# 智慧政务综合文本处理方法

“智慧政务”中的文本挖掘应用

# 摘要

随着物联网、云计算、移动互联网、Web 2.0 等新一代信息技术飞速发展，电子政务正由电子政府到“智慧政府”转变。而越来越多的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，对于以往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极大的挑战。因此自动从海量的留言数据中分类和挖掘有用的信息，对于提高政府的政务效率以及管理水平具有极大的推动作用。

针对于问题一：首先进行中文分词可以去除对有效信息造成的噪音干扰，利用去停用词策略能节省存储空间和计算时间，其次进行词语编码和文本向量化处理之后，用数据的 80%进行 TextCNN 训练得到分类模型，剩余的 20%进行检验数据，最后用 F-Score 对分类方法评价。

针对于问题二：首先基于 jieba 分词包，首先通过对照典生成句子的有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取。对于未登陆词（不在词典中的词）使用 HMM 进行新词发现。其次进行数据清洗提取有用的信息，利用层次聚类中的凝聚法进行信息分类。 根据自定义热度值：K=LR×RD×α+RR 对已处理好的信息进行热点信息的挖掘。

**关键字**: 智慧政务，文本挖掘，TextCNN，Jieba

# abstract

With the rapid development of new-generation information technologies such as the Internet of Things, cloud computing, mobile Internet, and Web 2.0, e-government is changing from e-government to "smart government." Increasingly, more and more online questioning platforms have become an important channel for the government to understand public opinion, gather peoples wisdom, and gather peoples popularity, which has brought great challenges to the work of relevant departments that used to manually divide messages and organize hot spots. Therefore, automatically classifying and mining useful information from massive message data has a great impetus to improve the governments government efficiency and management level.

For problem one: firstly, Chinese word segmentation can remove the noise interference caused by the effective information, and the use of the stop word strategy can save storage space and calculation time. Secondly, after word encoding and text vectorization, 80% of the data is used. TextCNN trains to obtain a classification model, and the remaining 20% is used to test the data. Finally, the classification method is evaluated with F-Score.

For problem two: First, based on the jieba word segmentation package, firstly generate a directed acyclic graph of the sentence by comparing the code, and then according to the selected mode, find the shortest path according to the dictionary to intercept the sentence or directly intercept the sentence. For unregistered words (words not in the dictionary), use HMM to discover new words. Secondly, the data is cleaned to extract useful information, and the information is classified using the aggregation method in hierarchical clustering. According to the custom heat value: $K = LR \times RD \times \alpha + RR$, the hot information is mined for the processed information.



**Keywords**: Smart Government Affairs, Text Mining, TextCNN, Jieba

# 目　录

# 1. 绪论

## 1.1 研究背景

（1）改革开始以来我国经济全面发展，人民生活水平不断提升但迎来更加复杂人口增长、食品安全、贫富差距等社会问题。政府只能通过转变行政方式才能更加积极有效应对日益复杂的社会矛盾。

（2）随着大数据、云计算、物联网、数据挖掘等技术不断出现和发展。不仅改变我们生活方式，同时也改变政府的服务模式。自从 IBM 公司提出“智慧地球”和 “智慧城市”的概念之后，赢得了全世界关注。而政府承担着治理城市的责任，其发展在建设城市变得尤为重要。 由此提出了智慧政务的概念，其目标在于实现智慧感知，为公众提供覆盖性、精准性的服务。

（3）随着越来越多的公民表愿意表达自己的诉求，而表达途径也是多样化的。尤其是在网络环境下，网络舆论、网络问政、信息公开等都考验着政府是否有全面把握信息的能力以及回应公众的能力，公民的力量改变政府行政方式和服务模式。

（4）国家相关政策的出台与支持。从十八届三中全会到 2016 的《政府报告中》都一直在强调要“创新行政管理方式”同年，国务院办公厅又转发了国家发展改革委等部门《推进互联网+政务服务信息惠民试点实施方案》的通知（以下简称《实施方案》），在该《实施方案》中，国务院部署了“互联网+政务服务”的总体思路、工作目标、主要任务、实施步骤与保障措施，旨

在解决群众“办事难”的问题，为智慧政务的发展提供了政策依据。

1.2　研究意义

目前，我国智慧政务的发展却存在诸多难题，而物联网和云计算等技术来收集群众问政留言并对其分类、分析，从而能够有针对性的解决群众的实际问题。

综上所述，本论文的研究意义在于:

（1）对于人民来说，在移动大数据和互联网时代下，公众更愿意通过各种渠道发布对公共服务的需求，政府根据智慧政务的留言分类进行针对性的回应，让公众感受到政府部门的服由公众向个性化服务转变。

（2）对于政府来说，政务变得更加智能化，服务呈现多元化，可以极大程度优化公共服务，提高行政效率。同时精简政府机构的工作，强化和规范自我管理。由于公众监督渠道的多元化。政府也能够通过多元化的途径收集群众留言信息，促进公民参与政务建设，实现以人为本的社会。

1.1 问题重述

（1）　在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提

供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。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。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

（2） 某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据附件 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，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。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。

## 1.2 技术路线

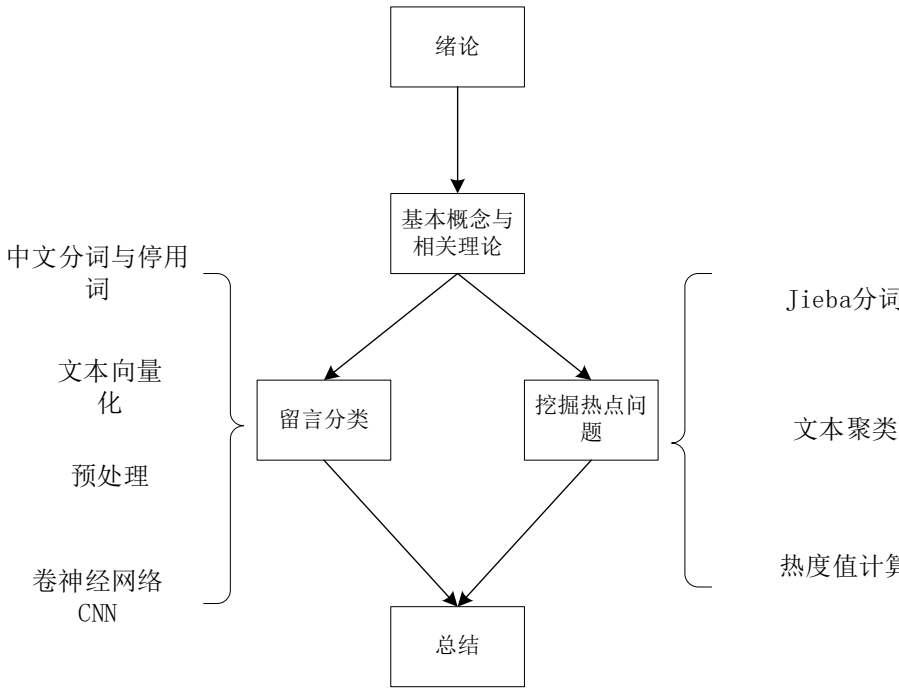



图 1 总体技术路线

# 2 基本概念与相关理论

围绕本文的主要研究问题及观点，本章对本文涉及的主要概念、理论进行解释分析，包括智慧政务的相关概念、卷积神经网络 CNN、Jieba 以及文本聚类等基本理论，为下文分析和构建智慧政务的文本挖掘奠定理论基础。

## 2.1 智慧政务

### 2.1.1 智慧政务概念

目前，国内外对“智慧政务”概念还没有较为明确的定义。国外与之相关的概念有”e-government”, ”smart government”, “smart city”等，其概念之间有许多相互交叉的部分。与此同时，学界也并没有较为严格的将“智慧政务”、“智慧政府”、“互联网+政务”等概念做出区分，但他们彼此之间的联系紧密，区别和分析他们之间的概念有助于我们了解智慧政务。因此，为了更好的研究智慧政务，将对 “智慧城市”、“智慧政务”、“智慧政府”、“互联网+政务”等几个概念进行阐述。

几个概念之间的关系如下图所示:

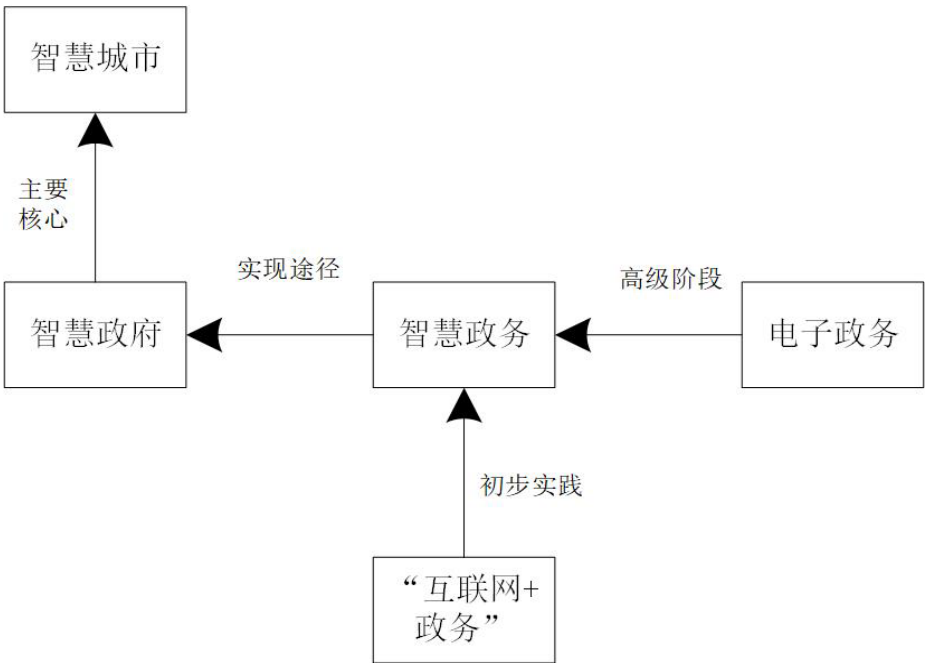



图 2 五种概念辨析图

表 1 智慧政务不同阶段比较

### 2.1.2 不同发展阶段的比较

| | 政府信息化前期 | 数字政务 | 移动政务 | 智慧政务 |
|---|---|---|---|---|
| 出现时间 | 1993-1999 | 1999-2003 | 2003-2010 | 2010-至今 |
| 政务载体 | 万维网 | 万维网 | Web2.0 | 实景网络 |
| 面向对象 | 面向政府内部 | 面向政府、柜台式 | 面向公众 | 面向公众 |
| 服务方向 | 内部单向 | 单向 | 单双向交叉 | 双向 |
| 主要技术特征 | 点对点 | 点对点 | 模块化、工具/方法明确 | 系统集成化、业务协同化、决策精确化、敏捷化、横向纵向有效整合 |

随着智慧政务不断形成于发展，其载体逐渐从万维网升级为实景网络；面向的服务对象逐渐从只面向政府内部转变为面向使用智慧政务的每一个个体；服务方式也逐渐从政府内部的单向转变为双向的服务；技术特点也从点对点变成了集约化、协同化。呈现出从低级到高级的发展，最后实现智慧政务的业务一致性、便捷性、智能型。下表（表 1）将对政府信息化前期、数字政务、移动政务、智慧政务的特点进行比较:

### 1.2 卷积神经网络 CNN

卷积神经网络是一类包含卷积计算且有一定深度结构的前馈神经网络（Feedforward Neural Networks）是

深度学习的代表算法之一。目前被广泛应用于计算机视觉中，随着词嵌入和深度学习技术的发展，很多人用在 NLP 中使用 CNN。如下图是关于 TextCNN 简单网络结构。

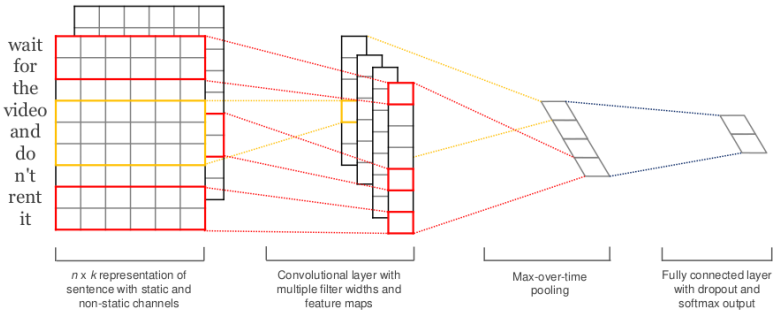



图 3 模型结构与两个通道

TextCNN 的详细过程原理图如下：

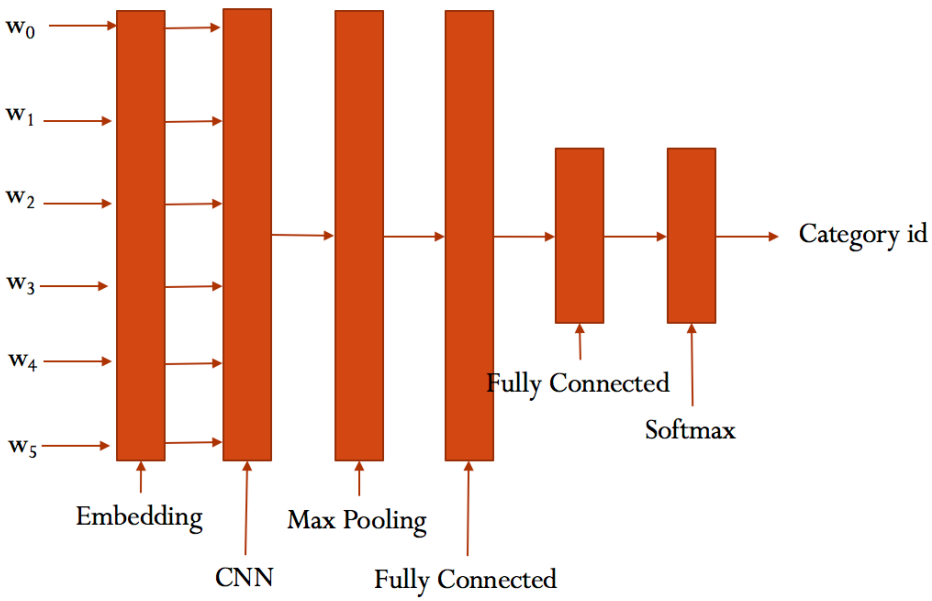



图 4 TextCNN 详细过程原理图

TextCNN 详细过程：

（1） 嵌入层(embedding layer)

第一层是图中最左边的 7 乘 5 的句子矩阵，每行是词向量，

维度=5，这个可以类比为图像中的原始像素点。

（2） 卷积层（Convolution Laye）

然后经过 kernel_sizes=(2,3,4) 的一维卷积层，每个 kernel_size 有两个输出 channel。

（3） 池化层（Pooling Layer）

第三层是一个 1-max pooling 层，这样不同长度句子经过 pooling 层之后都能变成定长的表示。

（4） 全连接层（Fully connected layer）

最后接一层全连接的 softmax 层，输出每个类别的概率。

## 2.3 Jieba 分词

Jieba 作为 Python 中文分词组件的翘楚，主要通过词典来进行分词及词性标注，两者使用了一个相同的词典。正因如此，分词的结果优劣将很大程度上取决于词典，但可以使用 HMM 来进行新词发现。

Jieba 分词包整体工作流程：

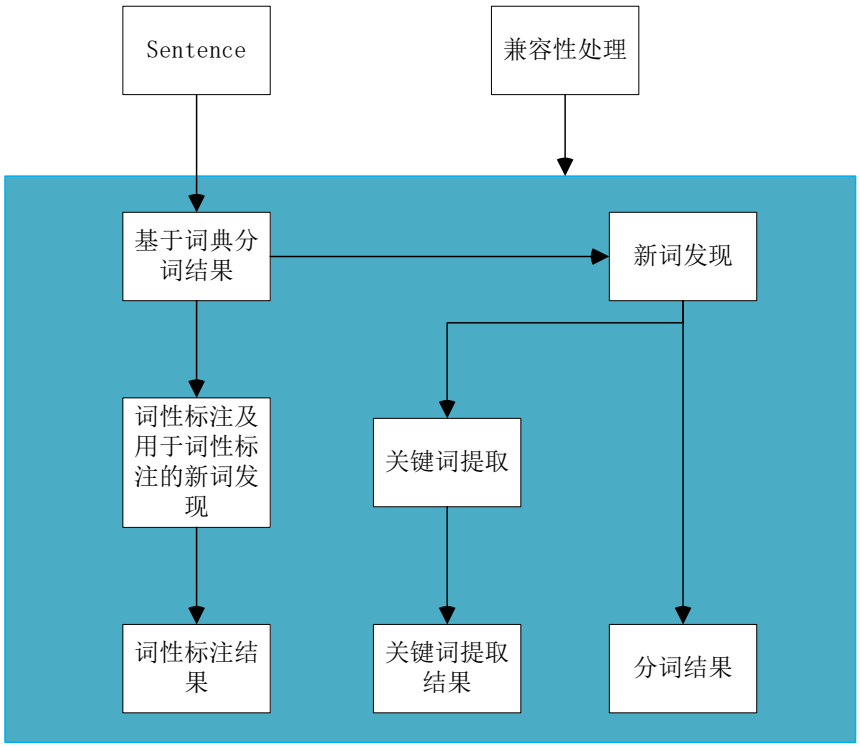



图 5 Jiaba 整体工作流程

## 2.4 文本聚类

层次聚类（Hierarchical Clustering）：就是要一层一层地进行聚类，可以从下而上地把小的 cluster 合并聚集，也可以从上而下地将大的 cluster 进行分割。所谓从下而上地合并 cluster，具体而言，就是每次找到距离最短的两个 cluster，然后进行合并成一个大的 cluster，直到全部合并为一个 cluster。整个过程就是建立一个树结构，类似于如下图:

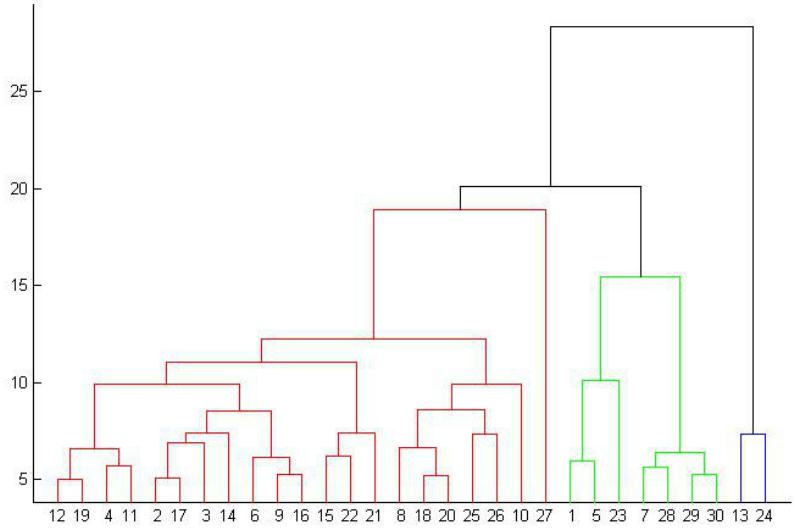



图 6 cluster 合并过程图

### 2.4.1 步骤

（1）移除网络中的所有边，得到有 n 个孤立节点的初始状态;

（2）计算网络中每对节点的相似度;

（3）根据相似度从强到弱连接相应节点对，形成树状图;

（4）根据实际需求横切树状图，获得社区结构。

### 2.4.2 依据

本次的聚类距离依据为杰卡德相似系数（Jaccard similarity coefficient），用于比较有限样本集之间的相似性与差异性。

系数的定义:

给定两个集合 A,B，Jaccard 系数定义为 A 与 B 交集的大小与 A 与 B 并集的大小的比值，定义如下：

J（A,B）=|A∩B|/|A∪B| =|A∩B|/(|A|+|B|-|A∩B|)

当集合 A，B 都为空时，J(A,B)定义为 1。

## 3 群众留言分类具体实现

### 3.1 实现步骤

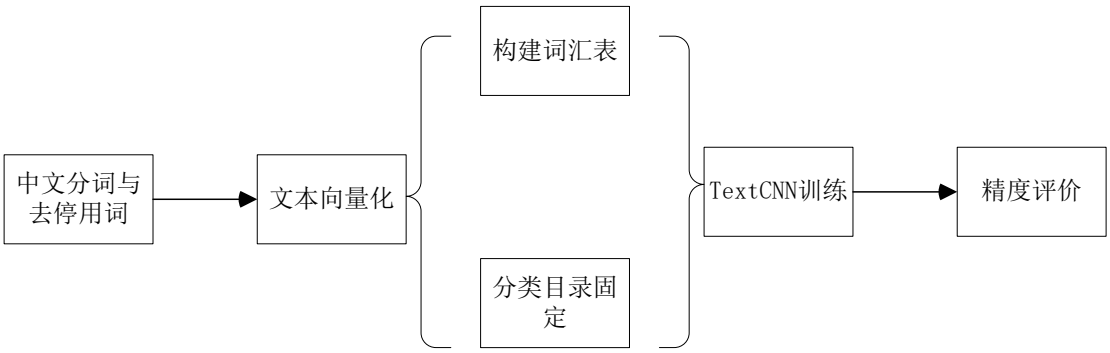



图 7 留言分类实现步骤

### 3.1 中文分词与去停用词

中文分词是将短文本切分成单独的词。例如“K 市中央新城小区电压低，电器都开不起”进行中文分词得到的词语列表为“[K 市 中央 新城小区 电压 低 电器 开不起 ]”。

（1） 基于词典分词算法

也称字符串匹配分词算法。该算法是按照一定的策略将待匹配的字符串和一个已建立好的“充分大的”词典中的词进行匹配，若找到某个词条，则说明匹配成功，

识别了该词。常见的基于词典的分词算法分为以下几种：正向最大匹配法、逆向最大匹配法和双向匹配分词法等。基于词典的分词算法是应用最广泛、分词速度最快的。很长一段时间内研究者都在对基于字符串匹配方法进行优化，比如最大长度设定、字符串存储和查找方式以及对于词表的组织结构，比如采用 TRIE 索引树、哈希索引等。

(2) 基于统计的机器学习算法

这类目前常用的是算法是 HMM、CRF、SVM、深度学习等算法，比如 stanford、Hanlp 分词工具是基于 CRF 算法。以 CRF 为例，基本思路是对汉字进行标注训练，不仅考虑了词语出现的频率，还考虑上下文，具备较好的学习能力，因此其对歧义词和未登录词的识别都具有良好的效果。常见的分词器都是使用机器学习算法和词典相结合，一方面能够提高分词准确率，另一方面能够改善领域适应性。随着深度学习的兴起，也出现了基于神经网络的分词器，例如使用双向 LSTM+CRF 实现分词器，其本质上是序列标注，所以有通用性，命名实体识别等都可以使用该模型，据报道其分词器字符准确率可高达 97.5%，算法框架如下图所示：

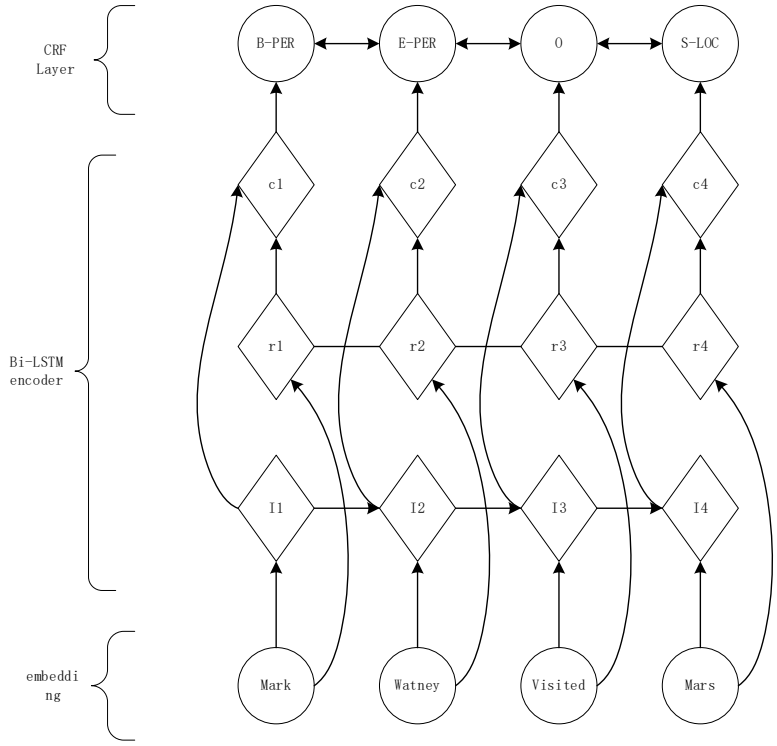



图 8 算法框架图

本文主要利用构建词典进行中文分词，因为文本中存在大量停用词会造成对于有效信息提取的噪音干扰。而本文进行去停用词方法既可以去除噪音也可以节省存储空间和计算时间。

3.2 文本向量化

将字符文本通过某种函数生成计算机可以理解的数值化数据。因为机器学习和深度学习的算法不能够直接处理含有字符文本的数据，所以在使用机器语言和深度学习时，需要将原始的文本内容用数值化的形式表示，即转换成数值向量。与此同时，不同的文本表现形式对模型结果的影响不同。因为 NLP 任务与算法模型不同，可以将字或词作为最小表示单元（本文中，以词作为文本表示的最小单元），进行数值向量化。

文本向量化大致分为两种：一种是 One-Hot 式编码和分布式表

示法（Distributed Representation）。

（1） One-Hot 式编码

又称“独热编码”。其实就是用 N 位状态寄存器编码 N 个状态，每个状态都有独立的寄存器位，且这些寄存器位中只有一位有效，就是只能有一个状态。One-Hot 编码是分类变量作为二进制向量的表示。这首先要求将分类值映射到整数值。然后，每个整数值被表示为二进制向量，除了整数的索引之外，它都是零值，它被标记为 1。例如，假设语料库中有三句话：

- 我爱中国
- 爸爸妈妈爱我
- 爸爸妈妈爱中国

首先，将语料库中的每句话分成单词，并编号：

1：我　　2：爱　　3：爸爸　　4：妈妈　　5：中国

然后，用 one-hot 对每句话提取特征向量：

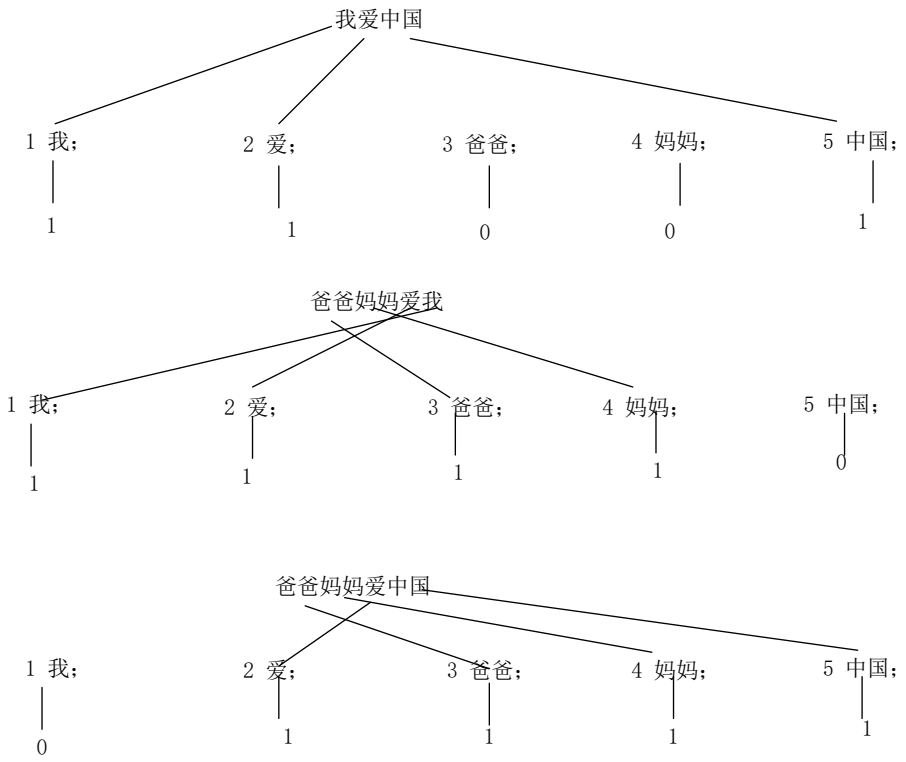



图 9 利用 one-hot 分类

所以最终得到的每句话的特征向量就是：

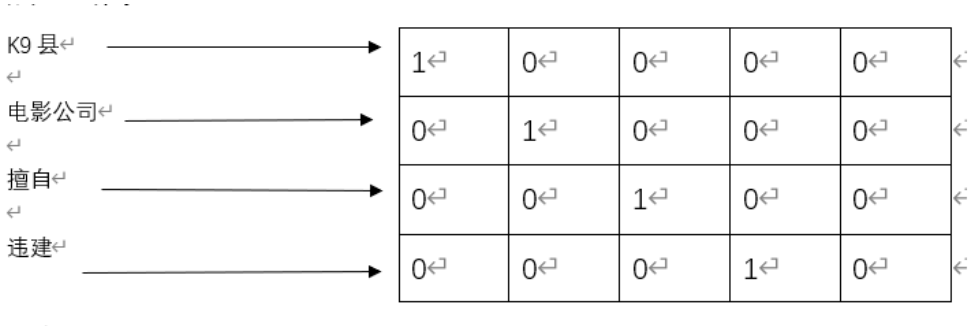
- 我爱中国 -> 1，1，0，0，1
- 爸爸妈妈爱我 -> 1，1，1，1，0
- 爸爸妈妈爱中国 -> 0，1，1，1，1

但这种方法有三个主要缺点，首先是 one-hot 是一个词袋模型，不考虑词与词之间的顺序问题，而在文本中，次的顺序是一个很重要的问题；其次 one-hot 是基于词与词之间相互独立的情况下的，然而在多数情况中，词与词之间应该是相互影响的；最后 one-hot 得到的特征具有离散型和稀疏性。

（2） 分布式表示法

与 one-hot 方法不同，分布表示法是基于“上下文相似的此，其语义也相似”，基本思想是利用统计学方法进行训练把句子中的每个字映射成 K 为的实数向量，如欧氏距离、余弦距离都是通过字与字的实数向量之间的距离去描述字与子之间的语义相似度，即相似语义中的字有相似数值向量。

现如今大多使用的是基于分布式表示法的文本表示模型。在本文的实验部分，例如 TextCNN 首先利用中文分词方法将“K9 县电影公司擅自违建”分为“K9 县“/”电影公司“/”擅自“/”违建”/”影院”，通过 Word2Vec 方式映射一个 5 维（维数可改变）词向量。如“”K9 县“->[1,0,0,0,0],”天气”->”[0,1,0,0,0]“,”违建”->[0，0，1，0，0]等等。如表二所示




| | | | | |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |

图 10 Word2Vec

3.3 进行 Convolution 卷积

Step1：例如分为“K9 县“/”电影公司“/”擅自“/”违建”对应的是 4*5 矩阵与卷积核做一个 point wise 的乘法然后求和，便是卷积操作：

feature_map[0] =0*1 + 0*0 + 0*1 + 0*0 + 1*0 + //(第一行)

0*0 + 0*0 + 0*0 + 1*0 + 0*0 + //(第二行)

0*1 + 0*0 + 1*1 + 0*0 + 0*0 + //(第三行)

0*1 + 1*0 + 0*1 + 0*0 + 0*0 //(第四行)

= 1

Step2：将窗口向下滑动一格(滑动的距离可以自己设置)," 分为" 电影公司"/"擅自"/"违建"/"影院" 对应的 4*5 矩阵 与卷积核(权值不变) 继续做 point wise 乘法后求和

feature_map[1] = 0*1 + 0*0 + 0*1 + 1*0 + 0*0 + //(第一行)

0*0 + 0*0 + 1*0 + 0*0 + 0*0 + //(第二行)

0*1 + 1*0 + 0*1 + 0*0 + 0*0 + //(第三行)

1*1 + 0*0 + 0*1 + 0*0 + 0*0 //(第四行)

= 1

feature_map 便是卷积之后的输出，通过卷积操作 将输入的 5*5 矩阵映射成一个 2*1 的矩阵，这个映射过程和特征抽取的结果很像，于是便将最后的输出称作 feature map。一般来说在卷积之后会跟一个激活函数，在这里为了简化说明需要，我们将激活函数设置为 f(x) = x

3.5 进行 channel 实验

| | | | | |
|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 |

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 |

图 11 Channel 1

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |

图 12 Channel 2

在 CNN 中常常会提到一个词 channel，上图的 channel1 和 channel2 便构成了两个 channel 统称一个卷积核，从这个图中也可以看出每个 channel 不必严格一样，每个 4*5 矩阵与输入矩阵做一次卷积操作得到一个 feature map. 在计算机视觉中，由于彩色图像存在 R, G, B 三种颜色，每个颜色便代表一种 channel。从论文实验结果来看多 channels 并没有明显提升模型的分类能力，七个数据集上的五个数据集但 channel 的 textCNN 表现优于多 channels 的 textCNN，如下图所示。

| *Model* | *MR* | *SST-1* | *SST-2* | *Subj* | *TREC* | *CR* | *MPOA* |
|---|---|---|---|---|---|---|---|
| CNN-rand | 78.1 | 45.0 | 82.7 | 89.6 | 91.2 | 79.8 | 83.4 |
| CNN-static | 82.0 | 45.1 | 86.8 | 93.0 | 92.8 | 84.1 | 88.7 |
| CNN-non-static | 80.5 | 48.0 | 87.2 | 93.4 | 93.6 | 84.3 | 89.5 |
| CNN-multiple | 81.1 | 47.4 | 88.1 | 93.2 | 92.2 | 85.0 | 86.4 |

表 2 Textcnn 实验

所以本文并没有采用 channel 进行留言分类。

### 3.6 max-pooling 处理

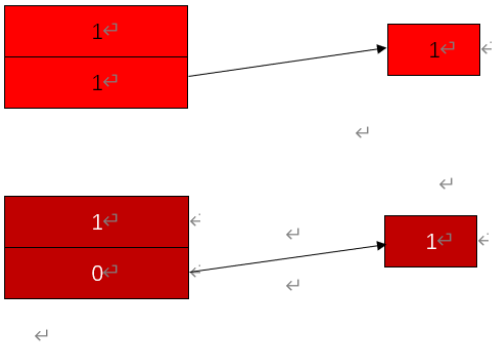



图 13 Max-pooling 说明

得到 feamap = [1,1] 后，从中选取一个最大值[1] 作为输出，便是 max-pooling。max-pooling 在保持主要特征的情况下，大大降低了参数的数目，从图五中可以看出 feature map 从二维变成了一维，

## 3.7 使用 softmax k 分类

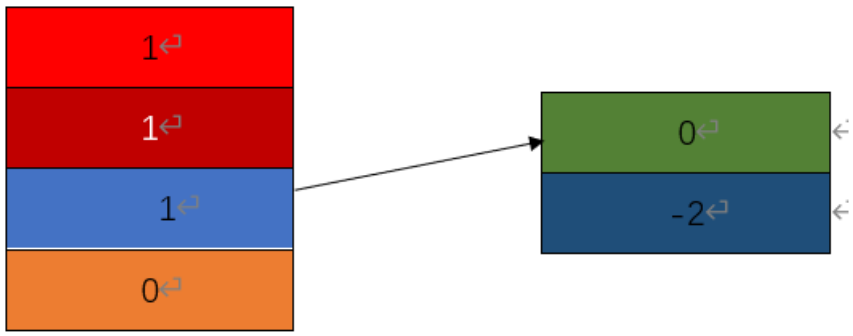



图 14 Softmax 示意图

如上图所示 我们将 max-pooling 的结果拼接起来，送入到 softmax 当中，得到各个类别比如 label 为 0 的概率以及 label 为-2 的概率。根据预测 label 以及实际 label 来计算损失函数，计算出 softmax 函数,max-pooling 函数，激活函数以及卷积核函数 四个函数当中参数需要更新的梯度，来依次更新这四个函数中的参数，完成一轮训练 。

## 3.4 分类评价

利用赛题给出 F-Score 对分类方法进行评价:

$$F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

其中 $P_i$ 为第 i 类的查准率，$R_i$ 为第 i 类的查全率。

故基于 TextCNN 的分类精确度如下图所示:

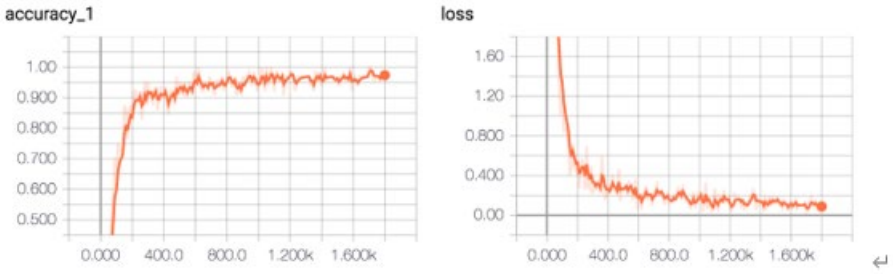



图 15 TextCNN 精确度图

# 4 热点问题挖掘具体实现

## 4.1 实现步骤

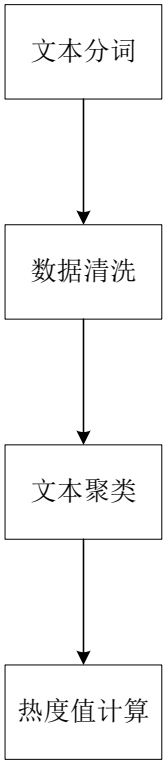



图 16 热点挖掘实现步骤

## 4.2 文本分词

### 4.2.1 jieba 分词

首先通过对照典生成句子的有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取。对于未登陆词（不在词典中的词）使用 HMM 进行新词发

现。

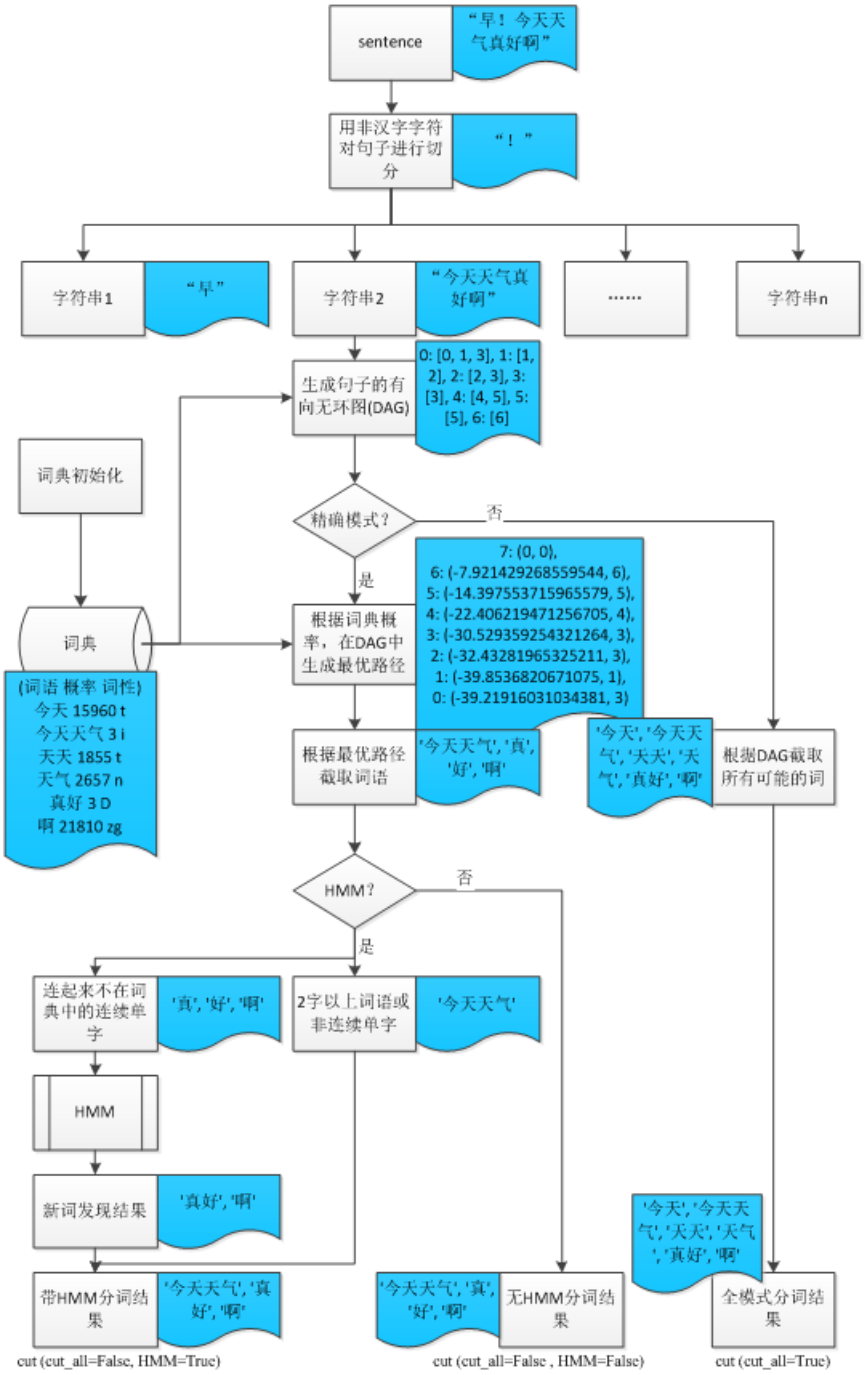



图 17 jiaba 分词过程

本文对每一个字符串进行图 17 分词过程，最后将切分的分词结果与非汉字部分依次连接起来，作为最终的分词结果。

4.2.2 HMM 实现流程

当文本中的字符串在词典中没有找到对应的单词，利用 HMM 模型进行新词发现。因为 HMM 属于模型的有向图 PGM,通过联合概率建模:

$$P(S,O)=\prod_{t=1}^{n}P(S_t|S_{t-1})P(O_t|S_t)$$

其中 S、O 分别表示状态序列与观测序列。

HMM 的解码问题为

$$arg^{max}_{S}P(S|O)$$

定义在时刻 t 状态为 s 的所有单个路径 s1 中的概率最大值为

$$\delta_t(s)=maxP(s_1^{t-1},o_1^t,s_t=s)$$

则有

$$\begin{gathered}\delta_{t+1}(s)\ =maxP(s_1^{t-1},o_1^t,s_{t+1}=s)\\=\ ^{max}_{s}P(s_1^{t-1},o_1^t,s_t=s)P(s_{t+1}|s_t)P(o_{t+1}|s_{t+1})\\=\ ^{max}_{s}[\delta_t(s')P(s|s')]P(o_{t+1}|s)\end{gathered}$$

## 2.5 数据清洗

### 2.5.1 停用词

根据自定义停用词利用 python 进行读入列表，将分词语句进行筛除。

### 2.5.2 同义词

自建同义词字典，用 python 读入成字典，

将相同意思的词语替换

2.5.3 关键词

jieba 分词中有两种不同的用于关键词抽取的算法，分别为 TextRank 和 TF-IDF。实现流程比较简单，其核心在于算法本身。

（1）TF-IDF

TF-IDF（词频-逆文本频率）是一种用以评估字词在文档中重要程度的统计方法。它的核心思想是，如果某个词在一篇文章中出现的频率即 TF 高，并且在其他文档中出现的很少，则认为这个词有很好的类别区分能力。

$$TF-IDF_{ij} = tf_{ij} * idf_i$$

其中：

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

式中，分子为 i 词在 j 文档中出现的次数，分母为 j 文档中所有字词出现的次数之和。

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|}$$

式中分子为语料库中的文件总数，分母为包含该词的文件数目。

jieba 分词中逆文档频率直接由词典读入。

（2）TextRank

TextRank 是一种用以关键词提取的算法，因为是基于 PageRank 的，利用投票机制对文本中重要成分进行排序。如果两个词在一个固定大小的窗口内共同出现过，则认为两个词之间存在连线。

TextRank 算法的得分定义为：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in \mathrm{In}(V_i)} \frac{W_{ji}}{\sum_{V_k \in Out(V_j)} W_{jk}} WS(V_j)$$

Vi：要提取的文本

Vj：能提取文本的视窗

WS(Vj)：Vj 的 TextRank 值

In(Vi)：全文本

Out(Vj)：存在要提取文本的视窗集

|Out(Vj)|：要提取的文本个数

多次迭代直至收敛，即可得到结果。在 jieba 分词中，TextRank 设定的词窗口大小为 5，将公式 1 迭代 10 次的结果作为最终权重的结果，而不一定迭代至收敛。

## 2.6 文本聚类

作为 NLP 领域最经典的使用场景之一，文本据类形成了如下的一般流程图。

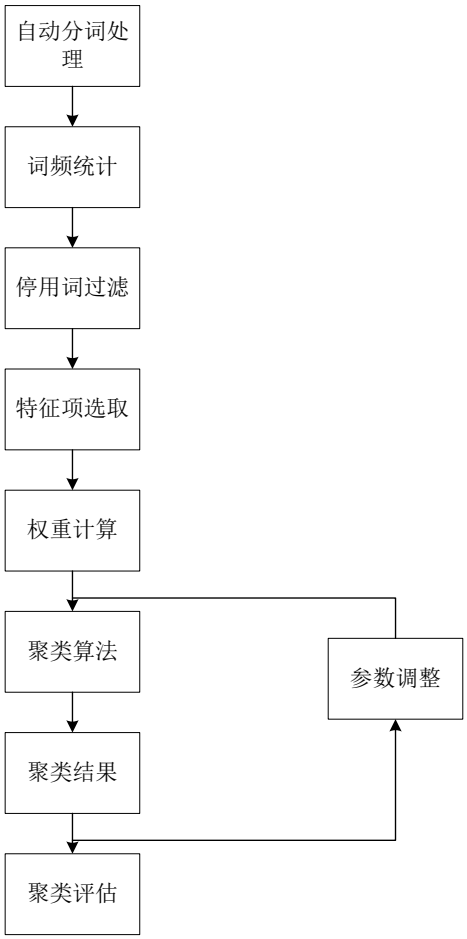



图 18 文本聚类的一般流程图

本文考虑到文本向量维度较高不利于聚类问题，采用增加 pca 将为的方法如图 19 所示。

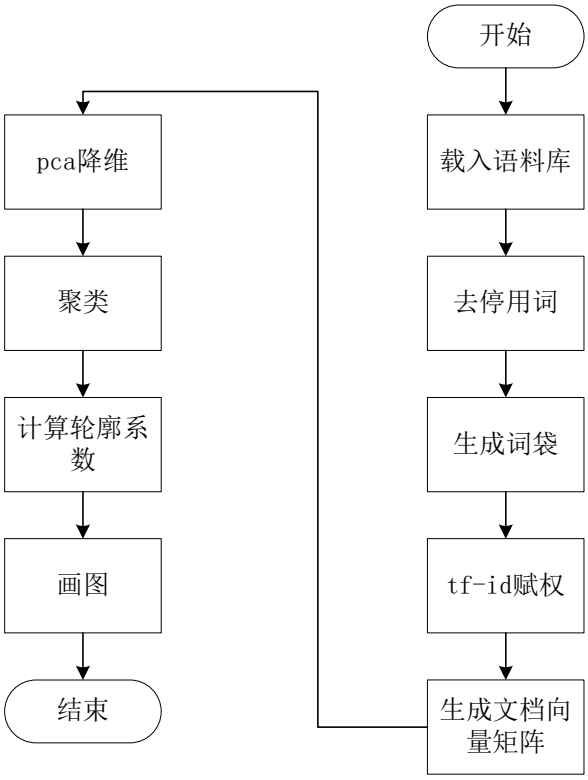



图 19 文本聚类的方法

## 2.7 设计热度值并实现

问题反应热度值是在网络问政平台中，衡量群众所反映的一个问题对群众生活的影响程度

热度值

$$K=LR\times RD\times \alpha+RR$$

其中 反映次数：α、反映人数：β、点赞数：L 反对数：U

赞同比：LR=L/(L+U)、真实反映人数：RR=L+β、反映分散度：RD=β/α

# 5 结语

本文的目的是通过数据挖掘和自然语言技术处理群众留言内容，减轻政府的工作压力，提高政府的工作效率及时解决人民生活问题。

# 参考文献

[1]刘洁.智慧政务APP应用问题与对策[J].合作经济与科技,2020(08):169-171.

[2]阳敏辉.智慧城市信息智能服务模式的构建[J].山西建筑,2020,46(07):195-197.

[3]王涛.“新型智慧政务”将迎来跨越式进步[J].中外管理,2020(Z1):30-32.

[4]杜鹏.国内政府智慧服务模型构建研究[J].经济研究导刊,2020(07):174-177.