2020 年全国泰迪杯
数据挖掘挑战赛

基于 Python“智慧政务”中的文本挖掘的应用

摘要

本文基于 Python 研究“智慧政务”中的文本挖掘及应用，通过所给附件内的留言情况建立分类模型，提取热点问题，并结合留言详情制定一套关于答复意见的质量评价方案。

针对问题一：题目要求建立关于留言内容的一级标签分类模型，那么首先读取数据（附件 2），对文本数据进行预处理（数据增强、中文文本分词、停用词过滤等），再将文本向量化，通过 TF-IDF 算法[1]构造文本特征向量集，再利用 LinearSVC 类[2]构造并训练一级标签分类模型，最后使用 sklearn.metrics[3]中的 F1(精准率与召回率的平衡)分数评估模型，为了使分类模型更为准确，需不断重复上述步骤，不断进行数据增强，以提高结果模型的精准率。

针对问题二：读取附件 3 中的数据，事先使用基于规则的方法（正则表达式处理），从“留言主题”中提取出命名实体（特定地点和特定人群），保存到文档，形成自定义词库。然后把文本数据转换为列表形式，通过 Gensim[4]形成语料库，使用 LsiModel 模型[5]算法，创建 TF-IDF 模型，并传入语料库来训练模型，得出 Tfidf 值，进而计算出相似度，然后对热度问题进行排序，最后通过其它一些基本方法将结果输出，并保存到相应文件中。

针对问题三：根据附件 4 中的“留言详情”和“答复意见”，从多个角度出发，对文本数据进行对比分析，制定一套关于答复意见的质量评价方案。

关键字：分类　TF-IDF　SVM 模型　LsiModel 模型 文本相似度 命名实体识别

[1] TF-IDF（term frequency－inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。

[2] SVM 可以用于分类、回归、异常检测。SVM 库中包括 SVC、LinearSVC 接口

[3] sklearn.metrics 分类指标

[4] Gensim 是一款开源的第三方 Python 工具包，用于从原始的非结构化的文本中，无监督地学习到文本隐层的主题向量表达。它支持包括 TF-IDF，LSA，LDA，和 word2vec 在内的多种主题模型算法，支持流式训练，并提供了诸如相似度计算，信息检索等一些常用任务的 API 接口

[5] LSI 一种简单实用的主题模型。LSI 是基于奇异值分解（SVD）的方法来得到文本的主题的

Text mining and application in ”smart government affairs” based on python

Abstract

This article is based on Python research on the text mining and application of "smart government affairs". A classification model is established based on the message situation in the attached attachment, hot issues are analyzed, and a set of quality evaluation schemes for answering opinions is formulated based on the message details.

Aiming at the problem of the first, The title requires the establishment of a first-level label classification model on message content, then first read the data (Annex 2) , the text data pre-processing (data enhancement, Chinese text segmentation, stop word filtering, etc.) , then the text is vectorized, the text feature vector set is constructed by TF-IDF Algorithm, the first level label classification model is constructed and trained by LINEAR SVC class, and finally sklearn. In order to make the classification model more accurate, it is necessary to repeat the above steps and enhance the data to improve the accuracy of the result model.

Aiming at the problem of the second, read the data in Annex 3, use a rule-based approach (regular expression) beforehand, extract named entities (specific places and specific people) from the "message topics", save them to the document, and form a custom thesaurus. Then the text data is transformed into a list form, the Corpus is formed by Gensim, the TF-IDF model is created by using LsiModel algorithm, and the TF-IDF model is imported into the Corpus to train the model, the TFIDF value is obtained, then the similarity is calculated, and then the heat problem is sorted, finally a some other basic methods will be output, and saved to the corresponding file.

Aiming at the problem of the third, According to the "message details" and "response comments" in Annex 4, the text data comparative analysis, the development of a set of quality evaluation plans for reply opinions was formulated.

Keywords: Classification TF-IDF SVM model LsiModel model Similarity Named Entity Recongition

目 录

1 问题重述

1.1 问题背景

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、凝聚名气的重要渠道，但伴随着各类社情民意相关的文本数据量不断攀升，收集到的社情民意繁多且杂乱，还可能存在恶意投递空白问题，也可能有人心急，多次投递重复的民意，不仅增加了工作人员的工作负担，还可能因为问题解决、反馈的不及时，给民众带来一定的困扰，这些都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极大的挑战。

为解决这一问题，我们可事先通过计算机自动处理文本，排除空的重复的民意。同时，在如今大数据、云计算、人工智能等技术的发展下，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。为此，我们可以利用自然语言处理和文本挖掘的方法，对群众留言进行分类，阶段性的提取热点问题。这不仅能够减轻工作人员的工作负担，而且能够及时解决民众的问题，并作出反馈，有效提升政府的管理水平和施政效率，更好的服务大众。

因此，“智慧政务”的文本挖掘应用对收集社情民意系统有重大意义。

1.2 问题描述

1）为解决目前好多电子政务系统依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题，参照附件 1，使用附件 2 的留言详情和一级标签这两列进行分析，建立一级标签分类模型。该模型的准确率较高。

2）为了及时发现热点问题，以帮助相关部门进行有针对性地处理，提升服务效率。根据附件 3 将某一时段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，给出相应热点问题对应的留言信息。

3）根据附件 4 给出的数据，从答复的相关性、完整性、可解释性等角度建立一套答复意见的质量的评价方案并实现。

2 问题分析

2.1 问题一的分析

2.1.1 分类模型的定义

根据留言详情所反映的文本数据特征，按照一定的划分体系进行自动分类。便于后续将留言分派给相应的部门处理。

2.1.2 问题分析

（一）主要任务

根据附件 2 所给数据，训练出关于留言内容的一级标签分类模型。

（二）需解决的问题

数据预处理：解决样本不均衡问题、清洗数据降低计算复杂度和系统开销；

获得分类文本的特征集向量：筛选出最具文本特征的词语并向量化；

建立训练分类器，评估分类模型。

2.2 问题二的分析

2.2.1热点问题的定义

某一时段内群众集中反映的某一问题可称为热点问题。关键点在于一段时间集中爆发的问题，即多人反映同一问题。

2.2.2 问题分析

（一）主要任务

根据附件 3 将某一时段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，按格式给出排名前 5 的热点问题及其具体留言信息。

（二）需解决的问题

问题识别：如何从众多留言中识别出相似的留言；

问题归类：把特定地点或人群的数据归并，即把相似的留言归为同一类问题，结果对应表 2；

热度评价：热度评价指标的定义和计算方法，对指标排名之后得出对应表 1。

（三）可能存在的难点

难点一：地点、人群的识别（表达多样化）

难点二：相似的计算复杂（特征多、两两之间计算相似计算量大）

2.2.3 文本相似性热度统计

根据需求可从不同维度进行统计：

1.分组不分句热度统计

根据某列首先进行分组，然后再对描述类列进行相似性统计。

2.分组分句热度统计

根据某列首先进行分组，然后对描述类列按照标点符号进行拆分，然后再对这些句进行热度统计。

3.整句及分句热度统计

对描述类列/按标点符号进行分句，进行热度统计。

4.热词统计

对描述类进行热词统计。

由于给出的文本数据相对较多，所以该题采用整句热度统计这层次对留言详情按标点符号进行分句，进行热度统计。

2.3 问题三的分析

为了能够规范对留言做出的答复意见，为了群众的问题得到解决，我们制定了答复意见评价方案。

3 符号说明

表 1 符号说明

符号	符号含义
TF	词频
stopword	停用词
IDF	逆文档频率
TP（True Positives）	正类判定为正类
TN（True Negatives）	负类判定为负类
FP（False Positives）	负类判定为正类，即“存伪”
FN（False Negatives）	正类判定为负类，即“去真”

4 模型建立与求解

4.1 问题一：群众留言分类

4.1.1 模型思想

将样本数据中的留言详情转为包含数字的词频向量，再转为 TF-IDF 向量，训练基于支持向量机的机器学习模型分类器。

（一）TF-IDF（Term Frequency-Inverse Document Frequency，词频-逆文件频

率)算法：

TF-IDF（term frequency－inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。TF 是词频(Term Frequency)，IDF 是逆文本频率指数(Inverse Document Frequency)。TF-IDF 是在单词计数的基础上，降低了常用高频词的权重,增加罕见词的权重。一个词语在一篇文章中出现次数越多, 同时在所有文档中出现次数越少（罕见）, 越能够代表该文章。

1) 计算词频

词频 (term frequency, TF)指的是某一个给定的词语在该文件中出现的次数。为了防止偏向长的文件。将其归一化(词频除以文章总词数)，公式如下：

$$TF_w = \frac{在某一类中词条w\ 出现的次数}{该类中所有的词条数目}$$

2) 计算逆向文件频率

逆向文件频率 (inverse document frequency, IDF)指的是如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。由总文件数目除以包含某一特定词语之文件的数目，再将得到的商取对数就可以得到这一词语的 IDF。公式如下：

$$\mathrm{IDF} = \log\left(\frac{语料库的文档总数}{包含词条w\ 的文档数+1}\right)$$

注：分母之所以要加 1，是为了避免分母为 0 的情况。

3) 计算词频-逆文件频率

词频-逆文件频率 (Term Frequency-Inverse Document Frequency, TF-IDF)指的是用某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，算出高权重的 TF-IDF。公式如下：

$$\mathrm{TF-IDF} = 词频（TF）\times 逆文档频率（IDF）$$

（二）支持向量机（Support Vector Machine，SVM）一对一方法

SVM 是基于统计学习理论，以结构风险最小化原则为理论基础，通过选择适当函数子集及该子集中的判别函数学习机的实际风险降至最小，保证了通过有限训练样本得到的小误差分类器对独立测试集的测试误差仍然小，得到一个具有最优分类能力和推广泛化能力的学习机。

一对一方法（one-against-one）[1]在 k-类训练样本中构造所有可能的二类分类器，每类仅仅在 k-类中的二类训练样本上训练，结果共构造 N=k(k-1)/2 个分类器。通过投票法组合这些二类分类器，得票最多的类就是新点所属的类。

4.1.2 模型建立

（一）流程图

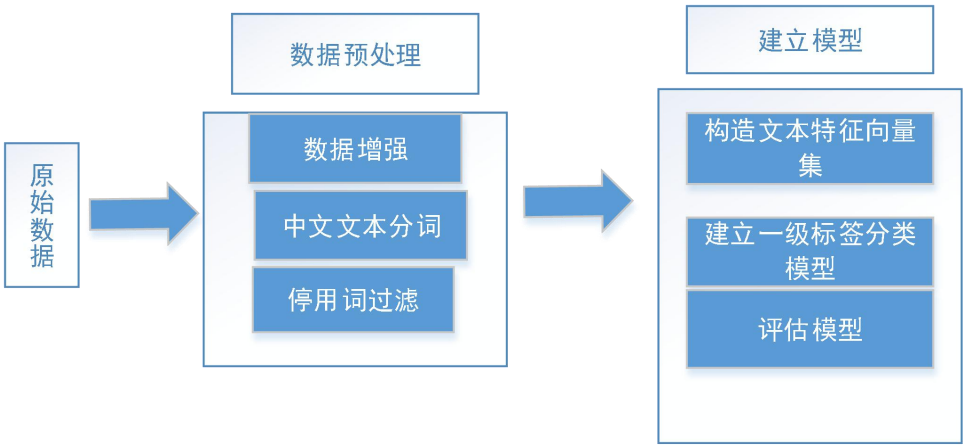

图 1 题目一的流程图

（二）数据预处理

本文把对文本数据的预处理分为四部分：

1) 分析样本数据

读取附件 2 数据进行分析,，得出数据总量有 9210，无空行空列。查看各一级标签的数据量发现数量不一致，分布不均匀。最多的一级标签数量有 2009，而最少才 613。

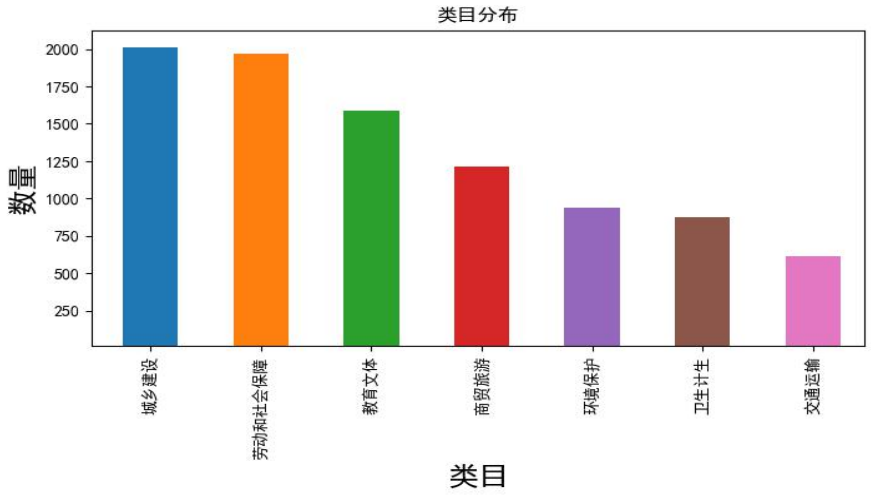

图 2 样本数据各一级标签的分布情况

2) 对样本数据进行数据增强

针对 1）发现的样本分布不均匀问题,根据一级标签对文本进行分类后，将不同分类的文本通过 EDA 算法进行同义词替换-新增-交换-删除-生成同义句[2] 进行不同程度的数据增强。这里采用的是 Synonyms 同义词库，生成相对应标签文件，如环境保护.xlsx、交通运输.xlsx 等。增强后的数据总量有 3 万多条，大部分数据量在 4000 左右，分布较为均匀。将生成相对应标签文件合并存储在 output.xls 中。

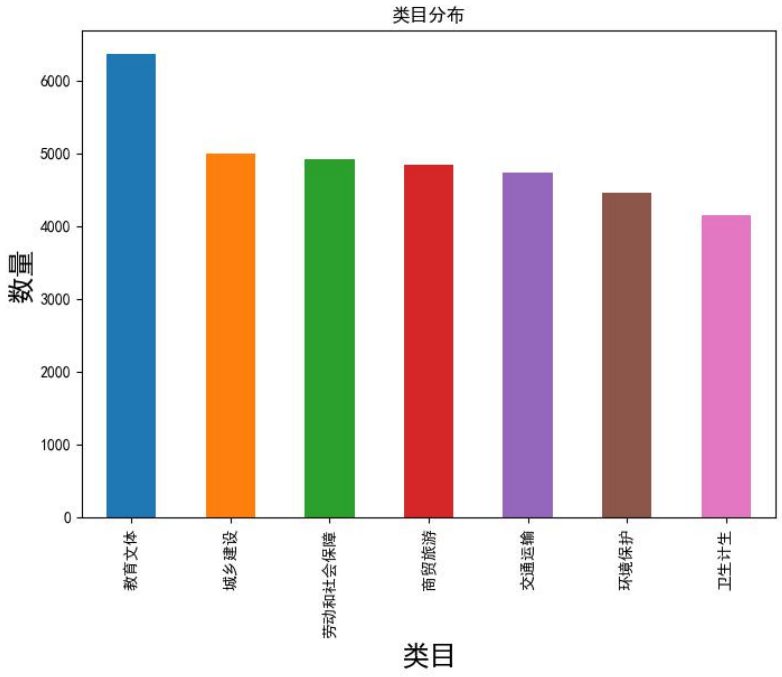

图 3 增强后各一级标签数据的分布情况

3) 对留言详情进行数据清洗

对 output.xls 中数据，运用正则表达式和过滤停用词（stop word），清洗掉对系统分析预测文本的内容没有任何帮助而会增加计算的复杂度和增加系统开销的符号和词语。

4) 中文分词

将清洗后的数据，运用 python 的中文分词包 jieba 进行分词。部分数据如图：

	一级标签	留言详情	cat_id	clean_review	cut_review
0	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t\t座落在月亮岛街道的时代倾城居民小区一二三期之...	0	座落在月亮岛街道的时代倾城居民小区一二三期之间的公共路段两边的人行道常年停满了社会制度车严重...	座落在 月亮 岛 街道 时代 倾城 居民小区 二三期 之间 公共 路段 两边 人行道 常年 ...
1	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t\t地处月亮岛街道时代倾城小区一二三期之间的公共...	0	地处月亮岛街道时代倾城小区一二三期之间的公共道路两边的人行道常年停满了社会车辆严重影响行人通...	地处 月亮 岛 街道 时代 倾城 小区 二三期 之间 公共 道路 两边 人行道 常年 停满 ...
2	交通运输	规划\t\t\t\t\t\n\t\t\t严重影响\t不知月亮岛街道时代倾城小区一\n之间的的...	0	规划严重影响不知月亮岛街道时代倾城小区一之间的的道路下的人行道常年停满了社会车辆行人通行社区...	规划 影响 不知 月亮 岛 街道 时代 倾城 小区 之间 道路 人行道 常年 停满 社会 车...
3	交通运输	\t\t\n\t\t\t\t\t地处月亮岛时代倾城小区二三期之间的公共两边人行道常年了社会车辆严...	0	地处月亮岛时代倾城小区二三期之间的公共两边人行道常年了社会车辆严重影响行人通行社区和交警也一...	地处 月亮 岛 时代 倾城 小区 二三期 之间 公共 两边 人行道 常年 社会 车辆 影响 ...
4	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t\t地处月亮岛街道时期倾城小区一二三期之间的公共...	0	地处月亮岛街道时期倾城小区一二三期之间的公共路段两旁的人行道常年停满了社会车受到影响行人通行...	地处 月亮 岛 街道 时期 倾城 小区 二三期 之间 公共 路段 两旁 人行道 常年 停满 ...

图 4 数据预处理后部分结果

（三）基于 TF-IDF 算法的文本特征提取并向量化

将预处理好的数据通过 sklearn.feature_extraction.text.TfidfVectorizer 方法提取文本的 TF-IDF 的特征值并将其向量化。这里我们使用了参数 ngram_range=(1,2),这表示我们除了抽取评论中的每个词语外,还要抽取每个词相邻的词并组成一个“词语对”,如: 词 1，词 2，词 3，词 4，(词 1，词 2)，(词 2, 词 3)，(词 3，词 4)。这样就扩展了我们特征集的数量,有了丰富的特征集才有可能提高我们分类文本的准确度。

（四）构造并训练基于 SVM 一对一算法的分类器

得到生成各个标签的 TF-IDF 向量后，这里用的是 scikit-learn 支持向量机算法库中的 LinearSVC 类实现构造并训练分类器[3]。

（五）对模型进行评价

调用 sklearn.metrics 中的 classification_report 函数和 accuracy_score 方法[3]，显示每个类的精确度、召回率、F1 值、分类准确率分数等信息以此来评估分类模型。

分类准确率分数：所有分类正确的百分比。

精确率：（Precision）真正正确的占所有预测为正的比例。公式如下：

$$\text{Precision} = \frac{TP}{TP + FP}$$

召回率（Recall）：真正正确的占所有实际为正的比例。公式如下：

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 值：精确度和召回率的调和平均值。公式如下：

$$\frac{2}{F_1} = \frac{1}{Precision} + \frac{1}{Recall}$$

4.2 问题二：热点问题挖掘

4.2.1 流程图

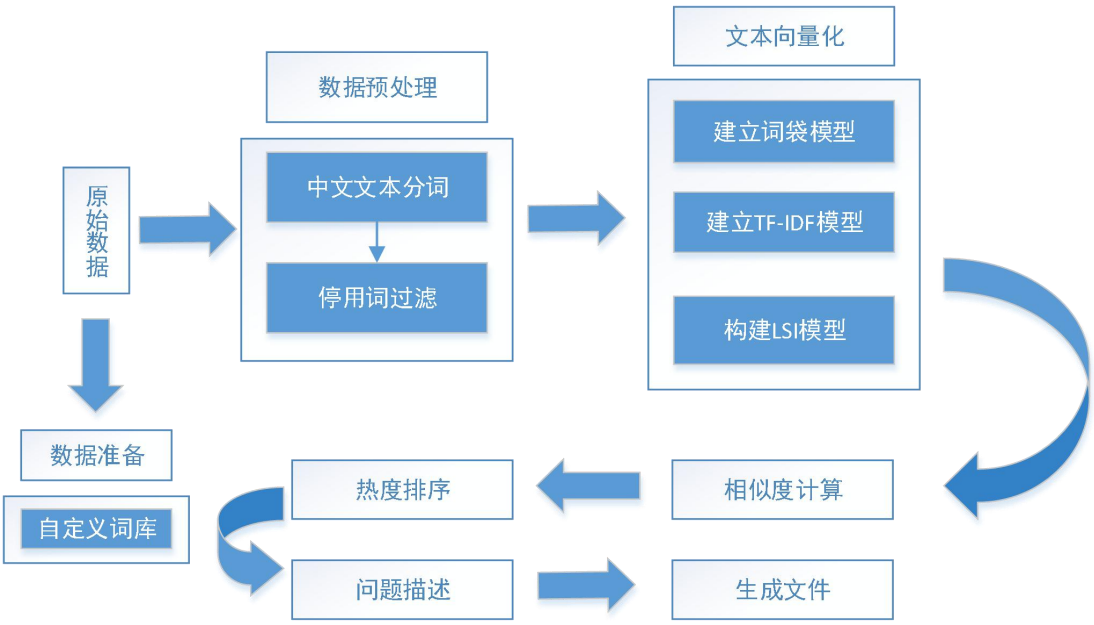

图 5 问题二流程图

4.2.2 算法描述

（一）命名实体识别

（1）命名实体识别（Named Entity Recognition，简称 NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。

（2）命名实体一般包括三大类（实体类、时间类、数字类），七小类（人名、机构名、地名、时间、日期、货币和百分比）。

（3）命名实体识别过程：确定实体的边界，即确定哪些词属于实体；确定实体的类别，即确定实体属于人名或者机构名等。

（4）主要方法

- 基于规则的方法

主要通过手工构造规则模板，选用特征，如关键字、标点符号、位置词、指

向词等，基于不同的规则权值进行判断。基于规则的方法性能上可解释性强，对于 badcase 的解决能力高，但构建规则库需要语言学专家且耗费时间长。

- 基于统计的方法

基于统计机器学习的方法主要包括：隐马尔科夫模型、最大熵模型、条件随机场等。 实际上是将命名实体识别转化为一个序列标注任务，这部分工具与分词及词性标注有一定的重合之处。

（二）Python+Gensim-文本相似度分析

使用 Gensim 计算文本相似度一般是先用 copora 模块把文档转为简单的稀疏矩阵；然后用 models 模块得到符合需要的向量模型；最后用 similarities 模块计算相似度。具体流程如图 6 所示。

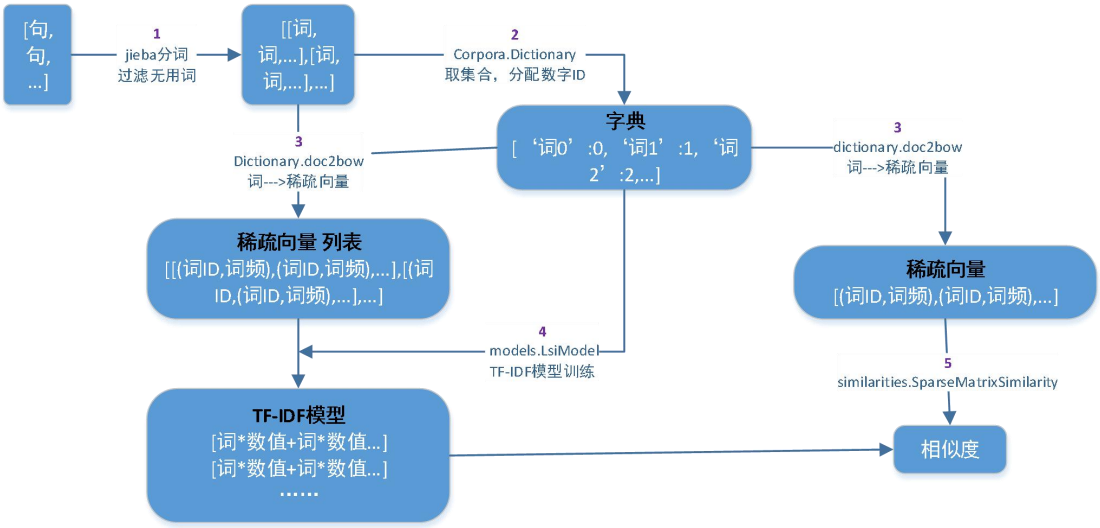

图 6 Gensim 使用流程图

4.2.3 解题方法及过程

（一）数据描述

通过观察所给数据，可以发现全部数据量共计 4326 条，且附件 3.xlsx 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。该题主要通过对数据进行相似度比较后再进行热度排序，由于文本数据存在大量复杂地名，所以可单独通过“留言主题”列将特定地点、特定人群等命名实体提取出来，形成自定义词库，以便提高中文分词的准确性，进而提高文本之间的相似性。同时，因已知文本数据包含留言主题及留言详情，两者可能存在一定的误差，为保证最终得到较为精准的热点问题，可分别通过对留言主题和留言详情进行相似度比较、热度排序后取二者较为精确的热点问题挖掘结果。（由于两者解决算法基本

一致，所以下面过程分析以“留言详情”为例）

（二）数据准备

根据已知数据（附件 3），通过命名实体识别基于规则的方法提取相应实体（主要以地名为主），并保存为文件，以便后续使用。具体过程如图 7 所示。

图 7 命名实体识别（地点识别）步骤

（三）数据预处理（中文分词）

在对热点问题进行挖掘分析之前，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。在附件 3 中，以中文文本的方式给出了数据。为了便于后续转换，先要对附件 3 中“留言详情”列进行中文分词。这里采用 Python 开发的一个中文分词模块——jieba 分词[4]的方法。jieba 分词用到的算法：

- 基于 Trie 树结构[5]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

同时，为节省存储空间和提高搜索效率，在处理文本之前将停用词进行过滤，且为提高分词的准确性，并导入自定义词典，将相应的命名实体合在一起。部分分词结果示例如图 8 所示。

```
0     [座落在，A市，A3区，联丰路，米兰，春天，G2，栋，320，一家，名叫...
1     [A市，A6区，道路，命名，规划，初步，成果，公示，文件，转化，正式，...
2     [系，春华镇，金鼎村，七里，组，村民，不知，相关，水泥路，到户，政策，...
3     [靠近，黄兴路，步行街，城南路，街道，古道，巷，一步，两，搭桥，小区，...
4     [A市，A3区，中海国际社区，三期，四期，蓝天，璞，洲，幼儿园，旁边，块...
5     [麓泉社区，麓谷明珠小区，栋，居民，近期，感觉，震惊，伤心，购房，签合同,...
6     [“，二高，一部，发出，非法，集资，打击，通知，中是，金融，犯罪，通知...
7     [一名，A市，地铁站，上班，安检员，中介，公司，介绍，上班，安检员，岗位...
8     [12，21，下午，17，52，分许，6路公交车，司机，座位，旁边，汽车...
9     [保利，麓，谷林语，桐梓坡路，麓松路，交汇处，地铁，凌晨，点，施工，噪音...
10    [下午，晚，高峰，五点，半左右，特立路，东四路，口时东，往西方，向车，越...
11    [宁静，我要，复习，迎考，大半，年底，商，空调，冰柜，外机，轰响，扰民...
12    [桐梓，坡，589，号，白鹤，咀，停车场，由聚，美龙楚，新能源，公司，...
13    [想，举报，A市，利保，壹号，公馆，项目，夜间，噪声，扰民，A市，利保...
14    [A7县，星沙街道，两个，孩子，上有老下有小，居民，居民，直斥，A市，地铁...
15    [北辰，D1，区，一名，业主，小区，住，改商，数量，达上，百家，安全隐...
16    [K3县，乡村，卫生室，处于，无证，行医，状态，情况，原因，卫生室，医疗...
17    [春华镇，一名，村民，接到，政府，通知，029，县，道旁，开，麻将馆，...
18    [书记，外地人，2018，年，毕业，落户，A市，申请，短期，出国，旅游,...
19    [退费，之日起，之内，退还，学习，费用，今日，超过，合同期，未，退费，...
20    [A市，A6区，月亮岛街道，乾源，国际，广场，停车场，违章，乱建，现象，...
21    [A7县，时代星城小区，幢，1321，室，非法，一家，家庭旅馆，住宅小区，...
22    [敬爱，佳兆业，水新，一期，小区业主，小区，43，栋，101，户，违建,...
23    [沙坪，街上，一家，无证，经营，理疗，馆，打着，免费，做，健康，理疗,...
24    [a，我姐，杜四娇，贵区，管辖，内德鸿，餐饮，打工，五年，2019，年,...
25    [长房云时代小区，月底，开盘，月初，摇号，选房，10，号，业主，观察，1...
26    [谢谢您，这份，非常感谢，松，雅西，地省，几个，小区，传销，记得，特别,...
27    [2019，年，12，18，上午，九点，A市，01，年，11，所下区，...
28    [黄兴路，步行街，仕弘，教育，A市，晚报，大厦，八楼，前程，优学，确保,...
29    [多年，A2区，迎新路，区政府，东门，万芙路，段，改装车，飙车，高考，期...
```

图 8 “留言详情”部分分词结果

（四）TF-IDF 文本相似度分析

（1）使用 gensim 中的 corpora 模块，将分词形成后的二维数组生成词典；

（2）将二维数组通过 doc2bow 稀疏向量，形成语料库（如图 9）；

```
语料库：
 [[(0, 1)], [(1, 1), (2, 1), (3, 1), (4, 1), (5, 2), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1),
(16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 2), (22, 2), (23, 1), (24, 1), (25, 2), (26, 1), (27, 1), (28, 1), (29, 1), (30, 1), (3
1, 1), (32, 1), (33, 1), (34, 1), (35, 1), (36, 1), (37, 1), (38, 1), (39, 1), (40, 1), (41, 1), (42, 1), (43, 1), (44, 2), (45, 2), (46,
1), (47, 2), (48, 1), (49, 1), (50, 1), (51, 1), (52, 1), (53, 1), (54, 1), (55, 1)], [(56, 1), (57, 1), (58, 1), (59, 1), (60, 1), (61,
1), (62, 1), (63, 1), (64, 1), (65, 1), (66, 1), (67, 1), (68, 1), (69, 1), (70, 1), (71, 1), (72, 1), (73, 1), (74, 3), (75, 1), (76, 1),
(77, 1), (78, 1), (79, 1), (80, 1), (81, 1)], [(58, 1), (67, 1), (68, 1), (69, 1), (71, 1), (74, 1), (82, 1), (83, 1), (84, 1), (85, 1),
(86, 1), (87, 1), (88, 1), (89, 1), (90, 1), (91, 1)], [(1, 1), (2, 1), (5, 2), (13, 2), (14, 1), (19, 1), (22, 1), (28, 4), (38, 2), (42,
2), (43, 2), (44, 1), (45, 1), (47, 1), (48, 3), (52, 1), (72, 1), (80, 1), (92, 1), (93, 1), (94, 2), (95, 1), (96, 1), (97, 1), (98, 1),
(99, 2), (100, 1), (101, 2), (102, 1), (103, 1), (104, 1), (105, 1), (106, 1), (107, 2), (108, 1), (109, 1), (110, 1), (111, 1), (112, 2),
(113, 1), (114, 3)], [(1, 1), (2, 1), (5, 1), (13, 1), (16, 1), (19, 1), (22, 1), (26, 2), (30, 3), (38, 1), (39, 1), (41, 3), (42, 1), (4
3, 1), (45, 2), (47, 1), (48, 3), (52, 1), (54, 1), (56, 2), (94, 2), (102, 2), (110, 2), (113, 1), (114, 1), (115, 1), (116, 1), (117,
2), (118, 1), (119, 1), (120, 1), (121, 1), (122, 1), (123, 1), (124, 1), (125, 1), (126, 1), (127, 1), (128, 1), (129, 1), (130, 3), (13
1, 1), (132, 1), (133, 1), (134, 1), (135, 1), (136, 1), (137, 1), (138, 1), (139, 1), (140, 1), (141, 1), (142, 1), (143, 1), (144, 1),
(145, 1), (146, 1), (147, 1), (148, 1), (149, 1), (150, 1), (151, 1), (152, 1), (153, 1), (154, 2), (155, 1), (156, 1), (157, 1), (158,
1), (159, 1), (160, 1), (161, 1), (162, 1), (163, 1), (164, 1), (165, 5), (166, 1), (167, 1)], [(28, 1), (38, 1), (42, 1), (43, 1), (48,
1), (52, 1), (94, 2), (98, 2), (110, 3), (117, 1), (119, 1), (136, 2), (154, 2), (168, 1), (169, 1), (170, 1), (171, 1), (172, 2), (173,
1), (174, 1), (175, 1), (176, 1), (177, 1), (178, 1), (179, 1), (180, 1), (181, 1), (182, 1), (183, 1), (184, 1), (185, 1), (186, 1)], [(3
0, 1), (59, 1), (63, 1), (80, 1), (148, 1), (187, 1), (188, 1), (189, 1), (190, 1), (191, 1), (192, 1), (193, 1), (194, 1), (195, 1), (19
```

图 9 语料库部分数据

（3）刚开始使用 TF 模型算法，后来更改为 LsiModel 模型算法，将语料库计算出 Tfidf 值；

（4）获取词典 token2id 的特征数；

（5）计算稀疏矩阵相似度，建立一个索引；

（6）读取 excel 行数据，通过 jieba 进行分词处理；

（7）通过 doc2bow 计算测试数据的稀疏向量；

（8）求得测试数据与样本数据的相似度。

根据上述得出的热度指数进行降序处理，即得到题目所求的热点问题。

由于最终结果是根据留言详情进行热度排序后输出，而题目要求按照特定格式给出排名前 5 的热点问题及其对应的留言信息，并保存为相应的文件，则后续还需利用算法概括对应的热点问题进行简单描述，填入对应表中，此处就便不予详细说明。

4.2.4 方法所存在的不足

不足点 1：使用基于规则的方法提取命名实体，方法虽简单，但并不能很好的从大量文本中提取出特定地点和特定人群，同时会影响到后面的相似度比较。

不足点 2：使用“Gensim-文本相似度分析”的方法对进行相似度计算，对大文本而言，相似度的相似性较低，需不断进行测试，提高相似性，才能更好的解决问题，将热点问题提取出来。

4.3 问题三：制定评价方案

（1）相关性

相关性是指两个变量的关联程度，也指在网络搜索中，关键字被搜索引擎收录所依据的相关指标。而本题的题干要求从相关性分析答复意见，根据是否与留言详情相关作为判断，只要留言详情中的关键字和答复意见的关键字相似，就可以判断答复意见的有效性。

拟操作：先在附件 4 分别提取留言详情和答复意见，分别对他们进行分词，去除停用词，统计每一列的词频，并把词频放在向量中，用 TF-IDF 模型计算相似度。

（2）完整性

包括两个方面，第一是格式完整性，第二是内容完整性。

格式完整性：开头要有尊称，要问好。要用标准词不能使用口头语，避免群众误会理解错误。结尾要附上时间。

内容完整性：首先先明确群众提出的问题，大致的复述一遍，让群众明白这是哪个问题的答复意见，还要表示已收到，对于询问相关的法律法规或对法律法规有所疑惑，可以直接在答复中给出相应的法律法规条例或仔细解析。对于环境

还是怀疑税务之类的，则说明已经反馈到相应的职能部门，正在等待处理，请群众耐心等候。

（3）合理性

回复的内容是否客观，没有激烈的言辞，能否让人理解。

5 结果分析

5.1 问题一结果分析

问题一主要通过 TF-IDF 算法和 SVM 方法对留言进行分类。通过结果可以发现方法中仍存在某些不足之处，从而导致结果产生偏差。例如：在使用 TF-IDF 算法过程中可能会因为前期的数据清洗的不到位导致特征值不准确，而特征值主要是用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度，使得分类的准确度降低；通常使用 SVM 方法实现构造并训练分类器，但该方法的执行需要大量的存储空间，若硬件配置较差，也会对结果产生一定的影响，甚至使得程序无法正常运行。综上所述，若想结果准确性较高，硬件、软件设备都要跟得上。具体结果如图 10 所示。

accuracy 0.9903279697529236

	precision	recall	f1-score	support
交通运输	0.99	1.00	1.00	1562
劳动和社会保障	0.98	0.98	0.98	1621
卫生计生	0.99	0.99	0.99	1372
商贸旅游	0.99	0.99	0.99	1599
城乡建设	0.99	0.98	0.98	1648
教育文体	0.99	0.99	0.99	2100
环境保护	0.99	1.00	1.00	1471
avg / total	0.99	0.99	0.99	11373

图 10 题一模型评价结果

5.2 问题二结果分析

通过对民众留言信息进行热度排序计数，选取热度排名前 5 的热点问题（结

果见表 2 排名前 5 的热点问题）并结合通过“留言主题”形成的词云图（见图 11）进行分析，可以发现现阶段的热点问题还是主要以环境污染、噪音污染等为主，总的来说就是与人们日常生活息息相关，这也说明目前政府对环境管理这方面所采取的措施仍有限，后期需对这方面加强管理力度，更好的服务民众。

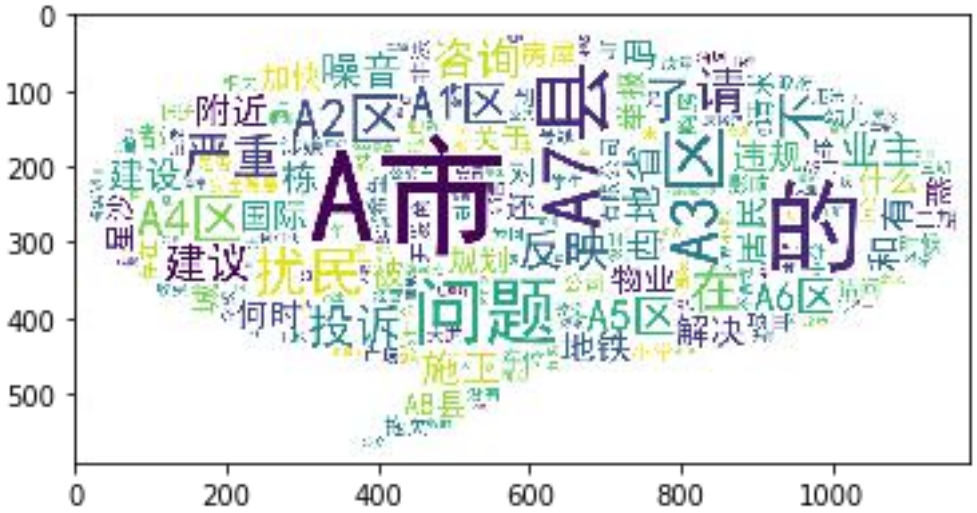

图 11 “留言主题”词云图

表 2 排名前 5 的热点问题

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	55	2019/07/11 至 2019/12/11	A 市伊景园滨河苑	开发商捆绑销售车位
2	2	52	2019/07/21 至 2020/01/15	A 市 A2 区丽发新城小区	小区附近搅拌厂灰尘噪音污染严重
3	3	21	2019/07/21 至 2019/12/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
4	4	12	2017/06/08 至 2019/11/27	A 市经济学院学生	学校强制学生去定点企业学习
5	5	12	2019/01/13 至 2019/05/22	A1 区辉煌国际城二期	居民楼下商铺违法开饭店

参考文献

[1] 胡国胜,钱玲,张国红.支持向量机的多分类算法[J].系统工程与电子技术,2006,28(1):127-132.

[2] 大漠帝国.文本数据增强二（EDA、同义词替换-新增-交换-删除-生成同义句).CSDN 博客,2019-04-27 19:32:13.

[3] -派神-.使用 python 和 sklearn 的中文文本多分类实战开发.CSDN 博客,2019-03-02 00:32:15.

[4]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(04):153-155.

[5]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.