基于自然语言处理技术的“智慧财务”文本挖掘

摘要

随着社会的发展，在快节奏的时代下，政府主要向微信、微博、市长信箱、阳光热线等网络问政平台了解民意、汇聚民智、凝聚民气，然而每一类社情民意相关的文本数据量不断扩展。因此，目前，社会治理创新发展的新趋势是建立基于自然语言处理技术的“智慧政务”系统，对附件中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进行分析研究，提取我们需要进行分析的部分进行深度挖掘和分析。

针对问题一：本为附件 1 中的结构化文本数据数值化处理，为附件 2 中非结构化文本去掉空行、X 序列、中文文本分词、过滤无意义的词等数据预处理，其次基于 TF-IDF 权重法提取 64000 个候选特征词，为了构造词汇-文本矩阵，而生成词袋，并且为了除同义词的影响，利用奇异值分解算法进行语义空间降维。根据文本内容，计算其欧式距离，完成基于 KNN 分类算法对各个留言描述进行分类。最后通过 F-score 分类器进行评估模型。

针对问题二：同样对其数据数值化处理后，通过热点问题的三要素进行研究挖掘，运用主成分分析法构建综合排名算法得出热点问题的排名。

针对问题三：为了将相关性、完整性、可解释性等描述量化，构建生成回答质量评价指标体系,通过感情分析进行数据答复意见，其中把情感分析中的维度转换为相关性、完整性、可解释性进行综合打分。最后用用回归模型来进行质量好坏的一个打分，构建这样一个评分函数。

关键词：TF-IDF；语义空间降维；F-score 分类器；主成分分析；感情分析

目录

一 问题的重述

1.1 问题背景

随着社会的发展，在快节奏的时代下，政府主要向微信、微博、市长信箱、阳光热线等网络问政平台了解民意、汇聚民智、凝聚民气，然而每一类社情民意相关的文本数据量不断扩展，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。

目前，社会治理创新发展的新趋势是建立基于自然语言处理技术的“智慧政务”系统，对附件中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进行分析研究，为提升政府的管理水平和施政效率有着非常重要的意义。

1.2 目标任务

问题一群众留言分类：处理网络问政平台的群众留言，工作人员通过附件 1 提供的内容分类三级标签体系对留言进行分类，以便后续将群众留言分派到相应的职能部门处理。为了避免工作人员根据经验处理分类，而造成的效率低，且差错率高等问题。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，最后使用分类的评判标准 F-score 分类器进行评价。

问题二热点问题挖掘：热点问题是某一时段内群众集中反映的某一问题，比如“XX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。由于及时发现热点问题，能让相关部门进行有针对性地处理，进而提升服务效率。根据附件 3 将热点问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

问题三答复意见的评价：针对附件 4 中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，构建指标来计算和评价。

二．问题分析

当今的大数据时代中，当前每月互联网中产生的数据总量都会翻一番，本文借助自然语言处理技术对“智慧财务”文本挖掘，这个问题被政府工作人员持续关注，现在我们通过题目要求学习分析建立相应的优化模型，尽可能的得出最优解。

2.1 对问题一的分析：

针对问题一，本题的总体架构及思路如下：

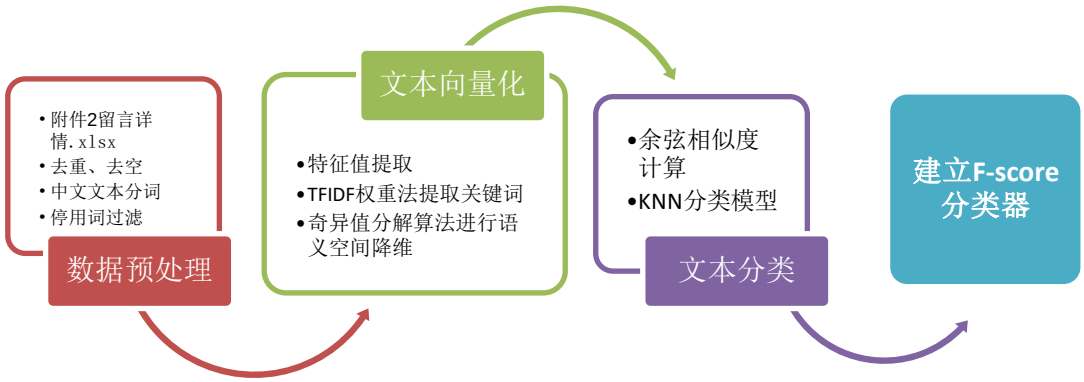

图 2.1 问题一流程图

步骤一：通过数据预处理，为附件 1 中的结构化文本数据数值化处理，为附件 2 中非结构化文本去掉空行、X 序列、中文文本分词、过滤无意义的词，进而优化模型。

步骤二：通过文本向量化，选择特征值提取，其次用 TFIDF 权重法来提取关键词，构造矩阵，最后为了除同义词的影响，利用奇异值分解算法进行语义空间降维，简化计算。

步骤三：文本分类，根据文本内容，计算附件 1 与附件二之间的欧式距离，再基于 KNN 分类算法对各个留言描述进行分类。

步骤四：建立 F-score 分类器，做混淆矩阵，用于描述一级标签分类模型，测试数据的性能，进行模型评估。

2.2 对问题二的分析：

针对问题二，关于热点问题挖掘，我们分为以下三个步骤进行：

（1） 问题识别：通过找问题的三要素（特定的时间、特定的地点、发生的问题）识别热点问题。其中我们的用到了命名实体识别。

（2） 问题归类：通过相似度计算，我们把相似的文本内容归类，方便接下来的量化评价指标。

（3） 热度评价：我们设计量化评价指标产生热度评价，运用主成分分析法按照表 1 的格式构建排名前 5 的热点问题，最后提交“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，最后提交“热点问题留言明细表.xls”。

2.3 对问题三的分析：

针对问题三，关于答复意见的评价，为了针对相关部对留言的答复意见，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，并尝试实现。其中有以下定义：

（1）相关性:答复意见的内容是否与问题相关

（2）完整性:是否满足某种规范

（3）可解释性:答复意见中内容的相关解释

为了将相关性、完整性、可解释性等描述量化，构建生成回答质量评价指标体系,实现面向群众需求的回答质量自动化评价和筛选，提高政府的服务质量，我们通过感情分析进行数据答复意见，其中把情感分析中的维度转换为相关性、完整性、可解释性进行综合打分。最后用用回归模型来进行质量好坏的一个打分，构建这样一个评分函数。

三．符号说明

P_i	第 i 类的查准率
R_i	第 i 类的查全率
D_k	第 k 个分类中文档的总数量
idf_{ik}	特征词t_i的反文档频率
tf_{ik}	特征词t_i在文档D_K中出现的总次数
$\tilde{x}_{ij}$	标准化指标

四．模型的建立与求解

4.1 问题一模型的建立与求解

4.1.1 问题一模型的建立

首先我们考虑实际问题，本问需要考虑数据预处理过程。因为附件二中留言详情数据量大，而且大数据时代对于数据的精度和有效性要求更为苛刻,因此数据的预处理过程必不可少,只有科学规范的预处理过程,才能使数据分析深层挖掘的结论更为合理可靠[1]。预处理流程图如下所示：

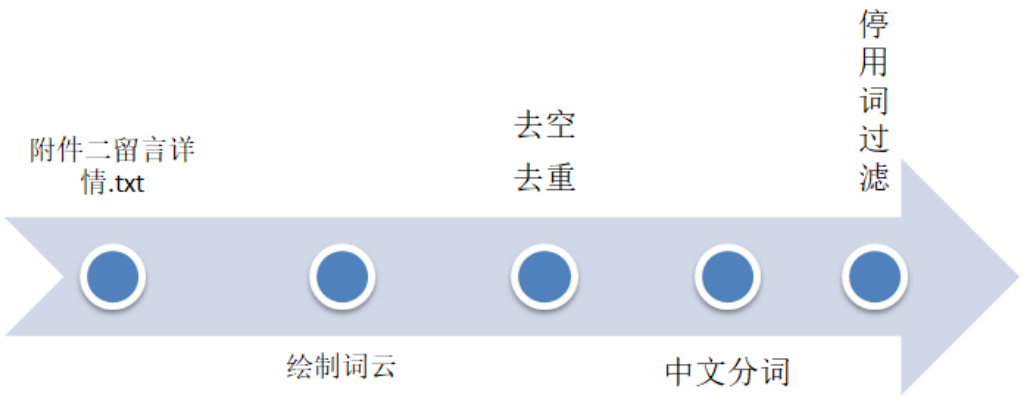

图 4-1 预处理流程图

其次，经过上述“附件二留言详情.txt”预处理后，即使去掉部分停用词，但还是存在大量词语，给下面需要做的文本向量化过程带来挑战，于是做特征抽取，其目的是在不改变文本原有核心信息的情况下，减少要处理的词数，以此来降低向量空间维数来简化计算，提高文本处理的速度和效率。本文用 TF-IDF 算法抽取特征词条，将权重按照从大到小的排序，抽取权重最大的前 64000 个特征词作为候选特征词。

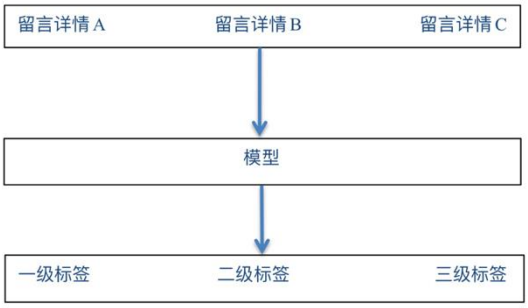

图 4-2 一级标签识别流程图

由于文本语义带来的词语交叉，我们需要语义空间降维，用奇异值分解基本原理保留相对大的奇异值，删去小的奇异值，从而可以对矩阵进行行与列的降维处理。

于是将无类别标记的留言详情根据不同的特征进行分类，为了发现数据中

未知的分类，使用相似度计算将具有一级标签属性或者相似属性的文本聚类在一起。从而建立起一级标签模型，群众就可以获得有效的回复，问题得到及时的解决，做一个“智慧”模型。

已知 K-means 算法是基于划分的聚类算法，即以空间中 k 个点为中心进行聚类，利用 Knn 算法找出最靠近 k 的对象。通过实验测试 k 的结果，逐次更新各聚类中心的值，直至得到最好的聚类结果。

根据 K 均值聚类，通过迭代的方法，则有[5]:

(1) K_1[初始化]：随机指定 K 聚类中心$(C_1 \quad C_2 \quad \dots \quad C_k)$；

(2) K_2[分配 X_i]:对每一个样本 X_i,找到离它近的聚类中心 C_v,并将其分配到C_v所标明类；

(3) K_3[修正C_w]:将每一个C_w移动到其标明的类的中心；

(4) K_4[计算偏差]: $D=\sum_{i=1}^{n}\left[\min_{r=1,\dots,K} d(X_i,C_r)^2\right]$

(5) K_5[D收敛]:如果D值收敛,则return$(C_1 \quad C_2 \quad \dots \quad C_k)$并终止本算法；否则,返回步骤$K_2$。

通过 K-means 分类得到聚类中心后，开始 Knn 算法，起步骤如下：

（1） 求距离：给定聚类中心，求解它与样本中的每个 TF-IDF 权重向量的距离。

（2） 找领边：确定离距离最近的 20 个样本，作为聚类中心的领边。

（3） 得分类：通过“python 根据 8 个语义关键词来对聚类进行分类，得出一级标签模型。

而以上提及的距离，本文采用的是欧式距离：

$$\mathrm{dis}(i,j)=\sqrt{\sum_{k=1}^{p}\quad (x_k-y_k)} \tag{4-1}$$

最后，面对题目中的多分类问题，我们采用 F-score 分类器来作为一级标签分类模型的评价指标，其中公式如下：

$$F=\frac{1}{\mathrm{n}}\sum_{i-1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{4-2}$$

而混淆矩阵分类模型的指标，属于 F-score 模型评估之一。混淆矩阵的数值可以看作分类器的结果。其中数值包括真正例、假正例、真正例、假负例。

（1） TN：真实值是负例，模型预测是负例的数量

（2） TP：真实值是正例，模型预测是正例的数量

（3） FN：真实值是正例，模型预测是负例的数量（第一类错误）

（4） FP：真实值是负例，模型预测是正例的数量（第二类错误）

表 4-1 混淆矩阵分类模型指标

	预测的类	负例	正例	合计
实际的类	负例	真负例（TN）	假正例（FN）	TN+FN
	正例	假负例（FN）	真正例（TP）	FN+TP
合计		TN+FN	FN+TP	

其中，精确率和召回率是二分类的常用评价指标，而本文为多分类问题，于是我们把文本的多分类为他转变成了多个二分类问题。以一级标签的类为正类，其他的类为负类。既而有以上四种情况。

4.1.2 问题一模型的求解

4.1.2.1 数据预处理

我们把这些文本数据的预处理分为五个部分：

(一)生成附件 2 留言详情.txt

对于附件二，提取部分留言详情，整成一个 txt 文件，另存为:“附件二留言详情.txt”。

(二)绘制词云

我们做了词频统计，画出词云图如下图，这样做的目的是词云可以反应群众的反馈关键信息。(详见:附件二词云.py)

图 4-2 词图

(三)去重、去空

对于附件二留言详情.txt：存在大量空行和长文本的无意义表达，通过python处理后如下所示：（详见：附件二去空去重.py）

图 4-3 去空去重图

(四)中文分词

中文分词是将汉语中连续的字序列切分成具有实际意义的最小单位的词语。汉语的历史悠久与博大精深决定了它的复杂性，提高分词算法的划分精度和划分速度，具有重要的理论意义和现实意义[2]。然而针对留言详情,在python中的“Jieba中文分词”算法中逆向匹配法的基础上提出双向最大匹配法。其中最大匹配算法主要是切分出单个字符串，然后和python中词库进行比对，如果匹配成功是一个词就记录下来，否则通过增加或者减少字符，继续匹配，直到匹配成功为止，如果该字符串不能增加或减少，则作为未登录词处理。

部分分词结果示例如下所示：（详见）

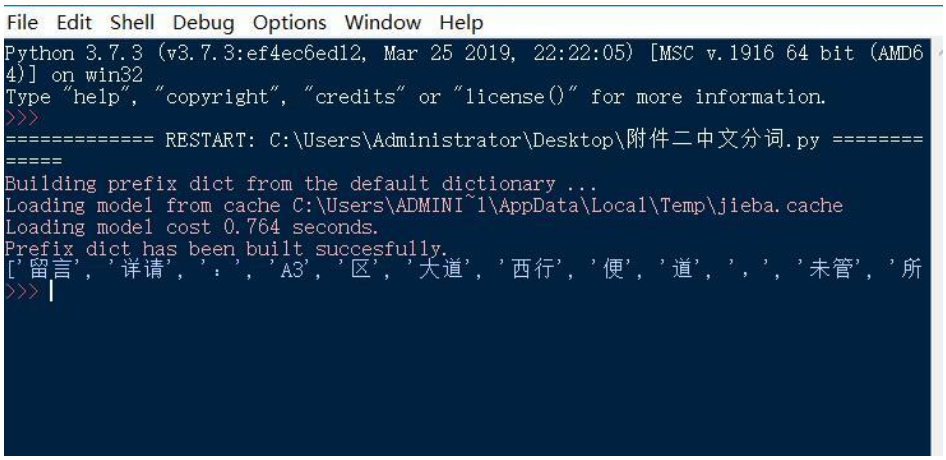

图 4-4 中文分词图

(五)停用词过滤

由以上的图可以看到，其中有着很多的标点和表达无意义的字词，对后续数据分析存在影响，因此接下来需要进行停用词过滤操作。通过 python 进行操作去除停用词后的部分结果示例如下图：（详见）

源数据：

“A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段”

分词之后：

“A3”“区”“大道”“西行”“便”“道”“未管”“所”“路口”“至”“加油站”“路段”“人行道”“包括”“路灯”“杆”“被”“圈”“西湖”“建筑”“集团”“燕子”“山”“安置”“房”“项目”“施工”“围墙”“内”“每天”“尤其”“上下班”“期间”“这”“条”“路上”“人流”“车流”“极”“多”“安全隐患”“非常”“大”“强烈”“请求”“文明城市”“A”“市”“尽快”“整改”“这个”“极”“不”“文明”“的”“路段”

停用词过滤：

“大道”“西行”“便”“道”“未管”“所”“路口”“加油站”“路段”“建筑”“集团”“燕子”“山”“安置”“房”“项目”“施工”“围墙”“内”“上下班”“期间”“路上”“人流”“车流”“多”“安全隐患”“大”“请求”“文明城市”“整改”“不”“文明”“路段”

图 4-5 停用词过滤示例图

4.1.2.2 文本向量化

（一）权重的计算

TF-IDF 是一种常用的权重计算方法，由于考虑了词频和反文档频率的影响，使得在少量文本中有较高的出现频率的词汇有较高的权重。我们利用分类中词频和文档频率之间关系。使用 TF-IDF 方法计算特征词在各分类中的权重[3]。

对于由 K 个分类组成的文档集合 D，有 D={D_i, …, D_k}, $k \in K$。而 n 个特征值组成的特征词集：

$$T = \{t_1, \cdots, t_i\}, i \in N \qquad (4\text{-}4)$$

计算 IDF 权重，即逆文档频率(Inverse Document Frequency)，需要建立词袋，用来模拟语言的使用环境。IDF 与特征性在文本中的分布呈正比，前者越大，分布越集中。特征词t_i在文档中的权重为：

$$W_{ik} = tf_{ik} \times idf_{ik} = tf_{ik} \times \log\left(\frac{D_k}{d_{ik} + 0.5} + 0.5\right) \qquad (4\text{-}5)$$

（二）文本的向量化表示

文本特征抽取将全部特征项筛选是 64000 个候选特征项，然后构建一个词袋，根据留言主题的特征项对应词袋中的位置，组成统一维数的向量：

$$C = (t_1, t_2, \ldots, t_n)$$

其中 C 为词袋集合，t_n 是每个词在向量中对应的位置。

这样留言评论主题可以参照词袋组成一维词向量，再通过 TF-IDF 向量化得到一个词汇—文本矩阵：

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1m} \\ w_{21} & w_{22} & \ldots & w_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ w_{41} & w_{42} & w_{43} & w_{44} \end{pmatrix} \end{array} \qquad (4\text{-}6)$$

生成文本向量的具体步骤如下：

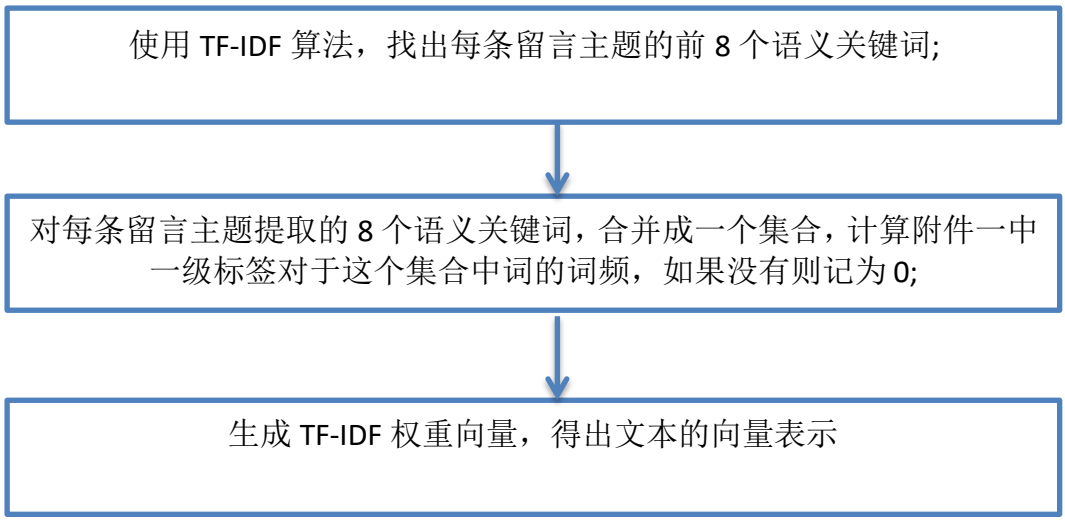

图 5-6 生成文本向量的流程图

（三）语义空间降维

留言详情文本信息中存在同义词和近义词等词语，文本语义也存在词语交叉等问题，便提出了词汇-文本矩阵的奇异值分解法。

我们删除对矩阵 A 影响较小的信息，去掉次要的字段，保留主要的字段，取 $A_{m\times n}$ 的相似矩阵 A_i。并对某一特征项为 n 的文本向量 t 进行奇异值分解得到[4]

$$A_{m\times n}=\begin{bmatrix} U_1SV_1^T & U_1SV_2^T & \cdots U_1SV_n^T \\ U_2SV_1^T & U_2SV_2^T & \cdots U_2SV_n^T \\ \vdots & \vdots & \ddots \quad \vdots \\ U_mSV_1^T & U_mSV_2^T & \cdots U_mSV_n^T \end{bmatrix} \tag{4-7}$$

$$t=t^{'}\sum U \tag{4-8}$$

又得出 t 在进行 k 维映射后得到的向量为：

$$t^{'}=tU_k^T\sum\nolimits_K^{-1} \tag{4-9}$$

4.1.2.3 文本分类

（一）K-means 聚类

由于留言详情给出了 条记录，去重后还有 条记录，如果把所有的留言详

情都用来挖掘分析,会占用空间和时间。为了方便计算，得出一级标签模型。从条记录数据中随机抽取 20000 条记录，通过 python 随机抽取(详见)，抽样结果保存在抽样样本.csv 文件里面。然后利用抽样样本进行分词、计算 TF-IDF 向量，并按照以下相应步骤使用 K-means 聚类：

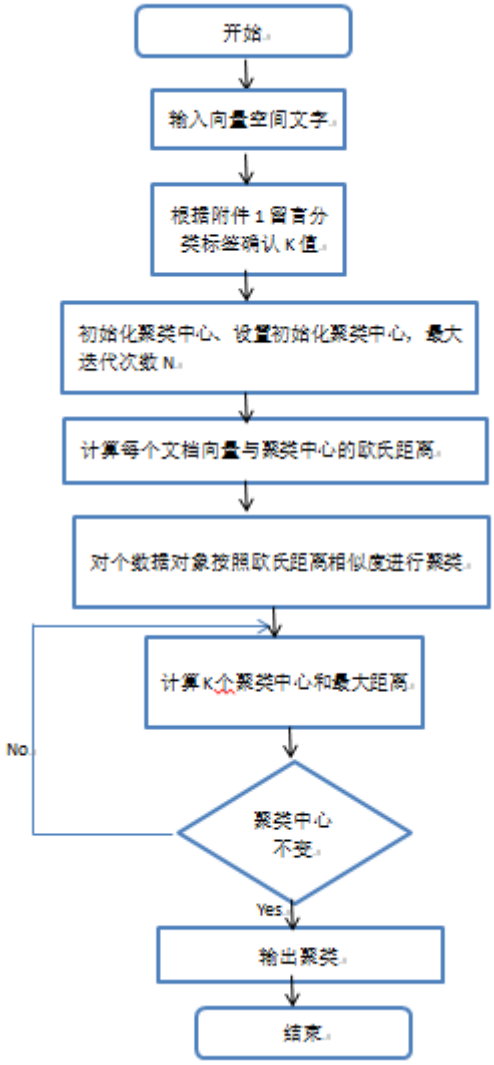

图 4-7 K-means 流程图

把样本按照 3 级标签进行分类(详见)得出来的三个聚类结果。

（二）Knn 最邻近分类算法

通过生成留言详情中的 TF-IDF 权重，为了衡量文本间相似度，通过欧氏距离进行相似度计算，对留言详情进行分类。采用 K-means 算法把留言详情分为 3 类。

根据以上内容，可以得出留言详情中的三级标签分类中，目前最多的是一级标签，我们也考虑到了三级标签中的从属包含关系，我们从后往前推建立了

这个一级标签分类模型，并且通过 F-score 分类器，评价我们的一级标签模型。

（1） 准确率与二分类相同，预测正确的样本在总样本的比例。

（2） 精确率是对于每个分类标签，分别计算 Precision 后取不加权平均

（3） 查全率是对于每个标签，分别计算 Recall，然后取不加权平均

（4） F-Score 要对每个分类标签，分别计算权，然后取不加权平均。将 2 个二分类评价的 TP,FP,FN 对应相加，计算 P 和 R，然后求得的结果如下：

		分类结果	
		0 （一级标签）	1 （非一级标签）
真实类别	0 （一级标签）	4714	146
	1 （非一级标签）	176	4770

图 4-8 混淆矩阵结果图

其中准确率（93.72%）、确度（88.61%）、召回率（96.44%）F 值（0.9602）虽然效果不是很好，可能是在预处理方面没有做好。

4.2 问题二模型的建立与求解

4.2.1 问题二模型的建立

在问题二中，题目给定了热点问题的概念，本文通过分析，定义了热点问题的三要素：特定的时间、特定的地点、发生的问题。针对三要素，结合所给的数据，用群众的问题描述、时间范围、地点人群描述热点问题。

（一）命名实体识别

所谓实体识别，就是将我们想要获取到的特定的时间、特定的地点、发生的问题，从附件 3 里面挑出来的过程。其中自然语言中的 NER 是一种序遵照序列标注问题的列标注问题，因此本任务提取 TIME、PERSON、LOCATION、ORGANIZATION 四种实体。以下列出来 BIOES 分别代表什么意思：

- B：Begin，开始
- I：Intermediate，中间
- E：End，结尾
- S：Single，单个字符
- O：即 Other，表示其他，用于标记无关字符

将“A3 区一米阳光婚纱艺术摄影是否合法纳税了么？”这句话，进行标注，结果就是：

[B-PER，E-PER，O，B-ORG，I-ORG，I-ORG，E-ORG，O，B-LOC，E-LOC，O，O，B-ORG，I-ORG，I-ORG，E-ORG，O，O，O，O]

同理附件三中根据输入的句子，预测出其标注序列的过程。

（二）相似度计算

该方法通过构建一系列汉语框架语义特征来表达每个问题的语义信息，进而使用最大熵模型相似度进行中文问题的自动分类。它决定量化评价指标的范围和方法，影响整个模型的性能。

（三）主成分分析算法

主成分分析法是一种降维的统计方法，设法将原来变量重新组合成一组新的相互无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上处理降维的一种方法[6]。

其步骤的流程图如下：

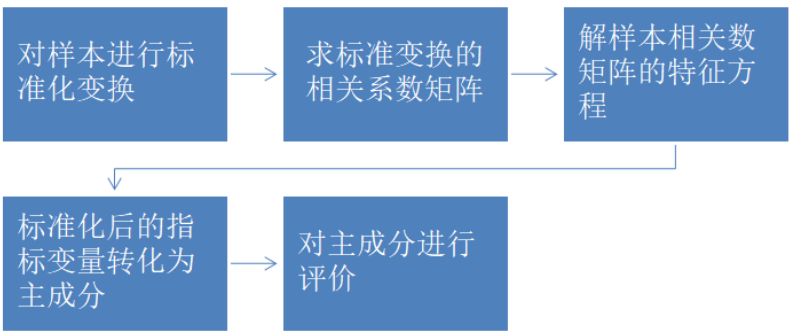

图 4-9 主成分分析法流程图

（1） 标准化处理：

$$\tilde{x}_{ij}=\frac{x_{ij}-\bar{x}}{s_j},(i=1,2,...,n;\ j=1,2,...m) \tag{4-10}$$

（2） 相关系数矩阵 R：

$$r_{ij}=\frac{\sum_{k=1}^{n}\bar{x}_{ki}\cdot\bar{x}_{kj}}{n-1},(i,\ j=1,2,...,m) \tag{4-11}$$

其中 r_{ij}=1， r_{ij}=r_{ji}， r_{ij}是第 i 个指标与第 j 个指标的相关系数。

（3） 计算特征值和特征向量：

计算相关系数矩阵 R 的特征值，与对应的特征向量。

（4） 选择主成分，计算综合评价值

$$b_j=\frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k}(j=1,2,3,...,m) \tag{4-12}$$

其中 λ_k 表示特征值， b_j 为第 j 个主成分的信息贡献率。

$$Z=\sum_{j=1}^{p}b_j y_j \tag{4-13}$$

通过以上计算其综合得分进行评价。

5.2.2 问题二模型的求解

（1） 问题的识别

通过我们实验研究，当前词能不能成为一个命名与他的前一个和后一个词有关，名词是命名实体，忽视动词，进而建立一个模板，关于定义特征词性。

（2） 问题的归类

为了使每一为群众的问题转化为向量，我们考虑到 Word2Vec，在这个数据量比较大、比较复制的情况下，我们采用 spark 训练。

（3） 热度评价

为了达到降维和属性独立的这两种目的，我们用到主分析发。用主成分分析对热点问题数据进行建模，最终将建模指标进行压缩，经统计，附件 3 中共有 4327 条数据，首先通过 excel 按照点赞数降序排列。因此其余 3049 个 0 点赞数的留言可以忽略不计，同样构造上述指标，最后利用 SPSS 对其进行综合排名。

A	B	C	D	E	F	G
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
208636	A00077171	区汇金路五矿万境K9县存在一系	2019/8/19 11:34:04	让狗咬人，请问有人对	0	2097
223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	益纳入配套入学，A3区	5	1762
220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	有案情消息总是失望，	0	821
217032	A00056543	关A市58车贷特大集资诈骗案保护	2019/2/25 9:58:37	小股东、苏纳弟弟苏	0	790
194343	A000106161	的A市58车贷案警官应跟进关注跟	2019/3/1 22:12:30	经侦并没有跟进市领	0	733
263672	A00041448	小区距长赣高铁最近只有30米	2019/9/5 13:06:55	回复到我如下问题：1、	0	669
193091	A00097965	富绿物业丽发新城强行断业主基	2019/6/19 23:28:27	只提供地摊上买的收据	0	242
284571	A00074795	他省尽快外迁京港澳高速城区段	2019/1/10 15:01:26	4、长浏高速出口，进	0	80
200667	A00079480	要把和包支付作为任务而不让市	2019/1/16 17:01:25	基层工作者也不理解，	0	78
262052	A00072424	月亮岛路沿线架设110kv高压线	2019/3/26 14:33:47	以上电力线路，应采用	0	78
226723	A00040222	三一大道全线快速化改造何时启	2019/9/15 15:31:19	改造，打通机场北通	0	66
281898	A00096623	房云时代多栋房子现裂缝，质量	2019/2/25 15:17:38	检站处理，陷入了与	5	55
272089	A00061602	于A6区月亮岛路110kv高压线的规	2019/4/9 17:10:01	西地省体操学校、西地	2	55
239595	A00057814	收回东六路恒天九五工厂地块，	2019/11/8 15:48:07	地区，这里潜力巨大，	0	44
279062	A00027836	议加大A7县东六线榔梨段拆迁力	2019/1/17 19:25:45	有土地（正钢机械厂	1	42
267630	A000100648	地铁3号线松雅湖站点附近地下	2019/5/22 23:37:38	越东四路和东四路西侧	0	42
209742	A00012969	区郝家坪小学什么时候能改扩建	2019/3/24 21:07:12	民小学、潇湾路小学、	0	41
239670	A00080329	东六线以西泉塘昌和商业中心以	2019/1/11 15:46:04	原置业有限公司厂房，	0	41
288398	A00053962	决出让星沙滨湖路以南，特立路	2019/2/11 14:09:40	之间有大量的闲置土地	5	40
257376	A909155	关于加快修建A市南横线的建议	2019/5/10 18:01:52	交通落后的现状，拉直	0	39
244178	A00057874	四号线北延线“同心路站”设在	2019/1/30 23:59:12	心路站设在雷峰大道	0	38
205217	A00040562	落实发放原A市七中01年后退休	2019/10/29 12:42:17	区下放教师待遇有关问	0	33
193286	A000103197	线A7县松雅西地省站西北方向	2019/4/17 11:13:12	路北侧穿越东四路和	0	32

图 4-10 部分点赞数降序排列图

4.3 问题三模型的建立与求

4.3.1 问题三模型的建立

（1）问答系统的总体架构及思路如下：

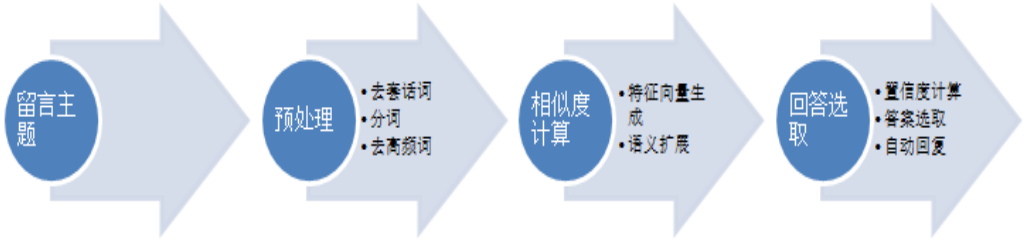

图 4-11 问题三系统架构及思路图

其中我们将群众对于答案的社会情感态度特征引入答案质量评价。另外，不同的群众受到认知、需求和兴趣这些特征影响，答案质量评价过程中还需要考虑用户自身的特征，使得筛选的答案更满足用户个性化需求[7]。

本题将群众社会情感和用户自身特征引入答案质量的评价，将评价指标分为 3 个维度，分别是相关性、完整性、可解释性。初步假设了 24 个评价指标，如下表所示：

表 4-2 混淆矩阵分类模型指标

维度	指标	解释及说明
答案文本特征	文本长度	答案文本包含的字符数。答案文本的长度越长,答案越丰富和完整
	关键词数量	答案文本中包含的关键词数量
	句子数量	答案文本中包含的句子数量
	停用词数	答案文本中包含的通用词数量,停用词数量越少,质量越高
	问题与答案耦合度	提问问题与答案之间的重叠部分,文本长度之比
	外部链接数量	答案文本中包含的超链接的数量
	段落数	答案文本的段落数
	问题答案长度比	问题长度与答案长度的比值
回答者特征	最佳答案数量	回答者的所有答案中被选为最佳答案的数量
	回答问题数量	回答者的所有回答的数量,表明回答者的经验和参与积极性
	用户权威性	回答者的社区等级(积分),表明专业程度和影响力
	提问数量	回答者提问问题的数量
时效性	答案的相对回答次序	答案在所有答案中的相对位置
	答案与问题生成间隔时间	回答时间与提问时间的间距
用户特征	用户学历水平	提问者的专业水平和学历程度
	用户提问数量	提问者以往提问问题的数量
	用户偏好与答案耦合度	用户的习惯,个人偏好信息需求与答案的关联性
	用户等级	提问者的权威性和影响力
社会情感	情感特征词数量	答案文本中包含的情感词的数量
	回答者情感态度	答案文本呈现出的回答者情感态度倾向性
	赞同数量	答案被赞同/支持的数量
	反对数量	答案被反对/踩的数量
	评论互动数量	答案被评论的数量
	关注关系	回答者与提问者的好友关系

（2）基于回归模型的生成回答质量自动化评价

由于需要我们预测将来群众们会发些什么群留言，那么这个任务便是回归任务。其中用到回归算法的评价指 SSE。

$$SSE=\sum_{i=1}^{n} w_i(y_i-\hat{y}_i)^2 \tag{4-14}$$

其中 SSE 越接近于 0，说明模型选择和拟合更好，数据预处理越成功。

4.3.2 问题三模型的求解

首先进行数据采集和预处理操作，通过留言主题匹配答复意见。其次面对群众提问数量为相同的，对于输出没有影响，可以不予考虑，只用于多个群众之间的比较分析。最后通过二八定律，即训练样本数为总样本数占八，测试样本数为总样本数占二。所以将附件四简化。

本题为了解决政府生成“智慧”的评价问题，针对存在的评价指标体系不全面、模糊性和缺乏个性化等问题，引入社会情感特征和群众特征维度，运用相似度构建答案质量评价指标体系。基于回归模型设计了答案质量自动化评价方法。应用结果表明本研究构建的评价指标体系和评价方法具有一定的合理性和有效性。但是研究仍然存在一定的不足，首先应用研究样本选取具有一定的局限性，话题内容也比较单一，没有进一步地将方法拓展到各个领域和类。

五．参考文献

[1] 周泉锡．常见数据预处理技术分析[J]. 通讯世界,2019,26(01):17-18.
[2] 吴巧玲.中文分词算法在自然语言处理技术中的研究及应用[J].信息与电脑(理论版),2011(12):39-40.
[3] 贺飞艳,何炎祥,刘楠,刘健博,彭敏.面向微博短文本的细粒度情感特征抽取方法[J].北京大学学报(自然科学版),2014,50(01):48-54.
[4] 黄章益,刘怀亮.一种基于语义的中文文本特征降维技术研究[J]. 情报杂志,2011(S2):123-125.
[5] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008(01):48-61.
[6] 李琼阳，田萍．基于主成分分析的朴素贝叶斯算法在垃圾短信用户识别中的应用[J]．数学的实践与认识，2019，49(01):136-140.
[7] 郭顺利，张向先，陶兴,等．社会化问答社区用户生成答案质量自动化评价研究——以"知乎"为例[J]．图书情报工作，2019(11).