基于 Bi-LSTM 分类与 K-means 聚类的“智慧”政务系统

摘要

近年来，微信、微博、市长信箱、阳光热线等网络问政平台凭借传递速度快、空间距离小、成本低廉等优势逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道。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。本文基于数据挖掘技术对群众留言进行分类，挖掘群众留言中的热点问题，有利于相关部门及时了解民情，解决民生问题。

首先我们对问题给与的数据进行预处理，具体操作为去重，利用 Python 中的 jieba 分词库进行分词，然后去停用词，得到相关词组数据。

针对问题一，我们首先将预处理后的数据分为训练集和测试集，利用 Tokenzier 分词器对词组数据进行 One-Hot 编码；我们把通过编码得到的词向量放进构造的 BI-LSTM 当中，得到最终的词向量表示；最后根据查重率，查全率以及 F-Score 对分类结果进行评估。

针对问题二，我们使用 TF-IDF 算法将词组转化为词向量，接着利用 K-means 聚类算法，将留言反映的问题进行归类；为了方便政务人员统计处理，我们给出一套合理热度评价指标，计算各类别问题的热度进行排序，取热度排名前五的问题，并利用基于词频统计的特征词抽取的方法为热点问题进行命名。

针对问题三，我们分别从相关性，可解释性与完整性对留言答复做出评价。通过文本相似度判定相关性；通过循环神经网络评价完整性；然后用一个解释性框架评定其可解释性。最终采用层次分析法确定各因素占比得出评分。综合评定后，我们提出了如下建议：一是安排专人每日检查，督促相关部门及时受理办理网民留言，二是加强网站平台建设，完善管委会网站及微官网功能，提高网民咨询类诉求响应时效，提升网站服务水平。

关键词：中文分词；BI-LSTM 神经网络；TF-IDF 算法；K-means 聚类；热点问题筛选；层次分析法

"Smart" government system based on Bi-LSTM classification and K-means clustering

Abstract

In recent years,WeChat, weibo, mayor's mailbox, sunshine hotline and other network political platform with fast transmission speed, small space distance, low cost and other advantages gradually have became an important method for the government to understand public opinion, gather people's wisdom, and condense people's spirit.The establishment of intelligent government system based on natural language processing technology is a new trend of social governance innovation and development.

Based on the data mining technology, this paper classifies the comments of the masses and excavates the hot issues in the comments of the masses, which is conducive to the relevant departments to timely understand the situation of the people and solve the people's livelihood problems.

Firstly, we preprocessed the data given by the problem,the specific operation was deduplication. And we used jieba word segmentation in Python to segment the words,then got rid of the words to get the related phrase data.

For the first problem, we divided the pre-processed data into training set and test set, and used Tokenzier word segmentation to conduct one-hot coding for the phrase data. We put the coded word vector into the constructed BI-LSTM to acheive the final word vector representation. Finally, the classification results were evaluated according to the recall rate, recall rate and f-score.

For the second problem ,we use tf-idf algorithm to transform the phrase into a word vector, and then use k-means clustering algorithm to classify the problems reflected by the message. In order to facilitate the statistical treatment of government officials, we give a set of reasonable heat evaluation index, calculate the heat of various problems to sort, take the heat of the top five problems, and use the method of word frequency statistics based on the extraction of feature words to name hot issues.

For the third question, we evaluate the response from relevance, interpretability and completeness. The correlation is determined by text similarity. The integrity was evaluated by cyclic neural network. Then an interpretative framework is used to evaluate its interpretability. Finally, the analytic hierarchy process is used to determine the proportion of each factor to get the score. After comprehensive evaluation, we put forward the following Suggestions: one is to arrange daily inspection by special personnel to urge relevant departments to timely accept and handle netizens' comments; the other is to strengthen the construction of website platform, improve the functions of the website of the administrative committee and the official website of the micro-official website, improve the response time of netizens' demands for consultation, and improve the service level of the website.

Key words: Chinese word segmentation; BI-LSTM; TF-IDF; k-means; clustering hotspot selection; Analytic hierarchy process;

目录

一、简介

1.1 文本挖掘概述

文本挖掘是指从文本数据中获取有价值的信息和知识，它是数据挖掘中的一种方法。文本挖掘中最重要最基本的应用是实现文本的分类和聚类，前者是有监督的挖掘算法，后者是无监督的挖掘算法。

本文主要使用文本挖掘中的自然语言处理技术对问题所给的留言进行挖掘，找到我们想要的内容。

1.2 挖掘目标

本文挖掘的目标是使用 pandas 读取群众留言信息表内数据，通过 jieba 中文分词等工具对其进行预处理，构造 BI-LSTM 神经网络、使用 K-means 聚类等算法，达到以下目标：

（1）利用文本分词对数据进行处理，使其成为结构化数据，并通过构建分类器，对群众留言进行分类，提高相应智能部门工作效率。

（2）对群众留言进行聚类，并构建合理的热度评价指标，对留言主题进行热度计算及排名，筛选出热度前五的热点问题，以此帮助相关部门有针对性的处理问题。

（3）从留言答复意见的相关性、可解释性以及完整性出发，建立一套评价方案，对留言答复进行评分。

1.3 挖掘意义

现今，网络逐渐成为民众反映民主民生问题的主要渠道。本文通过建立模型挖掘留言当中的问题，对群众留言进行分类，以提高相关部门工作效率，建立合理的热度评价指标筛选出民众最迫切解决的问题，及时并有针对性的解决民生实际问题。对于加深民众对政府的信任度以及提升政府的管理水平和施政效率具有极大的推动作用。

二、总体流程及其步骤

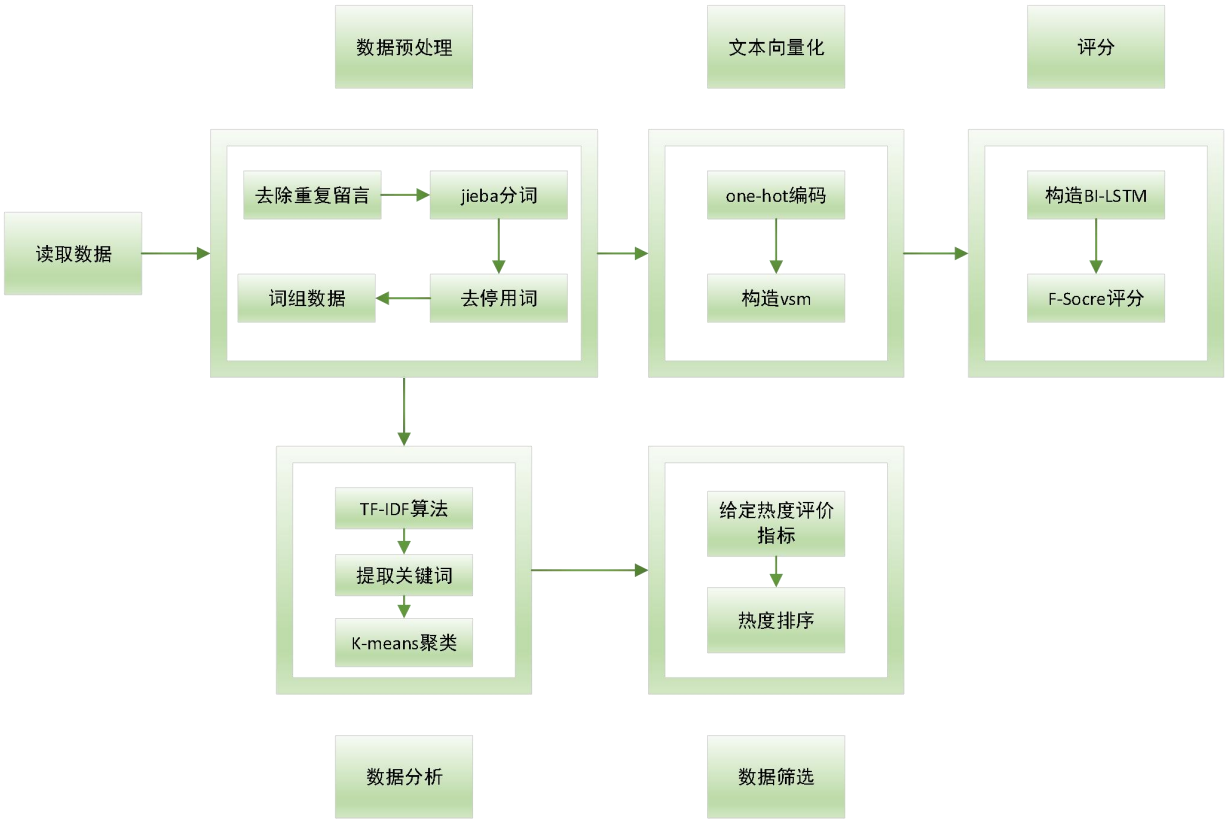


图 1　完整流程图

步骤一：数据预处理，对题中原始附件一二数据进行去重，然后借助 jieba 分词库进行中文分词，去停用词得到词组数据。

步骤二:对于处理后的附件一数据，我们采用 one-hot 编码构造 vsm，将文本数据转化为词向量；之后构造 BI-LSTM 神经网络建立文本分类器，最终对实验结果进行评估。

步骤三：对处理后得附件二数据，我们首先使用 TF-IDF 计算权重抽取关键词，然后使用 K-means 聚类方法对留言进行聚类；按照我们设定的热度评价指标，计算各类别问题的热度找出排名前五的问题。

步骤四：对处理后的附件四数据，我们分别从相关性，可解释性与完整性对留言答复做出评价。通过文本相似度判定相关性；通过循环神经网络评价完整性；然后用一个解释性框架评定其可解释性。最终采用层次分析法确定各因素占比得出评分。

三、数据分析与处理

3.1　数据分析

高质量的数据集是模型匹配和优化的基础，对整个数据集进行分析处理可以促进对数据的全面认识，从而更好的对数据进行编码，提高数据集的质量。

3.2 数据预处理

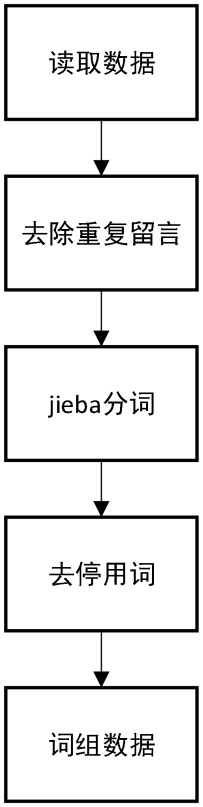


图 2　数据预处理流程图

3.2.1　jieba 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，从文本中提取词语时需要分词，本文采用 Python 开发的一个中文分词模块——Jieba 分词，对问题和回答中的每一句话进行分词。

jieba 分词：

（1）基于 *Trie* 树结构实现高效的词图扫描，找到字符串中所有可能的词条，构成一个有向无环图。

（2）每个词条对应一条有向边，利用统计的方法赋予对应的边长一个权值，找到起点到终点的最短路径（即基于词频的最大切分组合）

（3)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

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。

部分分词结果如图 2 所示：

```
0                                                [留言，详情]
1       [区，大道，西行，便，道，，，未管，所，路口，至，加油站，路段，，，...
2       [位于，书院，路，主干道，的，在水一方，大厦，一楼，至，四楼，人为，拆...
3       [尊敬，的，领导，：，A1，区苑，小区，位于，A1，区，火炬，路，，...
4       [A1，区，A2，区华庭，小区，高层，为，二次，供水，，，楼顶，水箱,...
                                ...
9206    [我们，夫妻，都，是，农村户口，，，大，的，是，女，9，岁，，，小...
```

图 3　部分分词效果

图 2 是没有停用词过滤的结果，会发现有大量无用的标点以及无意义的词，对后续挖掘有较大的影响，因此我们下面去停用词。

3.2.2　去停用词

停用词是指在信息检索中，为节省文本空间和提高搜索效率，在处理自然语言数据之前会自动过滤掉某些表达无意义的字或词，它们通常是一些单字，单字母以及高频的单词，比如中文中的“我、的、了、地、吗”等，英文中的“the、this、an、a、of”等。

对于停用词一般在预处理阶段就将其删除，避免对文本，特别是短文本，造成负面影响。这些停用词都是人工输入、非自动化生成的，生成的停用词会生成一个停用词表。

部分去停用词后的效果如图 3 所示：

	留言 详情
0	
1	大道 道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑 集团 燕子 山 ...
2	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水 电等 设施 烂尾 多年 ...
3	尊敬 领导 A1 区苑 小区 位于 A1 火炬 路 小区 物业 市程明 物业管理 有限公司 ...
4	A1 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味 ...
	...
9206	夫妻 农村户口 女 岁 岁 15 斤 治疗 两年 一级 脑瘫 女户 招郎 男方 两 兄弟 大...

图 4　部分去停用词后的效果

四、分析方法与过程

4.1 问题一分析方法与过程

4.1.1　流程图

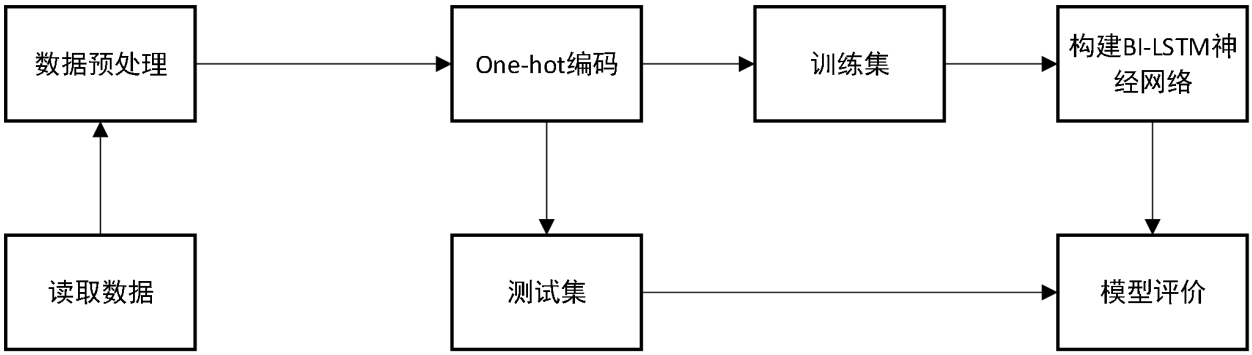


图 5　文本分类流程图

4.1.2　one-hot 编码

为了将语料输入神经网络进行训练，我们首先要将自然语言符号表示成计算机能够理解的数字形式。

一个自然的想法是把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为 ，只有一个维度的值为 1，这个维度就代表了当前的词。这就是独热编码形式（One-Hot）

在回归，分类，聚类等机器学习算法中，特征之间距离的计算或相似度的计算是非常重要的，而我们常用的距离或相似度的计算都是在欧式空间的相似度计算，计算余弦相似性，基于的就是欧式空间。而我们使用 One-Hot 编码，将离散特征的取值扩展到了欧式空间，离散特征的某个取值就对应欧式空间的某个点。将离散型特征使用 One-Hot 编码，会让特征之间的距离计算更加合理。

4.1.3 BI-LSTM 神经网络

为了尽可能的保证词组之间的语义联系，我们将词向量放入循环神经网。同时为了获得更多的上下文记忆信息，我们最终选择 BI-LSTM 网络。

4.1.3.1 RNN 和 LSTM

循环神经网络（Recurrent Neural Network,RNN）近年来由于其良好的性能代替深度神经网络（Deep Neural Network,DNN）成为主流自然语言处理建模方案，相对于 DNN，RNN 在隐层上增加了一个反馈，即 RNN 隐层的输入有一部分是前一级的隐层输出，这使 RNN 能够通过循环反馈看到当前时刻之前的信息，赋予了 RNN 记忆功能，能较好的表征上下文的语义。这些特点使得 RNN 非常适合用于对自然语言进行建模。如图 6 所示，所有的 RNN 都具有一种重复神经网络模块的链式形式, 在标准 RNN 中，这个重复的模块只有一个非常简单的结构，例如一个 tanh 层。

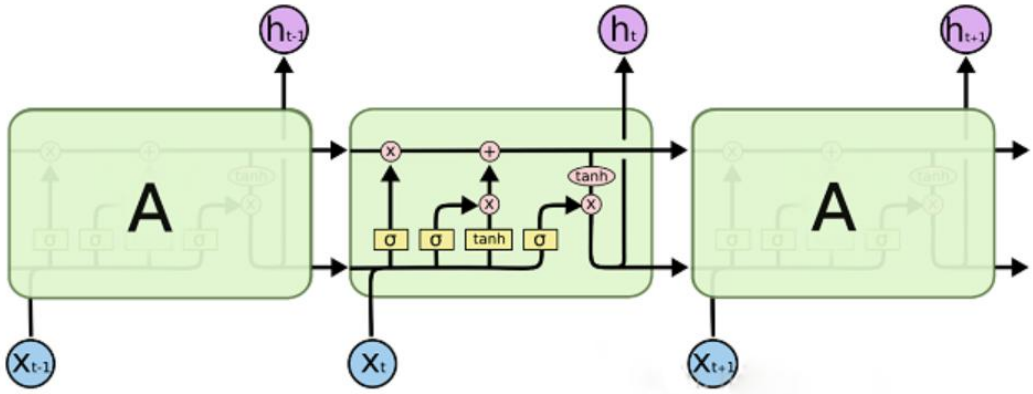


图 6　RNN 的结构

长短期记忆网络（Long Short-Term Memory，LSTM），是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。与标准 RNN 相比，LSTM 在算法中加入了一个判断信息有用与否的“处理器”，这个处理器作用的结构被称为细胞。一个细胞当中被放置了三扇门，分别叫做输入门、遗忘门和输出门如图 7 所示，这些精心设计的“门”结构实现了 LSTM 遗忘或增

加信息能力。一个信息进入 LSTM 的网络当中，可以根据规则来判断是否有用。只有符合算法认证的信息才会留下，不符的信息则通过遗忘门被遗忘。

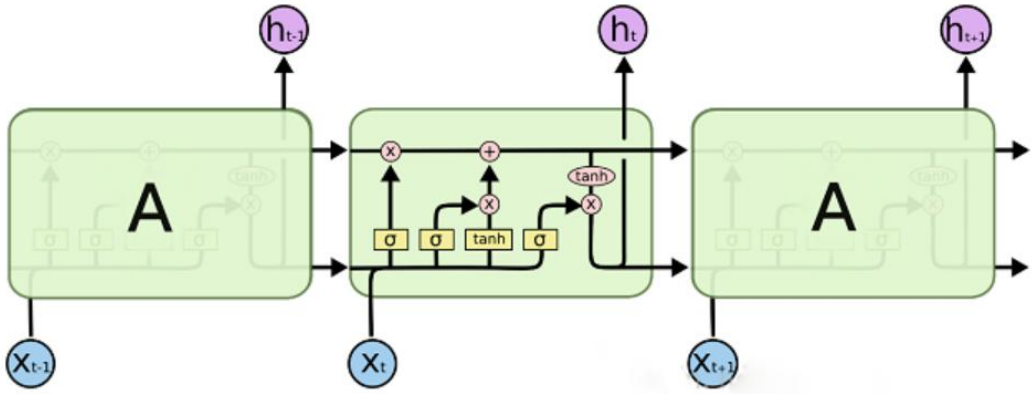


图 7　LSTM 细胞结构

某一步时间t的输入门（i_t）和遗忘门（f_t）都以输入变量（即我们的备选答案编码x_t），上一步时间步$t-1$　的输入向量（h_{t-1}）和偏置（b）作为输入，并通过激活函数得到响应值。

忘记门层：忘记门层决定了 LSTM 何时会从系统状态中丢弃信息。其公式为：

$$f_t = \delta(W_i \cdot [h_{t-1}，\ x_t] + b_f)$$

当我们在输入中看到新的主语时，可以通过使用忘记门层忘记旧的主语，提高语义准确度。

输入门层：输入门层确定了 LSTM 将要把什么样的信息保存在新的细胞状态中，其计算公式为：

$$i_t = \sigma(W_i \cdot [h_{t-1}，\ x_t] + b_i)$$

$$\bar{c}_t = \tanh(W_c \cdot [h_{t-1}，\ x_t] + b_c)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \bar{c}_t$$

输出门层：最终 LSTM 使用输出门层确定要输出的值，其计算公式为：

$$o_t = \sigma(W_o[h_{t-1}，\ x_t] + b_o)$$

$$h_t = o_t \cdot \tanh(c_t)$$

4.1.3.2 BI-LSTM

由于自然语言的语义复杂性，语义的重要信息可能出现在句首也可能出现在句尾。普通的 LSTM 模型可以解决 RNN 的遗忘问题，但是它只能做到前文理解后文，这样在某种情况下，LSTM 仍然可能会丢失部分句首或者句尾的重要含义。所以最终我们采用双向长短时记忆模型（Bi-LSTM）。

Bi-LSTM 可以将原来按序的输入转化成一正一逆的两股输入，通过两个 LSTM 单元组成一个新的双向 LSTM 单元，以尽量弥补纯问答对匹配方式对上下文信息考虑的不足，一定程度上解决 RNN 与单向 LSTM 的遗忘问题。

经过检验，BI-LSTM 模型较前述的 RNN 与 LSTM 模型的效果有明显的提升，它可以更完整的对句子的深层语义进行编码，中间编码向量既包含了句首的语义信息，也保留了句尾的语义信息，可以更好的表达句子的语义。结构如图 7 所示：

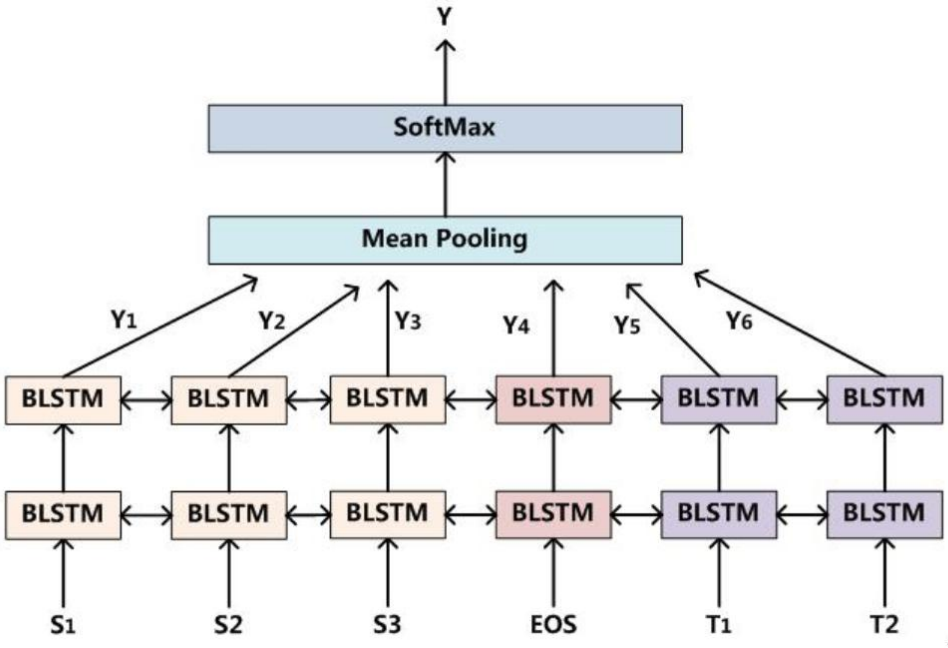


图 8　Bi-LSTM 结构图

4.1.4 分类评估

对于本文的模型，我们采用查重流程（精度）、查全率（召回率）、两者的调和平均数 F1-Score、准确率来评价我们模型的效果。

以上指标的详细定义如下：为了放便后面符号的说明定义一个混淆矩阵，如下表所示：

	相关	不相关
被检测到的	*TP*	*FP*
未被检测到的	*FN*	*TN*

表 1 混淆矩阵

- *TP*(True Positive): 正类项目被判定为正类
- *FP*(False Positive): 负类项目被判定为负类
- *FN*(False Negative): 正类项目被判断为负类
- *TN*(True Negative): 正类项目被判断为负类

(1) 查重率
查重率（精度)是衡量某一检索系统的信号噪声比的一种指标，即检出的相关留言量与检出的留言总量的百分比。

$$P = \frac{TP}{(TP + FP)}$$

（2）查全率
查全率是检出的相关留言量与检索系统中相关留言总量的百分比。衡量的是检索系统的召回率。

$$R = \frac{TP}{(TP + FN)}$$

（3）准确率
准确率是指正确检索出的有关无关留言数，占检索系统当中所有留言数的百分比。

$$A = \frac{TP + TN}{TP + FP + FN + TN}$$

（4）F-Score

分类的 F1 值就是查重率与查全率的调和平均值

$$F_1 = \frac{2PR}{P + R}$$

4.2 问题二分析方法与过程

4.2.1 流程图

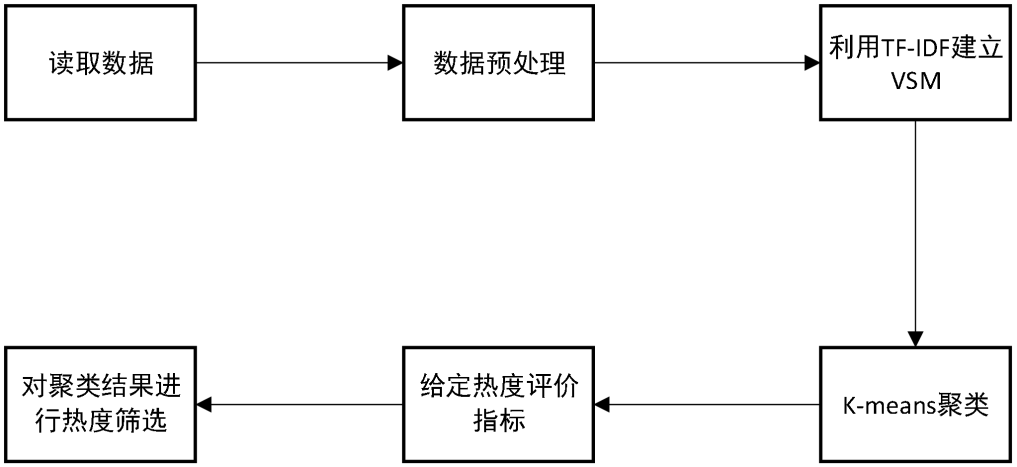


图 9　文本聚类流程图

4.2.2　TF-IDF 算法抽取关键词

在对留言进行数据预处理之后，需要将其转化为词向量。这里采用 TF-IDF 算法，将留言信息转化为权重向量。具体原理如下：

（1）计算词频（TF），即权重（Team Frequency)。

词频是指某个词在一个文本中出现的次数，与文本的主题相关。但要注意在特定的语言环境下都有许多特定的词不具备这种词性而应该被排除，如中文的“的”“地”、英文的“a”“an”等。

计算公式：

$$tf_{i,\ j} = \frac{n_{i,\ j}}{\sum_{k} n_{k,\ j}}$$

注：分子是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和。

（2）计算逆词频（IDF），即逆文档频率（Inverse Document Frequency)

IDF 越大，即某个词项在文本集合的多条文本中出现次数越多，该词项的区分能力越差。如果一个包含 1000 条文本的集合中，如果某个词项 A 在 100 条中都出现，而另一个词项 B 在 10 条中出现，则词项 B 比 A 具有更好的区分能力。

计算公式：

$$idf_i = \log(N/df(w_i))$$

其中 N 示文本集合中所有文本的总数，$df(w_i)$ 表示文本集合中有多少条文本出现了词项 w_i

（3）计算 TF-IDF 值，通常采用如下公式计算：

$$TF-IDF(w_i) = tf_j(w_i) \times \log(N/df(w_i))$$

实际得到 TF-IDF 的值越大，这个词在文本当中越重要。计算每个词的 TF-IDF 值，按照其大小进行排序，权重值越大的证明其为该文本的关键词。

4.2.3 文本聚类

4.2.3.1 文本相似度计算

文本之间是否相似的一个衡量标准是文本相似度。在聚类的过程当中，需要研究文本个体间的差异，即需要对文本相似度进行计算，根据得出的结果进行聚类。目前相似度计算的方法有基于字面匹配，计算向量之间的距离等。本文采用的是计算向量间距离的欧几里得距离计算方法。

令 $i = (x_1, x_2, \cdots, x_n)$ 和 $j = (y_1, y_2, \cdots, y_k)$ 是两个被 k 个数值属性标记的对象，则对象 i、j 之间的欧式距离可以定义为：

$$dis(i,j) = \sqrt{\sum_{p=1}^{k}(x_p - y_p)^2}$$

4.2.3.2 K-means 聚类

所谓文本聚类就是将无类别标记的文本信息根据不同的特征，使用相似度计算将具有相同属性或者相似属性的文本聚合在一起。本文中通过聚类的方法将具有相同表述含义的留言聚为一类，方便政务人员处理，做出统一答复。

K-means 算法属于无监督学习的一种聚类算法，其目的为在不知数据所属类别及类别数量的前提下，依据数据自身所暗含的特点对数据进行聚类，其实现原理如下：

假设有一个包含n个d维数据点的数据集$X=\{x_1、x_2、\cdots、x_n\}$，其中$x_i\in R^d$，K-means 聚类将数据集X组织为K个划分$C=\{c_i，i=1，2，\cdots，k\}$每个划分代表一类，每个类中有一个类别中心μ_i。选取欧氏距离作为相似性和聚类判断准则,计算该类中各点到聚类中心μ_i的距离平方和：

$$J(c_k)=\sum_{x_i\in c_k}\|x_i-\mu_i\|^2$$

聚类目标则是使得各类总的距离平方和最小，该平方距离和被称为偏差$J(C)$即：

$$J(C)=\sum_{k=1}^{K}J(c_i)=\sum_{k=1}^{K}\sum_{x_i\in c_k}\|x_i-\mu_i\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}\|x_i-\mu_i\|^2$$

其中：$d_{ki}=\begin{cases}1，& x_i\in c_i\\0，& x_i\notin c_i\end{cases}$，所以根据最小二乘法和拉格朗日原理，聚类中心$\mu_i$应该选取该类别中各数据点的平均值。

假设把数据聚成 K 个类别，算法具体描述如下：

（1）初始化：随机指定k个聚类中心；

（2）分配x_i：对每一个样本x_i求其到k个聚类中心的距离，找到离他最近的聚类中心μ_i，并将其分配到μ_i所标明类；

（3）修正μ_i：将每一个聚类中心移动到其标明类的中心；

（4）计算偏差：求出最小偏差值；

（5）判断偏差收敛性：若偏差收敛，则返回$(c_1，c_2，\cdots，c_k)$，并终止算法，否则返回步骤二；

流程图如图 9 所示：

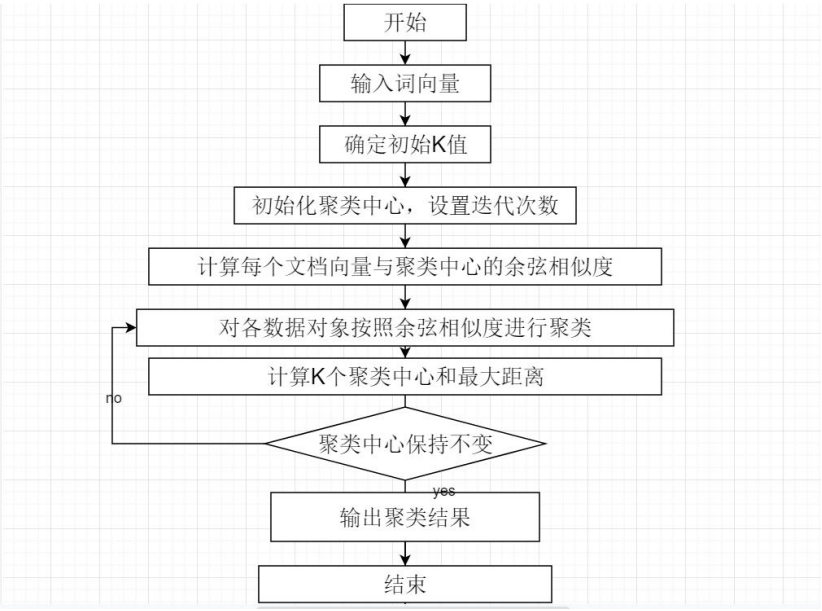


图 10 K-means 流程图

该算法要求在计算之前给出 K 值，经过多次实验比对结果后，我们设置本文的 K 值为 1000，由此得到聚类结果。

4.2.4 文本筛选

为了提高政务人员的工作效率，更清晰的了解到比较受群众关注的政务问题。我们给出一套合理的热度评价指标，根据这一指标，对问题二当中聚类的结果进行热度计算，提取热度排名前五的问题留言。

4.2.4.1 热度评价标准

本文基于留言特点，提出留言主题的热度评价因素——主题影响力，主题影响力体现在该主题内留言的数量以及单条留言内容的影响力总和。而对于单条留言的影响力可以分为直接影响力和间接影响力。

直接影响力——由于政务留言公开，所以直接影响力与群众对该留言的点赞数和反对数有关。

间接影响力——即留言反应的问题重要程度。

定义热度评价公式如下：

$$Inf(T)=\alpha\times\sum_{i=1}^{n}Inf(t_i)+\beta\times n,\quad i=1,2,\cdots,n$$

其中α、β为热度系数，满足$\alpha+\beta=1$，$\alpha>0$，$\beta>0$，$Inf(T)$为某主题的的影响T

力，n为该主题中留言条数，$Inf(t_i)$为单条相关留言的影响力。

现在考虑单条相关留言的影响力：

$$Inf(t) = Inf_D(t) + Inf_I(t)$$

其中$Inf(t)$为单条相关留言t的影响力，$Inf_D(t)$是单条留言的直接影响力，$Inf_I(t)$是单条留言的间接影响力，有：

$$Inf_D(t) = |supports| + |opposes|$$

其中α、β为系数，$|supports|$、$|opposes|$分别为点赞数和反对数

$$Inf_I(t) = a$$

其中a为留言重要程度，由具体留言内容所决定

4.3 问题三的分析与过程

4.3.1 流程图

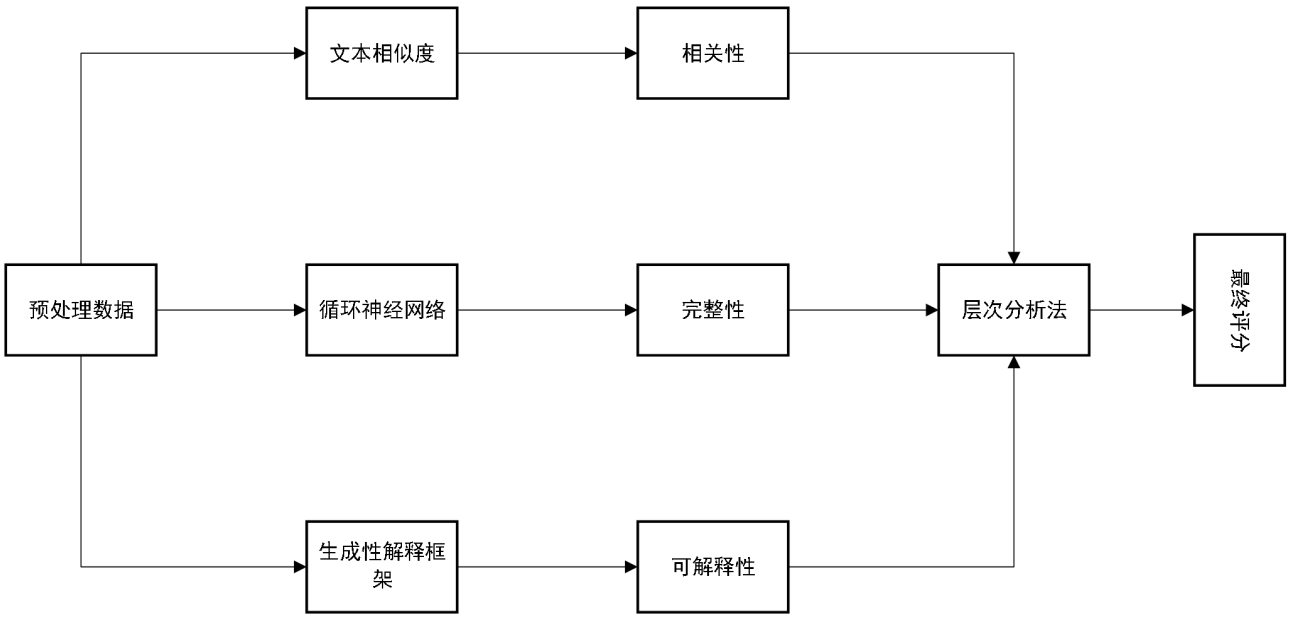


图 11

4.3.2 文本相似度

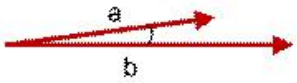
4.3.2.1 相似度度量（Similarity）

计算个体间的相似程度，相似度度量的值小，说明个体间相似度越小，相似度的值越大说明个体差异越大。

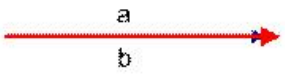
4.3.2.2 向量空间余弦相似度（Cosine Similarity）

对于多个不同文本或是短文本对话消息，要计算他们的相似度，我们通常将词组映射到向量空间，形成文本中文字和向量数据的映射关系。通过向量差异的大小来计算文本的相似度，一个常用方法，计算向量空间余弦相似度：

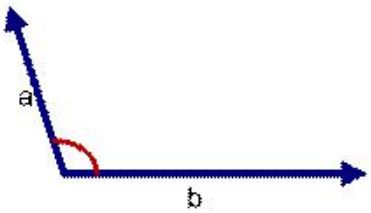
余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫"余弦相似性"。



上图两个向量 a,b 的夹角很小可以说 a 向量和 b 向量有很高的相似性。



极端情况下 a 和 b 可以重合，此时可以认为 a 和 b 向量是相等的，也即 a，b 向量代表的文本是完全相似的。



若如上图 a 和 b 向量夹角较大，或者反方向，那么说说 a 向量和 b 向量有很差的的相似性，或者说 a 和 b 向量代表的文本基本不相似。

4.3.3 生成性解释框架

4.3.3.1 基础性分类器和生成器

一个理想的模型应该同时提供预测结果及其解释，一个简单的生成解释方法即将文本向量输入到一个解释生成器中，以生成细粒度解释；

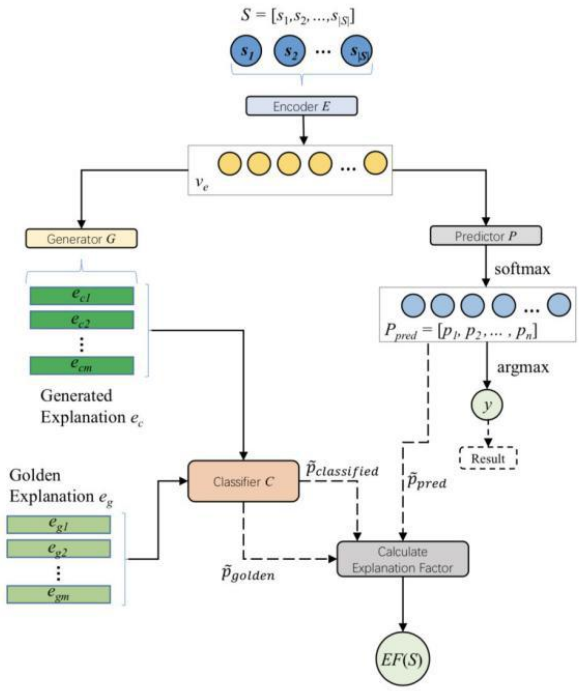


图 12　解释框架图

其公式如下：

$$v_e = Encoder([s_1, s_2, \cdots, s_{|S|}])$$

$$P_{pred} = Predictor(v_e)$$

$$y = arg\max(P_{pred,\ i})$$

$$e_C = f_G(W_G \cdot v_e + b_G)$$

其中，模型使用$soft\max$函数将分布概率转换为分类；

训练过程中，整体损失函数L包括两部分：分类损失Lp和生成损失Le

$$\mathcal{L}(e_g,\ S,\ \theta) = \mathcal{L}_p + \mathcal{L}_e$$

其中，θ代表所有参数

4.3.3.2 解释因子

上文提到的方法存在一个明显的缺陷：该方法无法在生成的解释和预测之间

建立合理解释，即解释和预测结果相互独立。为了生成更加合理的结果解释，此处使用一个解释因子来建立结果和解释间的联系。

一些情况下，细粒度信息比输入的原始文本序列更能反映整体结果。因此，本文预训练了一个分类器，该分类器直接将解释作为输入，并学习预测分类。

通过使用这一预训练分类器为文本编码器提供指导，使其能够生成一个含有更多信息量的向量。在训练过程中，首先使用解释生成器得到生成性解释，之后将向量输入到分类器中，得到预测结果的概率分布。

同时，我们将人们可接受的文本解释输入到分类器中，得到可接受的解释（golden explanation）的概率分布，公式如下：

$$P_{classsified} = soft\max(f_C(W_C \cdot e_c + b_C))$$

$$P_{gold} = soft\max(f_C(W_c \cdot e_g + b_C))$$

为衡量预测结果、生成的解释和可接受解释的生成结果之间距离，此处从 Pclassified ,Ppred ,Pgold 中分别抽取了真值，用于衡量最小风险训练预测结果和真值结果差异。

综上所述，我们得出解释因子定义如下：

$$EF(S) = \left|\widetilde{P}_{classified} - \widetilde{P}_{gold}\right| + \left|\widetilde{P}_{classified} - \widetilde{P}_{pred}\right|$$

4.3.4 循环神经网络

本文我们利用基于循环神经网络的语义完整性分析方法，通过判断句子是否语义完整，将长文本切分为多个语义完整句。

我们的模型基于双层的 Bi-LSTM 循环网络，首先，模型的输入为原始文本预处理后的次序列，将其映射为词向量，经循环窗口和欠采样处理后作为输入层，并通过分类器输出相应标签概率。

关于 Bi-LSTM 神经网络我们已经在 4.1.3.2 节中说明，此处不再重复。

4.3.5 层次分析法

4.3.5.1 内容

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。

层次分析法是将决策问题按总目标、各层子目标、评价准则直至具体的备投方案的顺序分解为不同的层次结构，然后用求解判断矩阵特征向量的办法，求得每一层次的各元素对上一层次某元素的优先权重，最后再加权和的方法递阶归并各备择方案对总目标的最终权重，此最终权重最大者即为最优方案。

4.3.5.2 原理

层次分析法根据问题的性质和要达到的总目标，将问题分解为不同的组成因素，并按照因素间的相互关联影响以及隶属关系将因素按不同层次聚集组合，形成一个多层次的分析结构模型，从而最终使问题归结为最低层(供决策的方案、措施等)相对于最高层(总目标)的相对重要权值的确定或相对优劣次序的排定。

五、相关结果

5.1 问题一结果

5.1.1 数据分析探索

读取预处理后的数据，利用 sklearn 库中的 split 方法将数据划分为训练集和测试集。

首先对数据集标签进行一个简单的探索，发现标签大致分布情况如下:

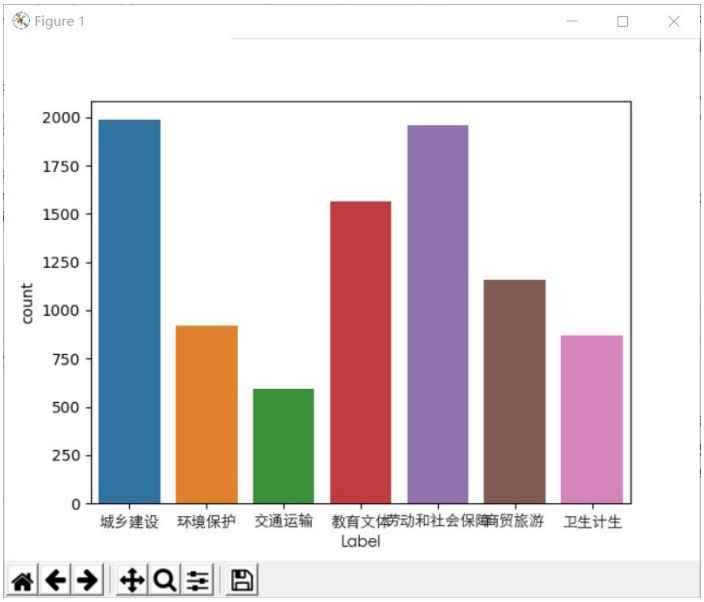


图 13 标签分布图

5.1.2 One-Hot 编码

采用传统的 One-hot 编码算法对语料进行词嵌入处理。（部分处理图 10 如下）：

```
('年', 1)
('县', 2)
('领导', 3)
('说', 4)
('公司', 5)
('学校', 6)
('工作', 7)
('元', 8)
('政府', 9)
('部门', 10)
====================
--------------------
('大道', 178)
('道', 325)
('未管', 1)
('路口', 110)
('加油站', 47)
('路段', 109)
('人行道', 111)
('包括', 370)
('路灯', 116)
('杆', 3)
```

图 14 部分处理图

5.1.3 BI-LSTM 模型建立以及训练

```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inputs (InputLayer)          (None, 600)               0
_________________________________________________________________
embedding_1 (Embedding)      (None, 600, 128)          640128
_________________________________________________________________
lstm_1 (LSTM)                (None, 128)               131584
_________________________________________________________________
FC1 (Dense)                  (None, 128)               16512
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
FC2 (Dense)                  (None, 7)                 903
=================================================================
Total params: 789,127
Trainable params: 789,127
Non-trainable params: 0
```

图 15 模型概况图

对模型开始进行训练：

```
6784/7241 [===========================>..] - ETA: 17s - loss: 0.1196 - accuracy: 0.9689
6912/7241 [===========================>..] - ETA: 12s - loss: 0.1208 - accuracy: 0.9689
7040/7241 [===========================>.] - ETA: 7s - loss: 0.1202 - accuracy: 0.9690
7168/7241 [===========================>.] - ETA: 2s - loss: 0.1190 - accuracy: 0.9694
7241/7241 [==============================] - 272s 38ms/step - loss: 0.1195 - accuracy: 0.9693
Epoch 10/10

 128/7241 [.............................] - ETA: 4:22 - loss: 0.0988 - accuracy: 0.9844
 256/7241 [>............................] - ETA: 4:26 - loss: 0.0790 - accuracy: 0.9805
 384/7241 [>............................] - ETA: 4:31 - loss: 0.0855 - accuracy: 0.9792
```

图 16 训练中

选择对训练集进行 10 次训练，并得出 10 次准确度：

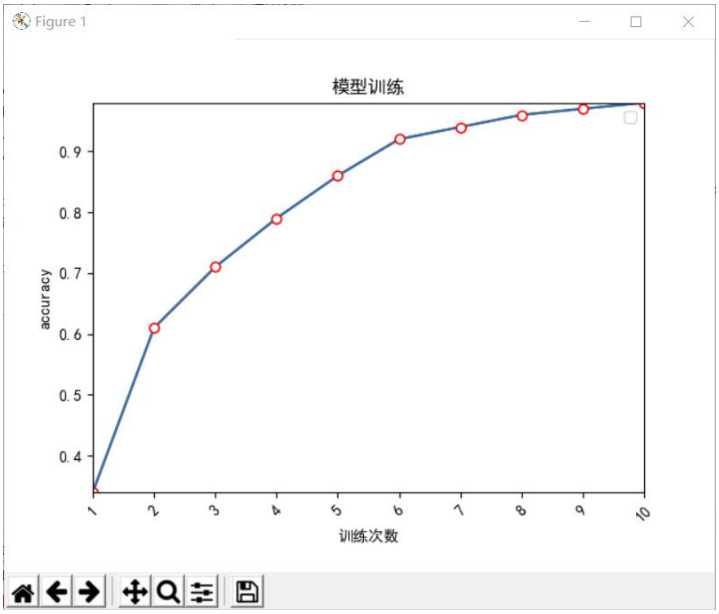


图 17

得出最终平均准确度为 0.84。

对于问题一，我们同样的实验了朴素贝叶斯算法，但是经过后期评估，发现朴素贝叶斯的 F1-socre 评分相对于 BI-LSTM 模型而言较低，所以我们采用

BI-LSTM 模型作为本次问题一的最终实验模型。

5.1.4.实验评估

5.1.4.1 混淆矩阵如图 18 所示

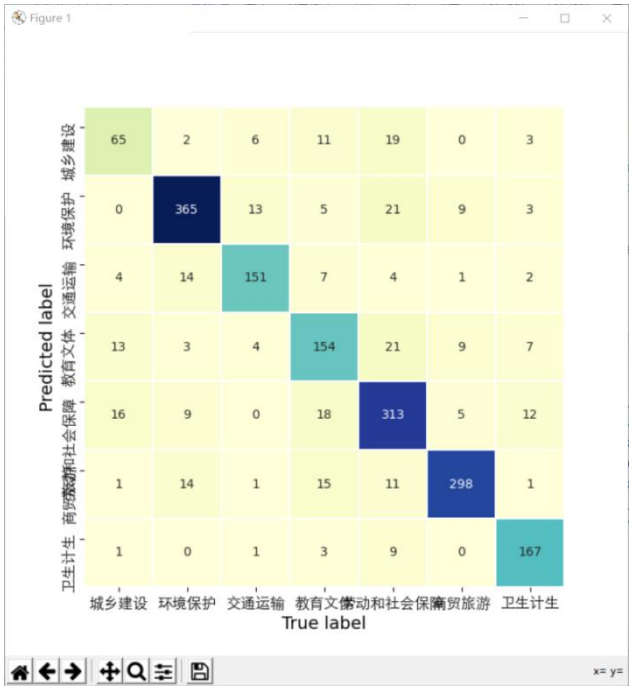


图 18 混淆矩阵

5.1.4.2 F-score 结果

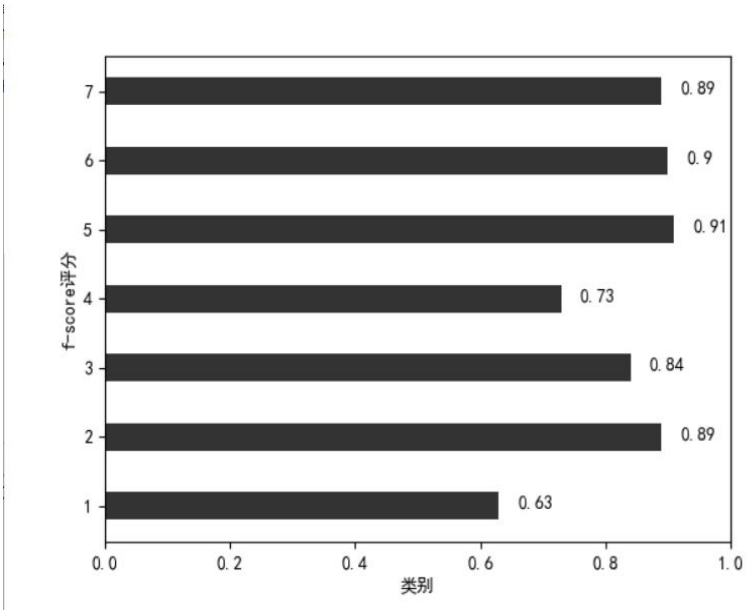


图 19 分类评价结果

通过上述评分结果，我们看到不同类别的准确度存在一定的差异，最高的达到 0.92，最低的为 0.61，之间的差距还是很大。

5.2 问题二结果分析

5.2.1 聚类结果

问题ID	留言编号	留言用户	留言主题
1	191579	A00017743	A市新奥燃气无法通过网络付费？
1	190087	A00052076	A市新奥燃气服务态度差
1	192869	A00051379	A市新奥燃气强行给我家换气表后，气的用量翻倍
1	205365	A00019782	A市营盘路279号居民楼后面一年四季粪水横流
1	218637	A0004691	A市润华燃气迟迟不给A6区新城国际花都四期业主开通燃气
1	231097	A00097918	A市宇冰燃气公司故意不退钱
1	250368	A00091108	A市新城国际都花物业公司限制业主买水
1	285322	A000103797	A市奥新燃气多收客户费用
2	252153	A00022081	A市跳马镇K10县村北冲组领导私自将集体土地第三方单位做垟
2	255276	A909219	再次希望领导“拯救”丽发新城小区居民
2	268449	A000100595	A2区通用时代小区居民希望尽快完成水改
2	281348	A909219	希望领导“拯救”丽发新城小区居民
3	283674	A00052059	希望A7县政府探索更多更公平的房屋拆迁补偿政策
3	255615	A00037136	反映A7县榔梨镇地铁6号线拆迁补偿问题
3	246915	A00037136	咨询A7县榔梨镇拆迁问题
3	197085	A00035481	A市耕地占用补偿标准是什么
3	218125	A00049650	咨询A7县北横线拆迁标准
3	228597	A00088587	咨询A2区修地铁拆迁补偿标准问题
3	241945	A00036943	A7县黄花镇黄花村征收拆迁，独生子女没有按标准补偿
4	208628	A00087687	呼吁A市撤销禁摩令
4	225138	A000107700	A7县新长海车库漏水什么时候可以停止！
4	266131	A000106805	呼吁停止在A7县泉塘街道小塘路设置110kv变电站

图 20 部分聚类结果

5.2.2 热度排名前五问题

热度排名	热度指数	时间范围	地点/人群	问题描述
1	2830	2019/1/7 - 2019/7/8	西地省 A4 区	58 车贷案件
2	2610	2019/1/15 - 2019/11/11	A 市五矿万境 k9 县	房屋建设存在安全隐患
3	1800	2019/2/15 - 2019/8/18	A 市学龄儿童	学龄儿童入学困难
4	1500	2019/4/10 - 2019/12/21	A2 区丽发新城	物业多方面存在问题
5	830	2019/8/23 - 2019/9/6	A4 区绿地海外滩	长赣高铁规划问题

表 2　热点问题表

完整热点问题表以及热点问题明细表见附件。

5.3 问题三结果分析

5.3.1 评分结果

通过利用 python 对该问题进行实现，最终我们得到了如下评分结果：

留言时间	留言详情	答复意见	答复时间	评分情况
2019/4/25 9:32:09	却以交20万保证	理费，在业主大会	2019/5/10 14:	0.57
2019/4/24 16:03:40	主意带来很大影响	且换填后还有三	2019/5/9 9:49	0.52
2019/4/24 15:40:04	是加大了教师的	教职工要依法签订	2019/5/9 9:49	0.53
2019/4/24 15:07:30	市，想买套公寓，	下（含），首次购	2019/5/9 9:49	0.42
2019/4/23 17:03:19	岭小学”，原“”的问题。公交站		2019/5/9 9:51	0.62
2019/4/8 8:37	已冲到右边，越	有说明卫生较差的	2019/5/9 10:0	0.38
2019/3/29 11:53:23	社区惠民装电梯的	政府办公室下发了	2019/5/9 10:1	0.55
2018/12/31 22:21:59	天寒地冻的跑	设施设备采购等	2019/1/29 10:	0.51
2018/12/31 9:55:00	到相关准确开工	检查后，西地省楚	2019/1/16 15:	0.42
2018/12/31 9:45:59	等地方做立体绿	要求完成了建设	2019/1/16 15:	0.54
2018/12/30 22:30:30	审批通过《温	地征收补偿款给原	2019/3/11 16:	0.43

图 21　部分评分结果

通过评分情况以及结果分析我们可以看到，相关部门的答复较为满意，从完整性、可解释性、相关性出发，群众问题普遍得到了解决，但还是存在部分问题，例如答复时间较久，对于紧急事件不能及时处理。

5.3.2 建议

通过对问题三的结果进行分析，我们提出了一定的建议：

（1）加强网站日常检查。安排专人每日检查，督促相关部门及时受理办理网民留言，加强留言办理质量监督检查；

（2）二是加强网站平台建设。完善管委会网站及微官网功能，部署智能问答系统，提高网民咨询类诉求响应时效，提升网站服务水平；

（3）三是加强市民热线建设。加大市民热线宣传推广，拓展市民诉求反映渠道，健全热线运行管理制度和联动响应机制，提高热线服务质量，多渠道解决市民诉求。

六、模型改进

6.1 问题一

由问题一的结果我们观察到各个类别的准确度存在较大的差异，我们思考是否是因为 One-hot 编码虽然方便易懂，但编码的维数由词典长度而定，过于稀疏，存在降维难问题，给计算造成了很大不便。

由此我们想能否采用其他的算法，例如 Skip_gram 算法，CBOW 算法，与本文的 One-hot 编码算法相比各类别之间的准确度差距是否会缩小。

6.2 问题二

问题二中的聚类效果不是非常理想，一方面可能是 K 值取值不当，另一方面可能存在的问题是数据集当中存在噪声维度。当噪声维度存在时 K-means 在聚类过程中仍旧将其看成是正常的特征维度进行利用，而不能加以区分。

因此我们了解到 WKmeans 聚类算法，这种方法在聚类时会给每个特征维度赋予一个权重，使得噪音维度的权重会尽可能的趋于 0，以此来去除对聚类结果的影响。而 WK-means 算法在新闻类数据集聚类中仍旧存在一个问题，即不能在不同簇中区分不同的有效维度。故而又出现了 EWK-means 聚类算法。依据上述说法，问题二可以朝此方向进行改进。

七、结论

微信、微博、市长信箱、阳光热线等网络问政平台凭借传递速度快、空间距离小、成本低廉等优势逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道。本文主要是利用题目所给的群众留言数据，对附件一二的数据进行基本的去重，中文分词以及去停用词处理后，将其转化为词向量；不同问题采用不同的处理方法，问题一我们通过 one-hot 编码，然后构造 BI-LSTM 神经网络建立文本分类器，对附件一数据进行分类；问题二采用基于权重的 TF-IDF 算法向量化附件二数据，之后采用 K-menas 聚类方法对留言进行聚类，同时给出一套合理的热度评价指标，筛选出留言当中排名前五的热点问题。

通过使用 python，在问题一中，我们成功对留言进行了分类，并通过实验评估验证其准确度；在问题二中，我们通过算法成功挖掘出了热度排名前五的主题，更好的帮助了行政人员了解群众最需要迫切解决的问题。

但是本次得到的分类结果，各个类别之间的准确率相差较大。而聚类效果也不是非常理想，可能是 K 的取值不够好，这也涉及到模型的一些改进问题。我们后期也会对文本挖掘做进一步的探讨和学习。

八、参考文献

[1] 何跃,帅马恋,冯韵.中文微博热点话题挖掘[J].统计与信息论坛.2014

[2] 马彦.大数据环境下微博舆情热点话题挖掘方法研究[J].现代情报.2014

[3] 王千,王成,冯振元,叶金凤.K-means 聚类算法研究综述[J].电子设计工程.2012

[4] 黄承慧,印鉴,侯昉.一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J].2011

[5] 於雯,周武能.基于 LSTM 的商品评论情感分析[J].计算机系统应用.2018

[6] 傅玳.多指标综合评价方法综述[J].知识丛林.2004

[7] 刘京卖野，刘新等.基于循环神经网络的语义完整性分析[J].计算机系统应用.2019

[8] SUN Ji-Gui，LIU Jie，ZHAO Lian-Yu,Clustering Algorithms Research[J].Journal of software.2008