“智慧政务”中的文本挖掘应用

摘 要

近年来，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。本文将基于数据挖掘技术对政务留言信息数据进行内在的信息挖掘，提取我们需要进行分析的部分进行深度挖掘和分析，探寻政务热点问题，对政务系统的答复完善性进行评估。

针对问题一：

1、特征选择　使用中文停用词库对附件 2 的原始数据进行简化，数据的简化包括 jieba 词库去重，用中文停用词库进行相应的删除。而后使用词频（TF）对简化后的文本进行特征选取

2、权重赋值　为优化文本空间，一是将一级分类各类别依次转换成自然数 。二是采用 Tf-Idf 算法，分别将一级分类和简化后的留言详情进行向量化后，对比高频词与低频词的权重与一级分类权重的拟合度

3、模型选择　选择 LSTM 模型，对留言详情的特征向量的测试集文本以及训练集文本采用 LSTM 模型进行不断地学习分类

模型评估　采用 F1-Score 进行分类方法的评估，并进行卡方分布的预测。

针对问题二：

1、对留言主题和留言详情分别进行去重，和停用词删除。并将它们转化为 TF-IDF 特征值

2、利用 K-means 对留言主题和留言详情的文本向量的特征权重分别进行聚类，根据手肘法找到最佳分类簇 K，进行一次聚类。

3、选择 Hanlp 来实现基于 HMM（隐马尔可夫模型）的命名实体识别，将留言详情的文本中地点/人群进行自动摘要。而后根据提取出来的地点，对出现同一地点的问题进行二次聚类。

4、TextRank 和 BM25 进行文本关键句摘要

5、构建话题热度公式模型，通过统计留言高频问题与问题点赞数以及时间跨度大小来排序

针对问题三：

1、基于 JS 散度和余弦距离进性文本相似度计算，构建相关性的评价体系。

2、对答复时间和提问时间进行处理，计算时间跨度，根据自定义的时间指标，构建时效性评价体系。

3、基于依存句法分析和 N-gram 构建答复的完整性评价体系。

4、基于 LDA 的主题模型解决文本可解释性问题，构建可解释性评价体系。

关键字： TF-idf 算法, LSTM 神经网络, 命名实体识别, LDA 主题模型

Application of text mining in intelligent government affairs

Abstract

In recent years, such as weibo, mayor mailbox, WeChat, sun hotline network asked ZhengPing gradually become the government to understand public opinion, pooling intelligence, condensed the important channel which the bull, the text of all kinds of public opinion related data volume rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government. Based on the data mining technology, this paper will carry out the internal information mining on the government message information data, extract the part that we need to analyze, conduct in-depth mining and analysis, explore the hot issues of government affairs, and evaluate the response perfection of the government affairs system.

For question 1:

1、The original data of attachment 2 is simplified by using Chinese stop word database. The simplification of data includes the duplication of jieba word database and the corresponding deletion of Chinese stop word database. Then word frequency (TF) is used to select the features of the simplified text

2、 The weight assignment is to optimize the text space. First, the first level classification is converted into natural Numbers. Secondly, tf-idf algorithm is adopted to vectorize the details of the first-level classification and the simplified comments, and then the fitting degree of the weights of high-frequency words and low-frequency words and first-level classification weights is compared

3、The LSTM model is selected for model selection, and LSTM model is used to continuously learn and classify test set text and training set text of feature vectors of message details

4、Model evaluation adopts f1-score to evaluate the classification method

For question 2:

1、delete the message subject and message details, and stop words. And convert them to tf-idf eigenvalues

2、use k-means to cluster the feature weights of the text vectors of the message subject and message details respectively, obtain the clustering center, and then classify the topic according to KNN.

3、select Hanlp to realize named entity recognition based on HMM (hidden markov model), and automatically summarize the location/population in the text of message details. In addition, word2vec based word similarity calculation model is used to combine similar problems

4、TextRank and BM25 for text key sentence summarization

5、build the topic heat formula model, and sort by counting the number of high-frequency questions and thumb up questions and the time span
For question 3:
1、Build a correlation evaluation system based on the text similarity calculation of JS divergence and cosine distance progression.
2、Deal with the response time and question time, calculate the time span, and build the timeliness evaluation system according to the customized time index.
3、Construct a response integrity evaluation system based on dependency syntax analysis and n-gram.
4、The theme model based on LDA solves the problem of text interpretability and constructs an interpretability evaluation system.

Key words: tf-idf algorithm ,LSTM neural network, named entity recognition, LDA theme model

目 录

1 挖掘目标

随着网络问政平台各类社情民意相关的文本数据量攀升，为减轻政务系统中，人工分类的压力，本文将采用利用自然语言处理和文本挖掘的方法，达到三个目标：

1、利用文本分词和长短期记忆网络(LSTM)的方法对非结构化的数据进行文本挖掘，采用无监督的学习方式对留言详情进行主题分类。

2、采用 TF-idf 将文本结构向量化，K-means 对文本特征向量聚类分析，并采用二分 K 均值算法，找到最佳分类簇。同时，选择 Hanlp 实现基于 HMM（隐马尔可夫模型）的命名实体识别，将留言详情的文本中地点/人群进行自动摘要。此外，采用基于 word2vec 的词语相似度计算模型来将同类问题合并。得到热点问题表和热点问题留言明细表。

3、对留言答复进行设定相关性，完整性，可解释性，以及时效性方面的评价

2 问题分析方法过程

2.1 问题 1 的流程图

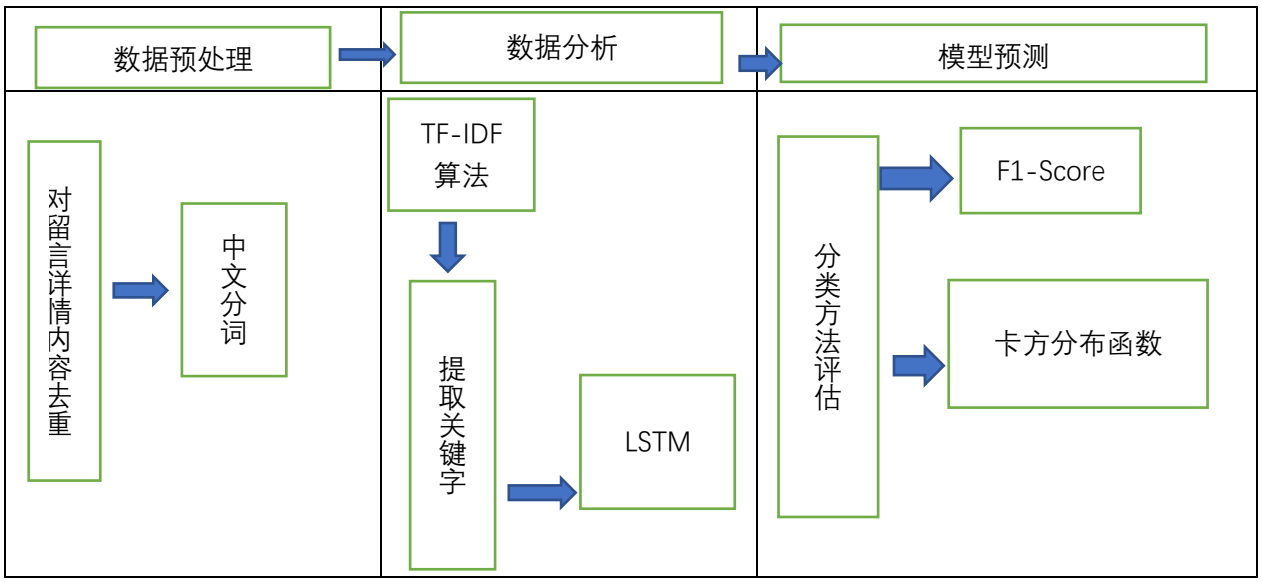

图 1：问题一流程

本用例主要包括以下步骤：

步骤 1：数据预处理，在题目给出的数据中，出现了很多重复的留言数据，在原始的数据上进行去重处理，之后进行文本的中文分词操作。

步骤 2：数据分析，在对留言详情分词后，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，找出每个留言内容的关键词，把留言内容信息转换为权重向量。使用 LSTM 进行无监督的学习进而自主预测分类。
步骤 3：模型评估，采用 F1-Score 对分类方法进行评估，采用卡方分布函数，测试词语对与词语间的拟合度

2.1.1 数据预处理

2.1.1.1 一级分类的数值化

在题目给出的数据中，出现和很多为空值的数据，因此我们第一步就是去除值为 NAN 的数据，我们首先对各个一级分类的频次做了统计

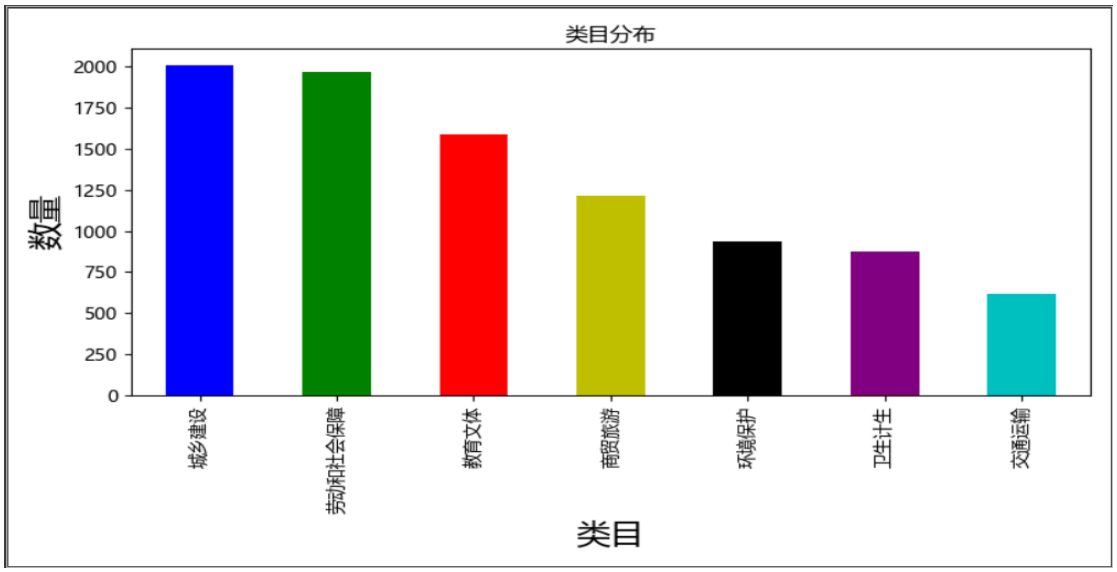

图 2：各个一级分类数量统计图

于是我们对一级分类的 id，也做了处理，将其转化为对应的 0-6 个自然数

表格 1：一级分类对应的 id 标签

Question_id	**一级分类**
0	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

2.1.1.2 对留言数据的去重和进行中文分词

在对留言信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。

在留言详情描述表中，以中文文本的方式给出了数据。为了便于转换，先要对这些主题描述信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，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)，同时采用了动态规划查找最大概率路径， 找出基于词频的最大切分组合，对于未登录词，采用了基于 汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。

而后引入中文停用词库，进行中文分词。

2.1.1.3 TF-IDF 算法

为了将文本特征结构化，我们这里采用 TF-IDF 的处理方式。

Tf-idf 算法目的是评估某个值对某个文本的重要程度，如果一个词或短语在一篇文章中出现的频率高，并且在文档的集合中出现频次低，则认为此词或短语具有很好的类别区分能力。

Tf-idf 由两个部分组成：词频（TF）指的是某一特定的词语在该文本中出现的频率；反文档频率（IDF）,即文本数量与某一个特定的词语在文本集中出现的次数的比值。

假设特征词 i 在文本 d 中出现的词频为tf_i（d）,n_i表示为含有特征词 i 的文本数，则 Tf-idf 函数为：

$$TFIDF_i(d) = tf_i(d) * ln\left(\frac{N}{n_i}\right) \qquad (1\text{-}1)$$

为了弱化极个别高频词对低频词抑制的作用，需要对以上 TF-IDF 值做归一化处理，其中 i 指特征词总数：

$$V_d = \sum_{i=1}^{l} \sqrt{（\mathrm{TFIDF_i}（\mathrm{d}））^2} \qquad (1\text{-}2)$$

2.1.1.4 生成 TF-IDF 向量

表格 2：附件 2 测试文本中文本 TF-IDF 部分数据

稀疏向量矩阵行列序号	在文档中的频率
(0, 30930)	0.1429165551642627
(0, 30643)	0.1429165551642627
(0, 47987)	0.1429165551642627
(0, 30937)	0.1429165551642627
(0, 19467)	0.1429165551642627
……	
(494, 48003)	0.05360450439942652
(494, 3269)	0.07648628079806415
(494, 3949)	0.11753421985839935
(494,22132)	0.05876710992919967

2.1.2 LSTM 进行无监督的学习分类

长短期记忆（long short-term memory,LSTM）是一种特殊的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。简单来说就是比普通 RNN 在更长的序列中有更好的表现。

LSTM 主要包括三个控件：

1、遗忘门：

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \quad (2\text{-}1)$$

由当前输入数据、上一时刻的隐藏状态（输出）一起做全连接并使用 sigmoid 制作出来。其负责与记忆细胞做元素乘确定哪些信息保留继续往后传导。

2、输入门:由当前输入数据、上一时刻的隐藏状态（输出）一起做全连接并使用 sigmoid 制作出来。其与候选记忆细胞做元素乘确定哪些新信息加入记忆细胞。

3、输出门:由当前输入数据、上一时刻的隐藏状态（输出）一起做全连接并使用 sigmoid 制作出来。其与经过遗忘门，和加入新信息之后的记忆细胞做元素乘，确定哪些信息作为当前时刻的输出。

2.1.3 模型评估

2.1.3.1 F1-Score 评估方法

F1 分数（F1-score）是分类问题的一个衡量指标。一些多分类问题的机器学习竞赛，常常将 F1-score 作为最终测评的方法。它是精确率和召回率的调和平均数，最大为 1，最小为 0。

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

2.1.3.2. 卡方分布模型预测

通过卡方分布我们可以获得词语对与主题间的拟合度分析，结果如下：

表格 3：一级分类对应的最佳拟合词语对

交通运输	. Most correlated unigrams: . 快递 . 出租车	. Most correlated bigrams: . 出租车 公司 . 出租车 司机
劳动和社会保障	Most correlated unigrams: . 退休 . 职工	Most correlated bigrams: . 劳务 派遣 . 最低工资 标准
卫生计生	Most correlated unigrams: . 医生 . 医院	Most correlated bigrams: . 受理 请问 . 乡村 医生
商贸旅游	Most correlated unigrams: . 导游 . 广告	Most correlated bigrams: . 资质 认定 . 检验 检测
城乡建设	Most correlated unigrams: . 小区 . 业主	Most correlated bigrams: . 汽车 检测站 . 年市 廉租房
教育文体	Most correlated unigrams: . 文化 . 学校	Most correlated bigrams: . 培训 机构 . 学校 老师
环境保护	Most correlated unigrams: . 环保局	Most correlated bigrams: . 环境 监测站

	. 污染	. 生态 破坏

2. 2. 问题 2 的流程图

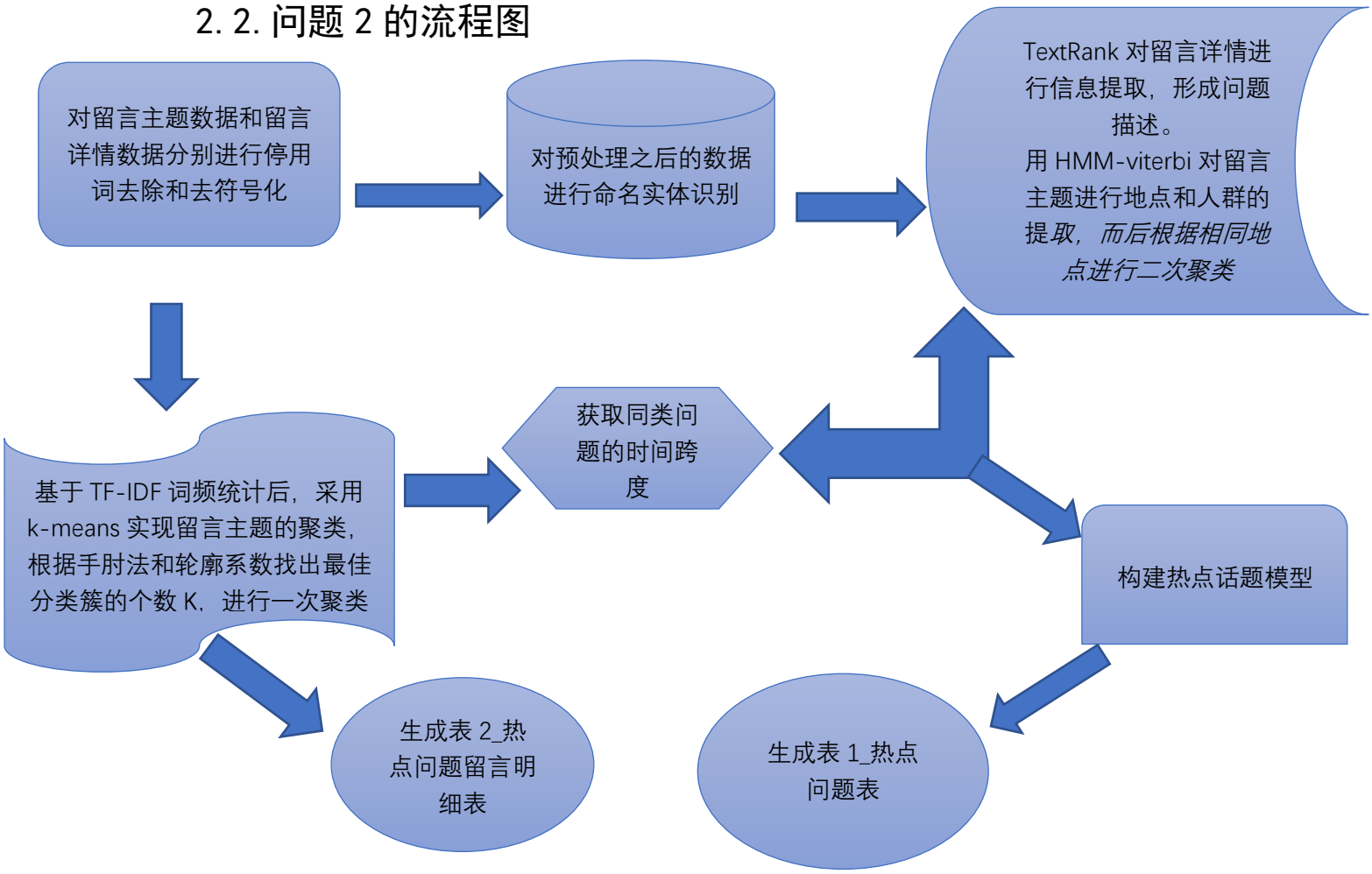

图 3：问题 2 的流程

2.2.1 数据预处理

2.2.1.1 对留言主题和留言详情的去重和进行中文分词

和针对问题 1 的数据处理一样，我们采用 jieba 分词和停用词库对原始数据进行处理。

如下图结果所示（详细数据见）：

表格 4：部分分词示例

A 市 经济 学院 体育 学院 变相 强制 实习
A 市 人才 APP 申请 购房 补贴 通不过
A5 区 劳动 东路 魅力 城 小区 临街 门面 烧烤 夜宵 摊
请 K3 县 乡村 医生 发 卫生室 执业 许可证
A5 区 劳动 东路 魅力 城 小区 一楼 夜宵 摊 污染 空气

2.2.1.2 生成 TF-IDF 向量

生成 TF-IDF 向量的具体步骤如下：

1、使用 TF-IDF 算法，找出每个留言问题描述的前 5 个关键词；

2、对每个留言问题描述提取的 5 个关键词，合并成一个集合，计算每个留言问题描述 对于这个集合中词的词频，如果没有则记为 0；

3、生成各个留言问题描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF）

2.2.2 K-means 算法

生成留言问题描述的 TF-IDF 权重向量后，根据每个留言问题的 TF-IDF 权重向量，对 留言问题进行分类。

K-mean 聚类的原理如下：假设有一个包含 n 个 d 维数据点的数据集 X = $\{x_1, x_2, x_3, x_4 \cdots x_i \cdots x_n\}$其中$x_i \epsilon R^d$ ，K-means 聚类将数据集 X 组织为 K 个划分 C = $\{c_k, i = 1,2 \ldots k\}$ 每个划分代表一个类c_k，每个类c_k有一个类别中心μ_i。选取欧式距离作为相似性和距离判断准则，计算该类内个点到聚类中心μ_i的距离平方和

$$J(c_k) = \sum_{x_i \epsilon c_k} ||x_i - \mu_k||^2 \qquad (2\text{-}3)$$

聚类目标是使各类总的距离平方和 J(c) = $\sum_{K=1}^{K} J(c_k)$ 最小

$$J(c) = \sum_{K=1}^{K} J(c_k) = \sum_{K=1}^{K} \sum_{x_i \in c_k} d_{ki} ||x_i - \mu_k||^2 \qquad (2\text{-}4)$$

其中，$d_{ki} = \begin{cases} 1, & 若 x_i \in c_i \\ 0, & 若 x_i \notin c_i \end{cases}$，所以根据最小二乘法原理和拉格朗日原理，聚类中心$\mu_k$，

应该取为类别c_k类各数据点的平均值。

K-mean 聚类的算法步骤如下：

1、从 X 中随机取 K 个元素，作为 K 个簇的各自的中心。

2、分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

3、根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。

4、将 X 中全部元素按照新的中心重新聚类。

5、重复第 4 步，直到聚类结果不再变化。

6、将结果输出。

K-mean 聚类的算法流程图如下：

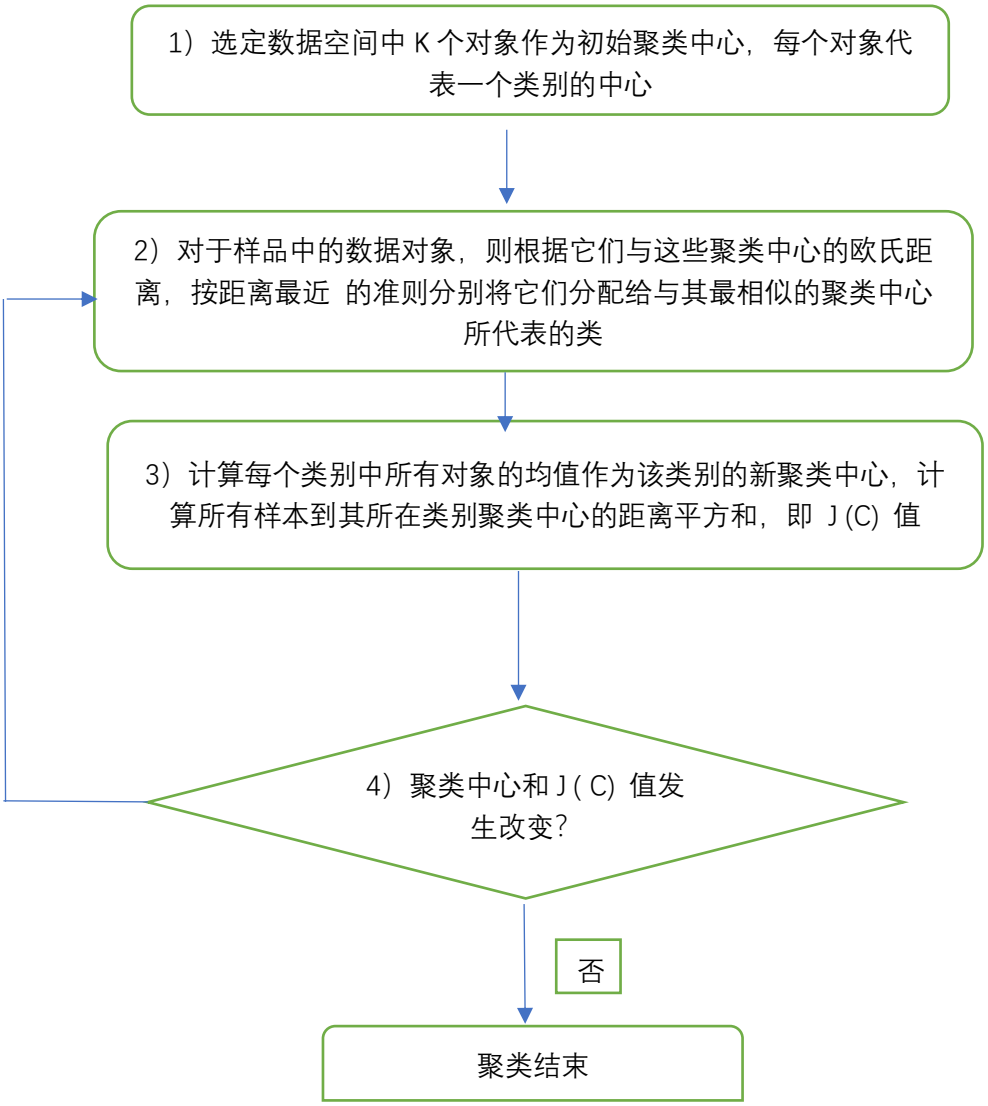

图 4：K-means 流程图

在 K-means 中，寻找最佳分类簇 K 是提高优化 K-means 性能的手段之一。本文选用手肘法和轮廓系数，寻找最佳分类簇。

手肘法的核心指标叫做 SSE(Sum of the squared errors,误差平方和)：

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \qquad (2\text{-}5)$$

2.2.3 命名实体识别

实体识别属于信息抽取技术中重要的一个关键任务，其主要作用是从大量的非结构化或半结构化的数据中提取出人们所关注的命名实体，包括人名，地名，团体组织等

针对问题 2，我们需要提取出留言主题中的事件地点主体以及事件人物主体，目前的实体识别主要是两种方法：一是基于规则集的命名实体识别，二是基于统计方法的命名实体识别。本文选择基于统计的命名实体识别。

2.2.3.1 HMM(隐马尔可夫模型) 地名识别

隐马尔科夫模型（Hidden Markov Model）是关于时序的概率模型，描述由一个隐含的马尔科夫链生成不可观测的状态序列，再由状态序列生成观测序列的过程。这种通过观测序列预测隐含的标记序列的问题叫做标注。

它用来描述一个含有隐含未知参数的马尔可夫过程。其难点是从可观察的参数中确定该过程的隐含参数。然后利用这些参数来作进一步的分析，例如模式识别。

以下是我们基于 HMM 命名实体识别实现的词性标注的示例：

表格 5：命名实体识别示例

原句： A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气
词性分析： A5 区/n 劳动/v 东路/ns 魅力/n 之城/ns 小区/n 一楼/n 的/u 夜宵/n 摊/n 严重/ad 污染/v 附近/f 的/u 空气/n

在 HMM 中，有 5 个基本元素：{N,M,A,B, π}

M:观察值的有限集合。在这里，是指每一个词语本身。

A:状态转移概率矩阵。在这里，是指某一个标注转移到下一个标注的概率。

B:观测概率矩阵，也就是发射概率矩阵。在这里，是指在某个标注下，生

成某个词的概率。

π:初始概率矩阵。在这里，是指每一个标注的初始化概率。

N:状态的有限集合。在这里，是指每一个词语背后的标注

而以上的这些元素，都是可以从训练语料集中统计出来的。然后，我们可以根据这些统计值，应用维特比（viterbi）算法，算出词语序列背后的标注序列。

注：这里我们采用的语料库是 MSR（微软语料库）和 PKU（北京大学语料库）

Viterbi 算法

维特比算法是一种动态规划算法用于最可能产生观测时间序列的-维特比路径-隐含状态序列，特别是在马尔可夫信息源上下文和隐马尔科夫模型中。术语“维特比路径”和“维特比算法”也被用于寻找观察结果最有可能解释的相关动态规划算法。例如在统计句法分析中动态规划可以被用于发现最有可能的上下文无关的派生的字符串，有时被称为“维特比分析”。

2.2.4 关键句提取

关于获取留言事件的问题描述，我们计划通过对附件 3 中的留言详情进行关键句处理来获取。

2.2.4.1 BM25

在信息检索领域中，BM25 是 TF-IDF 的一种改进变种。TF-IDF 衡量的是单个词语在文档中的重要程度，而在搜索引擎中，查询串(query)往往是由多个词语构成的。如何衡量多个词语与文档的关联程度，就是 BM25 所解决的问题。

形式化的定义 Q 为查询语句，由关键字 q1 到 qn 组成，D 为一个被检索的文档，BM25 度量如下:

$$BM25\,(D,Q)\ =\ \sum_{i=1}^{n} IDF(q_i) * \frac{TF(q_i,D)*(k_1+1)}{TF(q_i,D)+k_1*(1-b+b*\frac{|D|}{avgDL})} \qquad (2\text{-}6)$$

2.2.4.2 TextRank

有了 BM25 算法之后，将一个句子视作查询语句，相邻的句子视作待查询的文档，就能得到它们之间的相似度。以此相似度作为 PageRank 中的链接的权重，于是得到一种改进算法，称为 TextRank。它的形式化计算方法如下：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{BM25(V_i,V_j)}{\sum_{V_k \in Out(V_j)} Bu2s(V_k,V_j)} WS(V_j) \qquad (2\text{-}7)$$

其中，WS(Vi) 就是文档中第 i 个句子的得分，重复迭代该表达式若干次之后得到最终的分值，排序后输出前 N 个即得到关键句。

2.2.4 构建热度模型

热度模型是我们自定义的，为：

$$\text{热度} = \begin{cases} (\text{点赞数} + \text{反对数}) * \text{这类问题反映的次数} * \text{投诉天数}, & (\text{点赞数} + \text{反对数})! = 0 \\ \text{这类问题反映的次数} * \text{投诉天数}, & (\text{点赞数} + \text{反对数}) = 0 \end{cases} \qquad (2\text{-}8)$$

2.3 问题 3 的流程图

图 5：问题三评价体系

2.3.1 相关性的评价

在对答复意见进行评定时，我们首先解决的是文本相关性问题。答复是否与问题相关，是后续评价标准的基础。基于答复和问题分别属于两个不同

的分布这一特点，故而，我们采用 JS 散度对整体的答复相关度进行评估，之后选择基于余弦距离的语义相似度评定每一条回复的相关性。

2.3.1.1 JS 散度

JS 散度是度量两个概率分布的相似度，是基于 KL 散度的变体，解决了 KL 散度非对称的问题。

JS 散度相似度衡量指标：

现有两个分布P_1和P_2，其 JS 散度公式为:

$$JS(P_1 \parallel P_2) = \frac{1}{2}KL(P_1 \parallel \frac{P_1+P_2}{2}) + \frac{1}{2}KL(P_2 \parallel \frac{P_1+P_2}{2}) \quad (3\text{-}1)$$

其中 KL 表示 KL 散度又称相对熵，信息散度，信息增益。KL 散度是两个概率分布 P 和 Q 差别的非对称性的度量。在经典境况下，P 表示数据的真实分布，Q 表示数据的理论分布，模型分布。

定义如下：

$$DL(P \parallel Q) = -\sum P(x)\log 1/P(x) + \sum P(x)\log 1/Q(x) = \sum P(x)\log P(x)/Q(x) \quad (3\text{-}2)$$

对数函数为凸函数，所以 KL 散度的值为分复数。

KL 散度和 JS 散度度量的时候都有一个问题：如果两个分布 P,Q 距离较远，完全没有重叠的时候，KL 散度是没有意义的，在学习的时候，这就意味着在这一点的梯度为 0，即梯度消失了。

我们在测试文本中，得到评价的 JS 散度为 0.07707653410356628，表示整体答复大部分符合相关性的要求。

2.3.1.2 基于余弦距离的语义相似度计算

一个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近 1，夹角趋于 0，表明两个向量越相似，余弦值接近于 0，夹角趋于 90 度，表明两个向量越不相似。

基本思路是：

如果这两句话的用词越相似，它们的内容就应该越相似。因此，可以从词频入手，计算它们的相似程度。

步骤：

1、将数据映射为高维空间中的点（向量）

2、计算向量间的余弦值

取值范围[-1,+1]，越趋近于 1 代表越相似，越趋近于-1 代表方向相反，0 代表正交

$$\cos\theta = \frac{a*b}{\|a\|\|b\|} \tag{3-3}$$

$$\cos\theta = \frac{x_1x_2+y_1y_2}{\sqrt{x_1{}^2+x_2{}^2}+\sqrt{y_1{}^2+y_2{}^2}} \tag{3-4}$$

2.3.2 完整性的评价

在文本完整性的处理上面，我们选择两个层次的完整性进行处理。一个是广义上的完整性，表示内容完整性，另一个是狭义上的完整性，表示语义完整性。

2.3.2.1 基于相关性分析内容完整性

对于内容的完整性，我们采用基于之前的处理的相关性作为评价标准。相关性越强，说明答复和问题的内容拟合度越高，故而完整性就越好。

2.3.2.2 基于 N-Gram 判断句子是否通顺

基于我们对语义完整性的理解，我们想要解决的是，答复内容是否满足某种特定的语法结构，通俗的来说就是判断句子是否通顺的问题。

为满足这一需求，我们采用基于 N-Gram 判断句子是否通顺的方式进行处理，分为如下几个步骤：

首先加载分词和词性标注模型，加载训练数据，并对数据进行分词和词性标注，在句首句尾分别加上<s>和</s>作为句子开始和结束的标记。

对训练数据中标签为 0 的数据进行 1-gram 和 2-grams 的词性频率计数。

统计训练集中 2-Grams 概率的最小值，最大值，以及平均值，其中平均值将被用作判断句子好坏的阈值

接着计算测试集中每个句子基于 2-Grams 的除以字长的概率值，然后与由训练集计算得到的与其等字长类的平均概率值进行对比。如果训练集中没有找到与测试集中某个句子的等长的句子，则测试集中该句子概率值直接去总体训练样本计算得到的概率值进行对比。

关于 N-Gram 的原理，介绍如下：

N-Gram 是一种基于统计语言模型的算法。它的基本思想是将文本里面的内容按照字节进行大小为 N 的滑动窗口操作，形成了长度是 N 的字节片段序列。

每一个字节片段称为 gram，对所有 gram 的出现频度进行统计，并且按照事先设定好的阈值进行过滤，形成关键 gram 列表，也就是这个文本的向量特征空间，列表中的每一种 gram 就是一个特征向量维度。

该模型基于这样一种假设，第 N 个词的出现只与前面 N-1 个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。这些概率可以通过直接从语料中统计 N 个词同时出现的次数得到。

当 n=1, 一个一元模型（unigram model)即为 ：

$$P(w_1, w_2, \ldots . w_m) = \prod_{i=1}^{m} P(w_i) \quad (3\text{-}5)$$

当 n=2, 一个二元模型（bigram model)即为 ：

$$P(w_1, w_2, \ldots . w_m) = \prod_{i=1}^{m} P(w_i | w_{i-1}) \quad (3\text{-}6)$$

当 n=3, 一个三元模型（trigram model)即为:

$$P(w_1, w_2, \ldots . w_m) = \prod_{i=1}^{m} P(w_i | w_{i-2} w_{i-1}) \quad (3\text{-}7)$$

在给定的训练语料中，利用贝叶斯定理，将上述的条件概率值都统计计算出来即可。

2.3.3 时效性的评价

评价答复的及时性，我们设定了四个标准，如下：

表格 6：时效性等级评价标准

时间跨度	处理等级评定
7 天	S
15 天	A
30 天	B
更久	C

基于对附件四的时间处理很快就可以得出，关于时效性的评估。

2.3.4 基于 LDA 分类模型的可解释性的评价

可解释性的评价，体现在两个方面，一是模型可解释性，二是文本可解释性。为解决前者，我们采用基于 LDA 模型的方式，而对于后者，主要是强调解决是否有理可据的问题。

故而，我们选择对附件 4 的留言主题和答复详情两列分别与附件 1 的一级类目相关分类进行聚类，将问题与答复的主题拟合度，作为我们文本内容可解释性的标准。

基于 LDA 的可解释性评价流程如下图所

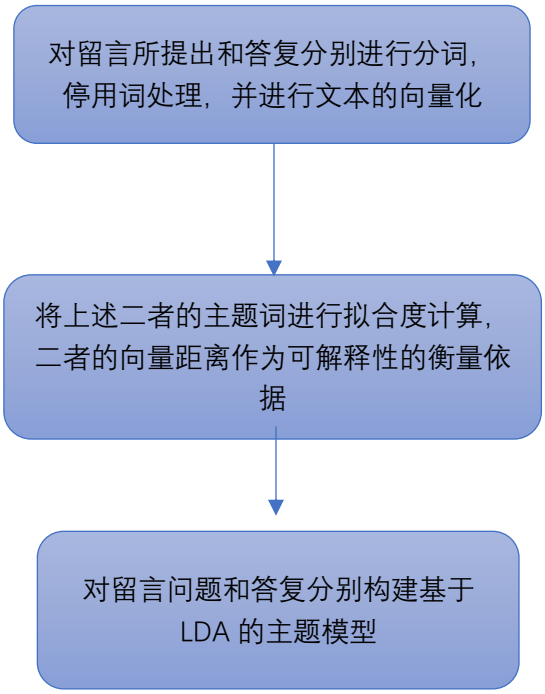

图 5：LDA 解决可解释性流程图

3 结果分析

3.1 问题 1 结果分析

3.1.1 LSTM 聚类分类结果

基于文本的向量化，以及训练和测试的数据集都准备好之后，接下来我们要定义一个的序列模型：

- 由于模型的第一层是嵌入层(Embedding)，它使用长度为 100 的向量来表示每一个词语
- SpatialDropout1D 层在训练中每次更新时， 将输入单元的按比率随机设置为 0， 这有助于防止过拟合
- LSTM 层包含 100 个记忆单元
- 输出层为包含 10 个分类的全连接层
- 由于是多分类，所以激活函数设置为 softmax
- 由于是多分类，所以损失函数为分类交叉熵 categorical_crossentropy

最终预测的文本为：

表格 7：部分一级标签预测文本

	留言详情	一级标签
6862	西地	劳动和社会
8028	请上级领导	商贸旅游
4096		教育文体
6433		劳动和社会
1323	敬爱的蒙牛	城乡建设
41	A6区奥新佳	城乡建设
5389	尊敬	劳动和社会
7064	本人	劳动和社会
2534	各位网友，	环境保护
498	海棠路山水	城乡建设

3.1.2 LSTM 模型的评估

对 LSTM 模型的评估如下：

表格 8：问题一模型评估

accuracy 0.9006578947368421				
	Precision	recall	f1-score	support
城乡建设	0.82	0.95	0.88	663
环境保护	0.96	0.93	0.94	310
交通运输	0.94	0.69	0.80	202
教育文体	0.94	0.94	0.94	525
劳动和社会保障	0.91	0.94	0.92	650
商贸旅游	0.90	0.83	0.86	401
卫生计生	0.95	0.86	0.90	289
Accuracy			0.90	3040
macro avg	0.92	0.88	0.89	3040
weighted avg	0.90	0.90	0.90	3040

基于上述表格显示，我们可以发现基于长短时记忆神经网络解决多分类问题，准确率可以达到 90%。并以较高的召回率和精确度来测试样本的分类归属问题。

3.2 问题 2 结果分析

3.2.1 对 K-means 聚类分类结果分析

基于对附件三进行文本分词及向量化之后，我们得到如下的分词短句：

表格 9：附件 3 留言主题部分分词

A3 区 一米阳光 婚纱 艺术摄影 合法 纳税
西地省 师大附中 校园 喇叭 太 扰民
举报 A1 区华海 3C 朝阳 菜市场 夜宵 摊 门店 口 摆满 桌椅
请问 A2 区西 牌楼 小区 有无 拆迁 计划
A4 区 大道 黄泥 扩 段 南北 约 1500 米 路 灯不亮
A5 区 万家 丽 中路 泰禹 家园 住 改商 改成 麻将馆 深夜 麻将

基于此，生成了 TF-IDF 文本向量矩阵(详情文件见附件留言主题--

TFIDF)。而后进行基于手肘法求最优 K 值的计算，得到如下结果：

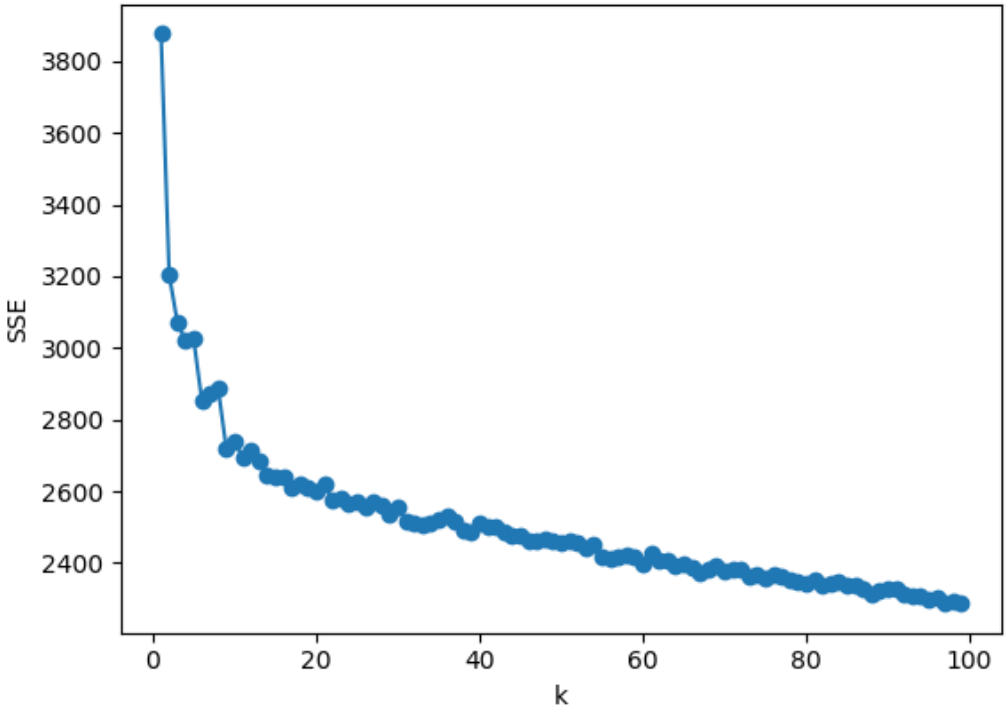

图 6：手肘法观察最优 k 值

从以上的手肘图可见，附件三大约分为 90 个簇的时候，对应的误差平方和（SSE）趋于平缓，说明分类簇为 90 个时 K-means 的聚类更合理。基于此我们得到 K-means 的分类效果（见附件 1）

3.2.2 对命名实体识别进行结果分析

在进行热点构造和人群地点提取时，我们选择直接采用 Hanlp 工具来进行地点的识别和基于规则的方式来提取地点，效果如下：

表格 10：地点/人群提取部分示例

地点:['A 市', '经济', '学院'] 词性分析: A 市/ns 经济/n 学院/n 寒假/t 过年/vn 期间/f 组织/v 学生/n 去/v 工厂/n 工作/v
地点:['A 市', '魅力', '之城', '商铺'] 词性分析: A 市/ns 魅力/n 之城/ns 商铺/n 无/v 排烟/vn 管道/n ，/w 小区/n 内/f 到处/d 油烟味/n

地点:['A7 县', '江背镇', '朱家桥', '社区'] 词性分析: A7 县/ns 江背镇/ns 朱家桥/ns 社区/ns 工厂/n 夜间/t 焚烧/v 工业/n 垃圾/n 存在/v 很多/m 年/q 了/y
地点:['A4', '区', '清水塘', '溪泉', '湾', '小学'] 词性分析: A4/nx 区/n 一/m 非法/b 驾校/n 与/p 清水塘/nz 溪泉/n 湾/ns 小学/n 之间/f 仅/d 隔/v 一道/d 栏杆/n
地点:['A7 县', '青山铺镇', '羊思坡'] 词性分析: A7 县/ns 青山铺镇/ns 羊思坡/ns 修路/v 补偿款/n 一直/d 没/d 发放/v

在命名实体识别过程中，我们采用的语料库是网络上存在的 PKU（中国人民日报）语料库和 MSR（微软亚洲研究所）的语料库进行模型的训练，故而基于该模型的热点问题测试集精确度存在一定的困难，但是地点识别的效果精确度可以达到 72%，能够满足地点实体识别的需求。其效果如下：

表格 11：命名识别实体准确率分析

NER （识别实体）	P	R	F1
ns	72.55	26.81	39.15
nt	80.14	30.71	44.40
avg.	79.01	30.11	43.60

而后得到表 2-热点问题留言明细表和表 1-热点问题表（见附件）

3.3 问题 3 结果分析

3.3.1 时效性评价

根据答复时间和提问时间的时间差，给定相应的评价标准（详见附件时效性.xls）

3.3.2 相关性评价

整体上，通过 JS 散度判断整体回复的相关性。局部上，采用余弦距离的中心思想分析每一条回复和提问的相关性。（结果如附件相关性.xls 所示）

3.3.3 完整性评价

在解决语义完整性上，采用 N-gram 判断语句是否通顺，得到的效果如下（结果如附件完整性.xls 所示），其中 0 表示不完整，1 表示完整。

3.3.4 可解释性评价

采用 LDA 提取主题词，效果如下：

```
Topic 138: 工作 符合 户籍 妻子 认定 诉求 扶贫 费用 肇事 赔偿 张员喜 家具 小准 花费 吴勤 儿媳 处理结果 事实 做好 交警大队
Topic 139: 工作 改革 事业单位 全额 公益 财政 拨款 编办 农机 省市 技术 做出 核定 机械化 差额 大力支持 财政局 服务 办法 释放
Topic 140: 政府 楼房 居民点 办公 人员 包括 危房 地基 老屋 垮塌 呼吁 受阻 实为 银田镇 来源 请教 外出 谢周 合适 希望
Topic 141: 详见 批准 人民政府 接受 大力支持 闲置 村支 无缝 名苑 交齐 充分利用 迹象 卫生间 建管 星沙人 解决方案 违规 标准 县政府 关系
Topic 142: 理解 支持 留言 网友 监督 后市 过渡 专人 线索 增长 省厅 整理 准入 确保 信箱 员工 瓶颈 发脾气 德政 服从安排
Topic 143: 谢谢 行人 资质 任意 不肯 滋味 咳嗽 胡军 增速 大树 至星 亚用 比赛 纯净 难不成 李林华 进驻 国土资源厅 实地 胚房
Topic 144: 回复 铺设 搅拌站 依法 娄双 油砂 工程 子堤 集体 无效 环保部门 损坏 引进 白改 顶楼 供应 货物 承包户 租赁 城堤
Topic 145: 修复 组织 损坏 开挖 施工 涉及 倾倒 经查 黄泥 修补 内脏 动物 闲置 经区 阻断 国标 论证 皮卡 只能靠 余亮
Topic 146: 有限公司 情况 科技 导致 来信 起重 知悉 生态 混杂 水电 利润 剖宫产 挖掉 清水 牛头 组合 瓷砖 添加 弄伤 借助
Topic 147: 尊敬 胡书记 感谢 召集 帮忙 时间 活动 洪灾 师德师 推送 钱财 交叉 迅伟 人力 时节 村团 有误 迁入 物品 雨水
Topic 148: 拆除 搭建 违建 督促 铁门 未办理 改变 协商 经县 楼顶 强力 对接 抗议 农合 公分 侵占 阳光房 城乡规划 报备 承建方
Topic 149: 设施 配套 现状 配备 红线 经理 功能 收到 密度 加盟费 可用 统筹 支部书记 宝贵 二手房 洪家 距离 认定 接到 配套工程
Topic 150: 整治 专项 执法 联合 秩序 城管 重点 工作 职能部门 取缔 查处 治理 乱象 部门 效果 打击 领导小组 常态 相结合 乞讨
Topic 151: 发电 稳定 群众利益 泥土 风力 饮用水 施塘 人数 报告 人口 甘益 村村 水源 引用 局于 异动 实地考察 周家 发现 人工
Topic 152: 情况 留言 分管领导 用于 采用 星沙 告知 询问 收到 栏杆 往返 办去 儿子 划设 提出 预防 原始 复印 艰苦 三环
Topic 153: 装修 建筑面积 公摊面积 毛坯 入住 包含 装饰 铲除 材料 花园 用于 设备 设计费 可向 墙体 保温层 提到 原则 实施细则 建设厅
Topic 154: 回复 困难 职能部门 为民 法规 调查 符合规定 理解 合并 改革 岸段 装修 文化娱乐 督办 烧坏 计划 调过去 证乙字 完美 足球
Topic 155: 咨询 连续 位置 提升 交通 返还 交通规则 户数 树立 物质 设施 房主 接入 全日制 无证 丢弃 地带 壕塘 领导 追究
```

图 7：LDA 提取部分主题词

图示中数值表示所对应文字出现的词频大小。而后进行主题间的文本特征数值化得到如下效果:

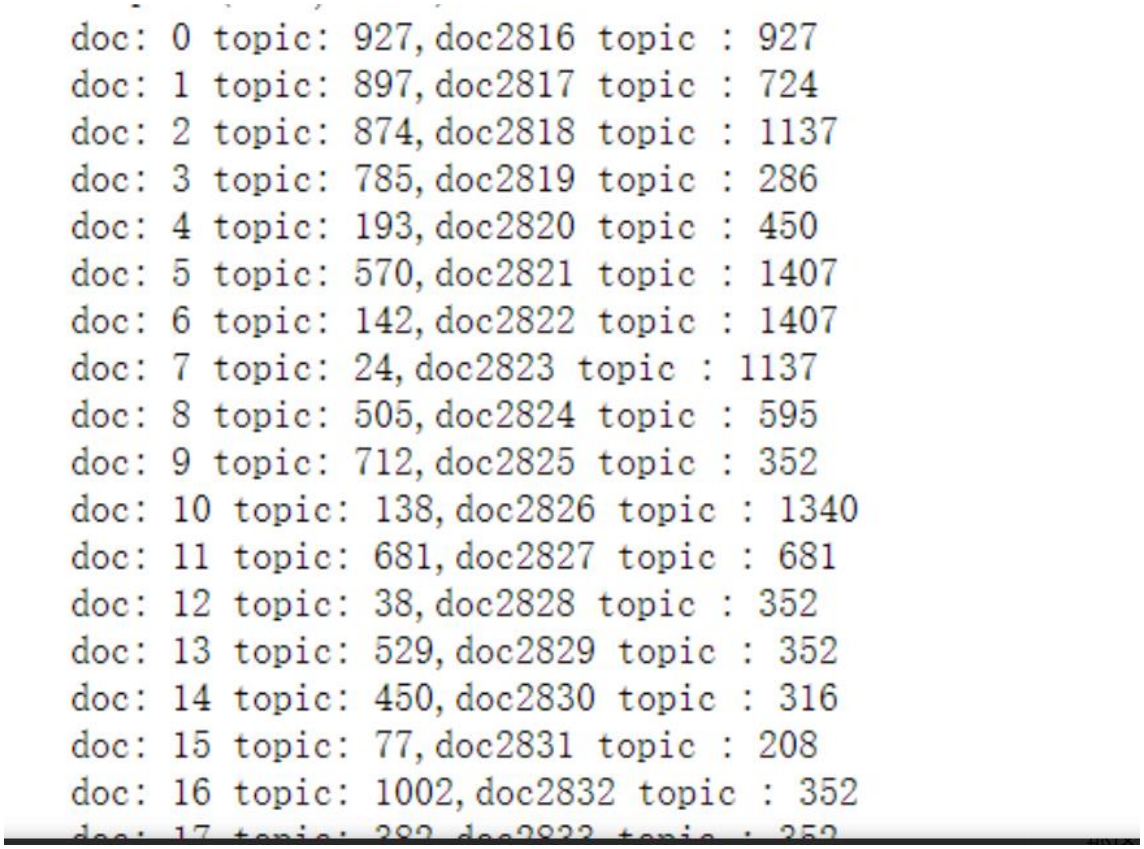

图 8：文本主题特征值图

前两列 doc—topic 对应的是问题特征值，后两列 doc—topic 对应是答复特征值。

二者进行拟合就得到了可解释性的依据如图所示：

	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	liuyan	huifu	new_doc	可解释性
0	2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	2019年4月以来位于A市A2区桂花坪街道的A2区公安分局宿舍区景蓉华苑出现了一番乱象该小区...	现将 网友 平台 问政 西地省 栏目 胡华衡 书记 留言 A2区 景蓉 花苑 物业管理 调查...	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	0.666667
1	2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	潇楚南路从2018年开始修到现在都快一年了路挖得稀烂用围栏围起一直不怎么动工有时候今天来台挖...	网友 A00023583 A3区 潇楚 南路 洋湖 段 修好 A3区 洋湖街道 高度重视 ...	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	0.062500
2	2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	地处省会A市民营幼儿园众多小孩是祖国的未来但民营幼儿园教师一直都是超负荷工作且收入又是所有行...	市民 同志 请加快 提高 民营 幼儿园 教师 待遇 来信 收悉 现 回复 改善 提高 民办...	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	1.000000

图 9：可解释性示例

4 结论

智慧政务的文本挖掘应用，对提高公务人员的办公效率起到了极大作用，同样也是将智慧智能融入政务办公系统的一大创新。本吻摒弃传统的繁杂的人工处理，采用自然语言处理的方式对文本留言数据进行处理。选用长短时记忆神经网络对留言数据进行一级分类的自动划分，并采用 HMM-viterbi 地点识别标注对热点事件进行地点实体挖掘，制定热度划分标准。此外，构建政务答复中的评价指标，分析政务系统处理中的优点与不足。

由分析结果数据显示可以看出，长短时记忆神经网络在分类问题上的效果较好，能够帮助我们解决智慧政务中问题归类的问题。在帮助相关部门进行有针对的处理解决问题上，命名实体识别的地点挖掘为这一需求提供了便利，从数据显示可以看出居民反映的普遍问题为市民对地铁违规的处理的不满以及对配套入学问题的关注。

在政务系统的答复评价中，答复的相关性与时效性呈良好趋势，但在答复完整性和解释性方面存在一定问题，同时也为今后的政务的改进提供了方向。

5 参考文献

[1] 赵琳瑛.基于隐马尔科夫模型的中文命名实体识别研究.西安电子科技大学.2007

[2]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

[3]朱志远.基于数据挖掘的网络招聘系统是设计与实现.电子科技大学.硕士学位论文.2013

[4]王千，王成，冯振元，叶金凤.K-means 聚类算法研究综述.2012

[5]王霞，孙界平，琚生根，胡思才等.基于段落内部推理和联合问题答案匹配的选择型阅读理解模型. 四川大学.2019

[6]吴妮，赵捧未，秦春秀.基于语义分析和相似强度的微博热点发现方法.西安电子科技大学.2015

[7]冯俊龙.基于文本分析的社交网络事件热度预测研究.哈尔滨工业大学.硕士学位论文.2018