智慧政务中的文本挖掘应用

摘 要

近年来，网络问政平台如雨后春笋，迅速发展，成为政府了解民意、汇聚民智、凝聚民气的重要渠道。因此，将自然语言处理和文本挖掘技术应用于智慧政务当中，是非常有意义的。本文主要内容如下：

针对问题一：首先将一级标签数值化，用数字 1~7 表示 7 种一级分类，再对留言主题和留言详情的文本数据进行去空去重、结巴分词、停用词过滤等操作。数据预处理之后，将留言主题和留言详情分别对应的词列表进行拼接，然后利用 TF-IDF 方法生成权重矩阵，最后利用支持向量机（LinearSVC）、多项式朴素贝叶斯、逻辑回归等三种模型对留言文本进行一级标签分类。

针对问题二：首先对留言主题和留言详情的文本数据进行去空去重、结巴分词、停用词过滤等数据预处理操作。再利用 TF-IDF 方法计算每个词的 TF-IDF 值，并进行排序，提取每一条留言数据中排名前 5 的关键词，同时生成对应的权重矩阵。最后通过 K-means 算法对留言文本进行聚类。采用 pyhanlp 工具对留言详情内容进行文本摘要，调用百度 API 识别留言主题和留言详情中的地点和人群。设计留言问题的热度评价模型，通过热度评价公式计算每一类留言问题的热度指数，并按热度指数进行排序，提取排名前 5 的留言类，将其作为热点问题，汇总得到 “热点问题表.xls”和“热点问题留言明细表.xls”这两个表格。

针对问题三：针对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、答复时间等角度对答复意见的质量制定一套评价方案。

关键词：多类分类,文本聚类,文本摘要,热度评价,答复质量评价

Text mining application in smart government

Abstract

In recent years, the network political platform has sprung up rapidly, becoming an important channel for the government to understand public opinion, gather people's wisdom and rally people's spirit. Therefore, it is very meaningful to apply natural language processing and text mining technology to intelligent government affairs. The main content of this article is as follows:

For question one: First of all, the first level of label numericalization, with the

number 1 to 7 for 7 first-level classification, and then the message subject and message details of the text data to empty weight, stop word filtering and other operations. After data pre-processing, the message topic and message details of the corresponding word list to stitch, and then use TF-IDF method to generate a weight matrix, and finally use the support vector machine (LinearSVC), polynomial simple Bayes, logic regression and other three models of the message text for a level label classification.

For question two: First of all, the text data of the message subject and message details to empty weight, word segmentation, de-checkword filtering and other data pre-processing operations. The TF-IDF method is used to calculate the TF-IDF value for each word and sort it, extracting the top 5 keywords in each message data, and generating the corresponding weight matrix. Finally, the message text is clustered by K-means algorithm. Using pyhanlp tool to summarize the content of message details, call Baidu API to identify the message topic and message details in the location and crowd. Design the heat evaluation model of message questions, calculate the heat index of each type of message question by the heat evaluation formula, and sort according to the heat index, extract the top 5 message classes, as a hot issue, summarized the "hot issue table.xls" and "hot issues message schedule.xls" two tables.

In response to question three: In view of the relevant departments to the reply to the message, from the relevance of the response, integrity, interpretability, response time and other points of view of the quality of the response to the response to the quality of a set of evaluation program.

Keywords: Multi-category classification, text clustering, text summary, heat evaluation, response quality evaluation

目录

1 引言

1.1 问题背景

随着网络应用的普及，出现了微信、微博、市长信箱等多种网络问政平台，网络问政对政府了解民情、汇聚民智有着积极的作用。智慧政务充分利用大数据、人工智能等新技术，对线上留言进行分类和热点整理，对民众的诉求进行科学的分析和判断，利于政府做出高效智能的回应和解决社会中存在的问题，并对其回应内容的价值做出评价，反映政策的运行效果，便于科学地改进决策[1]。

1.2 问题重述

(一)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。

(二)定义热点问题。根据附件 3 将某一时间内反映特定地点或特定人群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”。

(三)针对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，并尝试实现。

1.3 本文主要工作

本文基于问题中所给的分析角度，对每个问题附件中的数据进行分析和整理，并针对题目中要解决的问题进行建模和评价，具体步骤如下：

(1)数据预处理：对附件 2、3、4 中提供的留言主题、留言详情进行去空去重、分词、停用词过滤等操作，其中去重是针对留言详情相同的数据。

(2)文本分类：数据预处理之后，利用 TF-IDF 生成留言主题、留言详情的权重矩阵，利用线性支持向量机（LinearSVC）、多项式朴素贝叶斯、逻辑回归等三种机器学习模型对文本进行分类。

(3)热点问题挖掘：首先，利用 TextRank 方法提取每一条留言数据的 top5 的关键词，并用 TF-IDF 生成相应的权重矩阵；其次，通过 k-means 算法对留言文本聚类；然后对留言主题和留言详情进行命名实体识别和文本摘要等操作，得到与留言问题有关的“地点/人群”及“问题描述”；最后，构建热度评价模型，利用设计的热度公式计算每一类留言问题的热度指数，依据热度指数进行排序，提取排名前 5 的留言问题，作为热点问题。

(4)留言回复评价：基于答复意见的相关性、完整性、可解释性、答复时间等四个指标构建答复意见质量的评价模型。

总体流程图如图 1 所示：

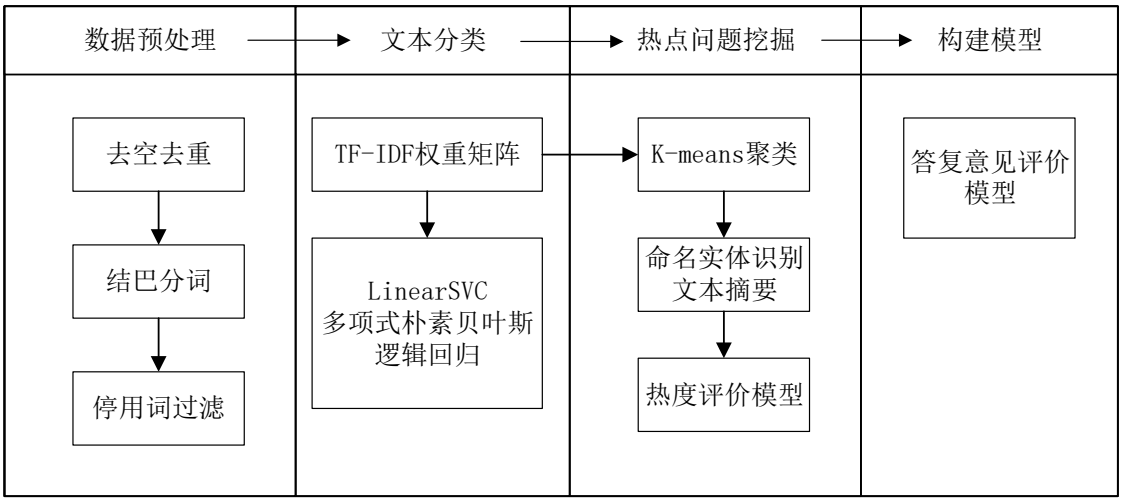

图 1 总体流程图

2 群众留言分类

对于问题一，首先对留言主题、留言详情、一级标签进行数据预处理，然后使用词频-逆向文本频率（TF-IDF）方法抽取文本特征并计算权重，最后通过线性支持向量机（LinearSVC）、多项式朴素贝叶斯、逻辑回归等三种模型对留言进行多类分类。本文根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

2.1 主要流程

问题一主要的流程如图 2 所示：

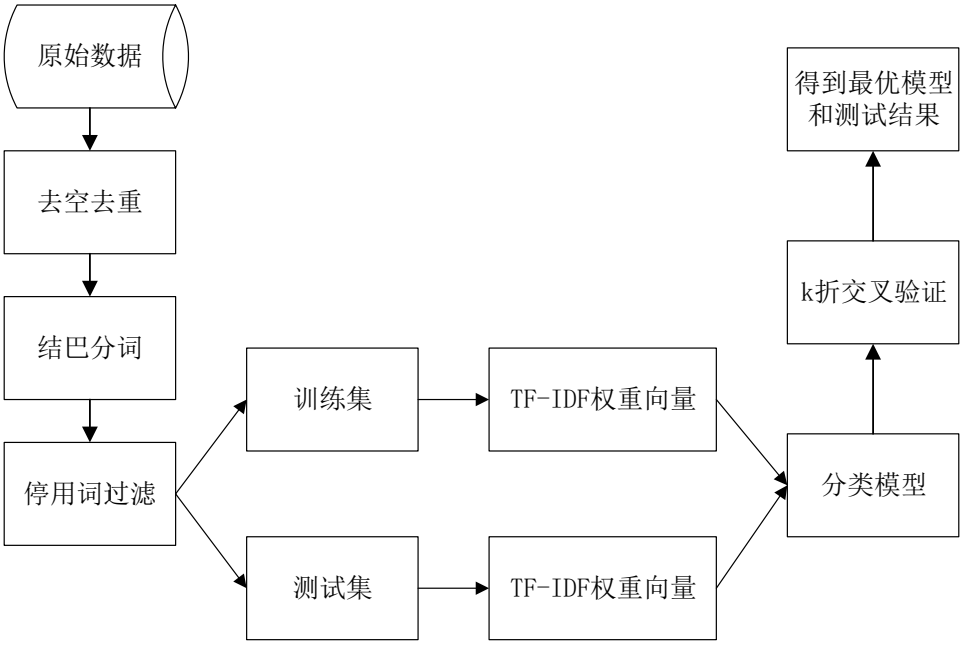

图 2 群众留言分类流程

根据图 2 中的群众留言分类流程，本题的主要研究内容如下：

导入附件 2 的数据,抽取其中的“留言主题”、“留言详情”、“一级分类”，并连接成二元表。使用 jieba 分词工具对留言主题，留言详情进行分词，去停用词得到数据预处理结果，然后将留言按 8.5:1.5 分为训练集和测试集，使用 TF-IDF 特征加权算法将训练集和测试集映射成相应的词向量矩阵。将训练集词向量矩阵放入分类模型中训练，得到训练模型，最后将测试集词向量矩阵放入训练模型，得到预测的一级分类结果，计算精准率、召回率、F1-Score 值，评估分类模型。

2.2 数据预处理

对附件二中的留言文本数据进行去空去重、中文分词、去停用词等数据预处理操作。

2.2.1 留言信息的去重

在附件 2 中，存在一些留言详情相同的留言，为节省计算时间和资源，去除相同项。去重前后留言总数对比情况如表 1 所示：

表 1 去重前后留言总数对比情况表

去重前留言总数	去重后留言总数
9210	9092

2.2.2 一级标签数值化

将附件二中的留言主题、留言详情、一级标签按行拼接，一级标签包含城乡建设、环境保护、劳动和社会保障等 7 个类别，用数字 1~7 依次表示这 7 类，将留言所属类别数值化，方便统计和分析。一级标签数值化结果如表 2 所示：

表 2 一级标签数值化结果表

一级分类	类别数值化
劳动和社会保障	1
城乡建设	2
教育文体	3
卫生计生	4
交通运输	5
商贸旅游	6
环境保护	7

2.2.3 中文分词

词是中文文本的最小单位，将长文本分割为词表，在词层面上处理文本数据，便于转换为词向量矩阵，可进一步提高模型的准确率。目前，常用的分词库有 jieba 分词、哈工大 NLP 分词等，本文选用的是 jieba 分词库，利用 jieba 分词处理留言文本得到分词后的词列表，jieba 分词的基本算法原理如下：

(1)基于 Trie 树结构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）。

(2)采用了动态规划查找最大概率路径，找到基于词频的最大切分组合。

(3)对于登录词,采用了基于汉字成词能力的 HMM 模型,使用了 Viterbi 算法。部分分词结果如图 3 所示：

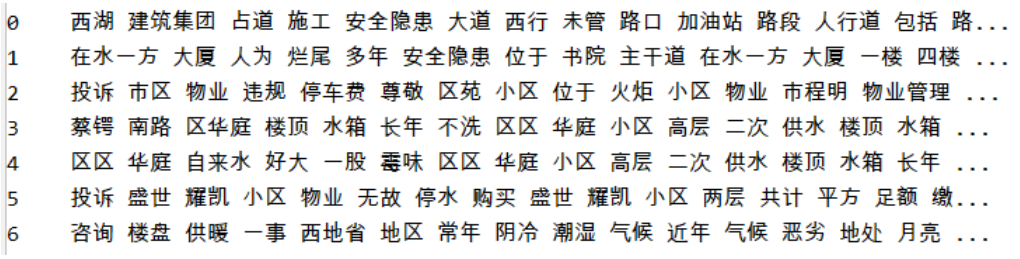

图 3　部分分词结果

2.2.4 去停用词

文本分类的质量相当一部分取决于文本词或字的重要性，过多的噪音不利于文本分类，这些噪音的主要载体是停用词。在尽量不改变文本原意的情况下，去除停用词，可以提高文本分类的质量。停用词，一般为留言文本中大量无意义的词，如“的”、“了”、“是”等语气助词、连词、介词及标点符号。去停用词不仅节省存储空间，并且对模型的训练效率也有提高。本文将常用的停用词表和适应本题目的自定义用户停用词表合成一个停用词表，对分词后的文本进行去停用词操作，生成词列表。

2.2.5 构建词频矩阵

在得到分词结果后，利用词袋模型构建出每条留言文本的词频矩阵。词袋模型的核心思想是将文本看作是一系列词语的集合，之后将集合中所有不重复的词语组成一个全局的词典库，因此文本中的所有词语都可以在词库中找到索引值，每个索引位置的取值为文档中此词语出现的次数。本文利用词袋模型将文本向量化，向量表示如下：

$$C = (t1, t2, \ldots, tn) \qquad (2\text{-}1)$$

其中 C 是词袋集合，t_n为每个词的特征值。

2.2.6 计算权重策略（TF-IDF）

在 VSM 模型中，一般采用 TF-IDF 来计算文档中每个特征词的权重。本文也采用这种方法,该方法基本原理如下：

在 TF-IDF 中，单词的重要性由两个因素共同决定，它与它在文档中出现的次数成正比，但它随着语料库中出现该词的频率越多而下降。

(1)词频

在某一文档中，词频（TF）是指词在文档中出现的次数，基本公式：

$$\text{词频（TF）} = \frac{\text{某个词在文本中出现的次数}}{\text{文本总次数}} \tag{2-2}$$

(2)逆文档频率

逆文档频率（IDF）是衡量单词总体重要性的指标，其值等于文档总数除以包含该单词的文档数量的商再取其商的对数，基本公式：

$$\text{词的逆文档频率（IDF）} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数+1}}\right) \tag{2-3}$$

(3)TF-IDF

某词在某文档中是高词频，而在整个文档集中，该词又是低文档频数，那么该词可以得到一个较高权重的 TF-IDF 值。因此，TF-IDF 有助于降低常见的词语特征。向量获得方式为首先统计出所有的词，把每个词当成向量的每一个维度，如果该文档中有某词，就在某词的维度上计算它的 TF-IDF 值；如果不存在某词，那么某词的维度上的值就为 0。用这种方式对所有的留言进行特征提取，提取的结果是一个稀疏矩阵。为了避免某词可能从来都没出现在所有的文档中，而导致被除数为零，一般分母用（包含该词的文档数+1）代替。 最后可以计算的出每个词的 TF-IDF 值为：

$$TF - IDF = TF \times IDF \tag{2-4}$$

使用 TF-IDF 方法计算词的权重，TF-IDF 公式如下：

$$W(t,d)=\frac{tf(t,d)\times\log(\frac{N}{n_t}+0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log(N/n_i+0.01)]^2}} \quad (2\text{-}5)$$

其中，$W(t,d)$为词 t 在文本 d 中的权重，而 $tf(t,d)$ 为词 t 在文本 d 中的词频，N 为训练文本的总数，n_t 为训练文本集中出现 t 的文本数，分母为归一化因子。利用 TF-IDF 计算权重之后，生成留言文本对应的特征矩阵。

2.3 留言文本分类

在得到留言文本的向量化表示之后，初步选择线性支持向量机 、多项式朴素贝叶斯[2]、线性回归[3]三种分类模型对留言文本进行分类，利用 k 折交叉验证的方法，对模型进行十次测试取其十次评估结果的平均值作为评估结果，评估结果如表 3 所示：

表 3 分类模型评估结果

MODEL NAME	F1-Score	Precision	Recall
LinearSVC	0.901124	0.901774	0.901124
MultinomialNB	0.898381	0.902586	0.898381
LogisticRegression	0.868551	0.876158	0.868551

LinearSVC 是 SVM 中的一种常用来处理多类分类问题的模型，LinearSVC 基于 o-v-r 算法，将多类分类转为多个二分类，利用多个二分类器和树模型，解决线性可分的多标签分类问题，由表 3 可知，LinearSVC 与 MultinomialNB、LogisticRegression 进行对比，其分类效果最优，所以我们选择 LinearSVC 作为最终的分类模型。

2.3.1 支持向量机算法

由于本题将线性支持向量机作为最终的分类模型，因此，简单介绍一下支持向量机的基本思想：大部分情况下，并不是全部分类样本都是线性可分的，因此第一步需要将原始空间转变为一个高维度空间，而这个转变的过程是通过选取一个合适的核函数完成的非线性转换，然后在这个新的空间中寻求最优解[4]。支持向量机基本原理如下：

当输入m个初始样本$(x_1, y_1), \cdots (x_2, y_2), \cdots (x_m, y_m), x \in R$，样本空间使用Rh来表示，空间维数使用h进行表示，这些所有样本被分为两类。所以分类问题一般情况下可以被描述成通过一定的方法搜索到可以把所有样本分为两类的最优超平面，这个超平面为：

$$w \cdot x + b = 0 \tag{2-6}$$

其中$w \cdot x$是内积，b是标量。实际上存在很多能够满足这样条件的平面，它们也能够将这些样本无误差地分为两类，此时对给定样本的误差值可以满足最低，但是其中只存在一个分类平面满足 SLT 中的结构风险最小化原则，这样的平面就是最优超平面，可以使得两类的间隔最大,最优分类平面在平面直角坐标系中的表示如图4所示：

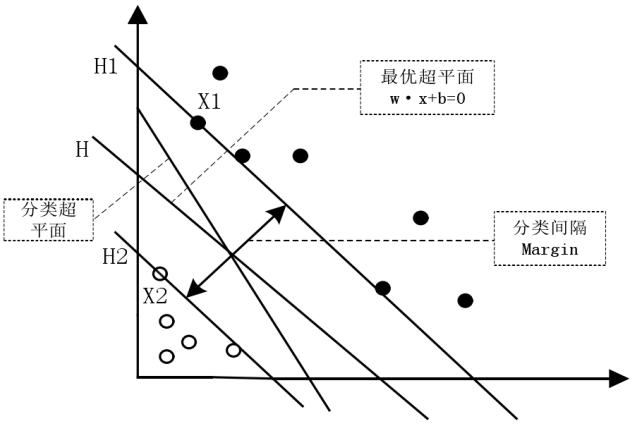

图4　最优分类平面

图 4 中，H1 平面上及其右边的点表示 w • x + b ≧1 的样本集，H2 平面上及其左边的点表示 w • x + b ≦-1 的样本集，H 为最优超平面，它使得每类距离超

平面最近的样本到超平面的距离之和最大。距离这个最优超平面最近的全部样本被称为支持向量，在图中表示为 H1 平面和 H2 平面上的所有样本，即满足 w • x + b =1 和 w • x + b =-1 的全部样本，两个平面之间的垂直距离叫做分类间隔（Margin），其表示公式如下：

$$\text{Margin} = \frac{2}{||w||} \tag{2-7}$$

最优超平面的目的就是在把所有样本非常准确地分成两类的同时，需要满足分类间隔最大，即保证分类器对给定样本误差最小的同时，还要满足置信水平最低，这样得到的分类器对于未知样本的分类误差很大程度上有可能满足最小。

但在本题中，原始的样本空间可能不存在将样本线性划分的超平面，即非线性分类问题。SVM 中的参数——核函数，将原始空间的样本映射在高维空间中，使其能在高维空间线性可分。常见的核函数有线性核函数、多项式核函数、径向基核函数、Sigmoid 核函数。我们采用的线性支持向量机利用的核函数就是线性核函数。

利用核函数实现非线性转换，如图所示：

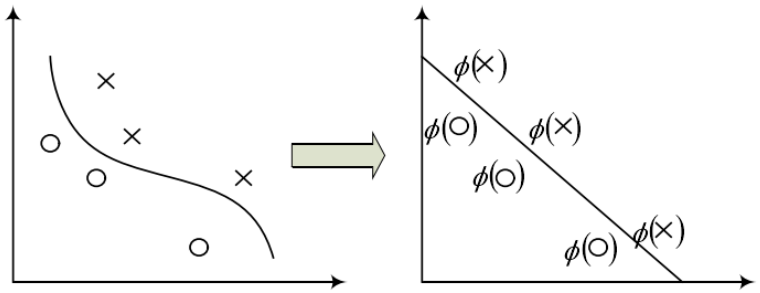

图 5 非线性转换

在图 5 中，左边是样本集在原始空间中的分类情形，右边是样本集映射到高维空间中的分类情形，这个转变的过程是通过选取一个合适的核函数完成的非线性转换，然后在这个新的空间中寻求最优解。这种转变过程可表示为：

$$x \to \emptyset(x) = (\emptyset_1(x), \emptyset_2(x), \cdots \emptyset_i(x) \cdots) \tag{2-8}$$

同时，经过非线性转换映射到新空间中的最优超平面方程相应地转化为：

$$W \cdot \emptyset(x) + b = 0 \tag{2-9}$$

将$\emptyset(\mathrm{x})$带入 SVM 原始模型，并将其转换为状态下的对偶问题，则样本 Xi，Xj 的内积计算为：

$$\mathrm{K}(x_i, x_j) = \emptyset(x_i) \cdot \emptyset(x_j) \tag{2-10}$$

其中，$\mathrm{K}(x_i, x_j)$为内积的核函数。

则 SVM 中的分类函数为：

$$\mathrm{f(x)} = \mathrm{sign}(\sum_{\mathrm{i,j=1}}^{\mathrm{n}} \alpha_i y_i (\emptyset(x_i)^T \emptyset(x_j)) + b) \tag{2-11}$$

由上述支持向量机原理可知，支持向量机（SVMs）利用核函数的特性完成非线性转换，是一种经典的解决二分类问题的模型，但本题中留言内容对应的一级标签类别共有 7 类，简单的二分类模型难以解决多类分类问题。基于将多分类转为多个二分类的思想，采用基于 1-v-1 (One-versus-One) 多类支持向量机算法[5]的 LinearSVC 模型对留言进行多类分类，其思想是在每两个类间训练一个分类器，则对于一个 k 类问题，需要有 k(k-1)/2 个分类函数。当对一个未知样本分类时，每个分类器都对其类别进行判断，并为相应的类别投票，最后将得票数最多的类别作为该未知样本的类别。

2.4 模型评测指标

精准率（查准率）、召回率（查全率）是衡量机器学习模型性能的基本评估指标，而 F-Scores 是综合考虑精准率和召回率的评估指标。通过这三项指标能比较准确的评估一个分类模型的好坏。以下是这三类指标的计算公式：

(1)召回率(Recall)Rj

$$R_j = \frac{TP_j}{TP_j + FP_j} \tag{2-12}$$

(2)查准率(Precision)P_j

$$P_j = \frac{TP_j}{TP_j + FN_j} \tag{2-13}$$

（3）F-Score

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_jR_j}{P_j+R_j} \tag{2-14}$$

2.5 评估结果分析

LinearSVM 模型预测精度达到 0.92，下表为 7 类一级标签相应分类结果的评估，从精准率、召回率、F1 值分析，只有一类的三个指标结果在 0.8 到 0.9 之间，其他类的三个指标结果在 0.9 以上。linearSVC 分类报告如下：

表 4　linearSVC 分类报告

	precision	recall	f1-score	support	预测精度
1	0.95	0.93	0.94	309	0.9200879765395894
2	0.90	0.92	0.91	290	
3	0.94	0.95	0.94	238	
4	0.90	0.94	0.92	124	
5	0.85	0.82	0.84	80	
6	0.91	0.89	0.90	175	
7	0.95	0.95	0.95	148	
accuracy			0.92	1364	
macro avg	0.91	0.91	0.91	1364	
weighted avg	0.92	0.92	0.92	1364	

3 热点问题挖掘

一般地，将某一时段内群众集中反映的某一问题可称为热点问题。及时发现热点问题，有助于相关部门进行有针对性处理，有效洞察政策落实情况，利于进一步完善相关政策,提高服务效率。在本节中，结合 TF-IDF 算法和 K-means 算法对留言进行聚类。然后对留言详情进行中文命名实体识别、文本摘要操作，将其结果分别作为留言问题的“地点/人群”和“问题描述”属性。最后针对留言问题的关注度、反映度、时间范围等三个方面，建立热度评价模型，基于该模型，提取热度排行 top5 的留言问题类。

3.1 留言聚类

3.1.1 数据预处理

使用 pyhanlp 工具分别提取每一条留言主题和留言详情重要性前五的关键字，然后用所有的关键字生成词袋模型，最后用 TF-IDF 算法得到特征矩阵。

3.1.2 基于 K-means 算法聚类

对附件 3 中的留言主题和留言详情文本进行数据预处理操作，得到特征矩阵之后，使用 k-means 模型对留言问题进行聚类。K-means 聚类的基本原理[5]如下：

算法是输入聚类个数 k，以及包含 n 个数据对象的数据，输出满足方差最小标准 k 个聚类的一种算法。k-means 算法将 n 个数据对象划分为 k 个聚类：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。k-means 算法的简要步骤：

（1）随机初始化 k 个点作为聚类质心。

（2）计算数据对象到质心的距离，并将数据对象分配到距离最近的一个簇。

（3）针对每一个聚类，计算簇中所有点的均值并将其作为新的质心。

（4）反复迭代，当数据对象距离最近簇的距离不再变化或满足一定的条件时，停止计算，返回 k 个聚类的质心。

K-means 聚类的流程如图 6 所示：

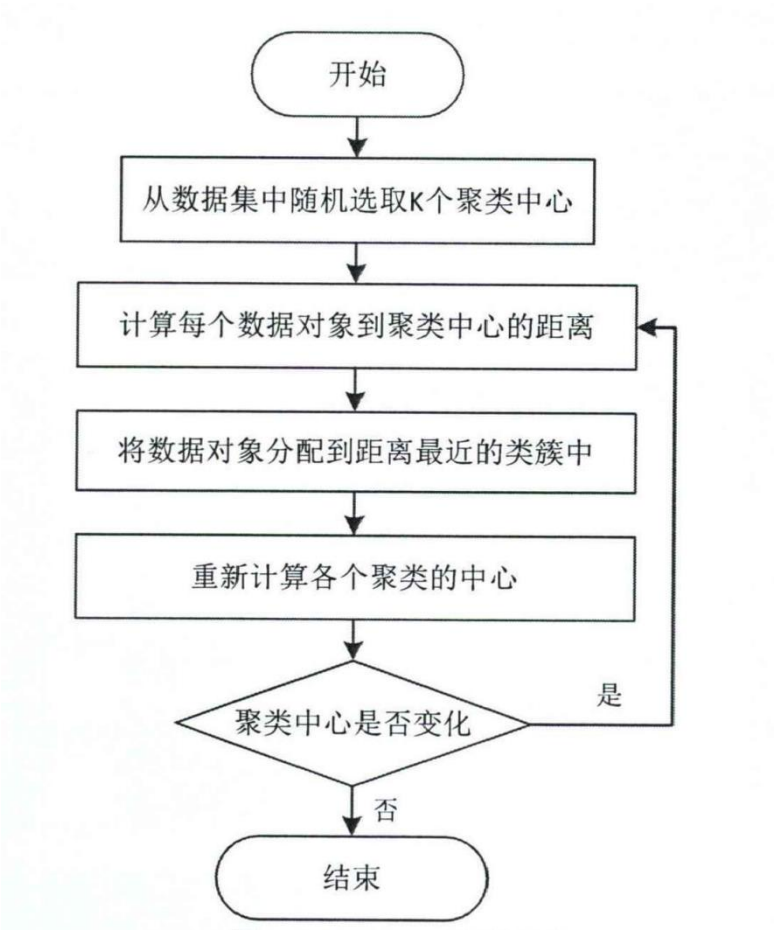

图 6　k-means 聚类流程

附件 3 中给出了 4326 条留言记录，经过了去空去重处理之后，样本留言记录还有 4325 条文本记录。由于文本记录的数量较多，若直接将留言主题和留言详情拼接后的 TF_IDF 特征矩阵放入 k-means 模型中进行训练，导致计算量大，且计算的效率不高，因此利用 TF-IDF 方法对关键词进行排序，提取 top 5 关键词，将 5 个关键词对应的特征矩阵放入 k-means 模型训练，共得到 2363 个留言问题类别，部分聚类结果数据如图 7 所示，全部数据见附件中的聚类结果.xlsx 文件。

	问题id	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数	分类	类留言总数
1648	1	227477	A0004924	A4区青竹	2019/11/1	A市A4区A	0	0	0	2
2317	1	243596	A0009851	A市青竹湖	2020/1/5	本人在A市	0	0	0	2
1242	2	217811	A0004454	A7县蟠龙	2020/1/2	蟠龙路废	0	0	1	2
4007	2	283298	A0003011	A市地铁6	2019/9/19	我是金澎	0	0	1	2
235	3	194022	A0004210	坚决反对	2019/7/8	我们是诺	1	0	2	6
1105	3	214546	A0007409	A市鹏基诺	2019/12/2	众所周知，	0	0	2	6
1631	3	226871	A000726	坚决反对	2019/8/16	反对理由	1	0	2	6
2680	3	252300	A0007409	A市鹏基诺	2019/12/2	尊敬的领	1	0	2	6
2685	3	252413	A000726	坚决反对	2019/8/16	爱尔眼科	0	0	2	6
2942	3	258186	A0004210	反对在A7	2019/7/8	我们是诺	3	0	2	6
1530	4	224465	A0003608	A7县黄星	2019/4/3	A7县黄星	0	0	3	1
159	5	192027	A0002022	反映A3区	2019/2/20	A3区A3区	0	0	4	4
750	5	206318	A0004136	反映A3区	2019/3/1	1。征收承	0	0	4	4
1394	5	220984	A0009614	A7县时代	2019/3/1	物业公司	4	1	4	4
4284	5	289772	A0004136	A3区A3区	2019/3/19	1、征收承	5	0	4	4
2823	6	255276	A909219	再次希望	########	尊敬的张	0	0	5	2

图 7　部分聚类结果

3.2 命名实体识别

命名实体识别(NER)是指识别文本中具有特定意义的实体，主要包括人名、地名。在问题二中，需要提取 top 5 的热点问题对应的地点和人群，本文调用百度 NLP 命名实体识别 API 对每一类热点问题的留言主题和留言详情文本的集合进行词性标注，并提取词性标注满足‘LOC’、‘ORG’、’PER’、’ns’、’nt’、’nr’其中一项的实体，存入实体列表，每一类热点问题对应实体共用一个 ID， 统计实体词频，选取词频最高的一个地点或一个人群实体，作为留言的“地点/人群”属性。识别的部分地点数据如图 8 所示：

```
 '保利麓谷林语',
 '地铁6号线',
 'A2',
 '长大建设集团福满新城彻夜',
 'A3区',
 '金星路'],
['A5区', '桃花苑二期', 'A5区', '汇金路', '五矿'],
['涉外经济学院', '商贸旅游职业技术学院', 'A市经济学院', 'A市经济学院', '市经济学院体育学院'],
['金毛湾', '卓越浅水湾']]
```

图 8 部分地点数据

3.3 文本摘要

利用 pyhanlp 工具进行文本摘要。将留言主题、留言详情的文本摘要作为问题描述。抽取式摘要建模为句子排序任务完成，句子排序任务是针对每个句子输出其是否是摘要句的概率，最终依据概率，选取 top k 句子作为最终摘要，在本文中，选取 top2 或 top3 句子作为问题描述。 部分文本摘要数据如图 9 所示：

```
['物业对未交装修押金的业主进行停水处理。那物业有何权利对我们二期全部房屋进行停水。我们质疑电梯硬件出现问题。',
 '而我们小区的群租房一直是无人问津的状态。物业说管不了。小区的保洁也是一直跟物业投诉然后也是一直无果。'],
['学校开始组织学生参加实习。当然学生是必须实习。学校要求学生必须去学校安排的几个点实习。我知道我们必须实习。',
 '学校开始组织学生参加实习。当然学生是必须实习。学校要求学生必须去学校安排的几个点实习。我知道我们必须实习。',
 '学校开始组织学生参加实习。当然学生是必须实习。学校要求学生必须去学校安排的几个点实习。我知道我们必须实习。',
 'A市经济学院强制16届电子商务跟企业物流专业实习。其中我们企业物流专业实习6个月。',
 '系里要求我们在实习前分别去指定的不同公司实训。我这的公司的工作内容和老师之前介绍以及我们专业几乎不对口。'],
['我是梅溪湖金毛湾的一名业主。A市教育局暂未将金毛湾楼盘纳入配套入学。', '请相关部门为百姓发声。严惩金毛。']]
```

图 9 部分文本摘要

3.4 热度评价模型

3.4.1 热度评价模型设计

为了挖掘出留言中的热点问题，对聚类之后的留言类，结合留言类的点赞数和反对数，时间，留言总数等设计了一个热度评价模型。

评价热度模型的设计，包括以下三个指标：

(1)群众对某一类留言问题的关注度

附件 3 中给出了留言的点赞数和反对数，考虑到点赞数和反对数较少，有的点赞数、反对数为 0，同时从群众关注的方面看，点赞数和反对数都可以反应群众的关注程度，从而体现留言问题的热度。假设某一类留言点赞数为 m、反对数为 n，统计所有留言的点赞数和反对数之和，存入集合 A 中，并提取集合中最大的数值，

记为 a。因此，$\frac{m+n}{a}$可作为考察群众对某一类留言关注度的指标。

(2)群众对某一类留言问题的反映度

某一类留言问题的总条数越高，表明群众在该类问题的反映积极程度越高，相应地，该类问题的热度也越高。假设某一类留言的总条数为 t，统计每一类留言问题的留言总数，存入集合 B 中，并提取集合中最大的数值，记为 b。那么，$\frac{t}{b}$可作为考察群众对某一问题反映程度的指标。

(3)某一类留言问题的持续时间长度

某一类留言问题的持续时间范围长度越长，表明该类问题在这段时间内没有得到很好的解决，因此该类问题也越应该要受到群众和政府的关注，将每一类留言的时间范围长度记为 h,统计每一类的留言的时间范围长度，存入集合 C 中，提取集合中的最大值，记为 c。那么，$\frac{h}{c}$ 可作为第三个指标。

自定义热度评价模型，根据热度指数对留言问题进行排名。考虑到有一些留言的点赞数、反对数为 0，因此将第二个指标的权重设置的高一掉，那么，以上三个指标对应的权重(1)、权重(2) 权重(3)的值分别设置为 0.4、0.3、0.3，综合以上所述热度评价公式如下：

$$\text{热度指数}=\frac{m+n}{a}\times \text{权重}(1)+\frac{x}{b}\times \text{权重}(2)+\frac{h}{c}\times \text{权重}(3) \qquad (3\text{-}1)$$

根据热度评价公式计算每一类留言的热度指数，并进行排序，提取排名前 5 的留言问题作为热点问题，并计算热点问题的时间范围。

3.4.2 表格汇总

将上述得到的 5 类热点问题的热度排名、问题 ID、热度指数、时间范围、地点/人群、问题描述等留言问题属性，存入附件“热点问题表.xlsx”，相关详情见附件“热点问题留言明细表.xlsx”。

部分结果如图 10、11 所示：

排名热度	问题id	热度指数	时间范围	地点/人群	问题描述
1	40	0.5034038	2019/01/23至2020/01/06	地铁6号线	晚上两三点左右还是施工的机器声和卡车的鸣叫声。严重影响周边居民的正常生活作息。
2	272	0.5001814	2019/01/21至2020/01/01	A4区	施工部门有夜间施工证。是不是有夜间施工证就可以夜间施工扰民。如果有夜间施工证和扰民都不处理。
3	1234	0.4473466	2019/03/22至2019/08/19	A5区	物业对未交装修押金的业主进行停水处理。那物业有何权利对我们二期全部房屋进行停水。我们质疑电梯硬件出现问题。
4	342	0.4395655	2017/06/08至2019/04/28	A市经济学院	A市经济学院强制16届电子商务跟企业物流专业实习。其中我们企业物流专业实习6个月。
5	739	0.4027853	2019/04/11至2019/09/25	金毛湾	我是梅溪湖金毛湾的一名业主。A市教育局暂未将金毛湾楼盘纳入配套入学。

图 10 热点问题表

问题id	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
40	259312	A00019079	A市地铁6号线桐梓坡地铁站交叉口低频率噪音严重扰民	2020/1/6 11:29:11	近2个月来，地铁6号线桐梓坡地铁站交叉口有设备发出怪异的低频率噪音，严重扰民，日夜不停。整个附近小区上空有这种低频	0	0

图 11 热点问题留言明细表

4 答复意见评价模型

对政府的答复意见做出合理的评价，有利于监测政府的答复效率，促进网络问政平台朝着良好态势发展，有益于群众问题得到高效解决。首先，评价答复意见要从相关性入手，即答复内容是否和留言问题相关，如果答非所问，那么答复就没有实际意义。然后，需要考虑答复意见的完整性，政府的答复是官方的，社会关注度高，一般答复具有统一的格式，如答复开头为问候语、答复引入语，结尾为感谢语、答复时间。再对答复的可解释性做出分析，一般具有可解释性的答复，会提到解决方案依据的法律或规定，若不是目前收到留言部门的管辖范围，会给出咨询其他部门的渠道或转给其他部门处理。最后，计算答复时间与留言时间的间隔，间隔越短，答复的时间效率就越高。因此，本节基于相关性、完整性、可解释性及答复时间构建答复意见评价模型，给这四项都分配 10 分，取总分的平均值为最终得分，并划分 ABCD 四个等级，等级 A 表示优，等级 B 表示良，等级 C 表示较差，等级 D 表示差。

4.1 基于相关性

答复意见的相关性即答复内容是否和留言问题相关，如果答非所问，那么答复就没有实际意义。合并留言主题和留言详情，通过对留言主题及详情、答复意见进行结巴分词、去停用词、word2vec 训练词向量等操作得到对应的特征词向量，最后计算留言主题及详情、答复意见对应的特征词向量间的余弦相似度，即为答复相关性的量化指标。

4.1.1word2vec 模型

word2vec 主要采用 CBOW(Continuous Bag-of-Words Model)和 Skip-Gram(Continuous Skip-Gram Model)两种模型[7]。无论是 CBOW 模型还是 Skip-Gram 模型，都是以 Huffman 树作为基础。Huffman 树中非叶节点存储的中间向量的初始化值是零向量，叶节点对应的单词的词向量是随机初始化的。CBOW 的目标是根据上下文来预测当前词语的概率，而 Skip-Gram 恰好相反，它是根据当前词语来预测上下文的概率。这两种方法都利用人工神经网络作为它们的分类算法。起初，每个单词都是一个随机 N 维向量，经过训练之后，利用 Skip-Gram 或者 CBOW 方法获得每个单词的最优向量。其原理图如图 12、13 所示：

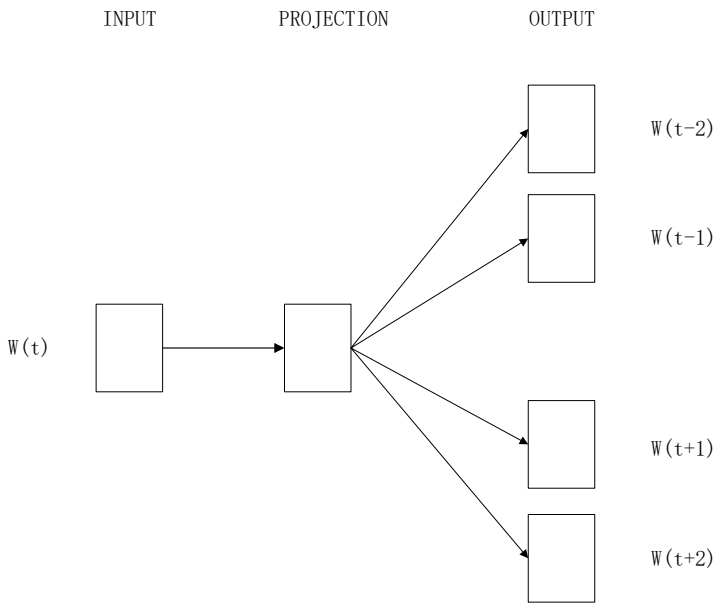

图 12　Skip-Gram 模型原理图

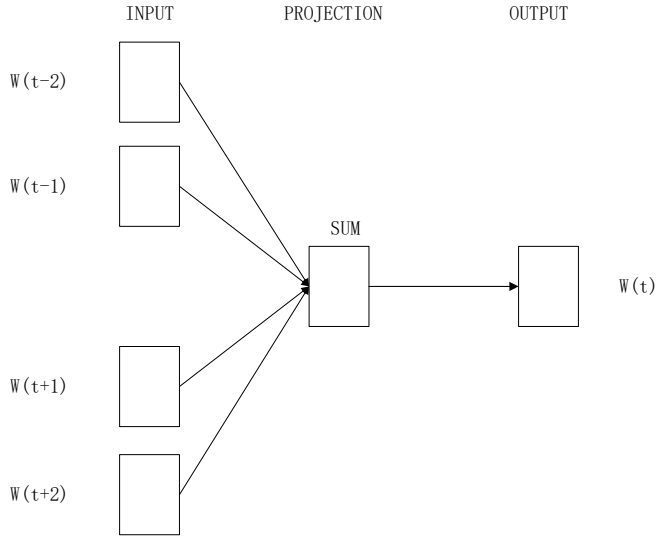

图 13　CBOW 模型原理

word2vec 模型在给定的语料库上训练 CBOW 和 Skip-gram 两种模型，然后输出得到所有出现在语料库上的词语的词向量表示。基于得到的词向量，可以计算词与词之间的相关性。考虑到语料库的选择，在训练 word2vec 模型时使用中文维基百科的语料，最终得到各个词的词向量，最后利用余弦相似度计算。

4.1.2 相关性的评分标准

将留言主题及详情、答复意见对应的特征词向量间的余弦相似度记为α，不同数值段的评分标准如表 5 所示：

表 5 相关性评分表

α	Score
α ≥ 0.6	10
0.50≦α＜0.60	9
0.40≦α＜0.50	8
0.30≦α＜0.40	7
0.25≦α＜0.30	6
0.20≦α＜0.25	5
0.15≦α＜0.20	4
0.10≦α＜0.15	3
0.05≦α＜0.10	2
α＜0.05	1

4.2 基于完整性

答复意见的内容是否完整取决于文本格式是否符合某种规范，因此对答复内容的一般格式做出分析。首先，对附件 4 中的答复意见文本内容进行粗略的观察，做初步分析，得出答复内容的一般格式为：

(1)开头部分：一般位于答复内容的第一个句子或前两个句子。基本格式依次为问候语、回复引入语，如“网友：您好！您于 3 月 4 日在平台书记问政发布的帖文已收悉，我镇高度重视，迅速调查核实，现回复如下”。

(2)结尾部分：一般位于答复内容的最后两个句子。祝福语、敬语“特此回复”出现的较少，感谢语和时间出现的较多，因此，基本格式依次为感谢语、时间，如“感谢您对我部工作的支持。 2018 年 11 月 9 日”。因此，完整的答复内容应同时包含问候语、回复引入语、感谢语、时间等具体内容。

4.2.1 主要步骤

(1)有无时间评分标准：对所有答复文本进行用正则表达式匹配留言回复文本尾部的时间，如果尾部有时间，则完整性评分加 2；若无时间，评分加 0。

(2)提取答复开头和结尾：过滤掉所有答复文本尾部的时间，按句号对所有答复文本进行分割，提取所有文本的第一个句子和最后一个句子。

(3)生成开头结尾关键词词库：对答复开头，答复结尾数据分别进行结巴分词，去停用词，生成两个词库，命名为答复开头词库，答复结尾词库。统计词频并排序，提取两个词库中的 top5 的关键词，将答复开头和结尾的关键词分别存入两个文件。

(4)答复开头和结尾部分评分标准：将所有留言回复的第一句与图 7 中开头关键词匹配，并将匹配的关键词个数记为 a；将所有留言回复的最后一句与图 7 中结尾关键词匹配，并将匹配的关键词个数记为 b。

评分标准如表 6 所示：

表 6 完整性评分标准

条件	得分
有时间	+2
无时间	+0
a>=3	+4
0<a<3	+2
a = 0	+0
b>=3	+4
0<b<3	+2
b = 0	+0

在答复详情开头和结尾部分分别提取排名前 5 的关键词如表 7 所示：

表 7 留言详情的开头和结尾部分的关键词

详情开头关键词	结尾关键词
您好	感谢您
网友	工作
收悉	支持
问题	我们
反映	理解

由表 7 中的关键词信息看出，过滤掉答复详情的时间后，提取其第一个句子和最后一个句子，并进行结巴分词、停用词过滤的数据预处理操作得到关键词词表，统计关键词的词频，按词频进行排行，提取 top5 的关键词。对于答复意见开头部分，其中“您好”是问候语、“网友”是对留言对方的称呼，“收悉”、“问题”、“反映”则表示留言中提出的问题已被相应部门收悉，这 5 个关键词可以从一定程度上，可以作为分析答复文本开头部分的完整程度的指标。同样的，对于答复意见结尾部分，其中“感谢您”、“工作”、“支持”、“我们”、“理解”这四个词可以体现政府相关部门对群众的感谢，群众对政府工作的支持和理解是政府工作人员的巨大精神支持，这五个关键词可以作为分析答复文本结尾部分完整程度的指标。

在得到每一条留言回复的完整性得分之后，对其总体完整性得分情况进行统计，绘制条形图，如下图：

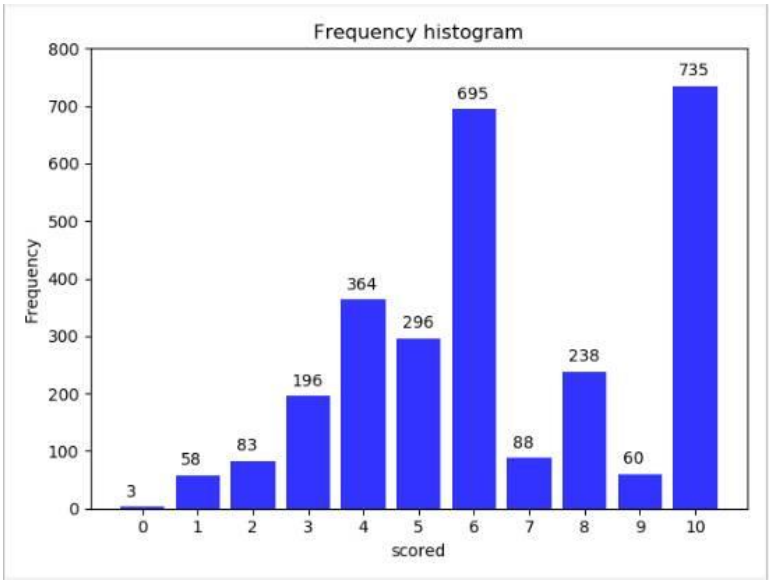

图 14　完整性得分情况

由上图可知，答复详情内容完整性得分为10的答复条数最多，表明政府做出的答复比较官方，答复内容及格式都比较完整。另外，答复数量为695的第二高峰的得分范围在5.5~ 6.5，而此范围正好为一个得分刚好合格的区间，此数据也反映了政府需要进一步完善在网络问政平台上的相关规定，不断促使答复详情内容和格式更加完整和规范。

4.3 基于可解释性

答复意见是否具有可解释性关键在于其内容是否对解决问题的方案提供了相关的解释，例如：依据相关的法律或规定提出解决方案；留言中的问题不属于当前收到留言的部门管辖，将该留言转交其他部门处理或提供其他部门咨询渠道。通过对答复意见内容进行文本语义分析，分析各类语素的使用频率，并建立政府答复意见专业高频语料表，以其中的top20关键词来反映答复的可解释性。其中，语素是指中文语言中最小的有意义的语言单位。用分词、词频统计等操作建立政府答复意见专业高频语料表。

该方法分为以下两个步骤：(1) 答复文本数据预处理及分析；(2)答复意见的可解释性评价标准。

4.3.1 答复文本数据预处理及分析

答复文本一般以句子或短文的形式出现，需要对其进行分词操作。使用“结巴”中文分词工具对答复文本进行分词，形成答复意见的语素库，然后利用汉语词频统计工具统计各项语素的频数与百分比。在词频分析中，利用TextRank算法[8]分析答复意见语素的重要性，其公式如下：

$$\mathrm{P}(V_i) = (1-\mathrm{d}) + \mathrm{d} \times \sum_{j \in \mathrm{ln}(V_i)} \frac{1}{|out(V_j)|} P(V_j) \qquad (4\text{-}1)$$

其中，P（V_i）代表语素i的重要性（PR值），d是阻尼系数，一般设置为0.85，In（V_i）是含有语素 i 的语素集合，Out（V_j） 是含有语素 j 中的语素集合，|Out

（V_j）| 是集合中元素的个数。通过 TextRank 算法实现答复意见的重要性由高到低进行排序。排序之后，过滤掉在分析完整性时采用的关键词，再提取 top20 的关键词，生成政府答复意见专业高频语料词表。

4.3.2 答复意见的可解释性评价标准

解释性不强的文本长度较短，有的答复本文只有时间、有的答复如“网友：您好！留言已收悉。”，此类答复不具有一定的可解释性。因此，设定答复意见文本长度阈值为 20，长度低于 20，文本可解释性不强。答复意见的可解释性评价标准如下：

1）答复意见文本长度大于等于 20，得分为 2，否则为 0；

2）出现 0 个关键词，得分为 0；

3）出现 1 个关键词，得分为 2 分；

4）出现 2 个关键词，得分为 4 分；

5）出现 3 或 4 个关键词，得分为 6 分；

6）出现 5 个及以上关键词，得分为 8 分。

4.4 基于答复时间

政府的答复时间和用户的留言时间之差，可衡量政府的答复在时间层面上的效率。计算各答复时间与对应的留言时间之差，存入文件中，设为集合 X，集合元素个数为 m,其中X_1，X_2，…X_m表示第一至第 m 条答复的时间与对应的留言时间之差。取出集合 X 的最小值 a 和最大值 b，其最大值和最小值之差为 D，将所有数据分为 10 组，组距为 d（取整数部分）。设定阈值为 100 天，若元素X_i值大于 100，则将该数据过滤,对应评分为 0。那么，第 n(n<10)个组就为 a+(n-1)d < X_n≤ a + nd，对应评分为 n。答复时间评价标准如下图：

表 8　答复时间评价标准

X_i	Score
X_i>100	0
$a \le X_i \le a+d$	1
$a+d < X_i \le a+2d$	2
$a+2d < X_i \le a+3d$	3
…	…
$a+(n-1)d < X_n \le a+nd$	n

4.5 评价模型分析

如图为答复内容基于完整性、相关性、可解释性、答复时间这四项评价指标各项得分部分数据表如图 15，全部数据见附件 result.xlsx。取各项得分之和的平均值作为总分 score，按总分划出 4 个等级，如表 9 所示：

ID	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复完整性	时间评判度	可解释性	相关性	总分	等级
0	2549	A00045581	A2区景蓉	2019/4/2	2019年4月	现将网友	2019/5/1	10	9	10	8	9.3	A
1	2554	A00023583	A3区潇楚	2019/4/2	潇楚南路	网友“A0	2019/5/9	10	9	6	7	7.8	B
2	2555	A00031618	请加快提	2019/4/2	地处省会	市民同志	2019/5/9	5	9	6	6	6	C
3	2557	A00011073	在A市买公	2019/4/2	尊敬的书	网友“A0	2019/5/9	6	9	6	3	5.4	C
4	2574	A0009233	关于A市公	2019/4/2	建议将“	网友“A0	2019/5/9	8	9	4	4	5.7	C
5	2759	A00077538	A3区含浦	########	欢迎领导	网友“A0	2019/5/9	10	7	0	4	4.9	C
6	2849	A00010080	A3区教师	2019/3/29	尊敬的胡	网友“A0	2019/5/9	10	6	8	8	8.4	A
7	3681	UU00812	反映A5区	2018/12/3	我做为一	网友“UU	2019/1/29	10	8	10	7	8.9	A
8	3683	UU008792	反映A市美	2018/12/3	我是美麓	网友“UU	2019/1/16	10	9	8	6	8.1	A
9	3684	UU008687	反映A市洋	2018/12/3	胡书记好	网友“UU	2019/1/16	9	9	6	8	7.8	B
10	3685	UU0082204	反映A2区	2018/12/3	我家住在	网友“UU	2019/3/11	10	3	6	5	6.6	B
11	3692	UU008829	A5区鄱阳	2018/12/2	胡书记：	网友“UU	2019/1/29	10	8	8	3	7.1	B
12	3700	UU00877	A4区万国	2018/12/2	尊敬的书	网友“UU	2019/1/14	10	9	6	7	7.8	B
13	3704	UU0081480	举报A市芒	2018/12/2	尊敬的领	网友“UU	2019/1/3	10	10	4	8	7.6	B
14	3713	UU0081227	建议增开	2018/12/3	建议增开	网友“UU	2019/1/14	10	9	6	4	6.9	B
15	3720	UU008444	关于A市新	2018/12/2	2016年下	网友“UU	2019/3/6	10	4	10	9	9.1	A

图 15 四项评价指标各项得分部分数据表

表 9　得分对应等级表

等级类别	score
等级 A	score ≥ 8.0
等级 B	6.0 ≤ score < 8.0
等级 C	3.0 ≤ score < 6.0
等级 D	0 < score ≤ 3.0

依据等级范围划分标准，统计每一个等级对应的留言回复总数并根据每一个等级中留言回复总数绘制饼图（如图 16）得到每一个等级中留言回复总数占所有留言回复的百分比。

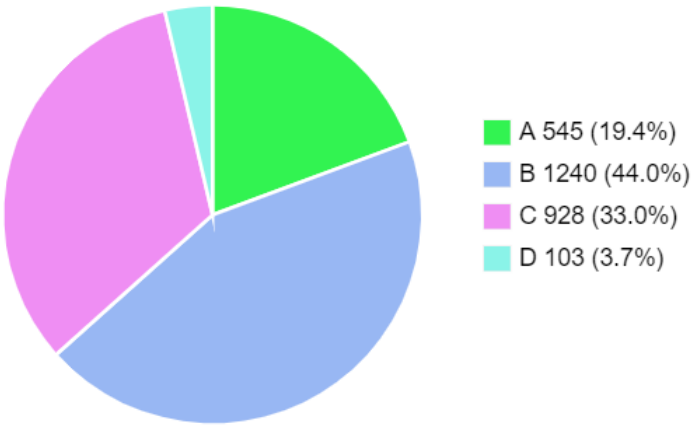

图 16　四类等级对应的留言总数所占比例

由图 16 中的数据看出，A 等级对应的留言总数占比为 19.4%，留言数为 545，B 等级对应的留言总数占比为 44.0%，留言数为 1240。二者的综合评价为优良，共占比 63.4%，已超过六成比例。等级 D 对应的留言总数占比仅为 3.7%,表明答复质量不高的数据仅占极少数。C 等级对应的留言总数占比为 33.0%，C 等级一般对应质量合格。从以上数据分析得出，政府做出的答复内容质量较高，可见政府做出了电子政务的创新举措，具有强大的活力，可预见，网络问政平台在新的时代背景下，有着越来越好的发展趋势。

5 结语

总结本次比赛，对于留言本文多类分类问题，本文先对留言主题和留言详情的内容文本做去空去重、结巴分词、停用词过滤等数据预处理操作，将词袋模型和 TF-IDF 方法配合使用，生成对应文本的权重矩阵，按 8.5:1.5 的比例将数据划分为训练集、测试集，通过线性支持向量机（LinearSVC）、多项式朴素贝叶斯、逻辑回归

等三种模型对其训练，通过十折交叉验证法，提高数据利用率，最后利用精准率、召回率、F1 值等评价指标，对分类模型结果做出评估，经三种模型的结果比较，其中支持向量机（LinearSVC）的分类效果最优。

对于热点问题的分析，通过 K-means 模型对留言进行聚类，但留言数据达到 4000 多条，计算复杂度高。因此，利用 textrank 算法提取每条留言文本中关键字排名前 5 的关键词，再生成权重矩阵，再通过 K-means 模型聚类，计算复杂度大大降低，聚类效果也有提升。

对于答复意见评价模型的设计，本文基于答复意见相关性、完整性、可解释性、答复时间等四个方面，建立答复意见评价模型。基于答复意见相关性，采取 Word2vec 方法生成词向量矩阵，再利用余弦相似度计算留言详情和答复意见的相似值，作为相关性分析的指标。基于答复意见完整性，答复意见的内容一般代表着政府的官方形象，因此，答复意见的内容一般具有某种官方固定的格式或规范。本文总结出，答复意见一般格式，包括问候语、回复引入语、正文、感谢语、答复时间等部分，其中正文部分放入可解释性里面分析，采取先识别时间，再提取关键词的方法，设定对答复意见一般格式的评价标准。基于答复意见的可解释性，利用 TextRank 方法提取其关键词，生成政府答复意见专业高频语料词表。基于答复时间，答复的时间越接近其对应留言的发布时间，则答复意见对留言问题的快速解决也更有作用。

6 参 考 文 献

[1].李超民，治理现代化视阈中的智慧政务建设．社会主义研究，2014(04)：第81-88页.
[2].蒋良孝，朴素贝叶斯分类器及其改进算法研究，2009，中国地质大学.
[3].雷金波,基于逻辑回归和支持向量机的设备状态退化评估与趋势预测研究,2008,上海交通大学.
[4].刘春雨，改进的支持向量机的理论研究及应用，2016，西北农林科技大学.
[5].杜圣东，基于多类支持向量机的文本分类研究，2007，重庆大学.
[6].王千等，K-means聚类算法研究综述．电子设计工程，2012．20(07)：第21-24页.
[7].唐明， 朱磊与邹显春， 基于 Word2Vec 的一种文档向量表示． 计算机科学，2016．43(06)：第214-217+269页.
[8].夏天， 词语位置加权 TextRank 的关键词抽取研究． 现代图书情报技术，2013(09)：第30-34页.