“智慧政务”中的文本挖掘应用

摘要

随着大数据时代的来临、文本技术的发展使得政务也变得越来越智能化。本文旨在基于赛题所提供的一系列数据，通过文本特征提取、数据分析与数据深入挖掘，对群众留言详情建立分类模型、挖掘热点问题、评价回复内容。

针对任务一，建立了基于LSVM的分类模型，通过TF－IDF（term frequency－inverse document frequency）算法提取文本特征，然后利用LSVM算法建立分类模型，最后利用F-Score进行模型的评价。通过评价本实验模型的精确度为0.913672，准确率为0.913616,，召回率为0.913672， f1值为0.913326。

针对任务二，首先采用jieba词性标注，然后查找文本中的人物地点，然后构建了DBSCN聚类器，对附件3中问题描述指标进行聚类，通过DBSCN聚类将附件3中的“留言主题”分成了45类。为了挖掘出数据中的热点问题，本文提出了热点问题挖掘算法，将点赞数、反对数、同类问题留言的数量、同类问题的时间跨度作为评价指标，通过权重计算得到热点指数，热点指数越大则问题的热度越高。通过该热点问题挖掘算法，本文挖掘了热点指数排名前五的热点问题。

针对任务三，本文从相似性、完整性、可解释性三个方面对答复意见进行评价。利用word2vec计算答复意见和留言详情两类文本的相似性，通过相似性计算能够知道两个文本之间的内容契合度。通过答复格式的划分，将答复意见划分成称呼和问候、重复声明留言中所表达的问题、对问题的解答、解答中涉及的法律法规和文件、如需进一步处理的联系方式五个部分，从而评价问题的完整性。

【关键词】LSVM；TF－IDF；F－Score；DBSCN聚类；热点问题挖掘算法

Abstract

With the advent of the era of big data and the development of text technology, government affairs are becoming more and more intelligent. Based on a series of data provided by the contest questions, this paper aims to establish a classification model for the details of the public message, mine hot issues and evaluate the reply content through text feature extraction, data analysis and in-depth data mining.

Aiming at task one, a classification model based on LSVM is established. The text features are extracted by TF-IDF (term frequency – inverse document frequency) algorithm, then the classification model is established by LSVM algorithm, and finally the model is evaluated by F-score. Through the evaluation, the accuracy of the experimental model is 0.913672, the accuracy is 0.913616, the recall is 0.913672, and the F1 value is 0.913326.

For task two, we first use the part of speech tagging of Jieba, then find the location of the characters in the text, and then build a dbscn cluster to cluster the problem description indicators in Annex 3. Through the dbscn cluster, the "message subject" in Annex 3 is divided into 45 categories. In order to mine the hot issues in the data, this paper proposes a hot issues mining algorithm, which takes the number of likes, anti logarithm, the number of messages of the same kind of problems, and the time span of the same kind of problems as the evaluation indexes, and obtains the hot issues index through weight calculation. The larger the hot issues index, the higher the heat of the problem. Through this algorithm, we mine the top five hot issues in the hot index.

For task three, this paper evaluates the response from three aspects: similarity, integrity and interpretability. Using word2vec to calculate the similarity between the two kinds of texts, we can know the content agreement between the two texts by similarity calculation. Through the division of reply format, the reply opinions are divided into five parts: address and greetings, questions expressed in repeated statement messages, answers to questions, laws, regulations and documents involved in the answers, and contact information for further processing, so as to evaluate the integrity of the questions.

【Key words】： LSVM; TF-IDF; F-score; DBSCN clustering; hotspot mining algorithm

目录

一、问题分析.....1
1.1 任务一.....1
1.2 任务二.....1
1.3 任务三.....1
二、数据准备.....3
2.1 数据清洗.....3
2.2 预处理.....3
2.2.1 分词.....4
2.2.2 去停用词.....4
2.2.3 格式化时间.....5
三、基于 LSVM 的群众留言分类模型.....6
3.1 文本特性提取.....6
3.2 基于 LSVM 的模型训练.....7
3.3 实验结果与分析.....9
3.3.1 F-Score 评价方法.....10
3.3.2 实验结果分析.....11
四、基于 DBSCN 聚类算法处理热点问题以及 JIEBA 地点人物提取.....13
4.1 首先设定保留词.....13
4.2 jieba 词性标注.....13
4.3 DBSCN 聚类算法.....14
4.4 热点算法的实现.....15
4.5 分析实验结果.....16
五、答复意见评价.....17
5.1 相关性.....17
5.1.1 预处理.....17
5.1.2 计算相似度.....18
5.1.3 Word2vec.....18
5.2 完整性.....18
5.2.1 答复格式划分.....19
5.2.1 划分答复.....20
参考文献.....22
附录.....23

一、问题分析

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。

赛题给出了四个附件数据，附件 1 提供了 517 条内容分类三级标签体系，附件 2 包括城乡建设、环境保护、交通运输、商贸旅游、教育文体、劳动和社会保障、卫生计生七个一级标签 9210 条群众留言数据，包括留言详情、一级标签等 6 个指标。附件 3 包含 4326 条群众留言数据，包含留言主题、留言详情、点赞数、反对数等 7 个指标。附件 4 包括 2816 条数据，包括留言时间、留言详情、答复意见等 7 个指标。

1.1 任务一

对于任务一要求的建立群众留言一级标签分类模型，其属于典型的文本多分类问题。在实际的分类中，要注意语义带来的词语交叉、数据不平衡带来的影响以及长文本的无意义表达。

1.2 任务二

任务二面向的问题是针对特定一段时间集中爆发的问题，多人反映同一问题，也就是热点问题的挖掘。对于这样一个热点问题我们给它的定义是某一时段内群众集中反映的某一问题可称为热点问题。如果可以及时发现热点问题，可以帮助政府部门进行有针对性地处理，提升服务效率。

1.3 任务三

开放性问题，探讨留言答复的质量评价方案。拟从三个方面对质量进行探讨。

一是答复的相关性，即答复中的回答内容是否针对留言中的问题，考虑采用word2vec 计算答复文本和留言文本间的相似性，并将其作为主要相关性的主要衡量指标。

二是答复的完整性，即答复是否满足某种规范，例如答复是否包括了对问题的解答，解答的依据，进一步的处理方式等等。如果答复包含了规范里要求的所有部分，则认为完整性较好，考虑使用基于规则的处理方法。

三是答复的可解释性，可解释性指答复内容中内容的相关解释。

二、数据准备

2.1 数据清洗

数据清洗主要包括两个部分：去掉文本标记符号和去掉重复数据。在文本数据中往往还有一些抬头信息、网页 HTML 标签信息、标点符号信息、阿拉伯数字等信息[1]。这类数据不仅会影响分类的结果，而且会给分类器造成负担，故这类符号在实验前必须要去除。赛题给出的附件 2 中的数据分布情况如图 1 所示：

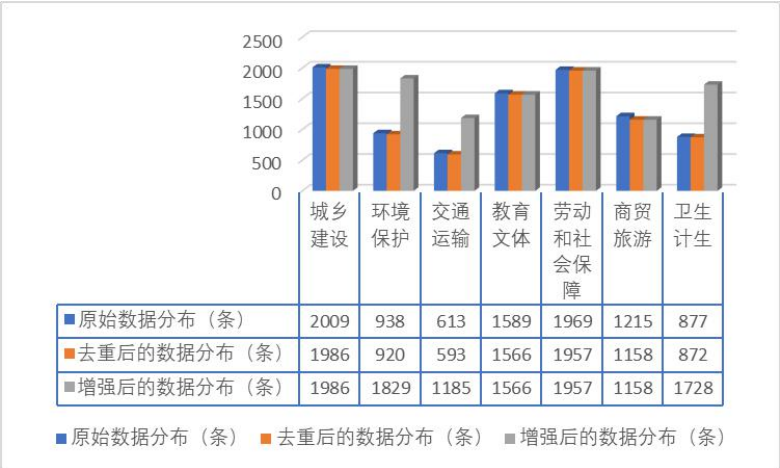


图 1 数据分布情况图

从分布结果来看，交通运输、环境保护、卫生计生这三类数据量远远少于其他类。因此本实验对附件 2 中的数据进行了增强，将交通运输、环境保护、卫生计生这三类数据中的“留言主题”新增为“留言详情”。增强后的数据分布如图 1 所示。附件 2 增强后的数据文件参考 clean. xlsx。

2.2 预处理

在文本数据中常常会有一些没有意义的词语，比如：“的”，“一个”，“这些”之类的噪声。为了避免这些噪声在文本分类时影响准确率和效率，因此需要对文本数据进行预处理。

中文文本预处理过程主要包括去除文本标记符、中文分词、去除停用词。

2.2.1 分词

由于中文构词方法和英文的不同，英文词与词之间直接用空格分离。但是中文文本字词之间没有明确的分割界限，都是连续的字词，句子和句子之间用标点符号隔开。分词结果的好坏直接影响分类模型的准确率。因此在文本分类中，分词时十分重要的技术。

本实验采用 jieba 分词作为本次实验的分词工具。Jieba 提供了三种分词模式：

1、精确模式，试图将句子最精确地切开，适合文本分析；

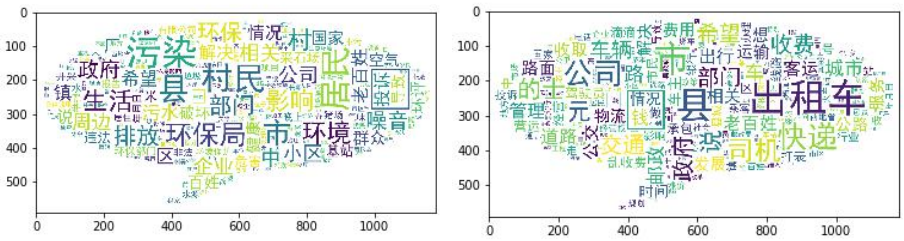
2、全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；

3、搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

本实验采用的是精确模式，对文本数据进行分词。

2.2.2 去停用词

停用词主要是去除一些对文本分类没有区分能力和意义的词，如叹词、语气助词、代词和副词等等。这些词不能为文本分类提供帮助，（如果没有去停用词）反而会增加文本的复杂度，消耗文本分类的时间。因此，去停用词是文本预处理一个较为重要的一步，同时也会影响文本分类的效率和准确率[2]。本实验的停用词表是根据文本特点，并在哈工大停用词表的基础上新建的。去停用词后，以词云的形式展示了语料的一些关键信息，图 X 给出了环境保护类别和交通运输类别的词云图，其他五类的词云图参见附录 1 ，附件 2 去除停用词后的数据文件为 clean__data. xlsx。



（a）环境保护类词云图　　（b）交通运输类词云图

图 2 语料关键信息词云图

2.2.3 格式化时间

附件 3 里面时间格式不一致，时间格式有 String 格式的%Y/%m/%d %H:%M:%S 还有 Datatime 格式%Y/%m/%d %H:%M:%S。这两个格式既无法直接相互比较，也没有办法自己和自己比较，所以要格式化时间，统一处理为时间戳。

时间戳既可以由两者转换，也可以相互计算，float 型的格式也方便后面考虑热点判断。

三、基于 LSVM 的群众留言分类模型

基于 LSVM 的群众留言分类模型具体的实验流程如下图所示。数据预处理过程参见第二章数据准备，本章节不再赘述。本章重点介绍文本的特征提取、LSVM 的模型训练以及模型评价。

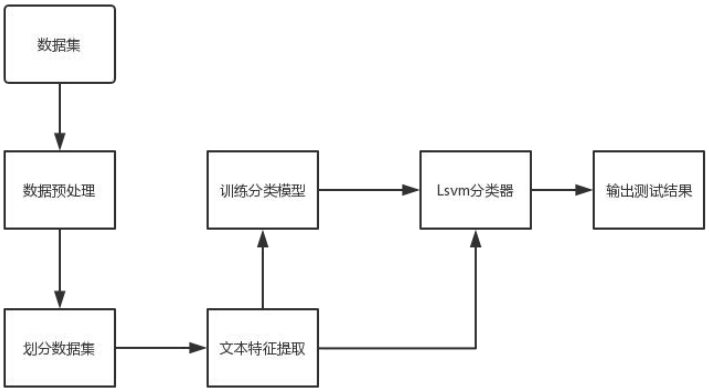


图 3 基于 LSVM 的群众留言分类模型具体的实验流程

3.1 文本特性提取

本实验文本特征提取方法用的是 TF-IDF（term frequency－inverse document frequency）。TF-IDF（term frequency－inverse document frequency）算法是由 Salton 等人在 1983 年提出的术语加权方案，在 2002 年，它被 Sebastiani 等人应用于文本挖掘任务[3]。TF-IDF 由两部分组成，词频(Term Frequency)和逆文本频率指数(Inverse Document Frequency)，通过统计学的方法来计算术语对于文本内容的重要性[4]。作为信息检索与数据挖掘的常用加权技术，TF-IDF 常被用来评估一个词语对于文本库中单个文件的重要程度[5]。词语在文本中出现的次数越多，词语重要性越强，但是随着在文本库其他文本中出现频率增加，词语重要性逐渐降低[6]。

对于上述两个问题，提出了 TF-IDF 的计算方法：

其中 TF（词频）的计算方法如下：

$$tf_{i,j} = \frac{count(n_{k,i})}{\sum_j count(n_{k,j})} \qquad (1)$$

式中：$n_{k,i}$ 是第 k 类文本中词表的第 i 个词语；$count(n_{k,i})$ 表示统计第 k 类文本中词表的第 i 个词语在第 k 类文本中的词频；$\sum_i count(n_{k,j})$ 则表示 k 类文本中词表的所有词语词频总和。

IDF（逆向文件频率）的计算方法如下：

$$idf_i = \log \frac{|d|}{|\{j \mid t_i \in d_j\}|} \qquad (2)$$

式中：d 表示训练样本的集合；$|d|$ 表示训练样本总数；d_j 表示训练样本中第 j 个样本；t_i 是使用训练样本构建的词表中第 i 个词语。

对于每个类别的词表，计算 TF-IFD 值，计算公式如下：

$$tf\_idf = tf_{i,j} \times idf_i \qquad (3)$$

将预处理过的数据集按照 4：1 的比例划分为训练集和测试集。利用 TF-IDF 算法提取训练集中七个类别的文本特征。TF-IDF 进行特征提取后的主成分分析（PCA）图如下图所示：

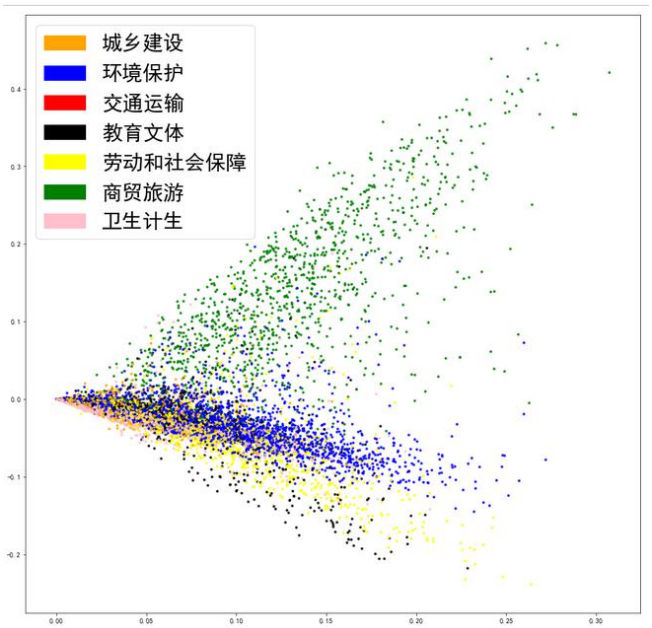


图 4　TF-IDF 进行特征提取后的主成分分析（PCA）图

3.2 基于 LSVM 的模型训练

支持向量机（support vector machines，SVM）是一种基于统计理论的分类算法，通过找到一个超平面，尽可能的将样例分开，即其决策边界是对样本求解

的最大边距的超平面[7]。如图，线性可分支持向量机：决策边界（实），间隔边界（虚），支持向量（红点）。

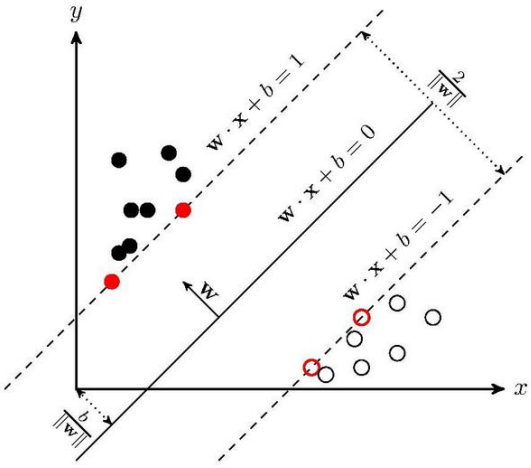


图 5 线性可分支持向量机

支持向量到超平面的距离为$\frac{1}{\|w\|}$，两个支持向量之间的距离为$\frac{2}{\|w\|}$。

目前 SVM 广泛应用于分类、回归、预测任务，而它的优势就在于对于非线性、高纬度的数据分类情况良好。当一个问题是线性可分的，优化问题只需要找到一条直线让本距离最大化位置即可。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM 的的学习算法就是求解凸二次规划的最优化算法。当处理多分类问题时，可以利用 SVM 模型将其转化成多个二分类问题进行求解。

本实验利用 LSVM 算法进行分类模型的训练，将训练样本表示为：

$$\{(x_i,x_i)\}_{i=1}^{l}, x_i \in R^n, y_i \in \{1,-1\} \qquad (4)$$

LSVM 的求解问题可转换成如下无约束线性优优化问题：

$$\min_w f(w)=\frac{1}{2}w^T w+C\sum_{i=1}^{l}\zeta(w,x_i,y_i) \qquad (5)$$

其中$\zeta(w,x_i,y_i)$表示损失函数，$C>0$表示惩罚因子，损失函数有以下的几种：

$$\zeta(w,x_i,y_i)=\begin{cases}\max(0,1-y_i w^T x_i), & L1-SVM\\ \max(0,1-y_i w^T x_i)^2, & L2-SVM\\ \log(1+\exp(-y_i w^T x_i)), & LR\end{cases} \qquad (6)$$

模型参数如表 1 所示。将训练好模型保存为 SVMModel.m。

表 1 模型参数

参数名称	备注
C	惩罚系数，用来控制损失函数的惩罚系数
Penalty	正则化参数，L1 和 L2 两种参数可选
loss	损失函数，有‘hinge’和‘squared_hinge’两种可选，前者又称 L1 损失，后者称为 L2 损失，默认是’squared_hinge’,其中 hinge 是 SVM 的标准损失，squared_hinge 是 hinge 的平方
dual	是否转化为对偶问题求解，默认是 True。
tol	残差收敛条件，默认是 0.0001，与 LR 中的一致。
multi_class	负责多分类问题中分类策略制定‘ovr’和‘crammer_singer’ 两种参数值可选，默认值是’ovr’,’ovr’的分类原则是将待分类中的某一类当作正类，其他全部归为负类，通过这样求取得到每个类别作为正类时的正确率，取正确率最高的那个类别为正类；‘crammer_singer’ 是直接针对目标函数设置多个参数值，最后进行优化，得到不同类别的参数值大小
fit_intercept	是否计算截距，与 LR 模型中的意思一致
class_weight	用来处理不平衡样本数据的，可以直接以字典的形式指定不同类别的权重，也可以使用 balanced 参数值。
verbose	是否冗余，默认是 False。
random_state	随机种子。
max_iter	最大迭代次数，默认是 1000。

3.3 实验结果与分析

本节主要讲述采用 F-Score 对本分类方法进行评价，然后通过将 LSVM 算法模型与逻辑回归、朴素贝叶斯、随机森林等传统的机器学习方法进行对比分析，从而说明我们模型的合理性与有效性。最后分析模型的最优参数调整过程以及各类参数对于模型的影响。

3.3.1 F-Score 评价方法

F-Score 是准确率（Precision）和召回率（Recall）的加权调和平均，常用于评价分类模型的好坏。利用模型进行文本预测时所有可能的情况如下表 2 所示：

表 2 模型进行文本预测

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

行表示预测的 label 值，列表示真实 label 值

TP：True Positive, 被判定为正样本，事实上也是正样本。

FP：False Positive,被判定为正样本，但事实上是负样本。

TN：True Negative, 被判定为负样本，事实上也是负样本。

FN：False Negative,被判定为负样本，但事实上是正样本。

在本实验中，数据集中有七个类别的数据，属于多分类问题。我们将其转换成多个二分类问题进行求解，用明确类和其他类表示。比如某一文本预测的结果为“城乡建设”，那么这就是一个明确类，其为正样本；如果预测结果不是“城乡建设”，那么其就属于其他类，该样本为负样本。

关于准确率（Precision）和召回率（Recall）的相关计算如下：

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \qquad (7)$$

acc 表示预测结果的精确度。

Precision（准确率）的计算公式如下：

$$p = \frac{TP}{TP + FP} \qquad (8)$$

其中 P 表示准确率。

Recall(召回率)的计算公式如下：

$$R = \frac{TP}{TP + FN} \qquad (9)$$

其中 R 表示召回率。

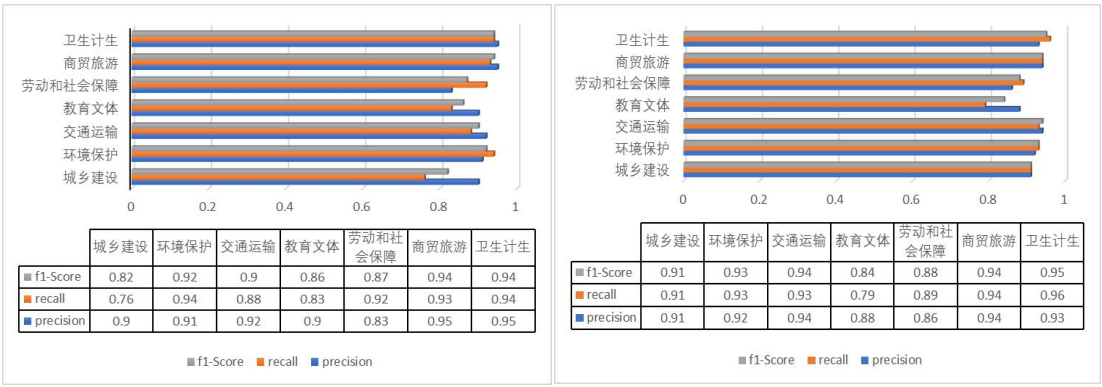
F_1 的计算公式为：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \qquad (10)$$

其中 P_i 表示为第 i 类的查准率，R_i 为第 i 类的查全率。通过 F_1 值评价模型的好坏。

3.3.2 实验结果分析

（1）数据增强前和增强后实验结果对比。在文本预处理阶段，我们发现数据存在很大不平衡，在未对数据进行增强时，利用 LSVM 算法训练出的模型进行 F-Score 评估后的结果为：acc=0.901712，precision=0.903491，recall=0.901712，f1=0.901481。数据增强后，模型评估结果为：acc = 0.913672，precision = 0.913616，recall = 0.913672，f1 = 0.913326。与数据未增强之前的结果相比，acc、precision、recall、f1 的数值都有所增加。其中准确度 acc 提高了 1.27%，f1 值提高了 1.31%。



	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生
f1-Score	0.82	0.92	0.9	0.86	0.87	0.94	0.94
recall	0.76	0.94	0.88	0.83	0.92	0.93	0.94
precision	0.9	0.91	0.92	0.9	0.83	0.95	0.95

	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生
f1-Score	0.91	0.93	0.94	0.84	0.88	0.94	0.95
recall	0.91	0.93	0.93	0.79	0.89	0.94	0.96
precision	0.91	0.92	0.94	0.88	0.86	0.94	0.93

（a）数据增强前的 LSVM 分类器效果　　（b）数据增强后的 LSVM 分类器效果

图 6　数据增强前后的分类效果对比

（2）LSVM 模型（数据增强后的模型）与朴素贝叶斯、BP 神经网络、逻辑回归模实验结果对比。通过构建以上三种传统机器学习算法模型，同样利用 F-Score 对这三个模型进行评估。各评价指标如表所示。本实验模型在精确度 acc、f1 值大于其他三个模型。其中本实验模型的 f1 值相比朴素贝叶斯模型的 f1 值提高了 3.57%，与逻辑回归模型的 f1 值相比提高了 1.07%；并且稍大于 BP 神经网络模型。

	acc	precision	recall	f1
LSVM	0.913672	0.913356	0.913672	0.913326
朴素贝叶斯模型	0.882121	0.883208	0.882121	0.881873
BP神经网络模型	0.911919	0.911705	0.911919	0.911482
逻辑回归模型	0.904032	0.90494	0.904032	0.903672

图 7　LSVM、朴素贝叶斯、BP 神经网络、逻辑回归模型效果对比

（3）模型参数调整过程。在 4.2 节中已经介绍了 LSVM 有关的参数，详情请参看表 1 。

通过调整 Penalty、loss、dual、multi_class、fit_intercept、class_weight（通过实验发现其他参数的改变不影响实验结果）的值可以得到不同的 f1 值，f1 的变化情况如图所示：

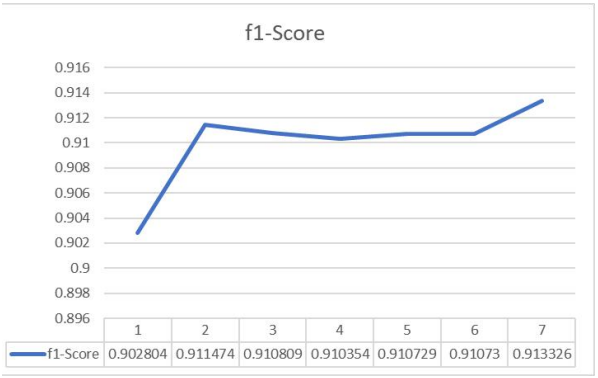


图 8 调整参数 f1-Score 变化曲线

当 f1 取到最大值，即 f1=0.913326 时，模型最优。模型最优时各个参数的值为：penalty='l2', loss='squared_hinge',dual=True,tol=0.0001,C=1.0, multi_class='ovr',fit_intercept=True, intercept_scaling=1, class_weight='balanced', verbose=0, random_state=None, max_iter=1000。

实验代码请参看附件：问题一：留言分类.html。

四、基于 DBSCN 聚类算法处理热点问题以及 JIEBA 地点人物提取

为了挖掘附件 3 中的热点问题，本实验采用 DBSCN 聚类算法先对问题进行聚类，然后通过制定热点评价指标进行热点问题挖掘，具体的实验过程如图 9 所示。

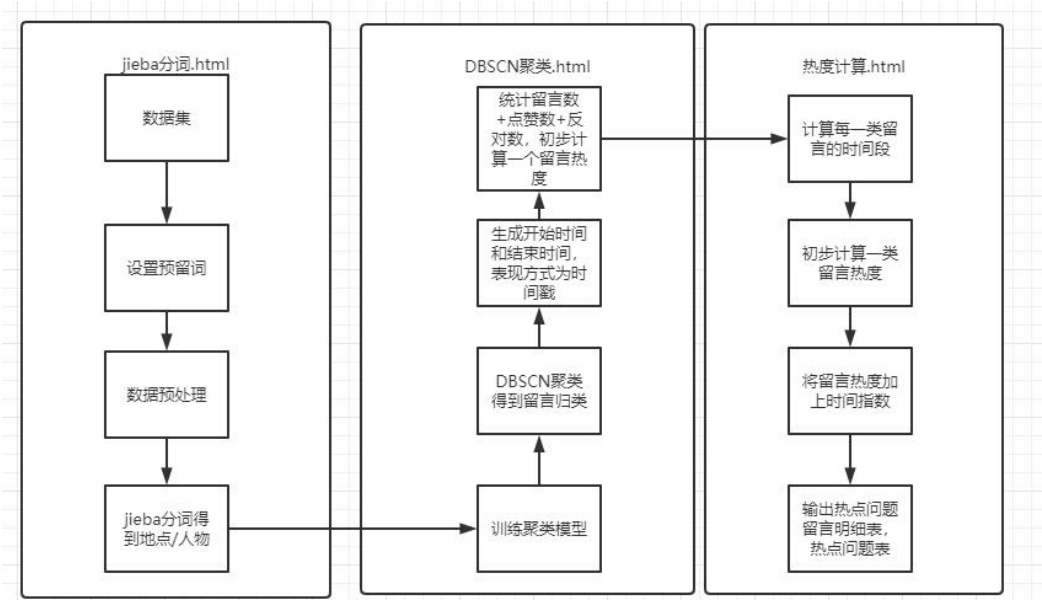


图 9 基于 DBSCN 聚类算法处理热点问题以及 JIEBA 地点人物提取流程图

4.1 首先设定保留词

因为 jieba 处理器虽然分词速度快，但是对字母和汉字的组合分词效果极差，比如：A 市，A2 市等，基本上是分不出来的，这里需要将这些字母和数字组合的城市名字提前添加到分词处理器里面。

4.2 jieba 词性标注

Jieba 一直是一个常用分词处理器，里面内置了词性标注，对比哈工大的 ltp，还是另外一个开源的 hanlp，效果上面区别不大，但是速度上面提升可能有两倍以上。经过观察，这里将地名标注为 nz、nr、ns 等词语抽取出来。对应上地点和人物的取样。主要操作代码在 jieba 分词.html。

4.3 DBSCN 聚类算法

在最初的实验过程中，本实验主要采用 k-means 和 DBSCN 两种聚类方法，通过对比两种聚类算法的结果发现，k-means 聚类其代码量比较多，同时计算 k 的算法又比较繁琐，不易理解，在很多方面不如 DBSCN 聚类，在实验的中 DBSCN 分类结果比 k-means 更优。

DBSCAN 的聚类定义很简单：由密度可达关系导出的最大密度相连的样本集合，即为我们最终聚类的一个类别，或者说一个簇[8]。

DBSCAN 的簇里面可以有多个核心对象。通过领域范围，让自己和其它簇密度可达，而最终的集合生成为一个 DBSCAN 聚类簇。

簇样本集合是怎么形成的？DBSCAN 使用的方法很简单，它任意选择一个没有类别的核心对象作为种子，然后找到所有这个核心对象能够密度可达的样本集合，即为一个聚类簇。然后继续选择另一个没有类别的核心对象去寻找密度可达的样本集合，这样就得到另一个聚类簇。不断运行直到所有对象都有自己属于的集合为止。

DBSCAN 主要有两个参数一个是范围，一个是范围当中最少点的个数。通过不断输入，发现范围设置越小越精细，当范围设置小于 0.9 时，聚类失败，因此本实验将范围设置为 0.9。聚类效果如图 1 0 所示。DBSCN 将不同类从 0～ n 排列，相同的语句属于同一个类，如图 9 给出的结果都属于第 44 类，如果是孤立点，那么类别统一设定为-1。设置结束后写入到 clean2.xlsx。具体代码在 DBSCN 聚类.html。

	问题ID	number	ID	user	thenm	time	content	opposition	favor	地点/人群	热点指数	结束时间	开始时间	时间范围
4224	44	4326	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	\n\t\t\t\t\t\n\t\t\t\t\t　书记您好，我是来自西地省经济学院体育学院...	9	0	A市	10	1.556444e+09	1.496914e+09	2017-06-08至2019-04-28
4223	44	4325	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	\n\t\t\t\t\t\t\t\t\t\tA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...	3	0	A市	4	1.556444e+09	1.496914e+09	2017-06-08至2019-04-28
4222	44	4324	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	\n\t\t\t\t\t\t\t\t\t\t\t\t各位领导干部大家好，我是A市经济学院的一名学生，...	0	0	A市	1	1.556444e+09	1.496914e+09	2017-06-08至2019-04-28
1891	44	1913	233759	A909118	A市涉外经济学院强制学生实习	2019/04/28 17:32:51	\n\t\t\t\t\t\n\t\t\t\t\t\t\t\t\t\t各位领导干部大家好，我是A市涉外经济学院的一...	0	0	A市	1	1.556444e+09	1.496914e+09	2017-06-08至2019-04-28
668	43	673	204531	A00082895	地绿A市城际空间站的建筑质量太差了	2019/9/24 8:02:57	\n\t\t\t\t\t\n\t\t\t\t\t\t\t\t\t\t位于A5区高铁新城的地绿A市城际空间站，号称...	0	1	A市	2	1.573260e+09	1.569209e+09	2019-09-23至2019-11-09

图 10　DBSCN 聚类效果

最后结果分出来 0 到 44，45 个类，其它为孤立点（不超过四个人提出相同问题）默认设定为-1，孤立点，如果特别分类没有意义，带-1 可以更好表达它的意思。

4.4 热点算法的实现

这里引入一个概念——时间戳，可以将时间变成一个 float 类型的变量，这样方便进行比较和计算。

为了让结果可观，将表格相同类别先进行同意和反对的计算，这里设定点赞和反对权重都是 0.2，一条留言权重设为 50，时间设定为现在时间到留言开始时间的差除以 100000 相当于每隔一天热度增加二，缺少问题是否完成，所以默认所有问题都没有完成，时间的差距越大，问题更加需要处理。这样计算问题热度并将它们进行排序，具体公式如下：

$$q_i = s_i \times 0.2 + o_i \times 0.2 + c_i \times 50 + \frac{t_{i,e} - t_{i,b}}{100000} \tag{11}$$

q_i 表示第 i 类问题热点指数，s_i 表示第 i 类问题点赞数，o_i 表示第 i 类问题的反对数，c_i 表示第 i 问题的总留言数，$t_{i,e}$ 表示第 i 类问题当前时间戳，$t_{i,b}$ 表示第 i 问题第一个留言时间戳。

具体代码在热度计算.html 文件中，这里将数据处理并返回要求的表格，处理有两个比较复杂的函数，时间段计算器（也就是将相同类的时间归到一起，找出最大最小的时间），第二个是相同类热点计算器（因为相同类热点相互有关系，这样必须全部处理完，达到结果才可以知道这一个类热点指数到底是多少）。

经过热点问题挖掘后，本实验挖掘出的热度排名前五的热点问题如图 11 所示，热点问题对应的问题留言明细如图 12 所示。

	热度排名	问题ID	热点指数	时间范围	地点/人群	问题描述
176	1	1	5760.398372	2019-01-03至2020-01-07	A市	请增加A市276路运行线路或增加发车频率
134	2	5	1616.270202	2019-07-07至2019-08-31	武广新城伊滨河苑	武广新城伊景园滨河苑违法捆绑销售车位,求解决
282	3	0	1211.074852	2019-11-13至2020-01-06	A市A2区新城	投诉小区附近搅拌站噪音扰民
2619	4	-1	696.770622	2019-08-19至2019-08-19	A市A5区汇金路K9	A市A5区汇金路五矿万境K9县存在一系列问题
167	5	2	871.445122	2019-01-06至2019-09-12	A3区西湖	反映A3区西湖街道茶场村拆迁问题

图 11　热点问题表

要注意把空白的主题替换成相同类的主题。

	问题ID	number	ID	user	问题描述	time	content	opposition	favor
243	1	193	192774	A00039164	A3区第十幼儿园数百名学龄前儿童即将无学可上	2019-06-13 12:20:43	\n\t\t\t\t\t\n\t\t\t\t\tA3区第十幼儿园，隶属于A3区教育集团，坐落...	0	2
277	1	208	193210	A00037344	A4区六号线楚雅路地铁施工严重扰民	2019-05-24 15:13:06	\n\t\t\t\t\t\n\t\t\t\t\tA市A4区楚雅路文昌阁地铁站最近一周整天从早...	0	1
279	1	1287	218575	A00063852	A3区中海业主的小孩都没办法上中海幼儿园	2019-09-06 10:48:53	\n\t\t\t\t\t\n\t\t\t\t\t中海国际社区三期内的幼儿园，业主孩子无法入学...	0	0
280	1	66	189856	A00073717	）	2019-07-03 11:53:35	\n\t\t\t\t\t\n\t\t\t\t\tA市保利麓谷林语小区，是A市的大型小区，居住...	0	1
250	1	908	210000	A00031618	请加快创建A市食品安全城市建设步伐	2019-01-10 14:32:19	\n\t\t\t\t\t\n\t\t\t\t\t地处时代倾城小区二期门面商铺常年违法销售不合...	0	0

图 12　热点问题留言明细表

4.5 分析实验结果

这里除了聚类算法比较慢，其他处理速度都是比较快的，本质上是 jieba 处理器比其他词性处理器速度上面的优势，然后是热度算法没有使用太复杂的算法，这样得到结果的速度大大提高。而聚类算法是任务二的核心，DBSCNk-means 需要指定聚类簇数 k，并且初始聚类中心对聚类影响很大。k-means 把任何点都归到了某一个类，对异常点比较敏感。DBSCAN 能剔除噪声，需要指定邻域距离阈值 eps 和样本个数阈值 MinPts，可以自动确定簇个数。

表 3 不同词性标注分类器速度表

分类器	处理相同一句话词性标注 时间单位：s
jieba	0.01799
thulac	8.15695
HanLP	2.44310
Ltp	0.09294

相比传统聚类算法 K-means 更加适合这个任务。具体体现如下：

1.传统的 K-means 算法需要指定多少簇，也就是 k，并且初始聚类中心对 k-means 聚类影响非常大，这样导致 k-means 将任何点都归到一个类，使得它对异常点比较敏感，而任务二有较多的孤立点，使用 k-means 可能比较慢。

2.DBSCAN 能去除噪声，它需要指定邻域距离阈值 eps 和样本个数阈值 MinPts，可以自动确定簇个数。

3.K-means 针对的是所有对象，而 DBSCAN 丢弃被认定是噪声的对象。

4.K-means 很难处理大小不同或者非圆形的簇。DBSCAN 可以处理这些簇，并且它可以不太受噪声和离群点的影响。

五、答复意见评价

针对相关部门对留言的答复意见，拟从相关性，完整性，可解释性三个角度对答复意见给予评价。

相关性是指答复意见是否对留言详情内的提问进行了针对性的解释和回复。拟采用文本相似度的方法对答复意见和留言详情的相关性进行量化计算。

完整性指答复意见是否满足答复规范。如答复意见应当包括对于留言问题的针对性解释；如有依据的法规，应当在答复意见中给出；同时，还应给出有关部门的联系方式。总体来说，完整性即代表回答需要有一定的格式和形式。

可解释性指答复内容中内容的相关解释。

本任务只给出方案与想法，并未实现。

5.1 相关性

考虑答复意见与留言的相关性，先对数据进行预处理，再使用 Doc2vec 模型来计算两者的相似度，并用相似度作为衡量相关性的主要指标，流程如图 13 所示。

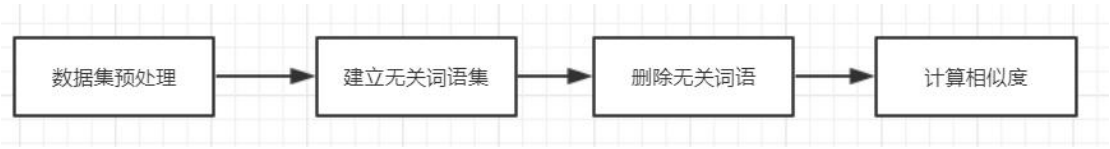


图 13 相关性分析流程

5.1.1 预处理

在计算相似度之前，需要对数据集进行预处理。预处理包括数据清洗，分词，去停用词等。（见第 2 章数据准备）

除了常规的预处理之外，需要去除留言和答复意见中的无关词语与习惯性短语，这与预处理中的去停用词是不同的，在本数据集中，会有“特异性”地出现一些高频的与留言或答复主题无关的习惯性用语。这些与主题没有太大关系的习惯性用语会干扰相似度的计算，使相似度降低。

例如，在留言中以极高频率出现了“您好”，“希望有关部门严查”，“希望有关领导重视”等等与留言内容主题并无直接关系的习惯性短语，在答复中又

高频率出现如“网友你好！”，“感谢对我们工作的关心”等等短语，句子。而这些内容会干扰相似度的计算。

对于去除与主题无关的习惯性用语，考虑采用基于词表的方法。

一种方法是人工归纳，标注一个习惯性用语词表。

二是基于频率的方法提取这些习惯性短语，一种可以考虑的处理方法是基于所有留言内容和回复内容建立词典，并分别进行词语和短语的频率统计，词频过高的词语和短语则极有可能是习惯性用语。

5.1.2 计算相似度

删除掉无关词语后，考虑使用 word2vec 计算两者相似度。文本相似度越高则留言与答复相关性越强。

5.1.3 Word2vec

word2vec 的基本原理如下。

word2vec 是谷歌于 2013 年推出的 NLP 工具，特点是将词语进行了向量化，并可以进一步挖掘文本之间的关系。Word2vec 主要采用 CBOW(ContinuousBag-of-Words Model)和 Skip-Gram(ContinuousSkip-Gram Model)两种模型[9]。

无论是 CBOW 模型还是 Skip-Gram 模型，都是以 Huffiman 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 CBOW 或者 Skip-Gram 方法获得每个单词的最优向量。

5.2 完整性

对于完整性的检测，考虑使用规则和制定权重来量化完整性。

5.2.1 答复格式划分

对数据集进行定性分析可知，一条答复的内容通常分为以下几个部分，如图14。

例如，数据集中有答复数据如下：

您好！您反映的问题已收悉，现将有关情况回复如下：B 市现行的专利资助政策依据是《B 市知识产权战略推进专项资金管理办法》（株财发[2016]20 号）文件，具体的资助内容为第三章内容。B 市现行的专利资助为一年内集中办理一次，采取申报制。每年只在一段时间内集中受理资助时间段内发生的专利资助事宜。资助时间段以每年的资助通知确定的时间为准。在资助工作启动后会将相应的专利资助通知在 B 市科学技术局网站进行公开，通知会将资助类型、标准、申报人需要准备的相关资料、资料报送方式及途径进行明确。申请人只需要按通知要求进行操作即可。我市的资助采取申请制，在规定申报时间段内不提交相关申请资料视为放弃相关资助。具体情况请咨询 B 市科技局专利管理运营科 0731-0000-00000000。

划分成五个部分为：

称呼，问候：您好！

重复声明回复的问题：<缺失>。

问题解答：现将有关情况回复如下：B 市现行的专利资助政策依据是《B 市知识产权战略推进专项资金管理办法》（株财发[2016]20 号）文件，具体的资助内容为第三章内容。B 市现行的专利资助为一年内集中办理一次，采取申报制。每年只在一段时间内集中受理资助时间段内发生的专利资助事宜。资助时间段以每年的资助通知确定的时间为准。在资助工作启动后会将相应的专利资助通知在 B 市科学技术局网站进行公开，通知会将资助类型、标准、申报人需要准备的相关资料、资料报送方式及途径进行明确。申请人只需要按通知要求进行操作即可。我市的资助采取申请制，在规定申报时间段内不提交相关申请资料视为放弃相关资助。

解答中涉及的法律法规和文件：B 市现行的专利资助政策依据是《B 市知识产权战略推进专项资金管理办法》（株财发[2016]20 号）文件。

进一步联系方式：具体情况请咨询 B 市科技局专利管理运营科 0731-0000-00000000。

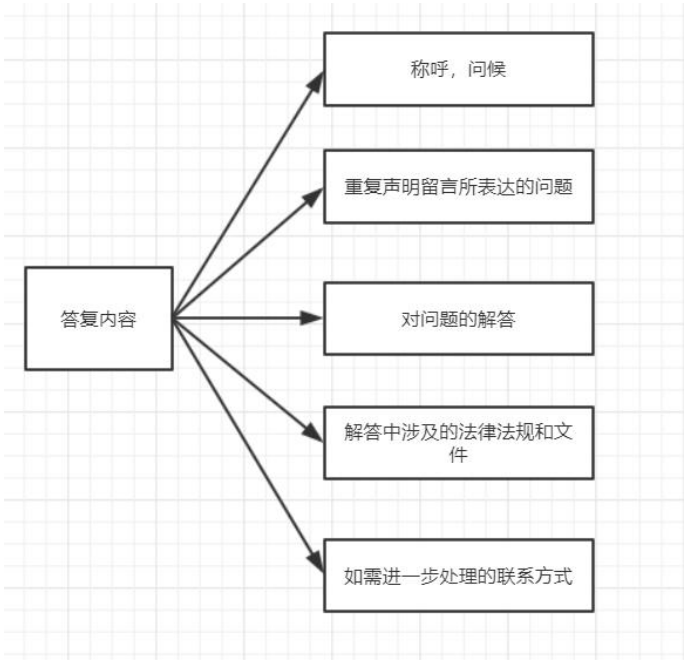


图 14 答复内容划分

一条答复若包含了以上全部五点，则认为完整性较好。若只包含部分内容，则考虑划分以上五部分内容的权重。如果检测到答复内容满足以上某几个部分，就根据对应权重进行计算，最后得出一个量化的分数。

5.2.2 划分答复

关于识别答复内容的方法，一种可行的方式是，首先对答复内容进行拆分，并归并到以上五个部分。对于以上五个部分，观察到，每个部分都会有相应的特征词语出现，如图所示。

人工对以上五个部分的特征词或特征句进行标注，得到特征词集。

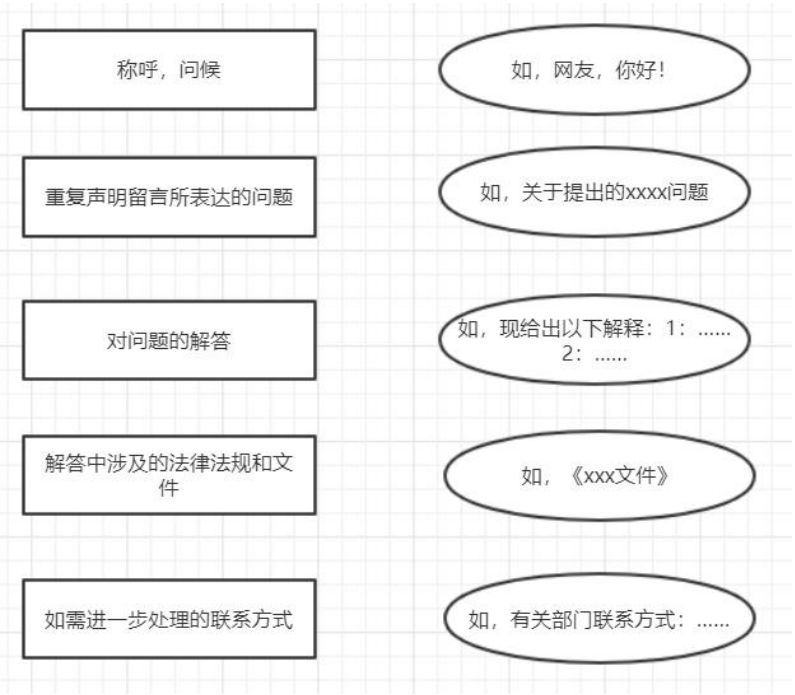


图 15 答复内容特征词集

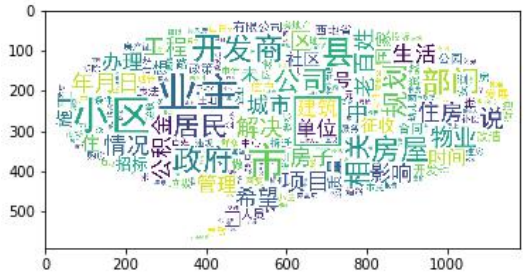
基于特征词句集对答复进行拆分，应注意五个部分的粒度是不同的。涉及到的法律法规和文件粒度为词或标点符号（书名号）；称呼，问候粒度为词语和句子。其他三者为句子。

故应从粒度较小的部分开始进行拆分归并，每当识别一部分后，将其删除，再识别下一部分。

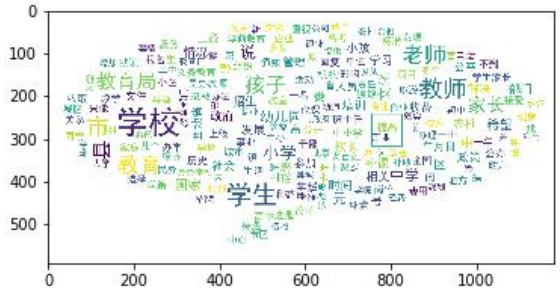
参考文献

[1]余本功,曹雨蒙,陈杨楠,杨颖.基于 nLD-SVM-RF 的短文本分类研究[J].数据分析与知识发现,2020,4(01):111-120.
[2]邢娟韬,白金牛.基于改进 ML-KNN 算法的文本分类研究[J].科技创新与应用,2020(09):25-26+28.
[3]杨孟英. 基于支持向量机的中文文本分类研究[D].华北电力大学,2017.
[4]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.
[5]杨锋.基于线性支持向量机的文本分类应用研究[J].信息技术与信息化,2020(03):146-148.
[6]贺心皓. 基于支持向量机的文本分类研究[D].成都信息工程大学,2019.
[7]陈利军,王畅.基于 DBSCAN 的地震电离层扰动异常数据检测方法[J/OL].地震工程学报:1-6[2020-05-07].
[8]薛兴荣,靳其兵.基于词典的文本极性计算及分类研究[J].网络安全技术与应用,2020(04):57-61.
[9]李晓, 解辉, 李立杰. 基于 Word2vec 的句子语义相似度计算研究[J]. 计算机科学, 2017, 044(009):256-260.

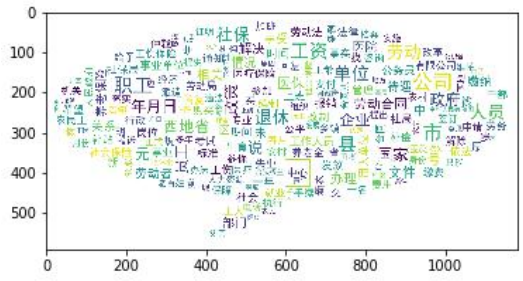
附录



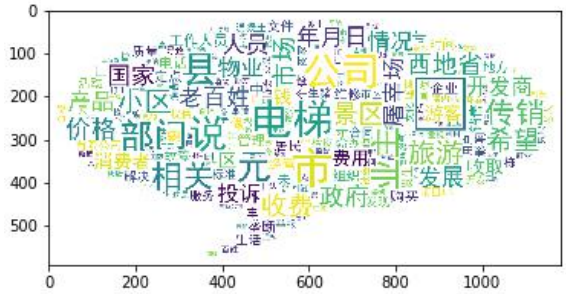
城乡建设词云图



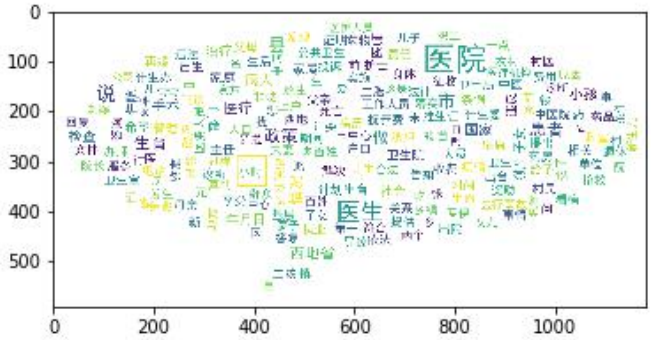
教育文体词云图



劳动和社会保障词云图



商贸旅游词云图



卫生计生词云图