“智慧政务”中的文本挖掘应用

摘要

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，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。所以需要建立基于自然语言处理技术的智慧政务系统，提升政府的管理水平和施政效率。

对于问题 1，为了解决人工处理数量庞大的留言分类问题，建立相应的分类模型减少认为工作量而提高部门工作的效率。首先进行文本预处理即对文本内容去重后，利用 jieba 将其分词，为了避免无意义词的干扰将文本去停用词。为了后续文本分类模型的建立，需要对文本数据量化，利用 TF-IDF 算法，得到每个词的权重大小即获取索沃德关键词，同时将其转化为向量的表达式。

人工分类留言内容是一项十分庞大的任务，为了准确有效的使每一条留言分到正确的部门，我们团队利用 SVM 分类算法，将一级标签作为分类的标签，找到训练集的分类策略，实现对测试集的标签标注。为了验证机器学习分类的准确性，我们利用 F-Score 评价该模型的准确度，通过计算可知最终测试出来的数据准确率高达 90%。

对于问题 2，为了挖掘某市的热点问题，促使政府各部门第一时间解决处理，在此我们通过对留言内容进行文本预处理等操作，后运用 TF-IDF 算法进行权重处理。观察数据可知热点问题的三要素分别为地点、人群、事件，所以对分词结果进行命名体识别，将地点、人群等专有名词提取出来。通过大规模的语料计算，模型的预测方法，将观测序列转换为标注序列。

由于提取得到的词性向量矩阵是高维数据，所以通过 T-SNE 方法对其降维处理。由 K-Means 分类得到聚类中心，利用 KNN 算法找出离各个聚类中心最近的元素，根据“少数服从多数”判定聚类中心所属类别。结合样本，分别找出聚类中心的近邻样本点所属的地点人群从而得到问题。定义评价指标，运用层次分析法分析分析指标权重，从而得到 5 个热点问题。

对于问题 3，为了对答复意见的质量给出一套评价方案，我们分别从答复的相关性、完整性、可解释性、时效性等角度去评判。

①相关性：将留言内容和留言回复特征词进行相似度计算。

②完整性：查看留言内容的关键词是否含有规范的回复格式。

③可解释性：查看相关部门是否给予相应的处理和解决办法。

④时效性：根据回复时间和留言时间计算相应的天数可以知道部门处理问题效率。

用余弦函数求两个计算相似度的文本的词频矩阵的余弦值，即为得到的 TF 系数，也就是两个文本的相似度。这里我们利用循环输出了每个对应的留言详情和答复意见的相似度。我们计算了相似度的平均值，达到了 0.8，可以认为答复意见较贴切留言。然后利用灰色关联分析模型，求得这四个因素对留言回复的影响大小，根据其影响力大小我们能具体的评判每一条留言的回复质量是否达标。

关键词： TF-IDF 算法　SVM　K-means 聚类　Knn　灰色关联分析

目录

1. 挖掘目标

1.1 问题分析

本次建模目标是利用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等数据，利用 jieba 中文分词工具对留言内容分词，利用去停用词去除无相关词语后，建立模型达到以下三个目标：

利用文本分词和去停用词等操作对附件 2 留言内容进行预处理，根据处理结果进行权重计算，结合已知的留言标签分类随机选出留言来进行训练，并通过测试集验证训练结果，解决留言分类以及标签标注问题，以便后续将群众留言分派至相应的职能部门处理，达到减少工作人员工作量、提高效率等目的。

根据附件 3 某一时段内反映特定地点或特定人群问题的留言，结合点赞数及涉及人群等评价指标，分析出当前的热点问题，帮助有关部门及时发现热点问题，并进行有针对性地处理，提升服务效率。

针对附件 4 相关部门对留言的答复意见，我们从答复的相关性、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。

1.2 分析方法与过程

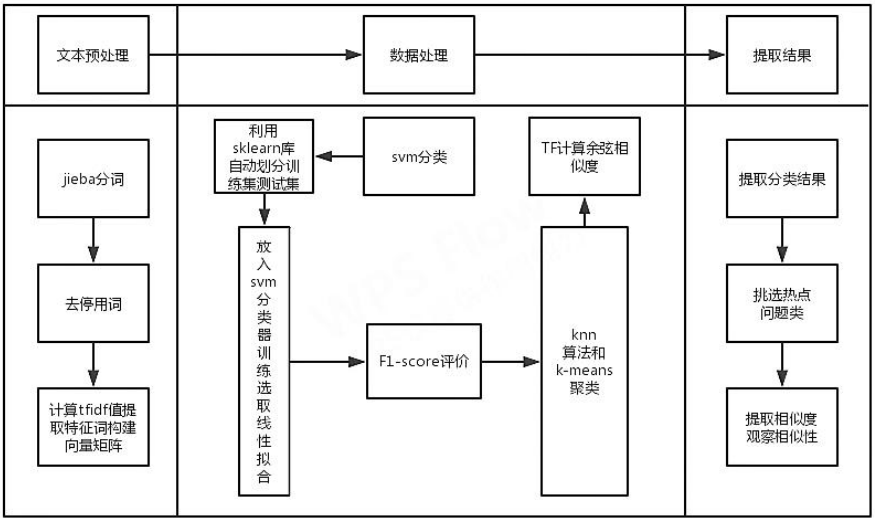

图 1：总流程图

2.问题 1 分析方法与过程

2.1 流程图

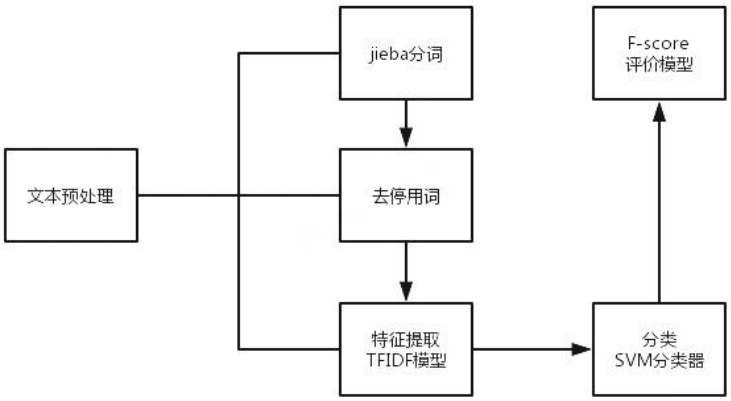

图 2：问题一流程图

2.2 文本预处理

2.2.1 留言内容的去重

在附件 2 给出的数据中，出现了很多重复的留言内容，重复留言会增加后续编程的工作量。考虑到部分留言相似程度极高，可是在某些词语的运用上存在差异，若是删除文字相近留言，则会出现误删的情况，去除这类留言显然不合适。因此，为了存留更多的有用语料，本节针对完全重复的语料下手，仅删除完全重复部分，以确保尽可能保留有用的文本留言信息。利用 python 编程去重后，留言由 9210 条变为 9052 条。

2.2.2 文本分词

为了后续方便计算机进行计算，我们需要通过分词的方式实现数据的量化。基于统计的分词方法是从大量已经分词的文本中，利用统计学习方法来学习词的切分规律，从而实现对未知文本的切分。

jieba 分词是基于前缀词典实现高效的词图扫描，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)，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，从而在中文分词方面达到很好的效果。分词是文本信息处理的基础环节，是将一个单词序列切分成一个一个单词的过程。准确的分词可以极大的提高后续计算的准确率。

2.2.3 去停用词

停用词一定程度上相当于过滤词，区别是过滤词的范围更大一些。去停用词是对连词

以及副词等无实际意义的词加以处理，提高实际词汇的权重。我们在分词的基础上导入一个自定义的词库，将词库中的词作为停用词，把留言内容中的没有意义的词删去。

2.2.4 绘制词云查看分词效果

进行数据预处理后，可绘制词云查看分词效果，词云会将文本中出现频率较高的“关键词”予以视觉上的突出。首先需要对词语进行词频统计，将词频按照降序排序，选择前100个词，使用wordcloud模块的WordCloud绘制词云，词云如图所示。

图3：分词词云

根据词云可以看出，对留言内容数据进行预处理后，分词效果较为符合预期。

2.3 TF-IDF计算权重

在对留言内容进行分词，去停用词后，需要把这些词语转化为向量，以供分类模型使用。这里采用TF-IDF算法，把留言内容转换为权重向量算法的具体原理如下：

第一步，计算词频，即TF权重

词频（TF）=某个词在文本中出现次数　　（1）

考虑到留言内容有长短之分，为了便于不用内容的比较，进行“词频”标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$\text{词频（TF）}=\frac{\text{某个词在文本中出现的次数}}{\text{文本的总词数}} \qquad (2)$$

第二步，计算 IDF 权重，即逆文档频率，需要建立一个语料库，用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率（IDF）}=\log\left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数}+1}\right) \qquad (3)$$

第三步，计算 TF-IDF 值

TF - IDF = 词频 × 逆文档频率 （4）

实际分析得出 TF-IDF 值与一个词在留言详情出现的次数成正比，某个词文本的重要性越高，TF-IDF 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职位描述表中文本的关键词。如图随机选取十个词语的 TF、IDF、TF-IDF 值如下：

	TF	IDF	TF-IDF
拆迁	172	6.522	1121.784
生二胎	43	7.645	328.748
环境影响	56	7.003	392.194
规划设计	59	7.573	446.806
民生问题	64	7.223	462.292
垃圾场	44	7.837	344.858
沟通	153	6.351	971.639
监督管理	94	7.291	685.267
公信力	54	8.243	445.131
财务	91	7.191	654.385
扰民	193	6.823	1316.903

图 4：TF-IDF 值

2.4 SVM 支持向量机算法

支持向量机是一类按监督学习方式对数据进行分类的广义线性分类器。在分类问题中给定输入数据和学习目标：$X=\{X_1,.....,X_N\}$，$y=\{y_1,.....,y_N\}$，其中输入数据的每个样本都包含多个特征并由此构成特征空间：$X_i=[x_1,.....,x_n]\in X$，而学习目标为二元变量 $y\in\{-1,1\}$ 表示负类和正类。

若输入数据所在的特征空间存在作为决策边界的超平面将学习目标按正类和负类分开，并使任意样本的点到平面距离大于等于 1：

$$w^{\mathrm{T}}X+b=0 \quad (1)$$

$$y_i\left(w^{\mathrm{T}}X_i+b\right)\geq 1 \quad (2)$$

则称改分类问题具有线性可分性，参数 分别问超平面的法向量和截距。满足该条件的边策边界实际上构造了 2 个平行的超平面作为间隔便捷来判别样本的分类：

$$w^{\mathrm{T}}X_i+b\geq+1,\Rightarrow y_i=+1 \quad (3)$$

$$w^{\mathrm{T}}X_i+b\leq-1,\Rightarrow y_i=-1 \quad (4)$$

所有在上间隔边界上方的样本属于正类，在下间隔边界西方的样本属于负类。两个间隔边界的距离 $d=\frac{2}{\|w\|}$ 被定义为边距，位于间隔边界上的正类和负类样本为支持向量。

我们在得到词语的 TF-IDF 权重后，根据已知的留言分类，将其分为训练集与测试集，运用 SVM 支持向量机算法,找到训练集的分类策略，实现对测试集的标签标注，从而实现对未知文本的留言分类。

2.5 F-Score 分类方法评价

利用 SVM 建立分类模型后，我们能得到机器学习后生成的每条留言内容对应的一级标签，为了检验建立分类模型的准确性，采用 F-Score 算法对其评价。假定 P_i 为第 i 类的精确率精确率是对测试集的结果而言的，是真正正确的占所有预测为正的比例，其代表对正样本结果的预测的准确程度、R_i 为第 i 类的召回率，是针对原样本而言的，是真正正确的占所有实际为正的比例。公式如下：

$$P_i=\frac{TP_i}{TP_i+FP_i} \quad (1)$$

$$R_i=\frac{TP_i}{TP_i+FN_i} \quad (2)$$

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \quad (3)$$

最终把统计出来每一类别的 P_i 和 R_i，代入 F 中求得 SVM 分类模型是否达到最优。通过计算

我们得到如下表所示：

	P	R
城乡建设	0.8571	0.9524
环境保护	0.9153	0.9772
交通运输	0.8965	0.9378
教育文体	0.9006	0.8987
劳动和社会保障	0.9367	0.9267
商贸旅游	0.8978	0.9983
卫生计生	0.9786	0.9356
	F=0.92821271	

图 5：F-Score

计算得出 F 值高于 90%，所以可以判定 SVM 模型对留言内容的分类达到了我们预期所要的结果。

3. 问题 2 分析方法与过程

3.1 流程图

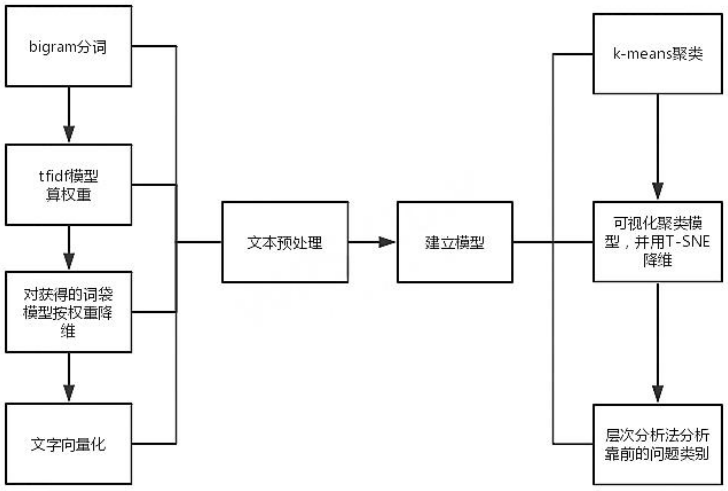

图 6：问题二流程图

3.2 命名体识别

为了查找出热点前 5 的留言内容，我们将附件 3 留言主题进行文本处理，基于问题 1 的文本预处理的操作，分别对留言主题进行 jieba 分词、去停用词、以及 tf-idf 权重的处理，再次不过多的加以叙述。然后对分词结果进行词性标注，通过大规模的语料计算，可以得到 HMM 的参数，然后结合模型的预测方法，将观测序列，也就是分词之后的句子转换为标注序列。jieba 的词性标注结果，对于每个词的词性，采用和 ICTCLAS 兼容标记法，其中 ns 表示地名，nt 表示机构团体，nz 表示专有名词。由于针对热点问题的研究，我们只需要知道相应的地点、人群以及时间，将这些词性的词提取出来并相应地导出权重，从而实现了将文字向量化的过程。词性标注如下图所示：

	index_content	word	nature	index_word
0	1	座落在	i	0
1	1	市区	n	1
2	1	联丰路	nt	2
3	1	米兰	ns	3
4	1	春天	t	4
...	...	...	...	...
702584	4326	当	t	153
702585	4326	儿童	n	154
702586	4326	顾问	nr	155

图 7：词性标注

3.3 T-SNE 降维

T-SNE 可以算是目前效果最好的数据降维和可视化方法之一，由于我们提取得到的词性向量矩阵是高维数据，所以在进行分类时无法准确的知道这个矩阵是否有很好的可分性由于同类之间间隔小、异类之间间隔大，所以可以通过 T-SNE 方法将矩阵投影到 2 维或 3 维空间中观察：如果在低维空间中具有可分性，则矩阵是可分的。

T-SNE 将数据点之间的相似度转化为条件概率，原始空间中数据点的相似度由高斯联合分布表示，嵌入空间中数据点的相似度由学生 t 分布表示。通过原始空间和嵌入空间的联合概率分布的 KL 散度（用于评估两个分布的相似度的指标，经常用于评估机器学习模型的好坏）来评估嵌入效果的好坏，即将有关 KL 散度的函数作为损失函数，通过梯度下降算法最小化损失函数，最终获得收敛结果。

3.4 K-Means 聚类

运用 T-SNE 降维处理后，根据词语的权重向量，在特定地点及特定人群方面对他们进行问题分类，这里采用 K-Means 算法分类。K-Means 是经典的划分聚类算法，算法的有点事时间复杂度低，聚类效果好。因此，利用 K-Means 算法经过向量化的特征词进行聚类，步骤如下：

①随机选择 K 个中心点；

②遍历所有数据点，把数据点划分到距离最近的一个簇类中；

③划分之后就有 K 个簇，计算每个簇类中点的平均值作为新的簇类中心点；

④重复步骤②和③，知道聚类中心不再发生变化，或是迭代次数达到设定的值。

对 K-Means 聚类中的 K 值得选择，可以依据基于误差平方和 SSE 的手肘法，计算公式如下：

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \qquad (1)$$

其中，C_i 是第 i 个簇，p 是 C_i 中的样本点，m_i 是 C_i 的质心即 C_i 中所有样本的均值，SSE 是所有样本的聚类误差，代表了聚类效果的好坏，在确定 K 的取值后，使用 K-Means 聚类算法对从信息技术文档中提取的特征进行聚类。具体操作流程如下图所示：

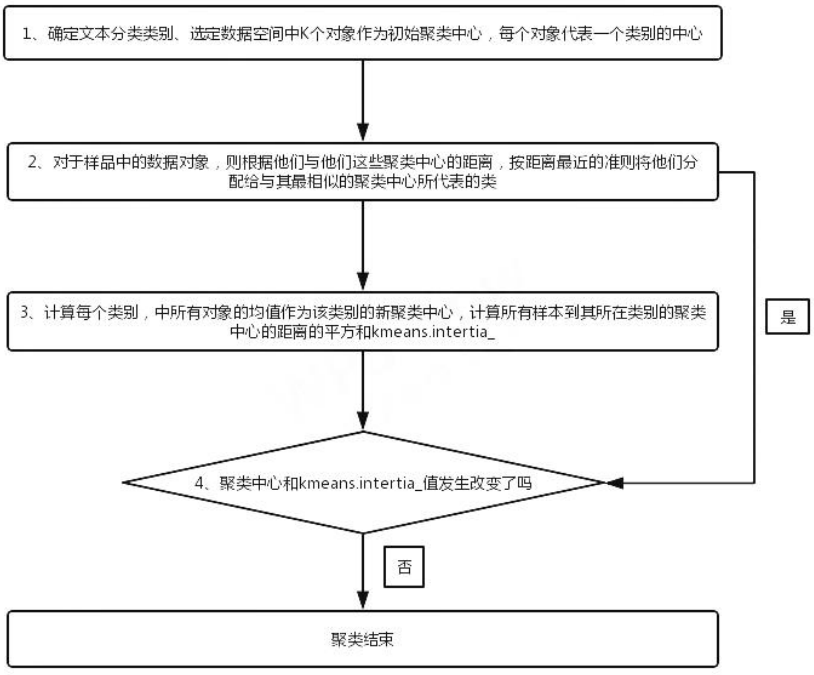

图 8：K-Means 算法流程图

3.5 knn 最邻近分类算法

由 K-Means 聚类有效性度量得到聚类中心后，利用 Knn 算法来确定初始聚类的中心点，进而实现对留言的分类。在对文本聚类后，我们得到了各类别的中心向量，表示为 $C_j(P_1,P_2,\cdots,P_n)$。由于向量组中存在一些值使向量范围较大，会造成文本分类结果不理想，所以要平衡各个属性对距离的影响，映射公式为：

$$X_i^{'}=(X_i-\min(\forall X_i))/(\max(\forall X_i)-\min(\forall X_i)) \tag{1}$$

这里 X_i 为向量 X 的第 i 个分量。将文本向量化为 n 维向量的形式，表示为 $T(P_1,P_2\cdots,P_n)$，根据距离找出相似的文本，计算文本相似度。具体步骤如下：

计算中心点的欧式距离，公式如下：

$$d=\sqrt{(x_A-x_i)^2} \tag{2}$$

其中，中心点坐标为 $A(x_i,x_i),B(x_j,x_j)$。

（2）计算文本间的相似度，公式为：

$$Sim(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{\sum_{k=1}^{M}W_{jk}^2}\sqrt{\sum_{k=1}^{M}W_{ik}^2}} \tag{3}$$

其中，d_i 为文本向量，d_j 为中心向量。

按照文本的相似度，对文本进行热点问题的分类。最终得到程序结果如下图所示：

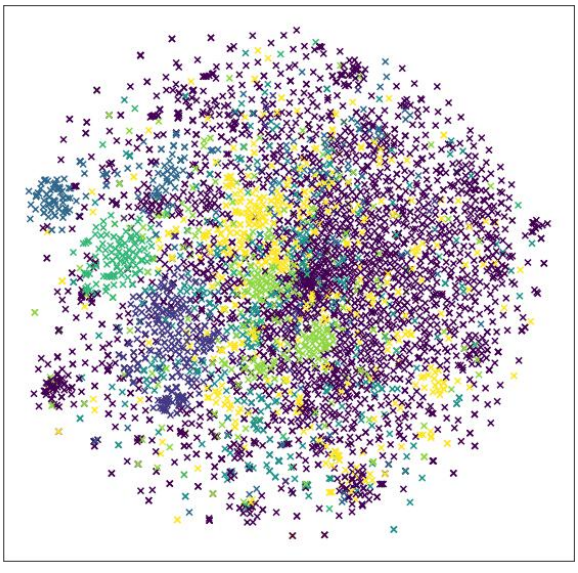

图 9：聚类分类图

上图为对留言详情分类的可视化结果图，存在允许范围内的误差，可以看出有的留言较其他留言分散，所以我们把统计数量多的留言作为一个评价指标。

3.6 基于层次分析法的评价指标权重计算

3.6.1 标注及描述

层次分析法中常用五个等级来区分指标的重要程度，分别为同样重要、稍微重要、较强重要、强烈重要、绝对重要。为了方便定量化，引入 1-9 比率标度法，其中 1、3、5、7、9 分别表示要素 i 与要素 j 相对比的五个等级，而 2、4、6、8 表示判断级之间的折衷值。

3.6.2 层次分析

层次分析法是将目标、准则和决策对象按照之间的相互关系分为最高层，中间层和最底层，我们经过分析画出层次结构图：

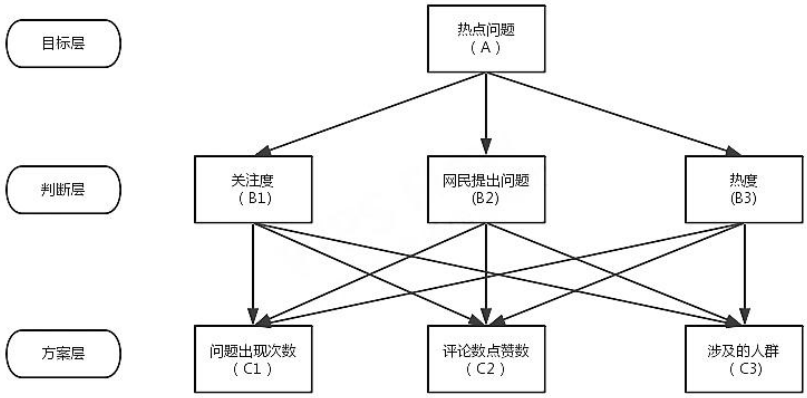

图 10：层次结构图

我们通过“1-9”标度法将评价指标进行两两比较，逐层得到权重排序结果，因而得到判断矩阵 P：

$$P=\begin{bmatrix} P_{11}\cdots\cdots P_{1\mathrm{m}} \\ p_{12} \\ \vdots \\ P_{1n}\cdots\cdots P_{nm} \end{bmatrix} \quad (1)$$

我们根据下式，运用方根法求解判断矩阵的特征向量：

$$\overline{\omega}(i=1,2,\cdots,\ n) \quad (2)$$

$$M_i=\prod_{j=1}^{n}P_{ij} \quad (3)$$

将 w_i 按照下式进行归一化处理：

$$W_i^{'} = \frac{\overline{\omega}}{\sum_{i=1}^{n}\overline{\omega_i}} \qquad (4)$$

得到判断矩阵的特征向量：$W^{'} = \left(W^{'}_1, W^{'}_2, \cdots, W^{'}_n\right)$

我们对矩阵进行了一致性检验，保证其准确性。

3.6.3 结论

我们利用 matlab 计算了各指标对总体排序的重要值，得到 C1，C2，C3 的重要度权重计算分别为：0.425,0.281,0.293。所以我们可以看出人们反应问题的数量是评价热点问题的重要指标，其次是涉及到的人群和点赞数。得到了如下图所示的前十名分类及排序。

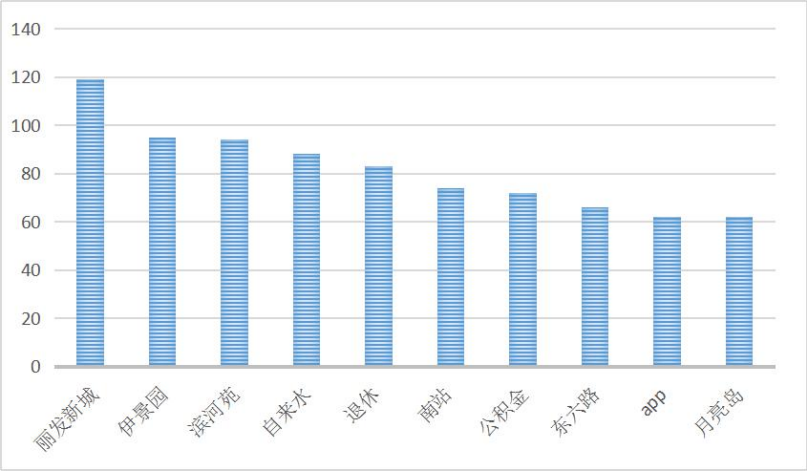

图 11：热点问题分类及排序

4. 问题 3 分析方法与过程

4.1 多维向量余弦相似度

4.1.1 相似性的求解

对留言内容和留言回复的进行相似性的求解，首先分别进行文本预处理，然后利用编辑距离计算、杰卡德系数计算、TF 计算、TFIDF 计算等计算文本相似度的算法，得出 TF 计算的相似度最接近正常的数值即留言回复的相关性。

4.1.2 夹角的余弦值

TF 计算是直接计算 TF 矩阵中两个向量的相似度，即求解两个向量夹角的余弦值：

$$\cos\theta = \frac{a \bullet b}{|a| * |b|} \quad (1)$$

运用 TF 计算，先把留言详情和答复意见读入 Python，然后对其进行文本预处理：结巴分词，去停用词，这里使用 sklearn 库中的 CountVectorizer 来计算句子的 TF 矩阵，再利用 numpy 库计算出二者的交集和并集，通过 CountVectorizer 的 fit_transform()将预处理好的文本转化为对应的词频矩阵，利用 np.dot()和 norm()方法分别获得向量的点乘积和模长，再利用余弦函数求两个计算相似度的文本的词频矩阵的余弦值，即为得到的 TF 系数，也就是两个文本的相似度。这里我们利用循环输出了每个对应的留言详情和答复意见的相似度。我们计算了相似度的平均值，达到了 0.8，可以认为答复意见较贴切留言。

4.2 灰色关联分析法

4.2.1 思路分析

由于我们已经计算出留言回复和留言内容之间的相似度，得到每条留言的相关性是否达到一定的标准的基础上，为了建立一套标准的针对答复意见的评价方案，在这里我们采用灰色关联分析模型，从留言回复的相关性、完整性、可解释性等多方面因素进行关联分析。

灰色系统理论提出了对各因素系统进行灰色关联度分析的想法，意图透过一定的方法，去寻求系统中各因素之间的数值关系。因为附件 4 中的留言回复存在对留言内容已经解决和未解决等问题，以及回复内容是否具有完整性、对出现的问题是否得到了相应的举措，还有针对回复的时间我们也可以得知政府有关部门效率。为了针对这几种因素来评定留言回复的质量，灰色关联度分析对于一个系统发展变化态势提供了量化的度量标准。

4.2.2 指标量化处理

由于我们的数据都是文本类型，所以需要先将文本进行处理，针对留言内容和留言回复都需要进行分词、提取特征词等方法如问题 1 文本预处理方法。

①相关性：将留言内容和留言回复特征词进行相似度计算。

②完整性：查看留言内容的关键词是否含有规范的回复格式。

③可解释性：查看相关部门是否给予相应的处理和解决办法。

④时效性： 根据回复时间和留言时间计算相应的天数可以知道部门处理问题效率。

4.2.3 灰色关联分析模型

（1）确定反映系统行为特征的参考数列和影响系统行为的比较数列

反映系统行为特征的数据序列，称为参考数列。影响系统行为的因素组成的数据序列，

称比较数列。

（2）对参考数列和比较数列进行无量纲化处理

由于系统中各因素的物理意义不同，导致数据的量纲也不一定相同，不便于比较，或在比较时难以得到正确的结论。因此在进行灰色关联度分析时，一般都要进行无量纲化的数据处理。

求参考数列与比较数列的灰色关联系数 $\xi(X_i)$

所谓关联程度，实质上是曲线间几何形状的差别程度。因此曲线间差值大小，可作为关联程度的衡量尺度。对于一个参考数列 X_0 有若干个比较数列 $X_1, X_2, \ldots\ldots X_n$，各比较数列与参考数列在各个时刻（即曲线中的各点）的关联系数 $\xi(X_i)$ 可由下列公式算出：其中 ρ 为分辨系数，一般在 0~1 之间，通常取 0.5。Δ 是第二级最小差，记为 $\Delta_{\min}$。 是两级最大差，记为 $\Delta_{\max}$。

为各比较数列 X_i 曲线上的每一个点与参考数列 X_0 曲线上的每一个点的绝对差值，记为 $\Delta oi(k)$。所以关联系数 $\xi(X_i)$ 也可简化如下列公式：

$$\xi_{oi} = \frac{\Delta(\min) + \rho\Delta(\max)}{\Delta_{oi}(k) + \rho\Delta(\max)} \qquad (2)$$

（4）求关联度

因为关联系数是比较数列与参考数列在各个时刻（即曲线中的各点）的关联程度值，所以它的数不止一个，而信息过于分散不便于进行整体性比较。因此有必要将各个时刻（即曲线中的各点）的关联系数集中为一个值，即求其平均值，作为比较数列与参考数列间关联程度的数量表示，关联度公式如下：

$$r_i = \frac{1}{N}\sum_{k=1}^{N}\xi_i(k) \qquad (3)$$

r_i 比较数列 x_i 对参考数列 x_0 的灰关联度，r_i 值越接近 1，说明相关性越好。

（5）关联度排序

因素间的关联程度，主要是用关联度的大小次序描述，而不仅是关联度的大小。将 m 个子序列对同一母序列的关联度按大小顺序排列起来，便组成了关联序，记为 $\{x\}$，它反

映了对于母序列来说各子序列的“优劣”关系。若 $r_{oi} \geq r_{oj}$，则称 $\{x_i\}$ 对于同一母序列 $\{x_0\}$ 优于 $\{x_j\}$，记为 $\{x_i\} \geq \{x_j\}$。

我们通过比较各因因素关联度的大小来判断待识别对象对研究对象的影响程度，将相关性、可解释性、完整性、时效性进行排序得到影响留言回复质量的重要程度，从而判定出一个标准的评价方案即相关性>时效性>可解释性>完整性。

5. 结论

由于现在人工处理留言分类问题存在工作量大效率低等影响，所以本次项目研究主要解决对群众留言分类等问题，使每一条留言能准确的分配到有关部门，能够让群众的问题得到及时的解决，同时也有助于让有关部门的领导关心和重视群众，倾听群众的声音，满足群众的诉求。本文运用了文本分析的算法，将文本进行量化后运用 SVM 支持向量机、K-Means 聚类和灰色关联等算法，对群众的留言进行数据处理来达到我们想要的结果。

通过结果我们分析得到，只有将留言分类化、热点化，群众问题才能得到高效及时地解决。同时也可以通过答复意见看出相关部门对问题解决的及时性和完整性。虚心听取群众的意见和建议，以真心换取群众的真心话。

6.参考文献

[1] 余以胜，韦锐，刘鑫艳.可解释的实时图书信息推荐模型.华南师范大学经济与管理学院.2019

[2] 徐文进，管克航，寻晴晴，许瑶，解钦.基于 KNN 算法的改进 K-means 算法.青岛科技大学.2019

[3] 毛郁欣，邱智学.基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究.浙江工商大学管理工程与电子商务学院

[4] 曹卫峰.中文分词关键技术研究.南京理工大学.2009

[5] 李学明，张朝阳，余维军.基于用户回复内容观点支持的的评论有用性计算.重庆大学.2016