C 题：

摘要：近年来，随着微信、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。本文将基于自然语言处理技术对互联网公开来源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。

针对问题一：本文将附件 2 中的所有数据进行去空、去重等处理后，进行中文分词和停用词过滤；利用 LSTM 模型进行预测得到结果。

针对问题二：本文将附件 3 中某一时段内反映特定地点问题的留言进行数据预处理，提取地点信息，分别运用 K-means、DBSCAN 和 LatentDirichletAllocation 三种聚类方法对所提取信息进行聚类分析，采取最优的聚类结果定义热度评价指标，结合评价指标给出评价结果。

关键词：LSTM；K-means 文本聚类；主成分分析；

Abstract: In recent years, as WeChat, Weibo, Mayor's Mailbox, Sunshine Hotline and other online political inquiry platforms have gradually become an important channel for the government to understand public opinion, gather people's wisdom, and gather people's popularity, the amount of text data related to various social conditions and public opinion has been increasing. It has brought great challenges to the work of the relevant departments that used to rely on manual labor for message division and hotspot sorting. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of a smart government system based on natural language processing technology has become a new trend in the development of social governance innovation, which has a great impact on improving the management level and efficiency of government Promote the role. Based on natural language processing technology, this article will carry out internal information mining on the public questioning message records of public sources on the Internet, and the relevant departments' responses to some of the people's messages, and extract the parts we need to analyze for in-depth mining and analysis.

Aiming at Problem 1: After all the data in Annex 2 are

processed, such as blanking and deduplication, Chinese word segmentation and stopword filtering are performed; LSTM model is used to make predictions to obtain the results.

Aiming at problem two: This article preprocesses the messages that reflect a specific location problem within a certain time period in Annex 3, extracts location information, and uses three clustering methods, K-means, DBSCAN, and LatentDirichletAllocation, to perform cluster analysis on the extracted information , Take the optimal clustering result to define the heat evaluation index, and give the evaluation result in combination with the evaluation index.

Keywords: LSTM ; K-means text clustering; principal component analysis;

目录

1.挖掘目标

网上留言信箱凭借其不受时间和空间限制、保护隐私、节约成本等优势，已成为市民们反馈问题和意见的主要渠道。

本次建模的目标是利用某平台收集自互联网公开来源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，其中包含结构化和非结构化文本数据， 在对留言信息进行基本的预处理、停用词过滤后，一方面根据留言信息的分类进行细分，采用长短期记忆（long short-term memory，LSTM）算法，对留言内容的一级标签进行分类；另一方面利用 K-means 聚类算法、DBSCAN 聚类算法和 LatentDirichletAllocation 对留言内容的地点进行聚类分析，采取最优的聚类结果定义热度评价指标，结合评价指标给出评价结果。

2.总体流程步骤

2.1 总体流程

本文的总体架构及思路如下：

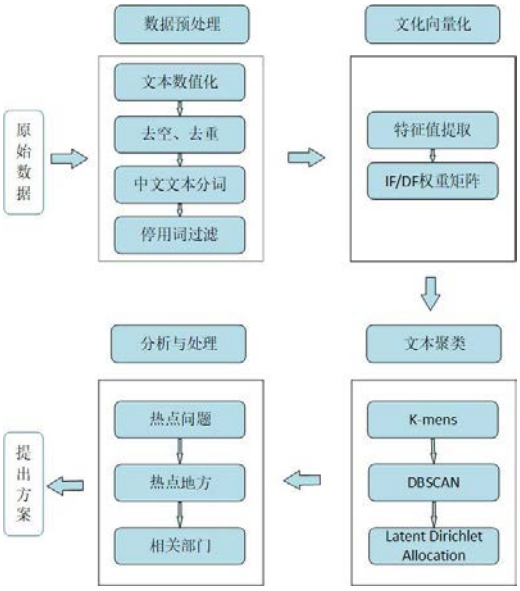


图 2.1 总流程图

步骤一：数据预处理，对附件 2 文本数据进行处理，去除重复项及空行、停用词过滤，以便后续分析；

步骤二：文本向量化，基于 TFIDF 权重法提取关键词，构造词汇-文本矩阵，对语义空间进行降维，去除同义词的影响，简化计算。

步骤三：文本聚类，根据文本向量，计算文档间的欧氏距离，再基于 K-means 聚类算法、DBSCAN 聚类算法和 LatentDirichletAllocation 对留言内容的地点进行聚类分析。

步骤四：分析与预测，构造算法判断热点问题、定义合理的热度评价指标，并给出评价结果。

步骤五：针对附件 4，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。

3.问题一分析方法与过程

3.1 数据预处理

通过观察所给数据，可以发现数据量比较大，且附件 2 中的字段大多为文本格式，需要将其量化成数值形式才能对其进行分析。且数据中可能会存在空行级重复的情况，如果不做处理对后续分析会造成影响，因此本文首先要对数据进行预处理。

（一） 去重、去空

对于附件 2，去除空行的问题描述文本完全一致的样本，去除后数据为 9209 条。

（二） 删除除字母、数字、汉字以外的所有符号。

删除后如图 3.1 所示：

	一级标签	留言详情	一级标签_id	clean_留言详情
375	城乡建设	\n\t\t\t\t\n\t\t\t\t您好！今天公积金中心突然颁布新政，异地公积金...	0	您好今天公积金中心突然颁布新政异地公积金贷款就不予办理那交了首付但是目前仍然在排队网签而来网...
5039	教育文体	\n \n G4县文物局，旅游局，都干了些什么。一个文化古城，不好好开发，只...	3	G4县文物局旅游局都干了些什么一个文化古城不好好开发只拿工资不干事以前西兰山乡政府内十几个唐...
7036	劳动和社会保障	\n \n 2020年度西地省城乡居民医保缴费啦！标准统一为250元/人，为...	4	2020年度西地省城乡居民医保缴费啦标准统一为250元人为什么不可以在网上缴费
8453	卫生计生	\n \n 领导同志：　　你好，我是J8县井坡乡的村民，我和老公都是同龄人，...	6	领导同志你好我是J8县井坡乡的村民我和老公都是同龄人按国家规定男的要满二十二周岁才可结婚女的...
8814	卫生计生	\n \n　　西地省M5市曾牧局三甲曾牧站站长邵海益违法生育二胎，小小孩已...	6	西地省M5市曾牧局三甲曾牧站站长邵海益违法生育二胎小小孩已经四五岁了但他为了掩盖事实的真相把...
987	城乡建设	\n\t\t\t\t\n\t\t\t\t　　K市K1区徐家井路4号,金元怡嘉住宅楼,...	0	K市K1区徐家井路4号金元怡嘉住宅楼2008年至2010年在没有取得法定建设手续的情况下擅自...
3938	教育文体	\n\n　　看了这篇文章有感https://baidu.com/　　作为一名教师...	3	看了这篇文章有感httpsbaiducom作为一名教师说出这样的话实在有损我们称之为老师这2...
4885	教育文体	\n \n 尊敬的领导：您好，我是一名在校大学生。2018年12月12日，一...	3	尊敬的领导您好我是一名在校大学生2018年12月12日一个自称爱学贷客服的人打来电话不仅能说...
1753	城乡建设	\n\t\t\t\t\n\t\t\t\t　　1，施工单位挂靠现象严重。人证严重不符，...	0	1施工单位挂靠现象严重人证严重不符2监理单位无证人员多人证严重不符
8062	商贸旅游	\n\t\t\t\t\n\t\t\t\t　　曾在5月21号发过投诉，至今也没有回信，...	5	曾在5月21号发过投诉至今也没有回信几个月过去房子的质量问题依旧没有解决国家扶持的棚户房就是...

图 3.1 部分留言删除除字母、数字、汉字后的结果

（三） 分词、停用词过滤

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[1]分词，对附件 3 中每一个岗位描述进行中文分词，jieba 分词用

到的算法为基于 Trie 树结构[2]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)；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。

为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words（停用词）。停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量极低，对文本标识无意义。

为了找出这些停用词，需要一些标准估计词的有效性。而高频词通常与高噪声值具有相关性。词语在各个文档中出现的频率可以作为一个噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应被看作是噪声词。

本文对留言结果中文分词后，将筛选出的停用词：的、了、是等加入停用词表，再利用停用词表过滤停用词，对留言的处理结果如图 3.2 所示：

	一级标签	留言详情	一级标签_id	clean_留言详情	cut_留言详情
0	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	0	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...	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑...
1	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	0	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏...
2	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...	0	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公...
3	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 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...
4	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 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...

图 3.2 停用词过滤后分词结果

3.2 TF-IDF 算法

在对留言信息分词后，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，把职位描述信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

词频（TF）= 某个词在文本中出现的次数 （1）

考虑到留言有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$\text{词频（TF）}=\frac{\text{某个词在文本中的出现次数}}{\text{文本的总词数}} \quad (2)$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率（IDF）}=\log\left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数}+1}\right) \quad (3)$$

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）× 逆文档频率（IDF） （4）

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本的关键词。

3.3 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，找出每个职位描述的前 5 个关键词；

（2）对每个岗位描述提取的 5 个关键词，合并成一个集合，计算每个岗位描述；

对于这个集合中词的词频，如果没有则记为 0；

（3）生成各个岗位描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF =词频（TF）×逆文档频率（IDF）　　（5）

3.4 留言内容的分类

3.4.1 基于 LSTM 的文本特征提取和分类模型设计

基于 LSTM 的文本特征提取和分类模型[3]，利用 LSTM 的特点，在考虑上文语境的情况下提炼出文本的特征向量。本文所用的 LSTM 模型将分类器与 LSTM 融合，在网络的最后一层使用 softmax 激活函数，直接将其作为文本分类器。基于 LSTM 的文本特征提取和分类模型是一个典型的端到端的学习模型，能够将特征学习和分类器的训练互相整合，避免了仅仅从词频角度进行文本特征提取的局限性。

作为一个利用 softmax 作为输出层归一化计算的分类模型，目标函数的设置常常是基于交叉熵损失函数设置的，因此基于 LSTM 的文本特征提取和分类模型的目标函数定义为：

$$L = \sum_{i=1}^{T} Y_i \log(y_i) \qquad (6)$$

其中，T表示语料库中的文本数量，Y_i表示当前文本类别的真实概率分布向量，y_i表示分类模型预测当前文本的概率分布向量，向量的维度均等于分类标签的个数。通过最小化目标函数，便能够训练得到分类模型。

模型的基本结构如图 3.3 所示：

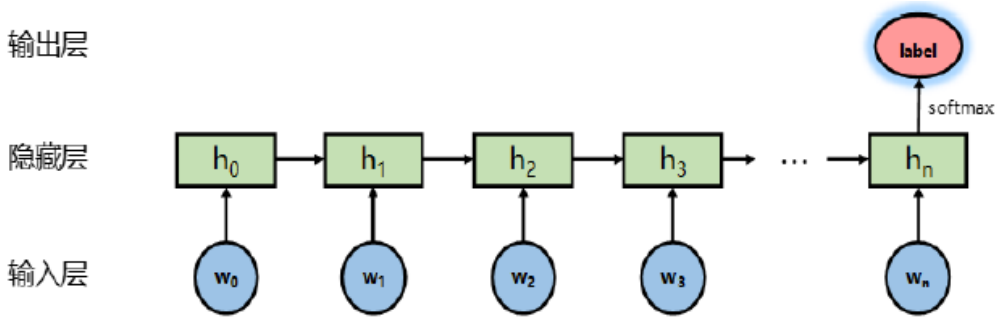


图 3.3 基于 LSTM 的文本特征提取和分类模型结构示意图

模型中 LSTM 单元内的计算过程如下：

$$h_n = \tanh\left(Uw_n + Vh_{n-1}\right) \qquad (7)$$

其中，w_n 表示t_n时刻下模型的输入数据，h_{n-1}表示t_n的前一时刻下模型隐藏层的状态。U, V 分别表示模型输入层和隐藏层的权重参数矩阵。

模型在输出层将会计算出分类标签的概率分布，其计算过程表示为：

$$y = softmax(h_n') = \frac{exp\left(h_{n(i)}'\right)}{\sum_{j=1}^{T} \exp\left(h_{n(j)}'\right)} \qquad (8)$$

$$h_n' = Wh_n \qquad (9)$$

模型采用 softmax 函数作为激活函数，T为类别标签的数量，W 表示模型输出层的权重矩阵。$h_{n(i)}'$表示向量 h_n'中第 i个分量值，向量长

度与分类标签的数量相等。softmax 函数作为输出层的激活函数能够得到一个概率分布，保证输出结果的每一个类别分量大于 0 并且和为 1。模型在t_0，t_1，t_2，t_3 ... t_n中的任一时刻都会得到一个概率分布作为输出值，然而在实际中认为只有完成对文本全部内容处理后得出的结果才有意义，因此模型只取文本输入完成之后得到的最后一项输出作为最终输出。

在训练过程中，最终的输出数据将和真实的概率分布求取交叉熵损失，表示为：

$$E = (Y, y_n) = -Ylog(y_n) \qquad (10)$$

其中，Y表示真实类别的概率分布，y_n表示模型预测出的类别的概率分布。

3.4.2 实验

本实验是基于深度学习框架 Tensorflow 上进行实现的。Tensorflow 是一种基于图计算的深度学习框架，其利用节点之间的数据流传递数据，并且在节点内完成计算。作为一个开源的深度模型框架，Tensorflow 整合了包括卷积神经网络，循环神经网络，以及长短时记忆模型在内的若干种模型。Tensorflow 框架的出现都使得深度学习模型的使用变得更加简易，方便，降低了应用深度学习模型的难度[3]。

评价指标：

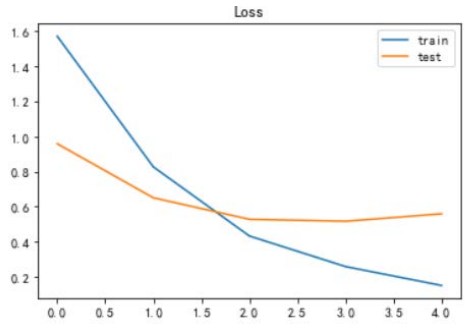
本文使用 F-Score对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \qquad (11)$$

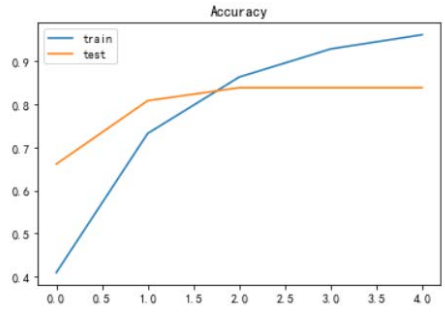
实验结果与分析：

本文 LSTM 模型结构的超参数设置为：LSTM 单元中隐藏层节点为 256，batch size 为 64，所有样本共进行 5 轮循环训练。

本文在数据加上进行了实验，结果如下图所示：



a）分类结果的 Loss 值



b）分类结果的准确率

预测结果如图 3.4 所示：

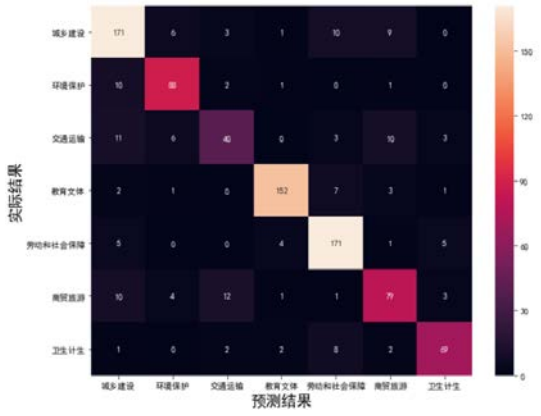


图 3.4 预测结果

4 问题一分析方法与过程

4.1 数据预处理

4.1.1 数据描述

通过观察所给数据，可以发现数据量比较大，且附件 3 中的字段大多为文本格式，需要对其进行信息提取才能对其进行分析。而附件 3：留言详情一列中有大量空行以及重复的情况，如果不做处理会对后续分析造成影响，并且描述文本信息存在大量噪声特征，如果把这些数据也引入进行分词、词频统计乃至文本聚类等，则必然会对聚类结果的质量造成很大的影响，于是本文首先要对数据进行预处理。

中文文本聚类相对于英文文本聚类来说，句子结构较为复杂，需要进行一系列的文本处理过程之后才可以变为计算机可理解的语言，才能进行下一步的分类操作[4]。文本处理过程是后续计算机自动分类的基础，可以将中文文字转化为计算机可识别的代码，是聚类计算的基础步骤。

4.1.2 文本预处理

我们把这些文本数据的预处理分为三个部分：

（一）提取信息

利用 python 中的 pandas 包对附件 3 中留言详情一栏的内容进行提取，运用 dropna 函数去除行首的数字和空格和冗杂描述的文本，最后将提取信息转化为 text 文件。提取信息如图所示：

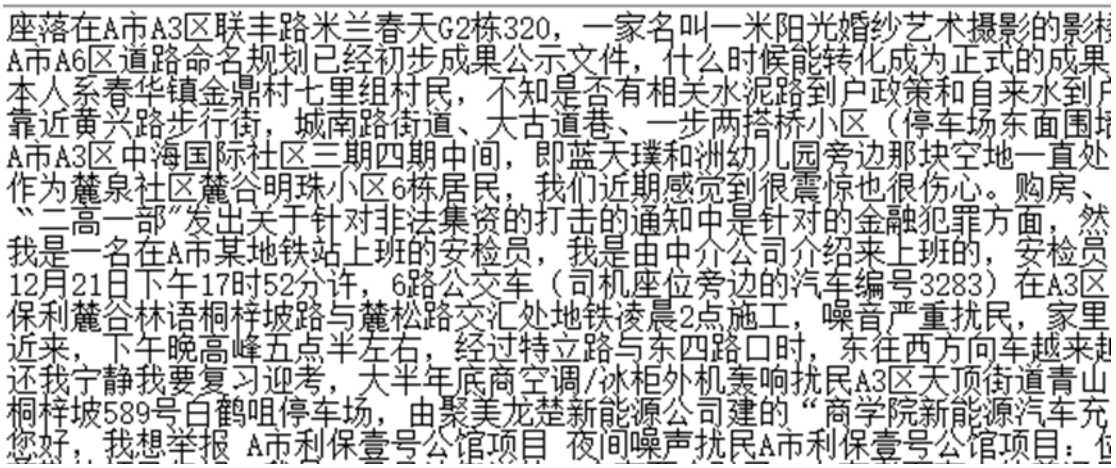


图 4 信息提取结果

（二）中文分词

中文分词操作是是文本预处理过程中必不可少的环节，是中文文本分类研究的基础。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[5]，对提取的留言描述进行中文分词，jieba 分词用到的算法：

- 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)
- 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。

（三）停用词过滤

为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words（停用词）。停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。

为了找出这些停用词，需要一些标准估计词的有效性。而高频词通常与高噪声值具有相关性。词语在各个文档中出现的频率可以作为一个噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应被看作是噪声词。因此用以下指标衡量词语的有效性：

a）词频（TF）

TF 是一种简单的评估函数，其值为训练集合中此单词发生的词频数。TF 评估函数的理论假设是当一个词在大量出现时，通常被认为是噪声词。

b）文档频数（DF）

DF 同样是一种简单的评估函数，其值为训练集合中包含此单词的文本数。DF 评估函数的理论假设是当一个词在大量文档中出现时，这个词通常被认为是噪声词。

本文选用 DF 方法筛选出如下停用词：我，有，的，了，是等。将筛选出的停用词加入停用词表，再利用停用词表过滤停用词，将分词结果与停用词表中的词语进行匹配，若匹配成功，则进行删除处理。筛选的停用词数据集结果示例如图：

图 5 停用词数据集

4.2 文本聚类

4.2.1 文本相似度计算

相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要研究文本个体将的差异大小，也就是需要对文本信息进行相似度计算，将根据相似特性的信息进行归类。目前相似度计算方法分为距离度量和相似度度量。本文采用的是基于距离度量的欧几里得距离计算文本间差异。

欧氏距离[6](Eucliden Distance)：

令i = $(x_1, x_2, \ldots, x_p)$和j = $(y_1, y_2, \ldots, y_p)$是两个被 p 个数值属性标记的对象，则对象i和j之间的欧式距离定义为：

$$\mathrm{dis(i,j)} = \sqrt{\sum_{k=1}^{p}(x_k - y_k)^2}$$

4.2.2 文本聚类

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，将有着各自特征的文本进行分类，使用相似度计算将具有相同属性或者相似属性的文本聚类在一起。这样就可以通过文本聚类的方法根据岗

位职责与任职要求，对不同职位进行分类。通过聚类方法，求职者可以结合自身状况更加快捷地获取相关信息资源。本文运用三种聚类方法比较获得最优结果。

4.2.2.1K-means 聚类原理

K-means 算法[7]是很典型的基于划分的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似性就越大。

K-means 算法的基本思想是：以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

假设要把样本集分为 k 个类别，算法描述如下：

1）适当选择 k 个类的初始中心；

2）在第 k 次迭代中，对任意一个样本，求其到 k 个中心的距离，将该样本归到距离最短的中心所在的类；

3）利用均值等方法更新该类的中心值；

4）对于所有的 k 个聚类中心，如果利用 2)-3)的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

4.2.2.2K-means 聚类结果

```
D:\Anaconda3\python.exe E:/text_clustering-master/DBSCAN/DbscanClustering.py
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\YUFENG~1\AppData\Local\Temp\jieba.cache
Loading model cost 0.889 seconds.
Prefix dict has been built successfully.
{0: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14], -1: [15, 16, 17, 18, 19, 20, 21], 1: [22, 23, 24]}

Process finished with exit code 0
```

图 6 聚类结果图

4.2.2.3DBSCAN 聚类原理

DBSCAN 是一种基于密度的聚类算法[8]，这类密度聚类算法一般假定类别可以通过样本分布的紧密程度决定。同一类别的样本，他们之间的紧密相连的，也就是说，在该类别任意样本周围不远处一定有同类别的样本存在。

通过将紧密相连的样本划为一类，这样就得到了一个聚类类别。通过将所有各组紧密相连的样本划为各个不同的类别，则我们就得到了最终的所有聚类类别结果。

具体过程如下：

1）检测数据库中尚未检查过的对象p，如果p为被处理(归为某个簇或者标记为噪声)，则检查其邻域，若包含的对象数不小于 minPts，建立新簇 C，将其中的所有点加入候选集N；

2）对候选集N 中所有尚未被处理的对象q，检查其邻域，若至少包含 minPts 个对象，则将这些对象加入N；如果q 未归入任何一个簇，则将q 加入C；

3）重复步骤 2)，继续检查 N 中未处理的对象，当前候选集N为空；

4）重复步骤 1)-3)，直到所有对象都归入了某个簇或标记为噪声。

4.2.2.4 聚类结果

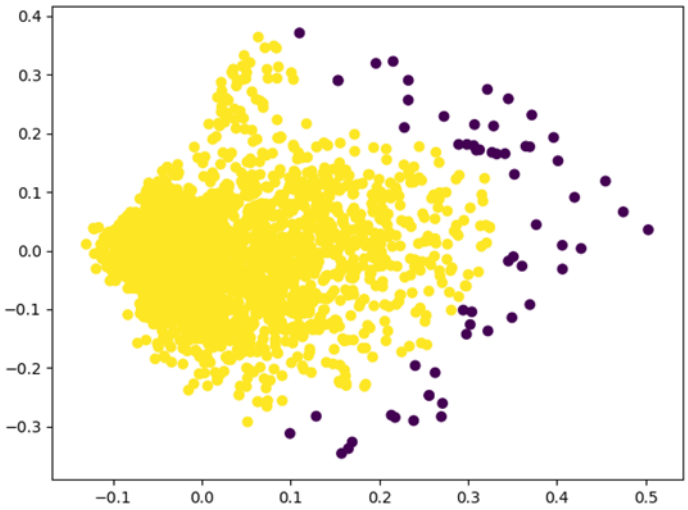


图 7 DBSCAN 聚类结果

4.2.2.5 LatentDirichletAllocation 聚类原理

LDA(Latent Dirichlet Allocation)主题模型是一种文档生成模型[9]，也是一种非监督机器学习技术，基于贝叶斯模型的一种方法。它认为一篇文档是有多个主题的，而每个主题又对应着不同的词。在 LDA 的理论当中一篇文档的主题构造过程如下所示，首先是以一定的概率选择文档当中的某个词，然后再在这个词下以一定的概率选出某一个主题，这样就生成了这篇文档的第一个主题。不断重复这个过程，就生成了指定的 K 个主题。

它的主要步骤：

1）构建词袋：对每篇文章进行分词处理（本文中使用的 jieba 库）。

2）进行统计词频，利用 CountVectorizer 得到所有文档中各个词

的词频向量和主题词袋构成的一个 list。

3）利用 LatentDirichletAllocation 进行 LDA 处理，设置需要分成的主题个数等参数，可以得到每篇文章属于每个主题的概率矩阵和分成每个主题中主题词的分布概率矩阵。

4.2.2.6 聚类结果

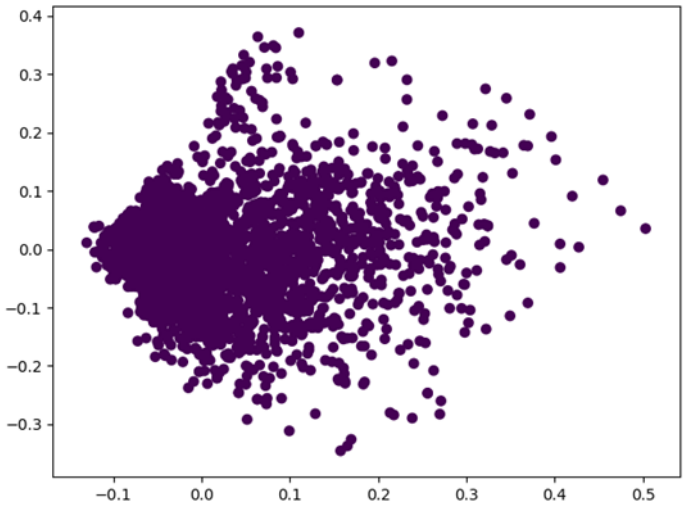


图 8 LDA 聚类结果

热门地域定义

“热门需求”是一个模糊的概念，所以我们首先要定义何为热门需求，本文通过分析，认为热门需求具备以下几个特征：

i. 在留言详情中出现频率较高

ii. 留言所属情况比较严重

iii. 所反映的问题属于当下的热门问题的

针对上述分析，结合所给数据，比较三种聚类方法的结果，最终采取 K-means 均值聚类方法，选取了排名前十类的地域。

5 总结

总结本次比赛，我们用传统的 LSTM 模型对论文中的留言分类进行预测；通过三种聚类方法对留言的地点进行聚类；分析了留言热度，定义热度评价指标，并对留言的答复意见制定了方案。虽然最后得到的结果不太好，也有一定的误差，但基本可以构造留言分类的模型。通过这次比赛认识到了在文本挖掘方面的不足，日后将对文本挖掘进一步探讨。

参考文献

[1]陶伟.警务应用中基于双向最大匹配法的中文分词算法实现[J]. 电子技术与软件工程, 2016(4).
[2]刘志.基于用户兴趣的协同过滤算法的广告推荐研究[D]. 昆明理工大学, 2014.
[3]王恰.基于 Attention Bi-LSTM 的文本分类方法研究[D]. 华南理工大学, 2018.
[4]崔嘉乐,姜明洋,裴志利,卢奕南.基于深度学习的文本挖掘研究[J].内蒙古民族大学学报(自然科学版),2016,31(05):403-407.
[5]张振亚,王进,程红梅,王煦法.基于余弦相似度的文本空间索引方法研究[J].计算机科学,2005(09):160-163.
[6]张跃,李葆青,胡玲芳,孟丽.基于K-Mean文本聚类的研究[J].中国教育技术装备,2014(18):50-52.
[7]郭艳婕,杨明,侯宇超,孟铭.基于相似性度量的改进DBSCAN算法[J].数学的实践与认识,2020,50(06):164-170.
[8]张涛,马海群.一种基于LDA主题模型的政策文本聚类方法研究[J].数据分析与知识发现,2018,2(09):59-65.