“智慧政务”中的文本挖掘应用

摘要

近年来，随着市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、解决民生的重要渠道，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。因此运用大数据、 云计算、人工智能等技术，建立基于自然语言处理技术的智慧政务系统对于改进社会治理、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 1，按照奇数偶数的顺序将附件 2 留言详情的内容分为训练集和测试集，随后采用 TextRank 算法通过自定义句子权重从留言详情中提取关键句并去除掉停用词和非中文字符。利用 jieba 中文分词工具对留言详情进行分词，同时利用 TF-IDF 算法将文本的词语向量化，使用朴素贝叶斯分类器对分词后的训练集进行文本标签分类训练，使用训练后的模型对测试集进行预测并根据 F-Score 的评价方法对分类结果进行评分，结果为 0.992，说明预测效果较好。

对于问题 2，首先对附件 3 的留言主题进行文本清理，根据问题加入了自定义的字典，从而保留关键地名。采用 jieba 里面自带的 pseg.cut 函数，进而有效识别留言主题中的地名、人名、机构团体名和动词等关键信息，继续计算分词后的留言主题的 TF-IDF 权重矩阵从而得到一个向量矩阵。接着小组更新了相关词性词语的权重，用于 kmeans 聚类。根据轮廓系数确定了合适 k 值为 65，即将留言主题分为 65 类，并通过四个重要指标：话题持续类内留言详情数、类内留言详情与主题相似度以及留言点赞反对数建立的熵值法综合评价模型，对 65 类问题进行热度评分排名并得到了排名前五的热点问题。最后根据隐狄利克雷分布（LDA）提取热点问题的主题并给出热点问题表和热点问题留言详情表。

对于问题 3，在对附件 4 的相关内容进行基础的数据预处理后，对政府部门答复的内容进行量化，量化核心围绕回复的内容以及回复的效用两个维度来评价答复意见的质量。使用“回复与问题的相关度”、“回复的信息量”、“回复的专业规范性”、“回复的时效型”、“回复的可读性”这五个指标并根据层次分析法建立政府答复质量评价模型，对答复意见作出评价并且应用到具体数据中。

关键词：TextRank　jieba 分词　TF-IDF 算法　kmeans 聚类 熵值法 层次分析法

目录

1.建模目标

本次建模的目标是根据自互联网公开来源的群众问政留言记录，以及相关部门对部分群众留言的答复意见，主要利用 jieba 中文分词工具对群众留言记录进行分词，然后采用朴素贝叶斯分类器、Kmeans 文本聚类、相关评分模型从而达到以下三个目标。

- 利用文本分词和文本分类的方法对文本数据进行文本挖掘，对群众留言进行分 类，以便后续将群众留言分派至相应的职能部门处理，最后根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型并利用 F-Score 对分类方法的评进行评价并希望有较高分数。
- 根据附件 3 的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使用 Kmeans 文本聚类的方法进行文本聚类，定义合理的热度评价指标，得出排名前五的热点问题及其详情，从而帮助政府及时有效的处理群众所关心的热点问题。
- 根据附件 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合理有效的评价方案。

2.分析方法与过程

2.1 问题 1 的分析方法与过程

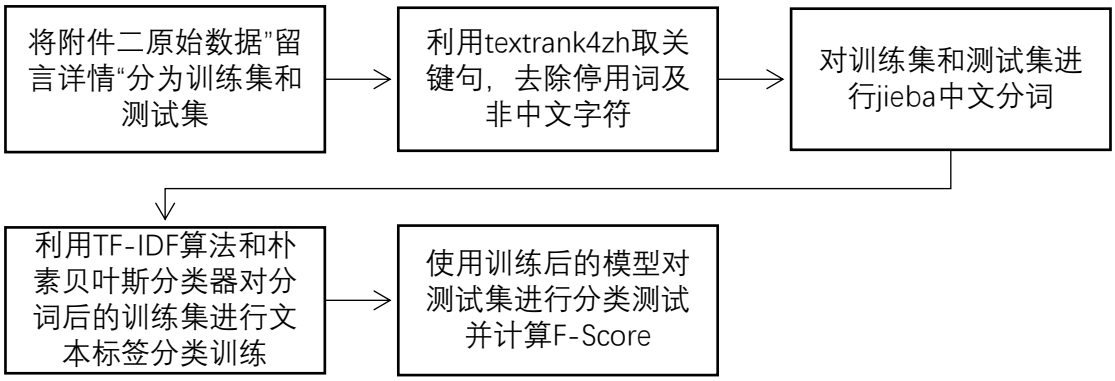

图 1.问题一分析流程图

2.1.1 数据预处理

根据题目所给出的附件二的数据，本小组分别按照奇数偶数的顺序分为训练集和测试集，因为本小组的目标是对留言进行分类，所以本小组只需要对留言详情分类即可，因为留言详情中包含大量的停用词和非中文字符，同时语句较长，所以本小组便采用 TextRank 算法取关键句并去除掉停用词和非中文字符，相关 python 代码见附件 question1.py 中[1][2]。

2.1.2 对训练集和测试集进行 Jieba 中文分词

在对预处理的留言详情进行文本分析之前，要先把非结构化的文本信息转化成计算机可识别的结构化文本信息，所以为了进行有效的中文文本分析，本小组对预处理的留言详情使用了 jieba 中文分词。本小组采用中文 jieba 分词的原因在于：jieba 的词典文件添加自定义词典很方便、jieba 词典的数量大于五千万，分词速度快、jieba 分词采用了动态规划查找最大路径并能找出基于词频的最大

切分组合。而对于未被录入的词语采用了基于汉字成词能力的 HMM 模型，能够高效的处理中文分词。

2.1.3 TF-IDF 算法实现

在对预处理的留言详情分词以后，本小组需要把这些词语转化可供分析计算的向量，于是本小组采用 TF-IDF 算法，把留言详情转为权重向量，以下是 TF-IDF 算法的原理。

（1） 计算词频也即是 TF 权重（Term Frequency）

词频（TF） = 某个词在文中出现的次数

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"标准化即：

词频（TF） = 某个词在文中出现的次数 ÷ 文章的总词数

或者：

词频(TF)=某个词在文中出现的次数 ÷ 该文出现次数最多的词的出现次数

（2） 计算逆文档频率（IDF）

本小组需要一个语料库（corpus），用来模拟语言的使用环境。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 表示对得到的值取对数。

逆文档频率（IDF）=log（语料库的文档总数÷（包含该词的文档数+1））

（3） 计算 TF-IDF

可以看到，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，就是计算出文档的每个词的 TF-IDF 值，然后按降序排列，取排在最前面的几个词。

TF-IDF=词频（TF）* 逆文档频率（IDF）

（4） 生成 TF-IDF 向量

首先使用 TF-IDF 算法找出每个留言的关键词，然后对每个留言的关键词

合并成的集合计算每个留言详情中对于这个集合中词的词频，如果没有则记为 0，生成各个留言的 TF_IDF 权重向量，计算公式即为：

2.1.4 朴素贝叶斯分类器的应用

2.1.4.1 朴素贝叶斯算法简介

朴素贝叶斯算法可以理解为贝叶斯定理 + 条件独立假设。贝叶斯定理也即是 P(A|B) = P(B|A) P(A)/P(B)；而条件独立假设指的是：解决分类问题时，会选取很多数据特征，为了降低计算复杂度，那么假设数据各个维度的特征相互独立。而算法的计算过程也可以用一句话概括：把计算“具有某特征的条件下属于某类”的概率转换成需要计算“属于某类的条件下具有某特征”的概率。

2.1.4.2 朴素贝叶斯分类算法应用步骤

第一步，本小组在 python 中决定使用的 sklearn 包的全称叫 Scikit-learn，它给本小组提供 3 个朴素贝叶斯分类算法，分别是高斯朴素贝叶斯（GaussianNB）、多项式朴素贝叶斯（MultinomialNB）和伯努利朴素贝叶斯（BernoulliNB）。本小组选用多项式朴素贝叶斯，这是因为本小组数据的特征变量是离散变量，符合多项分布，在文档分类中特征变量体现在一个单词出现的次数，或者是单词的 TF-IDF 值等。本小组将特征训练集的特征空间以及训练集对应的分类一级标签传递给贝叶斯分类器 clf，它会自动生成一个符合特征空间和对应分类的分类器。

第二步，在这里本小组采用的是多项式贝叶斯分类器，其中 alpha 为平滑参数。为什么要使用平滑呢？因为如果一个单词在训练样本中没有出现，这个单词的概率就会被计算为 0。但训练集样本只是整体的抽样情况，本小组不能因为一个事件没有观察到，就认为整个事件的概率为 0。为了解决这个问题，本小组需要做平滑处理。当 alpha=1 时，使用的是 Laplace 平滑。Laplace 平滑就是采用加 1 的方式，来统计没有出现过的单词的概率。这样当训练样本很大的时

候，加 1 得到的概率变化可以忽略不计，也同时避免了零概率的问题。当 0<alpha<1 时，使用的是 Lidstone 平滑。对于 Lidstone 平滑来说，alpha 越小，迭代次数越多，精度越高。本小组最终设置 alpha 为 0.001。

第三步，本小组使用训练集的分词创建一个 TfidfVectorizer 类，使用同样的 stop_words，然后用这个 TfidfVectorizer 类对测试集的内容进行 fit_transform 拟合，得到测试集的特征矩阵。接着本小组使用 predict 函数，传入测试集的特征矩阵，得到分类结果 。predict 函数做的工作就是求解所有后验概率并找出最大的那个。也即是本小组需要先利用朴素贝叶斯分类器对训练集进行一级标签分类，随后再对测试集进行分类，相关代码见附件 question1.py。

2.1.5 F-Score 评价

本小组最终采用了经过训练的朴素贝叶斯分类模型对测试集进行了一级标签分类并使用了 F-Score 对分类方法进行评价，F-Score 的相关公式为：

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

其中 P_i 为第 i 类的查准率， R_i为第 i 类的查全率。

本小组模型计算的 F 值和相关数值如下表所示：

表 1.F-Score 结果

模型精度	0.992
模型召回率	0.992
F1-Score	0.992

由表 1 可以看出本小组的模型精度和召回率都很高，F-Score 评价也很好，说明模型效果较为有效。

2.2 问题 2 的分析方法与过程

本小组解决本问题根据以下步骤：

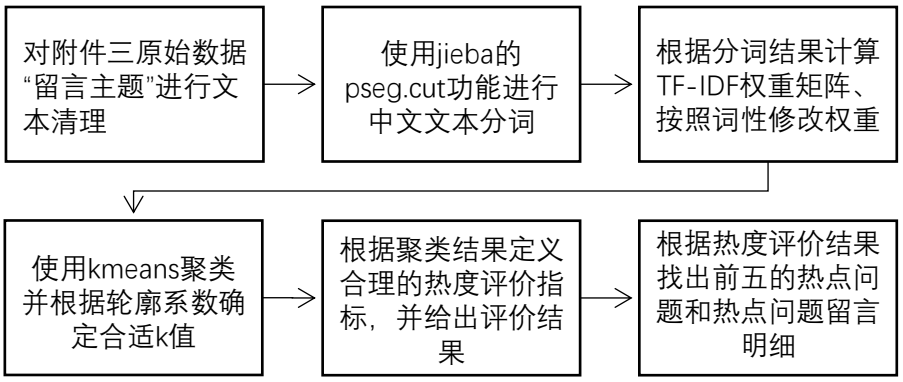

图 2.问题二流程分析

2.2.1 数据预处理

因为问题二要求本小组分析群众留言中的热点问题，所以本小组将继续采用 jieba 中文分词，为了有效分词和之后的聚类操作，本小组在分词之前进行了文本清理，但是和第一题不同，本小组根据问题要求去掉了除了大小写 A-Z 还有数字以外的非中文字符，从而保留关键地名，相关代码见附件 question2.py.

2.2.2 jieba 中文分词

根据问题二的要求本小组需要对留言主题进行词性标注，这里本小组采用 jieba 里面自带的 pseg.cut 函数，从而也就可以返回词还有词性，jieba.posseg.POSTokenizer(tokenizer=None) 可以新建自定义分词器，Tokenizer 参数可指定内部使用的 jieba.Tokenizer 分词器。jieba.posseg.dt 为默认词性标注分词器。它能标注句子分词后每个词的词性，采用和 ictclas 兼容的标记法。从而本小组能够有效识别地名、人名、机构团体名和动词等关键信息，相关代码见附件 question2.py,下图为本小组使用的部分词性对照表:

表 2.词性对照表

词性编码	词性名称	注解
Ag	形容词	形容词性语素。形容词代码为 a，语素代码 g 前面置以 A。
a	形成词	取英语形容词 adjective 的第 1 个字母。
vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起。
an	名形词	具有名词功能的形容词。形容词代码 a 和名词代码 n 并在一起。
f	方位词	取汉字“方
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母。
Ng	名语素	名词性语素。名词代码为 n，语素代码 g 前面置以 N。
v	动词	取英语动词 verb 的第一个字母
n	名词	取英语名词 noun 的第 1 个字母。
nr	人名	名词代码 n 和“人(ren)”的声母并在一起。
ns	地名	名词代码 n 和处所词代码 s 并在一起。
nt	机构团体	“团”的声母为 t，名词代码 n 和 t 并在一起。
nz	其他专名	“专”的声母的第 1 个字母为 z，名词代码 n 和 z 并在一起。
t	时间词	取英语 time 的第 1 个字母。

2.2.3 计算 TF-IDF 并调整权重

根据第一题的介绍，本小组继续计算 TF-IDF 权重矩阵来使得这些分词后的词语转化可供分析计算的向量，从而本小组就得到一个 weight（tfidf.toarray）的向量矩阵，也即是将 tf-idf 矩阵抽取出来,元素 a[i][j]表示 j 词在 i 类文本中的 tf-idf 权重，之后本小组对照词性表，重新修改权重，本小组对地名、动名词增加了权重，对一些语气词等等就减少了权重，经过试验本小组使得 Kmeans 的聚类预测更加准确。

2.2.4 使用 Kmeans 聚类并根据轮廓系数确定合适 k 值

2.2.4.1 Kmeans 算法简介

k 均值聚类算法（k-means clustering algorithm）是一种迭代求解的聚类分析算法，k 均值聚类是使用最大期望算法（Expectation-Maximization algorithm）求解的高斯混合模型（Gaussian Mixture Model, GMM）在正态分布的协方差为单位矩阵，且隐变量的后验分布为一组狄拉克 δ 函数时所得到的特

例。

2.2.4.2 Kmeans 聚类步骤

1. 先将数据分为 K 组，则随机选取 K 个对象作为初始的聚类中心
2. 然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。
3. 每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。
4. 终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

2.2.4.3 利用轮廓系数 - Silhouette Coefficient 确定合适的 k 值

对于一个聚类任务，本小组希望得到的簇中，簇内尽量紧密，簇间尽量远离，轮廓系数便是类的密集与分散程度的评价指标，公式表达如下：

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (2)$$

其中 a(i)代表同簇样本到彼此间距离的均值，b(i)代表样本到除自身所在簇外的最近簇的样本的均值，s 取值在[-1, 1]之间。 如果 S 接近 1，代表样本所在簇合理，若 s 接近-1 代表 S 更应该分到其他簇中。

本小组的判断标准是：轮廓系数范围在[-1,1]之间。该值越大，聚类越合理。si 接近 1，则说明样本 i 聚类合理； si 接近-1，则说明样本 i 更应该分类到另外的簇； 若 si 近似为 0，则说明样本 i 在两个簇的边界上。 所有样本的 s i 的均值称为聚类结果的轮廓系数，是该聚类是否合理、有效的度量。 使用轮廓系数(silhouette coefficient)来确定，选择使系数较大所对应的 k 值。sklearn.metrics.silhouette_score 函数中有对应的求轮廓系数的 API。根据本小组的试验结果，本小组计算了 K 从取 30 到 100 的时的轮廓系数，发现当 K=65

时，轮廓系数最大，所以本小组最终决定选取参数 K=65。

2.2.5 建立热度评价模型并进行热度问题评估

2.2.5.1 建立热度评价模型的指标体系

根据问题要求，本小组决定采用熵值法综合评价，在信息论中，熵是对不确定性的一种度量。信息量越大，不确定性就越小，熵也就越小；信息量越小，不确定性越大，熵也越大。根据熵的特性，本小组可以通过计算熵值来判断一个事件的随机性及无序程度，也可以用熵值来判断某个指标的离散程度，指标的离散程度越大，该指标对综合评价的影响越大。因此，可根据各项指标的变异程度，利用信息熵这个工具，计算出各个指标的权重，为多指标综合评价提供依据，所以再进行熵值法综合评价之前本小组需要先建立合适的指标体系，下面是本小组所选取的指标：

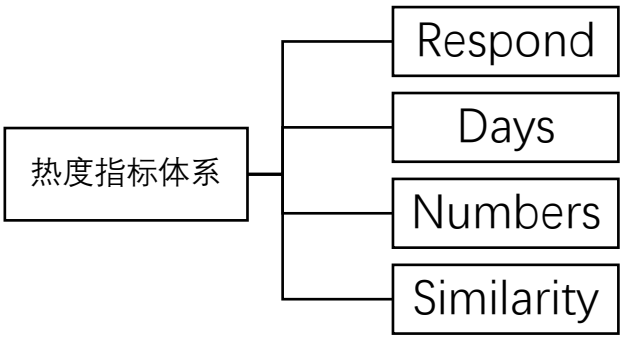

图 3.热度指标体系图

（1） 指标“Respond”

讨论量和互动量是表征问题热度的最直接指标，一般采用对各项数据加权的方式来计算。

问题讨论量、点赞量和反对量都是采用线性加权的方式来做，权重值的大小必然会有区别，这里根据现有内容计算指标权重，记为“Respond”。

（2） 指标“Days”

一个热点问题的留言时间跨度也是本小组需要考虑的内容，通常时间跨度越长也代表该热点问题收到的关注越多、问题越重要，所以是一项重

要指标，这里本小组计算一个问题类别的所有留言中最晚和最近的时间天数之差作为该问题类别的时间跨度，记为“Days”。

（3） 指标“Numbers”

各个热点问题包含的留言数量也是一项很重要的指标，它衡量了该类别问题在群众中的反响程度，留言数量越多说明群众的意见越大，所以本小组选取留言数量作为指标并记为“Numbers”

（4） 指标“Similarity”

聚类后，每一类主题内部有若干条留言的详情。由于主题的聚类并不是百分之百准确，因此类内每条留言与主题的相似度亦是热度的衡量的一个重要指标。

计算文本相似度的过程中，本小组借助了文本相似度库 gensim 完成步骤，具体过程如下：

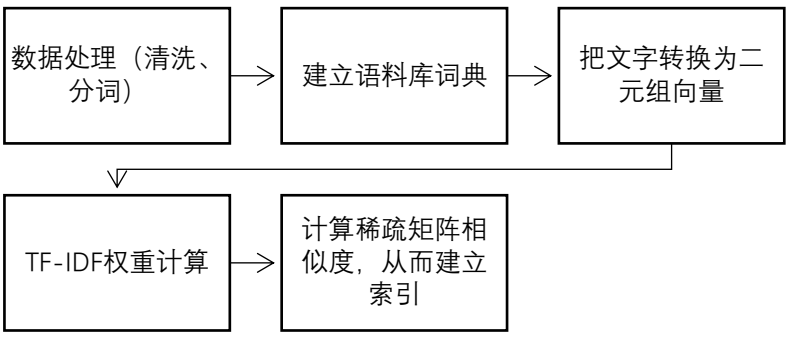

图 4.相似度计算流程图

其中，TF-IDF 的计算在问题一中分析过，而二元组向量对应的是（编号、频次数），可以对应于分词后的文档中的每个词，快速定位词语。因为相似度是数学上的概念，自然语言无法完成，因此，把文本转化为向量，利用余弦相似度方法计算是有必要的。

余弦相似度的公式如下：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \quad (3)$$

2.2.5.2 熵权法评价指标建立

熵权法赋权具体步骤如下：

（1）数据标准化

将各个指标的数据进行标准化处理。

假设给定了 k 个指标$X_1, X_2, \ldots, X_k$，其中$X_i = \{x_1, x_2, \ldots, x_n\}$。假设对个指标数据标准化后的值为$Y_1, Y_2, \ldots, Y_k$，那么，标准化公式为：

$$Y_{ij} = \frac{x_{ij} - \min(X_i)}{\max(X_i) - \min(X_i)}。 \quad (4)$$

（2）求各指标的信息熵

根据信息论中信息熵的定义，一组数据的信息熵为：

$$E_j = -\ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \quad (5)$$

其中$p_{ij} = Y_{ij} / \sum_{i=1}^{n} Y_{ij}$，如果$p_{ij} = 0$，则定义$\lim_{p_{ij} \to 0} \sum_{i=1}^{n} p_{ij} \ln p_{ij} = 0$。

（3）确定各指标权重

根据信息熵的计算公式，计算出各个指标的信息熵为$E_1, E_2, \ldots, E_k$。通过信息熵计算各指标的权重

$$W_i = \frac{1 - E_i}{k - \sum E_i} (i = 1,2, \ldots, k)。 \quad (6)$$

根据程序的计算，本小组得出各项指标的权重如下表所示：

表 3.热度指标及其权重

热度指标	指标权重
Similarity	0.278189359
Days	0.0270206
Numbers	0.321716391

Respond	0.373073649	

参数 K=65，所有留言被分为 65 类问题，根据之前建立的热度指标权重体系和熵权法的基本原理，本小组对 65 类问题进行了综合评价得分分析，下图是相关结果：

表 4.熵值法热点问题分析综合得分表

问题类别	Similarity	Days	Numbers	Respond	综合得分	Similarity 排名	Days 排名	Numbers 排名	Respond 排名	综合得分排名
12	15.766	0.155	19.925	19.310	55.156	1	1	1	1	1
52	1.980	0.061	2.922	12.583	17.545	2	3	2	2	2
64	1.842	0.060	1.895	0.415	4.212	3	10	3	7	3
49	0.860	0.060	1.503	0.839	3.262	4	4.5	4	3	4
4	0.654	0.060	0.612	0.115	1.440	5	4.5	5	15.5	5
27	0.307	0.055	0.249	0.762	1.373	8	17	10	4	6
36	0.324	0.052	0.317	0.599	1.291	6	23	7	5	7
61	0.320	0.072	0.340	0.146	0.877	7	2	6	10	8
50	0.185	0.060	0.113	0.501	0.859	20	7.5	24	6	9
19	0.260	0.051	0.211	0.143	0.665	14	25	13	11	10

2.2.6 热点问题主题提取

根据问题要求和表 4 的熵值法热点问题分析综合得分表，本小组利用对排名前五的热点问题进行了文本清理和 jieba 分词处理和关键词提取从而浓缩了聚类后的类别问题然后得出了排名前五的热点问题。

在上一步骤中，得到的每个分类，内容是由多个“留言主题”组成的。而本题的任务是需要获得最终热点问题的主题，因此小组需要从综合排名前五类中若然“留言主题”中提取出最终的主题。主题抽取有若干方法，目前最为流行的叫做隐狄利克雷分布（Latent Dirichlet allocation），简称 LDA[3]。

LDA 它是一种典型的词袋模型，及它认为一篇文档是由组词构成的一个集合，词与词之间没有顺序以及先后的关系。在 LDA 模型中一篇文档生成的步骤如下：

（1）从狄利克雷分布中取样生成文档 i 的主题分布θ_i

（2）从主题的多项式分布θ_i中取样生成文档 i 第 j 个词的主题$z_{i,j}$

（3）从狄利克雷分布β中取样生成主题$z_{i,j}$的词语分布$\emptyset_{z_{i,j}}$

（4）从词语的多项式分布$\emptyset_{z_{i,j}}$中采样最终生成词语$w_{i,j}$

其中，整个模型中所有可见变量以及隐藏变量的联合分布如下：

$$\mathrm{p}(w_{i,}z_{i,},\theta_i,\Phi|\alpha,\beta)=\prod_{j=1}^{N}p(\theta_i|\alpha)p(z_{i,j}|\theta_i)p(\Phi|\beta)p(w_{i,j},\theta_{z_{i,j}}) \qquad (7)$$

最终文档的单词分布的似然估计可以通过上式的θ_i以及Φ进行积分和对z_i进行求和得到：

$$\mathrm{p}(w_{i,}|\alpha)=\iint \mathrm{p}(w_{i,}z_{i,},\theta_i,\Phi|\alpha,\beta) \qquad (8)$$

接下来便可以继续用参数估计得出余下参数的估计值，再次不在赘述。而回到本问题上，python 中的 Sklearn 库中的 LatentDirichletAllocation()函数则可以快速完成此步骤，具体见附件 theme.py。

以下为表 5-热点问题表的展示：

表 5.热点问题表

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	55.1563	2017-06-08 至 2020-01-08	A市区经济学院	学生被学校强制要求到企业实习
2	2	17.54532	2019-01-01 至 2020-01-07	A4区珠江郦城	解决安置及拆迁问题反映
3	3	4.212349	2019-01-10 至 2020-01-07	A5魅力之城小区	夜间街道噪音和宵夜油烟扰民
4	4	3.261788	2019-01-01 至 2020-01-06	A市制药机械公司	涉嫌诈骗、拖欠工资
5	5	1.440007	2019-01-04 至 2020-01-06	A3区雨敞坪镇	通讯、公交发展存在问题

2.3 问题 3 的分析方法与过程

对于政府答复意见的评价，本小组从附件 4 获取相关内容后，进行基础的数据预处理。此后，关键的一步则是对政府部门答复的内容进行量化，而量化的前提是找出能够衡量相关机关对于市民提问答复质量优劣的指标。目前询问回答质量评价指标体系的构建一般基于用户和产品两种视角，用户视角从满足用户需求出发，而产品视角注重于信息产品的客观属性，如信息的完整性、可获得性等[4]。本小组认为询问回答质量评价属于这两个方面的综合。结合本体情况，小组确定从回复的内容以及回复的效用两个维度来评价答复意见的质量。

具体流程如下图：

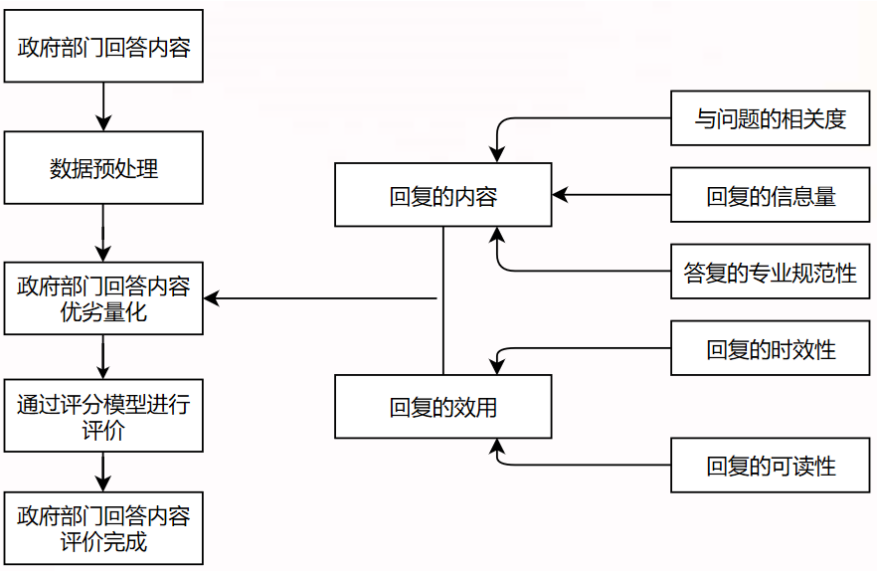

图 5.政府答复评价示意图

2.3.1 政府回复问题质量的特征

2.3.1.1 回复与问题的相关度

回复与问题的相关度一定程度地反映了政府的回答是否贴合题目的描述，本小组采用政府答复意见以及留言详情的相似度来衡量答复与问题的相关度。在此，本小组仍然采用文本相似度库 gensim，得出文本间的余弦相似度。

2.3.1.2 回复的信息量

信息量可理解为评论信息的含量，如句量和词量，也可表示为平均句长或单句长度。通常认为，评论的信息含量越大，评论的质量就越高[5]。本小组通过统计每条答复意见的中文字数和句数得到该答复意见的相关信息量。

2.3.1.3 回复的专业规范性

回复的专业规范性是指政府的答复内容是否尝试采用尽量多的机关公文书写词汇，根据《机关部门书写专用词汇》（见附件 txt 文件）中的词汇可以政府部门回复内容的专业度和回复质量，可以认为相关在回答语句中所用的词语和相关机关部门书写专用的词语的语义相似度越高，用到的规范词汇越多，则该评论的规范性则越高[6]。通过计算专用词汇表与答复意见的相似度，小组成员得到政府答复意见的专业度值。

2.3.1.4 回复的时效性

回复的时效性是指政府的答复时间是否与群众的提出问题时间相隔较近，时间相隔越近说明政府答复越及时，处理问题越高效，本小组采用计算答复时间日期与群众提出问题时间的天数之差这一指标来评价回复的时效性。

2.3.1.5 回复的可读性

政府答复意见的可读性反映了其是否易于阅读，能否很好被广大读者理解。关于中文评论可读性的研究，至今没有统一的定论和成型的算法，而英文可读性研巧相当成熟，本文则借鉴英文可读性的衡量公式自动化可读性指数（Automated Readability Index，ARI）的值表示[7]，即：

ARI = 4.71*(总字符数 / 总字数）＋0.5 *（总字数 / 总句数） - 21.43，

其中字符数可按照中文一个汉字相当于两个字符来计算，暂时获得评论的可

读性指数。

2.3.2 答复质量评价模型

2.3.2.1 层次分析法的简介和原理

根据问题要求，本小组决定采用层次分析法来建立答复质量评价模型，层次分析法是指将一个复杂的多目标决策问题作为一个系统，将目标分解为多个目标或准则，进而分解为多指标（或准则、约束）的若干层次，通过定性指标模糊量化方法算出层次单排序（权数）和总排序，以作为目标（多指标）、多方案优化决策的系统方法。

层次分析法的基本思路是将所要分析的问题层次化；根据问题的性质和所要达成的总目标，将问题分解为不同的组成因素，并按照这些因素的关联影响及其隶属关系，将因素按不同层次凝聚组合，形成一个多层次分析结构模型；最后，对问题进行优劣根据权重比较并排列。

2.3.2.2 层次分析法的步骤

1. 建立层次结构模型

将决策的目标、考虑的因素（决策准则）和决策对象按照他们之间的相互关系分为最高层（目标层）、中间层（准则层）和最低层（方案层），绘出层次结构图。

（1） 最高层： 决策的目的、要解决的问题。
（2） 最低层： 决策时的备选方案。
（3） 中间层： 考虑的因素、决策的准则。
（4） 对相邻的两层，称高层为目标层，低层为因素层。

层次分析法所要解决的问题是关于最低层对最高层的相对权重的问题，按此相对权重可以对最低层中的各种方案、措施进行排序，从而在不同的方案中做出选择或形成选择方案的原则。

2. 构造判断矩阵

层次分析法中构造判断矩阵的方法是一致矩阵法，即：不把所有因素放在一起比较，而是两两相互比较；对此采用相对尺度，以尽可能减少性质不同的因素相互比较的困难，从而提高准确度。

表 6.判断矩阵a_{ij}的标度方法

标度	含义
1	表示两个因素相比，具有同样重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素明显重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要
2,4,6,8	上述两相邻判断的中值
倒数	因素 i 与 j 比较的判断a_{ij}，则因素 j 与 i 的比较的判断 $a_{ji}=1/a_{ij}$

3.层次单排序及其一致性检验

对应于判断矩阵最大特征根λ_{max}的特征向量，经归一化（使向量中各元素之和为 1）后记为 W。W 的元素为同一层次元素对于上一层因素某因素相对重要性的排序权值，这一过程称为层次单排序。

本小组定义一致性指标:

$$CI=\frac{\lambda-n}{n-1} \tag{9}$$

当$CI=0$　有完全的一致性；

CI接近于 0，有满意的一致性；

CI越大，不一致越严重。

为了衡量 CI 的大小，引入随机一致性指标 RI：

表 7. 随机一致性指标 RI

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

本小组随后定义一致性比率：$CR = \frac{CI}{RI}$，一般认为一致性比率$CR < 0.1$ 时，A 的不一致程度在容许范围之内，有满意的一致性，通过一致性检验。可用其归一化特征向量作为权向量，否则要重新构造成对比较矩阵 A，对a_{ij}加以调整。

4. 层次总排序及其一致性检验

计算某一层次所有因素对于最高层（总目标）相对重要性的权值，称为层次总排序。

这一过程是从最高层次到最低层次依次进行的，根据本题的实际情况，本小组将指标进行如下层次构建：

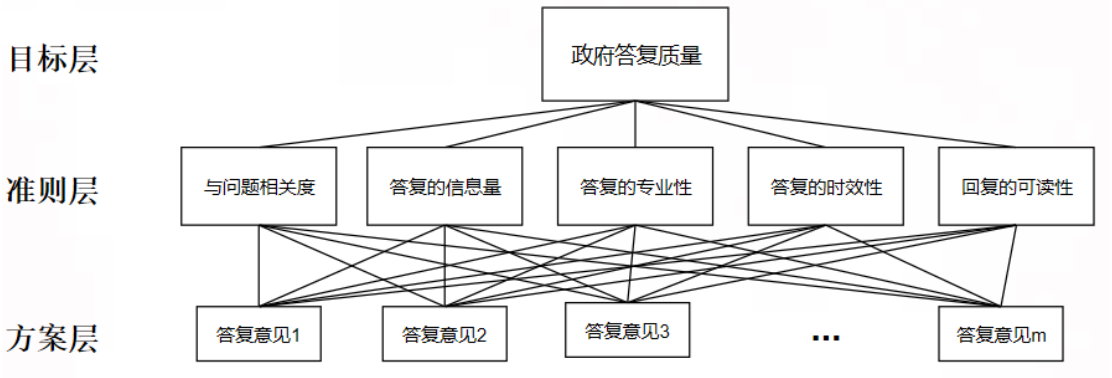

图 6. 层次分析法结构示意图

其中：

准则层 m 个因素$A_1 A_2, \ldots, A_m$,对总目标 Z 的排序为$a_1 a_2, \ldots, a_m$。方案层 n 个因素对上层 A 中因素为A_j的层次单排序为$b_{1j}, b_{2j}, \ldots, b_{nj}$ (j=1, 2, 3, ···, m)方案层的层次总排序(即 B 层第 i 个因素对总目标的权值为：$\sum_{j=1}^{m} a_j b_{ij}$为：

$$\begin{aligned} &B1:\ a_1b_{11}+a_2b_{12}+\cdots+a_mb_{1m}\\ &B2:\ a_1b_{21}+a_2b_{22}+\cdots+a_mb_{2m}\\ &\vdots\\ &Bn:\ a_1b_{n1}+a_2b_{n2}+\cdots+a_mb_{nm} \end{aligned} \qquad (10)$$

层次总排序的一致性比率为:

$$CR=\frac{a_1CI_1+a_2CI_2+\cdots\cdots+a_mCI_m}{a_1RI_1+a_2RI_2+\cdots\cdots+a_mRI_m} \qquad (11)$$

其中，当 CR<0.1 时，本小组认为层次总排序通过一致性检验。

代入本题获得的指标数据，以及加入小组成员讨论判断评分后，得出如下的评分以及结果：

代入本题获得的指标数据，以及加入小组成员讨论判断评分后，得出如下的评分判断矩阵 A 以及结果：

表 8.判断矩阵 A 所对应的指标的权重

	问题相关度	专业性	信息量	可读性	时效性
问题相关度	1	4	3	5	2
专业性	1/4	1	1/2	2	1/3
信息量	1/3	2	1	3	1/3
可读性	1/5	1/2	1/3	1	1/4
时效性	1/2	3	3	4	1

2.3.3 答复评价模型的应用

小组成员将上述指标评分模型应用到附件 4 当中，得出所有留言的一个评分（详情见附件“答复评分.csv”）。以下展示了其中一部分的评分结果：

表.政府部门答复评分表

留言编号	留言用户	留言主题	留言时间	答复意见	答复时间	评分
12278	UU00824	对 A 市地图、西地省地图出版社及 A 市地名委员会的希望	2014/5/23	网友：您好！留言已收悉	2014/6/5	0.4040

18413	UU008835	反对在 A7 县江背镇乌川湖村兴建涵洞	2015/12/30	网友“UU008835” 您好！来信收悉。现回复如下：经镇综治办…	2016/1/7	44.8082
88359	UU0082390	咨询 J9 县的油茶种植政策	2019/1/5	首先感谢您对 J9 县扶贫工作的支持与关注！针对您咨询…	2019/1/7	41.5786
37366	UU0082052	投诉 B9 市江源领航国际幼儿园乱收费	2019/6/18	您好！收到您反映的问题后，市教育局民办教育股立即对江源领航幼….	2019/7/1	20.9779

从答复评分表可以看出。留言编号为 12278 对应的答复意见评分较低，原因是答复的时间相对较长，且答复意见的字数过少，信息量不足，对留言的回应度更是很差。对比之下，得分较高的 18413 留言和 88359 对应的答复意见，则是信息量较大，回复字数在 1100 字以上。观察到两则答复的时效性，都是能够在 8 天内给出答复，而且答复的内容能够覆盖到留言主题以及留言详情。而留言编号为 37366 的答复则处于中等水平。

3.模型总结与展望

3.1 优点

1. 对于问题一，在进行文本分词之前，能够设计模型将长文本的关键句提出。另外，在筛选关键句时，还可以根据所需要的权重进行筛选，对于结果的优化有很大作用。
2. 对文本进行一级标签分类时，采取贝叶斯分类器，能够有较高的分类效率，并且最终有较高精度的输出。
3. 问题二在进行聚类前，对文本的权重根据地名、专有名词、动名词等词语属性进行权重的重新调整，确保了 kmeans 聚类时，保证地点人群等关键主题能对其有显著影响。
4. 利用熵权法对问题进行评分，进而得出热点问题时，充分考虑到了指标之间数量级的差异，能够经过处理，得出具有代表性数据。
5. 对于问题三，在评价指标设计时，考虑到了添加政府机关写作专用词字典，并通过分析与答复的吻合度，来展现政府回答的专业性，亦是一个

重要的评价维度。

3.2 缺点

1. 问题二中，对所有留言进行聚类后，某些类中仍然存在不相关的留言，精确性有待提高。
2. 对热点问题的主题提取过程中，输出的结果是以词语为单位，需要人为总结。

3.3 展望

- 在使用 textrank4h 来筛选关键句时，可以通过参考更多的资料，来确定最有效的权重。
- 使用 Kmeans 聚类时，可以再扩大 K 值可选取范围，进一步确定所选的 K 值为最优。
- 对答复问题的可解释性维度评价上，未来可以设计更好的模型，提炼出留言详情的关键问题，以及答复的关键主题，进行匹配分析。

4．参考文献：

[1] 杨玥,张德生.中文文本的主题关键短语提取技术[J].计算机科学,2017(S2):432-436.

[2] 彭世瑜.Python 编程：使用 textrank4zh、jieba、snownlp 提取中文文章关键字和摘要[EB/OL].(2019-04-23)[2020-5-1].

[3] -柚子皮-.主题模型 Topic Model：隐含狄利克雷分布 LDA [EB/OL].(2015-1-12)[2020-05-5].https://blog.csdn.net/pipisorry/article/details/42649657.

[4] 袁红,张莹.问答社区中询问回答的质量评价——基于百度知道与知乎的比较研究[J].数字图书馆论坛,2014(09):43-49.

[5] 胡泽. 在线问诊服务回答质量评价方法研究[D].哈尔滨工业大学,2019.

[6] 杨开平,李明奇,覃思义.基于网络回复的律师评价方法[J].计算机科学,2018,45(09):237-242.

[7] 郭银灵. 基于文本分析的在线评论质量评价模型研究[D].内蒙古大学,2017.