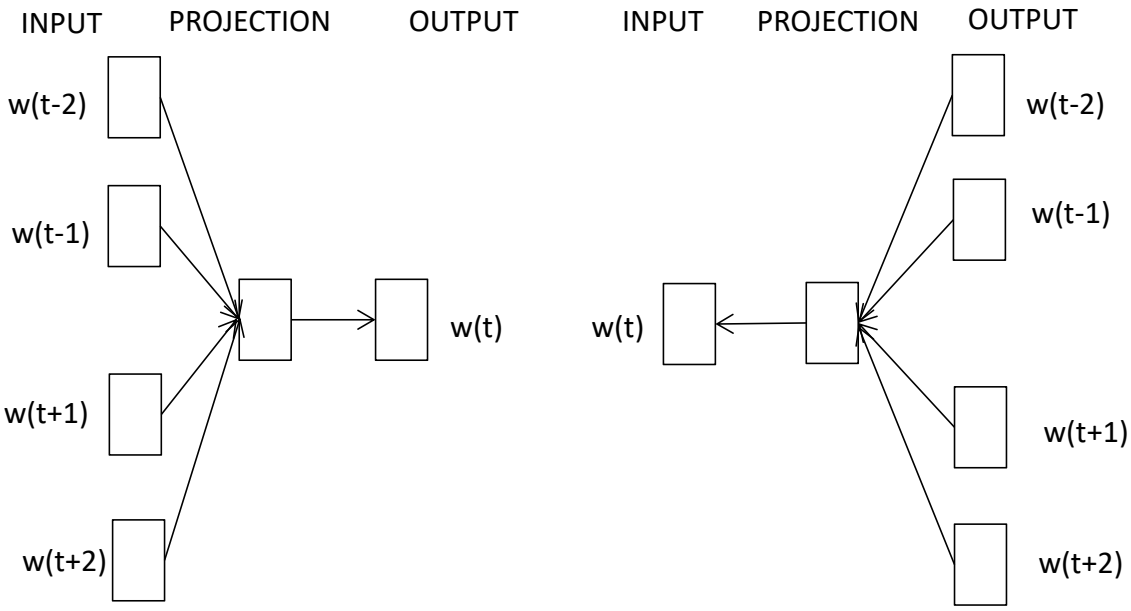



**图 5 CBOW 流程图**

### 2.1.4 标签分类的正确概率

针对附件给出的数据，通过上述步骤，对于留言内容的一级标签分类，利用 FastText 模型，使用如下公式：

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

计算出查准率和查全率分比为 85.9%、85.9%，取得了较好的结果。

## 2.2 问题二的研究方法与过程

### 2.2.1 数据预处理

**留言信息分词**

在对用户留言进行挖掘分析之前，需要先将非结构化的文本信息转化为计算机能够识别的结构化信息。在市民留言信息表中，以中文文本的方式给出了数据。为了方便进行转换，先要将这些用户留言进行中文分析，这里采用的是 python 中的中文分词包 jieba 进行分词。Jieba 采用基于前缀词典实现的高效词图扫描，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）。

在分词的同时，还采用了 TF-IDF 算法，从每条留言文本中抽取前五个关键词，这里采用 jieba 自带的语义库。

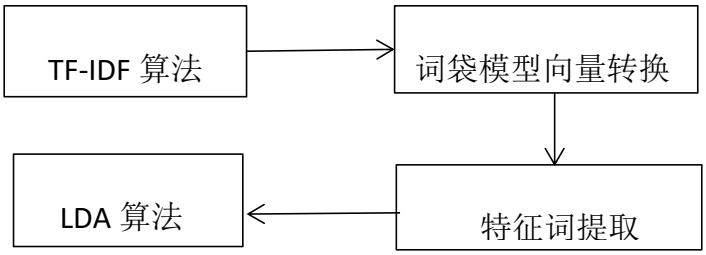
**去停用词**

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，这些字或词即被称为 Stop Words(停用词)。停用词主要包括英文字符、数字、数学字符、标点符号及使用频率特高的单汉字等。

在文本处理过程中如果遇到它们，则立即停止处理，将其扔掉。将这些词扔掉减少了索引量，增加了检索效率，并且通常都会提高检索的效果。此处采用 python 使用 jieba 进行去停用词。

### 2.2.2 数据分析

针对网络问政平台的群众留言热点问题，根据附件 3 给出的数据信息，进行数据分析，如下图：




**图六 数据分析流程图**

**TF-IDF 算法**

在对市民留言信息数据进行分析后，将这些词语转化为向量，以供接下来挖掘分析使用，这里采用 TF-IDF 算法，把市民留言信息转化为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）

词频（TF）=某个词在文本中出现的次数

考虑到市民留言有长短之分，为了便于不同市民的留言进行比较，将“词频”

进行标准化处理，除以留言信息的总词数或以该留言信息中出现最多的词语的出现词数，即：

$$\text{词频（TF）}=\frac{\text{某个词在留言中出现的次数}}{\text{留言的总字数}}$$

或

$$\text{词频（TF）}=\frac{\text{某个词在留言中的出现次数}}{\text{该留言出现次数最多的词的出现次数}}$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大时，此特征在留言中的分布越集中，则说明该分词在区分该留言内容属性能力越强。

$$\text{逆文档频率（IDF）}=\log\left(\frac{\text{语料库的留言总数}}{\text{包含该词的留言数}+1}\right)$$

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）

TF-IDF = 词频（TF）× 逆文档频率（IDF）

通过计算分析，得出 TF-IDF 值与一个词在市民留言信息表中出现的次数成正比，某个词在留言中的重要性越高，TF-IDF 值越大。因此，通过计算留言中每个词的 TF-IDF 值并进行排序，TF-IDF 值最大的即为要提取的市民留言信息表中留言的关键词。

**生成 TF-IDF 向量**

生成 TF-IDF 向量的步骤如下：

（1）使用 TF-IDF 算法找出每条留言的前五个关键词；

（2）对每条留言提取的五个关键词合并成一个集合，计算每条留言对于这个集合中词的词频，如无则计做 0；

（3）生成每条留言的 TF-IDF 权重向量，计算公式如下：

TF-IDF = 词频（TF）× 逆文档频率（IDF）

LDA 算法

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。可以用来识别大规模文档集(document collection)或语料库(corpus)中潜藏的主题信息。采用词袋(bag of words)的方法，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易于建模的数字信息。

对于语料库中的每条留言文本，LDA 定义了如下生成过程:

（1）对每一篇文档，从主题分布中抽取一个主题;

（2）从上述被抽到的主题所对应的单词分布中抽取一个单词;

（3）重复上述过程直至遍历文档中的每一个单词。

语料库中的每一篇文档与 T(通过反复试验等方法事先给定)个主题的一个多项分布 (multinomialdistribution)相对应，将该多项分布记为 $\theta$ 。每个主题又与词汇表(vocabulary)中的 V 个单词的一个多项分布相对应，将这个多项分布记为 $\phi$ 。

根据生成的语料库留言文本，用 python 进行分析得到用户留言信息表中潜藏的主题信息。

### 2.2.3 数据筛选

根据附件 3 所给数据，进行以下步骤的筛选：

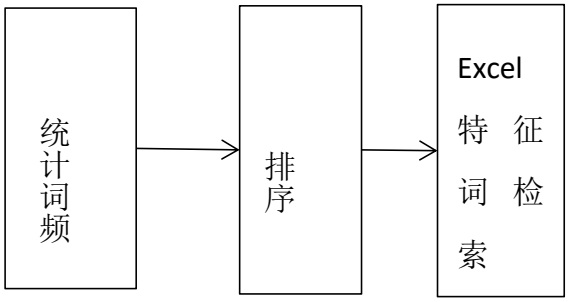



**图 7 数据筛选流程图**

此处采用 python 顺序统计各个词语出现的次数，并将最终统计数目进行排序，按词频高到低进行排序。通过 Excel 自带的筛选功能进行特征词的筛选，最

终筛选出五个热点问题进行下一步处理。

### 2.2.4 留言主题热度评价指标的建立

**构建原则**

指标体系的构建原则通常根据要求和对象的不同分为三个层面:指标选取层面一般采取客观性原则、系统性原则和敏感性原则。客观性原则是指指标体系的选择必须从客观实际出发，全面准确地反映留言的热点问题情况，克服因人而异的主观因素的影响。系统性原则是指指标体系的设计应从系统整体出发,能够包络形成留言热点问题的各个因子,各项指标间既相互独立又相互联系,共同构成一个有机整体。计算与操作层面一般采用数据的可得性和可操作原则，是指在设计指标体系时用较少指标反映较多的实质性内容，而且指标便于收集和量化针对留言热点问题这一特殊的研究对象。

在构建留言主题热度评价指标体系的过程中，除遵循以上基本性原则外,新增了趋势性原则和导向型原则。留言的热度是一个时刻变化的指标，趋势性原则就是体现留言热点问题的变化趋势；导向型原则是指该套指标体系的构建不仅要对留言进行检测,更是要为判断热点问题留言提供方向指导。

**构建评价体系**

构建留言主题热度综合评价指标体系表 1 如下：

| | 一级指标 | 二级指标 |
|---|---|---|
| 留言主题热度综合评价指标体系 | 地区特征热度影响力 | 热门城市 |
| | | 开发地区 |
| | 内容特征热度影响力 | 字数饱和度 |
| | | 出现及时性 |
| | 群众特征热度影响力 | 点赞数 |
| | | 反对数 |

**表 1 留言主题热度评价表**

## 2.3 问题三研究方法与过程

### 2.3.1 流程图

针对附件 4 相关部门对留言的答复意见，考虑到答复的相关性、完整性、可解释性等角度，对答复意见的质量给出了一套评价方案，过程如下：

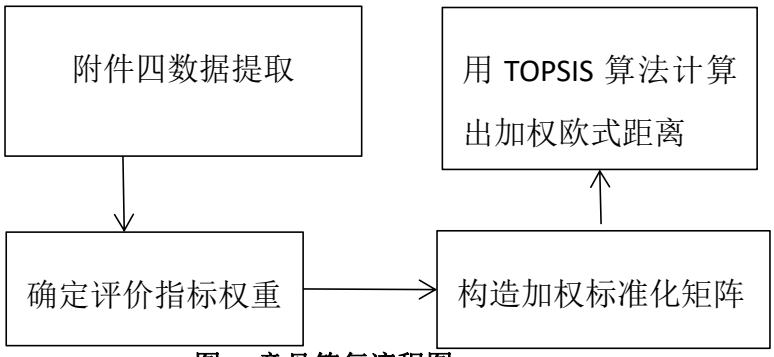



**图 8 意见答复流程图**

### 2.3.2 针对答复角度确定权重

随机筛选出五条留言信息做评价方法样本，针对不同的答复意见，从相关性、完整性、可解释性以及时效性四个角度作为答复意见的评价指标。

根据最小二乘法公式得出对应权重：

$$a=\frac{\sum xy-\frac{1}{N}\sum x\sum y}{\sum x^2-\frac{1}{N}\left(\sum x\right)^2}$$

$$b=\overline{y}-a\overline{x}$$

### 2.3.2 TOPSIS 综合评价法

TOPSIS 综合评价法即“逼近于理想值的排序方法”是根据有限个评价对象与理想化目标的接近程度进行排序，适用于多项目标、对多个方案进行比较选择的分析方法。

这种方法的中心思想在于首先确定各项指标的正理想值和负理想值，所谓正理想值是一设想的最好的值（方案），它的的各个属性值都达到各候选方案中最

好的值，而负理想解是另一设想的最坏的值（方案），然后求出各个方案与正理想值和负理想值之间的加权欧氏距离，由此得出各方案与最优方案的接近程度，作为评价方案的优劣标准。

**计算步骤**

（1）构造初始矩阵 $A$（$n$ 个评价指标，$m$ 个目标）

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

式中 $a_{ij}$ 表示第 $i$ 个目标的第 $j$ 项指标值（$1\le i\le m,1\le j\le n$）。

（2）由于各个指标的量纲可能不同，需要对原始数据进行标准化处理（归一化处理）。

$$A'=\begin{bmatrix} a_{11}' & a_{12}' & \cdots & a_{1n}' \\ a_{21}' & a_{22}' & \cdots & a_{2n}' \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}' & a_{m2}' & \cdots & a_{mn}' \end{bmatrix}$$

式中：

$$a_{ij}'=\frac{a_{ij}}{\sqrt{\sum_{i=1}^{m}a_{ij}^{2}}}$$

（3）构造加权标准化矩阵 $Z$

$$Z=A'W=\begin{bmatrix} a_{11}' & a_{12}' & \cdots & a_{1n}' \\ a_{21}' & a_{22}' & \cdots & a_{2n}' \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}' & a_{m2}' & \cdots & a_{mn}' \end{bmatrix}\begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix}$$

式中，$w_j$ 是第 $j$ 个指标的权重。

（4）根据加权矩阵判断正负理想解 $Z^{+}$，$Z^{-}$

$$z_j^+ = \begin{cases} max(z_{ij}), j \in J^* \\ min(z_{ij}), j \in J' \end{cases}, z_j^- = \begin{cases} min(z_{ij}), j \in J^* \\ max(z_{ij}), j \in J' \end{cases}$$

式中，$J^*$是效益性指标集，$J'$是成本型指标集。

（5）计算各个方案的到正理想点的距离$S_i^+$和负理想点的距离$S_i^-$

$$S_i^+ = \sqrt{\sum_{j=1}^{n}(z_{ij} - z_j^+)^2}, i = 1,2,\cdots,m$$

$$S_i^- = \sqrt{\sum_{j=1}^{n}(z_{ij} - z_j^-)^2}, i = 1,2,\cdots,m$$

（6）计算加权欧氏距离（即综合评价指数）

$$C_i = \frac{S_i^-}{S_i^+ + S_i^-}, i = 1,2,\cdots,m$$

# 第三章 结果分析

## 3.1 问题一结果分析

### 3.1.1 模型分类结果

当输入群众留言时，FastTest 会自动计算出他的主题是属于七个标签的哪一个以及他的概率。以下就举几个例子：

1、当输入“投诉 举报 M2 县 梓门桥镇 强制 居民 集中地 兴建 重度污染 氢化 洗金场”，概率排名前三的标签和概率如下表 2 所示：

| | |
|---|---|
| **环境保护** | **72, 55%** |
| **商贸旅游** | 12. 56% |
| **卫生计生** | 11. 89% |

**表 2 标签概率表（1）**

2、当输入“A1 东站 社区 通苑小区 业主 政府 提质 改造 项目 燃气项目 规定 缴纳 费用 管道” 概率排名前三的标签和概率如下表 3 所示：

| | |
|---|---|
| **城乡建设** | **98. 47%** |
| **商贸旅游** | 1. 37% |
| **环境保护** | 0. 10% |

**表 3 标签概率表（2）**

3、当输入“在 网上 得知 狂犬病 死亡率 近 100% 是 真的 在 预防 狂犬病 方面 那些 措施呢” 概率排名前三的标签和概率如下表 4 所示：

| | |
|---|---|
| **卫生计生** | **33. 78%** |
| **商贸旅游** | 19. 46% |
| **交通运输** | 15. 89% |

**表 4 标签概率表（3）**

### 3.1.2 F-score 评价分类求解

1、查准率：虽然准确率可以判断总的正确率，但是在样本不均衡的情况下，并不能作为很好的指标来衡量结果。

2、查全率：召回率是针对原样本而言的其含义是在实际为正的的样本中被预测为正样本的概率。准确率和召回率相互影响，理想状态下肯定是追求两个都高，但是实际情况下两者是相互制约的：追求准确率高，则召回率就低；追求召回率高，则通常会影响准确率。

3、F-score：一般来说准确率和召回率成负相关，一个高一个就低，如果两个都低的话，那么就一定有问题。所以$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$，从公式可以看出这个问题是查准率和查全率一样重要，权重的分配是一样的，所以最后计算出来的结果是 85.9%。$F_1$ 的值为 85.9%，属于较高，所以试验方法也比较理想。

我们使用了 FastText 模型进行文本分类，因为低频词的 N-gram 可以和其他词共享，所以低频词生成的词向量效果更好。而且我们可以叠加训练词库之外的单词的字符集 N-gram 向量，使得任然可以构建他们的词向量。而且这个模型更加专注于文本分析，在标签预测的问题上实现当下最好的表现。而且他还考虑了相似性，它的词嵌入学习能够考虑单词之间的相似，增加了准确性。

## 3.2 问题二结果分析

### 3.2.1 分析并选定特征词

首先从所给文件附录 3 中提取用户留言的主题部分（见附件留言主题），通过分词和去停用词，得到留言主题的特征词（见附件留言主题 cut），在 Python 中使用 LDA 聚类算法得出热点特征词以及其权重如表 5、表 6 所示：

| 地点/人群 | 权重 |
|---|---|
| 业主 | 14 |
| 路口 | 10 |
| 县星沙 | 9 |

| | |
|---|---|
| 滨河 | 8 |
| 西地省 | 8 |
| 东路 | 7 |
| 公交车 | 7 |
| 医院 | 7 |
| 幼儿园 | 7 |
| 地铁 | 6 |
| 有限公司 | 6 |
| 新城 | 6 |
| 学院 | 6 |
| 城市 | 6 |
| 公园 | 5 |
| 小学 | 5 |
| 星沙 | 4 |
| 区金麓 | 4 |

表 5 地点/人群特征词权重表

| 问题 | 权重 |
|---|---|
| 扰民 | 26 |
| 噪音 | 24 |
| 魅力 | 15 |
| 违规 | 14 |
| 解决 | 12 |
| 举报 | 10 |
| 房屋 | 9 |
| 劳动 | 8 |
| 经济 | 8 |
| 加快 | 8 |
| 施工 | 7 |

| | |
|---|---|
| 油烟 | 7 |
| 捆绑 | 7 |
| 车位 | 7 |
| 销售 | 7 |
| 物业 | 6 |
| 门面 | 6 |
| 经营 | 6 |
| 请问 | 6 |
| 发展 | 6 |
| 项目 | 5 |
| 安置 | 5 |
| 开发商 | 5 |
| 污染 | 5 |
| 安置 | 5 |
| 购房 | 5 |

**表 6 问题权重表**

### 3.2.2 Excel 中检索特征词

通过 Excel 自带的筛选功能进行特征词的筛选，经过多次筛选处理最终得到留言中的热点问题并进行整理总结（见附件热点问题表）如下表 7：

| 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 | 55 | 2019/11/02 至 2020/01/26 | A 区丽发新城小区 | A 区丽发新城小区附近搅拌厂对环境污染严重、噪音严重影响周边居民生活 |
| 2 | 2 | 54 | 2019/07/07 至 | 广铁集团职工 | 伊景园滨河苑强行捆绑消费要求员工购买 |

| | | | | | |
|---|---|---|---|---|---|
| | | | 2019/09/01 | | 车位 |
| 3 | 3 | 21 | 2019/07/21 至 2019/12/04 | A 市魅力之城小区 | 小区一层临街餐饮店油烟噪音扰民 |
| 4 | 4 | 16 | 2019/01/02 至 2019/12/02 | A 市引进人才 | A 市对外来人才实行购房及其他形式补贴 |
| 5 | 5 | 10 | 2019/01/06 至 2019/09/12 | 西湖道茶场 | A3 区西湖街道茶场五组如何规划拆迁项目 |

**表 7 热点问题表**

## 3.3 问题三结果分析

从所给文件附件四中随机抽取五条答复样本（见附件样本），通过对答复样本的相关性、完整性、可解释性、时效性等角度运用对数最小二乘法计算得权重如表：

| 答复意见 | 相关性 | 完整性 | 可解释性 | 时效性 |
|---|---|---|---|---|
| 1 | 1 | 0.7 | 0.4 | 1.2 |
| 2 | 0.8 | 0.5 | 0.2 | 0.5 |
| 3 | 1 | 0.8 | 0.5 | 1.2 |
| 4 | 0.6 | 0.8 | 0.8 | 0.5 |
| 5 | 1 | 1.5 | 1 | 0.7 |

**表 8 意见答复权重表**

解释：相关性和可解释性是成本型指标，完整性是效益型指标，时效性是区间型指标，最优范围是[0.5,1]，最差上限为 2（及发布问题两个月后才给予答复），最差下限为 0（及发布问题不加思索便给予答复）四个指标的权重通过相关性分析分别定为：0.4,0.3,0.2,0.1。

定义好各指标权重后用 TOPSIS 算法构造加权规范矩阵决策各属性的权重向量，根据加权矩阵确定正理想解与负理想解通过计算正负理想解的距离得出综合评价指数如表 9：

| 答复意见 | 综合评价指数 |
|---|---|
| 答复意见 1 | 0.7464 |
| 答复意见 2 | 0.4545 |
| 答复意见 3 | 0.7807 |
| 答复意见 4 | 0.3380 |
| 答复意见 5 | 0.7295 |

**表 9 意见答复评价表**

根据表中数据可得样本中第一三五条答复意见对留言问题有相对完善且及时的解释，而第二四条样本答复意见在相对性、完整性、可解释性、时效性等角度上不能做到针对留言问题而做出相对合理的答复。

# 第四章 模型评价与改进

## 4.1 优点

第一问因为数据比较庞大，我们使用了 FastText 模型，对标签进行编码，极大地缩小了模型预测的数量，而且简化了计算复杂度，从而缩减了运行时间。FastText 模型更加具有灵活性，利用了 H-Softmax 的分类功能，遍历分类树的所有叶结点，找到概率最大的 label。

第二问我们先对于主题进行 jieba 分词和词频统计的工作，分词是将句子精准切开，适合文本分析，而且 jieba 分词从一个字符到下一个字符比对是不需要遍历该节点的所有子节点的，更加减少了运行的时间；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，增加了分词的准确率。运用 LDA 算法归类出排名靠前的特征词，把主题转化为权重向量，更加直观的计算出热点问题。

## 4.2 缺点

FastTest 模型不共享参数，在输出空间很大的情况下泛化能力较差。

运用 jieba 分词，虽然算法简单，但是降低了召回率，而且占用的内存较大，自定义字典时，带空格的词不支持。

LDA 算法局限性大，受样本种类限制，投影空间的维数最多为样本数量 N-1 维，使结果只是范围而不是精确值。

由于才疏学浅，能力不足，加之时间和精力有限，还是存在一些不足之处：本题的结果应该是一个比较精确的值，但是我们团队所计算的是一个最优解范围，虽然误差范围很小但还是可以进一步论证与解决。类似如此的问题上，存在内容表述、论证上存在着些许不当之处，与自己的期望还相差甚远，有些问题还有待进行一步思考和探究。

## 4.3 模型改进

本题中使用了较多的分词，以及权重的计算，很多原本的字典并不适合这个

题目的运用，可以自己多思考，建立自己的字典，让分词的结果更加的准确，这样对于挖掘的分析将会更加的有力。

# 第五章 参考文献

[1]赵静，但琦. 数学建模与数学实验[M].北京：高等教育出版社，2014.

[2]孙海峰，郑中枢，杨武岳.网络招聘信息的数据挖掘与综合分析[R].北京林业大学，2017.

[3]梁昌明，李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报.2015(12),34(12):1278-1283.

[4]https://baike.so.com/doc/6744530-10482973.html