# 基于自然语言处理技术建立智慧政务系统


## 摘要：

随着大数据时代的到来，从文本中挖掘数据建立智慧政务系统逐渐成为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。因此，本文将根据自然语言处理技术以及文本挖掘的方法来建立智慧政务系统模型，以方便来处理政务相关问题。

针对问题一：首先对附件 2 中的文本数据进行预处理，包括数据清洗、利用 jieba 进行中文分词、添词典去停用词、绘制词云。然后将处理好的文本利用 IF-IDF 权重策略用向量表示，接着利用 python 工具建立高斯朴素贝叶斯模型来完成分类工作，并定义模型预测函数来判断文档或语句的类别。最后用 F-Score 对该分类方法进行评价。

针对问题二：首先对原始数据进行预处理，对得到的数据进行热点提取、热点集合。使用 LDA 快速且准确的对文本流进行分类得到主题。再将此作为热点挖掘的输出数据，利用 K-means 方法进行聚类。最后热点评价根据其留言数、点赞数以及反对数作为指标，再利用因子分析法进行热度验证。

针对问题三：我们从相关性、完整性以及答复效率三个角度进行评价。其中相关性利用 jaccard 相似度的原理来得到；答复效率通过对留言与答复的时间间隔进行求解；完整性通过 0 和 1 的准确率，最终计算出回复的正确率得到。



**关键词：jieba 分词　word2vec　IF-IDF 权重策略　高斯朴素贝叶斯模型　LDA　K-means　jaccard 相似度　因子分析**

# Based on NLP(Natural Language Processing) to establish a system to deal with intelligent government affairs


## Abstract:

With the advent of the era of big data, it has gradually become a new trend of social governance innovation and development to mine data from the text to build a smart government system, which plays a great role in promoting the management level and efficiency of the government. Therefore, based on natural language processing technology and text mining methods, this paper will establish a model of intelligent government system to facilitate the handling of government affairs related issues.

For problem 1: firstly,preprocess the text data in attachment 2, including data cleaning, Chinese word segmentation by jieba, adding dictionary to stop words, and drawing word cloud. Then the processed text is represented by vector using the if-idf weight strategy, then the gaussian naive bayesian model is established by using python tools to complete the classification work, and the model prediction function is defined to judge the category of document or statement. Finally, f-score is used to evaluate the classification method..

For problem 2: firstly, preprocess the original data, and extract and collect hot spots from the obtained data. Use LDA to quickly and accurately categorize text streams to get topics. This is then used as the output data of hot spot mining, and k-means method is used for clustering. Finally, the hot spot evaluation was conducted according to the number of comments, the number of thumb up and the number of opposition, and the factor analysis was used to verify the heat..

For problem 3:we evaluate them in terms of relevance, completeness, and response efficienty. The relevance is obtained by using the Jaccard similarty principle, the reply efficiency is obtained by calculating the time interval between the message and reply, and the integrity is obtained by caluating the intergrity is obtained by caluating the correct rate of the reply with the accuracy rates of 0 and 1.



**Keywords: jieba Word2vec IF-IDF Guassian Naïve Bayes LDA K-means jaccard Factor Analysis**

# 目 录

# 1. 引言

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，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和不便之处。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。

随着自然语言处理(NLP)以及大数据、 云计算等技术等人工智能领域的飞速发展,建立基于自然语言处理技术(NLP)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。

因此，本文将基于自然语言处理技术以及人工智能来构建用于处理智慧政务相关的模型。

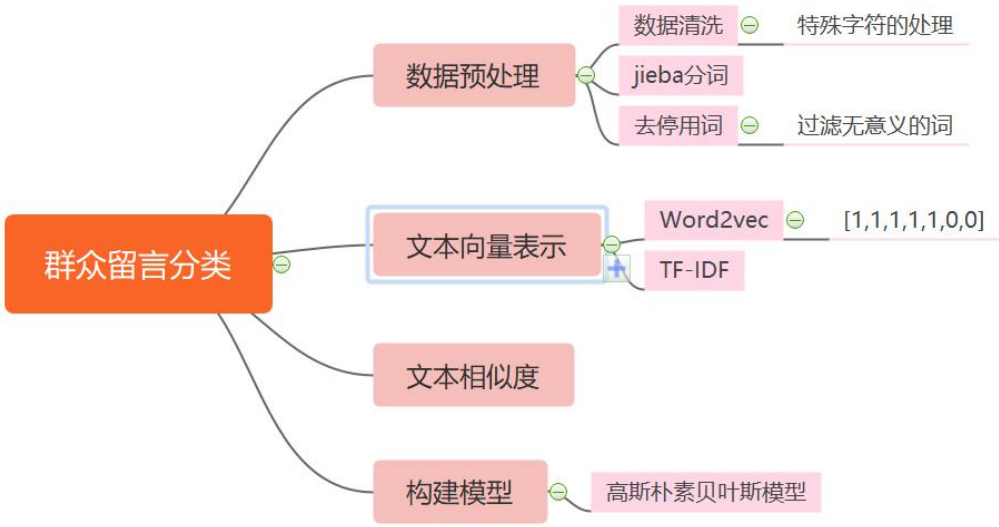
# 2. 问题分析

## 2.1 问题一的分析

问题一要求根据附件 2 给出的数据,建立关于留言内容的一级标签分类模型。

由于数据太多电脑运行受限，本文将通过对附件 2 中的关键内容进行提取并进行过抽样和欠抽样进行简化。通过 python 对抽取出的数据进行数据预处理，用 IF-IDF 权重策略将文本用向量表示，通过建立高斯朴素贝叶斯模型[1]来实现分类，并利用到了 F-Score 对分类方式进行评价。

问题一思路图如下：




图（1）

## 2.2 问题二的分析

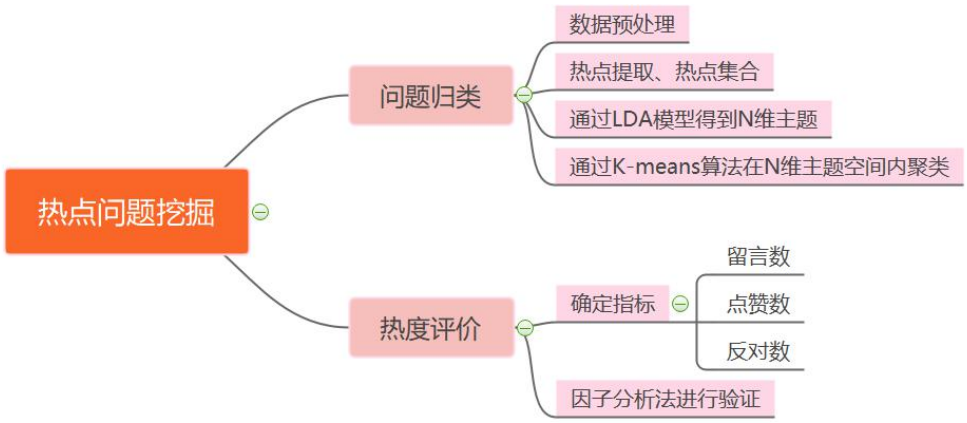
问题二要求根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。

问题二主要包括了两个子任务，问题归类以及热度评价。

其中问题归类就是把特定地点和人物的数据归并，即把相似留言归为同一类问题，并由此得到表 2。为了快速且准确的对文本流进行分类，使用 LDA 进行文本聚类，其输出为标记了主题的文本数据集，再将此作为热点挖掘的输出数据。最后利用 K-means 方法进行聚类。

对于热度评价，我们首先根据其点赞数和反对数确定指标，再利用因子分析法进行热度验证。

问题二思路图如下：




图（2）

## 2.3 问题三的分析

问题三要求根据针对附件 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。

对此问题，我们从相关性、完整性、可解释性以及答复效率四个角度进行评价。其中相关性通过对 python 工具的运用利用 jaccard 相似度的原理来得到，完整性和可解释性通过将答复分为解决和未解决两类，答复效率对留言与答复的时间间隔进行求解，利用简单的 python 间隔求法。


图（3）

## 3. 符号说明

| 符号 | 说明 |
| --- | --- |
| N | N 表示文本的总词汇数 |
| w | w 表示文章中出现的词 |
| $N_w$ | $N_w$表示词 w 在文本中出现的次数 |
| $D_w$ | $D_w$表示包含词 w 的文档数 |
| D | D 表示总文档个数 |
| t | t 表示某个词的 TF-IDF 值 |
| x | x 代表样本属性 |
| y | y 代表样本标签 |
| $\mu_{y_k}$ | $\mu_{y_k}$ 表示全部属于类$y_k$的样本中的变量$x_i$ *的均值* |

| | |
|---|---|
| $\sigma^2_{y_k}$ | $\sigma^2_{y_k}$表示全部属于类$y_k$的样本中的变量$x_i$*的方差* |
| D1 | D1 表示语料库 |
| M | M 表示语料库中文档的总数 |
| K | K 表示 k 个聚类中心 |
| $\boldsymbol{c_i}$ | $c_i$表示第几个中心 |
| dist | Dist 表示欧几里得距离 |
| $\boldsymbol{w_i}$ | $w_i$为旋转前或旋转后因子的方差贡献率。 |

(表 1)

# 4. 模型建立

## 4.1 数据预处理

由于数据内容太过庞大，电脑的运行效率过慢，这里我们采用抽样的方式对数据进行抽取，通过过抽样和欠抽样来进行。所谓过抽样就是通过增加少数类样本来提高少数类的分类性能。欠抽样是通过减少多数类样本来提高少数类

的分类性能。通过 python 工具得出每个类别的数量，如图：

```
In[5]: data['一级标签'].value_counts()
Out[5]:
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           938
卫生计生           877
交通运输           613
Name: 一级标签, dtype: int64
```

图（4）

根据图中的数据可在每一类别中取随机抽取 500 个样本进行计算。

### 4.1.1 数据清洗以及 jieba 分词

主要是对文本内容进行去特殊符号、去空格、去重以及利用 jieba 包对中文内容进行分词等处理。在这里我们利用 python 工具实现这一操作来得到去词后的内容。

示例：

```
>>>seg_list = jieba.cut("建议成立卫生基层工作领导小组")
>>>print ", ".join(seg_list)
建议, 成立, 卫生, 基层, 工作, 领导, 小组
```

### 4.1.2 添词典去停用词

在分词结束的的 python 程序中，将停用词包[2]导入，删除停用词。由于特有名词的存在使得分词的结果可能不尽人意，因此需将自定义的词典导入，得到进一步简化的准确率较高的分词结果。

### 4.1.3 绘制词云

词云图是显示文本结果、词频大小的有力工具，通过词云图的展示可以对文本内容数据分词后的高频词予以视觉上能给人造成视觉上的强调突出效果。对所需文本词汇的数量有了具体的表述，将高频词较为清晰的展现出来，使人一眼看到主要信息。通过对 python 工具下的 console 运用以及导入绘制的词云图轮廓可获得一级标签中七大类别的词云图。



图（5） 绘制词云图所用轮廓图

## 4.2 文本的向量表示

关于文本的向量表示，主要运用到了 TF-IDF 权重策略来实现。

TF-IDF 策略又称词频-逆向文件频率策略，是一种用于信息检索十分有效的文本加权技术。

$$\text{词汇频率：TF}=\frac{N_w}{N}$$

因此，词频表示一个词语与一篇文章的相关性，TF 越大，词 w 与文本的相关性越强。

$$\text{逆文档频率：IDF}=\log\frac{D}{D_w}$$

因此，逆文档频率表示一个词语出现的普遍程度。

经计算可得

$$\mathrm{TFIDF}(t_i,\ d_i)=\mathrm{TF}(t_i,\ d_i)\log\left[\frac{N}{N(t_i)}\right]$$

因此有

$$\mathrm{TFIDF}=\mathrm{TF}*\mathrm{IDF}$$

因此，我们可以根据此公式得到留言详情中每个词语的 TF-IDF，并于余弦定理相结合，便可得到与一级标签中的分类得匹配度。

利用 python 工具将预处理完成的数据利用 TF-IDF 模型转化成词频向量，最终转化成 TF-IDF 矩阵，来将文本数据放入模型。

## 4.3 高斯朴素贝叶斯模型的建立

高斯朴素贝叶斯模型（Guassian Naïve Bayes）[1]主要用于实现分类任务，弥补原始的朴素贝叶斯模型对连续变量的缺点。

其分类原理就是利用贝叶斯公式根据特征的先验概率计算出其后验概率，然后选择具有最大后验概率的类作为该特征所属的类，并且所有的特征之间是统计独立的。

统计独立：

假设某样本 X 有 a1, a2, a3, ……, an 个属性，那么如果有 P(X)=P(a1, a2, a2, a3, ……, an)=P(a1)*p(a2)*p(a3)*……*P(an)。则称这种现象为特征统计独立。

高斯朴素贝叶斯的本质就是先验为高斯分布的朴素贝叶斯，朴素贝叶斯表

达式为：

$$P(y|x)=\frac{P(y)P(x|y)}{P(x)}=\frac{P(y)}{P(x)}\prod_{i=1}^{d}P(x_i|y)$$

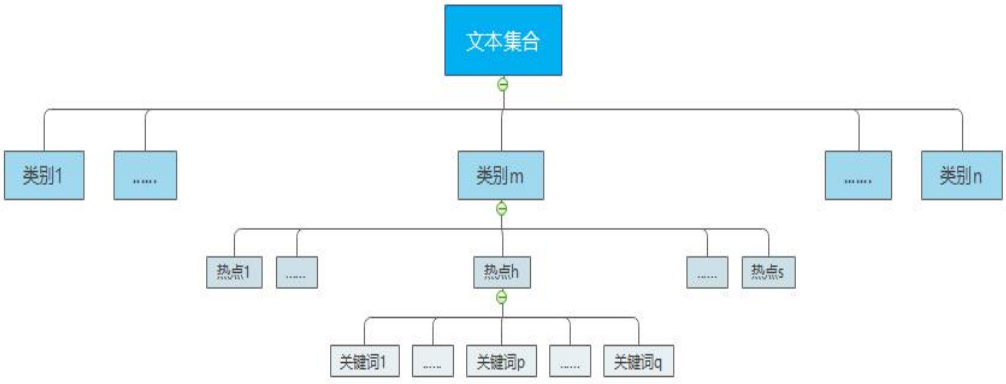
因此可得到，高斯朴素贝叶斯表达式为：

$$P(x_i=v|y_k)=\frac{1}{\sqrt{2\pi\sigma^2{}_{y_k}}}\exp\left(-\frac{(v-\mu_{y_k})}{2\sigma^2{}_{y_k}}\right)$$

最后将该模型利用 python 工具实现。

## 4.4 LDA 模型的建立

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。如图（4）：




图（6） LDA 的聚类模型

生成过程：对于语料库中的每篇文档，LDA 定义了如下生成过程(generativeprocess):

1.对每一篇文档，从主题分布中抽取一个主题;

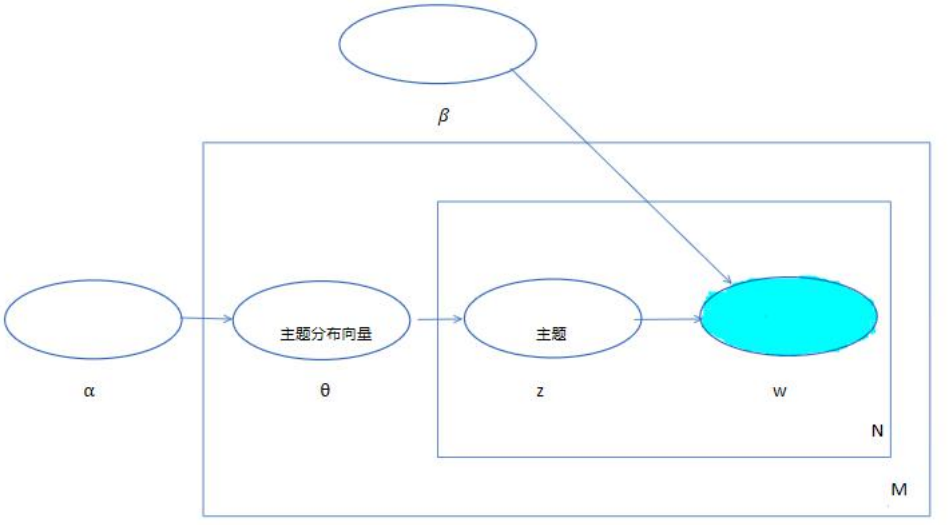
2.从上述被抽到的主题所对应的单词分布中抽取一个单词;

3.重复上述过程直至遍历文档中的每一个单词。

把这个模型的每一个参数都当作一个变量，每一个参数都有控制这个参数的参数。主题通式如下：

$$P(词语|文档)=\sum_{主题} p(词语|主题) * p(主题|文档)$$

其中，P(词语|文档)是一个一元的语言模型，包含 $p(词语|主题)$ 以及 $p(主题|文档)$，因此 LDA 的基本思想即是文档-主题-词语模型。对此我们首先选定一个主题向量，确定每个主题被选择的概率，在生成每个单词时，从主题分布向量θ中选择一个主题 z，按主题 z 的单词概率分布生成一个单词。如下图：




图（7） 主题向量概率模型

由该图可知 LDA 的联合概率为：

$$p(\theta,z,w|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\beta)$$

由该公式可计算边缘概率，得到：

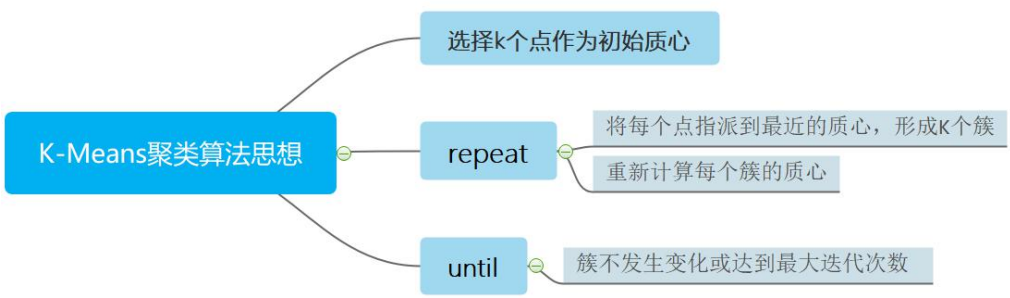
$$p(D_1|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left[ \prod_{n=1}^{N_d} \sum_{z_{d_n}} p(z_{d_n}|\theta_d) p(w_{d_n}|z_{d_n},\beta) \right] d\theta_d$$

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 θ(multinomialdistribution)相对应，每个主题又与词汇表(vocabulary)中的 w 个单词的一个多项分布ϕ相对应。

最后利用 python 语言将其实现。

## 4.5 K-means 聚类算法模型的建立

聚类分析[10]是在数据中发现数据对象之间的关系，将数据进行分组，组内的相似性越大，组间的差别越大，则聚类效果越好。




图（8） K-means 算法思想

这里的重新计算每个簇的质心，如何计算的是根据目标函数得来的，因此在开始时我们要考虑距离度量和目标函数。

考虑欧几里得距离的数据，使用误差平方和（Sum of the Squared Error，SSE）作为聚类的目标函数，两次运行 K 均值产生的两个不同的簇集，我们更喜欢 SSE 最小的那个。

$$\mathrm{SSE}=\sum_{I=1}^{k}\sum_{x\in c_i} dist(c_i,x)^2$$

我们可以对第 k 个质心$c_k$求解，最小化公式：对 SSE 求导，令导数=0，并求解$c_k$，如下所示：

$$\frac{\partial}{\partial c_k}SSE=\frac{\partial}{\partial c_k}\sum_{i=1}^{K}\sum_{x\in c_i}(c_i-x)^2$$

$$=\sum_{i=1}^{K}\sum_{x\in c_i}\frac{\partial}{\partial c_k}(c_i-x)^2$$

$$=\sum_{x\in k}2(c_k-x_k)=0$$

$$\sum_{x\in c_k}2(c_k-x_k)=0\overset{得到}{\Longrightarrow}m_kc_k=\sum_{x\in c_k}x_k\overset{得到}{\Longrightarrow}\frac{1}{m_k}\sum_{x\in c_k}x_k$$

如此可得到，簇的最小化 SSE 的最佳质心是簇中各点的均值。

该模型利用 python 语言实现。

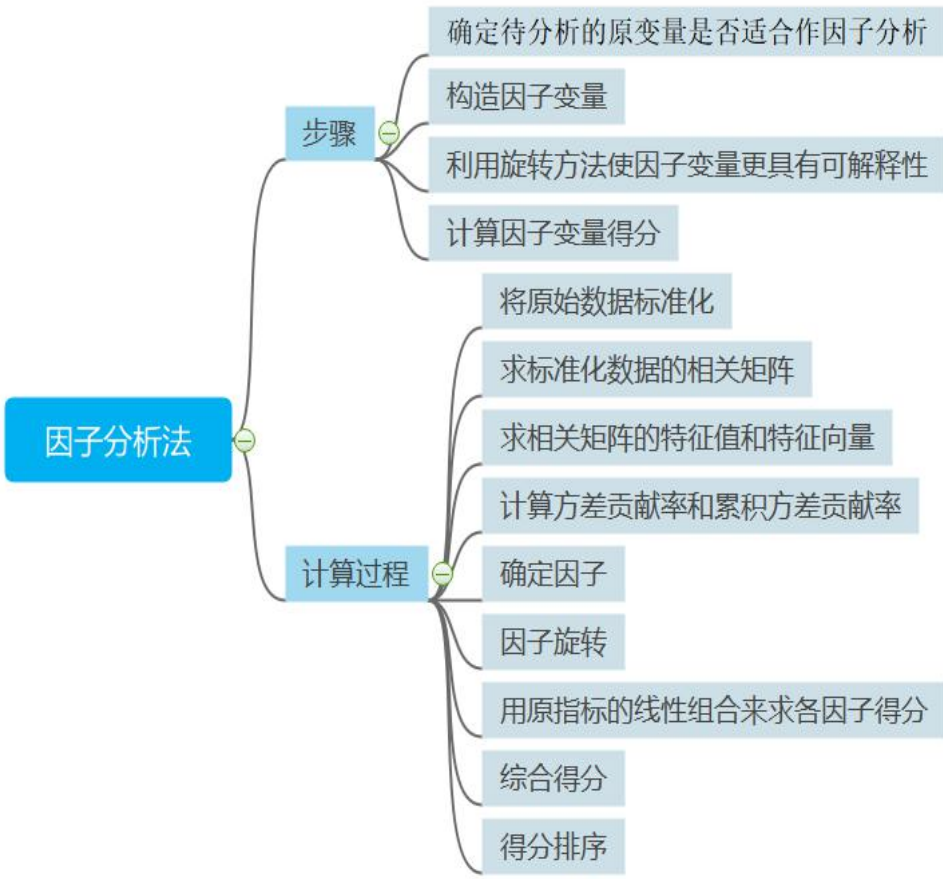
## 4.6 因子分析法

因子分析法是从研究变量内部相关的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。

应用 SPSS 25 将原始数据标准化并计算出相关系数矩阵，再进行 KMO 和 Bartlett 球形度检验，KMO 和 Bartlett 的检验结果显示，KMO 值>0.5，勉强适合做因子分析，显著性<0.5，说明指标间具有相关性，适合做因子分析。由相关系数矩阵求得其特征值与方差累计贡献。

因子分析模型建立后，还有一个重要的作用是应用因子分析模型去评价每个样品在整个模型中的地位，即进行综合评价。把留言数、点赞数、反对数三个指标提取为热度因子和内容因子，并按因子得分函数求得综合得分。由各因子的线性组合得到综合评价指标函数：

$$F = \frac{(w_1F_1 + w_2F_2 + \ldots + w_mF_m)}{(w_1 + w_2 + \ldots + w_m)}$$




图（9） 因子分析法思想

## 4.7 Jaccard 相似度

jaccard index 又称为 jaccard similarity coefficient 用于比较有限样本集之间的相似性和差异性。

定义：给定两个集合 A,B jaccard 系数定义为 A 与 B 交集的大小与并集大小的比值 $J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$，jaccard 值越大说明相似度越高。

当 A 和 B 都为空时，jaccard(A,B)=1；

与 jaccard 系数相关的指标是 jaccard 距离用于描述不相似度，公式为

$$d_j(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}=\frac{A\Delta B}{|A\cup B|}$$

例如：如果比较 X 与 Y 的 Jaccard 相似系数，只比较 xn 和 yn 中相同的个数，公式如下：

如集合 A={1,2,3,4};B={3,4,5,6};

那么他们的 J(X,Y)=1{3,4}/1{1,2,3,4,5,6}=1/3;

jaccard 相似度的缺点是值适用于二元数据的集合。虽然 jaccard 主要是在维度分析这样的稀疏向量中作用比较大，但是在文本相似度计算时也可用 jaccard。如下：

> 样本1：今天天气真好
> 样本2：今天天气不错

首先要做的还是分词：

A = [今天,天气,真好]

B = [今天,天气,不错]

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{2}{4} = 0.5$$

转化为01向量，用matlab验证下：

```
octave:7> X=[1,1,1,0;1,1,0,1]
X =

   1   1   1   0
   1   1   0   1

octave:8> D= pdist( X , 'jaccard')
D =  0.50000
```

图（10）

# 5. 模型求解

因为数据输出结果繁多，因此在正文中只显示部分输出结果截图，详细输出结果以及有关每个结果求解得 python 代码详情请见附件。

## 5.1 数据预处理结果

### 5.1.1 数据清洗与 jieba 分词

利用 python 中的 jieba 包对提取出的数据内容进行清洗与分词，部分结果如下：

```
In[6]: data_jieba
Out[6]:
1041    [ ,  , K, 市, 帝王, 广场, 7, 栋, 工程质量, 存在, 安全隐患,  , ...
1607    [我们, 是, A, 市, A5, 区, 桃花, 坡, 路, 58, 号庆德, 水韵, 山城...
1394    [ ,  , 建议, 城市, 的, 建设, 和, 规划, 中多, 给, 孩子, 们, 留点,...
855     [ ,  , 和平, 镇, 有, 个, 叫, 王, 才学, 的, , , 我们, 知道, 他,...
882     [尊敬, 的, 领导, ：, 您好, !, 在, 无, 任何, 公告, 的, 情况, 下, ...
879     [ ,  , 珠泉, 开发区, 港嘉, 娱乐城, 水压, 非常, 之, 小, 已经, 一年,...
606     [ ,  , 我们, 是, 西地省, E, 市, E7, 县, 的, 新, 就业, 人员, ...
1946    [自来水, 水质, 太差, 了, , , 黄褐色, 的, 用, 桶, 放, 一会, 底下, 就...
1395    [请求, 相关, 单位, 与, 领导, 能, 不能, 尽快, 落实, I6, 市路, 街道,...
1785    [ ,  , 楚发, 改价服, 〔, 2016, 〕, 144, 号, ″, 各, 服务, ...
739     [2018, 年, 5, 月, 8, 日, , , G, 市, 晚报, 刊登, 了, 由, G...
828     [J, 市内, 规化, 是, 一届, 领导, 换, 一次, 规化, 吗, ？, 现, 市内,...
903     [你好, , , 我, 是, 安任, 一名, 普通, 老百姓, , , 现向, 上级, 反映, ...
131     [ ,  , 投诉, 事项, ：, A3, 区德润园, 小区, 管理方, A, 市景, 鸿,...
801     [J9, 县, 沤, 江镇, 扶贫, 小区, （, 新, J9, 县, 车站, 旁, ）, ...
1606    [各位, 同仁,  ,  , 陆家, 渡, 大桥, 项目部, 在, 在建, 中, , , 项目...
```

图（11）　　数据清洗与 jieba 分词

从该图中可以看出数据得到了初步的清理，删掉了文中原有的\n、\t 等空格表示，并将每行内容进行分词。

### 5.1.2　添词典去停用词

在原有程序的基础上，通过导入停用词文档和自定义的常用词文档，对现有内容做进一步合理简化，部分结果如下：

```
Out[7]:
893     [富新, 国际, 因有, 长期, 欠, 物业费, 老赖, 合法, 资质, 物业, 种种, 导...
727     [F9, 市, 名流, 花园, 六幢, 户, 业主, 三月份, 围墙, 外厕, 刘初志, 私...
1471    [M1, 区, 黄, 泥塘, 办事处, 东来, 村关, 塘, 组涟, 钢汉, 大三, 六零,...
1195                [领导, 我想, 问下, K1区, 政务, 中心, 对面, 房子, 开, 户]
1119    [张, 局长, K1区, 虹彩, 燃气公司, 年初, 管理人员, 非法, 集资, 导致, 目...
1926    [领导, 网传, 国家, 取消, 棚户区, 改造, 政策, 请问, G6, 县道, 水南岸,...
896     [J10, 县城, 好多, 路灯, 坏, 不亮, 没人管, J10, 县, 大道, 漆黑, ...
1088                  [请问, K2区, 马坪, 农业, 开发区, 纳入, 总体规划, 文件]
1331    [L12, 市, 招标办, 西地省, 招标, 投标, 监管网, 发布, L12, 市, 第一...
1669    [鹿, 厅长, 市金茂府, 签定, 购买, 市金茂府, 方茂苑, 4304, 认购书, 付定...
1654    [厅长, 很大, 困扰, 解答, 长年, 在外, 打拼, 买房, 父母, 年纪, 打算, 定...
1191    [中午, 点, 搭乘, 河西, 开往, 河东, K市, K2区, 路, 公交车, 楚, M1...
1370    [L3, 县荔溪, 乡底, 坪, 村, 胡毛洞, 一组, 搬迁, 至底, 坪, 村组, 米长...
1549    [叶, 书记, 建二, 公司, 职工, 公司, 修公, 租房, 任意, 缩小, 面积, 面积...
1456    [请问, M, 市委书记, M, 市至, C, 市, 城际, 铁路, 开建, 2008, 提...
806     [西地省, 杨耀, 新能源, 科技, 有限公司, 法人代表, 胡耀中, 现, J, 市, J...
759     [G, 市, 晚报, 刊登, G, 市, 保障性, 安居工程, 领导小组, G, 市, 房地...
128     [车站, 路, 环线, 之间, 铁路沿线, 周边环境, 差, 城中村, 外地, 旅客, 市,...
570     [D3, 区, 五一路, 快乐, 购, 文体, 百货, 部, 天天, 摊位, 摆在, 人行道...
```

图（12）　　添词典去停用词

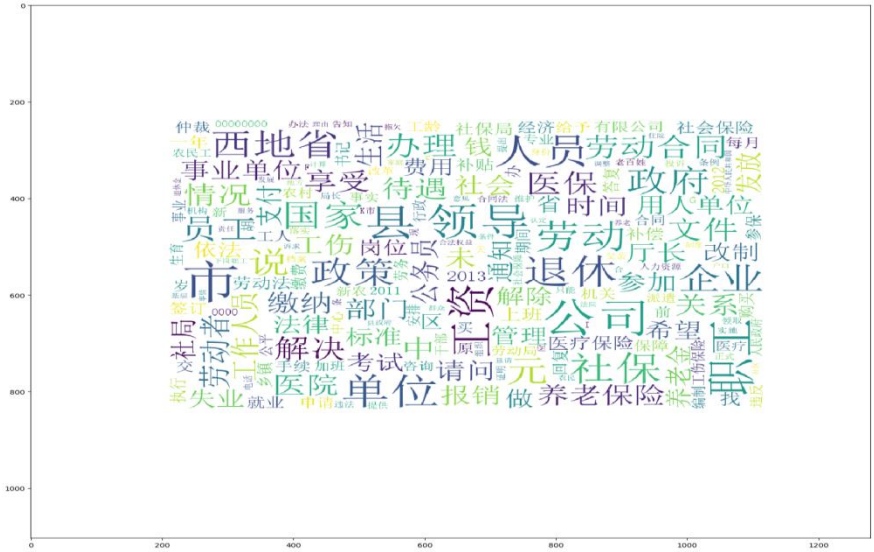
从该图中可以看出数据得到了进一步的合理清理，删掉了文中的一些停用词，并将特有名词放在一起，使词义合理。

### 5.1.3 绘制词云

利用 python 进行每一类别的词云绘制，图形如下：


图（13） 城乡建设

由图（13）可见，一级标签中“城乡建设”类别中出现的高频词汇主要有：业主、市、部门、开发商、县、房屋、居民、公积金、物业等。




图（14） 劳动和社会保障

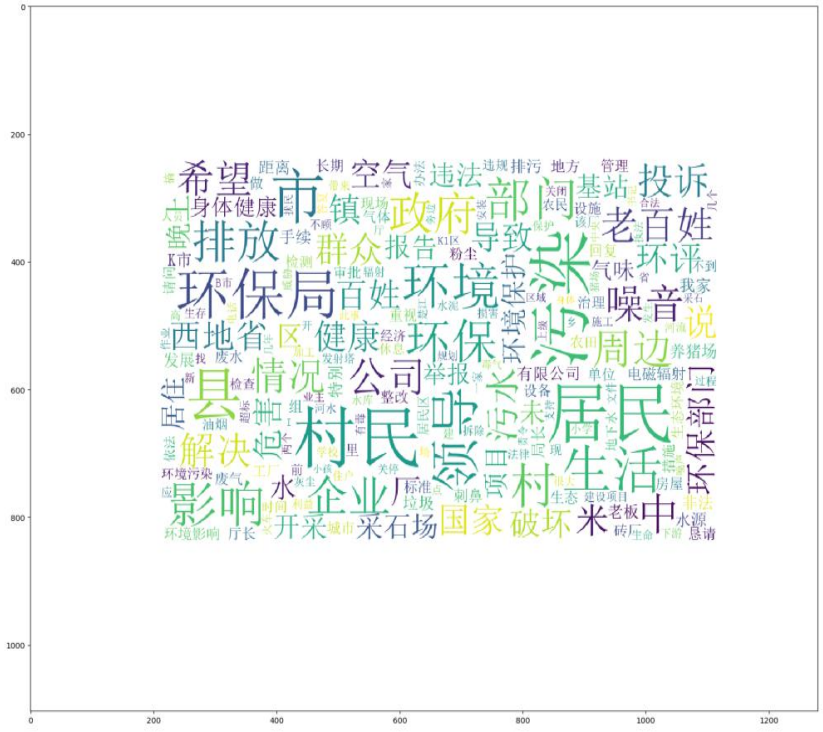
由图（14）可见，一级标签中“劳动和社会保障”类别中出现的高频词汇主要有：市、工资、单位、领导、公司、社保、职工、西地省、劳动、国家、医保、人员、劳动合同、退休等。


图（15）　　教育文体

由图（15）可见，一级标签中“教育文体”类别中出现的高频词汇主要有：学校、学生、孩子、教师、教育、老师、家长、领导、培训、补课、政策等。


图（16）　　商贸旅游

由图（16）可见，一级标签中“商贸旅游”类别中出现的高频词汇主要有：电梯、市场、部门、旅游、公司、景区、价格、元、老百姓、开发商、传销、领导、发展、收费等。




图（17）　　环境保护

由图（17）可见，一级标签中“环境保护”类别中出现的高频词汇主要有：村民、居民、环保局、环境、污染、噪音、环保、采石场、破坏、希望、政府、排放、空气等。


图（18）　　卫生计生

由图（18）可见，一级标签中“卫生计生”类别中出现的高频词汇主要有：医院、医生、政策、患者、家庭、检查、孩子、治疗、医疗、生育、医师、情况、卫生局、计划生育、卫生院等。


图（19）　　交通运输

由图（19）可见，一级标签中“交通运输”类别中出现的高频词汇主要有：出

租车、快递、司机、公司、车辆、收费、领导、费用、城市、百姓、公交、客运、运输、管理等。

## 5.2 文本的向量表示结果

利用 python 工具将预处理完成的数据利用 TF-IDF 模型转化成词频向量，最终转化成 TF-IDF 矩阵。其结果如下图：

```
In[2]: D_train
Out[2]:
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（20） 向量化的结果

```
In[3]: D_test
Out[3]:
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（21） 向量化的结果

## 5.3 模型结果及评价

### 5.3.1 高斯朴素贝叶斯模型结果

高斯朴素贝叶斯模型结果及评价如图：

```
In[2]: model=GaussianNB()
In[3]: model.fit(D_train ,一级标签_train)
Out[3]: GaussianNB(priors=None)
In[4]: f1=model.score(D_train ,一级标签_train)
In[5]: f1
Out[5]: 0.9966362325804902
```

图（22） 模型评价图

### 5.3.2 LDA 模型的结果

LDA 模型结果及评价如图：

```
l_model.print_topic(0)
Out[19]: '0.024*"A2区" + 0.015*"A3区" + 0.010*"扰民" + 0.009*"销售" + 0.009*"西地省" + 0.008*"噪音" + 0.008*"车位" + 0.008*"苑" + 0.008*"投诉" + 0.008*"新城"'
l_model.print_topic(1)
Out[20]: '0.010*"A6区" + 0.009*"业主" + 0.008*"中心" + 0.007*"扰民" + 0.007*"投诉" + 0.006*"购房" + 0.006*"物业" + 0.006*"A3区" + 0.006*"医院" + 0.006*"A9市"'
l_model.print_topic(2)
Out[21]: '0.016*"街道" + 0.014*"A4区" + 0.013*"咨询" + 0.012*"A3区" + 0.008*"安置" + 0.008*"A1区" + 0.006*"社区" + 0.006*"幼儿园" + 0.006*"村" + 0.005*"解决"'
l_model.print_topic(3)
Out[22]: '0.015*"A3区" + 0.015*"A5区" + 0.014*"扰民" + 0.011*"魅力" + 0.008*"咨询" + 0.008*"噪音" + 0.007*"西地省" + 0.007*"油烟" + 0.006*"门面" + 0.006*"一楼"'
l_model.print_topic(4)
Out[23]: '0.016*"建议" + 0.011*"路" + 0.009*"A1区" + 0.006*"A3区" + 0.005*"居民" + 0.005*"国际" + 0.005*"县星沙" + 0.005*"商铺" + 0.005*"公交车" + 0.005*"镇"'
```

图（23） LDA 模型结果

```
   ...: l_model.print_topic(0)
Out[19]: '0.024*"A2区" + 0.015*"A3区" + 0.010*"扰民" + 0.009*"销售" + 0.009*"西地省" + 0.008*"噪音" +
In[20]: l_model.print_topic(1)
Out[20]: '0.010*"A6区" + 0.009*"业主" + 0.008*"中心" + 0.007*"扰民" + 0.007*"投诉" + 0.006*"购房" + 0
In[21]: l_model.print_topic(2)
Out[21]: '0.016*"街道" + 0.014*"A4区" + 0.013*"咨询" + 0.012*"A3区" + 0.008*"安置" + 0.008*"A1区" + 0
In[22]: l_model.print_topic(3)
Out[22]: '0.015*"A3区" + 0.015*"A5区" + 0.014*"扰民" + 0.011*"魅力" + 0.008*"咨询" + 0.008*"噪音" + 0
In[23]: l_model.print_topic(4)
Out[23]: '0.016*"建议" + 0.011*"路" + 0.009*"A1区" + 0.006*"A3区" + 0.005*"居民" + 0.005*"国际" + 0.0
```

图（24）

需说明的是上述图中选取的类别是随机 n 是自定义的，将其分成 5 类，根据分析可知第一类主要是扰民和噪音问题；第二类是医院和购房问题；第三类是幼儿园安置问题；第四类是门面及油烟问题；第五类是商铺及居民出行问题。

### 5.3.3 答复的完整性结果

```
C:\Anaconda3\python.exe D:/Python作业/test.py
答复意见
0     18
1     83
dtype: int64
城乡建设回复率 0.6916666666666667

Process finished with exit code 0
```

图（25）

需要注意的是关于完整性是按照类别来进行表示的，该图中是对城乡建设的完整度进行计算从而得到成型建设的回复率。

### 5.3.4 答复的相关性及时间结果

```
C:\Anaconda3\python.exe D:/Python作业/句子相关性计算.py
请输入要判断的回复序号：8
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\Cyj\AppData\Local\Temp\jieba.cache
Loading model cost 1.555 seconds.
Prefix dict has been built successfully.
留言与回复的相关性为： 0.030927835051546393
留言时间与回复时间间隔了：28天12小时

Process finished with exit code 0
```

图（26）

需要注意的是序号是要进行自定义输入的，由结果截图来看是对序号为 8 的回复进行的判断，得到其相关性数值和时间间隔分别为上述所示。

# 6. 总结与展望

## 6.1 总结

模型优点：

高斯朴素贝叶斯模型既继承了朴素贝叶斯模型的优点，又减小了分类的误差率，可针对属性相关性较大的数据进行判断，并且与 TF-IDF 对文本进行向量化处理技术相结合便增加了预测和分类效果的准确性；LDA 模型对文本的语义进行分析和关键词语的挖掘；K-means 模型直观且结构形式简明，将文本的关键词进行较为准确清晰的分类。

模型缺点：

高斯朴素贝叶斯模型的相对简单性也给数据处理的效率和准确率带来一定误差；LDA 模型不适合用于对长度过短的文档进行研究，会导致产生的向量长度过于稀疏，并且模型没有考虑到单词间顺序的影响；K-means 模型在数据量很大的情况下不易获得，并且由于模型主要是根据计算点与中心质点的距离来得到关系，因此对不同关系的事物之间的联系状况存在一定影响。

## 6.2 展望

可通过对模型进行一定的优化来提高模型的准确性以此来减少误差。

高斯朴素贝叶斯模型的优化可通过增强模型的训练次数，增强迭代次数来确定超参数来更好地实现。

LDA 模型的优化可将 LDA 模型和循环网络联系起来，以此来解决文本中单词的顺序问题。

K-means 模型的优化尽可能使其收敛的速度得到加快，使用策略性来选择初始质心，尽量拉大距离更好的反应数据的分布。并且对其迭代过程进行一定的优化与改进，改进其复杂度，可通过不断更新类的质心来实现。

# 参考文献


[1] 高斯朴素贝叶斯模型（Guassian Naïve Bayes

https://www.360kuai.com/pc/95273ae7f9e11382c?cota=4&kuai_so=1&tj_url=so_rec&sign=360_da20e874&refer_scene=so_3

[2] 停用词包

https://pan.baidu.com/s/1u-Ob86VGVSk3vhnwf2S29w

[3] 泰迪课堂，《基于文本内容的垃圾短信处理》

https://edu.tipdm.org/classroom/119/courses

[4] 泰迪课堂，《自然语言处理技术》

https://edu.tipdm.org/classroom/119/courses

[5] 泰迪课堂，《深度学习原理及编程实现》

https://edu.tipdm.org/classroom/119/courses

[6] 《基于微信的社会舆论热点挖掘及分析模型研究》

http://www.doc88.com/p-9807408443218.html

[7] python 专业方向 | 文本相似度计算

http://www.jianshu.com/p/edf666d3995f

[8] 《基于新浪热门平台的微博热度评价指标体系实证研究》

https://wenku.baidu.com/view/e737267230b765ce0508763231126edb6f1a7663.html

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

[10] https://blog.csdn.net/taoyanqi8932/article/details/53727841

[11] Jaccard 相似度

https://blog.csdn.net/u012836354/java/article/details/79103099

[12] 相似度计算之 Jaccard 系数

https://blog.csdn.net/qq_34333481/article/details/84024513?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromBaidu-1&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromBaidu-1

[13] 因子分析法 https://baike.so.com/doc/3735268-3924517.html