“智慧政务”中的文本挖掘应用

摘要

“智慧政务”是电子政务发展的高级阶段，利用物联网、云计算、移动互联网、人工智能、数据挖掘、知识管理等技术，提高了政府办公、监督、服务、决策的智能化水平，而从海量信息中提取并分类有效信息的文本挖掘，起着至关重要的作用，数据的简化和整合，帮助政府实现管理的精细化、智能化、社会化，有利于相关部门进行即使有效地解决问题。

本文就该项目对以下三个任务进行讨论和分析：

任务一：我们将附件 2 的“留言主题”和“留言详情”合并的文本，作为参数 X，将对应的一级标签作为结果 Y。利用多项式朴素叶贝斯模型，让机器学习对留言内容的标签进行预测。

任务二：我们将附件 3“留言主题”和“留言详情”合并的文本放进词典里，通过 Gensim 对文本进行相似度计算，若相似度大于某个值，则将相应文本归为一类，将分类好的文本存成 Excel 表格，后续的热点问题表和热点问题明细表则通过读取上述表格，定义热度指数算法，找到热度指数排名前五的类别，将相应内容存为热点问题明细表，再通过热点问题明细表，找到各类问题有代表性的地点/人群和问题描述。将相应内容存为热点问题表。

任务三：本文分析附件 4 中的答复意见信息特性后，定义了答复意见质量评价指标，构建答复意见质量评价指标体系，然后选用层次分析法对答复意见质量进行评价，最后将得出的评价指标权重和留言者评分相结合，运用公式进行质量评分。

关键词：文本相似度，热度评议，评价指标

Abstract

"Smart government" is the advanced stage of e-government development. It improves the intelligent level of government office, supervision, service and decision-making by using Internet of things, cloud computing, mobile Internet, artificial intelligence, data mining, knowledge management and other technologies. Text mining, which extracts and classifies effective information from massive information, plays an important role. Data simplification and integration o help the government to realize the refinement, intelligence and socialization of management, and to help relevant departments to solve problems effectively.

The following are the paper discusses and analyzes of the project:

Task 1: We will combine the text of attachment 2 "message subject" and "message details" as parameter x, and take the corresponding primary label as result y. Using the multinomialnb model, then let NLP learning to predict the tags of the message content.

Task 2: We put the combined text of Annex 3 "message subject" and "message details" into the dictionary, and calculate the similarity of the text through Gensim. If the similarity is greater than a certain value, the corresponding text will be classified into one category, and the classified text will be saved into an Excel table, and the hot spot table and hot spot list will be determined by reading the above generated table .Through permutation, the top five categories of heat index will be found, and then, save the corresponding contents as the detailed table of hot issues. Then through the list of hot issues, we can find out the representative places / groups and problem descriptionsfrom from it. Last, save the corresponding content as the hot issue table.

Task 3: After analyzing the response information characteristics in Annex 4, we define the response quality evaluation index, build the response quality evaluation index system, and then use AHP to evaluate the response quality, and then ,combine the weight of the evaluation index and the score of the commenter. Finally, the formula can be used to evaluate the quality.

Key words: Text Similarity, Heat Review, Evaluation Index

目录

一、问题描述

1.1 问题描述

随着微信、微博、阳光热线等网络问政平台的兴起，各类社情民意相关的文本数据量不断攀升，大量的文本数据仅仅依靠人工划分不能达到可视的效率，不利于对“智慧城市”的建设。本文对如何处理并挖掘文本信息进行了研究，主要基于自然语言处理技术和文本挖掘。

对于所给的数据，我们要做的是根据附件 1 中的标签数据制作群众留言分类模型；接着是热点问题的挖掘，根据群众留言分类模型对留言进行分类，找出群众的主要问题，并对问题进行识别分类，定义合理的热度评议指标，计算文本相似度，得到热点问题表以及热点问题明细表；最后，对附件 4 中的答复意见进行评价，制作答复意见质量评价指标。

二、数据预处理

在数据挖掘的过程中，数据预处理往往是最基础、也是最重要的一步。数据预处理的结果影响后面文本的特征处理，文本挖掘以及数据分析。本文数据预处理主要分成三个部分：数据去重、统一时间格式、对“留言主题”、“留言详情”、“答复意见”等文字信息进行分词。

2.1 数据去重

在处理文本数据之前，我们考虑到所给的附件可能会出现重复数据的现象，为避免对后续的文本挖掘和数据分析造成影响，我们首先对附件中的所有信息进行去重的操作，剔除重复的部分。

2.2 统一时间格式

在分析数据时，我们发现附件 3 中“留言时间”这一列数据出现了时间格式不统一的问题，如图 1，会对后面研究时效性造成一定的影响，为便利接下来的文本挖掘，我们将所有数据出现的时间统一格式。

	A	B	C	D	E	F	G
25	188553	A00092239	老街上有无证理疗馆骗取老	2019/6/6 21:58:22	听了解的，他们什么松花粉，灵芝孢	0	0
26	188560	A00075321	鸿餐饮店拖欠工资，员工维	2019/10/6 15:01:36	告单位下达整改通知书是不是一纸空	0	0
27	188592	A00039456	云时代小区三期后面要建垃	2019/6/18 10:38:44	它都只有10多米，垃圾站朝向小区，	0	0
28	188665	A000106234	雅湖东方航标2栋2楼有传销	2019/3/26 0:18:26	寒嘘问暖一下后　就开始讲起来他们	0	1
29	188679	A000103846	台解决落实退休教师各项补	2019/12/19 12:48:05	的教育事业作出了贡献；现在情况有	0	5
30	188691	A00036601	市仕弘教育等培训机构涉嫌	2019/7/11 15:21:58	面试培训费近7000元，并信誓旦旦能	0	1
31	188774	A00048792	万芙路段经常有改装车飙车	2019/6/18 23:03:31	多次拨打A2区交警大队的电话（0000	0	0
32	188780	A00094754	A7县黄花镇梁坪村黄泥岭山	2019/10/20 14:59:13	林权办的工作人员拒不接收材料，无	0	0
33	188799	A000107349	城小区的孩子到泉塘小学上	2019/5/22 12:24:16	下学一般都是老人，并且随着二胎时	0	1
34	188801	A909180	河苑针对广铁职工购房的霸	2019-08-01 00:00:00	问题一个接着一个，首先未取得预售	0	0
35	188809	A909139	南路丽发新城居民区附近搅	2019/11/19 18:07:54	小区旁50米处建搅拌站，运渣车吵得	0	1
36	188820	A00028138	区旧城区棚户改造项目范围	2019/2/21 11:38:34	力，本人是A7县星沙街道杉仙岭社区	0	0
37	188829	A00026141	先锋派出所办个签证，拒收	2019/2/15 13:17:06	由于没带银行卡，我说用人民币支付	0	2
38	188856	A000104234	园路39号维也纳智好酒店有卖	2019/8/14 12:12:40	常在酒店一楼茶吧或二楼麻将房休息	0	0
39	188876	A00013435	县榔梨龙华安置区外围马路修	2019/2/26 11:29:49	一到下雨天裤子鞋子上都是泥巴　出	0	0
40	188887	A00085665	基凯旋门万婴格林幼儿园办营	2019/7/17 10:19:08	文件要求，我县全面开展城镇小区配	1	4

Sheet1

图 1

2.3 分词以及去停用词

文本数据庞大并且复杂，要进行文本的挖掘，首先要对所给信息进行删减，得到有效信息。在附件中，文本信息如“留言详情”、“答复意见”等，都含有无意义的词汇（如词汇“我们”、“谢谢”、“您好”以及标点符号“，”、“：”等），如图 2。我们对数据中的文字信息进行分词以及去停用词，获取了数据中的有效信息，如图 3。

	A	B	C	D	E	F	G
476		U0003167	取城市垃圾处理费不	2019/11/15 11:44:12	我们是梅家田社区辖区内的小区居民，我们每年都依法依规向小区物业公司交纳了城市垃圾处理费。我也认为，环卫局收取城市垃圾处理费是合法合规，并且有利于环境改善的，对此我们一直支持。但这个问题我也有几点意见向您提一下：一、据我所知，我们附近的有滨江新外滩小区、雅优翠园小区物业公司没有向业主代收城市垃圾处理费。两个小区的业主也从未交过城市垃圾处理费。二、同一个小区，有的业主坚持不交城市垃圾处理费，物业也无可奈何。以上两个问题确实影响到我们交城市垃圾处理费的意见，我们所在的物业公司也未给出满意的答复。但我咨询过相关专业人员，确实这是现实情况，环卫局也力不从心。但我希望，关于城市垃圾处理费的问题，加大宣传力度，以及执法力度，对拒不代收的小区物业公司、小区业主拿出实际办法予以处理。对一直积极配合代收城市垃圾处理费的小区，给予肯定及鼓励。对此一直拒交城市垃圾处理费的居民或业主，给予惩治办法。我希A6区市垃圾的管理以及处理更为科学，希望垃圾分类工作，越做越好，也希望A8县在确实环保达标的情况下，垃圾焚烧发电项目早日推动建设，让垃圾科学化处理妥当。谢谢。	城乡建设	
530		U0008488	A3区魏家坡小区脏乱差	2019/11/10 18:59:24	难闻，蚊叮虫咬，都没一个地方可以让人好好休息一下，小区的马路也是破烂不堪，每次路	城乡建设	

Sheet1

图 2

	2	投诉 市 A1 区兔 物业 造型	尊敬 领导 A1 区兔 小区 恒大 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公司
	3	A1 区 蔡锷 南路 A2 区华庭	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味
	4	A1 区 A2 区华庭 自来水	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味
	5	投诉 市 盛世 耀凯 小区	2015 年 购买 盛世 耀凯 小区 17 栋 楼 楼 两层 共计 平方 足额 缴纳 物业费 费用
	6	咨询 市 楼盘 供暖 一事	西地省 地区 常年 阴冷 潮湿 气候 近年 气候 恶劣 地处 月亮 岛 片区 近年 规划 楚江
	7	A3 区 桐梓 坡 西路 可可	尊敬 胡书记 您好 家住 市 A3 区 桐梓 坡 西路 可可 小城 居民 停水 小区业主 业委会
	8	C4 市 收取 城市 垃圾处理	梅家田 社区 辖区 小区 居民 依法 依规 小区 物业公司 交纳 城市 垃圾处理 费 环卫局
	9	A3 区 魏家坡 小区 脏乱差	尊敬 市政府 领导 您们好 市 A3 区 魏家坡巷 业主 多年 小区 脏 乱 差 社区 得不到
	10	市 魏家坡 小区 脏乱差	尊敬 市政府 领导 您们好 市 A3 区 魏家坡巷 业主 多年 小区 脏 乱 差 社区 得不到
	11	A2 区 泰华 一村 小区 第四	请求 依法 监督 泰华 一村 小区 第四届 非法 业主 委员会 涉嫌 侵占 小区业主 公共 资
	12	A3 区梅 溪湖 壹号 御湾	住 梅 溪湖 壹号 御湾 楼 2019 年 月份 住 每天晚上 停水 白天 水 很小 用水 高峰
	13	A4 区鸿涛 翡翠 湾 强行 入	尊敬 领导 你们好 市 A4 区 捞刀河 镇 彭家巷 社区 鸿涛 翡翠 湾 一名 业主 一名
	14	地铁 号线 施工 导致 市锦楚	地铁 号线 施工 导致 万家 丽路 锦楚 国际 星城 小区 三期 一个月 停电 10 来次 每次
	15	A6 区润 紫 郡 用电 解决	尊敬 领导 你好 A6 区润 紫 郡 业主 今年年初 小区 周边 竖起 一道道 高压线塔 筑起
	16	市锦楚 国际 新城 月份 停	市 A5 区 朝晖路 锦楚 国际 新城 三区 月份 一共 停电 次 每次 原因 停电 线路 卖
	17	A9 市 城区 南 西 片区	肯定 选择 A9 市 西南角 支持 A9 市 西南角 设 站 A8 县 南 北 中三向 融城 A9
	18	A6 区政府 加大 对滨水 新	尊敬 领导 A6 区 几年 发展 突飞猛进 城市道路 绿化 建设 却显 落后 滨水 新城 高楼
	19	A5 区楚府 线 几个 小区	A5 区楚府 线 包括 森林 雅苑 楚府 十城 天际 山庄 多个 小区 停电 短短 一周 停电
	20	调查 西地省 建望 集团 及	涂 愈 一名 建筑业 从业者 求助 请求 相关 部门 调查 西地省 建望 集团 及西地 省辉
	21	A2 区 山水 嘉园 栋 单元	市 A2 区 黄谷路 368 号 山水 嘉园 栋 单元 706 房 改建 成 户 出租 人员 消防
	22	A3 区 杜鹃 文苑 小区 外	市政府 市 交警支队 市 安监局 市 环保局 A3 区政府 市 A3 区 杜鹃 文苑 小区 业主
	23	建议 市 B 市 C 市 联合	市 B 市 C 市要 融城 交通 基础 长株 潭 城铁 开通 串联 市 河东 主城区 B 市

Sheet1

图 3

2.4 文本数据优化

数据之大之多给相关部门带来一定的麻烦，一方面不能确定出民意的具体要求，另一方面不能总结出热点要求，除了上文中提到的分词以及去除停用词，词云图也是解决该问题的一个有效办法。我们选取了“留言主题”这一组数据，做成如图 4 和 5 的词云图，得到留言中的热点问题。

图 4

图 5

三、任务一

3.1 机器学习了解

3.1.1 类别

在学习之前，我们了解到机器学习的方法有很多种，也有多种分类方法。

一、线性函数和非线性函数。

线性函数是满足线性定义的函数（斜直线），而非线性函数是不满足线性定义的函数（曲线），如图 6。

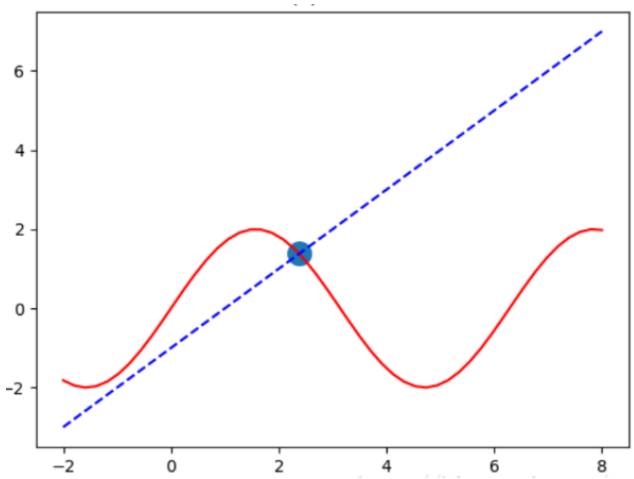


图 6

二、学习方式：监督学习、非监督学习、半监督学习等

监督学习：通过已有的输入数据与输出数据，找到两者之间的对应关系，生成一个对应的函数，将输入的数据映射到对应的输出数据。在监督学习类别中，有分支：分类（离散变量预测）和回归（连续变量预测）

非监督学习：对输入数据集进行建模半监督学习，综合利用原始数据，对原始数据找相似并分类。在非监督学习类别中，有分支：聚类（产生一组互相相似的集合）

半监督学习：通过一些有标签数据的局部特征和没标签数据的整体分布，去学习最优标注，从而提高学习准确率的方法。在半监督类别中，有分支：半监督分类，半监督回归，半监督聚类，半监督降维[1]。

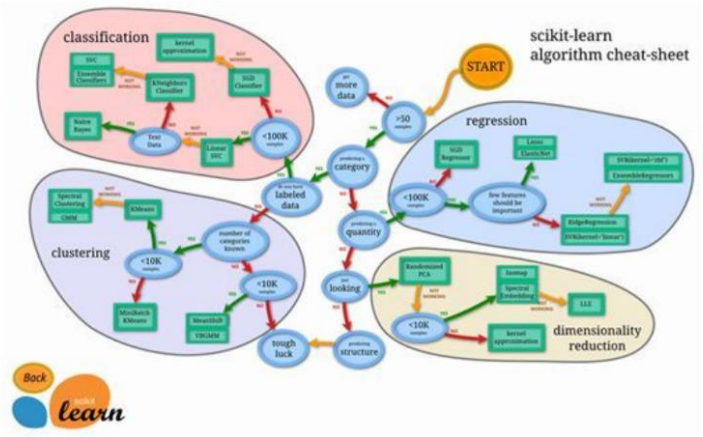


图 7.机器学习思维导图

3.1.2 分类问题

一、二分类/多分类/多标签
二分类：类别一共有两种，可用 0 和 1 分别表示两种类别。
多分类：表分类任务中有多个类别,每个类别只有一个标签。
多标签：分类任务中，有多个类别，每个样本有多个标签[2]。
二、sklearn 中支持的分类器
1、固有多类：朴素贝叶斯，LDA 和 QDA，决策树，随机森林，KNN。
2、支持多标签：决策树，随机森林，KNN。
3、一个-VS-一：sklearn.Svm.SVC。
4、One-VS-All：所有线性模型除了 sklearn.svm.SVC[3]。

3.2 朴素叶贝斯

3.2.1 朴素叶贝斯的介绍

朴素叶贝斯方法是以贝叶斯原理为基础，使用概率统计的知识对样本数据集进行分类，有着算法简单，分类有效率，误差小的特点，是广泛的分类模型之一。除此之外，又有以下三种常用的模型[4]：

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

高斯分布.png

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

多项式分布.png

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

伯努利分布.png

图 8.三个模型的算法

1、高斯分布的模型 (GasussianNB）
处理连续的特征变量，客观描述真实情况。
2、多项式分布的模型（MultinomiaLNB）：
常用于文本分类，特征是单词，值是单词出现的次数。
先验概率 P(c)=类 C 下单词总数/整个训练样本的单词总数。
类条件概率 P(tk|c)=(类 C 下单词 tk 在各个文档中出现过的次数之和+1)/(类 c 下单词总数+|V|)。
3、伯努利分布模型（BernoulliNB）
其特征是全局变量，值是布尔类型。
P(c)= 类 C 下文件总数/整个训练样本的文件总数
P(tk|c)=(类 C 下包含单词 tk 的文件数+1)/(类 C 的文档总数+2)

3.2.2 多项式朴素叶贝斯的流程

第一步：准备数据集，将数据集分为训练集和测试集。
第二步:特征工程，将原始数据变为特征。
第三步：利用数据训练模型。

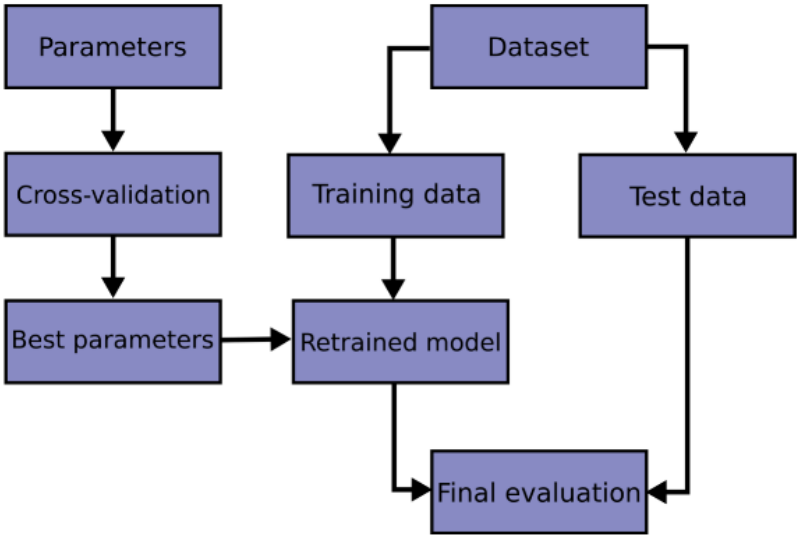


图 9.做训练模型步骤

3.2.3 关于任务一的代码解析

这里我们采用了上文提到过的多项式分布的朴素叶贝斯和多分类法，来对附件 2 的内容做模型。

第一步：先要配置环境，这里是我们需要配置的第三方库。

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report
from sklearn.metrics import precision_recall_fscore_support
```

图 10

第二步：准备数据：

附件 2 中一共有 9210 条数据，其中对一级标签有影响的内容是“留言主题”和“留言详情”，其它数据如“留言编号”、“留言时间”等对一级标签并没有影响，所以我们需要的数据有“留言主题”、“留言详情”和“一级标签”。考虑到一些用户的主题和留言是相互补的，如有的“留言主题”的内容只有半个括号、有些数据中的“留言主题”中提到了地点等有关信息，但却在“留言详情”中不再提及。为了不影响对一级标签的预测，我们有必要将去了停用词的留言主题和留言详情两者的字符串组合起来，作为条件读入。

第三步：训练模型：

在这之前，我们要先要了解其参数和指标：

1、CountVectorizer：将文本中的词语转换为词频矩阵。

2、Transform：使用该空间将文本数据转化为特征矩阵。

3、Fit：构建特征空间[5]。

4、Precision：准确率。

Precision 越高说明模型对负样本的识别能力强，与测量点方差有关。

$$p = \frac{TP}{(TP + FP)}$$

Precision

5、Recall：召回率。

Recall 越高说明模型对正样本的识别能力强。

$$R = \frac{TP}{(TP + FN)}$$

Pecall

5、f1-score：准确率和召回率的综合。

$$F1 = \frac{2PR}{P + R}$$

F1-score

7、Accuracy：精度宏。与测量点偏差有关。

$$A = \frac{TP + TN}{TP + FP + FN + TN}$$

Accuracy

8、support：支持的数据总数。

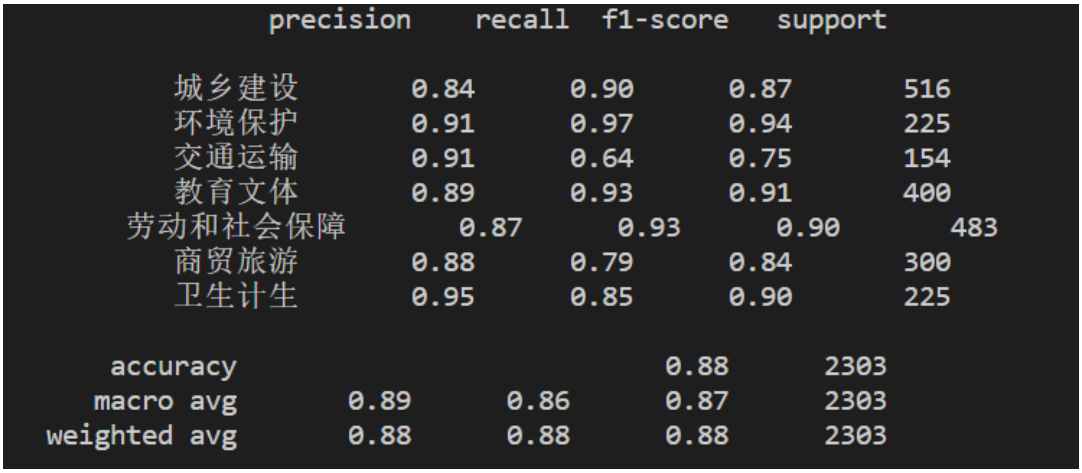
9、macro avg：平均数。

10、weighted avg：加权平均数。

接下来我们将读入数据的 75%的作为训练集，取剩余的 25%的数据做测试集。传入训练集和测试集参数：X_train, X_test, y_train, y_test。其中，X 对应着“留言主题”和“留言详情”，y 对应着 X 的一级标签。

由于这是有监督的训练模型，所以我们必须要同时输入 X_train_counts 和 y_train 这两个参数。用训练模型对测试集数据进行预测时，也要输入参数 X_test 和 y_test 这两个参数。

最后我们输出预测结果，如图 11。



	precision	recall	f1-score	support
城乡建设	0.84	0.90	0.87	516
环境保护	0.91	0.97	0.94	225
交通运输	0.91	0.64	0.75	154
教育文体	0.89	0.93	0.91	400
劳动和社会保障	0.87	0.93	0.90	483
商贸旅游	0.88	0.79	0.84	300
卫生计生	0.95	0.85	0.90	225
accuracy			0.88	2303
macro avg	0.89	0.86	0.87	2303
weighted avg	0.88	0.88	0.88	2303

图 11.任务 1 代码训练模型的结果展示

从图中展示的就是模型的学习成果。图中我们可以看到测试集中，一级标签个数只有 7 个，其准确率中，最准确的是卫生计生，达到了 95%；回召率中，最高的是教育问题已经达 88%。

F1-score 中，最准确的是教育文体，达到 91%，support 是测试集里一级标签对应的行数。在 accuracy 里，可知模型准确率为 88%，测试集的总条数为 2303 条。

最后两行分别是平均数和加权平均数的相关数据。

四、任务二

4.1 自然语言处理

4.1.1NLP 的介绍

要知道，机器是很难理解人们日常生活中使用的语言，那么如何才能让机器读懂我们的语言呢？这时，我们可以使用“自然语言处理”，即 NPL（Natural Language Processing），里面包含了所有用计算机对自然语言进行的操作。
在 Python 中，我们常用的对自然语言处理的第三方库有 NLTK、Gensim、Tensorflow 等。

4.1.2Gensim 的介绍

Gensim(generate similarity)是一个简单高效的自然语言处理 Python 库，用于抽取文档的语义主题（semantic topics）。Gensim 的输入是原始的、无结构的数字

文本（纯文本），内置的算法包括 Word2Vec，FastText，潜在语义分析（Latent Semantic Analysis，LSA），潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）等通过计算训练语料中的统计共现模式自动发现文档的语义结构。这些算法都是非监督的，这就意味着不需要人工输入，仅仅需要一组纯文本语料，任何纯文本（句子、短语、单词）就能采用语义表示简介地表达。

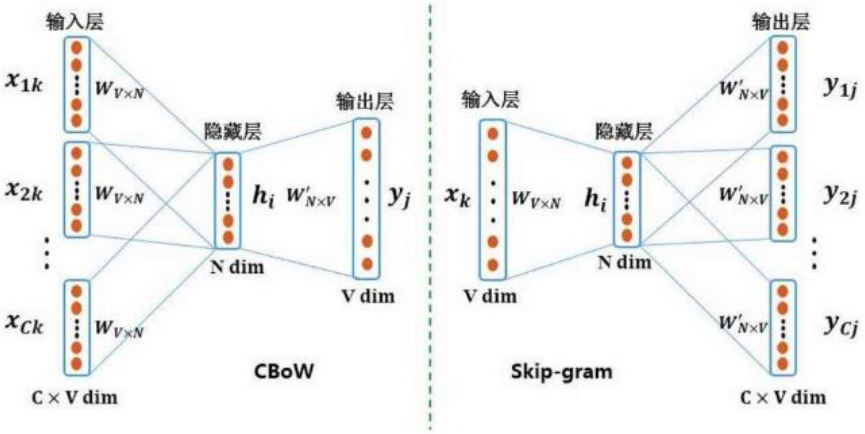


图 12.word2vec 的两种模型

而在此次项目中，我们主要利用了 Gensim 中的 corpora、models、similarities 来进行文本相似度的计算。首先我们利用 corpora 生成字典，形成语料库。接下来就是文本向量的转换，该步骤是 Gensim 的核心，通过挖掘语料中隐藏的语义结构特征，我们最终可以变换出一个简洁高效的文本向量，我们应用 models 初始化模型对象并进行文本向量的转换，最后我们利用 similarities 进行文本相似度的计算。

4.1.3 文本相似度的介绍

从信息论的角度阐明相似度与文本之间的共性和差异有关，共性越大、差异越小，相似度就越高；共性越小、差异越大，则相似度越低。相似度最大的情况是文本完全相同。同时提出基于假设推论出相似度定理，如下所示：

$$Sim(A,B) = \frac{logP(common(A,B))}{logP(description(A,B))}$$

图 13.相似度计算

其中，common(A,B）是 A 和 B 的共性信息，description(A,B)是描述 A 和 B 的全部信息，该公式表达出相似度与文本共性成正相关。

相似度一帮用[0，1]表示，该实数可以通过语义距离计算获得。相似度与语义距离呈反比关系，语义距离越小，相似度越高；语义距离越大，相似度越大，

我们通常用下面的公式来表示相似度与语义距离的关系：

$$Sim(S_A, S_B) = \frac{\alpha}{Dis(S_A, S_B) + \alpha}$$

图 14.相似度与语义距离

其中，$Dis(S_A,S_B)$表示文本 S_A，S_B 之间的非语义距离，α 为调节因子，保证了当语义距离为 0 时该公式的意义。

4.2 任务二代码解析

4.2.1 先要配置环境

这是我们需要配置的第三方库：

```
from gensim import corpora, models, similarities
```

图 15.

4.2.2 准备数据

附件 3 里一共有 4326 条数据，其中对于计算文本相似度相关的数据只有留言主题和留言详情，以为不同用户对问题的描述方法不同，有的用户在主题部分已经将问题的要点都描述出来，有的用户的问题要点都只写在了留言处，通过主题无法识别出用户要反映的问题。所以我们需要事先对附件 3 的留言主题和留言详情进行整合，为了避免一些无用的词对计算文本相似度的影响，我们要对文本进行分词和去停用词的处理。在这里，我们读入事先对这两列数据进行分词、去停用词和合并的 CSV 文件。

因为计算文本相似度后，目的是要生成一个对附件三内容进行分类的 Excel 表格，在查看附件 3 时，我们发现其时间格式有不同，因为最终计算热点问题，而热点与时间段有关，为了方便后续对时间的处理，我们要事先对时间格式进行统一。在这里，我们将时间统一为“%Y/%m/%d %H:%M:%S”的格式。

因为要对文本进行分类，我们选择的分类方法是通过计算文本相似度，若是相似的文本就在新增的问题 ID 列中加上与其他文本不同的 ID 号。因此，我们事先要对上述说的 CSV 文件中，新增一列，列名为“问题 ID”，然后都赋值为 0，以方便后续对文本的分类。

4.2.3 计算文本相似度

要找相似文本，我们得将上述处理后的文本放进词典，通过稀疏向量生成语料库，再通过 TF 模型算法计算出 TF 值，接着通过 token2id 得到特征数。最后将相似度按顺序放入一个列表里，最后得到一个二元列表。

做好上面的工作后，对文本进行可视的分类。我们的做法是先设定 ID=0，然后按顺序遍历整个文本，如果遇到对应文本的问题 ID 为 0，就 ID 就加 1，再接着上述提到的二元列表，如果相似度大于某个值，就都赋值同一个 ID 数值。以此类推，直到问题 ID 列里没有 0，就说明文本已经按问题 ID 分好类了。最后将问题 ID 列取出，变为列表。接着将问题 ID 列表与附件 3 中所有数据存成表格。在这里，我们给这个表格命名为"分类.xlsx"。

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	188006	A0001029	A3区一米阳光婚纱	2019/02/28 11:2	座落在A市A3区耳	0	0
2	188007	A0007479	咨询A6区道路命名	2019/02/14 20:(	A市A6区道路命名	0	1
3	188031	A0004006	反映A7县春华镇金	2019/07/19 18:1	本人系春华镇金	0	1
4	188039	A0008137	A2区黄兴路步行街	2019/08/19 11:4	靠近黄兴路步行	0	1
5	188059	A0002857	A市A3区中海国际社	2019/11/22 16:5	A市A3区中海国际	0	0
6	188073	A909164	A3区麓泉社区单方	2019/03/11 11:4	作为麓泉社区麓	0	0
7	188074	A909092	A2区富绿新村房产	2019/01/31 20:1	"二高一部"发出	0	0
8	188119	A0003502	对A市地铁违规用工	2019/05/27 16:(	我是一名在A市某	0	0
9	188170	A8801132	A市6路公交车随意	2019/12/23 08:5	12月21日下午17	0	0
10	188249	A0008408	A3区保利麓谷林语	2019/09/17 04:2	保利麓谷林语桐	0	0
11	188251	A0001309	A7县特立路与东四	2019/10/19 11:(	近来，下午晚高	0	0
12	188260	A0005348	A3区青青家园小区	2019/05/31 17:(	还我宁静我要复	0	0
13	188396	A0004758	关于拆除聚美龙楚	2019/04/15 16:2	桐梓坡589号白	2	1

图 16.分类.xlsx

4.2.4 做热点问题明细表

如何定义为热点？我们认为至少同一类问题出现的次数在两次以上。所以我们将同一类问题出现次数只有一次的数据删掉，这个可以通过计算不同问题 ID 出现的次数来过滤掉出现次数为一的数据来实现。接着我们将过滤后的数据放到 DataFrame 里。

接着我们开始为计算热点指数做准备工作。因为我们考虑的热点指数对应的参数有 3 个，分别是反映次数，持续时间，点赞数和反对数的和。为什么要对点赞数和反对数求和？因为无论是赞同还是反对，都反映了人群对这个问题的关注度，我们要求的热度的背后，终究是人们对该问题的关注度，这是我们要对两者进行求和的原因。所以，这一步的准备工作是从过滤后的数据中，得到同一类问题的反映次数，同一类问题的总的点赞数和和总的反对数的和，持续时间长度。

对于反映次数，我们要将过滤后的数据，按问题 ID 的从小到大的顺序，统计同一类问题的出现次数，并将他们按顺序存放的列表里。对于点赞和反对的求和，我们先要对这两列求和，将数据放到列表里，再对表格新起一列，让它等于上述求和的列表。对于时间持续长度问题，我们的做法是循环同类问题，将时间列变为时间戳，放到列表里，找到其中最大和最小的时间戳，令两者相减，因为时间戳是以秒为单位，为了方面后面对热点指数的计算，我们要将它变为以天为

单位，但是为了保证其准确性，我们直接用时间戳的极差除一天的秒数，即 86400 秒。

下一步就是计算热点指数。先前，我们得先定义热点指数的计算方法。我们认为，热点指数跟反映数，持续时间长度，关注度相关。那么该怎么定义算法呢？我们认为最主要的数据是反映数，因为如果一个问题反映的次数过少，说明人们对该问题的不太关心，问题的重要程度不足以让人群专门到该网站上去反映问题。其次是时间持续的长度，到底是要让反映次数除时间，来反映一段时间内问题的爆发性，还是要分开反映次数和时间，来反映一个问题在一段时间内持续被人群反映?我们最终觉得在短时间内突然出现的问题不足以反映人群对该问题的重视程度。而那些在一段时间内持续出现的问题，才能反映人群对其的重视程度。所以，我们决定将反映次数于时间持续长度分开计算。对于关注度，有些问题的的点赞数和反对数的和可以超过 2000，而有些问题的点赞数和反对数的和加起来也只有 0，可以很直观的看到人群对不同问题的关注程度。

下一个问题是是否对热点指数定义一个标准？我们对此进行了实验。

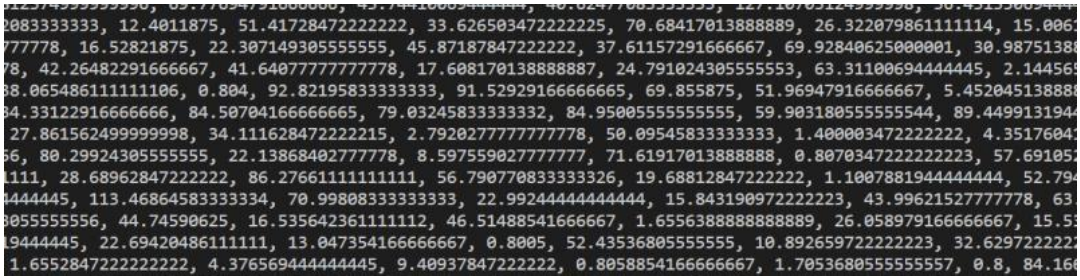


图 17.不定义热度指数

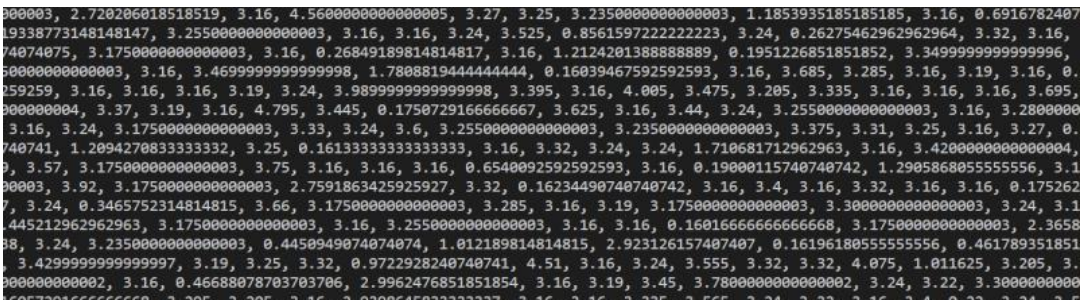


图 18.定义热度指数

从图中我们可以发现不定义一个标准的问题，就是无法通过看热度指数知道一个问题的热度大小。比如当你看见一个问题的热度指数为 10，接着你看到还有个问题的热度指数为 50，你以为这个问题是人们关注的热点问题，可是后面还有一个问题热度为 100 的。这时，只要我们定义了一个标准，就可以很直观的知道一个问题的热度大小。通过观察之前得到的分类表的数据，我们发现，同一类问题反映次数最多的是 42 条，最长的持续时间差不多超过了一年，一个问题受到最多的点赞数是 2097 条，点赞数超过 1000 的有两条，有不超过 10 条的问题的点赞数和反对数的和大于 200，其余的问题都小于 100。可见，小于 100 的条数占了大多数。所以，我们决定将同一类问题反映次数超过 50 的，就定义它得到反映数的满分，时间持续长度超过 30 天的，就定义它得到了时间的满分，同一类问题的总的点赞数和反对数的和超过 200 的，就定义它得到关注度的满分。我们先定义热度指数的满分为 10 分。最重要的参数是反映次数，其次是时

间和关注度。如何调整三者比重呢？我们通过多次试验，观察最终得出的排名，再调整比重，最终得出较佳的比重是：反映次数占 60%，时间持续长度占比 20%，关注度占比 20%。我们定义了一个专门计算热度指数的函数来计算热度指数。如图：

```
def redu(tiao,timeX,hot):
    N=tiao/50  #这里是定义条数为50.就得到条数的满分
    X=timeX/30  #这里是定义时间持续超过30天，就得到时间的满分
    H=hot/200   #这里是定义总点赞+反对数超过200，就得到关注度的满分
    if N>1:
        N=1
    if X>1:
        X=1
    if H>1:
        H=1   #使其不超过总满分的方法
    HOT=N*6+X*2+H*2    #满分为10分，其中条数占60%，时间占%20，关注度占%20
    return HOT  #返回热度指数
```

图 19.计算热度指数函数

下一步是通过热度指数，给数据做排名。因为最终我们只需要得到排名前五类的问题。所以这里，我们只通过热度指数得到排名前五的问题 ID 和其在之前过滤后的按 ID 号从大到小的关于不同类别问题的长度的列表中，对应的索引。将排前五是数据一次取出，并将其问题 ID 按排名顺序依次重新赋值为 1，2，3，4，5。将重新组合的前五的数据，按分类表的格式，存成热点问题明细表。到此，热点问题明细表就完成了。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	190337	A0009051	关于伊景园滨河苑捆	2019/08/	投诉伊景	0	0
1	191001	A909171	A市伊景园滨河苑协商	2019/08/	商品房伊	12	1
1	195511	A909237	车位捆绑违规销售	2019/08/	对于伊景	0	0
1	195995	A909199	关于广铁集团铁路职	2019/08/	尊敬的市	0	0
1	196264	A0009508	投诉A市伊景园滨河苑	2019/08/	A市伊景园	0	0
1	204960	A909192	家里本来就困难,还要	2019/08/	我是广铁	0	0
1	205277	A909234	伊景园滨河苑捆绑车	2019/08/	广铁集团	1	0
1	205982	A909168	坚决反对伊景园滨河	2019/08/	我坚决反	2	0
1	207243	A909175	伊景园滨河苑强行捆	2019/08/	您好！A市	0	0
1	209571	A909200	伊景园滨河苑项目绑	2019/08/	广铁集团	0	0
1	212323	A0002070	广铁集团要求员工购	2019/07/	尊敬的领	0	0
1	214975	A909182	关于房伊景园滨河苑	2019/08/	尊敬的领	3	0
1	218709	A0001066	A市伊景园滨河苑捆绑	2019/08/	伊景园滨	1	0

图 20.热点问题明细表

4.2.5 制作热点问题表

做这个表需要我们得到热点问题前五的排名，问题 ID，时间范围，地点/人群，问题描述。对于排名，因为在上一步，我们已经按照热度指数的排名，给问题 ID 分别赋予对应的新的问题 ID 号，所以排名可以等于问题 ID。热度指数在

上一步也已经得到关于前五的热度指数（按从大到小排列）的列表。时间范围和上一步一样，将同一问题 ID 的问题的时间按时间戳格式放进列表里，取出最大时间戳和最小时间戳，然后又将两个时间戳变为 "%Y/%m/%d" 格式，然后将两字符串以 时间-至-时间 的字符串格式放到一个列表里。对于地点/人群，通过观察数据，我们发现，一般在留言主题和留言详情里，大多数都是至少有出现[A-Z]市等，但是其描述地点方式有很多不同，有的是市-区-小区，有的是市-社区，有时是区-路，等等。无法直接按照一定格式取出地点/人群，但我们发现，大多数数据都是有[A-Z](市名)，所以我们用了一个投机取巧的方式，就是用 Re，找到里面从[A-Z]开始，到后面 7 个字的字符串，以此来作为地点/人群，对于问题描述，因为依据于所有归并一类问题的前提是其问题的相似性，所以我们认为，同一问题的任意一个主题在大多数情况下，都是具有代表性的，所以我们为了方便，选择用同一类问题的最后一条的主题作为热点问题的问题描述。最后，将数据合并，存成热点问题表。至此，任务 2 完成。

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	7.3	2019/07/07 至 2019/09/01	A市政府就广州局	投诉A市伊景园滨河院捆绑销售车位
2	2	4.47	2019/01/02 至 2019/11/08	A市经开区东四线	问问A市经开区东四线以西、新安路以南的规划
3	3	4.38	2019/04/10 至 2019/12/30	A7县政府在周边	反映A市地铁 3 号线松雅湖站点附近地下通道问题
4	4	4.36	2019/02/21 至 2019/08/13	A市A2区城南派	西地省富惠天下商务有限公司涉嫌经济诈骗
5	5	4.24	2019/01/08 至 2019/12/27	A市金晖优步花园	关于A市金晖优步花园相关问题的反映

图 21.热点问题表

五、任务三

数据质量是指在业务环境下，数据符合数据消费者的使用目的，能满足业务场景具体需求的程度体需求的程度。根据此次任务的需要，进一步参考信息质量评价指标并结合附件 4 的文本数据特征，提出答复意见质量评价方案。

5.1 答复意见质量评价指标

（1）完整性

在这里将完整性理解为对答复意见的文本数据是否完整的研究，信息必须具备一定的结构，一条答复意见中应包含对留言者问题的回应和对留言问题的相关措施，如果缺少了其中一项，结构的不完整导致表达的不完整。

（2）正确性

答复意见中所给的信息需客观、真实、准确，准确则需检索内容中的文字有无拼写与语法的错误、语序是否通畅，词语的涵义是否含糊等。内容的正确性才不会使留言者的判断产生影响。

（3）可解释性

在《数据整理实践指南》中用“4C”概念概括了数据分析，即完整性(Complete）、一致性(Coherent）、准确性(Correct)以及可解释性

（aCcountable），在此借用“4C”概念中的可解释性用于分析。可解释性本来指的是可以追踪到数据的来源，在此次分析中将可解释性理解为可读性，文本数据的用语是否简洁明了，答复意见的可读性可用自动化可读性指数 ARI，其计算公式如下，数值近似等于我们理解一段文字的最低程度[6]。

$$ARI = 4.71 *（总字符数/总字数）+ 0.5*（总字数/总句数）- 21.43$$

（4）相关性

相关性指的是内容与留言者目的相关，本文讨论的相关性主要为答复意见与相对应留言详情之间的相关度，答复的内容是否贴切于留言中所提出的问题进行答复，是否能满足留言者的需求。

（5）时效性

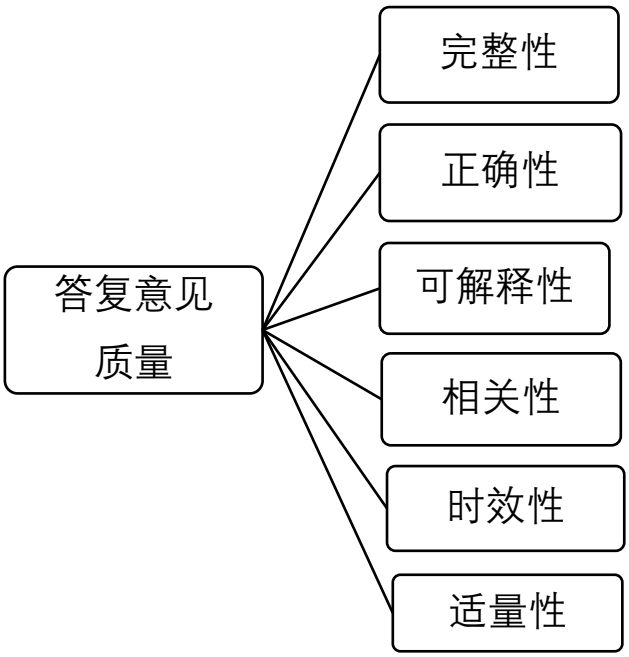
本文中的时效性针对留言时间与答复时间，如果答复时间与留言时间间隔过长，答复意见就会没有价值，在一定的时间限制内，答复意见对留言者的帮助才会有意义，通过计算二者间的时差反映出答复效率，时差越短，留言处理效率越高，体现答复意见的效用质量。

（6）适量性

过少的答复内容，会使留言者得不到自身所需的信息；然而过多的答复内容会导致文本冗余，留言者无法从中准确获取自身所需的信息。答复意见应简洁、适量，同时在必要部分有相应注释，便于留言者理解。

5.2 答复意见质量评价指标体系

遵循评价指标体系构建的科学性、系统性、客观性和发展性等原则，构建了针对关于留言的答复意见质量评价的指标体系，如下图所示：



	一级指标	指标定义
答复意见质量研究指标体系A	完整性B1	答复意见的内容与结构的完整性，是否缺失数据
	正确性B2	内容的客观真实性，是否有重复、模糊和错误数据
	可解释性B3	答复意见的表达的被理解性和可读性程度
	相关性B4	答复意见与留言者的留言需求之间的相关程度
	时效性B5	答复时间与留言时间之间的差距
	适量性B6	留言者对答复意见信息的利用率

图 22. 答复意见质量研究指标体系说明

5.3 评价方法

此次选用层次分析法对答复意见质量进行评价，这是一种定性和定量相结合的、系统化、层次化的分析方法，应用简单、方便理解、具有较强的实用性，对主观意向较强的评价更为适用。主要步骤如下：

1、构建判断/比较矩阵；

2、计算权重；

3、进行一致性检验。

5.3.1 构建判断/比较矩阵

比较第 i 个元素与第 j 个元素时，使用数量化的相对权重 a_{ij} 描述。设共有 n 个元素参与比较，$A=(a_{ij})_{nxn}$ 则称为成对比较矩阵。

成对比较矩阵中 a_{ij} 的取值可参考 Satty 的提议，按下述标度进行赋值。a_{ij} 在 1-9 及其倒数中间取值。

a_{ij} =1，元素 i 与元素 j 同等重要；

a_{ij} =3，元素 i 比元素 j 稍微重要；
a_{ij} =5，元素 i 比元素 j 较强重要；
a_{ij} =7，元素 i 比元素 j 强烈重要；
a_{ij} =9，元素 i 比元素 j 极端重要；
a_{ij} =2*n*，n=1,2,3,4，元素 i 与 j 的重要性介于 a_{ij} = 2n-1 与 a_{ij} = 2n+1 之间；

$$a_{ij} = \frac{1}{n}$$

n=1,2,……,9，当且仅当 a_{ij} =n。

元素i相比元素j	分值
同等重要	1
稍微重要	3
较强重要	5
强烈重要	7
极端重要	9
两相邻判断的中间值	2，4，6，8

图 23.分值划分

通过各因素之间的两两比较确定合适的标度。在建立层次结构之后，需要比较因子及下属指标的各个比重，为实现定性向定量转化需要有定量的标度，本文选用问卷调查的方法依照 1-9 比率标度法构造判断矩阵，如表所示。

评价	完整性	正确性	可解释性	相关性	时效性	适量性
完整性	1	2	3	5	3	2
正确性	1/2	1	3	2	5	5
可解释性	1/3	1/3	1	1/3	2	2
相关性	1/5	1/2	3	1	3	3
时效性	1/3	1/5	1/2	1/3	1	2
适量性	1/2	1/5	1/2	1/3	1/2	1

图 24.判断矩阵

5.3.2 权重计算和一致性检验

在构建判断矩阵时，有可能会出现逻辑性错误，如 A 比 B 重要，B 比 C 重要，但却又出现 C 比 A 重要。

一致性检验是指对判断矩阵确定不一致的允许范围，由最大特征值 λmax 是否等于 n 来检验判断矩阵是否为一致矩阵。具体的一致性指标用 CI 计算，CI 越小，说明一致性越大。CI=0，有完全的一致性；CI 接近于 0，有满意的一致性；CI 越大，不一致越严重。

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

RI 值可直接查表得出，随机一致性指标 RI 和判断矩阵的阶数有关，一般情况下，矩阵阶数越大，则出现一致性随机偏离的可能性也越大。

维度	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

图 25.一致性指标 RI

最终使用的检验统计量为检验系数 CR，公式如下：

$$CR = \frac{CI}{RI}$$

CR 值小于 0.1，则说明通过一致性检验，反之则说明没有通过一致性检验。如果数据没有通过一致性检验，此时需要检查是否存在逻辑问题等，重新录入判断矩阵进行分析[6]。最后得出一致性检验结果如下表所示。

项	指标权重	最大特征值	CI值	CR值	一致性检验结果
完整性	0.35	6.59	0.12	0.09	通过
正确性	0.26				
可解释性	0.1				
相关性	0.16				
时效性	0.07				
适量性	0.06				

图 26.一致性检验结果

5.3.3 质量综合评价

为降低操作的复杂性， 评价指标为：按不符合-2， 一般符合-4, 符合-6， 较为符合-8， 非常符合-10 进行评分。

评价过程中, 留言者通过依次浏览答复意见, 以自身的帮助程度为评分标准, 进行指标打分, 最后取留言者评分的平均值作为该指标的分数。

根据公式进行测评结果的计算，综合测评得分= $\sum_{i=1}^{n} BiWi$

其中 Wi 指的是第 i 个测评指标的权重 Bi 是指第 i 个测评指标的评价得分，前面已通过层次分析法对各项指标权重进行了计算, 因此通过公式可计算得到对答复意见的综合得分[7]。

六、模型优化

对于热度指数的计算方法，只用了简单的方法进行计算。但热点指数的背后有很多指标，需要精确的算法，这个有待改进。

对于热点问题表中的地点/人群，只利用文本相似度计算的话难通过地点和人群对问题进行分类，通过查看我们生成的热点问题明细表，我们发现，有问题 ID 为 5 的问题的主要内容都是关于房子装修的问题，但是其地点/人群有不同。如何取出地点和人群呢？我们想到了一个方法，通过 Re，找到大写字母及后面几个字的字符串，但是这样做是不正确的，万一用户没有写市名或区名的话，就会出现找不到地名的问题。

在后续的优化中，我们可以在分词，加入关于地点描述的多种格式的词放到词典里。这样，Jieba 分词就不会分开地点了。在这个题目里，关于市名，区名都是以字母和数字来取名，而在现实生活中，不可能有这样的情况出现，若要识别出地名，我们需要 Github 这个库，它可以精准识别出省市区路号楼等名字，这样就可以找到文本中的地名了。

对于问题描述，因为我们认为相似文本的主题基本上都有代表性，所以，我们只是取出任意一问题的主题作为一类问题的描述，但是会有的留言主题的描述不够准确，这样做就有可能取到不具有代表性的主题。这样的做法是冒险的，后续的优化中我们可以用到：TextRank。它可以自动对文本做摘要。我们只需要读入留言主题和留言详情作为文本即可。

七、总结

本文主要通过文本挖掘和自然语言等技术建立热度评议模型。

我们首先对数据进行预处理的操作，使得后续的数据的处理；接下来我们利用多项式朴素叶贝斯模型，让机器学习对留言内容的标签进行预测，形成留言分类标签；然后利用 Genism 对文本进行相似度计算，并形成表格，后续的热点问题表和热点问题明细表则通过读取上述表格来定义热度指数算法，制作热点问题表以及热点问题明细表；最后，我们制作答复意见质量评价指标，从完整性、正确性、可解释性、相关性、时效性、适量性的角度进行评定。

八、参考文献

［1］作者:千寻~·机器学习中的有监督学习，无监督学习，半监督学习的区别[DB/OL](2017-02-28)[2020-05-06]
https://blog.csdn.net/u011630575/article/details/58659372/
［2］作者：rocling·二分类、多分类与多标签问题的区别,对应损失函数的选择,你知道吗？[DB/OL](2019-04-09)[2020-05-06]
https://blog.csdn.net/rocling/article/details/89165463
［3］作者：米勒 111·sklearn 多分类多标签算法[DB/OL](2018-11-26)[2020-05-06]
https://blog.csdn.net/mixiaolemy/article/details/84529051
［4］作者:AI 算法工程师 YC 朴素贝叶斯的三个常用模型:高斯(GaussianNB)、多项式(multinomial model)、伯努利(Bernoulli model)（2019-11-14）[2020-05-06]
https://blog.csdn.net/qq_36134437/java/article/details/103065030
［5］作者：迷途未迷·文本特征提取之 CountVectorizer TfidfVectorizer 中文处理[DB/OL](2018-11-13)[2020-05-06]
［6]郭银灵.基于文本分析的在线评论质量评价模型研究[D].内蒙古:内蒙古大学计算机学院,2017:21-22.
［7］ 徐嘉徽.电子商务用户在线评论信息质量研究[D].吉林：吉林大学管理学院，2016.

九、关于该项目所用代码以及文件介绍

stoplist.txt : 停用词表
data_2.txt : 运行 task_0.py 后生成的文件 ， 运行task_1.py 时用到
data_3.txt： 运行task_0.py 后生成的文件 ， 运行task_2_1.py 时用到
data_3_4.txt : 运行 task_0.py 后生成的文件， 运行 task_3_1.py 时用到
data_3_3.txt : 运行 task_0.py 后生成的文件， 运行 task_3_1.py 要用到
task_0.py : 预处理
task_1.py : 任务1
task_2_1.py : 任务2的前一半
task_2_2.py : 任务2的后一半
task_3_1.py : 任务3的一部分
task_3_2.py ： 任务3的一部分
任务3的相关表格.docx
分类.xlsx : 运行 task_2_1.py 后生成的，运行task_2_2.py 要用到
热点问题表、热点问题留言明细表 是结果