“智慧政务”中信息的分析和挖掘

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，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。同时，随着大数据、云计算、人工智能等技术的发展，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。因此，利用自然语言处理和文本挖掘技术对“智慧政务”相关信息的研究具有重大的意义。

对于问题 1，群众留言分类。数据操作采用了 Pandas 的 DataFrame 数据结构对 Excel 文件读写和处理，数据的预处理用 Python 的中文正则表达式先提取英文字母和数字开头的地名单独提取，存入用户词典，再用经典的 Word2vec 工具的 Jieba 工具进行主题中文分词和详情关键词读取，采用了 TF-IDF 模型结合 LSI 模型对中文分词和关键词做向量化处理和文本的多分类模型的训练与预测。实验证明，经过优化的分词效果几乎达到人工分词的要求，细节方面还可优化；模型方面采用 TF-IDF 模型与 LSI 模型在用 70%，80%，90%等不同样本率进行比较 TF_IDF 模型分类效果与校本的数量有较大，LSI 模型对样本的要求不高，稳定性很好，在仅用主题进行分词的情况下，分类的准确率达到了 86%，结合详情关键词分词效果应该能达到 90%以上，而在 CountVector 模型和 LogisticRegression 分类模型下，单个一级分类的分类准确率仅有 73%。

对于问题 2，热点问题挖掘。数据的预处理主要是用问题 1 的技术和方法：正则表达式提词做用户词典，Jieba 分词，LSI 模型的向量化处理。数据预处理中有 2 点重点：1，抓住问题的时间地点主题三要素中的时间要素中时间要素进行时间段的划分处理，避免实验数据时间跨度比较大引起的热点问题失真。2，分词和关键词做 LSI 模型的向量化，采用向量余弦作距离进行文本相似度计算。热点问题的分类模型采用 K-Means 分类模型得到聚类中心，利用 Knn 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，寻找到所需的热点问题。实验工具主要使用 SPSSS 和 Python，根据问题 1 的分类模型对问题 2 分类，再在每

一个类别中进行问题归并，把相似的留言归为同一问题，最终比较得到排名前 5 的热点问题。实验效果经过人工比对，基本符合实际情况。

留言问题热度量化指标采用 AHP 层次分析法，先从留言发布者、留言详情、留言内容特征和群众反映四个维度选取留言内容特征热度影响进行定量评价。然后对群众留言内容特征从群众留言内容和表现形式两个角度,选取留言信息充实度、留言内容出现及时率和充实度三个二级指标进行反映。留言信息充实度用留言的长度来反映，留言字数越多其表达出的信息越充实,也越容易引发讨论。留言出现及时性主要反映留言出现的时间段，结合点赞率，反对率，评论率三方面。对各要素建立对比矩阵计算权重，根据总权重进行热度排序。问题 2 筛选出的前五个热点问题，使用 Matlab 软件按照留言热度评价指标体系所涉指标对五个热点问题进行数据采集与分析，说明评价指标还是有一定的可行性。

对于问题 3，答复意见的评价。运用 AHP 层次分析法答复意见评价指标进行分析先依据附件 4 制定了政府答复意见的文本模板，对该文本进行文本数据预处理，提取并整理初步的评价指标。评价指标主要有：一级指标有 2 项：答复质量、答复时效，二级指标有 3 项：语言运用、完整性、时效性，三级指标有 6 项：语言简洁程度、礼貌用语、留言情况调查、具体处理方案、收到答复时间、问题解决进度。根据三级指标六个标准建立矩阵进行加权归并，算出最终权重。因为时间关系，仅以示例说明评价指标方法的可行性。

关键词:正则表达式；JIEBA 分词；LSI 模型； K-Mmeans 聚类模型；KNN 算法；AHP 层次分析法

目录

1、挖掘目标

本次建模目标是利用网络智慧政务系统发布的群众留言相关的信息数据，利用jieba中文分词工具对群众留言描述进行分词、K-means聚类的方法及KNN算法，达到以下三个目标：

1)利用文本分词和文本聚类的方法对非结构化的数据进行文本挖掘，根据聚类结果，给群众留言合理归类。

2)再结合群众留言特点，及相似度计算，判断群众热点留言问题。

3)根据研究的群众留言情况、相关机构答复意见，建立答复意见评价指标，做出合理的答复意见评价。

2、分析方法与过程

本论文主要包括以下步骤:

步骤一：数据预处理，对题目给出的文本数据数值化处理，在原始的文本数据上去除重复词及空行、中文文本分词、停用词过滤，以便后续分析。

步骤二：数据分析，在对群众留言描述信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用TF-IDF算法，找出每条留言描述的关键词，把群众留言描述信息转换为权重向量。采用K-means算法对群众留言进行分类，利用KNN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，统计相关数据，分类筛选汇总，预测群众所关心的热点问题。

总体流程图如下：

图1：总体流程图

2.1 问题 1 分析方法和过程

2.1.1 问题 1 流程图

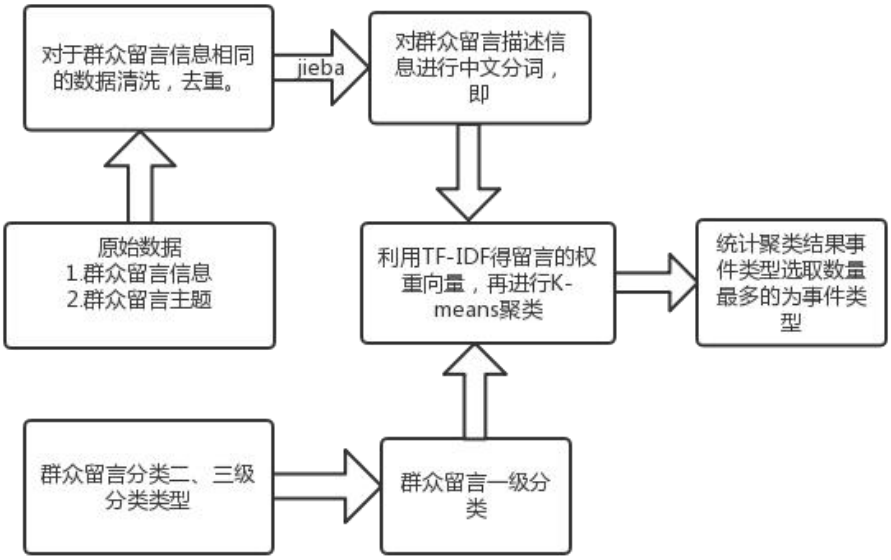

图 2：问题 1 流程图

2.1.2 数据预处理

数据描述：通过观察所给的全部数据，发现数据量很大，这些留言信息都存在于 excel 表格中，所以需要将其量化成数值形式才能便于我们进行分析。同时我们发现，这些留言详情中存有大量空行以及重复的情况，这就需要我们做一些数据处理，否则会对后续挖掘分析造成影响，这样把留言数据引入进行中文分词、词频统计乃至文本聚类等，会对聚类结果的效果造成一定的影响，所以本文首先在处理上要采取数据预处理。

2.1.2.1 群众留言信息表进行数据清洗

先将数据导入处理工具。使用文本文件存储+Python 操作的方式。然后观察

数据。这里包含两个部分：一是看元数据，包括字段解释、数据来源、代码表等等一切描述数据的信息；二是抽取一部分数据，使用人工查看方式，对数据本身有一个直观的了解，并且初步发现一些问题，为之后的处理做准备。

用 python 代码 excel_read_write ，将群众留言主题和留言详情按列提取出来，分别储存在文件留言主题附件 2（全部数据）_r.txt 和留言详情附件 2（全部数据）_r.txt 中。并要修改一些格式错误的数据，并去除不需要的字段，一些重复的群众留言评论的事件类型，直接删掉即可，在小规模数据上试验成功再处理全量数据，以避免处理不当所导至的数据丢失。进行数据清洗的数据保存在附件中。

2.1.2.2 对群众留言描述信息进行中文分词[1]

进行群众留言信息挖掘之前，先要把非结构化的文本信息转化为计算机能够识别的结构化信息。给出的数据都是以中文文本的形式出现，则需进行中文分词。jieba 是目前最好的 Python 中文分词组件，它支持 3 种分词模式：精确模式、全模式、搜索引擎模式；支持繁体分词；支持自定义词典。可使用 jieba.cut 和 jieba.cut_for_search 的方法进行分词。两者所返回的结构都是一个可迭代的 generator，可使用 for 循环来获得分词后得到的每一个词语（unicode），或者直接使用 jieba.lcut 以及 jieba.lcut_for_search 直接返回 list。

在这过程中涉及了能让汉字成词的 HMM 模型。在这过程中还需要对文本的停用词进行去除，停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据或文本之前或之后会自动过滤掉某些字或词，自行建立一个停用词库（stopwords.txt），在分词的基础上利用 python 语句来把文本中的停用词经 python 代码 temp.py 去除后储存在新的 txt 中。

附件 2 留言详情部分分词结果示例如图：

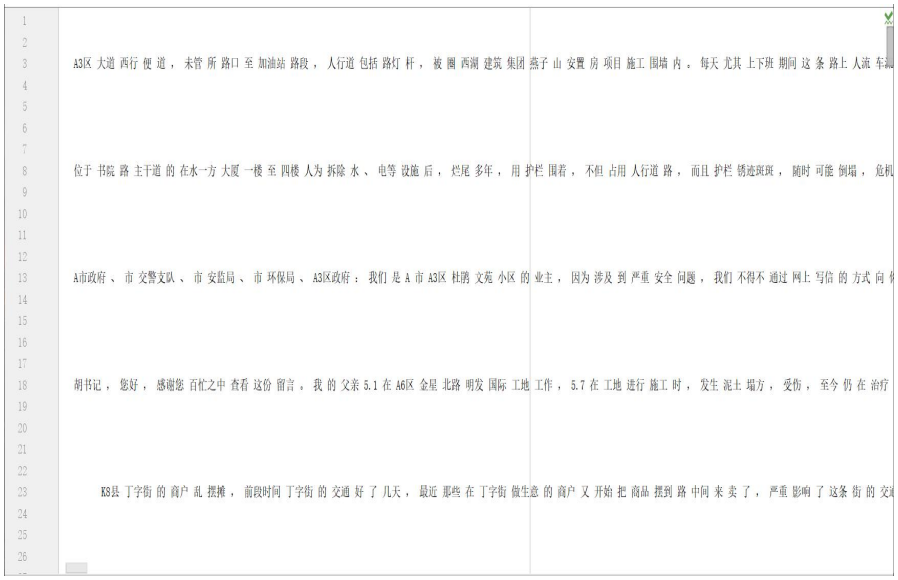

图 3：部分分词结果示例图

2.1.2.3 TF-IDF 算法[2]和生成 TF-IDF 向量[3] [4]

在完成对群众留言信息的分词后，把这些词语转化为向量，以便使用。则采用 TF-IDF 算法，这个算法用统计学语言表达，就是在词频的基础上，要对每个词分配一个"重要性"权重。最常见的词给予最小的权重，较常见的词给予较小的权重，较少见的词给予较大的权重，这个权重叫做"逆文档频率"，其英文为 Inverse Document Frequency，缩写为 IDF，它的大小与一个词的常见程度成反比。

知道了"词频"（TF）和"逆文档频率"（IDF）以后，将这两个值相乘，就得到了一个词的 TF-IDF 值。某个词对文章的重要性越高，它的 TF-IDF 值就越大。所以，排在最前面的几个词，就是这篇文章的关键词。TF-IDF 算法的具体原理为：

第一步，计算词频；

$$词频（TF）= 某个词在文章中出现的次数$$

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化。

$$词频（TF）= \frac{某个词在文章中的出现次数}{文章的总词数}$$

或者 $$词频（TF）=\frac{某词在文章中的出现次数}{该文章出现次数最多的词出现的次数}$$

第二步，计算逆文档频率；

这时，需要一个语料库（corpus），用来模拟语言的使用环境。

$$逆文频率（IDF）=\log（\frac{语言库中的文档总数}{包含该词的文档数}）$$

如果一个词越常见，那么分母就越大，逆文档频率就越小越接近 0。分母之所以要加 1，是为了避免分母为 0（即所有文档都不包含该词）。log 表示对取到的值取对数。

第三步，计算 TF-IDF；

$$TF - IDF = 词频（TF）\times 逆文档频率（IDF）$$

可以看出，TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。所以，自动提取关键词的算法就很清楚了，计算出文档中每个词的 TF-IDF 值，然后按照降序排列，去排在最前面的几个词，组成向量储存在文档中。

2.1.3 群众留言类型分类

根据 2.1.2 生成的 TF-IDF 权重向量[5]，对事件类型进行分类。主要采用 K-means 算法[6]把事件类型分为几类。

K-means 是一种最常见的聚类算法。算法输入为一个样本集或者为一点集，通过该算法可以将样本进行聚类，具有相似特征的样本聚为一类。针对每一个点，计算这个点距离所有中心点最近的那个中心点，然后将这个点归为这个中心点代表的簇。一次迭代结束之后，针对每一个簇类，重新计算中心点，然后针对每个点，重新寻找距离自己最近的中心点。如此循环，直到前后两次迭代的簇类没有变化。

k-means 算法是将样本聚类成 k 个簇（cluster），即随机选取 K 个聚类

质心点(cluster centroids)为 $\mu_1,\mu_2,\mu_3.....\mu_k \in R^n$。重复下面过程直到收敛，对于每一个样例 i，计算其应该属于的类，公式为；

$$c^{(i)} = \arg\min \| x^{(i)} - \mu_j \|$$

对于每一个类 j,计算该类的质心；

$$u_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^i}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

还要证明的是 K-means 完全可以保证收敛性。下面我们定性的描述一下收敛性，我们定义畸变函数（distortion function）如下：

$$J(c,\ \mu) = \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c(i)} \right\|^2$$

J 函数表示每个样本点到其质心的距离平方和。

这个算法的关键点是：找到距离自己最近的中心点和更新中心点。

这种算法的基本步骤为

第一步：选定要聚类的类别数目 k，选择 k 个中心点。

第二步：针对每个样本点，找到距离其最近的中心点（寻找组织），距离同一中心点最近的点为一个类，这样完成了一次聚类。

第三步：判断聚类前后的样本点的类别情况是否相同，如果相同，则算法终止，否则进入第四步。

第四步：针对每个类别中的样本点，计算这些样本点的中心点，当做该类的新的中心点，继续第二步。

2.1.4 KNN 最邻近分类算法[7]

由 K-Means 分类得到聚类中心，利用 Knn 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。根据向量空间模型，将每一类别文本训练后得到该类别的中心向量记为 $C_j(W_1,W_{2,}.....,W_n)$,将待分类文本 T 表示成 n 维向量的形式 T（$W_1,W_{2,}.....,W_n$），则文本内容被形式化为特征空间中的加权特征向量，即 D=D($T_1,W_1;T_2,W_2;.....T_n,W_n$),对于一个测试文本，计算它与训练样本集中每个文

本的相似度，找出 K 个最相似的文本，根据加权距离和判断测试文本所属的类别，具体算法步骤如下：

(1) 对于一个测试文本，根据特征词形成测试文本向量。

(2) 计算该测试文本与训练集中每个文本的文本相似度，计算公式为：

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

在该式中，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 值的确定一般先采用一个初始值，然后根据实验测试 k 的结果来调整 k 值。

(3) 按照文本相似度，在训练文本集中选出与测试文本最相似的 k 个文本。

(4) 在测试文本的 k 个近邻中，以此计算每类的权重，计算公式如下：

$$P(X,C_j)=\begin{cases}1, \text{ 若 } \sum_{d\in knn}Sim(x,d_i)y(d_i,C_j)-b\geq 0\\ 0, \text{其他}\end{cases}$$

在该式中，x 为测试文本的特征向量；Sim(x, d_i)为相似度计算公式；b 为阈值，有待于优化选择；而 y（d_i,C_j）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) 比较类的权重，将文本分到权重最大的那个类别中。

2.2 问题 2 分析方法与过程

2.2.1 问题 2 流程图

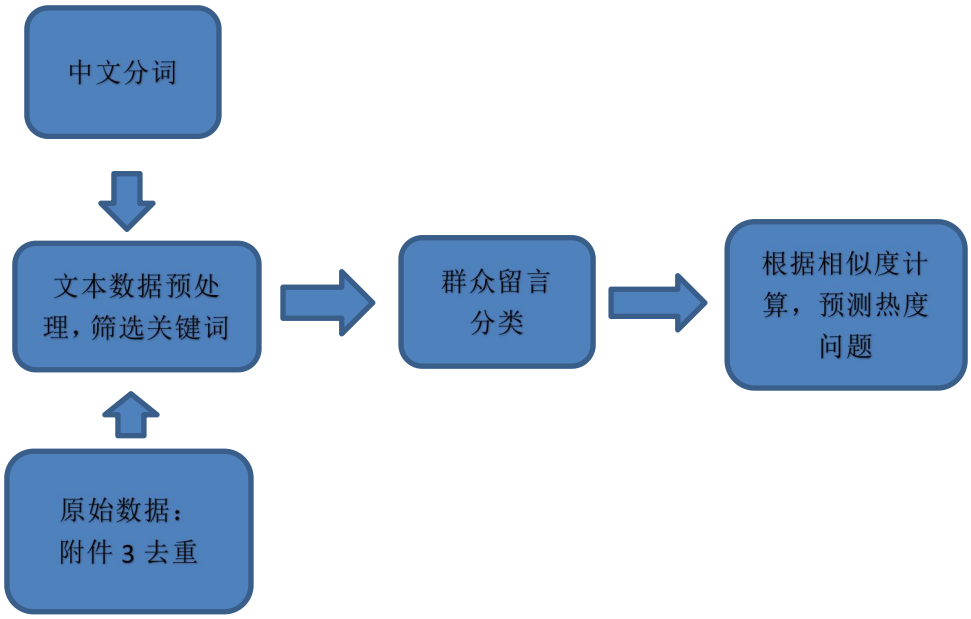

图 4：问题 2 流程图

2.2.2 数据预处理及对照筛选分析

2.2.2.1 数据预处理

（一）附件 3 去重[8]、去标点符号

对于附件 3：留言详情描述，存在大量空行和标点符号，先去除。

（二）对群众留言描述信息进行中文分词

由于中文文本的特点是词与词之间没有明显的界限，从留言描述中提取词语时需要分词，本文主要采用 Python 开发的一个中文分词模块——jieba 分词，对附件 3 中的每一条留言信息描述进行中文分词，jieba 分词用到的算法。Jieba 分词系统提供分词、词性标注、未登录词识别，支持用户自定义词典，关键词提取等功能。

附件 3 留言详情部分分词结果示例如图：

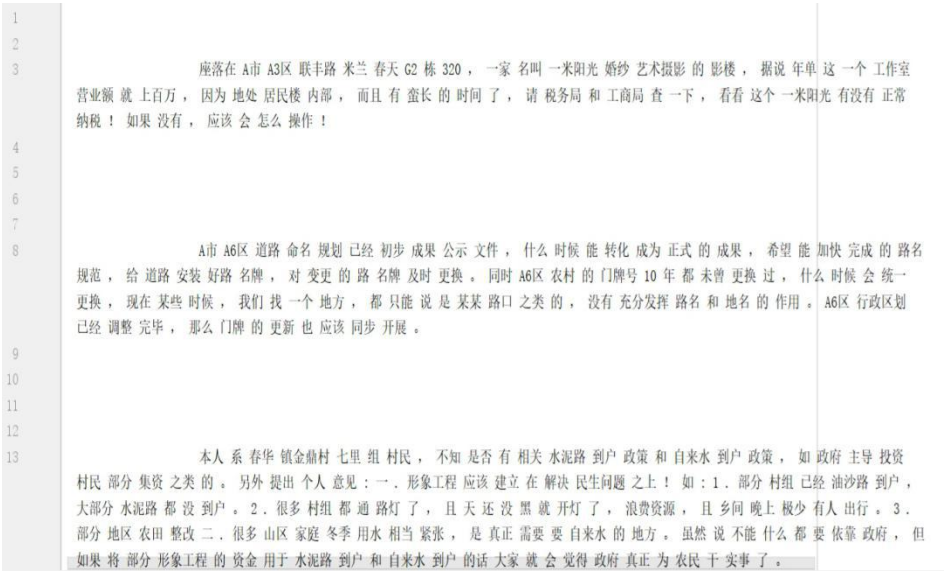

图 5：部分分词结果示例图

（三）停用词过滤

为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words（停用词）。为了分析文本数据和构建 NLP 模型，这些停用词可能对构成文档的意义没有太多价值。停用词有两个特征:一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。

为了找出这些停用词，需要一些标准估计词的有效性。而高频词通常与高噪声值具有相关性。词语在各个文档中出现的频率可以作为-一个噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应被看作是噪声词。因此用以下指标衡量词语的有效性:

1、词频(TF)

TF 是一种简单的评估函数，其值为训练集合中此单词发生的词频数。TF 评估函数的理论假设是当一个词在大量出现时，通常被认为是噪声词。

2、文档频数(DF)

DF 同样是一种简单的评估函数，其值为训练集合中包含此单词的文本数。DF 评估函数的理论假设是当一个词在大量文档中出现时，这个词通常被认为是嗓声词。

去除停用词后的部分结果示例如图：

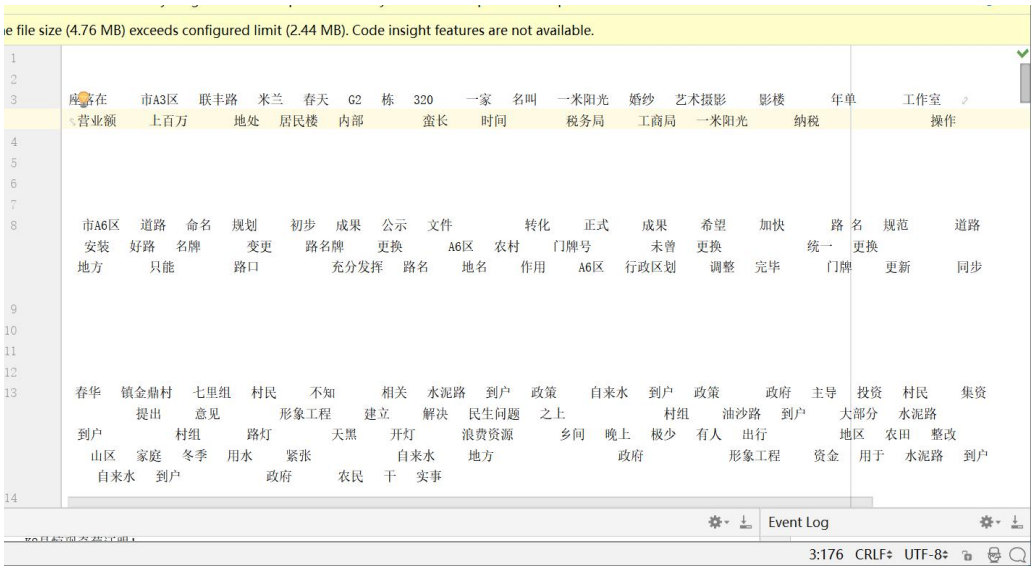

图 6：去除停用词后的部分结果示例图

2.2.2.2 数据对照筛选分析

（1）根据群众留言信息表（附件 3）结合上述方法进行分类筛选；

（2）对各条留言相似关键词出现的频率进行计算，通过排序得出排名前 5 的热点问题,从而得到“热点问题表.xls”；

（3）参照“热点问题表.xls”对应留言信息以及比对附件 3，得到“热点问题留言明细表.xls”。

2.2.3 构建热度评价指标体系[9]

1、构建原则

指标体系的构建原则通常根据要求和对象的不同分为三个层面:指标选取层面一般采取客观性原则、系统性原则和敏感性原则。客观性原则是指指标体系的选择必须从客观实际出发，全面准确地反映微博的热度情况，克服因人而异的主观因素的影响。系统性原则是指指标体系的设计应从系约整体出发,能够包络形成热点问题的各个因子，各指标间既相互独立又相互联系,共同构成一个有机整体。计算与操作层面一般采用数据的可得性和可操作原则，是指在设计指标体系时用较少指标反映较多的实质性内容，而且指标便于收集和量化。针对热点问题

这一特殊研究对象,在构建热点问题热度评价指标体系的过程中，除遵循以上基本性原则外，新增了趋势性原则和导向型原则。群众留言关注的问题热度是一个时刻变化的指标，趋势性原则就是体现热点问题的变化趋势;导向型原则是指该套指标体系的构建不仅要对留言热度进行检测，更是要为判断热点问题提供方向指导。

2、构建理论依据

留言问题热度是指某一时间地点反映某一事件信息，并引起群众对该信息的广泛关注和讨论的热烈程度，其实质上是一种信息传播活动。根据流行三要素和新闻传播学理论，本文拟从留言发布者、留言详情、留言内容特征和群众反映四个维度选取留言内容特征热度影响进行定量评价。

留言内容特征热度影响力。根据流行三要索理论可知群众留言内容特征对其热度具有很强的影响力。本文从群众留言内容和表现形式两个角度,选取留言信息充实度、留言内容出现及时率和充实度三个二级指标进行反映。留言信息充实度用留言的长度来反映，留言字数越多其表达出的信息越充实,也越容易引发讨论。留言出现及时性主要反映留言出现的新鲜程度，一手信息所受到的关注要多过二手信息。

据此构建出的留言热度评价指标体系，体系内涵如下：

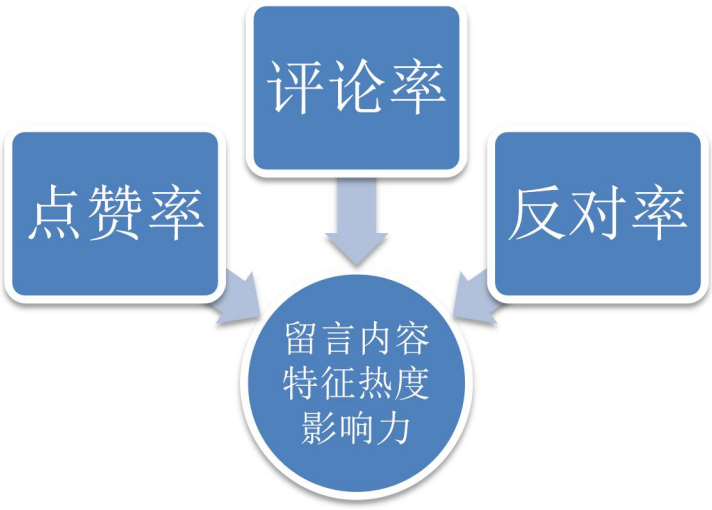

图 7:留言热度评价指标体系

2.3 问题 3 分析方法与过程

科学、全面、客观、系统的评价政府相关部门对市民留言的答复意见，是政府政务评估考核的重要环节。针对这一工作我们可尝试构建一个与此相关的答复意见评价指标体系，使政务工作更为精简明确、易操作。

2.3.1 构建答复意见评价指标的流程

第一步：以市民问题留言及政府答复意见的文本为模板（即附件 4），对该文本进行文本数据预处理，提取并整理初步的评价指标。

（1）利用 python 软件中 jieba 里的 TF-IDF 提取关键词方法，对答复意见文本进行读取、分词、去停用词、TF-IDF 关键词提取，提取了该文本中的前 1000 个关键词并降序排列。

取用结果中的前 100 个高频词汇（代码于附件文档问题三代码中），

结果如下：

收悉 回复 网友 问题 工作 进行 留言 反映 情况 相关 建设 西地省 支持 小区 有关 项目 业主 办理 监督 街道 学校 调查 要求 来信 理解 部门 规划 规定 管理 社区 关心 单位 处理 教育局 咨询 人员 问政 申请 公司 组织 学生 现场 投诉 政策 整改 施工 核实 调查核实 房屋 服务 道路 标准 教育 工作人员 收费 改造 答复 车辆 居民 登记 文件 参保 手续 垃圾 住房 教师 建议 政府 安置 加强 执法 城区 开发商 意见 依法 方案 缴费 存在 村民 安排 群众 城乡居民 符合 物业公司 中心 通知 提供 完成 招生 审批 联系 工程 关注 经营 部分 有限公司 农产品 解决 小学 基本 路段

并对这些词汇进行筛选，归类，如进行、反映、情况、相关、有关、核实、核实情况等可以将其视为对留言情况的调查，将建议、意见、方案、解决归纳为具体处理方案.....

（2）抽取文件中部分数据，进行人工查看进一步直观了解数据。从数据中概括关键词汇，关键点进行分类整合。如您好、感谢您、据查、相关、相关情况等

（3）综合前两点可以概括出回复数据中的关键点为：收到答复的时间、调查问题的情况、问题解决进度、具体处理意见的方案、礼貌用语、语言简洁程度、答

复的可实行度、答复可信度。

第二步：筛选评价指标并分类。

筛选的目的是为了剔除不合理指标，分类的目的是为了整合关联性强的指标。经过筛选和分类，最终得到比较合适的指标。根据关键点进行分类概括，剔除对评价影响不大的关键点： 答复可实行度、答复可信度，因为答复是由相关部门进行专业调查根据现行政策、法规给出的回复，所以答复的可实行度和可信度都较高，此项可忽略。余下的定为评价指标，根据评价指标间的相关性进行分类，比如收到答复的时间、答复中表示的问题解决进度可以归类为时效性，语言简短明确与礼貌用语可归类为语言运用，留言情况调查和具体处理方案可归类为答复的完整性。

第三步：确定评价指标体系。

初步分类后，根据评价指标的性质再综合归类，最终得到一级指标有 2 项，二级指标有 3 项，三级指标有 6 项。

一级指标包括：答复质量、答复时效，二级指标包括：语言运用、完整性、时效性 ，三级指标包括：语言简洁程度、礼貌用语、留言情况调查、具体处理方案、收到答复时间、问题解决进度。

参考神经网络法构建的答复意见评价指标如下图所示：

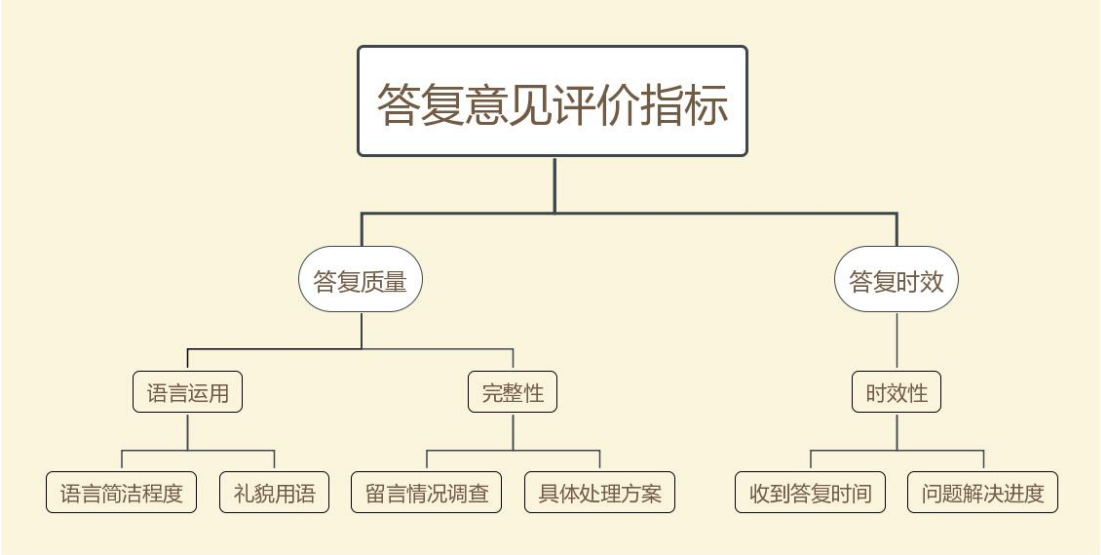

图 8：答复意见评价指标图

2.3.2 运用层次分析法[10]对答复意见评价指标进行分析测试

层次分析法，简称 AHP，是美国运筹学家萨蒂在 20 世纪 70 年代提出的一种实用的定性与定量分析相结合的系统分析和评价方法。将一个复杂的多目标决策问题作为一个系统，把目标分解为多个目标或准则，进而分解为多指标的若干层次。通过定性指标量化方法将难以量化的各种方案定量化，算出层次单排序（权数）和总排序，以作为目标、多方案优化决策的系统方法。它适用于对一组答复意见进行评价。

第一步：根据问题的性质和要求，提出一个总的目标。这一层次中只有一个元素，一般是分析问题的预定目标或理想结果，因此也称为目标层。将目标逐层分解为几个层次，建立层次结构模型。

第二步：构建准则层。这一层次中包含为实现目标所涉及的中间环节，该层次可以由若干个层次组成，包括所需考虑的准则、子准则。对同一层次的各元素关于上一层次某一准则的重要性进行两两比较并赋权，构造成对比矩阵。

设某层有 n 个因素 $x=\{X_1,X_2,...,X_n\}$，要比较其对上一层某一准则（或目标）的影响程度，确定在该层中相对于某一准则所占的比重。上述比较是两两因素之间的比较，比较时取 1-9 尺度。用 b_{ij} 表示第 i 个因素相对于第 j 个因素的比较结果，则：

$$A=(a_{ij})_{n\times n}=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

，A 称为成对比较矩阵。

根据比较的影响程度得出下列两两比较表格如下：

量化值 a_{ij}	因素 i 比因素 j
1	影响程度相同
3	影响程度稍强
5	影响程度强
7	影响程度明显强

9	影响程度绝对强
2、4、6、8	影响介于上述相邻等级之间
$1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\ldots,\frac{1}{9}$	b_j 与 b_i 之比 $a_{ij}=\frac{1}{a_{ij}}$

图 9：比较表

第三步：针对某一准则，计算各方面的相对权重（层次单排序），并进行一次性检验。定义一次性指标为 $\mathrm{CI}=\frac{\lambda-\mathrm{n}}{\mathrm{n}-1}$ 其中 λ 为最大特征根，n 为矩阵阶数。

随机一致性指标 RI 的数据如下：

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

图 10：随机一致性指标 RI 的数据表

当一致性比率 $\mathrm{CR}=\frac{\mathrm{CI}}{\mathrm{RI}}\langle 0.1$ 时，检验通过，可用其归一化特征向量作为权向量 w。

第四步：计算各层元素对总目标的合成权重，进行层次总排序。

设 B 层 m 个因素 $\mathrm{B}_1,\mathrm{B}_2,\cdots,\ \mathrm{B}_\mathrm{m}$，对总目标 A 的排序为 $b_1,b_2,\cdots,b_m$，C 层 n 个因素对上层 B 中因素为 B_j 的层次单排序为 $c_{1j},c_{2j},\cdots,c_{nj}$，　j=1,2，...，n

C 层的层次总排序为：

$$\mathrm{C}_1=b_1c_{11}+b_2c_{12}+\cdots+\mathrm{b}_\mathrm{m}c_{1m}$$

$$\vdots$$

$$\mathrm{C}_n=b_1c_{n1}+b_2c_{n2}+\cdots+b_mc_{nm}$$

即 C 层第 i 个因素对总目标的权值 W 为 $\sum_{j=1}^{m}b_jc_{ij}$

3、结果分析

3.1 问题 1 结果分析

3.1.1 文本的向量化表示结果

相关操作的详细程序见附件，图 11、图 12 是部分结果显示：

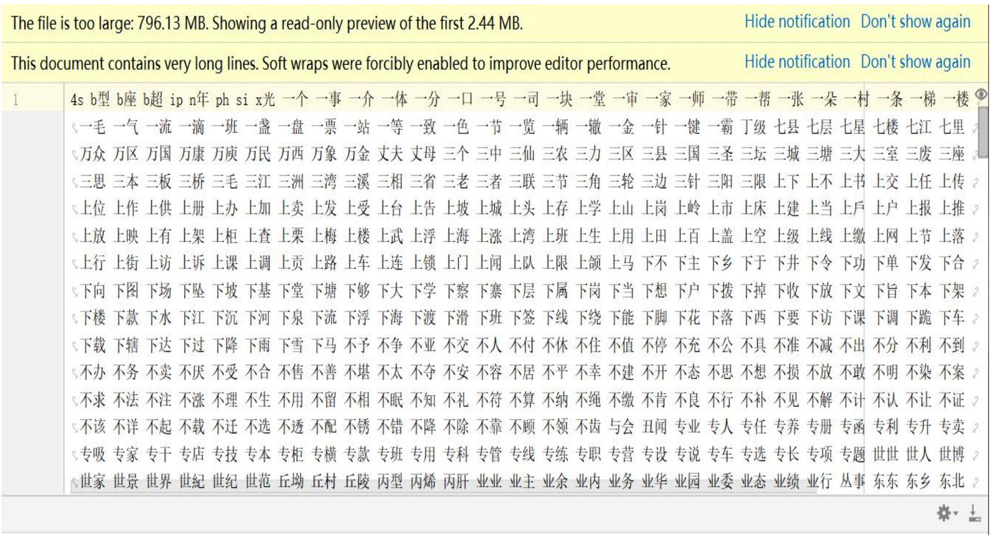

图 11：词汇-文本词频矩阵

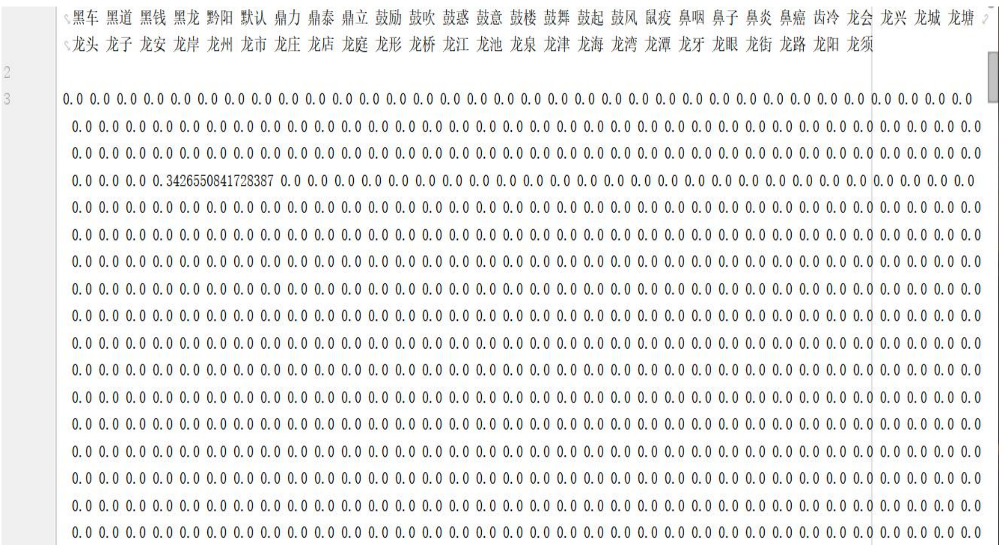

图 12：词汇-文本 TF-IDF 权重矩阵

3.1.2 K-means 聚类结果

通过去重后对文本进行分词，对每条留言详情描述每一段取关键词后由 K-means 分类得到聚类中心，然后利用 KNN 算法找出离七个聚类中心最近的前 5 个元素，根据“少数服从多数”判定聚类中心所属类别（程序见附件）。

部分聚类结果如下表所示：

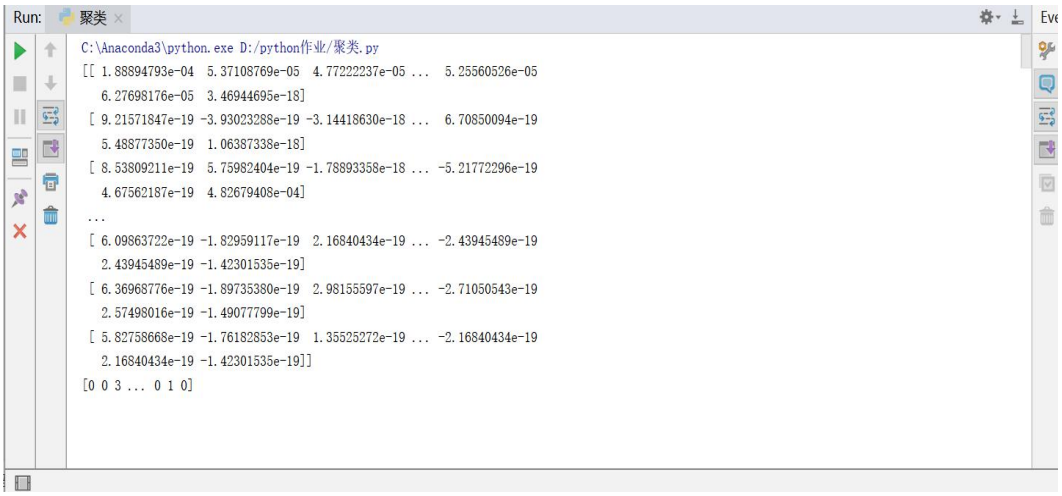

图 13：部分聚类结果图

其中 KNN 最邻近分类算法大致步骤如下：

1、算距离：给定聚类中心，计算它与样本中的每个 TF-IDF 权重向量的距离

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 5 个近邻归属主要类别，来对聚类中心进行分类

结合样本，分别找出七个聚类中心的 5 个近邻样本点所属的留言类型（程序见附件），结果如下表所示：

聚类中心	城乡建设	劳动和社会保障	教育文体	商贸旅游	环境保护	所属类型
第一个	3	2	0	0	0	城乡建设
第二个	0	1	3	0	0	教育文体
第三个	0	4	0	1	0	劳动和社会保障
第四个	1	0	0	3	0	商贸旅游
第五个	0	1	1	0	3	环境保护

第六个	4	1	0	0	0	城乡建设
第七个	3	2	0	0	0	城乡建设

图 14：KNN 分类表

从 KNN 分类表中可以看出：七个聚类中心可大致分为：城乡建设、劳动和社会保障、教育文体、商贸旅游、环境保护五个大类。

3.1.3 群众留言分类

结合上述分类统计，可以得到群众留言分类汇总图：

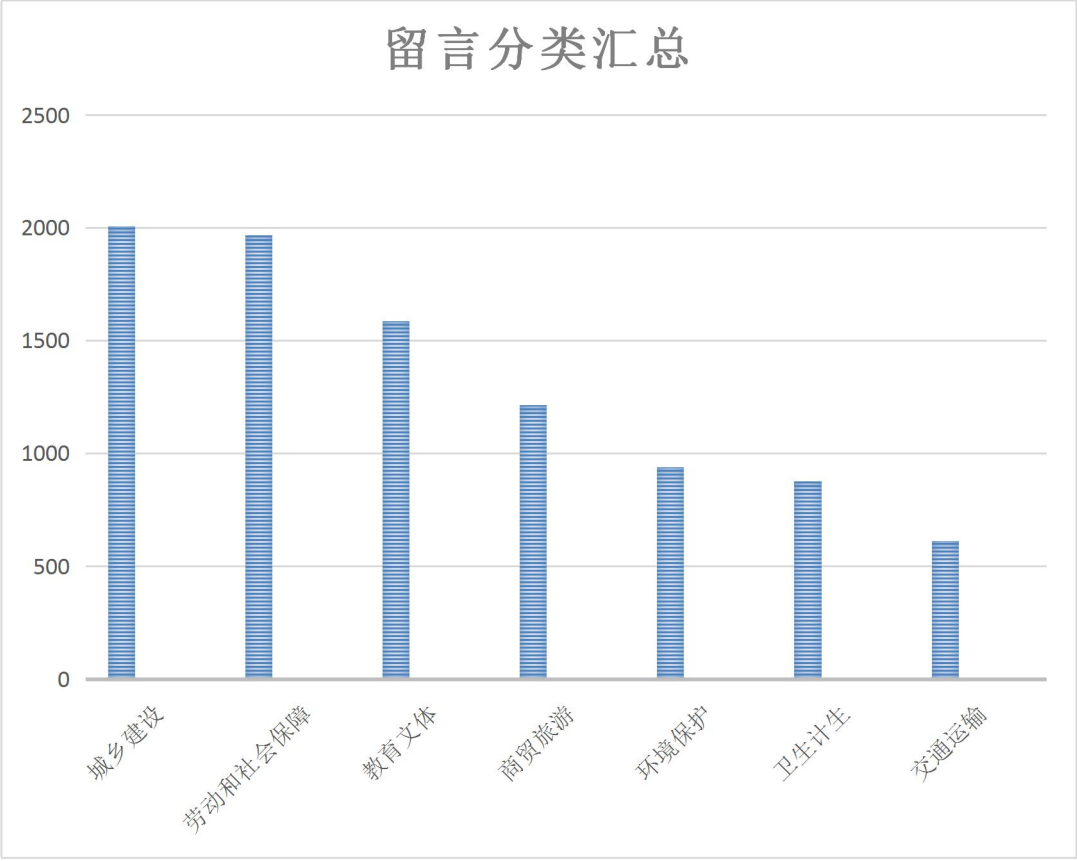

图 15：留言大类汇总图

由上表可以看出，群众留言问题集中在城乡建设、劳动和社会保障、教育文体等，这说明城乡建设是社会发展进步的特别关注问题，而民生基本保障是促使一切发展的前提，其次，教育为本，教育问题也是群众反映较多的。

3.2 问题 2 结果分析

3.2.1 对群众留言的分类

通过对每条群众留言所属大类的 15 个类别进行分类技术，得到交通运输类留言是最多的，其中交通运输类问题最突出的就是建设管理方面问题;第二是环境保护类留言，着重反映的是环境污染，涉及噪音扰民，油烟乱排等系列问题;第三是纪检监察类留言，强调干部作风，案件处理进度让民众堪忧；第四是城乡建设类留言，包括旧城改造、违章建筑、规划设计、安置补偿等，是贴近整个城市建设，值得关注的问题；再次是教育文体类留言问题，考虑到教育收费、学生负担，校园安全等问题.....

留言大类统计图如下：

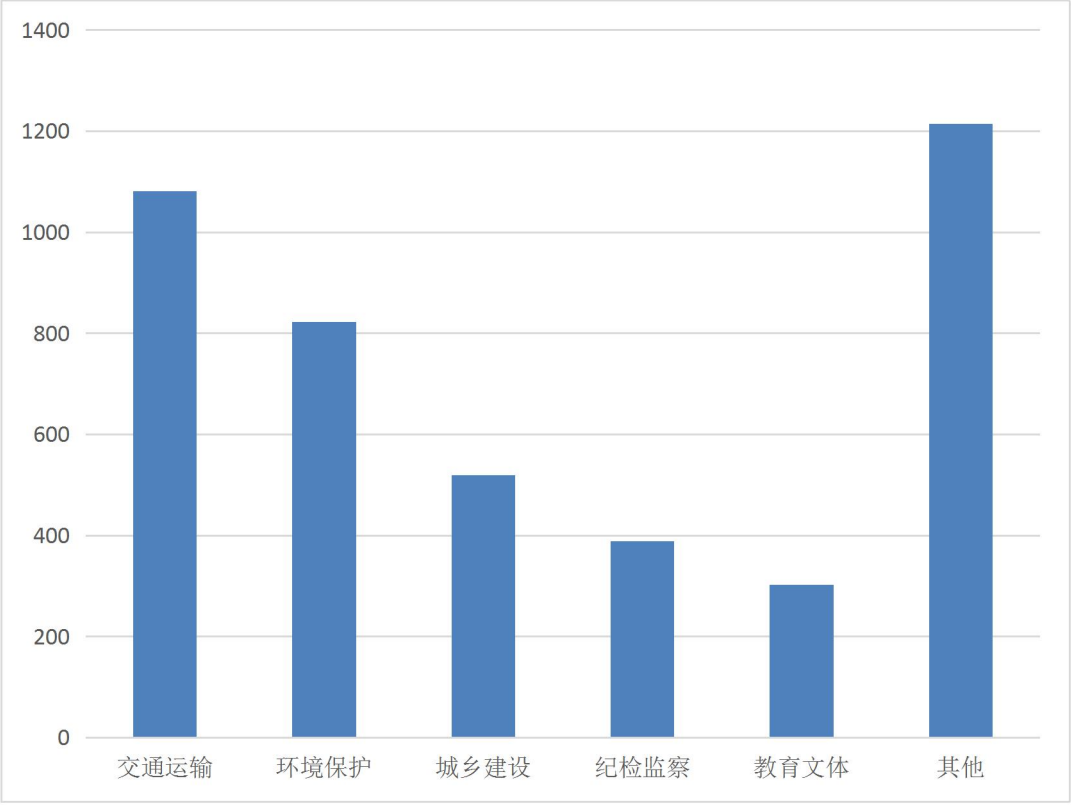

图 16：留言大类统计图

3.2.2 对热点问题的分析

通过分类之后，确定留言问题地点对各反映问题进行相似度比较（文本相似度见附件），综合排序计数，分析出在众多群众留言中排名前 5 的热点问题。

前 5 个热点问题占比如下：

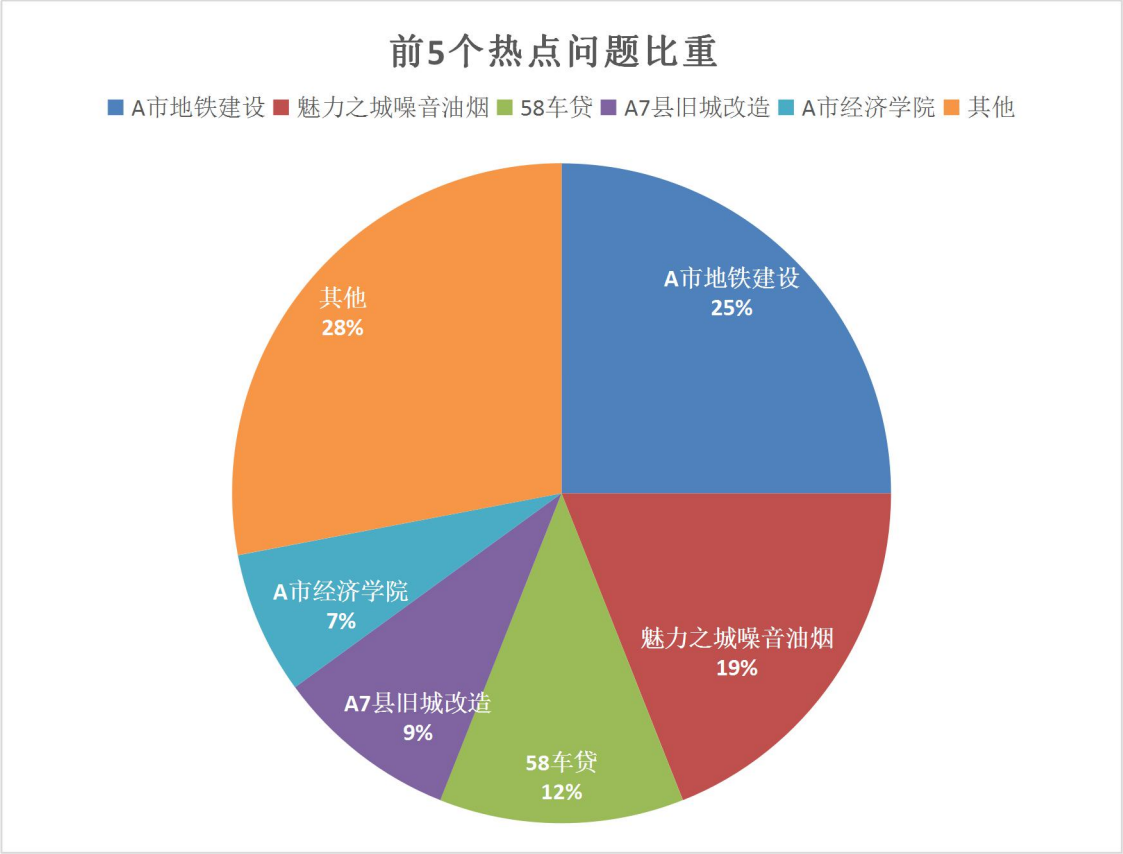

图 17：前 5 个热点问题比重

由上述图表结合“热点问题留言明细表.xls”可以看出，群众所关注的问题都是围绕着我们的生活圈。比如交通出行是否便利，怎样便利，以及建设施工的诸多问题及有效建议。而顺应时代的进步与发展，环境问题也日益突出，噪声污染，油烟排放各种扰民问题层出不穷。城乡建设是城乡管理的重要组成部分，小到群众住房，再到拆迁安置，大到规划重建，城乡布局等，群众留言的反映表明城乡建设是推动经济健康发展的保障，也是实现人民美好生活的载体。教育也是并且长期是一个值得关注的热点问题，收费合理与否，实习安排如何，就业问题等等.....总之，群众留言是一种良好揭露社会问题的方式，给党和政府了解民意，掌握舆情，及时沟通解决提供了便利。

3.2.3 热点问题热度评价指标

为进一步验证热点问题评价指标体系的合理性与客观性，从群众留言中筛选出的前五个热点问题，并按照留言热度评价指标体系所涉指标对五个热点问题进行数据采集，原始数据见下表：

热点问题	点赞率	反对率	评论率
A 市地铁建设	4%	0.3%	0
魅力之城噪音油烟	0.6%	0.6%	0
58 车贷	93%	0	0
A7 县旧城改造	0.9%	0.04%	0
A 市经济学院	0.04%	0.5%	0

图 18：原始数据表

由此观之，相较于非热点问题，这些问题具有相应的代表性。特别是点赞率极高的 A 市西地省 58 车贷案件处理问题，而结合“热点问题留言明细表.xls”可知此问题群众反响极大，引起社会关注。故此评价指标还是有一定的可行性。

3.3 问题 3 结果分析

实例分析[11]：

随机抽取文本中三组市民问题留言后的政府答复意见数据，利用所建的答复意见评价指标体系进行评价。抽取的数据如下所示：

甲组数据： 网友“A000100804”：您好！针对您反映 A3 区教师村小区盼望早日安装电梯的问题,A3 区住建局高度重视，立即组织精干力量调查处理，现回复如下：为了完善住宅使用功能，提高我区既有多层住宅居民的宜居水平，2018 年 6 月 7 日，A 市 A3 区人民政府办公室下发了《关于 A 市 A3 区既有多层住宅增设电梯实施方案》的通知。该方案明确了增设电梯条件及流程。详情可咨询政务服务中心住建局受理申请老旧小区加装电梯窗口，咨询电话：0000-00000000。感谢您对我们工作的关心、监督与支持。2019 年 4 月 2 日

留言时间：2019/3/29 11:53:23　　答复时间 2019/5/9 10:18:58

乙组数据： 网友“UU0081227”您好！您的留言已收悉。现将有关情况回复如下：261 路公交车全程 24 公里，配车 20 台，高峰期发车间距为 7-8 分钟/趟，

平峰为10-15分钟/趟，经查看近期发车时刻表，其发车间隔正常。由于驾驶员工作时间长，劳动强度大，造成车队驾驶员短缺，公司正在积极组织调配人员充实该线路运力，公司人事部正在积极进行驾驶员招募工作，条件具备后将增加该线路配车。感谢您对我们工作的支持、理解与监督！2019年1月8日

留言时间：2018/12/28 7:53:25　　答复时间：2019/1/14 14:33:17

丙组数据：网友“UU0081057”您好！您的留言已收悉。现将有关情况回复如下：根据对A5区石坝路马王堆路口交通标志标线实地考察后，市交警支队初步研究认为：根据《中华人民共和国道路交通安全法实施条例》第三十八条第一款第三项“红灯亮时，禁止车辆通行”之规定，左转弯信号灯为红灯时，禁止车辆越过停止线行驶。如您对处罚结果有异议，可向当地交警部门申请复议。感谢您对我们工作的支持、理解与监督！2019年1月9日

留言时间：2018/12/25 13:56:31　答复时间：2019/1/16 15:22:16

根据数据可得知：甲组答复意见语言较为简洁，有相关礼貌用语，留言情况调查较好，但是具体的处理方案和问题解决进度一般，收到答复时间也有41天。乙组答复意见语言较为简洁，有相应的礼貌用语，留言情况好具体处理方案和问题解决进度较好，收到答复时间较短仅为17天。丙组答复意见语言较为简洁，也有礼貌用语，留言情况调查和问题解决进度好，具体处理方案一般，收到答复时间为22天。

以下是评价排序答复意见时需要参考的六个标准：B1 语言简洁程度 、B2 礼貌用语、B3 留言情况调查、B4 具体处理方案、B5 收到答复时间、B6 问题解决进度。

三个需要评价排序的答复意见：C1 甲组、C2 乙组、C3 丙组

1、建立层次结构模型：上述六个标准都是在评价答复意见时应考虑的因素，其目的是要根据所建的答复意见评价指标体系，对随机抽取的甲乙丙三组答复意见进行评价并排序。据此可以建立的层次分析结构模型如下图：

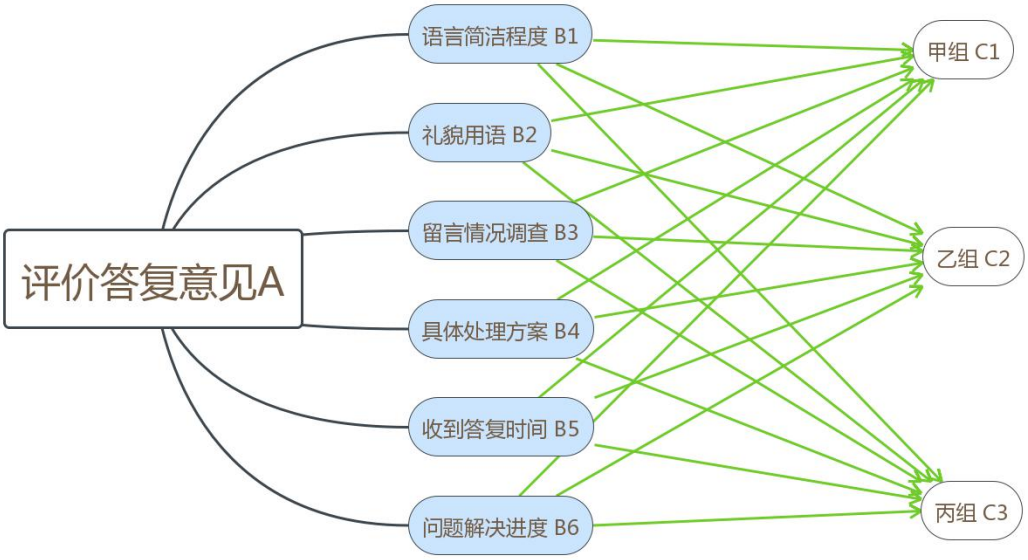

图 19：层次分析结构模型

2、通过两两比较准则层六个因素对目标层的影响程度、方案层对准则层的影响程度，并将其比较结果进行量化得到比较矩阵。构造出的成对比较矩阵如下表：

表一：

A	B1	B2	B3	B4	B5	B6
B1	1	2	1/7	1/7	1/3	1/5
B2	1/2	1	1/5	1/5	1/3	1/5
B3	7	5	1	1/3	3	1/3
B4	7	5	3	1	3	1/3
B5	3	3	1/3	1/3	1	1/5
B6	5	5	3	3	5	1

表二：

B1	C1	C2	C3
C1	1	1	1
C2	1	1	1
C3	1	1	1

表三：

B2	C1	C2	C3
C1	1	1	1
C2	1	1	1
C3	1	1	1

表四：

B3	C1	C2	C3
C1	1	1/4	1/4
C2	4	1	1/2
C3	4	2	1

表五：

B4	C1	C2	C3
C1	1	1/4	1/2
C2	4	1	3
C3	2	1/3	1

表六：

B5	C1	C2	C3
C1	1	1/7	1/5
C2	7	1	3
C3	5	1/3	1

表七：

B6	C1	C2	C3
C1	1	1/4	1/7
C2	4	1	1/3
C3	7	3	1

2、层次单排序，进行一致性检验：

用最大特征根对应的特征向量归一化后作为权向量 w，由 MATLAB 软件编程计算（所使用代码于附件文档问题三所使用代码中）结果如下：

表一：b ={0.0874 ， 0.0765 ， 0.3475 ， 0.5055 ， 0.1680 ， 0.7629}

CI = 0.1058　　CR =0.0853 <0.1　　$\lambda_{\max} = 6.5290$

表二、表三：c1、c2={0.4082 ， 0.4082 ， -0.8165}

CI =-2.2204e-16　　CR = -3.8284e-16 <0.1　　$\lambda_{\max} = 3.0000$

表四：c3={0.1656 ， 0.5257 ， 0.8344}

CI = 0.0268　　CR =0.0462 <0.1　　$\lambda_{\max} = 3.0536$

表五：c4={0.1999 ， 0.9154 ， 0.3493}

CI = 0.0091　　CR = 0.0158 <0.1　　$\lambda_{max}=3.0183$

表六：c5={0.1013 ， 0.9140 ， 0.3928}

CI = 0.0324　　CR =0.0559 <0.1　　$\lambda_{max}=3.0649$

表七：c6={0.1102 ， 0.3683 ， 0.9232 }

CI = 0.0162　　CR =0.0279 <0.1　　$\lambda_{max}=3.0324$

由 CR < 0.1 可知所有矩阵通过一致性检验。经过 MATLAB 软件编程计算将特征向量归一化得到相对应的权向量 W。由于表二、表三的矩阵数值都为一，所以可以忽略这两个比较项。（所使用代码于附件文档问题三所使用代码中）运算结果如下：

Wb ={0.0449　0.0393　0.1784　0.2595　0.0863　0.3917}

Wc3={0.1085　0.3446　0.5469}

Wc4 ={ 0.1365　0.6250　0.2385}

Wc5 = {0.0719　0.6491　0.2790}

Wc6 = {0.0786　0.2628　0.6586}

4、计算合成权重，进行层次总排序：

$$0.1784\begin{bmatrix}0.1085\\0.3446\\0.5469\end{bmatrix}+0.2595\begin{bmatrix}0.1365\\0.6250\\0.2385\end{bmatrix}+0.0863\begin{bmatrix}0.0719\\0.6491\\0.2790\end{bmatrix}+0.3917\begin{bmatrix}0.0786\\0.2628\\0.6586\end{bmatrix}$$

={0.09177074,0.38262023,0.44150903}

由该结果可知，甲乙丙三组的答复意见根据答复意见评价指标得出的评价为丙组答复最好，乙组其次，甲组稍差。

4、结论

总结本次数据挖掘赛，对“智慧政务”中信息的分析和挖掘，了解群众和社会相关问题，这对政府管理部门了解舆情，及时作出合理方案解决问题有重大意义，但同时这也是自然语言处理，文本挖掘的一个课题、一个难题。像传统的一些机器文本解读已经无法满足如此庞大的数据量信息的挖掘。而本文主要基于TF-IDF 权重法提取特征词，构造词汇-文本矩阵，再进一步通过 K-means 聚类算法和 KNN 最邻近分类算法，统计分析群众留言类别。

由分析结果可以看出，群众留言涵盖多个种类，涉及面广，有城乡建设、环境保护、交通运输、教育文体.....基于这些分类，我们采用相似度计算再次挖掘出群众留言所关注的热点问题。根据这些热点问题更进一步看出，在了解民生，处理社会问题时，群众留言极为重要。随着社会的进步，各种各样的问题也随之涌现，人们在享受当下生活的同时，对生活的要求也更加严格，希望营造出更为良好的环境以及氛围。及时倾听群众的声音，了解充分，有效沟通和解决是当下所有人的期望。

但反思所有流程，我们在数据细节处理上还是有所不足，最后得到的聚类结果准确度并不是那么高，与附件相比还是略有出入。通过查资料和共同探讨，推测可能是由于 k 均值算法在计算欧式距离时有一定误差。而资料上表明也可以构造基于文本相似度的聚类算法，这也涉及我们此文本挖掘模型的不足，后期仍需进一步翻阅资料，实践探讨。

5、参考文献

[1]周雄伟.python 中 jieba 进行中文分词.2018

[2]高扬.基于 LDA 主题模型的 TF-TDF 算法改进及应用.广西大学.硕士论文.2015

[3]cv_ml_dp，问本数据的向量化(TF-IDF)---样本集实例讲解+python 实现.2018

[4]ogghansi,基于 TF-IDF 对文本向量化.2017

[5]杨秀璋，基于 k-means 和 tfidf 的文本聚类代码的简单实现.2016

[6]翟东海，鱼江，高飞，于磊等.最大距离法选取初始簇中心的 K-means 文本聚类算法的研究.西南交通大学.2014

[7]张欣.KNN 算法原理解释.2018

[8]杨虎.面向海量短文文本去重技术的研究与实现.国防科学技术大学.2007

[9]梁昌明，李东强.基于新浪热门平台的微博热度评价指标体系实证研究.2015

[10]谢金星，姜启源，叶俊. 数学建模（第三版）高等教育出版社.2003

[11]濮长飞 .层次分析法应用于城市购房决策中的实例分析.2019