基于机器学习的政务信息分类模型

摘 要

本文旨在利用机器学习的分类算法，构建对“留言”和“答复”的分类模型。并基于统计方法构建评价指标。

任务一要求对留言进行分类，先对留言详情文本数据进行 TF-IDF 特征值提取，之后采用 SGD（随机梯度下降）构建分类模型。为提高分类效果，通过 EDA 增强数据，减少了数据不平衡对分类模型的影响，使用了小样本搜索最优参数和大样本局部调优的优化方式，减少了 SGD 算法最佳参数组合的搜索时间，使得 F-score 达到 0.99804，准确度达到 0.99837，接近理想的分类性能。

任务二要求对留言描述的问题进行聚类，找出热点问题并给出相应的热点评价模型。采用 single-pass 聚类算法，其中对于文本的相似度，采用余弦距离法测定。之后通过不断调整阈值直至 0.1，将留言分为 1205 个主题数。然后基于归一化后的留言数、点赞数、反对数、时间跨度和留言密度 5 个数据构建热度评价指标。按照热度评价指标大小，取出前五个主题，并用 python 的 textrank4zh 进行关键词句的提取，概况出对主题的简单描述。

任务三要求对留言的回复内容做出评价。这里从相关性与满意度两个方向，构建了与留言的相关性、时间满意度和内容满意度 3 个指标，其在评价“答复意见”的质量上有很强的参考价值。对相关性分析上采用潜在狄利克雷分布（LDA）和 Hellinger 距离，适合长文本数据的处理，得到比其他方法更好的效果，大部分留言与回复的相关度在 50%以上。之后根据回复的时间快慢和内容对用户的可参考性提出时间满意度和内容满意度，并在内容满意度得到分类结果达到 F-score 为 0.75436，准确度为 0.93971 的结果。

关键词：EDA；网格搜索；SGD；Single-pass；

目录

一、 背景

近年来由于互联网的不断发展，政府部门也在互联网上推行了自己的网络问政平台，并逐渐成为政府了解民意、凝聚民心的重要渠道。但相关政府部门也面临着数据量激增而带来的工作量和划分准确率的问题。随着人工智能、云计算和大数据等技术的发展，这些技术可以助力于政府部门，提升政府部门的管理水平和施政效率，从而提升人民的满意感和幸福感。

1.1 要解决的问题

(一)根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

(二)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。

(三)针对附件 4 相关部门对留言的答复意见，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。

二、 数据预处理

2.1 去除噪声

未经处理的文本数据自带一些对分析无用的特殊字符，还有一些字符会在不同字符编码格式互相转换时无法识别。对这些特殊字符予以删除。

2.2 自定义词库

由于赛题提供的数据为去敏数据，是利用 “C 区‘、”A 县“等字母序号代替真实地名的去敏方式，分词可能会把字母的区域单位分开，结果成不理想的：”A”、”县“，影响正确率。此外由于政务问题的特殊性，对于地名有较高的识别要求。虽然 jieba 分词可以识别新词的功能，但为保证更高的正确率，添加了清华大学创建的地点词库+所有可能出现的字母序列地名词库，保证较高的地名识别正确率。

2.3 分词

采用 jieba 精确模式。该模式较适合于文本分析。其采用以下算法

(一)基于前缀词典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)

(二)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

(三)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

2.4 去除停用词

停用词库汇总了百度停用词库、哈工大停用词库、四川大学机器智能实验室停用词库和中文停用词库。词库间相差并不大，为提升分词效果，采用汇总的停用词库。

2.5 词袋模型

使用词袋模型将文本内容转化成数值形式的特征向量。其算法流程如下

(一)在训练集中每一个出现在任意文中的单词分配一个特定的整数 id（比如，通过建立一个从单词到整数索引的字典）。

(二)对于每个文档，计算每个单词 w 的出现次数并将其存储在 X[i,j] 中作为特征 j 的值，其中 j 是在字典中词 w 的索引。

2.6 TF-IDF 算法

TF-IDF 是一种用于统计字词对于一个文件的重要度的计算方法。其计算方法为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,l}}$$

以上式子中$n_{i,j}$是该词在文件d_j中的出现次数，而分母则是在文件中所有字词的出现次数之和。

$$idf_i = lg\frac{|D|}{1 + |\{j: t_i \in d_j\}|}$$

其中

(一)|D|：语料库中文件总数

(二)$|\{j: t_i \in d_j\}|$：包含词语t_i的文件数目（即$n_{i,j} \neq 0$的文件数目）

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

三、 Python 库

使用到的 python 库有

(一) sklean：sklearn (scikit-learn) 是基于 Python 语言的机器学习工具。包含从数据预处理到建模及评估标准等一系列步骤所需的工具包。

(二) pandas：用于文件的读取、存储与计算。

(三) numpy：用于存储和处理大型矩阵。

(四) jieba：在中文分词领域使用较为广泛的分词组件。

(五) gensim：用于构建词袋模型、训练 LDA 模型和计算距离等。

(六) textrank4zh：用于关键句和词的提取。

(七) scipy:用于计算累积分布。

四、 任务一：群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是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。

任务一是利用附件 2 给出的数据构建一个分类模型。实质是一个有监督的分类任务，其含有多个类别，类别之间存在数据不平衡。在构建模型时，需要挑选适合多分类任务的算法以及减少数据不平衡现象。

4.1 评价方法

(一)F1 − Score值：

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

其中P_i为第i类的查准率，R_i为第i类的查全率。

(二)准确率Accuracy：

$$Accuracy = \frac{\sum_{i=1}^{n} R_i}{\sum_{i=1}^{n} L_i}$$

其中R_i为正确分成第i类的频数，L_i为第i类频数。准确率在类别数据较为均衡时可以做为一个参考。

4.2 解决思路

任务一是一个多分类任务，为解决多分类问题，我们采用 SGD（随机梯度下降）构建分类模型，其优点是较适合于多分类问题。另外考虑到类别之间存在数据不平衡，我们使用 EDA 技术增强数据，其主要优点是方法简单，性能优越。其次由于网格搜索的时间消耗大，我们采取小样本搜索最佳参数、大样本局部调优参数的参数搜素方法,其极大减少了最佳参数搜索时间,使模型性能趋于最佳。

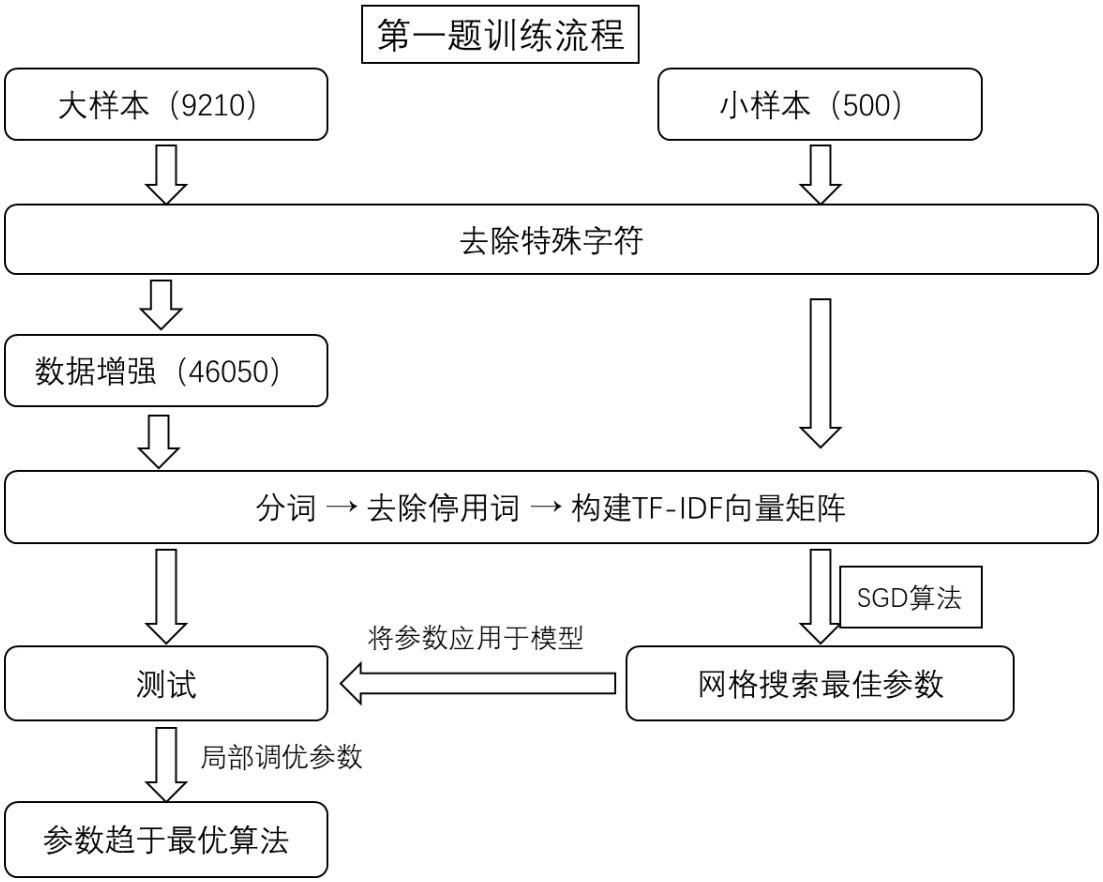

图 1 任务一训练流程图

4.3 基于 EDA 的数据增强

附件 2 给出数据的一级标签共有 7 个分类，每个分类的频数分布并不一致，频数最大的分类（城乡建设）占 21.81%，频数最小的分类（交通运输）占 6.66%。在划分训练集与测试集时，由于数据不平衡，可能会带来如下问题：

(一)抽样样本对模型效果影响过大。例如训练集的划分可能会抽取过少的“交通运输“类，而过多的抽取”城乡建设“类，导致分类效果不佳。

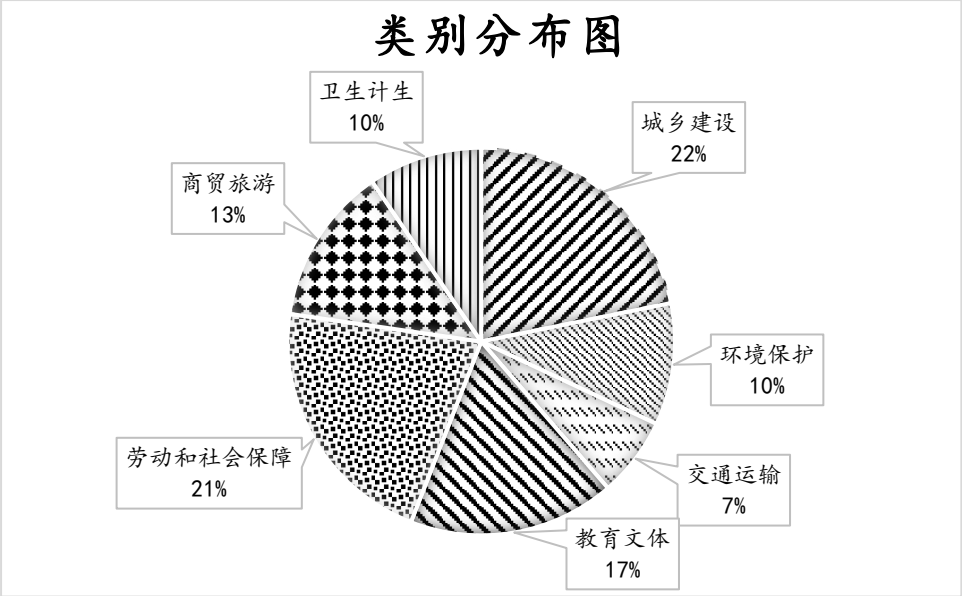

图 2 类别分布图

为解决该问题，我们使用 EDA（easy data augmentation），即简单数据增强，增强数据。

简单数据增强包括四种方法：

(一) **同义词替换(Synonym Replacement, SR)**：从句子中随机选取 n 个不属于停用词集的单词，并随机选择其同义词替换它们；

(二) **随机插入(Random Insertion, RI)**：随机的找出句中某个不属于停用词集的词，并求出其随机的同义词，将该同义词插入句子的一个随机位置。重复 n 次；

(三) **随机交换(Random Swap, RS)**：随机的选择句中两个单词并交换它们的位置。重复 n 次；

(四) **随机删除(Random Deletion, RD)**：以 p 的概率，随机的移除句中的每个单词；

由于长句比短句有更多的单词，所以长句在保持原有的类别标签的情况下，能吸收更多的噪声，因此我们采用基于句子长度来变化改变的单词数。

$$n = \alpha l$$

其中 n 表示改变的单词数，α 表示改变句子的百分比，l 表示句子的长度。其中$\alpha = p(RD)$。

附件 2 提供了 9210 条数据，根据作者推荐，设置参数为

$$\alpha（修改比例）= 0.1，n_{aug}(增强数) = 4$$

结果生成原始数据的 5 倍数据$9210 \times 5 = 46050$。

EDA 的方法简单，实现的性能很优越。实验结果表明，平均情况下，仅使用 50% 的原始数据，再使用 EDA 进行数据增强，能取得和使用所有数据情况下训

练得到的准确率。另外 EDA 是首次将数据增强用于 NLP 的技术。相较于另一些数据增强技术：语境增强、噪声、GAN 和反向翻译，EDA 能以更小的代价，实现差不多的性能。

EDA 虽然简单，但是也有其缺点。首先是 EDA 有可能降低模型的性能，因为 EDA 在增强过程中，改变了语义，但原始标签仍存在，从而产生了标签错误的句子。其次算法开销较大，时间复杂度与样本量 N 和句子长度 l 呈线性关系。

4.4 基于 SGD (随机梯度下降)的分类模型

任务一是一个“一对多“的多分类问题。随机梯度下降（SGD）是一种简单但非常有效的方法，在处理大规模的稀疏数据集上有较高的效率，且合适于处理多分类问题。其表达式为：

Loop{

for i = 1 to m ,{

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right)x_j^{(i)} \quad (for\ every\ j)$$

}

}

随机梯度下降的优点：

- 效率。
- 易于实施（有很多优化代码的机会）。

随机梯度下降的缺点：

- 需要很多的超参数，例如正则化参数和迭代次数。
- SGD 对特征缩放很敏感。

为克服随机梯度下降的缺点，通过网格搜索方式查找最优参数。网格搜索即列举参数列表，在参数列表中穷举所有可能的参数组合进行运算，找出最佳参数组合。其优点是可以找出列举的最佳参数组合，但是缺点也很明显，穷举是非常耗时的。为解决耗时问题，我们采用了小样本搜索最佳参数及大样本局部调优的优化过程。这样做可以极大缩短最佳参数搜索时间。由于样本量不一致，小样本的最佳参数可能不是大样本的最优状态。此时通过大样本下小范围对 SGD 关键参数进行网格搜索，找到大样本下的局部最优。这样的方式能够在较短时间内找出与最优相接近的结果。

最佳参数搜索

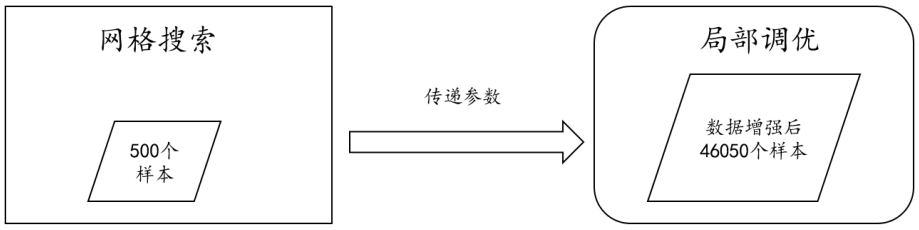

图 3 最佳参数搜索流程图

(一)在小样本下搜索得到的最佳参数为

表 1 小样本参数搜索结果

SGDClassifier 参数

loss	squared_epsilon_insensitive
penalty	l2
alpha	1e-3
n_iter_no_change	22
shuffle	Ture
tol	None
fit_intercept	False
random_state	42

抽样

test_size	0.2
random_state	3

将小样本（500）最佳参数应用于附件 2（9210）进行训练得到的 F1-score 值、Accuracy 值如下

$$F_1(1) = 0.89627,\ \text{Accuracy}(1) = 0.90553$$

(二)将小样本（500）最佳参数应用于附件 2 增强后的数据集（46050）进行训练得到的 F1-score 值、Accuracy 值如下

$$F_1(2) = 0.93995,\ \text{Accuracy}(2) = 0.94201$$

(三)用附件 2 增强后的数据集（46050）搜索得到的局部最优参数为

表 2 大样本局部参数搜索结果

SGDClassifier 参数

loss	perceptron
penalty	l2

alpha	1e-3
n_iter_no_change	22
shuffle	Ture
tol	None
fit_intercept	False
random_state	42
max_iter	1000
average	Ture
eta0	8
leaning_rate	invscaling

抽样

test_size	0.2
random_state	3

用调整后参数训练得到的 F1-score 值、Accuracy 值和混淆矩阵如下

$$F_1(3) = 0.99804，\ \text{Accuracy}(3) = 0.99837$$

表 3 混淆矩阵 Confusion matrix

	0	1	2	3	4	5	6
0	597	1	0	4	0	0	0
1	0	1944	2	0	1	1	0
2	0	0	863	0	0	0	0
3	2	0	0	1230	1	1	0
4	0	1	0	0	2039	0	0
5	0	1	0	0	0	1568	0
6	0	0	0	0	0	0	954

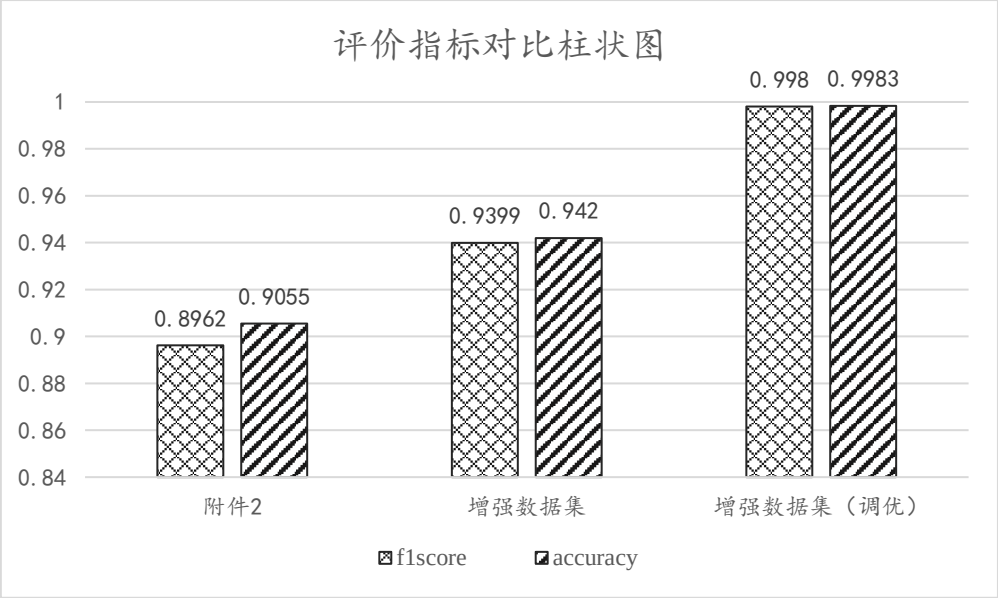

图 4 评价指标对比柱状图

如图所示，通过网格搜索方式，F1score 值被调优至 99.80%，准确度也与 F1score 值相近，模型趋于理想状态，说明模型已具备良好的性能。

五、 任务二：热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率

任务二是根据附件 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，并按格式给出排名前 5 的热点问题，和具体留言信息。

5.1 解决思路

任务二要解决的问题是:找出一段时间内集中爆发的热点问题。解决办法是，先通过 single-pass 聚类算法对留言详情进行聚类，得到多个主题，之后统计每一个主题下的所有留言数据的留言数，点赞数，反对数，时间跨度，并各自分配一个权重，从而得到每个问题的热度评价指标。根据热度评价指标对所有问题进行排序，得到排名前 5 的热点问题。

5.2 基于 single-pass 算法的主题聚类模型

任务首先要对留言详情进行聚类，识别同主题的留言。

Single-pass clustering，中文名一般译作“单遍聚类”，它是一种简洁且高效的文本聚类算法。在文本主题聚类中，Single-pass 聚类算法比 K-means 来的更为有

效。Single-pass 聚类算法不需要指定类目数量，可以通过设定相似度阈值来限定聚类数量。

Single-pass 聚类算法同时是一种增量聚类算法（Incremental Clustering Algorithm），每个文档只需要流过算法一次，所以被称为 single-pass，效率远高于 K-means 或 KNN 等算法。它可以很好的应用于话题监测与追踪、在线事件监测等社交媒体大数据领域，特别适合流式数据（Streaming Data），比如微博的帖子信息，因此适合对实时性要求较高的文本聚类场景，该问题就是 single-pass 聚类算法的最佳使用场景。

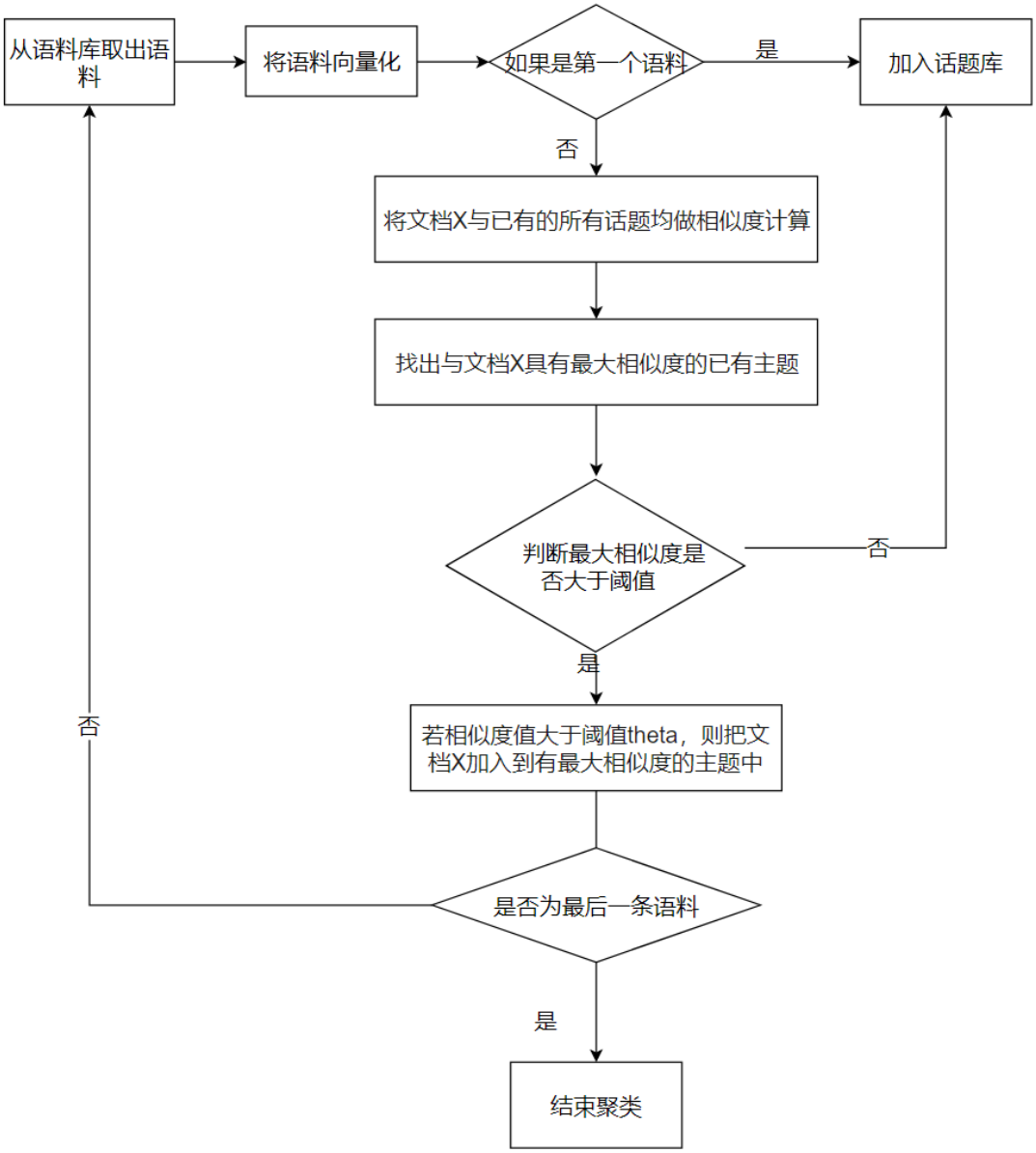

图 5 Single-pass 算法流程图

其算法流程如下：

(一)以第一篇文档为种子，建立一个主题；

(二)基于词袋模型将文档 X 向量化；

(三)将文档 X 与已有的所有话题均做相似度计算，可采用欧氏距离、余弦距离。

(四)找出与文档 X 具有最大相似度的已有主题；

(五)若相似度值大于阈值 theta，则把文档 X 加入到有最大相似度的主题中，跳转至 7；

(六)若相似度值小于阈值 theta， 则文档 X 不属于任一已有主题， 需创建新的主题类别，同时将当前文本归属到新创建的主题类别中；

(七)聚类结束，等待下一篇文档进入 。

Single-pass 算法由不同的阈值分为不同的话题，阈值越大，算法越精确，留言之间被分为一类的可能性越小，从而分成的话题数越多。由于文本数据之间相差较大，因此文本间的相似度阈值不应过大，应通过不断调整阈值，并人工查看聚类效果，可以得出当阈值在 0.1 附近时，分类效果较好。

阈值	聚类个数	阈值	聚类个数
0.05	501	0.2	2457
0.1	1205	0.25	2881
0.15	1891	0.25	2881

表 4 Single 算法阈值与聚类个数表

5.3 提取关键词/句

每一个主题分类中的文档数据都是围绕这一个话题讲述，某些词肯定会多次出现，将文本数据通过结巴分词后，用 TF-IDF 算法计算出其特征值矩阵，接着取出其中出现频率最高的前 10 个词，作为该主题的关键词。同理，将文本数据切分为一个个完整的句子，之后计算各个句子的相似度，并把意思相近的句子归为一类，最终取出意思相近的句子数量最多的前三个句子作为关键句。通过该算法可以快速的了解一个主题的主要内容。

具体算法的实现借助 python 库中的 textrank4zh 包。调用其中的 TextRank4Keyword 功能进行关键词提取，调用其中的 TextRank4Sentence 功能进行关键句提取。

最终各个主题的关键词和关键句保存到 result_list.txt 文件中。其中一个主题的结果如图所示：

【主题索引】:30
【主题声量】: 62
【主题关键词】: 车位,职工,销售,捆绑,广铁集团,购房,购买,政府,铁路职工,开发商
【主题中心句】 :

投诉：武广新城片区伊景园滨河苑为广铁集团的定向商品房，在未取得预售资格强行逼迫职工缴纳 18.5 万购房款且不签购房合同，工地停工 1 年多，现广铁集团又强烈要求捆绑购买 12 万车位一个，不买就取消购房资格，捆绑车位销售明显与中央文件及法律法规不符，但辛辛苦苦一辈子的老百姓房都买不起，要贷款，何来钱买车位

A 市武广新城片区下的伊景园.滨河苑是广铁集团铁路职工的定向商品房，之前已经统一交了 18.5 万的认购款，但没有正规的合同，现在集团下发文件，强制要求职工再交 12 万的车位费，不交就取消购房资格，捆绑销售车位

商品房伊景园滨河苑项目是由 A 市政府办牵头为广铁集团铁路职工定向销售的楼盘，作为集团的一名退休员工，我深深感觉到政府对铁路员工的关怀，在广铁辛苦几十年，住的是职工原来配置的福利房，遇到这样一个利好消息十分激动，现在到了缴纳认购款的时候，却被告知除了要缴纳 18.5 万的认购款还要缴纳 12 万元的钱买车位，不然就取消购房资格

图 6 result_list.txt 内容其一

5.4 热度评价模型

要想建立热度评价模型就需要得到每个主题的热度评价指标。用$i(1,2,\cdots,n)$表示主题的序号，用H_i表示第i个主题的热度。一个主题的热度主要与留言数，点赞数，反对数，时间跨度以及留言密度有关。这些指标定义如下：

(一)留言数

留言数为某一主题包含的所有留言的数量 。每个主题下面都有多个留言，留言数越多表示，该主题越重要，即热点评价指标和留言数成正相关关系。需要考虑的问题是一个用户可能会对同一个主题短期内进行多次留言，会使一个主题的留言数大量增加，影响热度的评定，因此需要剔除一个较短时间内用户对同一个主题的多次留言，只保留一个用户留言作为有效留言，之后统计一个主题中所有的有效留言的个数。

(二)点赞数

点赞数为某一主题包含的所有留言的点赞数之和。每个留言都会有人观看，如果其他人表示认同就会对该主题点赞，点赞数越多，就表示认同该主题的人越多，日常生活中也遇到过该主题人越多，即热点评价指标与点赞数成正相关关系。此外虽然会出现重复留言，但是点赞是所有用户对一个主题的评价，用户观看的是同一主题下的不同留言，重复留言对其影响较小。因此假设：用户对同一主题只进行一次点赞。最终一个主题的总点赞数是该主题下所有留言的点赞数之和。

(三)反对数

反对数为某一主题包含的所有留言的反对数之和。人们在看留言时候，也会

不认可该留言，认为该留言所说主题有偏差，该主题的热度会削弱，即热点评价指标与反对数成负相关关系。此外进行类似点赞数的假设：用户对同一话题只进行一次反对。最终一个话题的反对数是对该话题的所有留言的反对数之和。

(四)时间跨度

时间跨度为某一主题下留言的最晚留言时间与最早留言时间的差。表示一个话题出现的时长，时间跨度越大，表示话题出现的时间越久。即热点评价指标与时间跨度成负相关关系。

(五)留言密度

留言密度为单位时间内留言的数量，留言密度=留言数/时间跨度，留言密度越大表示单位时间留言的数量越大，话题热度越大，即热点评价指标与留言密度成正相关关系。

分别用$x_{i1},x_{i2},x_{i3},x_{i4}$表示第$i$个主题的留言数、点赞数、反对数和时间跨度，则$x_{i1}/x_{i4}$表示第$i$个主题的留言密度。但是由于这几个指标单位不统一，因此需要对$x_{i1},x_{i2},x_{i3},x_{i4}$进行无量纲化处理，我们使用的是 min-max 归一化方法：

$$x' = \frac{x - min}{max - min}$$

得到无量纲化处理后的$x'_{i1},x'_{i2},x'_{i3},(x_{i1}/x_{i4})'$。之后分别给$x'_{i1},x'_{i2},x'_{i3},(x_{i1}/x_{i4})'$分配权重$\emptyset_1,\emptyset_2,\emptyset_3,\emptyset_4$。得到热度评价指标的公式：

$$H_i = \emptyset_1 x'_{i1} + \emptyset_2 x'_{i2} - \emptyset_3 x'_{i3} + \emptyset_4 (\frac{x_{i1}}{x_{i4}})'$$

基于接近现实原则，我们给不同的权重分别设置为：

表 5 权重设置列表

$\emptyset_1$	0.8
$\emptyset_2$	0.09
$\emptyset_3$	0.05
$\emptyset_4$	0.06

最后得出的排名前 5 的热点问题表见附件

六、 任务三：答复意见的评价

任务三从“相关性“和“满意度”两个维度去评价答复意见的质量。

6.1 相关性

相关性是指答复意见对留言详情的相关度，即答复意见是否在回答留言详情的问题。如果答复意见在回答留言详情的问题，那么两个文本之间应该是相似的，或可以说两个文本在描述同一个主题。相关性的问题就可以通过计算两个文本的相似度来判断其相关性。

目前计算文本相似度的有基于字符串的方法、基于语料库的方法和基于世界知识的方法。基于字符串的方法原理简单、易于实现，但不足的是将字符或词语作为独立的知识单元，并未考虑词语本身的含义和词语之间的关系，且对于大部分基于字符串的方法不适合用于长文本。一般来说，留言详情与答复意见属于长文本，且留言详情与答复意见更多的是“主题相关”、“语义相关”，在字符上可能并未表现太多的关系。基于字符串的方法很难找到合适的算法。

为解决长文本问题和语义相关问题，我们采用潜在狄利克雷分布（LDA）。LDA 适用于长文本，其基于相似词语可能属于同一主题的理论，保证了文本的语义。计算距离的方法使用较为流行的 Hellinger 距离。Hellinger 距离的输出范围为 [0,1] ，越接近 0 表示两者之间越相似。

LDA 模型使用“附件 2”、“附件 3”的“留言详情”与“留言主题”作为语料库。基于附件 1 的“一级分类“构建 15 个主题。我们使用 gensim 库的 Lda 方法构建 LDA 模型。模型参数如下

$$\text{num\_topics} = 15, \text{minimum\_probability} = 1e-8, \text{其余为默认}$$

随后基于 Hellinger 的方法计算附件 4 留言详情与答复意见的距离。以计算出来的 Hellinger 距离表示两个文本的相关性。为便于理解，我们调整了算法结果。把相关性改为：

$$\text{相关性} = 1 - \text{Hellinger 距离}$$

即越接近 1 文本之间越相似。

我们对相关性进行描述统计如表 5，可以看到平均数和中位数分别 0.5095 和 0.5234，说明大部分文本的相关性是在 0.5 左右。偏度为-0.351，说明大部分的文本相关性是大于 0.5 的。峰度为-0.240，说明数据分布较为平缓。

表 6 相关性描述统计表

描述				
			统计	标准误差
相关性	平均值		0.5095	0.00300
	平均值的95%置信区间	下限	0.5036	
		上限	0.5154	
	5% 剪除后平均值		0.5133	
	中位数		0.5234	

	方差	0.025	
	标准差	0.15919	
	最小值	0.06	
	最大值	1.00	
	全距	0.94	
	四分位距	0.21	
	偏度	-0.351	0.046
	峰度	-0.240	0.092

构建的 LDA 模型缺点是用于构建语料库的附件 2 和附件 3 可能并未包含完整的 15 个一级分类。多出来的类别将可能会对距离计算产生负面影响。

6.2 满意度

满意度是对于用户角度去观察的。从留言系统的用户角度看，当用户的留言得到回复时，首先会关注是否已经解决了提出的问题，或者是否已经在解决，随后会关注其解决的过程。另外在一份关于酒店回复策略对潜在用户的回复满意度调研中，其研究结果发现：相比于程序化的回复方式，潜在顾客对有针对性回复的满意度更高；潜在顾客对于快速的回复满意度要高于缓慢的回复。对此影响满意度的因素可以分为两个部分：关于内容上的满意度，和时间上的满意度

于是我们基于满意度的特性，提出两个指标：内容满意度、时间满意度。

6.2.1 内容满意度

内容满意度我们针对 2 个特性：针对性、详细度。由于其在文本中难以去计算衡量，于是我们从结果的角度去思考，基于内容满意度的两个特性，手动把附件 4 的答复意见分成“强”、“中”、“弱”3 个不同满意程度。然后使用 SGD（随机梯度下降）算法构建分类模型。该方法的优点是在实质上把评价问题转化成分类问题，易于理解和实施起来简单，可扩展性强。

关于 3 个程度的评判标准如下：

- **“强”**：内容针对性强，解决程度高，内容详细。
- **“中”**：内容针对性不强，解决程度不高，或没有具体的解决过程。
- **“低”**：内容少，包含过多程序性的语句，解决程度低，或没有具体的解决过程。

最后的手动分类结果如下，“弱”和“中”分别占 5.22%和 6.96%，“强”占 87.82%。

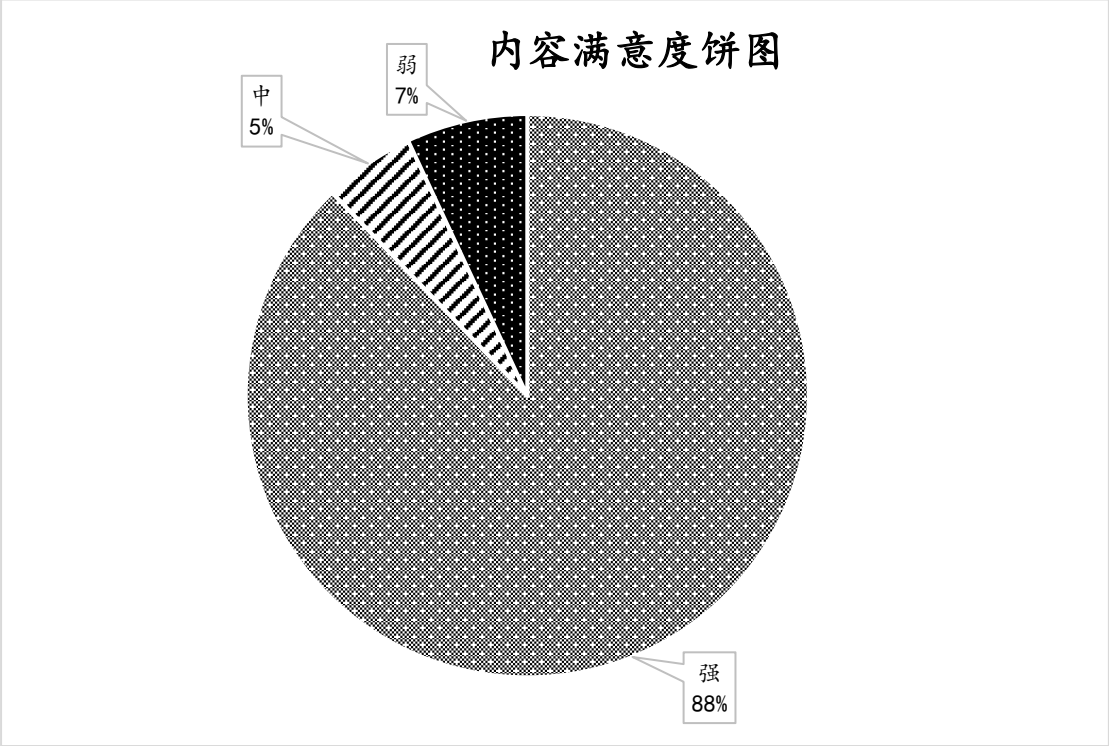

图 7 内容满意度饼图

由于数据不平衡在使用 SGD 构建模型时，我们提高了对“弱”和“中”的权重。参数设置如下：

表 7 任务 3 SGD 模型参数列表

SGDClassifier 参数	
loss	perceptron
penalty	l2
alpha	1e-3
n_iter_no_change	22
shuffle	Ture
tol	None
fit_intercept	False
random_state	42
max_iter	1000
average	Ture
eta0	8
leaning_rate	invscaling
class_weight	{'弱':0.30,'中':0.5}
抽样	

test_size	0.2
random_state	4

得到的 F1score 值和 Accuracy 值为：

$$F_1(4) = 0.75436，\ \text{Accuracy}(4) = 0.93971$$

该方法的缺点是：在手动分类过程中，带有主观性，且难以达到较高的分类效果。

6.2.2 时间满意度

从用户角度来说，回复得越快越好。我们假设回复满意度遵循反 S 曲线。那么时间满意度的表达就可以映射到正态分布曲线上，以其在正态分布曲线上的累计分布表示时间满意度。

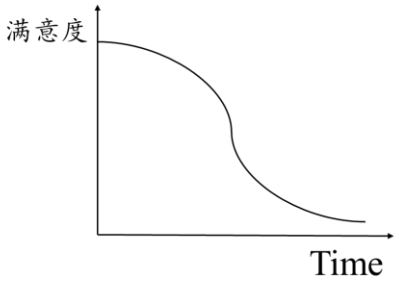

图 8 满意度随时间变化曲线图

我们首先计算每个留言时间与答复时间相差的秒数。然后对所有的秒数用 z-score 标准化的方法进行数据预处理（使用方法为 sklean.scale）。随后用 scipy 的 norm 方法计算每个每个值在期望为 0，标准差为 1 的标准正态正态分布中的累计分布。以累计分布作为评价时间满意度的指标。为便于理解，我们调整了算法结果。把时间满意度改为：

$$\text{时间满意度} = 1 - \text{累计分布}$$

即越接近于 1，时间满意度越大。

用标准正态分布的累计分布来评价，优点是其表现出答复时间越接近留言时间满意度上升得越快的特点，以及在接近 1 和 0 的两端，增速放缓。这样的动态规律，较符合现实意义。

七、 参考文献

[1] 韩世依，张钰晖，马云山，涂存超，郭志芃，刘知远，孙茂松，THUOCL：清华大学开放中文词库，2016.

[2] Jason Wei，Kai Zou，EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks，2019.2.

[3] NLP 中数据增强的实现

[4] 陈二静，姜恩波，文本相似度计算方法研究综述，2017.

[5] 田泽民，酒店管理回复策略对潜在顾客的回复满意度影响，2016.

[6] 文本挖掘从小白到精通（十）--- 不需设定聚类数的 Single-pass

八、 附件

表 8 热点问题表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	30	0.943341	2019/04/17 至 2019/09/01	A 市武广新城片区下的伊景园.滨河苑	广铁集团在未取得预售资格强行逼迫职工缴纳 18.5 万购房款且不签购房合同,现广铁集团又强烈要求捆绑购买 12 万车位一个,不买就取消购房资格
2	35	0.755543	2019/01/14 至 2019/12/20	A 市 A3 区咸嘉湖街道白鹤咀社区高心麓城小区业主	巴罗比幼儿园被投诉后,被教育局责令你园停止办学,但园方无视政府部门行政处罚,不愿移交本该属于全体业主的配套幼儿园给区教育局,仍继续在招生、教学,并广泛散布今年 9 月幼儿园将转成普惠制幼儿园谣言迷惑广大业主,严重损害了广大业主的利益
3	31	0.705432	2019/06/25 至 2020/01/26	A 市暮云街道丽发新城社区	搅拌站,水泥厂产生噪音和扬尘等环境污染问题,严重影响了周边居民的正常生活
4	67	0.547291	2019/01/05 至 2019/12/24	诺亚山林小区	反对在诺亚山林小区门口建医院
5	114	0.48152	2019/01/02 至 2019/12/25	社保与公积金缴存都在 A 市的普	住房公积金政策存在不合理的地方

				通职工	