基于文本挖掘的“智慧政务”留言处理应用

摘要

近年来，随着人民生活水平的提高，人民可以通过不同方式（如微信、阳光热线等)网络问政平台提出自己的疑问、建议。这些方式逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，因此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。

对于问题一，利用 jieba 库对第三方中文文本分词，使用 Python 中的统计中文文档 n-gram 的出现次数，调用 scikit-learn 中朴素贝叶斯类库建立模型。将数据分成两部分，利用测试集检验模型是否正确。

对于问题二，利用层次凝聚聚类法，将附件 3 中的留言主题进行文本聚类，将聚类结果进行筛选，选择类型最多的留言主题，给出排名前 5 的热点问题。再通过 EXCLE 中的筛选，将留言详细选择出来，给出相应热点问题对应的留言信息。

对于问题三，首先对数据进行文字预处理，通过 tf-idf 算法分别得到群众留言与相关部门答复两个部分的特征值以及对应权重。将两组的权重进行相关性分析；同时对相关部门答复的关键词进行 K-medoids 聚类分析比较，以及对关键词制作词云做到关键词可视化，最终得到评价方案。

关键词：tf-idf 算法　K-medoids 聚类　层次凝聚聚类法　朴素贝叶斯模型

Message processing application of "intelligent government" based on text mining

Abstract:In recent years, with the improvement of people's living standards, people can put forward their own questions and suggestions through different channels (such as Wechat, sunshine hotline, etc.) . These methods have gradually become an important channel for the government to understand public opinions, gather people's wisdom and gather people's opinions. Therefore, the establishment of an intelligent government system based on natural language processing technology has become a new trend in the innovative development of social governance, it can promote the management level and administration efficiency of the government greatly.

For the first question, using the Jieba Library to segment the third-party Chinese text, using the statistics of the number of Chinese document n-gram in Python, calling the naive Bayes class library in scikit-learn to build the model. The data is divided into two parts and the test set is used to verify the correctness of the model.

For the second question, using hierarchical clustering method, we cluster the message topics in Annex 3, select the message topics with the most types, and give the top 5 hot topics. Through the filter in Excle, the message is selected out in detail, giving the corresponding hot topic corresponding message.

For the third question, first of all, the text preprocessing of the data, through the tf-idf Algorithm to get the crowd message and the response of the relevant departments of the two parts of the eigenvalue and the corresponding weight. The weight of the two groups was analyzed, and the key words of the relevant departments were compared by K-medoids cluster analysis, and the key words cloud was visualized.

Keywords: tf-idf K-medoids clustering, hierarchical clustering, Bayesian cognitive science

目录

1 符号说明

序号	符号	含义
1	|D|	留言主题中的文本总数
2	$n_{i,j}$:	某个词在文本中出现的次数
3	$\sum_k n_{k,j}$	文本的总次数
4	$\|\{ j:t_i \in d_j\}\|$	包含该词的文本数
5	$P(X_j = x_{jl}\|Y = C_K)$	第 k 个类别的第 j 维特征的第 l 个取值条件概率
6	m_k	训练集中输出为第 k 类的样本个数
7	λ	为一个大于 0 的常数，常常取为 1，即拉普拉斯平滑。
8	dist	簇的邻近度
9	（C_i, C_j）	第 i 个簇与第 j 个簇的距离

2 挖掘目的

本次目标是利用来自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见数据，利用 jieba 中文分词工具对留言、回复进行分词、层次凝聚聚类、K-medoids 聚类的方法以及 tf-idf 算法，达到以下三个目标：

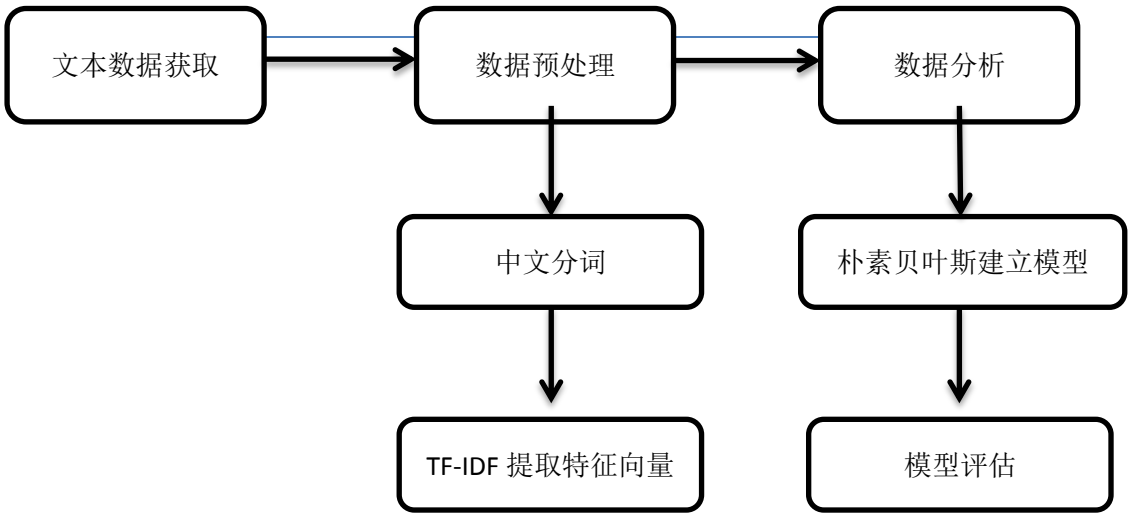
（1）根据提供群众留言数据，建立模型方便对群众留言进行分类，以此提高工作效率。

（2）根据提供的群众留言数据，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，根据问题的反映次数归为热点问题，求出排名前 5 的热点问题。

（3）根据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

3 问题 1 分析方法与过程

3.1　问题 1 流程图



图：问题 1 流程图

3.2 文本数据获取

读取文本数据后对数据进行处理需要调用 python 的第三方库来实现。调用 python 的 os 库，它是一个 Python 的系统编程的操作模块，可以处理文件和目录这些我们日常手动需要做的操作；调用 jieba 库对第三方中文文本分词；调用 sklearn.feature_extraction.text 库统计中文文档 n-gram 的出现次数；调用 scikit-learn 中朴素贝叶斯类库建立模型；其次调用 sklearn.externals 模块保存训练模型，并在测试集进行使用；最后调用 time 库获取系统时间并格式化输出功能。

3.3 对数据预处理

3.3.1 对留言主题进行中文分词

在对“智慧政务”中的文本挖掘分析之前，先要把非结构化的的文本信息转化为计算机可以识别的结构化信息，即对数据进行正则表达，在附件 2 的留言主题描述中，以中文文本的方式给出了数据，为了方便转换，我们先对文本进行中文分词，这里采用 python 中的中文分词包 jieba 进行分词，生成汉字中所有可能成词的情况，但解析出来的文本中有很多无效的词，比如“的”,“和”及一些

标点符号等，这些我们不想在文本分析的时候引用，因此再对文本引入常用的停用词，得到有效的分词再将其做成词云效果如下图所示：

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

对文本分词后可以由词云效果图看出一级标签的关键词，其中，城乡建设主要反映了关于，小区，建议等问题；环境保护反映了污染，严重，公司，排放等问题；交通运输反映出租车，快递等；教育文体反映教师，中学，学校等；劳动和社会保障反映职工等；商贸旅游反映传销，收费等；卫生生计放映医院，咨询，西地省等问题。通过这些文本分词的得到的关键词可以有效，快捷的将留言分派至相应的职能部门处理。

3.3.2 TF-IDF 提取特征向量

（1）TF-IDF 原理及步骤

对留言主题分词后,需要把这些词语转换为向量来对数据进行分析使用，使用 TF-IDF 算法，对分词进行 TF-IDF 权重向量提取特征向量。TF-IDF 算法的基本原理及步骤如下：

第一步：计算词频，即 TF 权重

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

第二步：计算 IDF 权重

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

如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用

$$1 + |\{j : t_i \in d_j\}|$$

第三步：计算 TF-IDF 值

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$$

实际分析得出 TF-IDF 值与一个词在留言主题描述表中文本出现的次数成正比，某个词文本的重要程度越高，TF-IDF 的值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的文本的关键词。

（2）提取特征向量权重

部分特征向量权重如下图所示：

```
(0, 271)	0.006989077617241079
(0, 325)	0.006989077617241079
(0, 330)	0.006989077617241079
(0, 321)	0.006989077617241079
(0, 121)	0.006989077617241079
(0, 300)	0.07687985378965187
(0, 299)	0.055912620937928634
(0, 301)	0.055912620937928634
```

通过调用 sklearn.feature_extraction.text 包，对已分词文本进行向量处理，对每一类文本进行进行特征向量的提取，统计出每一条留言主题的权重,有效的表示一个特征词在文本中区分文本属性的重要程度。

3.4 数据分析

3.4.1 建立朴素贝叶斯模型

（1）模型的分析

MultinomialNB 假设特征的先验概率为多项式分布，即如下式:

$$P(X_j = x_{jl}|Y = C_K)=\frac{x_{jl}+\lambda}{m_k+n\lambda}$$

MultinomialNB 参数比 GaussianNB 多，但是一共也只有仅仅 3 个。其中，参数 alpha 即为上面的常数 λ，如果你没有特别的需要，用默认的 1 即可。如果发现拟合的不好，需要调优时，可以选择稍大于 1 或者稍小于 1 的数。布尔参数 fit_prior 表示是否要考虑先验概率，如果是 false,则所有的样本类别输出都有相同的类别先验概率。否则可以自己用第三个参数 class_prior 输入先验概率，或者不输入第三个参数 class_prior 让 MultinomialNB 自己从训练集样本来计算先验概率，此时的先验概率为 $P(Y=c_k)=m_k/m$。总结如下：

fit_prior	class_prior	最终先验概率
false	填或者不填没有意义	$P(Y=C_k)=1/k$
true	不填	$P(Y=C_k)=m_k/m$
true	填	$P(Y=C_k)$=class_prior

在使用 MultinomialNB 的 fit 方法或者 partial_fit 方法拟合数据后，我们可以进行预测。

（2）模型的建立

在使用 MultinomialNB 的 fit 方法或者 partial_fit 方法拟合数据后，我们可以进行预测。预测结果如图所示。我们知道由贝叶斯模型可以发现模型的正确分类率并不高，只有卫生，教育，环境正确分类，其他文本分类并不准确。为了提高模型的准确率，还应对模型进行优化处理。

['劳动']
交通测试.txt
['环境']
劳动测试.txt
['卫生']
卫生测试.txt
['教育']
商贸测试.txt
['卫生']
城乡测试.txt
['教育']
教育测试.txt
['环境']
环境测试.txt

图：贝叶斯模型预测结果

3.4.2 模型评估

F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

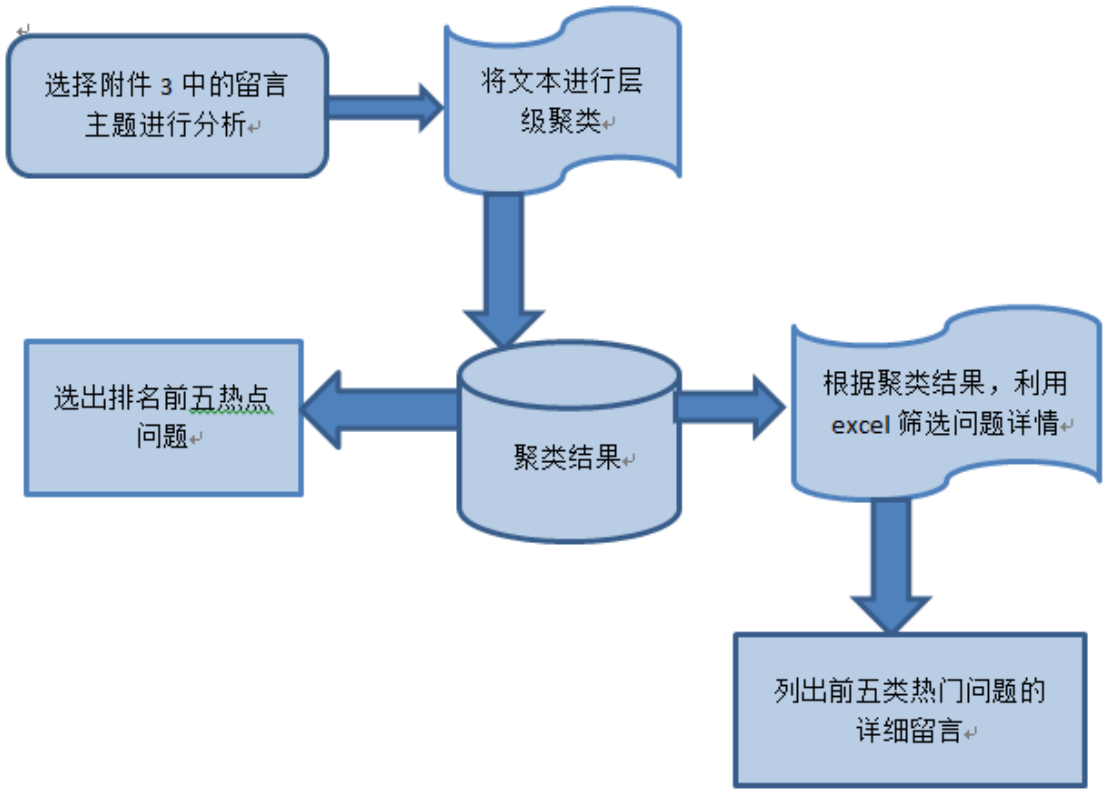
在做文本分类的时候遇到了一些问题：在根据公司业务范围的描述对公司进行行业分类时最终的分类准确率不高，利用各种算法如 KNN、SVM 或 Boosting 等都无法获取很好的效果，正确分类率只有 60%左右。总结的原因可能是生成的 VSM 过于稀疏和文本描述的信息过于宽泛，特征性不强，于是又加入了 lDA 主体模型来做，但效果仍不明显。

4 问题 2 分析方法与过程

4.1 问题 2 分析

问题二需要找出前五类热点问题，则找出文本出现最多的部分。我们采用文本聚类中的层次凝聚聚类法，将留言主题通过聚类找到相似文本，得到大类。在通过 excel 对得到的大类进行筛选，总结得到排名前五的热点问题。最后将留言详细选择出来，给出相应热点问题对应的留言详情信息表。

4.2 问题 2 流程图



图：问题 2 流程图

4.3 数据筛选

在本题中，我们首先筛选附件 3 中的留言主题进行文本的聚类，主题中起到总结留言详情的作用（表达出留言者的主要问题），可以提高效率。

4.4 层次凝聚聚类法

层次聚类的合并算法通过计算两类数据点间的相似性，对所有数据点中最为相似的两个数据点进行组合，并反复迭代这一过程。这个算法具有，可以通过设置不同的相关参数值，得到不同粒度上的多层次聚类结构；不需要事先设定好类的个数，常常可以用作其他聚类算法的前期探索的优点。

层次聚类的合并算法是通过计算每一个类别的数据点与所有数据点之间的距离来确定它们之间的相似性，距离越小，相似度越高。并将距离最近的两个数据点或类别进行组合，生成聚类树。而当两个数据点进行组合成后，用一个新的代表点，替换原有的两个数据点，再进行下次组合。层次凝聚聚类法具体算法步骤如下：

（1）将训练样本集中的每个数据点都当做一个聚类；

（2)计算每两个聚类之间的距离，将距离最近的或最相似的两个聚类进行合并；

（3）重复上述步骤，直到得到的当前聚类数是合并前聚类数的 10%，即 90%的聚类都被合并了；当然还可以设置其他终止条件，这样设置是为了防止过度合并。（4）取距离阈值 T，当 D(n)的最小分量超过给定值 T 时，算法停止，所得即为聚类结果。或不设阈值 T，一直将全部样本聚成一类为止，输出聚类的分级树。输出聚类的分级树。

其中计算距离的三个公式分别为：

(1) 最小距离（单链(MIN)）:定义簇的邻近度为不同两个簇的两个最近的点之间的距离.

$$d_{\min}(C_i, C_j) = \min \quad dist\,(C_{i,}, C_j)$$

(2) 最大距离（全链(MAX)）:定义簇的邻近度为不同两个簇的两个最远的点之间的距离。

$$\mathrm{d_{max}(C_i, C_j) = max\ \ dist\ (C_i, C_j)}$$

(3) 平均距离（组平均）：定义簇的邻近度为取自两个不同簇的所有点对邻近度的平均值。

$$d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{C_i} \sum_{C_j} dist\ (C_i, C_j)$$

在本题中，我们将数据定为附件 3 中的留言主题，将文本导入 Python 中。将文本数据转化成为列表，利用 CountVectorizer 函数提取列表中关键词出现 min 到 max 次的关键词，将关键词转化成词篇矩阵为下一步的聚类做准备。在聚类的过程中，利用距离 corr（x，y）相关系数，来刻画二维随机变量两个分量间互相关联程度。使用 Ward 聚类预先计算的距离定义链接矩阵（即最小距离公式原理）。最后用 plt 函数画出聚类图，进行下一步结果分析。

4.5 问题 2 结果分析

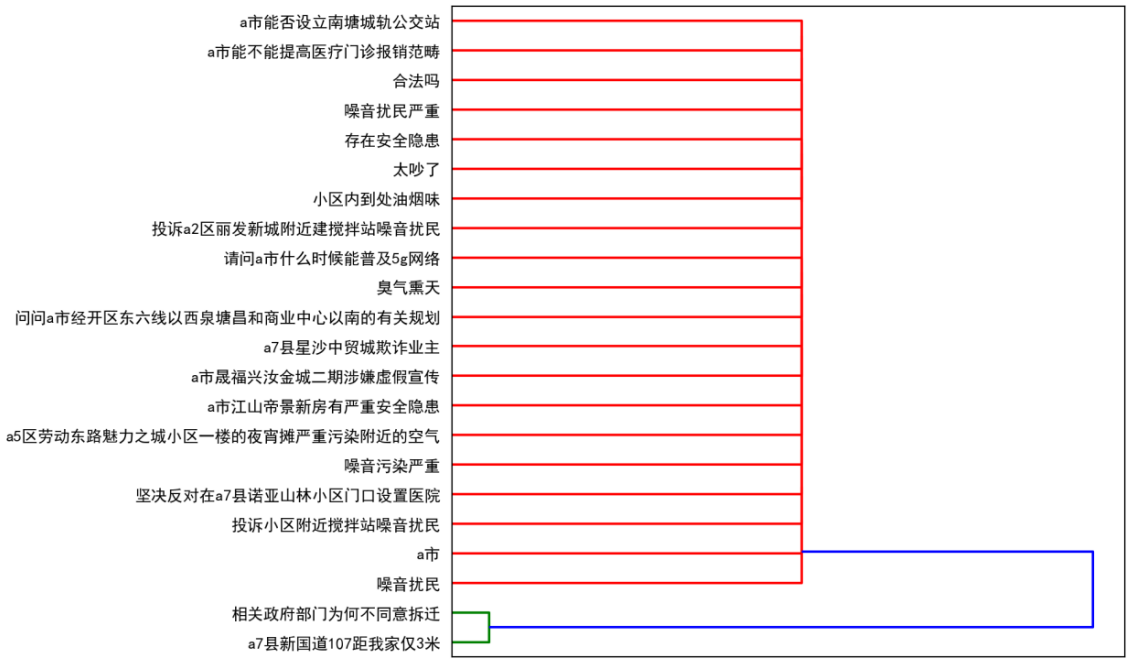
4.5.1 热门问题划分

通过使用层次凝聚聚类法在 Python 里编程，得到如图所示聚类结果。红色部分由于文本数据点距离相近，被划分成一个类型。

根据 Python 运行结果（如图所示），在 excel 中进行关键词（地点、事件）筛选，再根据留言数量选出排名前五热点问题，前五热点问题的地点为 A2 区丽发新城、A 市万科魅力之城小区、A 市江山帝景小区、a7 县诺亚山林小区、A 市

经开区。具体问题为 A2 区丽发新城附近搅拌站污染扰民、A 市万科魅力之城小区附近餐饮店油烟噪音扰民、A 市江山帝景小区新房有安全隐患、a7 县诺亚山林小区门口设置医院问题、A 市经开区有关规划咨询。

整理前五热点问题表格如下表所示：



图：Python 聚类运行结果

热度排名	问题 ID	时间范围	地点/人群	问题描述
1	1	2019/11/2 至 2020/1/26	A2 区丽发新城	小区附近搅拌站污染扰民
2	2	2019/7/21 至 2019/12/4	A 市万科魅力之城小区	小区附近餐饮店油烟噪音扰民
3	3	2019/2/26 至 2019/11/12	A 市江山帝景小区	新房有安全隐患
4	4	2019/7/8 至 2019/12/23	a7 县诺亚山林小区	小区门口设置医院
5	5	2019/1/2 至 2019/2/11	A 市经开区	经开区有关规划咨询

表：前五热点问题

4.5.2 热点问题明细

根据热门问题划分，我们将五类问题筛选出来，在 excel 中再次进行筛选观察是否有误，整理得到下表热点问题明细（部分截图）。

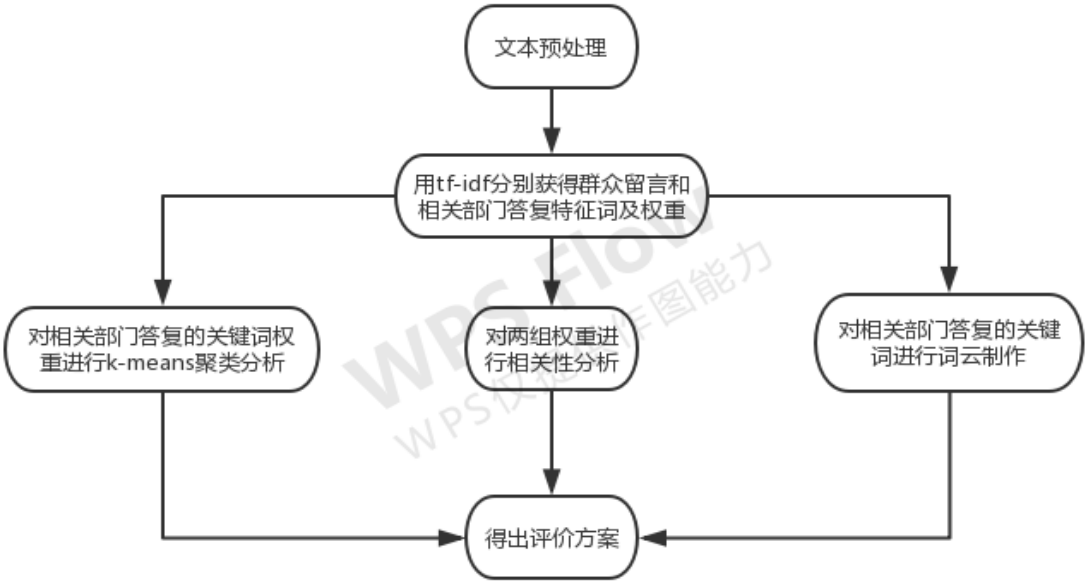
问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	267050	A909227	灰尘污染的A2区丽发新城附近已扰乱居	2019-11-02 10:18:00	音，小区居民不能正常休息，灰尘导致	0	0
1	264944	A0004260	区丽发新城附近修建搅拌厂噪音、灰尘污	2019-11-02 14:23:11	堂而皇之修建搅拌厂，请问环保部门、	0	0
1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-11-13 11:20:21	百米的地方建搅拌站。可想而知，一个	0	0
1	281943	A909216	投A2区丽发新城小区附近仍存在非法搅拌	2019-11-15 08:56:46	害小区居民的身心健康！请相关部门彻	0	0
1	225217	A909223	区丽发新城附近修建搅拌厂严重影响睡	2019-11-15 09:17:36	区附近建一搅拌站,每天尘土飞扬,噪音	0	0
1	243692	A909201	丽发新城小区附近的搅拌站噪音严重扰民	2019-11-15 11:23:21	且搅拌站的灰尘极大，都飘到小区里来	0	2
1	255008	A909208	投诉小区附近搅拌站噪音扰民	2019-11-18 12:23:22	别的地方搬过来的，体会最深的就是噪	0	0
1	272447	A909206	投诉小区附近建设搅拌站	2019-11-20 19:12:22	吵死人了，连晚上都不能好好休息，严	0	0
1	258378	A00084226	新城社区附近搅拌站修建严重影响居民	2019-11-23 00:00:00	了社区环境，还严重影响了我们业主的	0	0
1	261072	A909207	投诉小区附近搅拌站噪音扰民	2019-11-23 23:12:22	回复，在居住区建立搅拌站水泥厂是否	2	9
1	281348	A909219	希望领导“拯救”丽发新城小区居民	2019-11-24 00:00:00	睡眠受到影响，更有多名业主发现家里	0	0

表：热点问题明细表截图

5　问题 3 分析方法与过程

5.1 问题 3 流程图

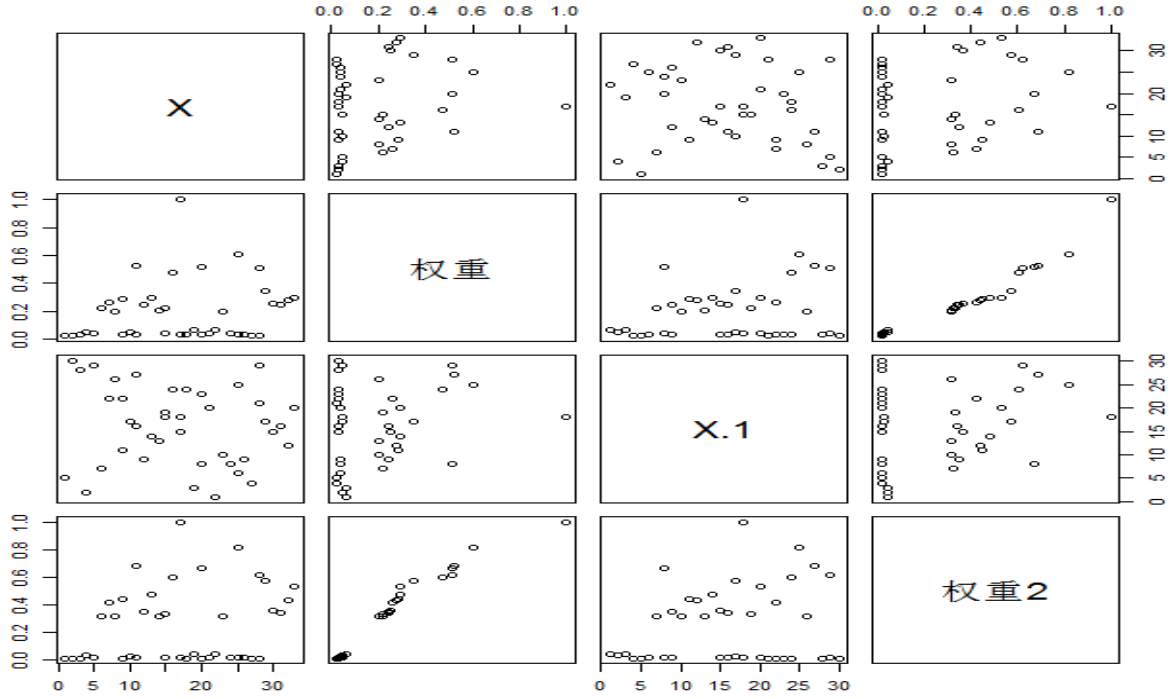
我们以留言与答复的相关性、答复的完整性和答复的可解释性为评价答复好坏的指标。



问题 3 流程图

5.2 相关性分析

对于相关性的分析，首先我们先用 TfidfTransformer()分别获得对留言和答复中每个单词的 tf-idf 权值，然后利用 tfidf.toarray()将分词中的 tf-idf 矩阵抽取出来进行了文本与向量的转换，获得每个关键词的权重，再利用每个关键词的权重进行相关性分析，从下图中可以看出，留言的关键词权重与答复的关键词权重成正相关，且相关性较大。



图：关键词权重与答复的关键词权重相关性比较图

5.3 完整性分析

对于完整性的分析，我们利用 Python 的 jieba 包和 Wordcloud 做出对于答复文本的关键词词云，通过词云的展现，我们发现“工作”，“情况”，“进行”，“回复”，“收悉”，“相关”，“反映”等词出现的次数最多，再结合我们对附件 4 中的数据的初步浏览，我们发现，大部分答复都是先进行调查，熟悉情况后，再进行合理答复的，有些答复根据情况的不同，留有不同相关部门的联系电话，或者有相关条令的普及与解说，这大大的丰富了答复的完整性，使群众获得了有效的回复和问题的及时解决。

图：相关部门答复关键词词云

5.4 可解释性分析

对于可解释性的分析，我们用了 K-medoids 聚类算法来判断每条答复之间的差异性，差异性越大可解释性就越大。我们先用 sklearn 里提供的算法包中的 CountVectorizer()函数对关于留言和答复的文本文档数据集转化成单词矩阵，然后利用 TfidfTransformer()统计每个词语的 tf-idf 权值，用 fit_transform()对文本分词进行了词频矩阵转换，转换后的词频矩阵用 vectorizer.get_feature_names()获取词袋模型中的所有词语，再将 tf-idf 矩阵从词袋模型中取出来，就能得出相关部门对群众留言的答复中的关键词及其权重，以 for 循环遍历每类文本的 tf-idf 词语权重，得出可视化聚类图

1、K-medoids 算法描述

在 K-medoids 算法执行过程中，可以通过随机的方式选择初始质心，也只有初始时通过随机方式产生的质心才是实际需要聚簇集合的中心点，而后面通过不断迭代产生的新的质心很可能并不是在聚簇中的点。如果某些异常点距离质心相对较大时，很可能导致重新计算得到的质心偏离了聚簇的真实中心。

2、K-medoids 算法步骤

（1）随机选择 k 个代表对象作为初始的中心点

（2）指派每个剩余对象给离它最近的中心点所代表的簇

（3）随机地选择一个非中心点对象 y

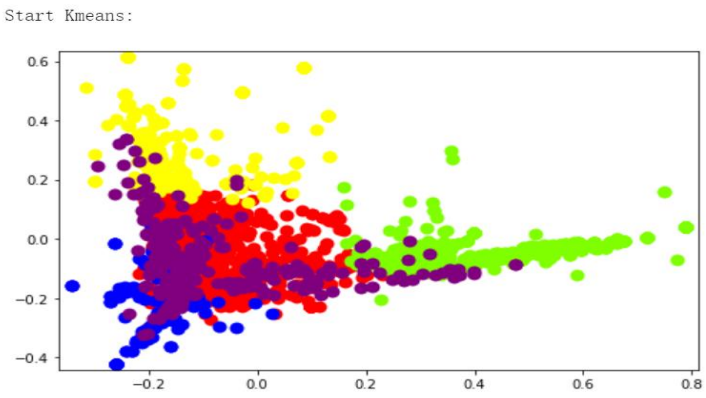
（4）计算用 y 代替中心点 x 的总代价 s

（5）如果 s 为负，则用可用 y 代替 x，形成新的中心点

（6） 重复(2)(3)(4)(5)，直到 k 个中心点不再发生变化.

5.5 问题 3 结果分析

我们用提取出来的关键词的权重数据拟合分类器模型，多次随机的选择中心点训练 K-medoids，选择效果最好的聚类结果，如下图所示。从图中可以看出我们选了 5 个簇，每个簇的分散点都比较多，这说明相关部门给的答复意见大多数不同，差异性很大，可以看出大部分答复人员做到了具体问题具体回答的做法。图中还有一些部分是密集的，经过我们对群众的留言分析，我们发现有部分留言所提出的问题几乎相似，于是，我们推断出图中密集的部分应该是相关人员对这些相似留言的密集答复处理。



图：K-means 聚类结果

5.6 评价方案

我们以上面研究的群众留言和对应的相关部门的答复的相关性、完整性和可解释性的相关规律来做出以下对于答复的评价方案：

（1）五星：群众留言与对应的相关部门答复的相关性显著+相关部门对问题进行调查，并给出合理的解决办法，留下联系方式+相关部门对回答每个问题的答复的差异性大。

（2）四星：群众留言与对应的相关部门答复的相关性显著+相关部门对问题进行调查，并给出合理的解决办法+相关部门对回答每个问题的答复的差异性大。

（3）三星：群众留言与对应的相关部门答复的相关性显著+相关部门对问题进行调查，没有给出合理的解决办法+相关部门对回答每个问题的答复的差异性大。

（4）二星：群众留言与对应的相关部门答复的相关性显著+相关部门对回答每个问题的答复的差异性大。

（5）一星：群众留言与对应的相关部门答复的相关性显著+相关部门对回答每个问题的答复的差异性小。

6.结论

对“智慧政务”中的的文本挖掘应用进行分析研究，了解社会和行业的相关需求特点与趋势，对政府了解民意、汇聚民智、凝聚民气有重大意义。同时，这也是文本分析的一个难题，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。本文通过 TFIDF 权重法提取特征词，构造词汇-文本矩，再由层次凝聚聚类法筛选出留言信息的热点问题，并对留言信息答复进行评价。

由分析结果可以看出小区居民对污染，噪音，新房安全，医院设置，经开区咨询规划等问题尤为关注。定义热门问题，可以发现热门问题的答复情况来对总体进行评价。

统计相关部门对不同留言信息的答复情况，得出答复的其相关性，完整性与可解释性都较为完善，说明政府的管理水平和施政效率有一定的基础，同时也为提升服务起到了推动作用。

参考文献

[1] 刘建平.sklearn 朴素贝叶斯类库(naive_bayes)使用小结.2018.04.17

参考网址

[1] TF-IDF 及其算法.2016.09.05.
https://blog.csdn.net/sangyongjia/article/details/52440063

[2] 手把手教你做文本挖倔：
https://zhuanlan.zhihu.com/p/30499198?utm_source=QQ_article_bottom

[3] 文本挖掘：手把手教你分析携程网评论数据 ：
https://blog.csdn.net/Nicholas_Liu2017/article/details/76021332

[4] 文本层次聚类：
https://blog.csdn.net/weixin_43718084/article/details/90313132

[5] 机器学习笔记（3）——使用聚类分析算法对文本分类：
https://blog.csdn.net/leaf_zizi/article/details/82684921

[6] 多个文本对比相似度分析：
https://blog.csdn.net/lwq123free/article/details/84255038

[7] 月上贺兰　R 语言-文本挖掘：
https://www.cnblogs.com/luhuajun/p/8654854.html