# 第八届“泰迪杯”
# 全国数据挖掘挑战赛

## “智慧政务”中的文本挖掘


### 摘要

近年来，通过网络问政平台汇聚的各类社情民意不断攀升，累积的大量文本数据给留言划分和热点整理等相关部门的人工工作带来了巨大挑战。在此背景下，本文就将利用大数据和自然语言处理的相关技术，对智慧政务有关的三个问题展开研究。

对于问题 1，首先通过群众编号对群众留言主题表进行去重，得到不重复的留言主题信息。然后利用jieba 中文分词工具对留言主题描述信息进行留言主题一级、二级、三级分类，同时通过 TF-IDF 算法提取每个留言详情描述的前 5 个关键词。其次使用TF-IDF算法获得每个消息中描述的TF-IDF权向量，使用F-Score对TF-IDF权向量进行聚类，获得5个质心。然后分别求出距离各个质心最近的几个留言主题分类信息，结合留言主题信息表的字段，同时根据KNN 算法，为各个类加上留言一级标签。最后分别对各个留言主题标签类型进行统计分析，建立关于 留言详情的标签分类模型。

对于问题 2，在问题一的基础上，首先对 留言详情进行TF-IDF算法的特征提取，同时生成TF-IDF向量，并通过向量对文字进行聚类。然后用 word2vec 词向量模型对文字进行爬取，同时用K-Means聚类，对爬取出来的 留言详情进行聚类。然后对于热点问题，通过对时间、地点、文本关键词聚类来分析，在 留言详情中使用 TextRank 算法，在本文中得将 留言详情进行重复研究。其次调用 TextRank 算法进行计算，得到达到字数要求的句子的排序。对于权重最高的句子，可以认为就是该处热点的主话题。最后通过点赞数和反对数对留言问题进行热点排序，选取前五名，作为政府重点关注的留言问题。

对于问题 3,首先针对附件 4 相关部门对 留言详情给出的答复意见，从留言答复的相关性、完整性、可解释性等角度对政府的答复意见的质量给出一套评价方案，同时用余弦相似性、TF-IDF算法和三元法对留言答复进行评价，最后政府对留言的答复的可能性进行尝试实现。



关键词：去重　中文分词　F-Score 聚类　 KNN 算法　TF-IDF 算法　word2vec 词向量模型　三元法

# Text mining in "intelligent government affairs"

## Abstract

In recent years, all kinds of social conditions and public opinions gathered on the political platform through the Internet have been on the rise, and the accumulated mass of text data has brought great challenges to the manual work of relevant departments such as message division and hot spot sorting.In this context, this paper will make use of big data and natural language processing technology to study three issues related to smart government.

For question 1, first of all, through the mass number to the mass message topic table to carry on the de-duplication, get not repeated message topic information.Then, jieba Chinese word segmentation tool is used to classify the description information of message subject into level 1, level 2 and level 3, and the first five keywords of each message detail description are extracted by tf-idf algorithm.Secondly, tf-idf algorithm is used to obtain the tf-idf weight vector described in each message, and f-score is used to cluster the tf-idf weight vector to obtain 5 centroid.Then, the classification information of several message subjects closest to each center of mass is calculated respectively, combined with the fields in the message subject information table, and according to the KNN algorithm, a message level label is added to each class.Finally, the paper makes a statistical analysis on the tag types of each message topic and establishes a tag classification model for the message content.

For problem 2, on the basis of problem 1, firstly, feature extraction of tf-idf algorithm is carried out for the message content, while tf-idf vector is generated, and the text is

clustered through the vector.Then, word2vec word vector model is used to crawl the text, and k-means clustering is used to cluster the message contents retrieved from the crawl.Then, for hot issues, through the clustering of time, place, text keywords to analyze, the use of TextRank algorithm in the message content, the message content will be repeated in this paper.Next, TextRank algorithm is called to calculate the order of sentences that meet the word length requirement.The sentence with the highest weight can be regarded as the main topic of the hot topic.Finally, thumb up number and objection number are used to sort the hot spots of message questions, and the top five are selected as the key message questions concerned by the government.

For question 3, the first of message content is given in annex 4 related department reply, reply message from the Angle of the relevance, integrity and interpretability of the opinions of the government's response is given a set of quality evaluation scheme, at the same time using cosine similarity, TF - IDF algorithm and three element method to evaluate a message reply, finally the possibility of government responses to the comments to try to achieve.

Key words: de-weighting Chinese word segmentation KNN algorithm tf-idf algorithm word2vec word vector model ternary method

# 目录

# 1. 挖掘目标

本次建模目标是利用集互联网公开来源的群众问政留言记录数据，利用jieba中文分词工具对留言详情进行逐级分类、F-Score评价方法、word2vec模型及 KNN 算法等，达到以下三个目标：

1）采用文本分割和文本聚类的方法对非结构化数据进行文本挖掘。根据聚类结果，结合留言详情、留言主题；对留言主题进行逐级分类。

2）根据留言主题一级、二级、三级分类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给出相应热点问题对应的留言信息。

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。

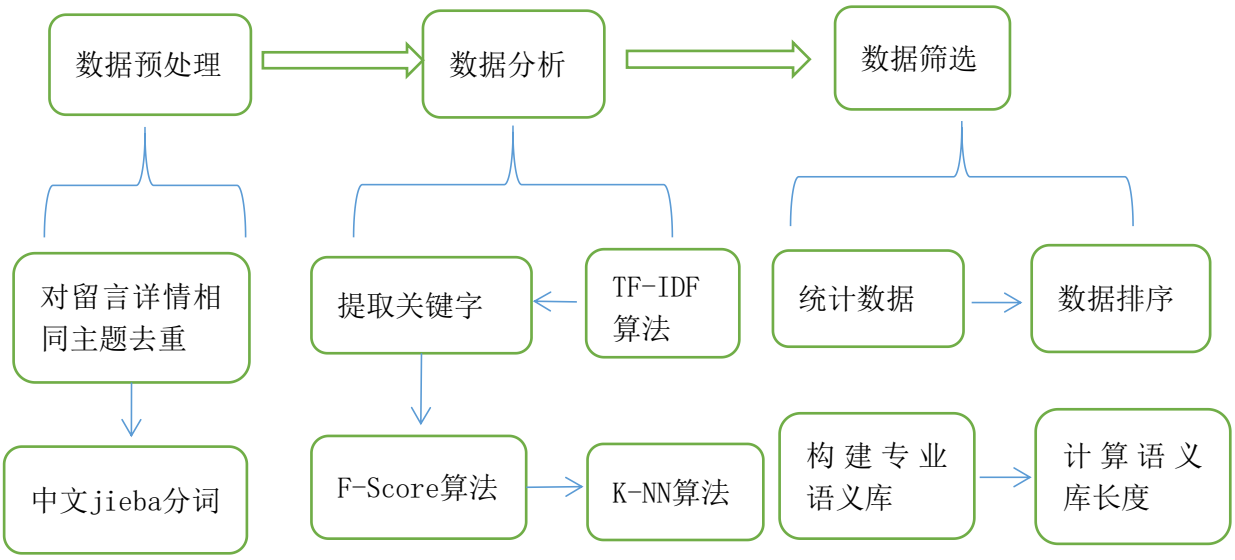
## 1.1. 挖掘思路

针对问题一，通过多分类模型对留言主题进行一级分类，然后用F-Score算法来对分类模型进行评价;针对问题二，首先提取出每条留言详情的时间、地点和事件关键字，用相似度计算各条留言之间的相似度，用量化评价指标来选出前五，作为热点问题；针对问题三，通过TF-IDF算法来评价留言答复的完整性、相关性和可解释性。

通过特殊字处理和正则表达式来进行留言主题和留言详情清洗，清洗后便于关键字提取；用jieba分词对留言主题和内容分词和词性确定；同时用去停用词库来过滤无意义的字词；word2vec算法来进行文本表示；k-means聚类算法来聚类，TextRank自动文摘要对 留言详情进行热点选择和排序；对 留言详情和留言答复之间建立相似度模型，给出评价方案。

# 2. 分析方法与过程

## 总体流程图




总体流程图

本用例主要包括如下步骤：

步骤一：数据预处理，在题目给出的数据中，出现了很多重复的留言数据，在原这里使用 TF-IDF 算法，找出每个留言主题描述的关键字，把留言主题信息转换为权重向量。开始数据上对进行再处理，在此基础上进行中文分词。

步骤二:数据分析，经过消息主题信息分割后，需要将这些词转换成向量进行挖掘分析。采用F-Score算法对留言主题进行分类，利用K-nn算法找出与各中心相似的元素，根据个数多的判定所属类别。

步骤三：数据筛选，相关数据统计，分类，筛选，汇总，找出对应的留言主题的模型分类，对于分类后的关键词需要特征提取，主题指派，分析出相应的分布，然后确定热点问题和热度指数。

步骤四：使用步骤1的结果来构建一个专业的语义库。通过计算到语义库的距离，

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。

# 2.1. 问题1分析方法与过程

## 2.1.1. 流程图

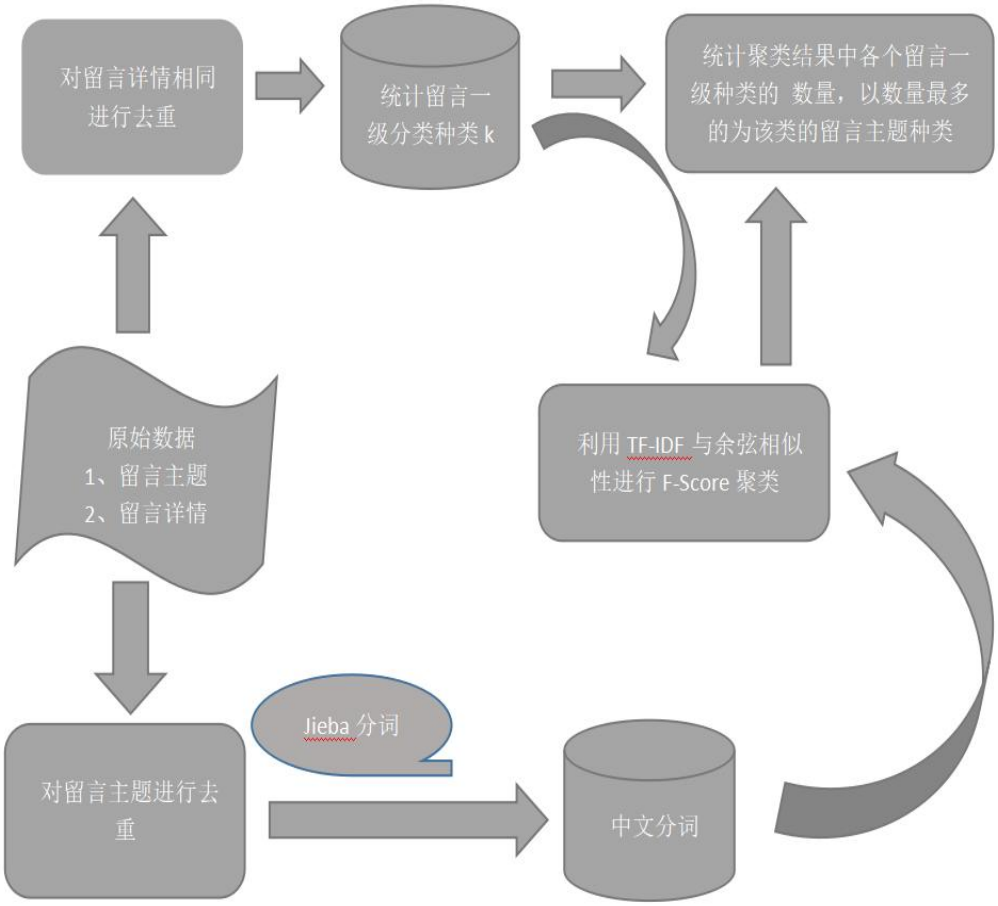



图2：问题一流程图

## 2.1.2. 数据预处理

### 留言主题的去重、去空

在主题给出的数据中，有许多重复的消息数据。考虑到政府回复留言详情时可能每天都会对留言详情进行更新，因此在去重的时候应该取更新时间最晚的记录，去掉历史记录。考虑到python中的dictionary保存数据，键是相同的，值是最后更新的值。因此在读取数据时，按时间升序把留言主题信息内容作为键， 把整个留言详情

信息作为值保存在 value 中。最后，字典的内容可以写入留言主题中。同时会有重复的留言详情，会将干扰分析的问题，采取直接过滤的方法，从文本中删除。

## 对留言主题进行中文分词

在挖掘和分析消息主题之前，必须将非结构化文本信息转换成计算机能够识别的结构化信息。在附件的表格中，数据是中文的。为了便于转换，消息主题信息为中文分词。这里，使用python的中文分词包jieba进行分词。Jieba同义词典采用基于字典实现高效单词前缀映射扫描,生成所有可能的中文句子中的词的情况有向无环图(DAG),同时通过动态编程找到最大概率路径,找出最大的基于词频的分割组合,对未知的词,采用HMM模型基于汉字的能力到一个词,可以更好的实现中文分词效果。同时，采用TF-IDF算法提取每个消息主题描述中的前五个关键字。这里采用了jieba自己的语义库。

## TF-IDF 算法

留言主题信息分割后，需要将这些词转换成向量进行挖掘分析。这里采用TF-IDF算法，将留言主题信息转化为权向量。TF-IDF算法的具体原理如下:
第一步是计算词频，即： TF权(词频)。

词频（TF）=某个词在留言主题中出现次数

考虑到留言主题分为长度和长度，为了便于比较不同的留言主题，“词频”的标准是除以留言主题的总字数或留言主题中出现频率最高的字数：

$$\text{词频（TF）}=\frac{\text{某个词在留言主题中的出现次数}}{\text{留言主题的总词数}}$$

或

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

第二步是计算IDF权值，即逆文档频率。建立一个语料库来模拟语言的使用环境是很有必要的。IDF越大，该特征在文本中的分布越集中，说明分割对留言主题属性的区分能力越强。

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

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

TF-IDF=词频（TF）×逆文档频率（IDF）

实际分析表明，F-Score值与留言主题表的留言主题关键字出现的次数成正比，留言主题关键字的重要性越高，TF-IDF值越大。计算留言主题每个关键字的TF-IDF值，排序，最多的时候是提取留言主题表中的留言主题关键词。

## 生成 TF-IDF向量

生成TF-IDF向量的具体步骤如下:
(1)使用TF-IDF算法查找每个留言主题的前五个关键词;
(2)将从每个留言主题中提取的五个关键词组合成一个集合，并针对每个留言主题细节计算集合中单词的词频。
(3)为每个留言主题级别描述生成TF-IDF权向量，计算公式如下:

TF-IDF=词频（TF）×逆文档频率（IDF）

## 2.1.3. K-NN 最邻近分类算法

根据F-Score对聚类中心进行分类，使用k-nn算法找出与每个中心相似的元素，根据元素个数确定类别。根据向量空间模型，对每个类别的留言主题进行训练，得到每个类别的中心向量记为$C_j(W_{1,}W_{2,}\cdots\cdots,W_n)$，将待分类留言主题T表示成n维向量的形式T（$W_{1,}W_{2,}\cdots\cdots W_n$）。对于一个测试文留言主题，计算它与训练样本集中每个留言主题的相似度，找出K个最相似的留言主题，根据加权距离和类别判断测试留言主题的类别。具体算法步骤如下：

（1）对于各个测试留言主题，根据特征字形成测试的留言主题词向量。

（2）计算该测试留言主题与训练集中每个留言主题的留言主题相似度，计算公式为：

$$\mathrm{Sim}\left(d_{i,}d_j\right)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{\sum_{k=1}^{M}W_{ik}^2}\sqrt{\sum_{k=1}^{M}W_{jk}^2}}$$

式中，$d_i$为测试文本的特征向量，$d_j$为 $j$类的中心向量；$M$为特征向量维数：$W_k$为向量的第$k$维。k值的确定一般采用初值，然后根据k的实验测试结果调整k值。

(3)根据留言主题相似度，从训练留言主题集中选择与测试留言主题最相似的k个文本。

(4)在测试留言主题的k近邻中，各类的权重按下式计算:

$$P\left(X,C_j\right)=\begin{cases}1,若\sum\limits_{d\subset Knn} Sim\left(x,d_i\right)y\left(d_i,C_j\right)-b\geq 0\\ 0,其他\end{cases}$$

式中，$x$为测试留言主题的特征向量；$Sim\left(x,d_i\right)$为相似度计算公式；$b$为阙值，有待于优化选择；而$y\left(d_i,C_j\right)$的值为1或0，如果$d_i$属于$C_j$,则函数值为1，否则为0。

（5）比较各类的权重，将留言主题分到权重最大的那个类别中来判断。

## 2.1.4. 分类方法评价

## F-Score方法如下：

生成留言详情的 TF-IDF 权重向量后，根据每个留言主题的 TF-IDF 权重向量，对留言主题进行分类。这里采用F-Score算法把职业类型分成 6 类，分为农村农业、民政、科技与信息产业、经济管理、纪检监督、环境保护、国土资源、党务政治、劳动和社会保障、城乡建设、教育文体、卫生计生、交通运输、商贸旅游、政法十五大类，

F-Score聚类的原理如下：

在深度学习中，精确率(Precision)和召回率(Recall)是常用的评价模型性能的指标，从公式中可以看出，两者没有很大的关系，但是在实践中，它们是相互制约的。我们都希望模型的准确性和召回率高，但是当召回率高的时候，召回率往往较低;当召回率高的时候，召回率往往较低。往往需要对模型的精确率和召回率做出取舍：

例如在一般的检索任务中，在保证查全率的同时，尽可能提高查全率的准确率;在癌症检测、财务造假任务中，在保证准确率的同时，尽可能提高召回率。

很多时候，我们需要综合权衡这两个指标，从而得出一个新的指标，F-Score，这是一个考虑精度和召回率的调和值。

$$\text{F-Score}=(1+\beta^2)=\frac{Precision\cdot \text{Recall}}{\beta^2\cdot Precision+\text{Re}call}$$

其中,B作为参数表示的是准确率和召回率的相对权重，B大于1时,表示准确率比召回率更加重要B小于1时,表示召回率比准确率更重要;B等于1时,表示准确率和召回率重要程度相同。当分数β等于1时，就变成了F1-Score，其中召回率和准确率都很重要，权重是一样的。当我们认为精度更重要时，我们将空白调整为小于1，如果我们认为召回更重要，我们将β值调整为大于1。本文中取β=0.98。

## 2.1.5. 分析留言主题类型和初步定义留言分类

对附件3根据F-Score聚类方法和K-nn最邻近分类得出5个点和每个点周围100个主题类型关键词，根据这些留言详情关键词对照附件所属的类型，包括农村农业、民政、科技与信息产业、经济管理、纪检监督、环境保护、国土资源、党务政治、劳动和社会保障、城乡建设、教育文体、卫生计生、交通运输、商贸旅游、政法十五大类，统计数量最多的即为目前政府最需要解决的问题类型，并定义相关问题所属领域。

## 2.2. 问题2分析方法与流程

### 2.2.1. 数据清洗

将留言详情表中的点赞数和反对数为0的删除，将留言时间年月日用series来分开，形成单独的类，再讲类组合进行排序分类，将 留言详情按照时间分类，达到便于按照时间和内容聚类。

留言详情中需要清除多余的空白符；同时为了清理这些和问题事件无关的词汇，可以利用正则表达式将其匹配，并替换为空字符。

对于留言详情的聚类，多数标点符号也会影响聚类效果，所以需要将留言详情的标点符号全部删除以便剩下的中文、英文和数字能很好的反映留言详情。

### 2.2.2. 数据提取

#### jieba 分词的留言详情分词及重要词汇提取

对于清理后的留言详情，需要对留言详情进行分词操作，在分词前考虑分词工具分词效果可能不是十全十美,分词的结果可能分错误,所以本文通过观察 留言详情分词后的结果，自定义一个用户词词典，将分词工具的未登录词，即无法识别的词汇加入用户词词典，这样在分词前，先让分词工具对用户词词典进行分析，再使用分词工具便不会出现错误分词的现象。分词工具分词的同时，是能给词加上词性的。这个很方便本文提取想要的相应词性的词。因为形容词、副词、助词等对 留言详情的特征贡献不大，所以在内容特征提取的过程中，只提取名词、动词、时间、地点、简称等词性的词语，以便能更好的分辨这封留言。不过，留言详情提取这些词，也不一定能达到目的，为了更好的聚类效果，需要设置了停用词词典,即在分词之后,把在停用词词典中不需要的词在分词后的词中去除；同时，本文还使用消歧词典来将一个多个同一意思的词转化为同一个词；本文也

创建了一个保留单字词典，只保留在词典中的单字，把其他所有的单字都去除，因为单字可能都没有特别大的意义，比如“是”、“会”等等。

## TF-IDF 的 留言详情特征提取

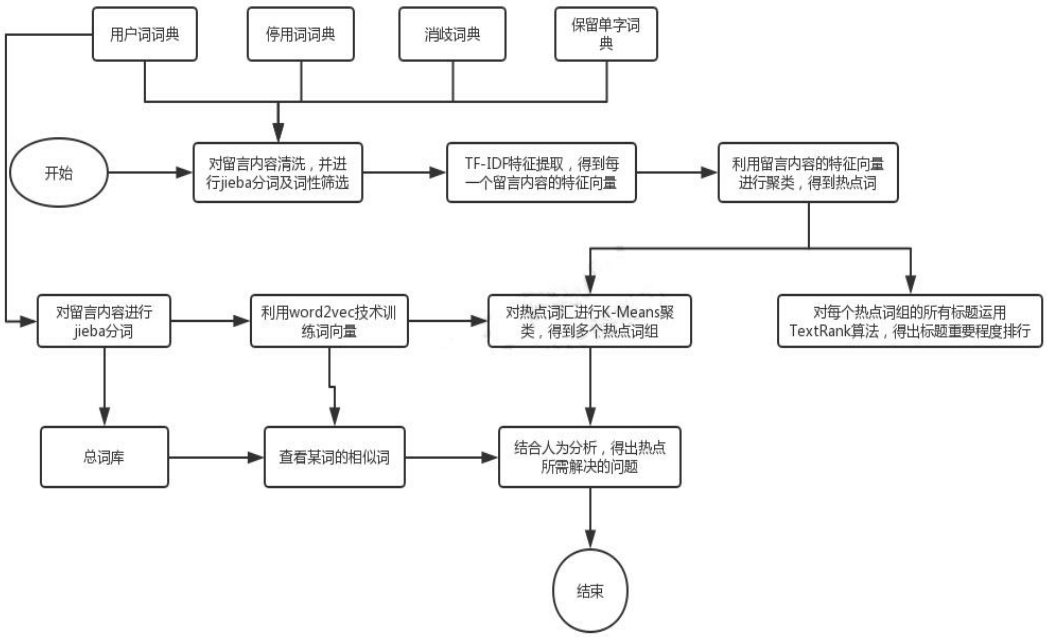
从上一步分词并筛选重要词之后，用剩下的这些字来进行特征提取，能更好反映留言详情的特征。因为留言详情中普通的词很多，但其重要性可能都比较小，本文使用 TF-IDF 来提取特征，这种特征提取能更好的留言详情特点。

通过 TF-IDF 特征提取之后，在所有的留言详情中，每篇留言详情都有一个向量标识。向量上的每一个值都是一个字的 TF-IDF 值。其向量获得方式为首先统计出所有的词，把每个字当成向量的每一个维度，如果该文档中有某词，就在某词的维度上计算它的 TF-IDF 值；如果不存在某词，那么某词的维度上的值就为 0。用这种方式对所有的留言详情进行特征提取，提取的结果是一个稀疏矩阵。

## 2.2.3. 问题2分析及流程图

## 对留言详情聚类

本文主要通过文本挖掘技术进行留言详情热点问题分析，把从附件中的留言主题单独列出来，通过对留言详情的聚类，得到留言热点；再对热点进行分析，通过对与热点相关的词汇进行聚类，我们可以让人民、行业或组织参与到热点中来。主要涵盖的内容如图所示：

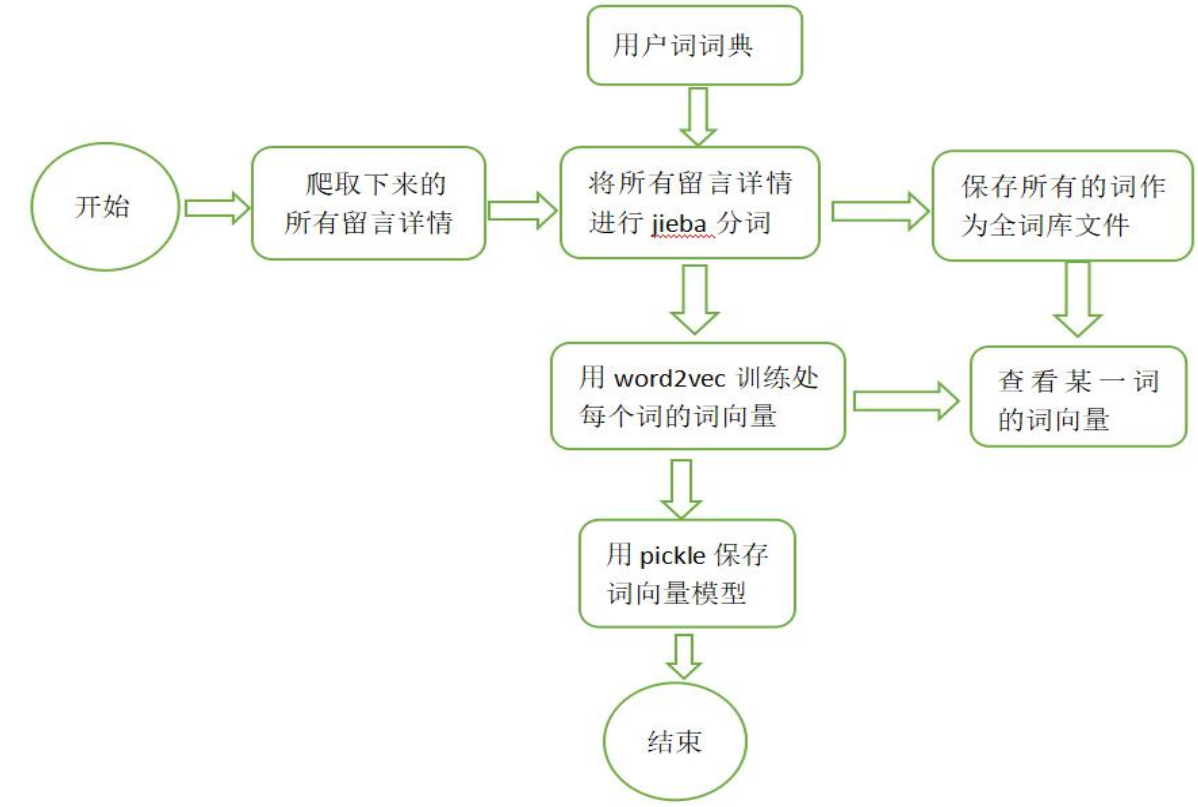


由图所见，问题2主要研究的内容为：
1、去重、时间过滤消息的内容,然后jieba消息内容的内容文本分词和词性标注,并过滤出名词、动词,称为“词性,如分词使用自定义用户词之前提高分割的准确性字典,使用后停止词词典,消除歧义,保留词词典过滤掉与主题无关的词的准确性和聚类精度，建立每个词的信息同义词典，利用 TF-IDF 特征提取之后对留言进行聚类，并对每个类的大小进行排序；
2、针对聚类后的每一类留言详情，为了得到该处热点的话题信息，还需要提取它们5个的标题，利用 TextRank 算法，对留言详情的重要程度进行排序，用重要性最高的留言详情关键字来描述该处热点的话题；
3、对所有的留言详情进行 jieba 分词，并训练出 word2vec 词嵌入模型，然后对聚类后的每一类留言详情，提取它们的内容分词后的结果，运用 word2vec 模型得到每个词的词向量，再利用 k-Means 聚类算法进行相近词聚类。

## 对时间聚类

用df_dt=pd.to_datetime(data_time.留言时间,format="%Y/%m/%d")代码对时间进行分类后聚类。从而达到时间聚类，删除间隔时间较长的 留言详情，留下时间较集中的 留言详情。达到对留言时间的聚类。

## 2.2.4. 训练 word2vec 词向量模型



本文运用了爬取下来的大约 1000 条留言详情数据进行分词，接着进行训练，得到词向量模型，并保存下来。保存模型的运用 pickle 模块，该模块实现了基本的数据序列和反序列化。pickle 的序列化操作常常用来保存程序运行中的对象到文件中，永久存储；通过 pickle 的反序列化模块，又可以在程序中创建之前保存成文件的对象。

## 2.2.5. K-Means聚类

K-mean 聚类的原理如下：

假设留言详情有一个包含$n$个$d$维数据点的数据集$X=\{x_1, x_2, \cdots\cdots x_i, \cdots\cdots x_n\}$，其中

$x \in R_i$，K-means聚类将留言详情数据集$X$组织为$K$个划分C={1,2,3，……，k}

每个划分代表一个类$C_k$，每一类$C_k$有一个类别中心$\mu$，选取欧式距离作为相似

性和距离判断准则，计算该类内个点到聚类中心$\mu$的距离平方和:

$$J(C_k)=\sum_{x_i \in C_k} x_i - \mu_k$$

聚类目标是使各类总的距离平方和$J(C)=\sum_{k=1}^{k} J(C_k)$最小，

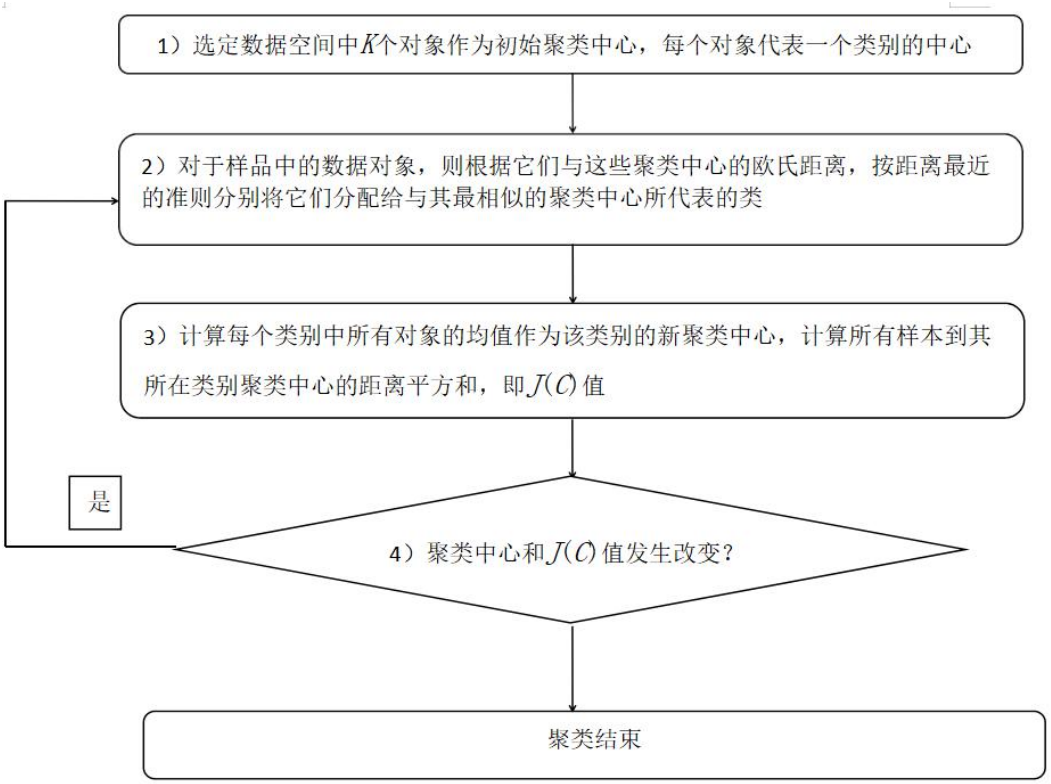
$$J(C)=\sum_{k=1}^{k} J(C_k)=\sum_{k=1}^{k}\sum_{x_i \in C_i} x_i - \mu_i=\sum_{k=1}^{k}\sum_{i=1}^{n} d_{ki} - \mu_i$$

其中，$d_{ki}=\begin{cases}1, \text{若} x_i \in C_i \\ 0, \text{若} x_i \notin C_i\end{cases}$，因此根据最小二乘法和拉格朗日原理，聚类中心$\mu_k$

应该取为类别$C_k$类每个留言详情数据点的均值。

k-means聚类算法步骤如下:

1. 从X中随机选择K个元素作为K个簇的中心。
2. 计算剩余留言详情元素到K个组中心的相位异质性，并将这些元素划分为相位异质性最低的组。
3. 根据聚类结果，对K个聚类的中心进行重新计算，取聚类中所有留言详情元素各维度的算术平均值。
4. 根据新的中心重新聚集X中的所有元素。
5. 重复步骤4，直到集群结果不再发生变化。
6. 输出结果。k-均值聚类算法流程图如下:




## 2.2.6. 热点 留言详情排行

### TextRank 的自动文摘算法

TextRank 是一种基于图的用于文本的排序算法，基本思想来自于 Google 的PageRank 算法。类似于网页的排名，对于词语可得到词语的排行，对于句子也可得到留言详情的排名，所以 TextRank 可以进行关键词提取，也可以进行自动文摘。其用于自动文摘时的思想是：将每个句子看成 PageRank 图中的一个节点，若两个留言详情之间的相似度大于设定的阈值，则认为这两个留言详情之间有相似联系，对应的这两个节点之间便有一条无向有权边，边的权值是相似度，接着利用 PageRank 算法即可得到留言详情的得分，把得分较高的留言详情作为文章的摘要。

TextRank 算法的主要步骤如下：

（1）预处理：分割原文本中的句子得到一个留言详情集合，然后对留言详情进行分词以及去停用词处理，筛选出候选关键词集。

（2）计算留言详情间的相似度：在原论文中采用如下公式进行计算留言详情 1 和留言详情 2 的相似度：

$$\text{留言详情的相似度}=\frac{\text{两个留言详情都出现的词的数目}}{log(\text{留言详情1中的词的数目})+log(\text{留言详情2中的词的数目})}$$

对于两个留言详情之间的相似度大于设定的阈值的两个留言详情节点用边连接起来，设置 其边的权重为两个留言详情的相似度。

（3）计算句子权重：

设阻尼系数为$\lambda$，从留言主题中选取两个句子，留言详情句子1的相似度为$A_i$，

留言详情句子2的相似度为$B_i$，

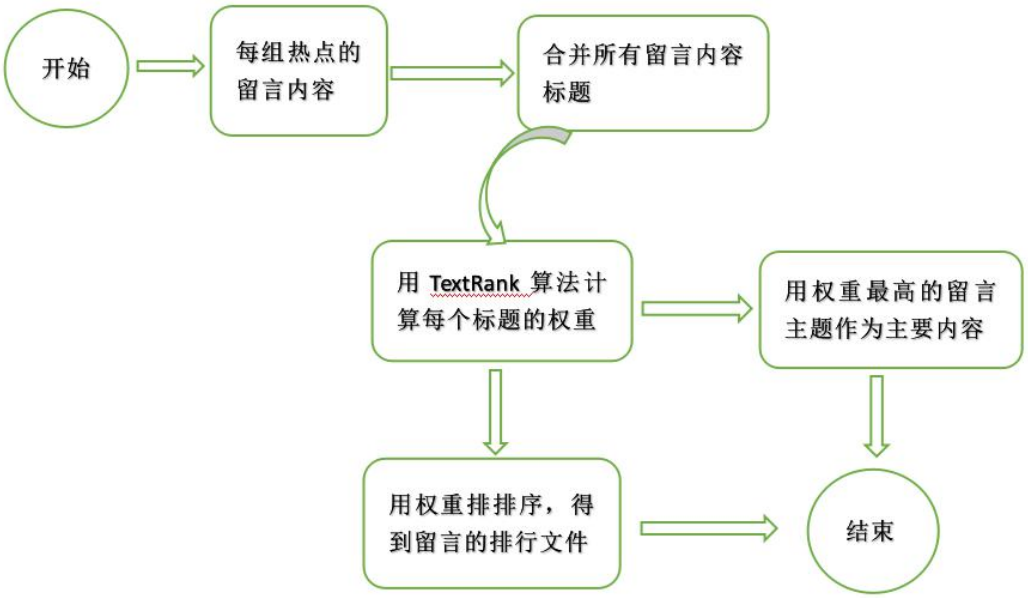
$$句子权重=（1-\lambda)+\lambda\times\sum_{B_i}\frac{(A_i+B_i)\times\vec{B}_i}{\sum_{i=1}(A_i+B_i)}$$

由公式可多次迭代计算直至收敛稳定之后可得各留言详情的权重得分。

（4）形成文摘：将留言详情按照留言详情得分进行倒序排序，抽取得分排序最前的几个留言详情作为候选留言详情热点摘句，再依据字数或句子数量要求筛选出符合条件的留言详情组成文摘。

## 留言详情的热点排行

因为先前已经获得每条留言详情所在的类别，可以按照类别进行数量统计，从而获得各个类别留言详情数量的排行，将数量最多的留言详情所在的组当成最热的热点，第二次之，以此类推。 利用 TextRank 算法进行热点话题进行排序的过程如图所示：



为了能使用 TextRank 算法，在本文中得将 留言详情进行重复研究。然后调用 TextRank 算法进行计算，得到达到字数要求的句子的排序。对于权重最高的句 子，可以认为就是该处热点的热点话题。

## 2.3. 问题3分析方法及过程

针对附件 4 相关部门对留言的答复意见，本问题从答复的相关性、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。

### 2.3.1. 答复完整性

**三元法：**

实体关系抽取是信息抽取的一项重要工作。该任务的输入是多结构文本数据，包括:结构化的infobox信息框、半结构化的表和非结构化的自由文本。结构化的表和非结构化的自由文本。该任务的输出是一个实体关系，它可以表示为三元组(实体1、关系、实体2)。对于结构化和半结构化数据，可以直接分析关系三元组。目前，实体关系提取的研究主要集中在非结构化文本的实体关系提取上。所收集的三元组可用于建立知识库，对问答系统、语义网、机器翻译等建立汉语语义知识库具有重要意义。爬取百度百科全书和互动百科网站数据，提取其结构化部分，输入关系三元实体1、关系词、实体2、gt，构建中文语义知识库。当知识库中给定关系词的频率大于某一阈值时，将关系词视为高频关系词，反之视为低频关系词。在高频关系词的提取方面，将其转化为序列标注问题。高频关系词对应知识库中丰富的关系三元组。这些三元组可以使用评分策略来标记文本中的候选句子，并自动构建训练语料库。利用关键字匹配策略从待提取的条目页面中定位待提取的句子，在训练条件下用随机场模型对提取的部分进行标记，然后根据标签的结果提取关系三元组。本实验比较了候选句的不同选择策略，并在准确性和召回率方面给出了不同的建议。该方法利用领域知识和规则提取低频关系词的实体关系，有效地避免了低频关系词不能自动标记为训练语料库的问题。在提取关系词之前和之后确定实体的类别，并扩展表示关系的关键字数据库。利用实体类的同义词典数据，根据文本中实体对和关键字共现的策略，提取相应的关系三元组。另外，利用关联分析的方法来学习规则，可以挖掘出非常丰富的关联词模板。利用Word2vec训练词向量来判断和提取中文实体关系。利用谷歌开源工具包word2vec，结合百度百科全书的文本数据，对矢量词进行学习，并通过实验评估矢量词的效果。根据词向量，学习要提取的关系词对应的关系矩阵，利用关系矩阵训练分类器，将实体关系提取转化为二分问题。通过三元法从 留言详情与答复两类句子中进行频率分析，三元分析值高的为完整率高，三元分析值低的为完整率低，从而对留言答复进行评价。

### 2.3.2. 答复相关性

余弦相似性原理：

余弦相似性表示两个向量的相似程度，当向量是二维时，可表示为两条线夹角的余弦值，两条线之间的夹角越小，其余弦值越接近1，两个向量越相似，余弦相似性计算公式在n维也成立，假设A和B是两个n维向量，A是［$A_1,A_2\cdots\cdots A_n$］，B是［$B_1,B_2\cdots\cdots B_n$］，则A与B的夹角 θ 的余弦计算公式如图所示：

$$\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

根据上述余弦相似性原理和公式，进行实践判定两个留言详情的相似程度。具体思路是：
1、通过对留言详情进行分词并统计词频，可将留言详情转化为词频向量来表示，入上述公式计算得到两个留言详情的相似性评价值；
2、对留言详情进行分词，分词采用的是Zhparser，PostgreSQL扩展；
2、统计两个留言详情中出现的所有词；
3、列出两个留言详情中所有词；
4、根据两个留言详情中所有词，统计每个留言详情中相应词的词频,生成词频向量；
5、根据两个留言详情的词频向量和余弦相似性计算公式，计算两个留言详情的余弦相似性，判定两个留言详情的相似性；
6、通过对留言详情进行分词并计算词频，将留言详情转化为词频向量来表示，根据余弦相似性原理计算两个向量的相似程度，以此判定两个留言详情的相似程度。

### 2.3.3 答复可解释性

通过答复的完整性和相关性，人为分析该方案能否可行，完整性和相关性都比较高的政府可以采取方案，针对现实情况来进行措施。

## 3.1. 问题1结果分析

### 3.1.1. 聚类中心分类结果

将聚类中心分类结果进行去权分解，提取5个关键字，然后根据F-Score对聚类中心进行分类。利用KNN算法找到离每个聚类中心最近的前5个元素，根据“少

数服从多数”来确定聚类中心的分类。KNN算法的大致步骤如下：
1、算距离：给定聚类中心，计算其与样本中的各TF-IDF权向量的距离；
2、找邻居：圈定距离最近的15个样本，作为聚类中心的近邻；
3、做分类：根据这5个近邻归属的主要类别，来对聚类中心进行分类结合抽样样本，分别找出5个聚类中心的5个近邻样本点所属的留言主题类别。
从KNN分类表可以看出：5个聚类中心可分为：农村农业、民政、科技与信息产业、经济管理、纪检监督、环境保护、国土资源、党务政治、劳动和社会保障、城乡建设、教育文体、卫生计生、交通运输、商贸旅游、政法十五大类。

### 3.1.2. 留言主题领域分类

从留言主题一级分类中分为二级分类，由留言主题聚类关系可得：
例如：从城乡建设中分为安全生产、城市建设和市政管理、城乡规划、村镇建设、工程质量、国有土地上房屋征收与补偿、建筑市场、住房保障与房地产、其他。
从二级分类中结合留言主题，进行三级分类，能够确切知道每条留言反映的确切问题，从而对留言进行更好的分类。

### 3.1.3. 分类方法的评价

用三种分类模型进行选择，三类分类模型为SVC、LinearSVC、SGDClassifier，三者进行选择、叠加，使分类模型的准确率提高。做出的结果如图：

```
classification report
              precision    recall  f1-score   support

           1       0.80      0.88      0.84       396
           2       0.89      0.84      0.86       191
           3       0.89      0.81      0.85       123
           4       0.91      0.88      0.89       316
           5       0.86      0.91      0.89       412
           6       0.85      0.75      0.80       219
           7       0.86      0.84      0.85       185

    accuracy                           0.86      1842
   macro avg       0.87      0.84      0.85      1842
weighted avg       0.86      0.86      0.86      1842

confusion matrix
[[348  12   4   8  10  10   4]
 [ 18 160   0   3   7   3   0]
 [ 14   0 100   0   5   4   0]
 [ 12   1   3 278  17   4   1]
 [ 12   2   0   9 376   0  13]
 [ 21   3   4   7  11 165   8]
 [  8   2   1   2   9   8 155]]
```

## 3.2. 问题2结果分析

留言详情聚类是一个比较大的过程，需要进行数据清理、分词、特征提取、聚类等操作。本问题实现了对留言详情的聚类，在分词的时候使用了 jieba 分词工具，而特征值提取的时候，则通过 TF-IDF 提取留言详情特征，最后使用基于密度的聚类算法进行聚类，并最终把聚类的标签在文件中用 label 列指出。

### 3.2.1. 热点 留言详情排行

总体留言详情热点的排行结果如图所示：

| 问题id | 留言编号 | 留言用户 | 留言主题 | 留言详情 | 点赞数 | 反对数 |
|---|---|---|---|---|---|---|
| 1 | 214282 | A909209 | 市丽发新城小区附近搅拌站噪音扰民和污染环境 | 天天吵天天吵，烦死了不仅吵还臭！说好 | 0 | 0 |
| 1 | 188801 | A909180 | 投诉滨河苑针对广铁职工购房的霸王规定 | 的问题一个接着一个，首先未取得预售资 | 0 | 0 |
| 1 | 188809 | A909139 | A市万家丽南路丽发新城居民区附近搅拌站扰民 | 在小区旁50米处建搅拌站，运渣车吵得人 | 0 | 1 |
| 1 | 189950 | A909204 | 投诉A2区丽发新城附近建搅拌站噪音扰民 | 到百米的地方建搅拌站。可想而知，一个 | 0 | 0 |
| 1 | 190108 | A909240 | 丽发新城小区旁边建搅拌站 | 扬尘严重影响几千名学生的健康，很多业 | 0 | 1 |
| 1 | 190523 | A00072847 | 市丽发新城违建搅拌站，彻夜施工扰民污染环境 | 音污染严重；3、搅拌站几百米外就是小 | 0 | 0 |
| 1 | 190802 | A00072636 | A市丽发小区建搅拌站，噪音污染严重 | 带来了巨大的粉尘，严重影响居民健康； | 0 | 0 |
| 1 | 203393 | A00053065 | 新城小区侧面建设混泥土搅拌站，粉尘和噪音污 | 带来了巨大的粉尘，严重影响居民健康； | 0 | 2 |
| 1 | 208285 | A909205 | 投诉小区附近搅拌站噪音扰民 | ，关闭门窗还是有很大噪音，吵得不能入 | 0 | 24 |
| 1 | 208714 | A00042015 | 区丽发新城附近修建搅拌站，污染环境，影响生 | 内空气质量和声环境质量急剧下降，我们 | 0 | 4 |
| 1 | 213464 | A909233 | 投诉丽发新城小区附近违建搅拌站噪音扰民 | 违建大型搅拌站。该搅拌站的设备太吵了 | 0 | 0 |
| 1 | 213930 | A909218 | 区丽发新城附近违规乱建混凝土搅拌站谁来监管 | 小区居民强烈呼吁政府和有关职能部门， | 0 | 0 |
| 1 | 214282 | A909209 | 市丽发新城小区附近搅拌站噪音扰民和污染环境 | 天天吵天天吵，烦死了不仅吵还臭！说好 | 0 | 0 |
| 1 | 216824 | A909214 | 站大量加工砂石料噪音污水影响丽发新城小区 | 查了解，这些严重扰民的噪音是从位于看 | 0 | 0 |
| 1 | 217700 | A909239 | 丽发新城小区旁的搅拌站严重影响生活 | 绿心范围内搬迁到丽发新城小区旁边不到 | 0 | 1 |
| 1 | 231136 | A909204 | 投诉A2区丽发新城附近建搅拌站噪音扰民 | 建搅拌站。距离上次投诉已经过去一个月 | 0 | 0 |
| 1 | 235362 | A909215 | 街道丽发新城小区附近水泥搅拌站非法经营何时 | 的扬尘肆虐，严重危害居民身体健康！我 | 0 | 0 |
| 1 | 238212 | A909203 | 丽发新城小区附近建搅拌站合理吗？ | ！小区作为居民区应是一个安静的好环 | 0 | 0 |
| 1 | 239336 | A909213 | A市A2区丽发新城小区遭搅拌站严重污染 | ，该厂成日运作，离居民区非常近！附近 | 0 | 0 |
| 1 | 239648 | A909211 | 市A2区丽发新城小区附近搅拌站明目张胆污染环 | 还有灰尘颗粒！都不敢开窗透气了，赶 | 0 | 0 |
| 1 | 243692 | A909201 | 丽发新城小区附近的搅拌站噪音严重扰民 | 并且搅拌站的灰尘极大，都飘到小区里来 | 0 | 2 |
| 1 | 244335 | A909135 | 暮云街道丽发新城社区搅拌站灰尘，噪音污染严 | 商把特大型搅拌站，水泥厂从绿心范围 | 0 | 0 |

具体见表2；

由图可见，每条留言详情都通过类别大小进行排序，在留言详情列表中，通过增加一列“问题id”来代表的热度，热度从1开始由高到低增大。不是热点的留言详情问题id标志为0。 这种方法确实能很好的得到一个热点的留言详情，也能很好的从标题的排序中得出话题对该处热点的贡献。

通过统计方法对总体留言详情进行热点排行；利用 TextRank 算法进行热点留言详情的排序。结合这两种方法，可以发现各处热点并得到各处热点所涉及的内容。

### 3.2.2. 留言详情top5

通过 word2vec 技术训练了爬取下来的所 留言详情的文本内容，从而训练出每个词的词向量，而且越相近的词，词向量之间的距离也越小，距离可以通过字向量之间的余弦相似度、欧式距离等计算得到。随后可以使用训练出来的字向量模型转化留言详情中的词汇，对每一处热点中的留言详情文本的词汇进行聚类。通过这种词汇聚类的方式，可以获得与热点有联系的相关人或事物的集合。整理出前五的热点问题，如图：

| 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 | 0.821 | 2019/11/18至2019/12/25 | A2区丽发新城小区 | 搅拌站噪音污染严重 |
| 2 | 2 | 0.735 | 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 |
| 3 | 3 | 0.728 | 2019/06/19至2019/12/20 | A2区 | 垃圾臭气熏天 |
| 4 | 4 | 0.725 | 2019/02/18至2019/09/21 | A市 | A市交通问题 |
| 5 | 5 | 0.713 | 2019/03/08至2019/09/25 | A市 | 购房补贴问题政策 |

通过热点指数的排序，可以看出某一段时间内某区域的显著问题，便于政府作出相应的对策。

热点排名为1的“在2019年11月18日到2019年2月25日时间段内，A2区丽发新城小区的搅拌站噪音污染严重”政府应对搅拌站扰民问题作出相应的解决措施；

热点排名为2的“在2019年8月8日到2019年9月4日时间段内，A市A5区魅力之城小区的小区临街餐饮店油烟噪音扰民”，政府可以对该餐馆进行实地检查，如若为真，对该餐馆进行惩罚；

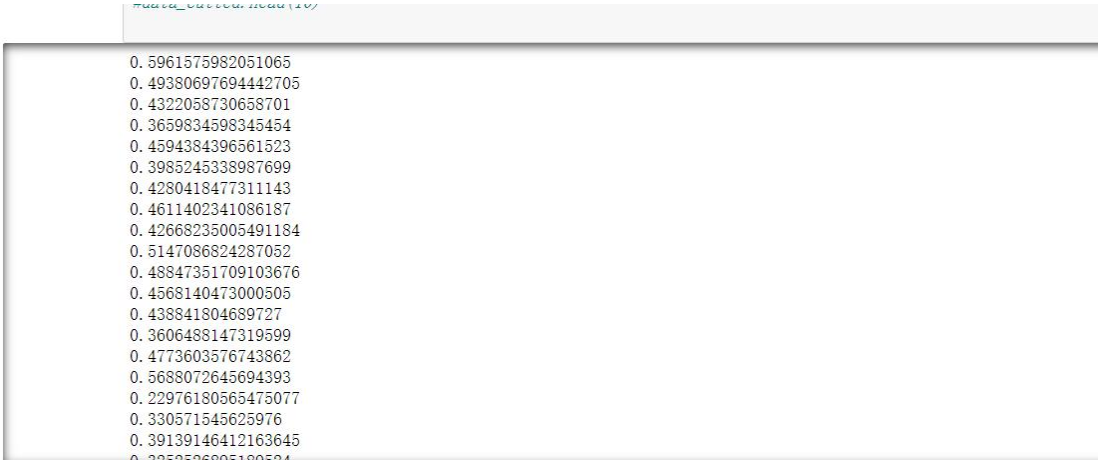
热点排名为3的“在2019年6月19日到2019年9月4日时间段内A2区内垃圾臭气熏天”，政府可以出台相应的环境法规，用奖罚制度规范人们更好的扔垃圾和保护环境；

热点排名为4的“在2019年2月18日到2019年9月21日A市交通问题问题显著”，政府可以出台相应的交通法规，规范交通秩序和派遣交警维持道路秩序等；

热点排名为5的是“在2019年3月8日到2019年9月25日A市购房补贴问题政策的缺陷问题严重”，政府可以注重这个问题，对政策作出修改，以便老百姓更好购房。

## 3.3. 问题3结果分析

通过对留言答复的分析进行评价，我们用余弦相似度模型的概率来确定，概率高则认为该留言回复完整性和相关性较高，概率较低则认为该留言回复的完整性和相关性较低，从而对留言回复进行客观的评价，同时形成一套评价方案，结果如图所示：



# 4. 结论

对政府留言中“智慧任务”进行分析研究，了解社会对政府的需求特点与趋势，对政府有重大意义，同时也是文本分析的一个课题、一个难题。传统的文本解读已经不能满足数据量庞大的政府留言信息。本文采用根据F-Score聚类方法和K-nn最邻近分类,统计目前政府最需要解决的问题类型,深入分析社会的问题，对政府作出改变来更好服务人民。

由分析结果可以看出，政府最需要解决的五个问题，对各个领域，出现的留言问题进行计数，从而定义热门的问题领域，可以看出通过排序得出各大留言情况，并定义热门的留言问题。

对政府给出的留言答复中，我们通过对留言答复的完整性、相关性和可解释性对留言答复进行评价。对政府留言回复作出更好的解释。

# 5. 参考文献


[1] 沈斌.基于分词的中文文本相似度计算研究.天津财经大学.2006

[2] 网络招聘信息的数据挖掘与综合分析-优秀论文 - 道客巴巴

[3] 基于word2vec和TF-IDF算法实现酒店评论的个性化推送-百度

[4] nlp使用算法和平台 - CSDN博客

[5] K-means在关键词聚类中的尝试__CSDN博客

[6] 田长波.融合PAM主题模型的领域历史沿革信息抽取方法.内蒙古师范大学.2016

[7] 在留言等文本信息,焦点问题如何挖掘挖掘? - 知乎

[8] 深度学习F2-Score及其他(F-Score)__CSDN博客

[9] jieba分词过滤停顿词、标点符号及统计词频_知乎

[10] 陈明华.语音合成系统中自动分词技术的研究.哈尔滨理工大学.2009

[11] 陈捷.基意见词的隐性产品特征提取方法研究及应用.东华大学 .2016

[12] 张雯.TextRank算法的改进及在政法全文检索系统中的应用.广西大学.2015

[13] 成松松.基于平均词频的文本特征提取方法.《计算机应用与软件》 .2016

[14] 张琦.网站数据完整性检测工具的设计与实现.中国海洋大学.2014

[15] 张慷.《一种基于文本先分类再聚类的互联网热点信息发现方法》.《兰州工业高等专科学校学报》.2013

[16] Python中的Pickle模块实现了基本的数据序列与反序列化-百度

[17] 张龙凯文本摘要问题中的句子抽取方法研究.《中文信息学报》. 2012

[18] python NLP总结_Python_zwwh

[19] 张旭.民意调查在政府施政中的应用及其制度化研究.湖南大学.2016

sxq的博客-CSDN博客

[20] 朱江基于金融本体库的热点分析研究.北京工商大学.2012

[21] TF-IDF与余弦相似性的应用(二)-百度

[22] 文本分析-词频与余弦相似度_技术博客-CSDN博客

[23] 张爽.数学专业英语辅助写作系统.吉林大学硕士.2017 张爽
[24] 王子慕.一种利用TF-IDF方法结合词汇语义信息的文本相似度量方法研究.吉林大学 2015