“智慧政务”中的文本挖掘应用

摘要

随着互联网的普及，中国的网民数量急剧增加，网络成为非常便捷、流行的交流平台。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，伴随着大数据、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。

本文以 C 题的三个问题以及四个附件的数据为核心，使用文本挖掘、文本聚类等技术，基于 python 和 Excel 等工具，对问题展开了研究：

首先是对群众留言分类的问题，即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，随着群众逐渐依赖网络留言，这样传统的工作模式存在工作量大、效率低，且差错率高等问题。我们基于朴素贝叶斯算法，通过配合使用贝叶斯分类器，建立了关于留言内容的一级标签分类模型。

其次是处理热点问题，即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。我们使用了 BIRCH 聚类算法得到相似的文本，并结合词频和赞同与反对数，定义了计算热度指数的公式，并得到热度前 5 的热点问题表与热点问题留言明细表。

最后是对答复意见的质量给出一套评价方案。准确地评价方案可以客观真实地对答复意见给与评价，有利于工作部门的回访并了解实际工作情况方便做出改进。本文根据附件四的内容，采用计算文本相关性的方法，同时结合回复的时间，构建出一套评价系统。

关键词： 文本挖掘 文本聚类 贝叶斯分类 机器学习

目录

一．问题重述

1.1 问题的背景

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。如何利用自然语言处理技术来减轻管理人员的压力和快速处理市民的需求值得被更多人去研究。

1.2 问题的重述

1.2.1 问题 1

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。使用 F-Score 对分类方法进行评价建立关于留言内容的一级标签分类模型能够很好地提高分类的效率。

1.2.2 问题 2

某一时段内群众集中反映的某一问题可称为热点问题，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。我们根据附件 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。

1.2.3 问题 3

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，并实现对回复的评价。

1.3 需要解决的问题

1.建立分类模型对留言内容的一级标签进行分类。
2.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，给出排名前 5 的热点问题以及相应热点问题对应的留言信息。
3. 针对有关部门对留言的答复意见，给出一套对意见的评估方案。

二．问题的分析

2.1 问题一的分析

首先因其文本内容的多样性，我们将一级标签数据进行数字化来对文本内容进行分类，在数据处理时进行去重和停用词的常规处理提高机器学习的准确性和效率。然后通过相关函数得到所有有效的词，将这些与训练集和测试集词频矩阵构成的词作为新文本预测时的特征。最后利用朴素贝叶斯算法，通过数据模型的结果比较来建立以留言处理+留言详情为预测数据源的数据模型，经模型检验得吻合度为 85%。

2.2 问题二的分析

对于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。要找到文本之间的相似度，然后进行文本聚类分析。BIRCH 算法比较适合于数据量大，类别数 K 也比较多的情况。通过 BIRCH 聚类算法得到相似的文本，并结合词频和赞同与反对数，定义了计算热度指数的公式，并得到热度前 5 的热点问题表与热点问题留言明细表。

2.3 问题三的分析

对于评价回复质量的评估，关键的地方在于找到评估指标的选取。本文首先选取五组数据作为分析基础,将回复时常以及回复相关性作为评估留言的重要指标。然后通过对时间列表以及文本相关性的归一化处理完成数据模型的构建，即其得分越高其答复的质量越高。最后进行模型检验，验证评估指标选取的可行性和准确性，结果与预期吻合良好。

三．模型假设

1. 假设不同一级标签的文本之间没有联系，是相互独立的。
2. 假设回复留言的时间只与工作效率有关。
3. 假设留言的点赞数和反对数与其关注度成正比。
4. 假设附件中留言的标签准确。

四．符号说明

表 1：符号说明

符号名称	符号含义
Score	答复意见的得分
s_time	答复所用时间的归一化得分
s_topic	答复意见与留言主题相关性
s_detail	答复意见与留言详情相关性
s_all	答复意见与主题结合详情的总相关性

五．模型一的建立与求解

5.1 表格预处理

5.1.1 赋予标签

我们将一级标签的数据按下表进行数字化以方便后面的计算，得到以下表 2

表 2：一级标签符号说明

一级标签	数字符号
城乡建设	1
环境保护	2
交通运输	3
教育文体	4
劳动和社会保障	5
商贸旅游	6
卫生计生	7

5.1.2 分组与整合

我们先把附件 2 中以下三组标题单独提取出来，分别是留言主题，留言详情，一级标签。

为了考虑不同类型的数据对第一题结果的影响，我们将分别考虑“留言主题”、“留言详情”、“留言主题+留言详情”作为数据源进行机器学习求出预测最准的方式。

我们通过 Excel 的“&”函数，将表格整合如下表 3：

表 3

留言主题	一级标签		留言详情	一级标签		留言主题留言详情	一级标签
A市西湖建	1		A3区大道西行便道	1		A市西湖建筑集团占道施	1
A市在水一	1		位于书院路主干道	1		A市在水一方大厦人为烂	1
投诉A市A1	1		尊敬的领导：A1区	1		投诉A市A1区苑物业违规	1
A1区蔡锷南	1		A1区A2区华庭小区	1		A1区蔡锷南路A2区华庭楼	1
A1区A2区华	1		A1区A2区华庭小区	1		A1区A2区华庭自来水好大	1

5.1.2 数据处理

1. 去重处理

通过 Excel 表格的清除重复项功能，将每列的重复数据清除，这样进行机器学习时，相同数量的训练集留言可以相对包含更多的特征词，使机器学习更加高效。

2.停用词处理

在机器学习的过程中，为了降低留言文本信息中无意义的字词或符号如：的了、人民、末、啊、阿、哎、哎呀、哎哟、唉、俺等对结果产生的不必要的影响，同时减少机器计算的维度，我们将一些字词存入 TXT 文件中，在对留言分词前读取改文件中的停用词。

3.文本特征提取与选择

在我们先将分词完的文本保存起来，通过 vectorizer.fit_transform( )可以计算出每个词语出现的次数，通过 vectorizer.get_feature_names()函数，我们可以得到所有有效的词，我们将这些与训练集和测试集词频矩阵构成的词保存到 TXT 文件中，作为新文本预测时的特征、

5.2 文本机器学习

5.2.1 朴素贝叶斯算法的基本原理

朴素贝叶斯（Naive Bayesian）是基于贝叶斯定理和特征条件独立假设的分类方法，它通过特征计算分类的概率，选取概率大的情况，是基于概率论的一种机器学习分类（监督学习）方法，被广泛应用于情感分类领域的分类器。

5.2.2 贝叶斯公式

设Ω为试验 E 的样本空间，A为E的事件，如果有k个互斥且有穷个事件，即 $B_1, B_2, ..., B_k$ 为 Ω 的一个划分，且 $P(B_1)+P(B_2)+...+P(B_k)=1$，$P(B_i)>0$ $(i=1,2,...k)$，则：

$$P(B_i \mid A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i)P(A \mid B_i)}{P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + \cdots + P(B_k)P(A \mid B_k)}$$

2.2.3 朴素贝叶斯算法基本原理

朴素贝叶斯法利用贝叶斯定理首先求出联合概率分布，再求出条件概率分布。这里的朴素是指在计算似然估计时假定了条件独立。基本原理可以用下面的公式给出：

$$P(Y \mid X) = \frac{P(Y)P(X \mid Y)}{P(X)}$$

其中， $P(X \mid Y) = P(X_1, X_2, \ldots X_m \mid Y) = P(X_1 \mid Y)P(X_2 \mid Y)\ldots P(X_n \mid Y)$

$P(Y \mid X)$ 叫做后验概率， $P(Y)$ 叫做先验概率， $P(X)$ 叫做似然概率，P(X)叫做证据。

2.2.4 基于文本机器学习的朴素贝叶斯算法

先验概率

$$P(C = c) = \frac{\text{属于类}c\text{的文档数}}{\text{训练集文档总数}}$$

条件概率

$$P(w_i \mid c) = \frac{\text{词}w_i\text{在属于类}c\text{的所有文档中出现次数}}{\text{属于类}c\text{的所有文档中的词语总数}}$$

1.条件概率表示的是词 w_i 在类别 c 中的权重

2.条件概率独立性假设，丢失了词语的位置信息，在文本表示上来说，就是它失去了语义信息。可以通过 ngram 的特征来减少损失，但是也不能有效解决语义上的损失。

3.先验概率和条件概率的计算都利用了最大似然估计。它们实际算出的是相对频率值，这些值能使训练数据的出现概率最大。

5.3 建立模型

5.3.1 模型比较

为了获得 F 值最大且均衡性最好的预测模型，我们根据训练与预测数据源的不同，分别建立以下四个模型、

1. 以留言主题为预测数据源的数据模型
2. 以留言详情为预测数据源的数据模型
3. 以“留言主题+留言详情”为预测数据源的数据模型
4. 将每条数据源使用上述三个模型分别预测，当结果不一致时，取 f-score 最大的结果。

为了减少模型结果偶然性，我们采用了不同的数据进行测试得到的结果如下表 4 所示：

表 4：每个模型对应 F-score 分数

测试集	训练集	模型 1 的 F	模型 2 的 F	模型 3 的 F	模型 4 的 F
1000	500	0.72	0.75	0.79	0.78
1000	1000	0.74	0.76	0.80	0.79
2000	500	0.75	0.80	0.81	0.81
3000	1000	0.74	0.79	0.83	0.82
4000	4000	0.87	0.90	0.91	0.91

本文使用算法得到其中一次 F-score 结果，如图 1-图 4 所示。

	precision	recall	f1-score	support
1.0	0.84	0.88	0.86	732
2.0	0.99	0.70	0.82	253
3.0	0.91	0.71	0.80	205
4.0	0.87	0.86	0.86	682
5.0	0.82	0.97	0.89	1170
6.0	0.93	0.81	0.86	478
7.0	0.92	0.81	0.86	480
accuracy			0.87	4000
macro avg	0.90	0.82	0.85	4000
weighted avg	0.87	0.87	0.86	4000

图 1：模型 1 对应 F-score 分数

	precision	recall	f1-score	support
1.0	0.82	0.93	0.87	732
2.0	0.98	0.83	0.90	253
3.0	0.95	0.60	0.73	205
4.0	0.92	0.92	0.92	682
5.0	0.87	0.98	0.92	1170
6.0	0.96	0.79	0.87	478
7.0	0.98	0.86	0.91	480
accuracy			0.90	4000
macro avg	0.92	0.84	0.87	4000
weighted avg	0.90	0.90	0.89	4000

图 2：模型 2 对应 F-score 分数

	precision	recall	f1-score	support
1.0	0.84	0.95	0.89	732
2.0	0.99	0.88	0.93	253
3.0	0.96	0.65	0.78	205
4.0	0.93	0.93	0.93	682
5.0	0.89	0.98	0.93	1170
6.0	0.97	0.82	0.89	478
7.0	0.99	0.89	0.93	480
accuracy			0.91	4000
macro avg	0.94	0.87	0.90	4000
weighted avg	0.92	0.91	0.91	4000

图 3：模型 3 对应 F-score 分数

	precision	recall	f1-score	support
1.0	0.84	0.95	0.89	732
2.0	0.99	0.86	0.92	253
3.0	0.96	0.64	0.77	205
4.0	0.93	0.93	0.93	682
5.0	0.89	0.99	0.93	1170
6.0	0.96	0.81	0.88	478
7.0	0.98	0.88	0.93	480
accuracy			0.91	4000
macro avg	0.93	0.86	0.89	4000
weighted avg	0.92	0.91	0.91	4000

图 4：模型 1 对应 F-score 分数

我们得出在此题中 F 得分最高和均衡度分析最好的模型是 3 号模型，因此选取了 3 号模型。

5.3.2 模型建立

我们将数据集划分为 6500 条随机训练集，1000 测试集和 1000 条未经处理的新数据集，通过朴素贝叶斯的原理。

通过上文中数据处理的方法对训练集和测试集进行数据处理，得到其分词文件并将其转化为词频矩阵后我们调用 MultinomialNB().fit 函数建立模型。

得到模型后，我们使用 TXT 文件将文本特征词保存才来，同时为了提高模型使用的效率以及可重复使用性，我们引入 joblib 库，将训练完的模型 model 以及矢量化的词频矩阵 vectorizer 进行保存。

5.3.3 模型检验

通过代入 1000 条随机测试集进行检验，我们得到的模型评价如图 5 所示：

Name: 7, Length: 1000, dtype: object

	precision	recall	f1-score	support
1.0	0.79	0.91	0.85	201
2.0	0.88	0.95	0.91	147
3.0	0.95	0.52	0.67	120
4.0	0.93	0.92	0.93	197
5.0	0.83	0.96	0.89	164
6.0	0.73	0.73	0.73	100
7.0	0.95	0.79	0.86	71
accuracy			0.85	1000
macro avg	0.87	0.83	0.83	1000
weighted avg	0.86	0.85	0.85	1000

图 5：测试集检验结果图

如上图所示，根据计算得出的 F 值为 0.85。

5.4. 模型通用性检测

我们将除训练集和测试集以外的模型带入到训练好的模型中，得到的结果如下：

	precision	recall	f1-score	support
1.0	0.65	0.86	0.74	141
2.0	0.81	0.89	0.85	121
3.0	0.94	0.26	0.41	65
4.0	0.98	0.95	0.96	329
5.0	0.82	0.91	0.86	149
6.0	0.83	0.79	0.81	143
7.0	0.88	0.73	0.80	52
accuracy			0.84	1000
macro avg	0.85	0.77	0.78	1000
weighted avg	0.86	0.84	0.84	1000

图 6：新数据检验结果图

从图 6 对结果的分析看出，即使是预测未经处理的新数据，模型仍然可以有较

好的准确率，可见模型具有一定的准确性和通用性。

六．模型二的建立与求解

6.1 数据预处理

6.1.1 去重压缩

我们打开的附件 3 的数据，发现了留言主题是有重复的。为了提高数据的有效性，本文进行了多次的去重处理。先利用 Python 的去重函数把附件 3 的主题数据去重，数据量从 4326 条变成了 4209 条。这样进行数据分析时，使数据分析更加准确。

有些词句是多重叠词的形式，比如“很好很好很好很好很好很好很好很好很好很好很好很好很好很好很好”和“加油加油加油加油加油加油加油加油”，这类词语对本文计算情感倾向的得分影响较大，本文在获取数据时设计了一个文本机械压缩函数，从而使得获取的文本短而精炼，如“很好很好很好很好很好很好很好很好很好很好很好很好很好很好很好”可以压缩为“很好”，提高了算法的效率和准确性。

6.1.2 分词统计词频

本文把留言主题的内容根据 Jieba 分词可以得到留言主题中关于地区的词频，其中部分地区的评论数据如表 5 所示：

表 5：关于地区的词频

地区	评论数
A7	662
A3	429
A2	259
A4	235
A1	206
A5	169
西地省	155
A6	143

地铁	110
幼儿园	86
A8	82
新城	78

6.2 模型准备

6.2.1 文本相似度的计算

在将文本表示成向量之后，衡量文本之间相似度的方法主要可以通过计算向量之间的距离来实现，距离越小，相似度越高。除此之外，也可以通过集合之间的相似度计算方法来衡量。

6.2.2 文本聚类算法

文本聚类算法是话题发现的基础。一般来说，描述同一事件或话题的文本相似度较大，而描述不同事件或话题的文本相似度较小。基于这个假设，可以将文本相似度作为文本之间距离的度量，将那些相似度较高的文本聚合在一起，实现信息的有效组织和管理。我们常用的是划分据类：划分聚类是将文本的集合划分到事先指定个数的类中，有模糊划分和确定性划分两种方式。k-means 是目前最为常用的划分聚类的方法。k-means 作为一种非常常用的聚类算法，思路简单清晰，聚类效果也较好。但是基于附件 3 中的数据量比较大，这里我们选取了 BIRCH 聚类算法。BIRCH 算法比较适合于数据量大，类别数 K 也比较多的情况。它运行速度很快，只需要单遍扫描数据集就能进行聚类。

6.3 分类

6.3.1 详细地区分类

由于 A7 区的频数是地区中最多的。本文选出留言主题或者留言详情中含有 A7 关键字的评论。在这些数据进行词频统计，得到 A7 地点对应的具体地点词频

如表 6 下：

表 6：详细地区的词频数

详细地区	评论数
星沙	135
泉塘	38
幼儿园	27
楚龙街道	22

6.3.2 文本聚类为同一问题

本文以星沙关键词作为例子，首先选取了留言主题或者留言详情含有星沙关键词的评论。通过 BIRCH 聚类算法，把 134 条聚类为 105 条评论。根据聚类的个数得到以下部分数据如表 7 所示：

表 7：聚类算法结果

留言主题	聚类个数
A7 县星沙中贸城欺诈业主、拖欠业主资金不退还	8
A7 县星沙镇星沙派小区违法违建	5
A7 县星沙一桥从早堵到晚	4
A7 县星沙恒大翡翠华庭小区内幼儿园收费是否合理？	3

6.3.3 时间序列统计

根据计算同一主题的最早时间与最晚时间，可以得到发表该主题的时间范围，如表 8 所示。

表 8：时间范围与点赞与反对个数

留言主题	时间范围	点赞与反对总个数
A7 县星沙中贸城欺诈业主、拖欠业主资金不退还	2019/3/25 至 2019/3/26	5
A7 县星沙镇星沙派小区违法违建	2019/10/15 至 2019/10/17	12
A7 县星沙一桥从早堵到晚	2019/8/7 至 2019/8/8	15

A7 县星沙恒大翡翠华庭小区内幼儿园收费是否合理？	2019/1/30 至 2019/2/15	0

6.3.4 计算热度

本文根据时间周期内的评论数 n_c，点赞与反对数 n_l，词频个数 n_w。根据

$$score = 0.4n_c + 0.2n_l + 0.4n_w$$

从而得到每个评论对应的热度分数，为了与其他评论对比分析，所以采用了最大最小归一化处理，使得热度指数最大为 1。

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

6.4 获取热点问题表与热点问题留言明细表

根据热度指数排序可以得到以下热点问题表如表 9 所示。

表 9：热点问题表

问题 ID	热度指数	时间范围	地点/ 人群	问题描述
1	0.95	2019/4/10 至 2020/1/26	A2 区丽发新城附近	投诉 A2 区丽发新城附近建搅拌站噪音扰民
2	0.85	2019/1/9 至 2019/9/12	A3 区西湖街道茶场村	A3 区西湖街道茶场村五组何时启动拆迁?
3	0.79	2019/7/8 至 2019/12/23	A7 县诺亚山林小区	坚决反对在 A7 县诺亚山林小区门口设置医院
4	0.76	2019/7/24 至 2019/9/05	A3 区梅溪湖看云路一师润芳园小区临街	A3 区梅溪湖看云路一师润芳园小区临街门面烧烤夜宵摊扰民
5	0.73	2019/3/26 至 2019/7/15	A1 区东成大厦	请求公开 A1 区东成大厦申报建设成本资料

根据文本聚类算法找到相对应的详细留言详情，部分数据如表 10 所示：

表 10：详细留言详情表

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	265342	A000106707	…	2019/8/4 17:10	…	0	5
1	193091	A00097965	…	2019/6/19 23:28	…	0	242
3	210107	A00042107	…	2019/7/8 10:38	…	0	14
5	203369	A00027934	…	2019/3/26 19:05:00	…	0	0

七．模型三的建立与求解

7.1 数据选取

本文选取了附件四的以下信息留言主题，留言详情，答复意见，留言时间，答复时间。

7.2 指标选取

7.2.1 回复时长

本文利用 Excel 的 DAY 和 Hour 函数，计算出每个留言从留下到答复的时间，我们规定回复时间越少对应的答复质量就越高。我们以小时为单位，最短时间的留言用了 0 小时 30 分便被回复，而最长的留言 767 小时才被回复。

7.2.2 回复相关性

我们使用 python 的 gensim 库计算留言和答复之间的相关性。我们规定与留言主题和留言详情的相关性越高的答复质量就越高。我们将计算的结果进行归一化处理。我们分别计算了 3 种相关性，即留言主题与回复的相关性，留言详情与

回复的相关性，留言主题与留言详情整体与回复的相关性。

7.3 评价模型建立

7.3.1 模型建立

通过对时间列表以及文本相关性的归一化处理，最终答复意见的评分为：

$$Score = 25(s\_time + s\_topic + s\_detail + s\_all)$$

该模型可将一条答复意见在 0~100 分的区间内进行打分评估，得分越高的答复意见我们认为其质量就越高。

7.3.2 模型检验

以其中一条留言为例，如表 11 所示。

表 11：其中一条留言

留言主题	投诉 B 市买云望体育新城小区众多问题
留言详情	我于 2017 年购买云望体育新城小区的房子，合同约定交房日是 2019 年 3 月 30 日，由于开发商和承建方扯皮导致工程延期到现在，现收到开发商 B 市云望房地产开发有限公司于今年 8 月 30 日收房的通知，现将投诉问题列出如下：1，延期交房违约金 0.1 万分之，全国最低了，一套 60 万的房子一天违约金只有 6 块钱，请 B 市的主管部门能管一管这种欺诈老百姓的开发商吗?请保障普通百姓的合法权益;2，收房通知上面写明，交房必须预交 1 年的物业费，但是在我们的购房合同上并没有写明必须要预交一年物业费，请问开发商这捆绑交房条件，是否合法?业主能否拒绝?3，物业费高达 1.8 块一平米，但是我们小区是和建材市场混建的，只有 3 栋住宅，容积率极高，绿化几乎没有，开发商这种高定价，主管部门能否协调(业主代表和开发商已经协商过，但是无果)?4，按现在的工程进度，月底肯定达不到交房条件，开发商说交房后再搞绿化和一些配到设施，请问他们这样做合法吗?能通过政府部门验收吗?
答复意见	“UU0081201”您好！您通过平台《问政西地省》的留言收悉，已转有关部门调查核处，如有相关情况将及时反馈，谢谢！2019 年 8 月 6 日尊敬的网友：您好，您反映的问题已收悉。现回复如下：1、延期交房的违约赔偿标准，按照合同约定执行。如果任何一方对约定的标准有异议，可以通过司法途径维护自己的权益。2、

	根据株发改发 201927 号关于印发《B 市城区物业服务收费实施意见》的通知第一条第四款物业服务费按月计收，经双方约定可以预收，但最长预收期限不得超过十二个月。3、前期物业服务开发商及物业公司是通过招投标程序，物业服务公司按照中标价格实行的物业服务费标准。1.8 元/m2 属于四级物业服务收费。具体服务内容及收费标准小区收费处须有公示。4、开发商符合购房合同约定的交房条件即可交房。商品房合同约定通常是：交付时须取得主体工程竣工验收合格，既在我局质安站指导、监督下的工程竣工“五方验收”合格;绿化和配套设施验收属“合并验收”范畴。2019 年 8 月 8 日
留言时间	2019/8/6 11:12:58
答复时间	2019/8/6 22:20:40
时长	11

我们将文本信息带入文本相关性模型中得到以下结果，如表 12 所示

表 12：相关性模型结果

留言文字类型	相关性
留言主题	0.18731715
留言详情	0.7106317
留言主题与详情	0.7244719

代入模型，最终该留言得分为：65.19 分

八．模型的评价与改进

8.1 模型的优点

1. 将数据分别代入算法训练将结果进行对比，故得到的分类模型更为准确

2. 代入数据前对数据进行了去重、压缩、去停用词等处理，使模型的效率得以提高。

3. 预测模型训练时将数据随机代入，减少模型建立的偶然性。

4. 建成后有测试集和完全未处理新数据分别代入检测，使模型的灵活性和实用性得以保证。

8.2 模型的缺点

1. 建立分类模型时只使用了贝叶斯算法，未与其他算法比较，可能模型的质量会受到算法本身局限性的影响。
2. 建立评价留言答复模型时，基于 gensim 的文本相关性计算中，可能忽视了同义词近义词的影响，降低模型的精确性。

九．参考文献

[1]邓海龙.Python 词向量训练与应用技术解析[J].语料库语言学,2019,6(02):88-109+116-117.

[2]李玥.机器学习的分类、聚类研究[J].电脑知识与技术,2020,16(04):161-162.

[3]石凤贵.基于机器学习的垃圾短信识别应用[J].电脑知识与技术,2020,16(03):202-204.

[4]何伟. 基于朴素贝叶斯的文本分类算法研究[D].南京邮电大学,2018.