智慧政务背景下网络留言信息的文本挖掘和政府职能评价

摘要

智慧政务是“互联网+电子政务”下的产品，利用云计算、人工智能以及数据挖掘等技术，简化群众办事环节、提升政府行政效能、畅通政府服务渠道，解决群众“办事难”等问题。

通过政府意见箱里的网络留言，可以快速挖掘民生民意，克服了人工处理民众意见时分类难、错误率高等问题，同时挖掘出某一时间段的热点问题，并对政府工作职能的效率进行评价。

针对问题一，首先利用 Jiebe 分词和 TF-IDF 算法将附件 2 的留言详情进行数据预处理形成特征空间，然后利用 MLP、SVM、模型融合等方法对比分类效果，并使用混淆矩阵的热力图直观展示，最后通过 F1-micro 计算得到最优结果为 MLP，值为 0.9186。

针对问题二，热点问题的挖掘，在数据预处理这一步骤中的分词使用了 HanLP 的 python 接口以及 TF-IDF 转化为文本向量，再利用 PCA 提取出前 k 个特征词项，将长文本转化为短文本；然后使用 pyHanLP 库中的文本聚类将同一问题归为一类，最后使用命名实体识别提取附件 3 中反应问题的特殊地点和人群；最后基于每类热点问题的留言数量、点赞数和反对数的差自定义一个热度评价方法，按热度指标从高到低排列。

针对问题三，使用 LDA 分别提取问题和答复的主题，计算主题分布的相似度；利用句法分析提取句子的成分结构，将完整性量化；运用正则表达式，人工提取答复意见中关键的字词和符号。最后，以层

次分析法的三种方式求取权重，得到稳健的评分结果。

关键词：Jieba 分词、TF-IDF、MLP 分类器、pyHanLP 分词、命名实体识别、文本聚类、LDA、句法分析、正则表达式、层次分析法。

Abstract

Smart government affairs is a product under the "Internet + e-government". It uses cloud computing, artificial intelligence and data mining technologies to simplify the masses 'work links, improve the government's administrative efficiency, smooth the government's service channels, and solve the masses' "difficult to handle" issues.

Through the online message in the government suggestion box, people's livelihood and public opinion can be quickly mined, which overcomes the problems of difficult classification and high error rate when manually processing people's opinions. .

For problem one, first use Jiebe word segmentation and TF-IDF algorithm to preprocess the message details of Annex 2 to form a feature space, and then use MLP, SVM, model fusion and other methods to compare the classification effects, and use confusion matrix heat map to visually display Finally, the best result obtained by F1-micro calculation is MLP, with a value of 0.9186.

For the second problem, the mining of hot issues, the word segmentation in the data preprocessing step uses the HanLP Python interface and TF-IDF to convert into a text vector, and

then uses PCA to extract the top k feature terms and convert the long text into Short text; then use the text clustering in the pyHanLP library to classify the same problem into one category, and finally use named entity recognition to extract the special locations and crowds in Annex 3 that respond to the problem; finally, based on the number of messages and likes of each type of hot issue The difference between the logarithm and the antilogarithm defines a heat evaluation method, which is arranged according to the heat index from high to low.

For question three, LDA was used to extract the topic of the question and answer, respectively, to calculate the similarity of the topic distribution; use syntactic analysis to extract the composition of the sentence structure to quantify the completeness; use regular expressions to manually extract the key words and answers in the reply opinion symbol. Finally, the weights are obtained in three ways of analytic hierarchy process, and a stable score result is obtained.

目录

图录

表录

第 1 章 挖掘背景

1.1 挖掘背景

随着大数据时代的发展，微信、微博、投诉信箱等网络问政平台成为政民沟通的主要渠道，物联网、移动互联网、数据挖掘等技术的兴起，为“智慧政务”在政府了解民意、汇聚民智、凝聚民气的作用提供了弹跳点。

而人民群众有关网络问政、发表社会舆情的投诉意见快速攀升，由于以往主要是人工进行留言划分和热点整理，这给相关部门带来了极大的工作量且耗费时间长、错误率较高。同时，网络文本包含大量冗余信息干扰分类准确度，相关问题的辨识度低，自然语言处理和文本挖掘在“智慧政务”上的应用是新社会的发展趋势。

此题需要根据给出的表格内容，将其中的文本转换为计算机可以理解的语言，从中提取出具有特征的句子，在分类问题上按照已知的分类组合进行有监督学习，其次在热点整理问题上，从文本中提取有效的事件，最后对比投诉意见和回复两个文本，对政府的回复质量和效率作出总结。

随着自然语言处理和机器学习、深度学习的日益成熟，如何处理长、短文本提取有效文本，已成为目前电子政务亟待解决的一个领域。构建基于文本内容的模型和评价指标，提高“智慧政务”处理效率。

1.2 挖掘目标

大数据时代背景下，加快推动智慧政务系统建设已经是社会发展的新趋势，对提升政府办事效率以及紧密联系政府与民众之间的关系都有极大助力作用。本文针对给出的群众留言记录以及相关部门对部分群众留言的答复意见。采用 Jieba 中文分词工具、MLP、pyHanLP 库的 K-Means 聚类以及 LDA、句法分析等方法，来完成以下三个目标：

（1）根据附件 1 里的一级至三级分类的信息，将附件 2 的留言内容与分类标签使用监督学习，建立分类模型，并根据 F1 值对分类效果评价。即，利用 Jieba 中文分词和 TF-IDF 方法来对留言详情进行文本预处理，再运用线性支持向量机

建立关于留言内容的一级标签分类模型，最后使用 F-Score 对模型进行评价。

（2）根据附件 3 里的留言内容，进行数据预处理之后，通过库中的 k-Means 聚类以及自定义评价方法得到热点问题，同时给出评价结果。

（3）根据附件 4 的内容，将答复意见分别作 LDA、句法分析和正则表达式解析进行相关性、完整性和可解释性的研究。

第 2 章 问题分析

2.1 问题一的分析

对附件 1 的留言详情和留言主题共同进行数据预处理和模型构建，对留言详情进行分词操作，将特殊字符和影响语义的常用词删去，再对其进行特征选择，将具有代表性的特征抽取，并用向量文本表示，最后将数据集划分为训练集和预测集使用有监督学习的分类器，将留言划分为附件 1 的一级分类标签。模型的评价使用 F1 值，通过混淆矩阵将分类结果可视化，并计算出 F1 值，选择 F1 值最高的分类器即效果最好的。

2.2 问题二的分析

对附件 2 的留言主题和留言详情进行命名实体识别，使用 HanLP 分词获取句子词性标注，将留言内容的时间、地点、人物、事件提取出来。将提取出来的句子进行相似度计算，距离相近的归为一类。最后根据点赞数和反对数对热点问题赋予热度指标。本题采用了五种模型和模型融合六种方法进行分类对比。

2.3 问题三的分析

相关性：无论答复是否解决了问题，只要与留言内容相关就具有相关性。使用 LDA 主题模型，对需要匹配的两个文本计算主题分布，之后计算相似度即可。完整性：建立一套标准，使用 Stanford Parser 进行句法分析，确定一个句子的主谓宾成分。可解释性：评价政府的留言回复是否能够针对留言有相关解释，通过对附件 4 的观察，利用正则表达式选取合适的符号和字词进行匹配，匹配度越

高可解释性越好。

2.4 技术流程图

整体解题步骤流程为：

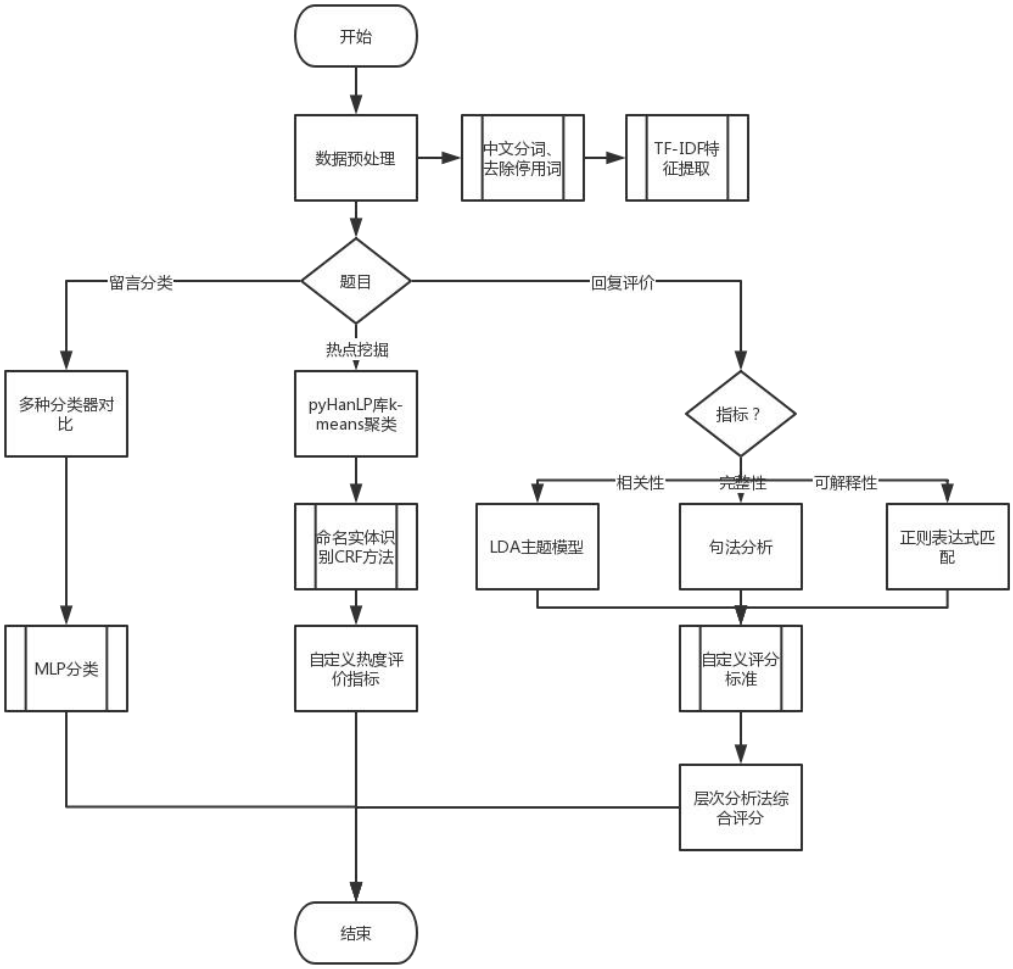

图 2-1 本文技术流程图

第 3 章 文本数据预处理

本文数据预处理主要分为 3 个部分，主要针对第一题和第二题的解答，结巴

分词在第一题中使用，pyHanLP 分词在第二题中使用。

由于给的数据中的一到两个属性是文本数据，则需要取其中的文本数据进行处理。为了将文本编辑为计算机理解的语言，要将句子转换为以 0-1 表示的向量，这样初步处理之后，利用有监督学习将文本分类，利用无监督学习将相似文本聚类，整体流程如下：

3.1 文本分词

中文分词是中文自然语言处理的一个非常重要的组成部分，也是一个基础步骤。因为中文句子没有词的界限，所以在进行自然语言处理时，根据词性不同的划分往往得到的结果不一样，分词效果也不一样。

3.1.1 结巴分词

第一题为留言详情进行标签分类，这里采用了使用最广泛、分词速度最快 python 里的 jieba 库对文本分词，jieba 库分词深度最深，对地名、人名的分词效果很差，但由于本题地名等对分类影响较小，则结巴分词对分词效果影响不大。

但需注意的是，对待同一字符串存在多个分词结果，如“中华人民共和国”可能为一个词，但分词细粒度却为“中华/人民/共和国”，又如一个字既可以与前文组成词也可以与后文组成词等等，还有一些词语可能是新词还未录入分词词典的。

（1）在进行 Jieba 分词之前，首先将所有数据一级标签与留言详情相对应，将一级标签用具体的数字代替，存在 category_id_df 中，便于之后的处理计算。针对附件 2 中的所有留言内容，先考虑到留言内容中存在非中文字符，故先进行去标点、去空和去英文数字的操作。

（2）在完成以上的基础上，调用 Python 的中问问分词包 Jieba 来进行分词。为节省存储空间和提高搜索效率，对留言内容进行去停用词（出现频率高且对信息含量的词），例如我、的、了、呢等，最终得到分词的结果并存储在 data 中。

随后利用 TF-IDF 算法来提取关键词。

3.1.2 pyHanLP 分词

由于第二题是热点分类，需要从文本中将时间、地点、人物、事件等提取出来，如果使用最传统广泛的词典分词，由于分词深度较深以及地点人名等分词词典中没有，则考虑加入自定义词典，将“魅力之城”、“经济学院”、“A 市”等地名手动输入 txt 文件；由于数据量的限制，手动输入过多并且未必可以“遍历”所有地名、人名，因此不使用结巴分词。

命名实体识别应运而生。命名实体识别是自然语言处理应用中的重要步骤，它可以检测出实体边界，还检测出命名实体的类型，是文本意义的基础，本题我们采用其中 HanLP 的 python 接口方法将地名、人名等提取出来。

pyHanLP 定义了不同的词性，依次可以自定义多种分词规则和模型，也可以加入自定义词典，但因数据量的限制和输入地名的不确定性和这里选择 pyHanLP 自带的默认分词方法效果就很好，则不选择自定义词典的方法而选择其自带的词典即可。

3.2 去除停用词

中文分词之后，出现很多词语、字和特殊字符，不仅不易审阅，而且处理麻烦，还有不少连词、副词等影响后续建模，所以这部分关键在于停用词表的维护，有利于将干扰项剔除，保留关键词和语义。目前已有的停用词表有哈工大、百度停用词表等，本题除根据已有的停用词表外，还按照文本内容特有的字符、词汇适当加入原有未包含的停用词。

本题将附件里留言内容的不重要字符和字词剔除处理之后，只剩下一些必要的字词，为使后续处理这些字词更加方便，我们将其转换为 m×n 的向量，其中每行为一个列表。接下来将这些列表转换为数字表示。

3.3 文本向量表示——TF-IDF 算法

文本的向量表示是基于向量空间模型的方法，将文本表示成实数值分量所构

成的向量，向量中的每个分量对应一个词项，相当于将文本表示成空间中的一个点。将文本转换成向量不仅可以训练分类器，而且可以计算向量之间的相似度可以度量文本之间的相似度。本题使用 TF-IDF（词频-逆文档频率）计算方式。其主要思想是：如果某个词或短语在一篇文档中出现的频率 TF 高，并且在其他文档中很少出现，则认为此词或短语具有很好的类别区分能力，适合用来分类；换言之，如果包含某个词的文档越少，IDF 越大，这说明该词具有很好的类别分区能力。

TF（Term Frequency，词频），表示在某一文档 d_j 里的词语 t_i 来说，词语 t_i 的频率为：

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

IDF（Inverse Document Frequency，逆向文件频率），即包含词语 t_i 的文档越少，IDF 越大，说明词语 t_i 在整个文档集层面上具有很好的类别区分能力。表示为总文件数除以包含该词语的文件数，再将得到的商去对数：

$$\mathrm{idf}_i = \log \frac{|D|}{|\{j:t_i \in d_i\}|}$$

最后，tf-idf 的公式为：

$$\mathrm{tf\text{-}idf} = \mathrm{tf}_{i,j} \times idf_i$$

通过 TF-IDF 算法将处理后的词项按照一定权重转换为数字表示，它可以体现文本中的重点词汇，如“交通”一词在一文本中出现频繁且在其他文档中出现的次数少，因而该算法突出主要词语、一直次要词语的作用，可以判断其属于“交通运输”一类。

由于担心存在数据排序的影响，故打乱了数据的顺序，然后对数据按照 4:1 的比例划分训练集与测试集。本文通过 TF-IDF 方法对数据进行向量化，将每个词在该类文本的 TF-IDF 值来作为量化的结果，分别对训练集和测试集采取相同的操作，提取得到各个一级分类的得分前 100 个关键词。明显可以看到卫生计生的得分较高的关键词有：医院、医生、患者等；而旅游商贸的关键词有：旅游、游客、食品等；劳动和社会保障的关键词有：劳动、单位、工资等等；其他类别不再一一赘述。

图 2-2 各类标签的文本词云图

第 4 章 标签分类模型及评价

本题使用附件 1 和附件 2，基于附件 2 的留言详情和已经分好的类别标签进行有监督学习，目前已有的分类器有很多，本题将从五个模型 RF、XGBoost、

lightGBM、MLP、SVM 和模型融合投票法六个角度选取最 F1 值最高的分类器，其中 MLP 分类器和模型融合的效果最好。

附件 2 数据集有“留言编号”、“留言用户”、“留言主题”、“留言时间”、“留言详情”和“一级标签”六个属性，其中“留言主题”为短文本，“留言详情”长短文本混杂，“一级标签”作为分类的标准。这里采用“留言主题”和“留言详情”两个属性共同处理的方式。

4.1 标签分类模型

4.1.1 RF 分类器

RF 即随机森林，是一个包含多个决策树的分类器，其输出的类别是由个别树输出的类别的众数而定。随机森林(RF)要建立了多个决策树(DT)，并将它们合并在一起以获得更准确和稳定的预测。随机森林是集成学习的一个子类，它依靠于决策树的投票选择来决定最后的分类结果。

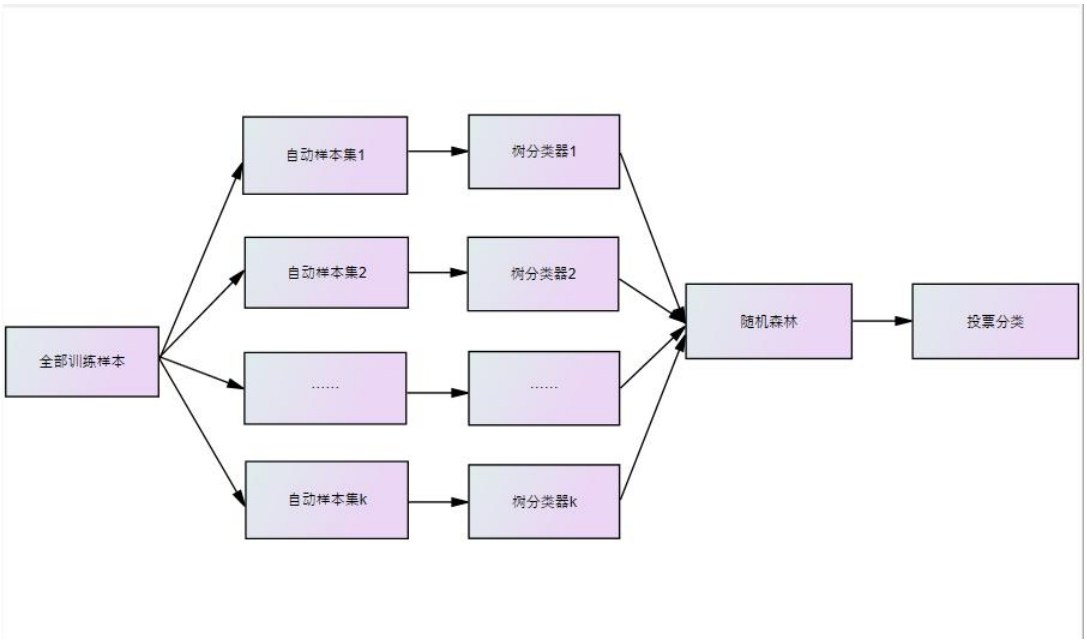

图 4-1 随即森林原理

假设 N 表示训练用例（样本）个数，M 表示特征数目，随机森林的构建过程如下：

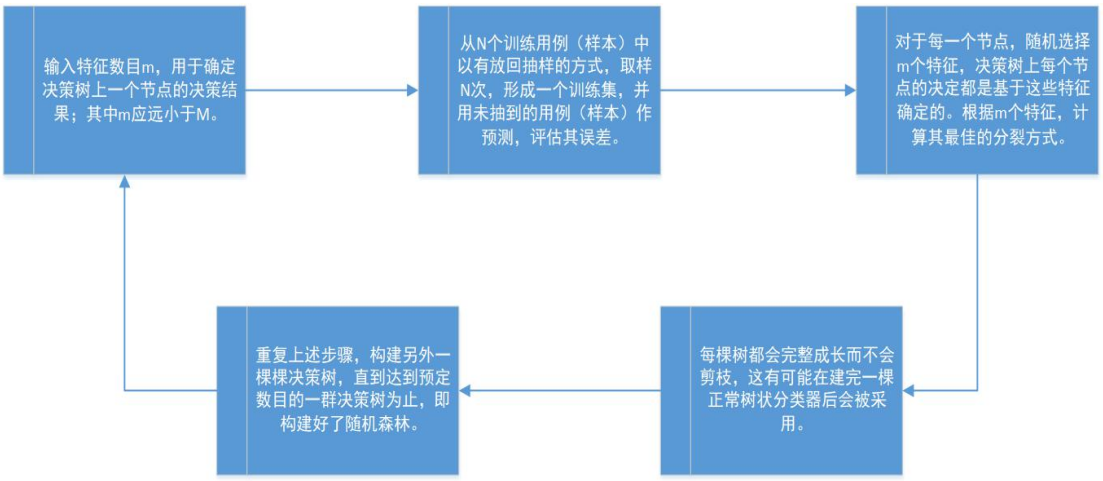

图 4-2 随机森林构建

其一，随机森林的数据选取采用有放回的抽样，构造子数据集，解决了各分类标签的数据不平衡问题；其二，随机森林中子树的每一个分裂过程仅从所有的待选特征中随机选取一定的特征，再在随机选取的特征中选取最优特征。这样能使得随机森林中的决策树都能够彼此不同，提升系统的多样性，从而提升分类性能，同时解决了多分类问题带来的难度；其三，引入随机性，随机森林不容易过拟合且具有良好的抗噪声能力，解决了文本向量与文本不相关的离群点问题。

尽管如此，当随机森林的决策树个数很多时，其训练空间和时间较大，其中还有很多不好解释的地方与黑盒模型相像。

最后的分类结果为：

使用randomforest

```
[[376   9   2  12  19   6   2]
 [ 24 164   0   5   1   2   0]
 [ 26   3  73   3   3   4   1]
 [  7   0   0 309  11   2   0]
 [  8   0   1   6 374   2   6]
 [ 42   3   1   4   9 163   2]
 [  5   0   0   1  17  10 124]]
0.8593919652551575
```

图 4-3 RF 分类效果

4.1.2XGBoost 分类器

该算法思想就是不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数，去拟合上次预测的残差。当我们训练完成得到 k 棵树，我们要预测一个样本的分数，其实就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数，最后只需要将每棵树对应的分数加起来就是该样本的预测值。

XGBoost 目标函数定义为：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

目标函数由两部分构成，第一部分用来衡量预测分数和真实分数的差距，另一部分则是正则化项。正则化项同样包含两部分，T 表示叶子结点的个数，w 表示叶子节点的分数。γ 可以控制叶子结点的个数，λ 可以控制叶子节点的分数不会过大，防止过拟合。其算法流程如图：

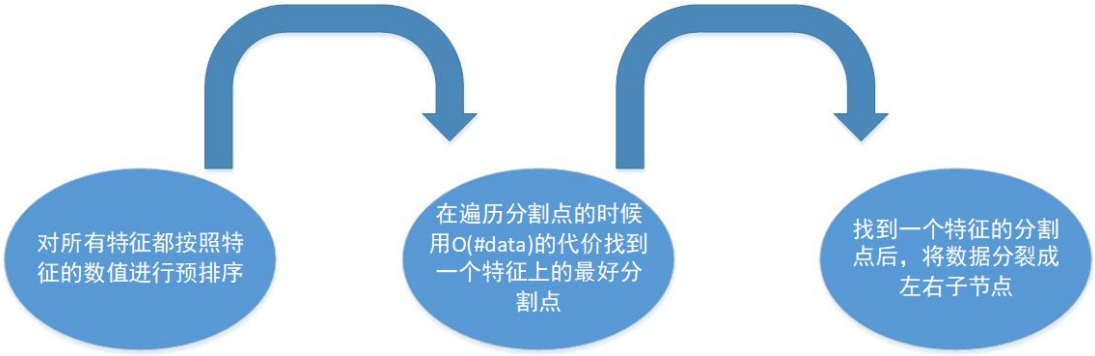

图 4-4 XGBoost 算法流程图

最后的分类结果为：

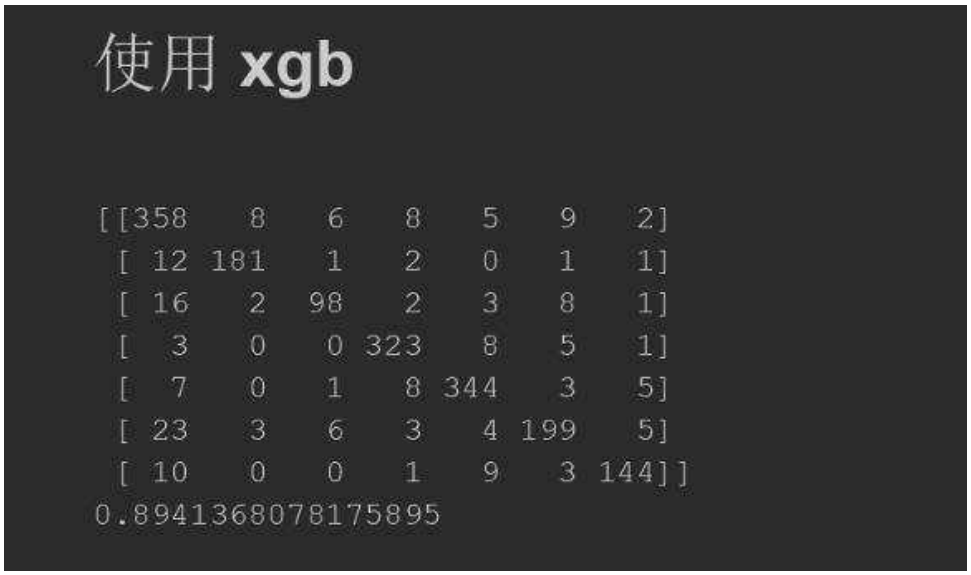

图 4-5 XGBoost 分类结果

4.1.3 lightGBM 分类器

lightGBM，基本原理与 XGBoost 一样，使用基于学习算法的决策树，只是在框架上做了一优化（重点在模型的训练速度的优化）。最主要的是 LightGBM 使用了基于直方图的决策树算法，基本思想是先把连续的浮点特征值离散化成 k 个整数，同时构造一个宽度为 k 的直方图。在遍历数据的时候，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。

最后的分类结果为：

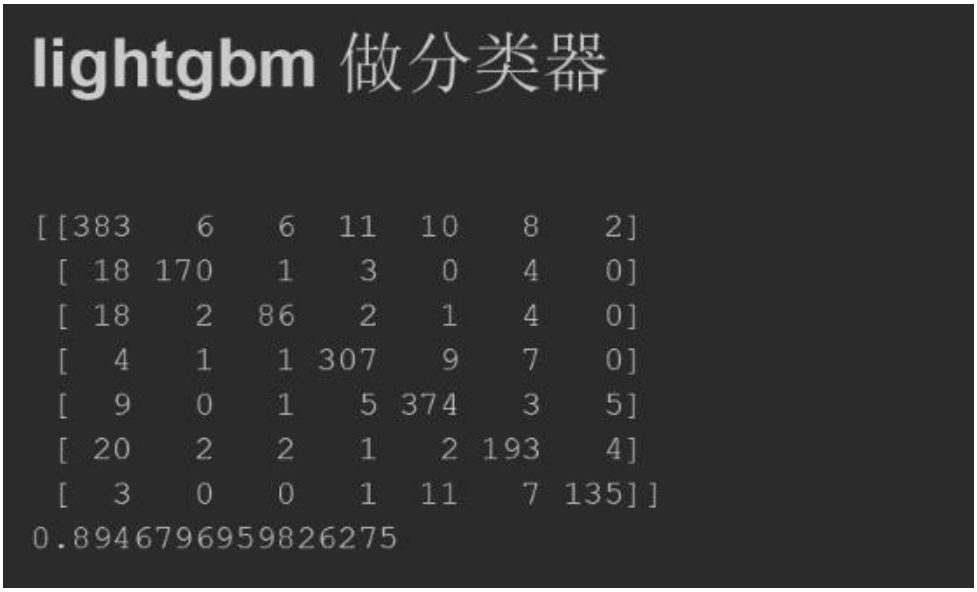

图 4-6 lightGBM 分类结果

4.1.4 非线性 SVM 分类器

基于本题文本向量特性，SVM 分类方法选择非线性。对于输入空间中的非线性分类问题，可以通过非线性变换将它转化为某个维特征空间中的线性分类问题，在高维特征空间中学习线性支持向量机。

其输入维训练数据集 $T=\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$，其中 $x_i\in\Re^n, y_i\in\{+1,-1\}, i=1,2,...,N$；输出为分离超平面和分类决策函数。

最后的分类结果为：

```
Prefix dict has been built successfully.
[[349  10   6   2   8  13   1]
 [ 10 187   0   0   0   1   0]
 [ 11   1 120   0   1   1   0]
 [  7   0   0 305   6   8   1]
 [  8   0   0   6 380   0   4]
 [ 17   2   5   3   4 186   4]
 [  7   0   0   0   8   6 154]]
0.9125950054288816
```

图 4-7 SVM 分类结果

4.1.5 MLP 分类器

多层感知机（MLP）也叫人工神经网络，它包含输入层、输出层和至少一个的隐层，即三层的结构，多层感知机层与层之间是全连接的（全连接的意思就是：上一层的任何一个神经元与下一层的所有神经元都有连接）。多层感知机最底层是输入层，中间是隐藏层，最后是输出层，如图所示：

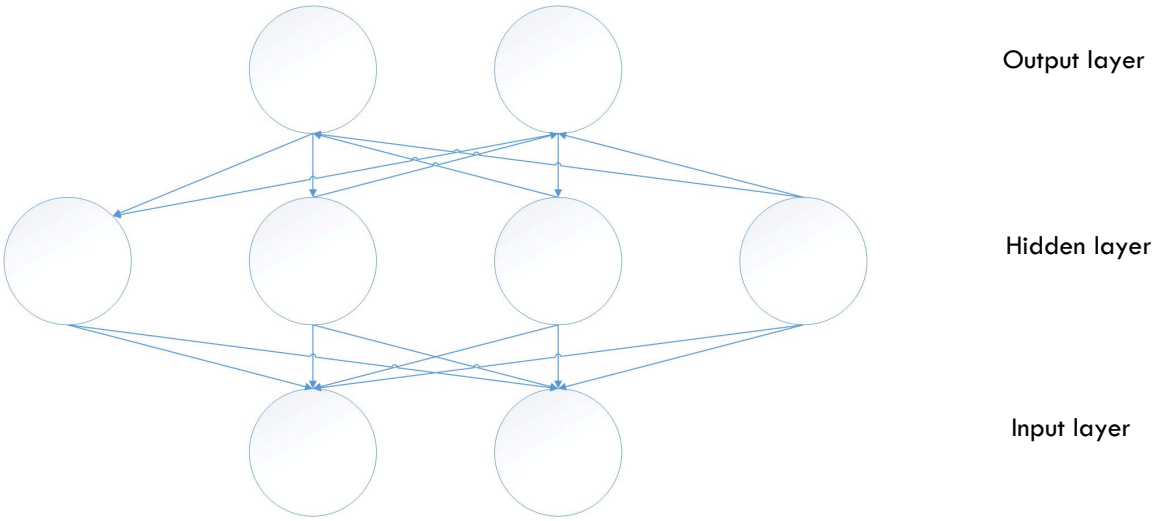

图 4-8 MLP 原理图

输入层输入数据预处理之后的 n 维向量，则有 n 个神经元，设输入层用向量 X 表示，则隐藏层的输出即 $f(W_1X+b_1)$，W_1 是权重，b_1 是偏置，函数 f 可以是 sigmoid 函数或 tanh 函数：

$$\text{sig}moid(a)=\frac{1}{1+e^{-a}}$$

$$\tanh(a)=\frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$$

隐藏层到输出层可以看成一个多类别逻辑回归，即 softmax 回归，所以输出层是 $\text{soft}\max(W_2X_1+b_2)$，$X_1$ 表示隐藏层的输出 f。将上述三层总结起来得：

$$f(x)=G(b^{(2)}+W^{(2)}(s(b^{(1)}+W^{(1)}x)))$$

因此，MLP 的所有参数就是各个层之间的连接权重和偏置，根据本题使用梯度下降法求解最佳参数：

图 4-9 SGD 算法

最后的分类结果为：

图 4-10 MLP 分类结果

4.1.6 模型融合

这里采用模型融合的 Voting 法策略，该投票法针对分类模型，多个模型的分类结果进行投票，少数服从多数。融合模型的目的就是结合不同子模型的长处，已达到互补短处的目的。这里采用了 lightgbm+svm+mlp+xgboost+random forest（voting），由于设备限制，没有调节参数，效果不是很好，最终比较得到单模型效果 mlp 最佳，可以达到接近 0.92 的 F1—score。

4.2 注意事项

本题的分类是任务基于附件 1 的三个级别标签，需要注意的是使用几级标签

可以使分类器更好地分类，更好地解决文本语义相同词语却不同的问题，由于多分类任务又不同于二分类，因此如何选用模型提升多分类的 F1 值是另一难点。此外，附件 2 给出的文本数据长、短文本夹杂，有些长文本对本题无意义的句子过多，以及解决不同类别的文本数不平衡问题，有助于本题 F1 值的提高。

4.2.1 最优算法选择

多标签分类面对每类标签数据样本量不平衡的问题，采用 k 折交叉验证，可以缓解数据量不均衡的问题，采用网格搜索算法进行最优模型参数选择，对于 MLP 的激活函数和最大特征量进行调整，利用 TF-IDF 抽取特征，采用了五个单模型进行分类任务，也尝试了利用投票策略进行模型融合，最终多层感知机+TF-IDF 效果最好。

4.2.2 多分类问题带来的难度

一般使用最多的是二分类模型，基于此题有七个分类标签，提升了分类难度，因此使用多模型融合参与投票进行分类，有助于提高分类准确率，而且增强了数据的鲁棒性。

4.2.3 数据不平衡带来的影响

附件 2 给出了七种分类标签，共 9210 条数据，但各类数据并不平衡，如下图所示：

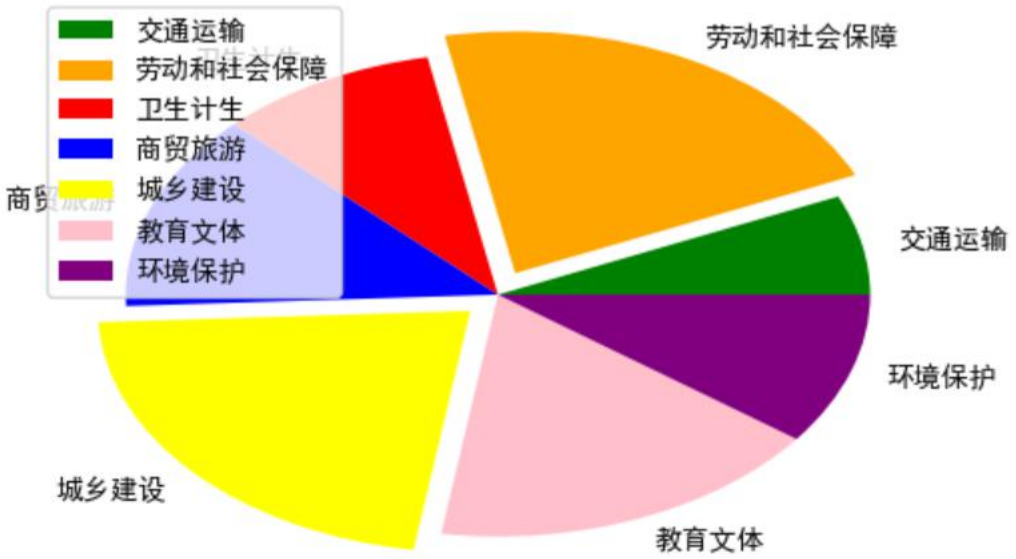

图 4-11 附件 2 数据分布

解决数据不平衡问题有数据增强和交叉验证两种方法，由于做数据增强需要网络爬虫、手动做数据和重复采样等方法，过于繁琐并且容易降低准确率，因此本题使用交叉验证的方法解决数据不平衡问题。

交叉验证，即将原始数据随机分为两组，一组做为训练集，一组做为验证集，利用训练集训练分类器，然后利用验证集验证模型，记录最后的分类准确率为此分类器的性能指标，下图为交叉验证的工作原理：

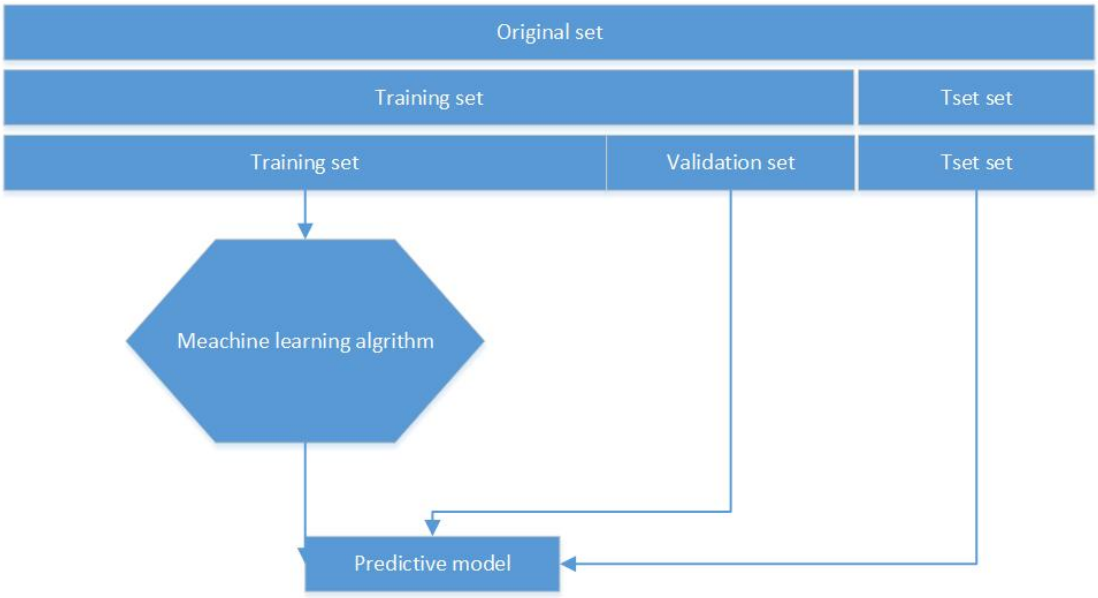

图 4-12 交叉验证工作原理

4.2.4 长文本的无意义表达过多

基于附件 2 的留言详情，其中的内容有长文本也有短文本，仔细阅读长文本可发现其有一大部分是套话、无意义的话，是文本特征的噪声，即干扰项。为解决此问题，使用 TF-IDF 文本的向量表示方法，将特征文本词项依据其重要程度转化为用数字表示的向量，即 n×k 矩阵，将向量的每一行按从大到小的顺序排列，将比较重要的词项的数学表示放在前面，选取前 k 个特征，即可取得长、短文本中重要的词项，将不重要的舍弃。

4.3 F1 值评价

对于一个多分类的任务，我们通常知道哪些分类容易预测，哪些困难。而且我们对于不同的预测的代价也是不同。这里使用混淆矩阵将各个分类器的效果直观表示。

RandomForest 的混淆矩阵：

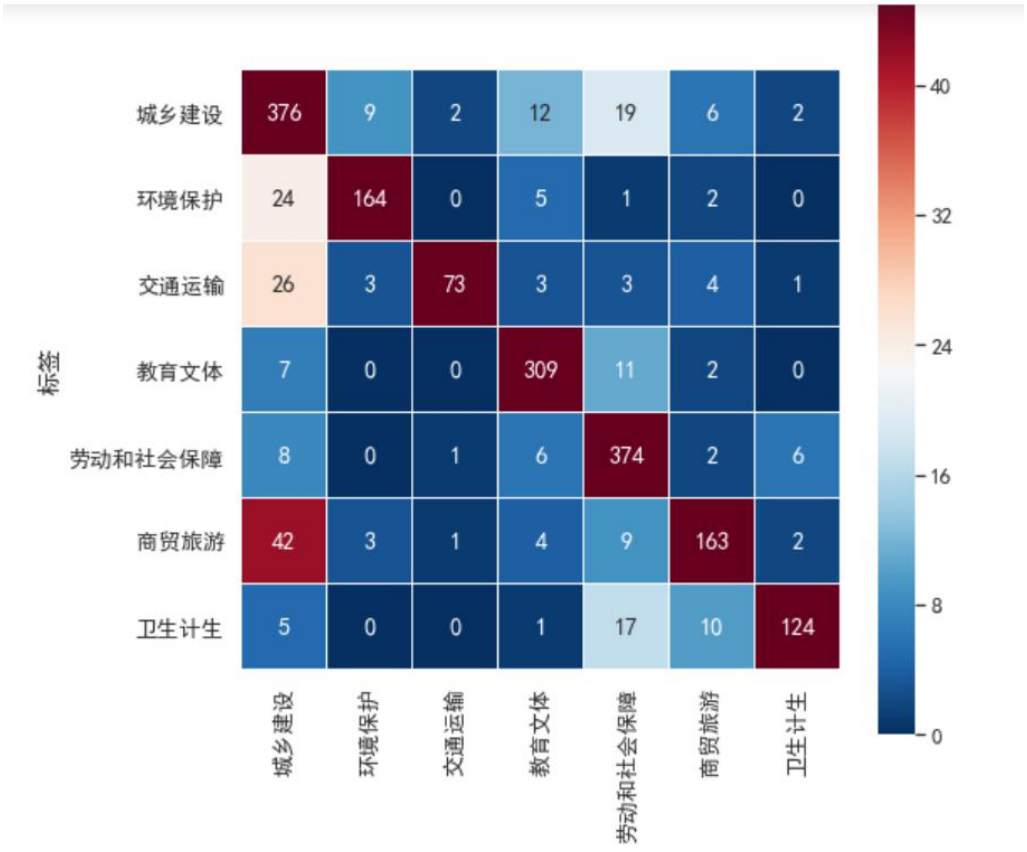

图 4-13 RF 混淆矩阵热力图

XGBoost 的混淆矩阵：

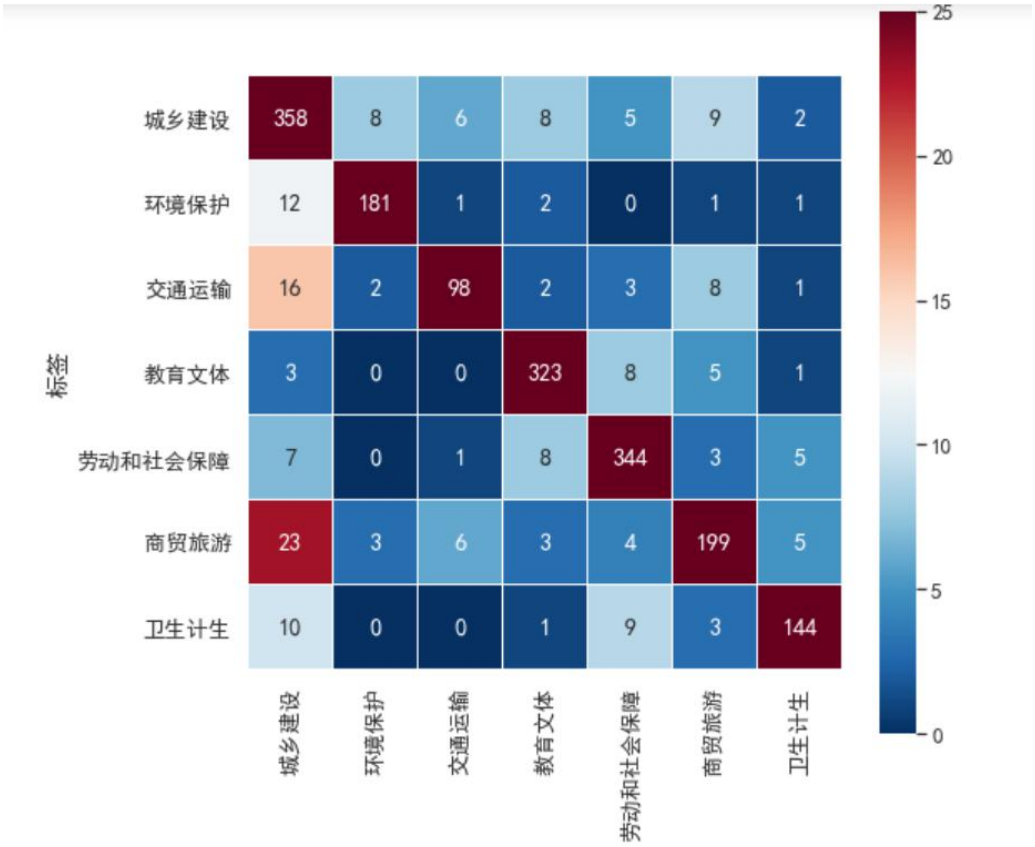

图 4-14 XGBoost 混淆矩阵热力图

lightGBM 的混淆矩阵：

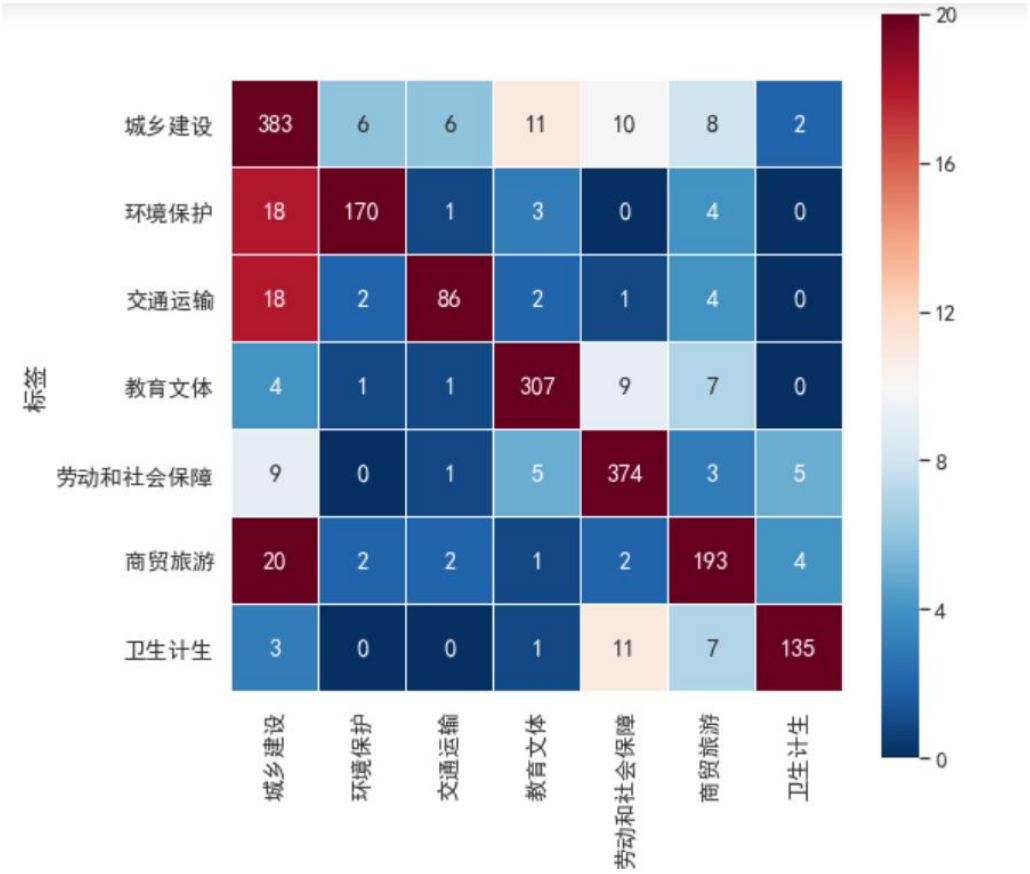

图 4-15 lightGBM 混淆矩阵热力图

SVM 的混淆矩阵：

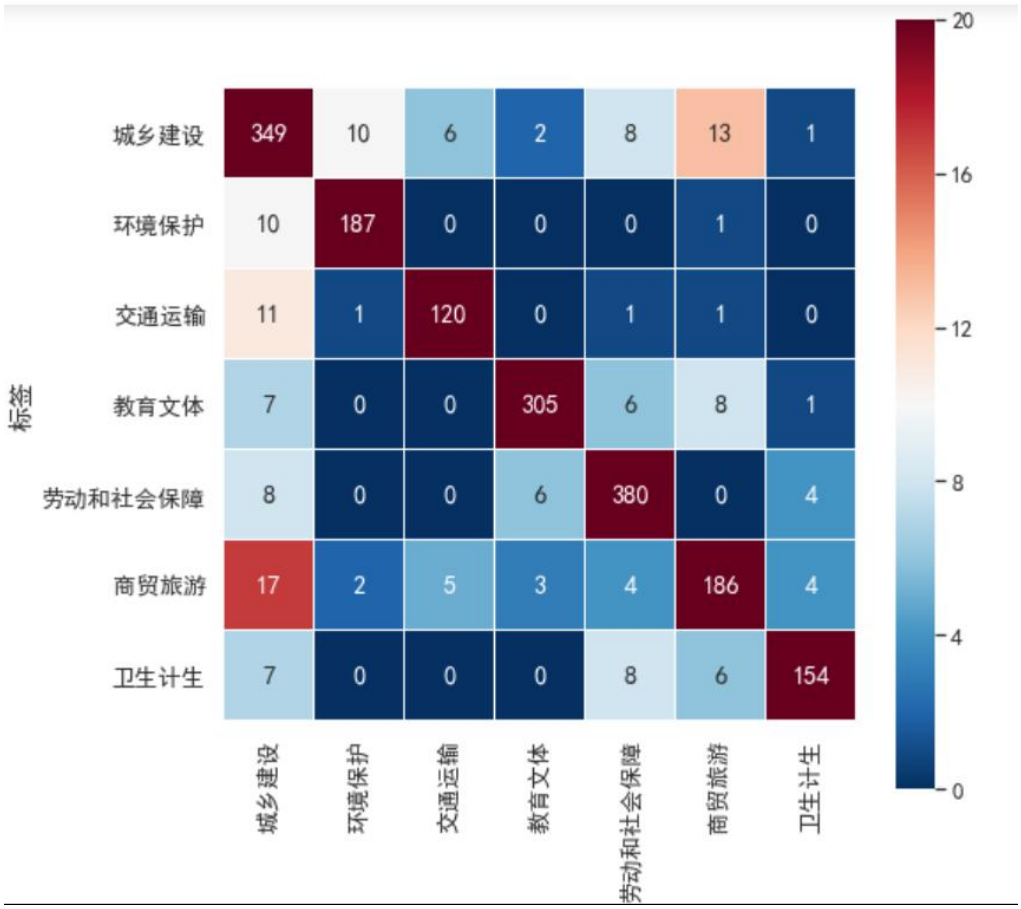

图 4-16 SVM 混淆矩阵热力图

MLP 的混淆矩阵：

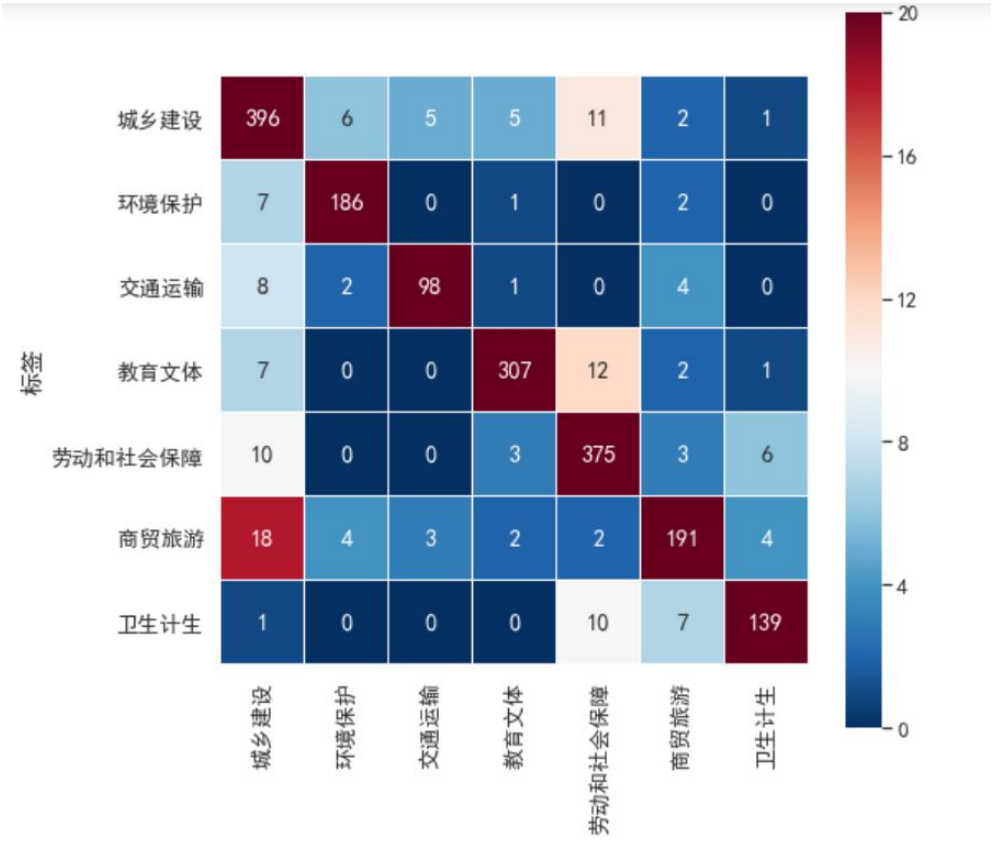

图 4-17 MLP 混淆矩阵热力图

本题为多分类问题，有关多分类问题的 F1 值计算分为 micro 和 macro 两种方法，由于本题数据不平衡，而 F1-micro 方法适用于类别数据不一样的情况，而 F1-macro 并不考虑各类别的数据量多少。令：

TP_{i}：真实类别为正例，预测类别为正例；FP_{i}：真实类别为负例，预测类别为正例；FN_i：真实类别为正例，预测类别为负例；TN_{i}：真实类别为负例，预测类别为负例。则其查准率、查全率和 F1 值的计算公式为：

$$\mathrm{precision}_{\mathrm{mi}} = \frac{\sum_{i=1}^{7} TP_i}{\sum_{i=1}^{7} (TP_i + FP_i)}$$

$$recall_{mi} = \frac{\sum_{i=1}^{7} TP_i}{\sum_{i=1}^{7} (TP_i + FN_i)}$$

$$F_{1,mi} = 2\frac{\mathrm{precision}_{\mathrm{mi}} \times recall_{mi}}{\mathrm{precision}_{\mathrm{mi}} \times recall_{mi}}$$

RandomForest 的 F1 值：

$$F_{1,\mathrm{RandomForest}} = 2\frac{precision_{mi} \times recall_{mi}}{precision_{mi} + recall_{mi}} = 0.89468$$

light GBM 的 F1 值：

$$F_{1,lightGBM} = 2\frac{precision_{mi} \times recall_{mi}}{precision_{mi} + recall_{mi}} = 0.89468$$

XGBoost 的 F1 值：

$$F_{1,XGBoost} = 2\frac{presicion_{mi} \times recall_{mi}}{presicion_{mi} + recall_{mi}} = 0.89414$$

MLP 的 F1 值：

$$F_{1,MLP} = 2\frac{precision_{mi} \times recall_{mi}}{precision_{mi} + recall_{mi}} = 0.91857$$

第 5 章 热点问题挖掘

本题基于第三章数据预处理中 pyHanLP 分词、去除停用词和 TF-IDF 文本表示进行后续操作。根据已有留言详情中特定地点、特定人群的问题提取，并按照附件 3 给出的点赞数和反对数建立一套热度评价指标将整理出的热点问题按照热度指标进行排列。先采用 pyhanlp 的文本聚类，利用肘部法则，然后在再进行命名实体识别（NER），由于采用的是 pyhanlp 语料库里面的地名数据，所以需要自定义词典，添加特殊地名，这样可以提高识别准确率。

整体的流程图如下所示：

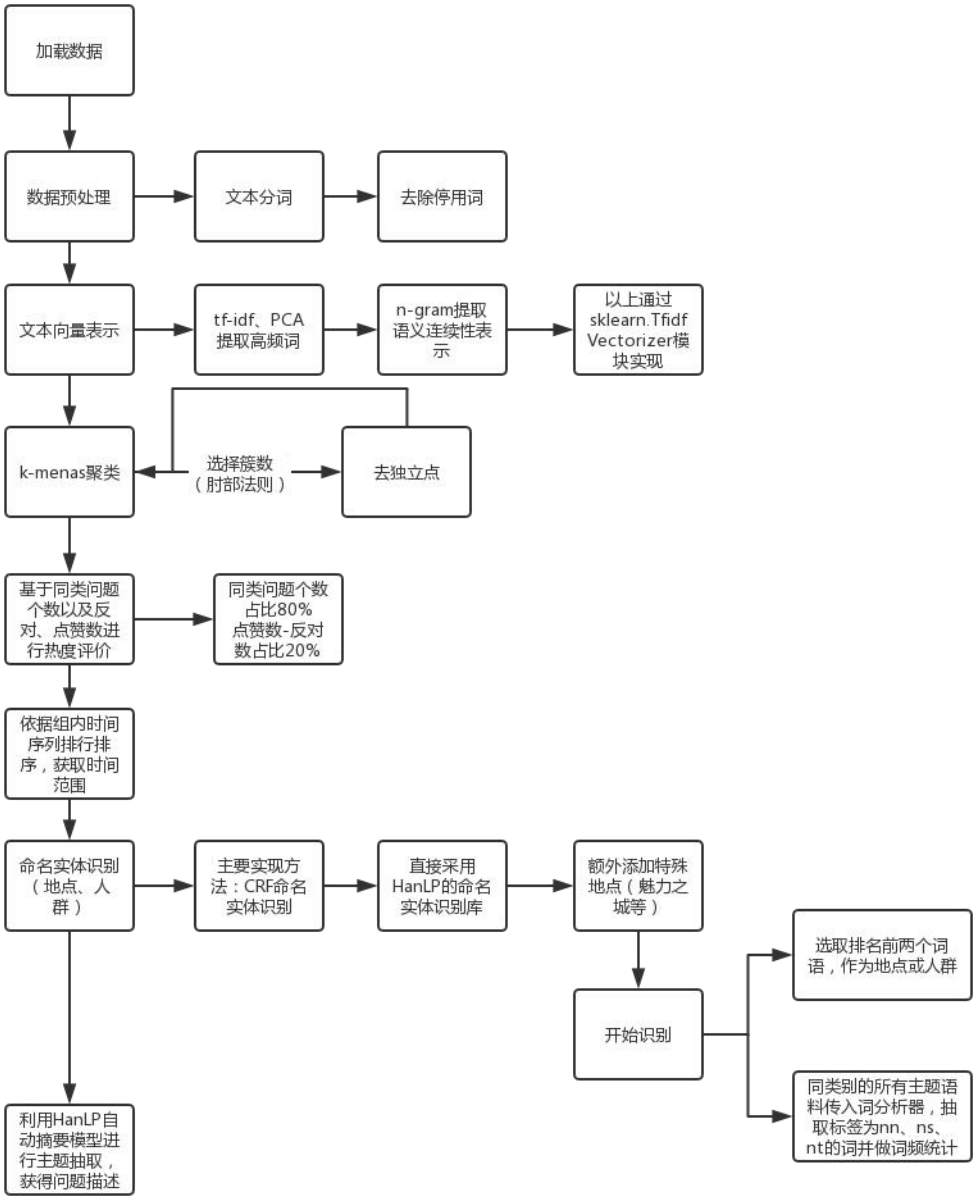

图 5-1 热点挖掘流程图

在进行“识别相似留言”步骤之前，命名实体识别 HanLP 提取和 TF-IDF 提取重要词项已在第三章给出。接下来从第二步中 PCA 特征提取开始研究。

5.1 PCA 特征提取

把文本数据集使用 TF-IDF 算法转化为向量之后，由于文本词项过多导致数据特征较多，因此考虑使用 PCA 降维，即减少数据特征的个数，使每个点到他们对应直线上的投影点的距离，使得距离的平方最小。

PCA 算法首先需要计算协方差$\sum$ ，即：

$$\sum = \frac{1}{\mathrm{m}}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^{T}$$

第二步计算协方差的特征向量，使用 svd（即奇异值分解），最终得到 3 个矩阵，矩阵 U 是我们需要的 n×n 矩阵：

$$[U,S,V] = svd(\sum)$$

第三步，将向量转换为 k 维，即提取前 k 个向量，此时 U 矩阵降到 k 维为：

$$U\_reduce = [u^{(1)},u^{(2)}..u^{(k)}]$$

得到一个 n×k 矩阵，将其转置得到 k×n 矩阵，再乘以特征 X（n×1），最后得到 k 维向量：

$$z^{(1)},z^{(2)}...z^{(k)}$$

本题使用 python 中自带的库训练即可得到 PCA 降维之后的主要特征。

5.2 余弦相似度

因为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，相比距离度量，余弦相似更加注重两个向量在方向上的差异，而非距离或长度上。余弦相似度的计算公式为：

$$\mathrm{simi}larity = \cos(\theta) = \frac{A\cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

在本题中 TF-IDF 算法将文本特征词项转换为向量后，为空间中的点，因此使用余弦相似度，它对绝对的数值不敏感，更多的可以用于留言详情中的关注点来区分相似度和差异，同时修正了投诉用户间可能存在的度量标准不统一的问题。

5.3 相似问题归类

本题不同于第一题已经有分好类的标签，本题需要将没有标签的文本数据集

按照特定地点、特定人群相同的一套标准将相似问题归到一类，因此使用无监督学习的算法，即聚类。基于已有数据集的类别个数不清楚，且每个类别的文本数也不清楚，因此传统的机器学习算法不能应用于此题。

本题采用 pyHanLP 库中的一个聚类算法——repeated bisection 算法，该算法的聚类模块可以接受任意文本作文档，也可以接受列表作为输入；另外，该算法性能和运行速度优于 HanLP 库中的其他聚类算法。

repeated bisection 算法，通过给准则函数的增幅设定阈值 β 来自动判断 k，解决了已有数据集分类个数不确定的问题。此时算法的停机条件为，当一个簇的二分增幅小于 β 时不再对该簇进行划分，即认为这个簇已经达到最终状态，不可再分；当所有簇都不可再分时，算法终止，此时产生的聚类数量就不再需要人工指定了。

5.4 热度评价指标

本文选定的属性是每类留言用户数量、点赞数减去作为评价指标，但由于部分点赞数过高导致不同热点问题之间得分差异过大，故通过调节不同属性的比例来缓解误差。本文对于留言用户数量、点赞数减去反对数的比例为 8：2，由此降低点赞数带来的影响，最终得到一个问题得分的排名。

首先对于点赞数和反对数的公式如下：

$$\mathrm{t}_i = \sum_{j=1}^{n} (U_{i,j} - F_{i,j})$$

其中，t_i 为第 i 类热点问题的点赞和反对的支持度指标，$U_{i,j}$ 为第 i 类热点问题中第 j 个文本的点赞数，$F_{i,j}$ 为第 i 类热点问题的第 j 个文本的反对数；令 s_i 为每类问题的留言数量，最后的热度指标结果为 W_i，则其综合热点评价公式为：

$$W_i = 0.8s_i + 0.2t_i$$

最后排序结果见附件，下表为排名前五的热点问题：

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	0.1400	2019/5/5 至	K9 县 汇金路	A 市 A5 区汇金路五矿万境 K9

		48711	2019/10/10		县存在一系列问题
2	2	0.117478286	2019/2/27 至 2019/12/31	金毛湾 A 市	反映 A 市山水湾孩子上学问题
3	3	0.101993773	2019/1/16 至 2019/7/8	西地省 美国	不要让 A 市因为 58 车贷案件而臭名远扬
4	4	0.054955373	2019/1/7 至 2019/9/15	候家塘派出所 5 区公安分局	请书记关注 A 市 A4 区 58 车贷案
5	5	0.044931231	2019/8/23 至 2019/9/6	临路 绕城	A4 区绿地海外滩二期业主被噪音扰得快烦死了

表 6-1 热点问题

热点明细表如下图所示:

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	208069	A0009443 6	A5 区五矿万境 K9 县的开发商与施工方建房存在质量问题	2019-05-05 13:52:50	本人是 A5 区洞井街道汇金路五矿万境 K9 县 25 栋的业主，本人要投诉五矿地产及施工方 1、五矿万境 K9 县宣传作为五矿地产与 A 市 A5 区政府合力打造的七大重点项目之一，项目介绍（来源五矿万境 K9 县公众号），做为重点项目，其房屋质量骑在堪忧，宣传与政府打造的重点项目，A5 区政府是否参与了一定的监督？2、24、25 栋与 2017 年 12 月 30 日交房，交房时水电未通正式水电，至今交房一年半，还是未通正式水电，导致家中电器一致处于低电压或电压不稳中运营，导致电器损坏，入住不到一年，期间多次停电停水，扰乱了业主正常生活秩序，更可怕的期间因电路问题，导致车库起火（详情看图），庆幸火势不大，未造成人员伤亡和财产损失。3、房屋质量问题：渗水、开裂；交房不到一年，入住不到半年，承重墙就发现大面积裂缝（电视背景墙、主卧、次卧都属于承重墙体-详情看图），还有大部分业主都反馈家中渗水，导致漆面发霉！这种房屋质量是否符合与政府打造的重点项目？！这种房屋质量是否能符合五矿世界 500 强的地位？！这种房屋质量是否能符合五矿在西地省的龙头企业？！这种房屋质量是否符合五矿央企的性质？！4、25 栋负一楼车库问题；第一点：一楼与负一楼（车库）无楼梯连接，只能通过电梯及户外平台一处公用楼梯上下，这样的设计是否通过消防检查？第三点提到的车库起火，当时本人就在地下车库，起火导致停电电梯停运，无法通过电梯通往一楼，只能通过户外的楼梯通往一楼，作为业主我也是才知道那个楼梯通往一楼，如是外来人员找不到出口不会被活活呛死？第二点：车库漏水问题，以为是简单的滴水吗？ 不是是漏水！是想瀑布一样的漏水，漏就漏水，知道水是漏在哪吗？？漏在电箱上，供应这整栋楼的电	0	2

					箱上！，去年整改一回，说整改好了，前几天 A 市大雨，又开始漏了！！是不是漏水导致大量人员触电身亡，开发商才会重视？政府才会派人督促？如果发生了这样的事情,那逝去的生命谁来祭奠？5、以上所有问题，均通过 12345 市长热线、物业等途径反馈，最主要的问题是开发商家大业大，对业主采用拖字诀，爱理不理，业主也没任何途径与开发商沟通上！		
1	2086 36	A0 00 77 17 1	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019-08-19 11:34:04	我是 A 市 A5 区汇金路五矿万境 K9 县 24 栋的一名业主,我们小区一开始的定位是一个高端别墅小区，实行人车分流管理，物业也标榜 37 度五星级服务,到目前为止小区群租房泛滥成灾，有直接十几个工人租住毛坯房的,也有二房东将毛坯房租过来擅自改成好多格子间，加装 n 块电表的，也有毛坯房直接搬煤气罐入户做饭的，有穿个内裤到处晃悠的，有没装烟机打开门做饭导致整个楼道烟味呛鼻的，我们小区高层大部分都是刚需买房，或者是因为结婚，或者是因为给孩子读书，这些群租房的存在给我们的安全带来极大的隐患以及不确定性,就好像埋在小区的不定时炸弹,说不定啥时候就爆了。为什么 A 市处理了那么多群租房，而我们小区的群租房一直是无人问津的状态，多次投诉，社区说管不了，物业说管不了，12345 也说管不了，到底有谁能来管管，有谁能保证我们人民群众的安居？难道非得等群租房起火爆炸了或者是出现了严重的治安事故才有人来管么？为什么这么大的隐患存在投诉无数次都无人问津无人处理的情况,每个部门都在推卸责任，三房两厅被改成五个单间的格子间，请问是谁批准的，这一波二房东有组织有计划的侵占我们业主的正当权益，给我们业主埋下了炸弹请问是谁允许的？我们小区也曾发生过狗咬人,请问有人对养宠物的情况进行过排查吗？我们小区黄谷路路口每天车辆通行不计其数，数量非常庞大，然而孩子们上学这是必经的路口，每天大人过马路都胆战心惊的，更何况孩子呢？为什么这样一个危险的路口没有人行天桥或者地下通道？五矿万境 K9 县小区作为一个封闭式的高端小区现在物业管理及其混乱，小区门禁几乎没用，保安见人就开门从不询问，跟物业投诉过多次，却始终无果，小区的保洁也是一直跟物业投诉然后也是一直无果，对于一个这样的物业，收取着高额的物业服务费却几乎没尽到物业该尽的责任,这样一个物业谁评的五星？谁能来监管,谁能保证我们业主的基本安全和卫生？关于之前一直强调的人车分流的小区,现在小区儿童的休闲区域各种快递车外卖车以及业主的电动车都往上面开,孩子的安全谁来保障？小区将电动摩托车充电桩安装在地下室中间,地下室阴暗潮湿视线不好制动也不好，极容易发生交通事故，这样安排合理？电摩充电一旦发生火灾小区的车辆是不是都要遭殃？24 栋 25 栋每个单元消防楼	0	2097

					梯都只到一楼,在负一楼和一楼中间唯一的楼梯在小区地下车库中间，然后一旦发生火灾地下室的人该如何逃生？盼请回复，还正政府出面帮我们解决这些实际问题,只有让我们安居了我们才能乐业。		

表 5-2 热点明细表

第 6 章 答复意见的评价

6.1 相关性

常见的文本向量化方法有**向量空间模型**(Vector Space Model,VSM)、**布尔模型**(Boolean Model,BM)和**统计语言模型**(Statistical Language Model,SLM)。VSM 是一种词袋模型,它将每一个文本看作是装在袋子里的词,忽略了其中的顺序，采用**词频-逆文档频率**(Term Frequency-Inverse Document Frequency,TF-IDF)来计算权重，成为目前比较流行的文本表示方法。

Doc2bow 方法主要用于实现典型的词袋模型,忽略了答复语句的语序和语法,将答复文本看作是独立单词的集合,互不干扰，dow2bow 中用一组无序的单词来表示答复文本。在本题中,将分词后的结果集放入到构建的字典对象中,并建立字典中单词与编号之间的映射,在计算完每一个单词的词频后,使用 doc2bow 来统计所有的词频,并以稀疏矩阵的形式输出。

LDA（Latent Dirichlet Allocation）是一种概率主题模型，在 2003 年由 Blei,David M 等人提出。LDA 是一种典型的词袋模型,其核心思想认为文档是由词构成的一个集合,而文档中的词是没有顺序的独立个体,一篇文档中可以包含多个主题,每一个主题又由词生成。LDA 中词生成过程的数学表达式如式所示：

$$\boldsymbol{P(w|d) = P(w|t) \times P(t|d)}$$

其中,$P(w|d)$表示答复意见 d 生成词语 w 的概率;$P(w|t)$表示词语 w 在主题 t 下的概率;$P(t|d)$表示生成话题 t 的概率。在主题向量的生成过程中,假设每个主题的 $P(wld)$相等,则选择 $argmax(P(wld))$,即使词语生成概率最大的主题作为该词所属主题,累加后可得到新的主题向量 $P(tld)$。根据新的主题向量对

$P(wld)$进行更新迭代,最终可以得到稳定的主题向量。

余弦距离（Cosine Distance）使用两个向量夹角的余弦值作为衡量两个个体间差异的大小。欧氏距离衡量空间各点的绝对距离，跟各个点所在的位置坐标直接相关；而余弦距离衡量的是空间向量的夹角，更加体现在方向上的差异，而不是位置。相比欧氏距离，余弦距离更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题。本题使用余弦距离计算留言与答复意见的相似度，余弦相似度计算公式如下:

$$sim(i,\ p) = \frac{\sum_{j=1}^{n} i_j \times p_j}{\sqrt{\sum_{j=1}^{n} i_j^{\,2}} \times \sqrt{\sum_{j=1}^{n} p_j^{\,2}}}$$

其中,$sim(i,p)$表示留言 i 和答复 p 之间的相似度,通过计算留言 i 和答复 p 之间对应特征向量的余弦夹角来判断两者之间的相似度,余弦值越接近 1,夹角就越接近 0 度,两个向量则越相似。

1.LDA 关键步骤：

符号说明

符号	解释
$\boldsymbol{S_W}$	类内散布矩阵
$\boldsymbol{S_B}$	类间散布矩阵
$\boldsymbol{n_i}$	属于 i 类的样本个数
$\boldsymbol{m}$	所有样本的均值
$\boldsymbol{x_i}$	第 i 个样本
$\boldsymbol{D_i}$	第 i 类样本
$\boldsymbol{m_i}$	第 i 类样本的均值
$\boldsymbol{c}$	样本类别数

表 6-1 符号说明

（1）对 d 维数据进行标准化处理（d 为特征数量）

（2）对于每一类别，计算 d 维的均值向量

$$m_i = \frac{1}{n_i} \sum_{x \in \boldsymbol{D_i}}^{c} x_m$$

（3）构造类间散布矩阵$\boldsymbol{S_B}$以及类内散布矩阵$\boldsymbol{S_W}$

$$S_w = \sum_{i=1}^{c} S_i$$

$$S_i = \sum_{x \in \boldsymbol{D_i}}^{c} (x - m_i)(x - m_i)^T$$

$$S_B = \sum_{i=1}^{c} (m - m_i)(m - m_i)^T$$

（4）计算矩阵$\boldsymbol{S_W}^{-1}\boldsymbol{S_b}$的特征值以及对应的特征向量

（5）选取前 k 个特征值所对应的特征向量，构造一个 $d \times k$ 维的转换矩阵 W,其中特征向量以列的形式排列

（6）使用转换矩阵 W 将样本映射到新的特征子空间上

2.算法流程

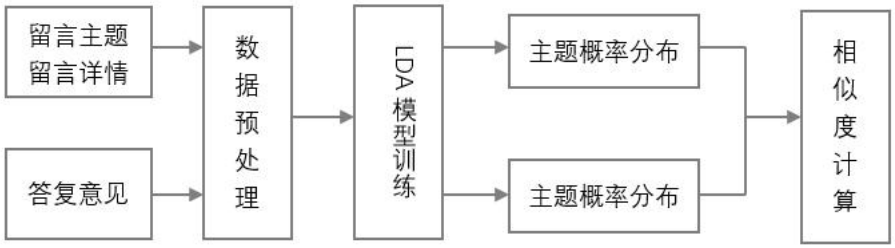

图 6-1 LDA 算法流程

3.代码实现思路

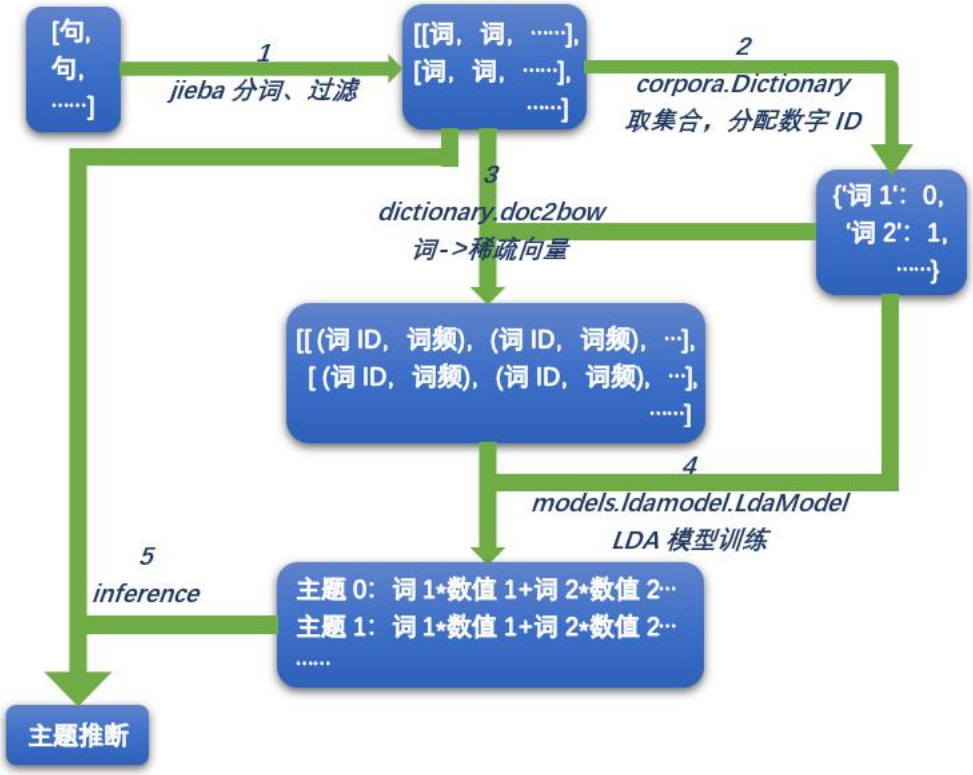

图 6-2 LDA 代码实现思路

4. 以第一条记录为例：

Step1：导入数据。

留言主题	留言详情	答复意见
A2 区景蓉华苑物业管理有问题	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，该小区的物业公司美顺物业扬言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了(原水电在小区买，水 4.23 一吨，电 0.64 一度) 所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交 20 万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各	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”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2 区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于 2019 年 4 月 10 日至 4 月 27 日以“意见收集方式”召开了业主大会，经业委会统

业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪?	计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5 月 5 日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。

表 6-2 数据示例

Step2：数据预处理。首先合并留言详情和留言主题，便于后续与答复意见比较。接着，运用正则表达式，只匹配出英文、数字和汉语。其次，jieba 分词将每段话构成一个单词列表；最后，filter 过滤出需要单词。得到的结果如下所示：

留言详情+留言主题：['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', '的', '物业公司', '美顺', '物业', '扬言', '要', '退出', '小区', '因为', '小区', '水电', '改造', '造成', '物业公司', '的', '高昂', '水电费', '收取', '不了', '原', '水电', '在', '小区', '买水', '423', '一吨', '电', '064', '一度', '所以', '要', '通过', '征收', '小区', '停车费', '增加收入', '小区', '业委会', '不知', '处于', '何种', '理由', '对', '该', '物业公司', '一再', '挽留', '而', '对', '业主', '提出', '的', '新', '应聘', '的', '物业公司', '却', '以交', '20', '万', '保证金', '不能', '提高', '收费', '的', '苛刻', '条件', '拒之门外', '业委会', '在', '未', '召开', '全体', '业主大会', '的', '情况', '下', '制定', '了', '一', '高昂', '收费', '方案', '要', '各', '业主', '投票', '而', '投票', '不', '采用', '投票箱', '只', '制定', '表格', '要', '物业公司', '人员', '这一', '利害关系', '机构', '负责', '组织', '对', '投票', '业主', '隐私权', '没有', '任何', '保护', '还', '对', '投', '反对票', '的', '业主', '以', '领导', '做', '工作', '等', '方式', '要求', '改变', '为', '同意', '票', '这种', '

投票', '何来', '公平', '公正', '公开', '面对', '公安干警', '采用', '这种', '方式', '投票', '合法性', '在', '哪', 'A2', '区景蓉华苑', '物业管理', '有', '问题']

答复意见：['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', '调查核实', '情况', '向', '该', '网友', '答复', '如下', '您好', '首先', '感谢您', '对', '我们', '工作', '的', '信任', '和', '支持', '关于', '您', '在', '平台', '栏目', '给', '胡华衡', '书记', '留言', '反映', 'A2', '区景蓉', '花苑', '物业管理', '有', '问题', '的', '情况', '已', '收悉', '现将', '我们', '调查', '处理', '情况', '答复', '如下', '经', '调查', '了解', '针对', '来信', '所', '反映', '的', '小区', '停车', '收费', '问题', '景蓉华苑', '业委会', '于', '2019', '年', '4', '月', '10', '日至', '4', '月', '27', '日以', '意见', '收集', '方式', '召开', '了', '业主大会', '经', '业委会', '统计', '超过', '三分之二', '的', '业主', '同意', '收取', '停车', '管理费', '在', '业主大会', '结束', '后', '业委会', '也', '对', '业主', '提出', '的', '意见', '和', '建议', '进行', '了', '认真', '疏理', '归纳', '并', '进行', '了', '反馈', '业委会', '制定', '的', '停车', '收费', '标准', '不', '高于', '周边', '小区', '价格', '针对', '来信', '所', '反映', '的', '物业公司', '去留', '问题', '5', '月', '5', '日', '下午', '辖区', '桂花', '坪', '街道', '牵头', '组织', '社区', '物业公司', '业主', '委员会', '业主', '代表', '的', '会议', '区', '住房', '和', '城乡', '建设局', '也', '参加', '了', '会议', '在', '综合', '各', '方面', '的', '意见', '后', '辖区', '桂花', '坪', '街道', '区', '住房', '和', '城乡', '建设局', '已', '要求', '业委会', '依法', '依规', '召开', '业主大会', '根据', '业主大会', '的', '表决', '结果', '再', '执行', '相应', '的', '程序', '再次', '感谢您', '对', '我区', '工作', '的', '理解', '和', '关心', '2019', '年', '5', '月', '9', '日']

Step3：文本的向量化表示。首先，引入 corpora 模块里面的 Dictionary 类，对该语料库建立字典。接着，基于 doc2bow 方法对不同词频的单词进行统计，形成稀疏向量。最后，将整个语料库基于 TF-IDF 加权表示文本，将数值映射到[-1，1]之间，完成对文本的向量化抽取表示。

Dictionary:{'G': 0, '为了': 1, '了': 2, '党和政府': 3, '创业': 4, '发展': 5, '和': 6, '固定': 7, '场所': 8, '城乡': 9, '实在': 10, '小摊': 11, '小贩': 12, '工作': 13, '市': 14, '我们': 15, '摆摊': 16, '残疾人': 17, '没有': 18, '生活': 19, '由于': 20, '的': 21, '社会': 22, '社会青年': 23, '给': 24, '给与': 25, '请求': 26, '一下': 27, '也': 28, '予以': 29, '他们': 30, '你': 31, '你们': 32, '你好': 33, '公正执法': 34, '十九': 35, '双百': 36, '双联': 37, '反馈': 38, '可以': 39, '城市': 40, '大': 41, '安排': 42, '对': 43, '市场': 44, '希望': 45, '并': 46, '应当': 47, '开展': 48, '当地': 49, '想法': 50, '所说': 51, '摊点': 52, '整理': 53, '文明执法': 54, '是': 55, '澄清': 56, '研究': 57, '禁止': 58, '第二': 59, '管理': 60, '精神': 61, '经营': 62, '要': 63, '这个': 64, '进入': 65, '进行': 66, '首先': 67}

Step4：运用 LDA 主题模型进行训练。

第一条留言的 sentences：['的', '投票', '小区', '物业公司', '要', '业主', '对

', 'A2', '在', '而', '采用', '业委会', '高昂', '区', '这种', '收费', '方式', '了', '制定', '水电', '的', '小区', '物业公司', '投票', '对', '业主', '在', '要', 'A2', '制定', '这种', '方式', '而', '了', '该', '业委会', '高昂', '水电', '采用', '区', '的', '小区', '物业公司', '要', '投票', '在', '业主', '对', 'A2', '收费', '水电', '区', '该', '了', '高昂', '采用', '制定', '这种', '方式', '业委会', '的', '小区', '物业公司', '要', '投票', '业主', '对', 'A2', '方式', '了', '水电', '该', '区', '这种', '业委会', '采用', '高昂', '在', '而', '制定', '的', '小区', '物业公司', '要', '业主', '投票', '对', 'A2', '在', '该', '业委会', '而', '高昂', '水电', '区', '方式', '采用', '制定', '收费', '了']

第一条回复的 sentences：['的', '业主', '了', '业委会', '和', '问题', '反映', '月', '业主大会', '也', '停车', '在', '对', '5', '意见', '所', '情况', '来信', '区', '进行', '的', '业委会', '和', '业主大会', '反映', '了', '意见', '月', '问题', '对', '在', '5', '业主', '年', '停车', '经', '情况', '调查', '坪', '收费', '的', '月', '和', '业委会', '业主', '在', '问题', '了', '5', '反映', '情况', '对', '停车', '业主大会', '意见', '建设局', '住房', '物业管理', '所', '向', '的', '反映', '业主', '了', '业主大会', '业委会', '在', '和', '问题', '停车', '意见', '对', '月', '4', '5', '来信', '情况', '区景蓉', '现将', '住房', '的', '和', '业主大会', '业委会', '了', '反映', '在', '问题', '业主', '意见', '停车', '月', '情况', '对', '5', '平台', '书记', '物业公司', 'A2', '小区']

Step5：余弦相似度计算。

第一条记录留言与答复的相似度为 0.391074561685699

6.2 完整性

句法分析是自然语言处理（natural language processing, NLP）中的关键底层技术之一，其基本任务是确定句子的句法结构或者句子中词汇之间的依存关系。

句法分析分为**句法结构分析**（syntactic structure parsing）和**依存关系分析**(dependency parsing)。以获取整个句子的句法结构或者完全短语结构为目的的句法分析，被称为**成分结构分析**（constituent structure parsing）或者**短语结构分析**（phrase structure parsing）；另外一种是以获取局部成分为目的的句法分析，被称为**依存分析**（dependency parsing）。本题即使用句法结构分析判断答复意见的完整性。

句法结构分析，识别句子的主谓宾、定状补，并分析各成分之间的关系。通过句法结构分析，分析语句的主干，以及各成分间关系。对于复杂语句，仅仅通过词性分析，不能得到正确的语句成分关系。句法结构分析的标注如下：

关系类型	Tag	Description	Example
动宾关系	VOB	直接宾语，verb-object	我送她一束花（送-->花）
主谓关系	SBV	subject-verb	我送她一束花（我<--送）
状中结构	ADV	adverbial	非常美丽（非常<--美丽）
定中结构	ATT	attribute	红苹果（红<--苹果）

表 6-3 句法结构分析标注

HanLP 是一系列模型与算法组成的 NLP 工具包，目标是普及自然语言处理在生产环境中的应用。HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点;能提供词法分析（中文分词、词性标注、命名实体识别）、句法分析、文本分类和情感分析等功能。本题留言与答复的完整性基于 HanLP 实现。

列号	列名	含义
1	ID	当前词在句子中的序号
2	FROM	当前词语或标点
3	LEMMA	当前词语（或标点）的原型或词干，在中文中，此列与 FORM 相同
4	CPOSTAG	当前词语的词性（粗粒度）
5	POSTAG	当前词语的词性（细粒度）
6	FEATS	句法特征，在本次评测中，此列未被使用，全部以下划线代替。
7	HEAD	当前词语的中心词
8	DEPREL	当前词语与中心词的依存关系

表 6-4 符号说明

以第一条记录的部分答复意见内容为例：

HanLP.parseDependency("您好，首先感谢您对我们工作的信任和支持……")，得到的结果将句子按句法结构分词。

制定完整性评价规则。关于礼貌的判断，若得到的结果存在礼貌用语，则加20分。关于句法结构的判断，若存在动宾关系或主谓关系，则加30分；若存在状中结构或定中结构，则加10分。但是，若不存在主谓关系，关于句法结构的加分直接为0。根据该完整性评价规则，第一条记录的答复意见有80分，以1为量纲后的得分为0.8。

6.3 可解释性

在可解释性的评价规则中，人工创建规则的语料库，例如['根据'，'《.+》'，'法律','完善','调整','提高','核实','要求','管理','收悉','经查','调查','处理']。根据语料库中的关键词，判断答复意见是否“引经据典”，使答复有据可依。

6.4 层次分析法评价模型

符号	解释	详细说明
M	目标层	本题目标层为最佳答复意见
C	准则层	$C_i(i=1,2,3)$分别表示相关性、完整性、可解释性
P	方案层	$P_i\left(i=1,2,3,\cdots\right)$表示各条答复意见
λ_{max}	最大特征值	判断矩阵（成对比较矩阵）的最大特征值
CI	一致性指标	$CI=(\lambda_{max}-n)/(n-1)$
RI	平均随机一致性指标	
CR	一致性比例	$CR=CI/RI$

表 6-5 层次分析法符号说明

层次分析法的具体步骤：

Step1：**建立答复意见评价指标，构建层次结构图。**本题将问题分为三个层次，最上层为目标层M，即最佳答复意见(O)，最下层为方案层P，即各条答复意见$P_i\left(i=1,2,3,\cdots\right)$，中间层为准则层$C$，包括相关性$C1$、完整性$C2$、可解释性$C3$三个指标，并构建出了层次结构图。

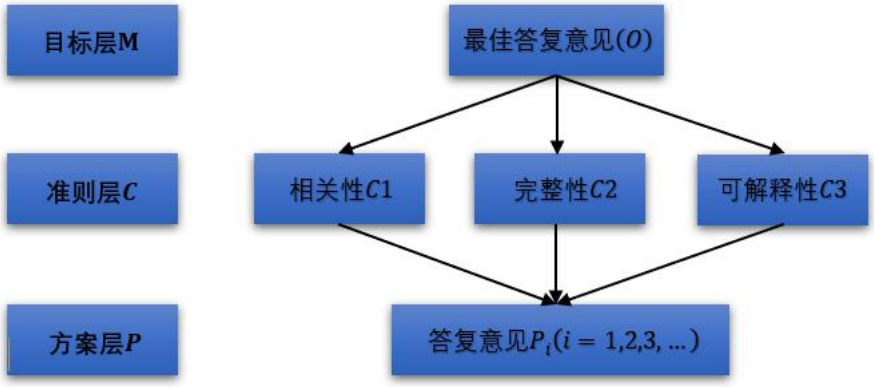

图 6-3 层次结构图

Step2：**构造判断矩阵。**通过查询网络资料，将准则层 C 的三个元素 $C1$、$C2$、$C3$ 两两比较，得到判断矩阵 $M-C$。

M	$C1$	$C2$	$C3$
$C1$	1	2	3
$C2$	1/2	1	1
$C3$	1/3	1	1

表 6-6 判断矩阵

Step3：**一致性检验。**在使用判断矩阵求权重前，必须进行一致性检验。一致性比例的公式如下：

$$CR = CI/RI$$

其中 CI 为一致性指标，计算公式如下：

$$CI = \frac{(\lambda_{max} - n)}{(n-1)}$$

RI 为平均随机一致性指标，可查表 6·4·2 得到。

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49

表 6-7 平均随机一致性指标

如果 CR<0.1，则可认为判断矩阵的一致性可以接受，否则需要对判断矩阵进行修正。

Step4：**计算权重。**为了保证结果的稳健性，分别用算术平均法、几何平均

法、特征值法求权重。使用三种方法计算权重能避免采用单一方法产生的误差，使结果更全面、更有效。

假设判断矩阵 $M-C$=$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & 2_{23} \\ a_{31} & a_{32} & a_{32} \end{bmatrix}$

（1）算术平均法的权重向量公式为

$$\omega_i = \frac{1}{3}\sum_{j=1}^{3}\frac{a_{ij}}{\sum_{k=1}^{3} a_{kj}} \quad (i = 1,2,3)$$

（2）几何平均法的权重向量公式为

$$\omega_i = \frac{\left(\prod_{j=1}^{3} a_{ij}\right)^{\frac{1}{3}}}{\sum_{k=1}^{3}\left(\prod_{j=1}^{3} a_{kj}\right)^{\frac{1}{3}}} \quad (i = 1,2,3)$$

（3）特征值法求权重：

第一步，求出判断矩阵 $M-C$ 的最大特征值λ_{max}及其对应的特征向量；

第二步，对求出的特征向量进行归一化即得到所需权重。

权重	算术平均法	几何平均法	特征值法
$C1$	0.5485	0.5499	0.5499
$C2$	0.2409	0.2402	0.2402
$C3$	0.2106	0.2098	0.2098

表 6-8 权重比

Step5：评价答复意见。由于三个指标量纲相同，可直接根据指标权重求出答复意见的得分进行评价。例如，第一条答复意见在算术平均法的权重下，得分为 0.5485*0.3911+0.2409*0.7273+0.2106*0.8000=0.5582，在几何平均法的权重下，得分为 0.5499*0.3911+0.2402*0.7273+0.2098*0.8000=0.5576，在特征值法的权重下，得分为 0.5499*0.3911+0.2402*0.7273+0.2098*0.8000=0.5576，取三者平均值，最终得分为 0.5578。

参考文献

[1]吴疆,刘欢,董婷.基于标准数据集的分类器融合学习模型[J].微型电脑应用.

[2]邱德钧,冯霞.谓词逻辑视角下 HanLP 中文分词中对歧义的处理[J].科学经济社会.

[3]陈曙东,欧阳小叶.命名实体识别技术综述[J/OL].无线电通信技术.

[4]刘京麦野. 基于循环神经网络的语义完整性分析[D].湘潭大学.

[5]刘宇鹏,栗冬冬.基于 BLSTM-CNN-CRF 的中文命名实体识别方法[J/OL].哈尔滨理工大学学报.

[6]衡宇峰,李俊,彭望龙,黄元稳,房冬丽.基于语义分析的政策法规智能审核研究与实现[J].通信技术.

[7]Romain Atangana, Daniel Tchiotsop, Godpromesse Kenne, et al. EEG Signal Classification using LDA and MLP Classifier.

[8]艾楚涵，姜迪，吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术.