第八届“泰迪杯”数据挖掘挑战赛 C 题

【摘要】在处理群众留言分类的问题上，本文采用了数据增强、摘要提取和数据清洗等方式对数据进行了相关处理，运用了 AdaBoost 算法+单层决策树的算法对分类模型进行了优化；在处理热点问题挖掘的问题上，本文采用了文本相似度分析、词云图分析等方法，同时结合了 Stack Overflow 公式进行热度分析；在处理答复意见评价的问题上，我们对答复意见的相关性和完整性进行了评价。

【关键词】数据增强、机器学习、文本相似度处理、数据清洗

【Abstract】In dealing with the problem of mass message classification, this paper uses data enhancement, abstract extraction and data cleaning to process the data, optimizes the classification model by using adaBoost algorithm and single-level decision tree algorithm, and uses text similarity analysis, word cloud analysis and so on to deal with hot issues, and combines the Stack Overflow formula for heat analysis, and the question of dealing with the evaluation of responses. We evaluated the relevance and completeness of the response.

【 keywords 】 Data enhancement, machine learning, text similarity processing, data cleaning

一、背景及意义

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。

附件给出了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。请利用自然语言处理和文本挖掘的方法解决下面的问题。

二、系统总体设计

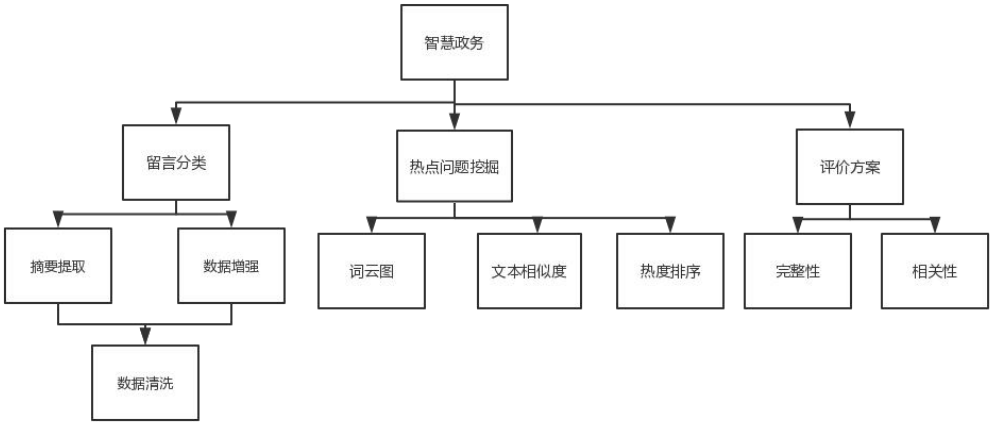

图 1：系统总体设计框架

三、群众留言分类

3.1 技术框架

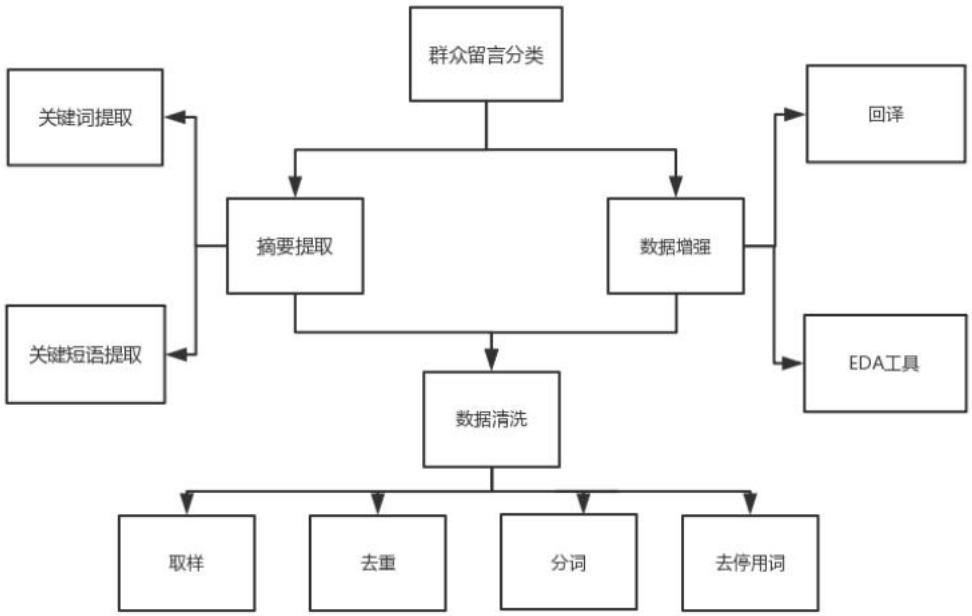

图 2：问题一技术框架

3.2 提取关键词、关键短语及摘要

在对附件二留言详情这一列数据做数据处理时，存在以下难点：文本信息量大，机器处理时间长；与留言主题不相关的信息较多，如“各位领导：你们好！”、“A 市是一座历史名城，是一座具有幸福感的城市”这些信息等，会对分类模型的精确率和召回率产生影响。基于以上出现的问题，我们需要对留言详情的文本数据进行提取摘要的处理。

我们选择了基于 PageRank 的 TextRank 算法用于完成对于文本的关键词、关键短语及摘要提取的需求。在提取关键词时，将原文本拆分为句子，在每个句子中过滤掉停用词（可选），并只保留指定词性的单词（可选），由此可以得到句

子的集合和单词的集合，然后通过 TextRank 算法计算出每个单词的重要性，然后选取若干重要的单词作为关键词。在提取关键短语时，若原文本中存在若干个关键词相邻的情况，那么这些关键词可以构成一个关键短语。在提取摘要时，将每个句子看成一个节点，若两个句子之间有相似性，认为对应的两个节点之间有一个无向有权边，权值是相似度，然后调用算法计算每个句子的重要性，最后提取重要性较高的句子作为摘要。

在完成提取摘要的算法后，对附件二的留言详情数据进行处理，针对处理的结果不断优化模型，以期最后得到的摘要结果可以尽可能的表达留言详情的主要信息。

部分留言详情的摘要提取	
留言详情	摘要
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，不但占用人行道路，而且护栏锈迹斑斑，随时可能倒塌，危机过往行人和车辆安全。请求有关部门牵头处理。	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，不但占用人行道路，
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，水是我们日常生活必不可少的用品，霉是一种强致癌物，我们住在这里连基本的健康保障都没有，请政府街道各领导重视起来，也请环保部门来检测，还我们一个健康安全的基本生活环境！	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，我们住在这里连基本的健康保障都没有。

图 3：部分留言详情的摘要提取

3.3 数据增强

为了提高分类模型的精确性，就必须要对已有的原始数据进行数据增强。比较简单的数据增强方式是进行同义词替换，但是仅仅做同义词替换的话无法真正达到数据增强的需求，因为在大多数情况下做完同义词替换后的两个文本相似度极高，并没有实现所谓的数据增强。因此我们采用了 EDA 的数据增强方法，EDA 包括了同义词替换、随即插入、随机交换和随机删除这几个处理方式。

在调用 EDA 工具时，根据原始数据量的不同要修改 num_aug 参数（每一条语料将增强的个数）和 alpha 参数（每一条语料中改动的词所占的比例），因为当训练数据很小时，模型更容易过拟合，这时建议多生成一些数据增强的样本。当训练数据很大时，大量增加数据增强样本可能没有帮助，因为模型本身可能已经能够泛化。因此我们根据多次的实验，针附件二中的数据结构我们将 alpha 的值设置为 0.2，对于不同一级分类标签下的数据设置了不同的 num_aug 参数，使得每种类型的数据在分布平均的情况下尽量达到模型拟合的目的。

除此之外，我们还采取了回译的方法，目的是在保持原意的前提下增加或移除单词并重新组织句子的能力。在回译的过程中，我们调用了 python 中的 translate 包实现相关的算法、并调用了 transformer 进行机器学习。

数据增强前后的对比				
原句	增强后的 1 号语句	增强后的 2 号语句	增强后的 3 号语句	增强后的 4 号语句
A 市西湖建筑集	A 基本建设西湖建	A 市西湖建筑集团占	A 市大中城市西湖	A 市西湖建筑集团

团占道施工有安全隐患	筑集团占道施工有安全隐患	道施工存在安全隐患	建筑集团占道施工有安全隐患	违规施工安全存在隐患
A 市在水一方大厦人为烂尾多年，安全隐患严重	A 市在水一方大厦人为烂尾多年，安全隐患非常严重	A 市在水一方大厦人为烂尾很久有安全隐患严重	A 市在水一方大厦人造烂尾多年，安全隐患严重	A 市在水一方大厦人有安全隐患严重，为烂尾多年
投诉 A 市 A1 区苑物业违规收停车费	A 市 A1 区投诉苑物业违规收停车费	投诉 A 市 A1 区苑物业管理违规收停车费	投诉 A 市一了百了，A1 区苑物业违规收停车费	投诉 A 市 A1 区苑物业违规散场收停车费

图 4：数据增强前后的对比

3.4 数据清洗

在完成了提取摘要和数据增强的工作后，就开始对数据做下一步的清洗工作。在数据清洗的过程中，我们通过调用了 python 中的 re、pandas、jieba、numpy 等库对数据进行了取样、去重、分词、去除停用词等操作，最后得到了以分词形式储存的数据。

部分数据的分词结果	
序号	分词结果
A 县红树湾小区物业频繁业主的维修基金红树湾小区的物业，近段在上频繁支取业主维修基金，维修价格市场价很多倍、而后没有时间、问题也并未解决，屡次、频繁出、频繁动用维修基金的怪圈、按法定程序办事栋一单元维修电梯有征询签字外，其余款项没有任何明细公示，业主委员会也并未征询广大业主意见、监督，进驻也并未到分之二业主同意这种情况下，公共利益容易变成部分谋取私人的温床，小区物业费使用、业主公共利益所得维修基金大肆使用未有正常公示征询法定的现象	县 红树湾 小区 物业 频繁 业主 维修 基金 红树湾 小区 物业 近段 频繁 支取 业主 维修 基金 维修 价格 市场价 倍 时间 并未 解决 频繁 频繁 动用 维修 基金 怪圈 法定程序 办事 栋 单元 维修 电梯 征询 签字 外 款项 明细 公示 业主 委员会 并未 征询 业主 意见 监督 进驻 并未 分之二 业主 同意 情况 公共利益 谋取 私人 温床 小区 物业费 业主 公共利益 所得 维修 基金 大肆 未有 公示 征询 法定 现象
反映 J 市威恒水泡制品跨国公司经消防计划性审核无度动土消防、安全生产违规反映-J 市威恒水泡制品跨国公司（J2 区许家洞镇倪华快餐馆天井）1、经消防计划性审核无度动土，该行为违拗了《中华民国医师法》第十二枝条的规定 2、未论进行消防计划性立案、未尽消防立案验收手续，竣工验收消防立案、未按要求设置环形消防车道、未按要求设置活动消防设施	J 市威恒 水泡 制品 跨国公司 消防 计划性 审核 无度 动土 消防 生产 违规 J 市威恒 水泡 制品 跨国公司 J 区 许家 洞镇 倪华 快餐馆 天井 消防 计划性 审核 无度 动土 违拗 中华民国 医师法 第十二 枝条 未论 消防 计划性 立案 未尽 消防 立案 验收 手续 竣工 验收 消防 立案 未 设置 环形 消防车 道 未 设置 活动 消防设施

图 5：部分数据的分词结果

3.5 机器学习

解决方案：安装开源的机器学习库；按照一定比例抽样将样本集划分为训练集和测试集；将文本分词并向量化后，得到词汇表中每个词在各个文本中形成的词向量；利用 TF-IDF，即“词频-逆文本频率”做进一步预处理，用 CountVectorizer 类向量化之后再调用 TfidfTransformer 类进行预处理；采用 adaboost+单层决策树

算法构建机器学习模型，抽取训练样本对模型进行训练，最后对测试集做出预测；算出每个一级标签的 f1-score，取每个标签的均值，得到 f-score 的值。

构建完运行环境后，利用 CountVectorizer()函数将处理好的分词向量化，用 TfidfTransformer()函数对训练集数据进行 tf-idf 预处理，TF 也就是我们前面说到的词频，我们之前做的向量化也就是做了文本中各个词的出现频率统计，IDF 可以凸显出样本中某个分词的重要性，概括来讲， IDF 反应了一个词在所有文本中出现的频率，如果一个词在很多的文本中出现，那么它的 IDF 值应该低，IDF 的基本公式如下：

$$IDF\ (X) = \log \frac{\mathrm{N}}{\mathrm{N}\ (X)}$$

我们的目标是让预测结果和真实结果尽可能符合，我们首先采用了朴素贝叶斯算法，f-score 的值只能达到 0.67 左右，因此我们采取了优化算法，即 AdaBoost 算法+单层决策树，adaboost 是一种将弱分类器和多个实例构建成一个强分类器的集成方法，单层决策树是一个简单的决策树，仅基于单个特征来做决策，这棵树只有一次分裂过程，在 adaboost 的框架中，对每次迭代：利用 build_stump()函数找到最佳单层决策树并加入到单层决策树数组，计算 alpha，计算新的权重向量 D，如果错误率 error_rate=0.0，则退出循环。

代码实现后将训练集代入到模型中得到 **F-score** 的值约为：0.92，各标签的得分如下：

	precision	recall	f1-score	support
交通运输	0.97	0.92	0.94	402
劳动和社会保障	0.89	0.88	0.89	396
卫生计生	0.95	0.94	0.94	423
商贸旅游	0.86	0.94	0.90	361
城乡建设	0.90	0.87	0.88	409
教育文体	0.91	0.92	0.92	408
环境保护	0.95	0.95	0.95	401
accuracy			0.92	2800
macro avg	0.92	0.92	0.92	2800
weighted avg	0.92	0.92	0.92	2800

图 6：各标签的训练得分表

四、热点问题挖掘

4.1 技术框架

基于前一问的研究基础，我们能够整体做到相关问题描述的数据清洗，对各个语料进行准确分词。但想要从众多纷杂问题中挖掘到热点问题，我们采取下列

方法：

（1）词云图；我们选择在分词的基础上借助词云图直观的找到较为突出的几个热点问题。

（2）关键词提取；以此对问题进行重点化分类，提取出相关问题。

（3）TF-IDF；继而借助 TF-IDF 相似度算法对相关问题进行相似度处理。

（4）热度排序；其次按照题目提供的地点、人群、点赞量进行统一分析，进行几个热点问题的热度排序。

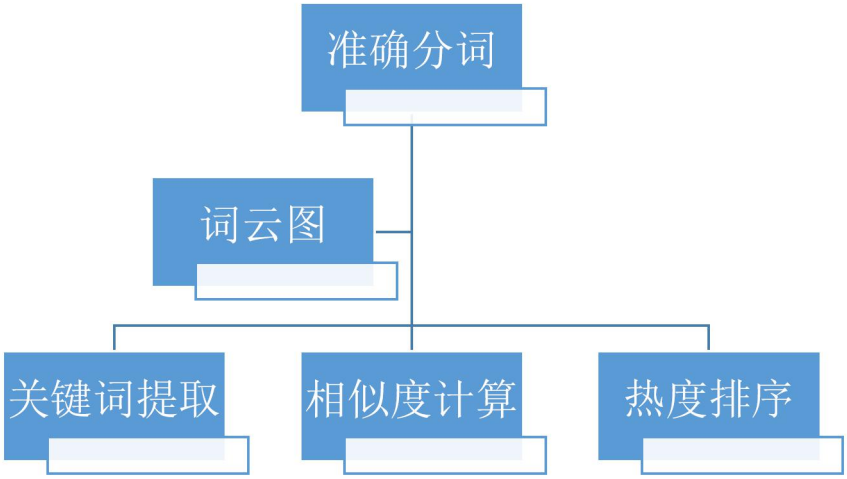

图 7：问题二技术框图

4.2 问题分类

4.2.1 去除停用词

由于问题描述中存在有许多无效信息，特指在文本中出现频率很高，但实际意义又不大的词。这一类主要包括了语气助词、副词、介词、连词等，通常自身并无明确意义，只有将其放入一个完整的句子中才有一定作用的词语。如常见的“的”、“在”、“和”、“接着”之类。为了不影响后续问题分类效果，要去除这类分词。

由于我们处理的问题描述语句均为日常用语，不包含较为特殊的停用词，所以我们借助网上最常用的停用词表，应用到处理过后的分词文件中，通过匹配将分词文本中的停用词去除。得到问题相关的有效分词文件。

4.2.2 绘制词云图

绘制词云图主要借助于 python 中 matplotlib 集成库进行图形绘制。“词云”对文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，文本中频次越高的词在词云图中显示越大。借助专有词表，遍历我们已知分词文本的每一个分词，统计词频。绘制出相应的问题词云图。具体结果如下图

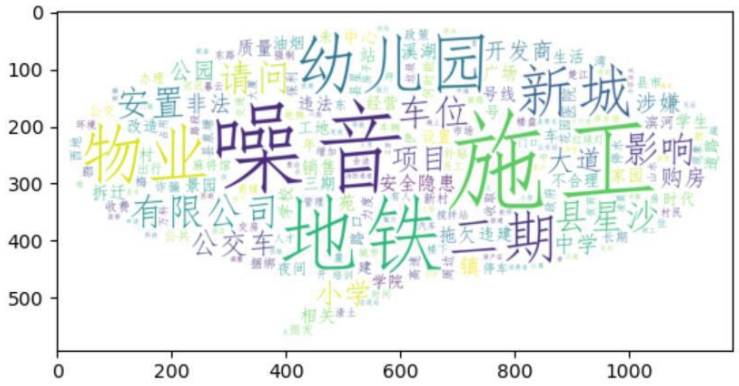

图 8：词云图

4.3 问题分类

4.3.1 关键词提取

由于代码实现提取更为繁琐，因此我们借助所得问题词云，直观得到最主要的几大问题。噪音、地铁、施工、幼儿园、物业，是观测到的几大问题关键词。

4.3.2 问题分类

首先我们根据获取的热点问题关键词，利用 excel 检索功能从上千条数据中检索出来相应问题的完整性描述，进行归纳统一分析。同时进行每一类热点问题的总体性描述。

4.3.3 提取目标问题

由于我们要对热点问题进行详细的留言统计，因此直接运用检索到的热点问题相关的留言主题查询到留言详情。

借助 dataframe 的相关运算操作，将读取到的留言详情写入到新的热点问题留言明细表中。其中主要借助与嵌套循环结构，遍历整个数据库，找到需要的留言详情。

4.4 相似度分析

由于需要处理的数据量比较大，所以依据 **4.3 问题分类**中按关键词对全部数据进行问题分类，相同相似问题归为一类。因此在后续处理中需要对分组过后的问题进行相似度处理，并进行排序。具体实现步骤如下：

相似度分析实现	
1、	首先通过 pandas 获取 excel 信息，通过 jieba 分词进行处理，jieba 分词要首先自定义词典以及排除信息，这样效果会差异很大，然后形成一个二维数组。
2、	使用 gensim 中的 corpora 模块，将分词形成后的二维数组生成词典
3、	将二维数组通过 doc2bow 稀疏向量，形成语料库
4、	后来更改为：LsiModel 模型算法，将语料库计算出 Tfidf 值。

5、	获取词典 token2id 的特征数
6、	计算稀疏矩阵相似度，建立一个索引
7、	读取 excel 行数据，通过 jieba 进行分词处理
8、	通过 doc2bow 计算测试数据的稀疏向量
9、	求得测试数据与样本数据的相似度

在求得各个问题的相似度后，借助其相似度进行排序，部分排序结果如下表

留言主题	频次
关于 A7 县恒基凯旋门万婴格林幼儿园办普惠园的咨询	4
A 市美联幼儿园什么时候能招生	4
A 市保利西海岸配套幼儿园迟迟不开园	4
A2 区天悦幼儿园的社团收费太贵了	3
咨询 A4 区万国三期幼儿园的学位问题	3
何时发布 2019 年 A7 县普惠性幼儿园清单及收费标准	3
A7 县星沙恒大翡翠华庭小区内幼儿园收费是否合理？	3
请 A 市提高民办幼儿园的师资待遇水平	2
A3 区中海国际社区幼儿园无法满足实龄幼儿就读需求	2
A3 区高心麓城小区配套幼儿园无证办学且拒绝转成普惠制幼儿园	2
咨询 A7 县金科时代中心小区配套幼儿园的问题	2
A4 区君悦幼儿园为什么一直不改为普惠或公办？	2
A8 县喻家坳中心幼儿园突然砍掉学生上课时间	1
投诉 A3 区颐美兰亭幼儿园整个环境有问题	1
A6 区时代倾城万婴幼儿园多收取学费，退费困难的投诉	1
A7 县深业睿城幼儿园迟迟不开园	1

图 9：部分相似度排序

4.5 综合考虑相似度、时间、点赞量

为了保证热度计算的准确性，在进行相似度处理后，我们还综合考虑了问题出现的时间，点赞量等几个因素。借助 Stack Overflow 排名算法进行热度排序。

$$\frac{\frac{M*(Y-N)}{5}}{((Q\text{age}+1)-(\frac{Qage-Qapdated}{2}))^{1.5}}$$

公式解读：

我们的问题热度

(1)和该问题出现的总次数成正比，次数越多，热点度越高

(2)和该问题本身被评价的好坏正相差成正比，点赞数越多，反对数越少，热点度越高，说明这是个热点问题

(3)和该问题提出的时间与所选时间成反比，这个问题被提出的越早，热点度越低。

(4)和该问题提出时间与所选时间的差值成反比，这个问题如果距所选时间很久，那么热点度降低。

通过算法对相似度排序后的每个数据计算热度指数。最后根据热度指数再次排序得出所有问题中热度最高的 5 个问题（即排名前 5 的热点问题）。具体结果见“热点问题表.xls”

为简化数据处理，方便理解，对公式中涉及的变量进行声明：

M：	问题数量
Y：	问题点赞量
N：	问题反对量
Qage:	所选日期距离问题发表的时间
Qapdated:	所选日期距离最后一个回答的时间

实际计算中所选日期为:　2020.05.07

五、答复意见的评价

5.1 技术框架

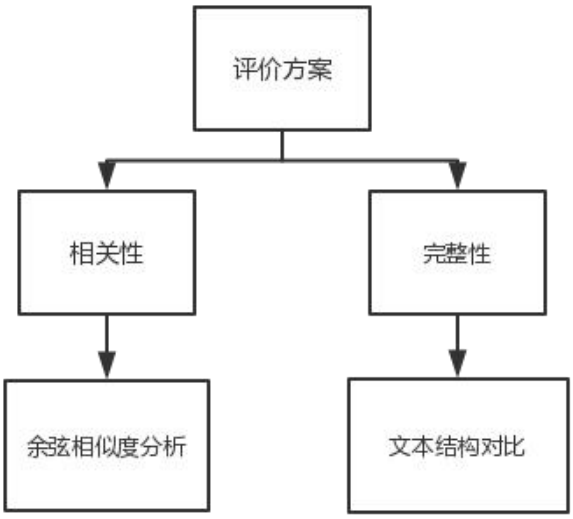

图 10：问题三技术框架

5.2 相关性分析

先对数据进行分词、清洗、增强等步骤，后对每个问题及答复意见进行相似度对比，采用余弦相似度 cosine_similarity，计算出每条评价的相似度百分比，值越大就表示越相似，相关性也就越强，余弦距离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，我们可以把它们想象成空间中的多条线段，都是从原点（[0, 0, ...]）出发，指向不同的方向，通过线与线之间夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似。相比欧氏距离，余弦距离更加注重两个向量在方向上的差异。余弦相似度的计算公式如下：

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

5.3 完整性分析

通过对样本数据的分析，我们发现一条完整的回复无非体现在开头，正文和结尾的叙述上，开头出现网友、你好等词语，正文出现回复、办法、文件、方案等词语，结尾出现年月日等词语，我们认定该回复为一条相对完整的回复，利用关键词匹配等技术，如若上述三个板块出现我们预先设定的词语，则为该回复的完整性评分增加 0.01，评分越高，该回复越完整。

部分答复意见评价结果		
留言用户	相关性评分	完整性评分
A00045581	0.234742767	0.02
A00023583	0.067936622	0.02
A00031618	0.030460385	0.02
A000110735	0.045291081	0.01
A0009233	0.015249857	0.01
A00077538	0.06172134	0.01
A000100804	0.084270097	0.03
UU00812	0.068599434	0.01

图 11：部分答复意见评价结果

六、模型评价及模型优化

本文建立的模型基本可以解决上述问题，但是依然存在以下不足：在数据增强模块依然存在大量相似度极高的数据，没有真正达到数据增强的目的；在数据清洗时存在语义歧义的问题；本文针对文本相似度处理所采用的算法计算量太大，且精度不高等；对于答复意见的评价方案尚不承受，未来将尝试其它的算法进行优化。

参考文献

[1]Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Association for Computational Linguistics, 2004.

[2]基于文本内容的敏感词决策树信息过滤算法[J].邓一贵,伍玉英.计算机工程. 2014(09)

[3]基于决策树的网页敏感词过滤系统设计[D].李伟.西北农林科技大学.2018