“智慧政务”中的文本挖掘应用

摘 要

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。

针对问题 1，我们首先将附件 2 中一级标签的七个类别分别用数字代替，然后从每个类别中随机抽取 600 条数据构成新的数据集，利用 jieba 分词对留言详情进行分词、去停用词。其次利用 TF-IDF 算法计算留言详情的 TF-IDF 值，然后将 TF-IDF 值划分为训练集和测试集。最后使用多分类 Logistic 算法对训练集进行建模，并在测试集上验证，最终模型的 F-Score 值为 88%。

针对问题 2，我们首先利用 Excel 将附件 3 的数据按点赞数降序排列，然后使用 Pyhton 中的 drop_duplicates 函数对留言详情去重，利用 jieba 分词对留言详情做分词、去停用词处理，再利用 TF-IDF 算法计算留言详情的 TF-IDF 值，通过 TF-IDF 值计算留言详情间的相似度。分别找出与点赞数超过 100 的问题相似的问题，保存在 data 文件夹下。最后利用 K-Means 算法对剩下的留言详情聚类，计算每个簇的平均相似度，如果簇内平均相似度大于 0.2，则将其存放到 Excel 中，并保存在 data 文件夹下。定义热度指数，分别计算 data 文件夹下各类问题的热度指数，最终得到排名前 5 的热点问题。

针对问题 3，从三个方面量化答复意见：对留言主题、留言详情进行分词、去除停用词、计算词频、通过语料库建立词典；将答复意见进行 jieba 分词、去除停用词、通过 doc2bow 转化为稀疏向量从而得到新语料库；将新语料库通过 tfidfmodel 进行处理，得到 TF-IDF 值；通过 token2id 得到特征数、稀疏矩阵相似度，从而建立索引，得到最终相似度结果；对答复建议进行情感分析——礼貌性用语、回复格式、具体细节进行量化；答复时间是否及时回复；综合以上三个因素对答复意见进行打分。

关键词：Jieba 分词；TF-IDF 算法；Logistic 算法；K-Means 聚类；情感分析

Application of text mining in "intelligent government affairs"

Abstract

In recent years, with WeChat, such as weibo,mayor mailbox, sun hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government.

Firstly, level 1 label in attachment 2 of the seven categories using Numbers instead of separately, and then randomly selected 600 data from each category a new data set, message details to make use of jieba participle word segmentation, to stop words, reuse of TF - IDF algorithm to calculate the message details tfidf value, then put the tfidf value is divided into training set and test set, the final classification using multiple Logistic algorithm modeling was carried out on the training set, and tested on the test set, the final model of F - Score value of 88%.

Secondly, We first used Excel to arrange the data in annex 3 in descending order of thumb up number, then used the drop_duplicates function in Pyhton to deduplicate the message details, used jieba word segmentation to do word segmentation and stop and stop word processing for the message details, and then used tf-idf algorithm to calculate the tf-idf value of the message details, and calculated the similarity between the message details through the tf-idf value. Find out the similar problems with thumb up number over 100 and save them in the data folder. Finally, k-means algorithm is used to cluster the remaining message details to calculate the average similarity of each cluster. If the average similarity within the cluster is greater than 0.2, it will be stored in Excel and saved in the data folder. Define the heat index, calculate the heat

index of various problems under the data folder respectively, and finally get the top 5 hot issues.

Thirdly, Quantified the response from three aspects: dividing words for the subject and details of the message, removing the stop words, calculating the word frequency, and establishing a dictionary through corpus; Jieba word segmentation, the removal of stop words, and the conversion of doc2bow into sparse vector to obtain the new corpus; The new corpus was processed by tfidfmodel to obtain the tf-idf value. The similarity of feature number and sparse matrix is obtained by token2id, and then the index is established to obtain the final similarity result. Emotional analysis of response Suggestions -- quantification of polite language, response format, and specific details; Whether the reply time is timely; Based on the above three factors, the replies were rated.

Keywords：Jieba participle，TF-IDF algorithm，Logistic algorithm，K-Means clustering，Sentiment analysis

目 录

1. 挖掘背景与研究现状

1.1 挖掘背景

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，也是智慧城市建设中的重要课题之一。

1.2 研究现状

孙海锋等[1]运用网络文本分析和数据挖掘技术对网络招聘信息进行分析与挖掘，得出各个职业类型对应的专业领域，预测出相关职位的需求，给在校大学生的就业规划提出了可行性的建议。李春林等[2]以新能源汽车的评论数据为研究对象，利用文本挖掘方法进行情感分析，并挖掘出用户对新能源汽车的关注点，对商家改进和服务销售具有积极意义。文本数据挖掘与我们的生活息息相关，它所涉及的应用领域十分广泛，具有重要作用。

2. 问题 1 分析方法与过程

2.1 流程图

这里，我们通过对问题进行详细的分析，得到如下处理流程图。

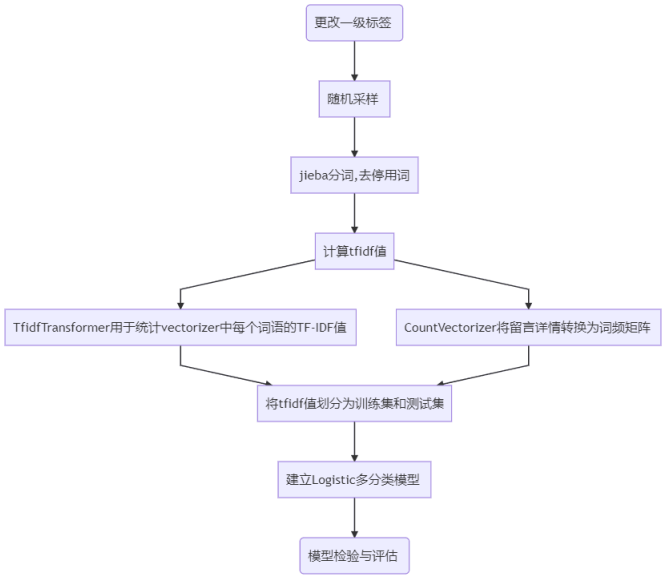

图 1：问题 1 处理流程图

2.2 更改一级分类标签

在附件 2 的数据中，一级分类标签的值全为中文，这不利于之后分类模型的建立。因此将标签用 0,1,2,3,4,5,6 代替，更改后的标签如下表所示。

表 1：标签对应表

一级标签	Label
城乡建设	0
环境保护	1
交通运输	2
教育文体	3
劳动和社会保障	4
商贸旅游	5
卫生计生	6

2.3 随机采样

在附件 2 中，每类标签下的问题数目都是不一样的。由图 2 的类目数量分布图可以看出，环境保护、卫生计生和交通运输类的数量远小于城乡建设和劳动和社会保障的数量，而机器学习算法就是从大量的数据集中通过计算得到某些经验，进而判定某些数据的正常与否。不均衡数据集显然少数类的数量太少，模型会更倾向于多数集。因此，直接用附件 2 的数据进行建模是不合适的。于是，从这七个类中随机抽取 600 条数据构成新的数据集，并保存到 data_new.xls 中。

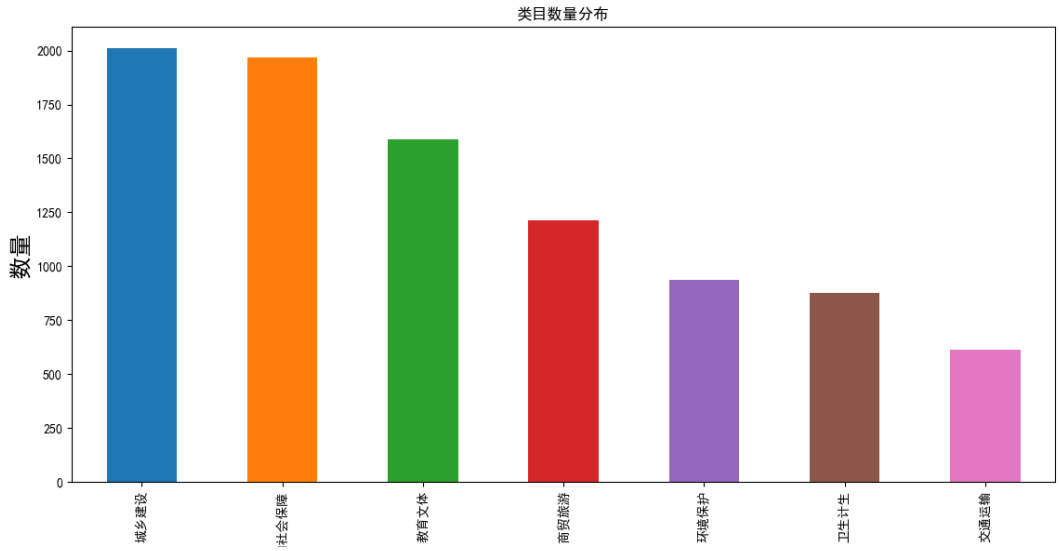

图 2：类目数量分布图

2.4jieba 分词、去停用词

利用 Python 中的 jieba 中文分词库对留言详情进行分词。jieba 分词主要是基于统计词典，构造一个前缀词典；然后利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置，构造一个有向无环图；通过动态规划算法，计算得到最大概率路径，也就得到了最终的切分形式。

语气助词、副词、介词、连接词等，通常自身并无明确的意义，只有将其放入一个完整的句子中才有一定作用，如常见的“的”、“在”之类的词。因此，jieba 分词之后，需要去除这些无意思的词，即停用词。

2.5 词云图绘制

在 jieba 分词、去停用词之后，已经将留言详情切分成了词语或者词，为了更加直观的显示每个类别中比较重要的词语，分别绘制每个类别的词云图。词云图是对文本数据中出现频率较高的关键词予以视觉上的突出，形成“关键词的渲染”，就类似云一样的彩色图片，从而过滤掉大量的文本信息，使人一眼就可以领略文本数据的主要表达意思。代码 Word_cloud.py 中自定义了一种绘制词云图的函数，通过调用 wordcloud()函数就可以绘制出各个类别的词云图。图中词语越大，代表在城乡建设类中出现该词语的频率越高。图 3 为城乡建设的词云图，由图 2 可以看出，开发商、业主、小区、建设在城乡建设类别中出现的频率较高。

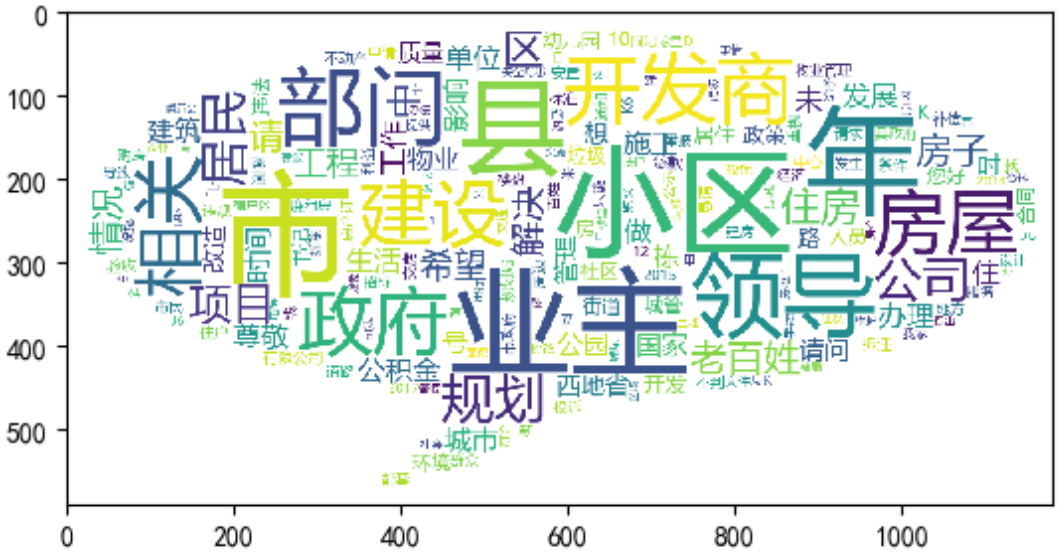

图 3：城乡建设词云图

2.6 计算 TF-IDF 值

在对留言详情进行建模之前，要将非结构化的数据转化为结构化的数据。首先将分词、去停用词之后的留言详情转换为词频矩阵，这里采用 Python 中

CountVectorizer 函数计算词频矩阵。其次利用 TF-IDF 算法将词频矩阵转化为 TF-IDF 值。TF-IDF 算法原理如下：

第一步：计算词频 TF（term frequency）：

$$\text{词频（TF）}=\text{某个词在文章中出现的次数} \qquad (1)$$

考虑到文章有长短之分，为了便于不同文章的比较，这里进行“词频”标准化

$$\text{词频（TF）}=\frac{\text{某个词在文章中出现的次数}}{\text{文章的总词数}} \qquad (2)$$

第二步：计算逆文档频率 IDF（inverse document frequency）

$$\text{逆文档频文（IDF）}=\log\left(\frac{\text{文档总数}}{\text{出现该词的文档数}+1}\right) \qquad (3)$$

如果一个词越常见，那么分母就越大，逆文档频率就越小、越接近 0。分母之所以要加 1，是为了避免分母为 0（即所有文档都不包含该词）。log 表示对得到的值取对数。

第三步：计算 TF-IDF 值

$$TF\text{-}IDF=TF\times IDF \qquad (4)$$

可以看到，TF-IDF 值与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。

2.7 模型建立与评估

计算出 TF-IDF 值之后，已经成功将非结构化的数据转化成了结构化的数据，那么可以通过 TF-IDF 值来构建分类模型。首先将留言详情的 TF-IDF 值划分为训练集和测试集，其中训练集所占比例为 0.7，测试集所占比例为 0.3。其次用多分类 Logistic 算法在训练集上拟合，在测试集上验证。Logistic 算法原理如下：

Logistic 回归属于概率型非线性回归模型，是研究分类观察结果（Y）与一些影响因素（X）之间关系的一种多变量分析方法。假设对一个试验样本在一组自变量作用下所发生的概率用 P 表示，则该事件不发生的概率为 1-P，发生概率与不发生概率之比记做“优势”，对其取自然对数，则得到 Logistic 函数：

$$F(p)=\ln\left(\frac{p}{1-p}\right) \qquad (5)$$

则 Logistic 回归模型为：

$$F(p)=\ln(\frac{p}{1-p})=b_0+b_1x_1+b_2x_2+.....+b_mx_m \qquad (6)$$

其中，b_0 为常数项，b_1，$b_2.....b_m$ 称为回归系数。从式(6)中可以看出，当 p 在(0,1)之间变化时，对应的 $F(p)$ 在（-∞，+∞）之间变化，这样自变量可在任意范围取值。同时根据式(6)得到 x_1，$x_2.....x_m$ 可在任意范围取值。根据式(6)又得到：

$$Y=\frac{ea}{(1+ea)} \qquad (7)$$

其中，$a=b_0+b_1x_1+b_2x_2+.....+b_mx_m$。

对于二分类问题常用的评价指标是精确率（Precision）和召回率(Recall)，通常以关注的类为正类，其他类为负类，分类器在数据集上的预测或者正确或者不正确，有 4 种情况：

TP：True Positive，把正类预测为正类；

FP：False Positive，把正类预测为负类；

TN：True Negative，把负类预测为负类；

FN：False Negative，把正类预测为负类。

在混淆矩阵中如表 2 所示。

表 2：混淆矩阵表示的预测情况

真实值 预测值	Positive	Negative
正	TP	TN
负	FP	FN

精确率(Precision)：精确率是指在预测为正类的样本中真正类所占的比例：

$$P=\frac{TP}{(TP+FP)} \qquad (8)$$

召回率(Recall)：召回率是指在所有的正类中被预测为正类的比例：

$$R=\frac{TP}{(TP+FN)} \qquad (9)$$

对于多分类问题，通常使用 F-Score 对分类方法进行评价：

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \qquad (10)$$

其中， p_i 和 R_i 分别为第 i 类的精确率和召回率。

使用多分类 Logistic 算法在测试集上的 F1 值为 0.88，为了进一步检验该模型的有效性，又分别用随机森林、高斯贝叶斯等算法来拟合模型，结果如表 3 所示。由表 3 中各类算法对比可以看出，使用 Logistic 算法进行拟合要优于其他算法。

表 3：算法对比

Model	Precision	Recall	F1-Score
Logistic	0.88	0.88	0.88
GaussianNB	0.71	0.72	0.71
RandomForest	0.84	0.83	0.83

3. 问题 2 分析方法与过程

3.1 流程图

这里，我们通过对问题进行详细的分析，得到如下处理流程图。

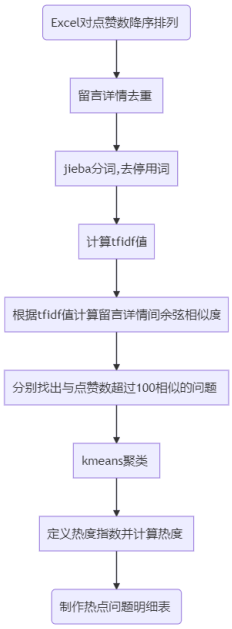

图 4：问题 2 处理流程图

3.2 将附件 3 的数据按点赞数降序排列

点赞数是衡量热点问题的重要指标，通过点赞数的多少可以间接反映人们对这一问题的关注度，同时为了避免点赞数多但是反映该问题人数较少而导致热点问题被忽略情况的出现，首先用 Excel 对附件 3 的数据按点赞数的多少进行降序排列。表 4 为点赞数超过 100 的留言主题，由表 4 可以初步看出，群众对 A 市

58 车贷案和 A 市 A5 区汇金路五矿万境 K9 县存在一系列问题的关注度较高。

表 4：留言主题及相应点赞数

留言主题	点赞数
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2097
反映 A 市金毛湾配套入学的问题	1762
请书记关注 A 市 A4 区 58 车贷案	821
严惩 A 市 58 车贷特大集资诈骗案保护伞	790
承办 A 市 58 车贷案警官应跟进关注留言	733
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？	669
A 市富绿物业丽发新城强行断业主家水	242

3.3 留言详情去重

在附件 3 给出的数据中，存在一些相同的留言详情，即一个群众可能多次反映同一问题，而同一问题的多次反映情况对于挖掘热点问题而言是没有帮助的。因此利用 Python 中 drop_duplicates 函数对留言详情去重，去重前的数据共有 4326 条，去重后的数据有 4225 条。

3.4 留言详情分词、去停用词、计算 TFIDF 值

与问题 1 类似，在这里同样需要对留言详情做分词、去停用词处理，为了便于之后的分析，还需利用 TF-IDF 算法将非结构化的数据转化为结构化的数据。

3.5 计算留言详情间的相似度

将留言详情转化为 TF-IDF 值之后，留言详情已经映射到向量空间，可以通过 TF-IDF 值计算留言详情间的余弦相似度，得到相似度为 4225×4225 的对称矩阵，存放在 similarity 中。

对于多个不同的文本或者短文本对话消息要来计算他们之间的相似度如何，一个好的做法就是将这些文本中的词语，映射到向量空间，形成文本中文字和向量数据的映射关系，通过计算几个或者多个不同向量的差异的大小，来计算文本的相似度。这里用到的是向量空间余弦相似度(Cosine Similarity)。

向量空间余弦相似度(Cosine Similarity)是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。余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫“余弦相似度”。

假设有 a，b 两个高维向量，其中 $a=(x_1,x_2,.....,x_n)$，$b=(y_1,y_2,.....,y_n)$。那么 a，b 的余弦相似度为：

$$Similarity = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \qquad (11)$$

3.6 筛选出与点赞数超过 100 的问题相似的问题

为了避免点赞数高而反映人数少的问题被忽略情况的出现，这里筛选出与点赞数超过 100 的问题相似的、相似度超过 0.25 的问题。

3.7K-Means 聚类

将与点赞数超过 100 的问题相似的问题筛选出之后，对剩下的留言详情进行 K-Means 聚类，设置初始簇中心 k =150。K-Means 聚类的基本思想是将每一个样本分配给最近中心（均值）的类中，具体的算法至少包括以下四个步骤：

(1)从 n 个数据对象随机选取 k 个对象作为初始簇中心；

(2)计算每个簇的平均值 并用该平均值代表相应的簇；

(3)计算每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分；

(4)转步骤(2)，重新计算每个(自变化)簇的平均值。这个过程不断重复，直到某个准则函数不再明显变化或者聚类的对象不再变化为止。

去重后的数据还有 4225 条，其中很大一部分数据都是反映不同的问题，K-Means 聚类后输出的结果中，并不是每个簇的问题都是相似度极高的，而我们只需要反映相同问题的簇，于是定义平均相似度（Mean-Similarity）来衡量每个簇的优劣，平均相似度定义如下：

$$MeanSimilarity = \frac{\sum_{i=2}^{n} similari[1][i]}{n} \qquad (11)$$

其中 similarity 为之前计算的相似矩阵,n 为簇内数据的长度,similarity[1][i]表示簇内第一条数据与第 i 条数据的相似度。

经过计算如果一个簇 MeanSimilarity 的值大于 0.25，则可认为该簇问题的相似度极高，并把它保存到 Excel 中，以该簇第一条数据的留言主题命名，如：请书记关注 A 市 A4 区 58 车贷案.xlsx ,存放在 data 目录下。

3.8 热点问题挖掘

3.8.1 热度指数定义

经过上述步骤，已经成功将相似度极高的问题都存放在了 data 文件夹下。

接下来则是计算各类问题的热度指数，由于各类问题的数目相对于所有数据来说太少了，于是给每一类问题赋一初始热度$b=50$，热度指数显然与各类问题的点赞数、反对数和问题数量有关，由于问题的反对数都很少，可忽略不计，于是给出的热度指数（score）定义如下：

$$score=\frac{\frac{length}{length\_total}+\frac{agree}{agree\_total}}{2}+b \qquad (12)$$

其中$length$为 data 文件夹下某类问题的长度，$length_total$l 为 data 文件夹下所有问题的长度，$agree$为某类问题的点赞数，$agree_total$为 data 文件夹下所有问题的点赞数，b为常数 50。

3.8.2 计算各类问题的热度

定义了热度指数（score）之后，就可以通过热度指数计算各类问题的热度，然后将各类问题按热度值的大小降序排列，图 5 为计算后排名前五热点问题的主题，代码存放在热点问题.py 中。

问题主题	hot
A市A5区汇金路五矿万境K9县存在一系列问题.xlsx	81.33
请书记关注A市A4区58车贷案.xlsx	76.16
A市基础设施建设刻不容缓.xls	66.19
A市伊景园滨河院捆绑销售车位.xls	64.43
A2区丽发新城小区噪音扰民.xls	63.98

图 5：排名前五的热点问题

3.8.3 制作热点问题明细表

热点问题明细表相比热点问题还多了问题 ID 列，在各类问题的 Excel 表中新增问题 ID 列，并将其复制，比如“A 市 A5 区汇金路五矿万境 K9 县存在一系列问题.”的热度值排名第一，则将该类问题的问题 ID 值全部复制为 1，“请书记关注 A 市 A4 区 58 车贷案”的热度值排名第二，则将该类问题的问题 ID 值全部复制为 2，依次类推，图 6 为制作完成后的部分热点问题明细表，代码存放在热点问题明细表.py 中，完整的热点问题明细表存放在热点问题明细表.xlsx 中。

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	208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019/8/19 11:34:04	我是A市A5	0	2097
1	234086	A00099869	A市五矿万境K9县房子的墙壁又开裂了	2019/6/20 9:30:44	五矿万境K	0	6
1	198961	A000103957	反映A5区圭塘路五矿万境水岸路段拥堵及执法不为问题	2019/11/11 16:30:39	尊敬的领导	0	3
1	208069	A00094436	A5区五矿万境K9县的开发商与施工方建房存在质量问题	2019/5/5 13:52	本人是A5区	0	2
1	215507	A000103230	A市五矿万境K9县存在严重的消防安全隐患	2019/9/12 14:48	预交房23栋	0	1
1	252650	A00010531	A市五矿万境K9县交房后仍存在诸多问题	2019/9/11 15:16:02	尊敬的相关	0	0
1	262599	A000100428	A市五矿万境K9县房屋出现质量问题	2019/9/19 17:14:49	我是西地省	0	0
1	275491	A00061339	A市五矿万境K9县负一楼面积缩水	2019/9/10 9:10:22	关于五矿万	0	0
1	283732	A00021495	A市五矿万境水岸三期违规建设使用垃圾站	2019/1/15 10:29	我们是A市	0	0
2	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	尊敬的胡书	0	821
2	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	胡书记：您	0	733
2	226265	A000106448	恳请A市经侦公正办理58车贷案件，还我们受害人一个公道	2019/5/28 15:08:51	唐局长，您	0	3
2	254532	A000106062	A市58车贷恶性退出立案近半年没有发过一次案情通报	2019/1/14 22:08:20	背景：58车	0	3
2	214238	A00061787	请问A4区公安派出所对58车贷一案办案的进度如何了	2019/1/20 22:28:40	标题：恳请	1	2
2	272413	A000106062	西地省A市58车贷恶性退出，A4区立案已近半年毫无进展	2019/1/14 20:23:57	西地省58车	0	2
2	198854	A000106735	A2区余易贷平台涉嫌诈骗，群众合法维权被强行扣押	2019/8/13 16:32:05	书记您好：	0	1
2	218132	A000106090	再次请求过问A市58车贷案件进展情况	2019/1/29 19:15:49	尊敬的胡书	0	0
2	234320	A000106592	不要让A市因为58车贷案件而臭名远扬	2019/7/8 17:16:57	胡书记：您	0	0
2	272858	A00061787	A市58车贷恶性退出案件为什么不发布案情进展通报？	2019/1/16 23:21:21	唐局长，您	0	0

图 6：制作完成后的部分热点问题

4. 问题 3 分析方法与过程

根据附件 4，答复意见的评分受到三个因素影响：相关性[3]、情感分析[4]、是否及时回复。综合以上三个因素，得出答复意见的综合得分（总分 100 分）。

4.1 相关性

相关性得分是量化答复意见与留言详情、留言主题之间关联性的指标。将留言主题、留言详情进行分词、去除停用词、计算词频、通过语料库建立词典；将答复意见进行 jieba 分词、去除停用词、通过 doc2bow 转化为稀疏向量从而得到新语料库；将新语料库通过 tfidfmodel 进行处理，得到 tfidf；通过 token2id 得到特征数、稀疏矩阵相似度，从而建立索引，得到最终相似度结果（附件：第三题答复意见总分.xlsx）。

相似度在 0 和 1 之间，相似度越大说明答复意见和留言详情、留言主题与越相似。

4.2 情感分析

情感分析是用来分析答复意见是否规范得体的方法。通过定制一系列的格式字典和规则，对文本进行段落拆解、句法分析，计算感情值，最后通过感情值来作为文本的情感倾向依据。根据已经构建的字典，对答复意见进行分词、去除停用词，抽取字典对应的词语，计算出答复意见的情感得分。字典包括程度词(附件 degree.csv)、否定词(附件 not.csv)、格式词(附件 format.xlsx)。

例如：“网友 xxxxx，您好！非常感谢你为我们提出的宝贵意见。”

得分：（8） + （10）+ 2×（10） + （8）=46

如以上例子，一句话会根据字典进行打分，若词语前出现程度词则乘以程度词相关权重，若出现否定词，则乘以-1，最后累加求和，得到这句话的情感值。同时将情感得分上限设置为 150，情感值越高，说明答复意见越有礼貌、格式越标准。最后将情感得分的范围设置在 0 和 1 之间（附件：第三题答复意见总

分.xlsx）。

4.3 是否及时回复

一条留言的答复时间与留言时间之差大于 30 天，认为未在规定时间内答复，则扣除总分的 5%；一条留言的答复时间与留言时间之差小于等于 30 天，则不扣分。

4.4 计算答复意见总分

按以下公式计算出答复意见总分（附件第三题答复意见总分.xlsx），得分范围 0 到 100 之间，得分越高说明答复意见越好。

$$答复意见总分=\begin{cases}\dfrac{情感得分+相关性得分}{2}\times 100 & 30天内答复\\ \dfrac{情感得分+相关性得分}{2}\times 100\times 95\% & 30天后答复\end{cases} \quad (13)$$

	A	B	C	D	E	F	G	H	I	J	K
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	情感得分	相关性得分	是否及时回复	答复意见总分
2	2549	A00045581	区景蓉华苑物业管理有[	2019/4/25 9:32:09	司却以交20万保证理费，在业主大会与		2019/5/10 14:56:5	1	0.537171841	是	76.86
3	2554	A00023583	楚南路洋湖段怎么还没	2019/4/24 16:03:40	生意带来很大影响，且换墙后还有三		2019/5/9 9:49:10	0.483333333	0.833333373	是	65.83
4	2555	A00031618	提高A市民营幼儿园老师	2019/4/24 15:40:04	更是加大了教师的任教职工要依法签订		2019/5/9 9:49:14	0.685333333	0.383597612	是	53.45
5	2557	A000110735	富能享受人才新政购房	2019/4/24 15:07:30	市，想买套公寓,以下（含），首次购		2019/5/9 9:49:42	0.743333333	0.475651473	是	60.95
6	2574	A0009233	市公交站点名称变更的	2019/4/23 17:03:19	岭小学”，原“岑”的问题。公交站		2019/5/9 9:51:30	0.596666667	0.951189756	是	77.39
7	2759	A00077538	3区含浦镇马路卫生很	2019/4/8 8:37	巴冲到右边，越是有说明卫生较差的是		2019/5/9 10:02:08	0.608666667	0.746003747	否	64.35
8	2849	A000100804	坡师村小区盼望早日安	2019/3/29 11:53:23	社区惠民装电梯的民政府办公室下发了		2019/5/9 10:18:58	0.659333333	0.377964467	否	49.27
9	3681	UU00812	东澜湾社区居民的集体	2018/12/31 22:21:59	，天寒地冻的跑好及设施设备采购等工		2019/1/29 10:53:0	0.86	0.382682979	是	62.13
10	3683	UU008792	阳光住宅楼无故停工以	2018/12/31 9:55:00	到相关准确开工作检查后，西地省基		2019/1/16 15:29:4	0.713333333	0.326637298	是	52.00
11	3684	UU008687	顺路洋湖壹号小区路段	2018/12/31 9:45:59	等地方做立体绿化要求完成了建设，		2019/1/16 15:31:0	0.593333333	0.53086549	是	56.21
12	3685	UU0082204	区大托街道大托新村违	2018/12/30 22:30:30	局审批通过《温室耕地征收补偿款给原		2019/3/11 16:06:3	0.656666667	0.450815856	否	52.61
13	3692	UU008829	阳村D区安置房人防工程	2018/12/29 23:27:51	置房地下室近两万防发[2014]7号文件要		2019/1/29 10:52:0	0.593333333	0.474341631	否	50.71
14	3700	UU00877	段请求修建一座人行天	2018/12/29 11:55:34	量从小区开车出主行具体选址，招标		2019/1/14 14:34:5	0.593333333	0.193649173	是	39.35
15	3704	UU0081480	A市芒果金融平台涉嫌	2018/12/28 17:18:45	关政府部门的大猜，已由银盆岭派出		2019/1/3 14:03:07	0.593333333	0.160046101	是	37.67
16	3713	UU0081227	建议增开A市261路公交	2018/12/28 7:53:25	上！天寒地冻，期驶员工作时间长，		2019/1/14 14:33:1	0.713333333	0.233549684	是	47.34
17	3720	UU008444	与披塘路交叉路口通行	2018/12/27 15:18:07	ttps://baidu.c路路口两端各拆除2		2019/3/6 10:26:14	0.863333333	0.530716896	否	66.22
18	3727	UU0081194	桐梓坡路益丰大药房	2018/12/27 1:55:21	各种理由拒绝退货的信息进行投诉信息		2019/1/3 14:02:47	0.656	0.670820415	是	66.34
19	3733	UU008706	在A市梅溪湖开办一个居	2018/12/26 16:51:40	建议在艺术中心,梵湖二期金菊路与雪松		2019/1/14 14:32:4	0.593333333	0.707106769	是	65.02
20	3747	UU008201	3区中海国际社区一期	2018/12/25 19:35:12	就施工，严重影响位由于需要夜间连续		2019/1/8 16:19:16	0.686666667	0.866025448	是	77.63
21	3755	UU0081681	卡、医保卡、居民健康	2018/12/25 16:23:27	可以尽快合一。让均，需三方或三方以		2019/1/4 15:48:23	0.7	0.242535621	是	47.13
22	3756	UU0081681	一卡通尽快支持手机n	2018/12/25 16:19:49	苹果等手机都无间请关注潇洒支付公		2019/1/4 15:49:44	0.785333333	0.27975142	是	53.25
23	3760	UU0081500	泉水村塘下组土地征	2018/12/25 14:40:13	国家行政机关进行了土地补偿协议，并		2019/1/8 16:18:00	0.825333333	0.417126	是	62.12
24	3762	UU0081057	大队纠正电子交通警	2018/12/25 13:56:31	车辆和行人通行，第三十八条第一款第		2019/1/16 15:22:1	0.593333333	0.433526427	是	51.34
25	3777	UU008162	线北段在楚江北路上	2018/12/23 21:47:34	频发。如果8路纠您好，非常感谢您对		2019/1/29 10:50:3	0.833333333	0.214397848	否	49.77
26	3788	UU0081604	商业住房贷款转公积金	2018/12/21 11:01:00	否能在A市办理商贷持非本中心的缴存，		2019/1/3 14:00:47	0.656666667	0.866025388	是	76.13
27	3791	UU008694	劳动东路-机场高票）	2018/12/20 17:28:09	市国际会展中心3完成约800米路基，		2019/1/4 15:47:36	0.593333333	0.226778671	是	41.01
28	3797	UU008765	西湖街道茶场村公路	2018/12/20 11:16:07	A3区山景区西大	资计划调整，该项目	2019/1/3 13:59:33	0.653333333	0.391311884	是	52.23
29	3838	UU0082119	新江洋湖集体资产的	2018/12/15 15:17:53	多亿好远，这笔协的西地省洋兴置业		2019/1/4 15:44:31	0.733333333	0.46358633	是	59.85
30	3848	UU008233	佳兆业云顶小区筹建的	2018/12/14 14:29:25	这样操作的。梅东为A市A3区那么好市		2018/12/29 15:05:1	0.64	0.436435759	是	53.82
31	3871	UU0082278	在A市怡海星城楼盘建么	2018/12/12 8:57:13	路一直没有修好,小学，中学属于青雅		2019/3/15 15:40:0	0.653333333	0.509175062	否	55.22
32	3877	UU00840	路254号汇富中心前的	2018/12/11 15:35:40	，后面才想到在路|与该车实际停车时		2019/1/4 15:45:01	0.643333333	0.483720958	是	56.35
33	3878	UU008355	有限公司收取服务费、	2018/12/11 15:23:04	，富吉又说在与|“3•18”“5•20”“		2019/3/15 15:39:4	0.782	0.48152864	否	60.02
34	3906	UU0081202	举报有公司骗取加盟费	2018/12/8 12:16:24	说实话。如果县|详细信息及证据，请		2018/12/27 9:23:0	0.781333333	0.113402307	是	44.74
35	3910	UU0081955	市岸海保利西物业的违	2018/12/7 20:56:34	是与开发商签，|，而并非您所称的3		2018/12/29 15:02:1	0.593333333	0.50514102	是	54.92
36	3913	UU0081274	A区珠江俪城公园一号	2018/12/7 14:20:19	持续骚扰楼上业|板明确表示会充分考		2018/12/27 9:21:3	0.66	0.32844311	是	49.42
37	3944	UU0081707	市交警部门在太平路安	2018/12/4 16:11:03	转弯的车辆的视线移交，交警部门无		2018/12/27 9:17:1	0.593333333	0.077096626	是	33.52

图 7：答复意见总分

答复意见评价体系综合考虑相关性、情感分析、是否及时回复三因素，能够综合性的对答复意见进行打分，对无实际意义的答复意见有很好的效果（图 8）。

	A	B	C	D	E	F	G	H	I	J	K
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	情感得分	相关性得分	是否及时回复	答复意见总分
2	25918	UU0081182	关云段A1区南路路灯及	2014/4/21 1:40:59	条A1区南路A市段	“UU0081182”	14/5/28 15:11:	0	0	否	0.00
3	37459	A00039732	市带小孩打疫苗要带什	2019/1/13 1:56:01	带什么证件呢？是	2019年1月14日	19/1/14 16:06:	0	0	是	0.00
4	37482	A00062705	89市规划一个校车接送	2018/12/7 18:48:01	是多的交通事故！	2018年12月12日	18/12/13 18:53:	0	0	是	0.00
5	41610	UU0081530	盛名城小区旁的一公司	2019/9/29 12:11:52	司北边就是正在建	你好！2019年10月10日	19/10/17 12:35:	0	0	是	0.00
6	114346	UU0081119	城锦豪雍景园小区商品	2013/6/3 13:05:49	。但购房补充合同	已收悉	013/7/5 16:47:4	0.056666667	0	否	2.69
7	6556	UU0081320	打狂犬疫苗报销比例是	2018/3/20 15:19:47	，打狂犬疫苗报销	已收悉	18/3/28 16:05:	0.056666667	0	是	2.83
8	30019	UU0081515	准备全款购买二手房事	2016/11/3 10:00:17	房，房产局资金监	已收悉	16/11/22 12:25:	0.056666667	0	是	2.83
9	11927	UU0081626	工矿棚户区改造项目申	2014/10/2 19:53:03	市轴承厂工矿棚户	网友：您好！留言已收悉	14/11/5 16:44:	0.11	0	否	5.23
10	11974	UU0083334	汽车西站至金洲大道公	2014/9/16 8:41:49	下班都搞得心惊胆	网友：您好！留言已收悉	14/10/20 9:45:	0.11	0	否	5.23
11	12142	UU0082324	古曲南路交叉路口建设	2014/7/14 10:09:40	段是一个下坡路，	网友：您好！留言已收悉	14/8/27 16:11:	0.11	0	否	5.23
12	12152	UU0081903	路运输驾驶员接受远程	2014/7/11 19:15:26	老乡都可以通过网	网友：您好！留言已收悉	14/8/13 17:50:	0.11	0	否	5.23
13	12163	UU0082207	或915路公交调整经过	2014/7/9 13:43:44	A市大学一来可以	网友：您好！留言已收悉	14/8/13 17:43:	0.11	0	否	5.23
14	12165	UU0082271	站成为烂尾楼，请求政	2014/7/9 10:02:23	直无法收房。	网友：您好！留言已收悉	14/8/13 17:39:	0.11	0	否	5.23
15	12289	UU0081164	泉塘社区的“当代和城	2014/5/17 12:41:32	位组长为他当选，	网友：您好！留言已收悉	14/6/16 16:59:	0.11	0	否	5.23
16	12383	UU0087957	建议取消绕城高速收费	2014/4/13 17:26:33	生，每天来往均要	网友：您好！留言已收悉	14/5/16 15:57:	0.11	0	否	5.23
17	12387	UU0081236	金洲新区高新安置房小	2014/4/10 13:38:43	经有三年多了，我	网友：您好！留言已收悉	14/5/16 15:55:	0.11	0	否	5.23
18	12415	UU0081509	区含浦镇白鹤社区的用	2014/3/27 17:16:36	导这些难道都不是	网友：您好！留言已收悉	014/5/9 17:28:0	0.11	0	否	5.23
19	12458	UU00854	镇头镇连山村交通不	2014/3/7 22:39:35	不断的扩宽提质，	网友：您好！留言已收悉	14/4/14 12:13:	0.11	0	否	5.23
20	11985	UU0082294	上十万人出行非常不方	2014/9/10 14:29:51	盲得小区居民没办	网友：您好！留言已收悉	014/9/26 9:38:3	0.11	0	是	5.50
21	12031	UU0082194	交2路车（A市晚报-黄	2014/8/23 18:52:08	去黄花镇，城乡之	网友：您好！留言已收悉	014/9/4 16:30:2	0.11	0	是	5.50
22	12189	UU0089485	制黑路综合管理农村道	2014/7/2 9:02:52	讲价钱的机没好	网友：您好！留言已收悉	14/7/30 17:25:	0.11	0	是	5.50

图 8：无实际意义的答复意见

5. 参考文献

[1] 孙海锋，郑中枢，杨武岳. 网络招聘信息的数据挖掘与综合分析[EB/OL]. (2017-4-17)[2020-05-07]. https://www.doc88.com/p-0744910690230.html.

[2] 李春林, 冯志骥. 基于文本挖掘的新能源汽车用户评论研究[J]. 经济与管理科学, 2020, (4): 148-151.

[3] 章志华, 陆海良, 郁钢. 基于 TFIDF 算法的关键词提取方法[J]. 信息技术与信息化, 2015(08): 158-160.

[4] 侯艳钗. 基于词语权重的中文文本分类算法的研究[D]. 河北工业大学, 2011.

[5] 杨涛. 面向海量文本的分类算法研究[D]. 齐鲁工业大学, 2016.

[6] 齐向明, 孙煦骄. 基于语义簇的中文文本聚类算法[J]. 吉林大学学报(理学版), 2019, 57(05): 1193-1199.

[7] 毛郁欣, 邱智学. 基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J]. 中国信息技术教育, 2020(08): 99-101.

[8] 刘惠, 赵海清. 基于 TF-IDF 和 LDA 主题模型的电影短评文本情感分析——以《少年的你》为例[J]. 现代电影技术, 2020(03): 42-46.

[9] 曾小芹, 余宏. 基于 Python 的商品评论文本情感分析[J]. 电脑知识与技术, 2020, 16(08): 181-183.