“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。

本文借助自然语言处理技术，应用文本挖掘方法，以人工智能从大数据中获取有价值的信息，从而让政府公务人员得以更轻松地迎接所面临的巨大挑战。

对于问题一，将群众留言分类，我们首先对数据进行预处理，接着采用朴素贝叶斯的方法来对每一个数据的内容进行分类。获得分类结果之后，将之与附件中的分类结果用 F-score 分类评价方法来评价分类效果。

关于问题二，对数据中的热点问题的挖掘，我们首先对数据分词和去停用词处理，再用 K-Means 聚类方法进行文本聚类，并根据赞成和反对人数综合制定热度评价指标，最终以此指标输出热度前五的问题。

针对问题三，评价答复意见，我们通过计算答复内容与留言内容的文本相似度来衡量相关性；通过答复意见的文本长度和是否包含某些特定词语来衡量完整性和可解释性。最后对数据分配不同权重计算综合指标。

关键词：文本挖掘；朴素贝叶斯；TF-IDF；Kmeans 算法；

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that used to rely mainly on human to divide messages and sort out hot spots.

This paper uses natural language processing technology and text mining method to obtain valuable information from big data by artificial intelligence, so that government officials can more easily meet the huge challenges.

For problem one, we first preprocess the data and then classify the content of each data by naive Bayes method. After the classification results are obtained, the classification effect is evaluated by F-score classification evaluation method.

On the second problem, to mine the hot issues in the data, we first deal with the data segmentation and de stop words, then use k-means clustering method to cluster the text, and according to the number of pros and cons, we make a comprehensive heat evaluation index, and finally output the top five heat problems with this index.

In response to question 3, we evaluate the response opinions. We measure the relevance by calculating the text similarity between the response content and the message content; we measure the integrity and interpretability by the text length of the response opinions and whether it contains some specific words or not. Finally, the comprehensive index is calculated by different weight of data distribution.

Keywords: text mining; naive Bayes; TF-IDF; K-Means algorithm

目录

一、 挖掘目标

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。

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。

1.2 挖掘目标

请利用自然语言处理和文本挖掘的方法解决下面的问题

第一题：留言内容分类，根据给定的标签对留言进行分类。

第二题:热度问题挖掘,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，并列出热度前五的话题。

第三题：根据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，并尝试实现。

二、 数据处理

数据处理主要是对附件 2 中“留言主题”、附件 3 中“留言详情”、附件 4 中的“留言详情”和“答复意见” 做分词和去停用词处理，对数据的文本表示、特征提取和相似度分析起重要作用。三个问题的具体数据处理过程和结果在其对应的模型构建中会较为详细地介绍。

三、 模型构建及评价

3.1 群众留言分类模型

3.1.1 基本思路

首先对“留言详情”文本数据进行分词和去停用词处理，再进行文本权重计算，特征提取，以训练数据建立模型，并给出模型评价，最后根据模型评价结果对模型进行优化，反复操作，直至达到理想结果

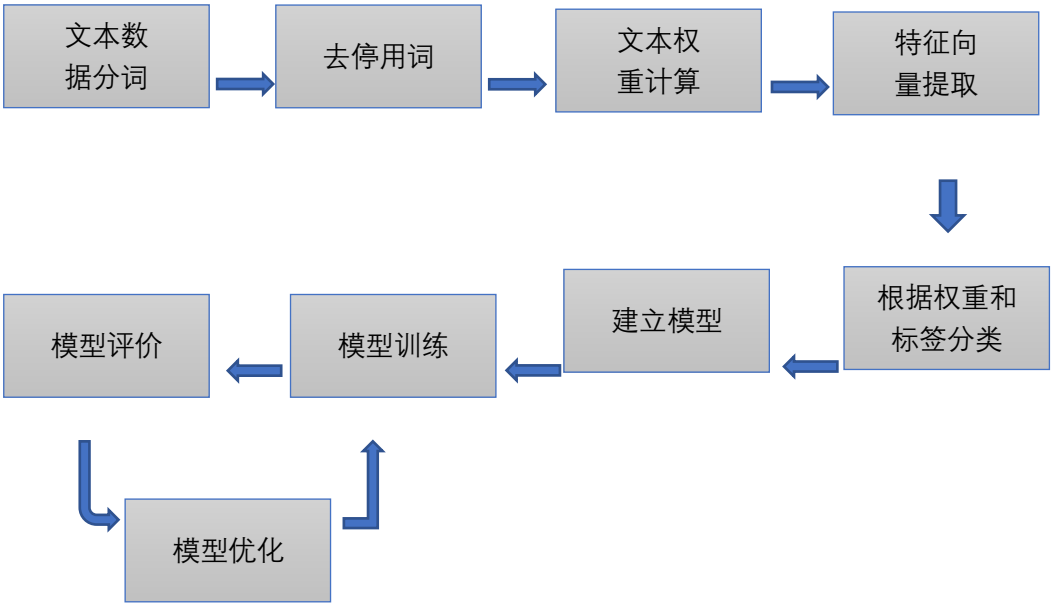


图 3.1.1 模型建立流程图

3.1.2 文本权重计算

(1) 中文分词和去停用词

a) jieba 分词

基于自然语言的 jieba 分词处理库属于概率语言模型分词。概率语言模型分词的任务是：在全切分所得的所有结果中求某个切分方案 S，使得 P(S)最大。

原理：对于给定的文本数据，通过正则获取连续的字符，划分为短语列表，对每个短语使用 DAG(查字典)和动态规划，得到最大概率路径，对 DAG 中那些没有在字典中查到的字，组合成一个新的片段短语，使用 HMM 模型进行分词，也就是作者说的识别未登录词。

jieba 支持三种分词模式。全模式：把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；精确模式：试图将句子最精确地切开，适合文本分析；搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。目标是对“留言详情”内容，根据“一级分类”中的标签进行分类，所以适合使用 jieba 中的精准模式

b) 去停用词

由于一些虚词、副词等的存在，会影响分词以后的字、词频统计，使得真正起到作用的实词占比降低，影响权重分配，所以要建立合适的停用词表，并对数据进行去停用词化。

图 3.1.2-1 是分词和去停用词处理过后的文本数据，可以看出停用词被去除，文本被分为词语，对后续的分类起到铺垫作用。

	分类	留言详情	cat_id	clean_review	cut_review
26	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t A3区阳光丽城附近垃圾场，白天偷偷将垃圾...	0	A3区阳光丽城附近垃圾场白天偷偷将垃圾运过来晚上趁夜色进行焚烧特别是10点以后外面烟雾弥漫味...	A3 区 阳光 丽城 附近 垃圾场 白天 偷偷 垃圾 运过来 晚上 夜色 进行 焚烧 特别...
10	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 易市长您好，感谢您的阅读。十二五期间，非...	0	易市长您好感谢您的阅读十二五期间非省会地级市的轨道交通规划与建设已经席卷而来截止到2016年...	易 市长 您好 感谢您 阅读 十二五 期间 非 省会 地级市 轨道交通 规划 建设 已经 席...
16	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t胡书记，这冬天又到了，A市这湿冷的冬天真是受...	0	胡书记这冬天又到了A市这湿冷的冬天真是受不了该太冷了被子感觉都潮潮的洗的衣服都难干每天早上起...	胡书记 冬天 A 市 湿冷 冬天 真是 受不了 该 太冷 被子 感觉 都 潮潮 洗 衣服 都...
17	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的市委市政府：A市是一座历史名城，是一座...	0	尊敬的市委市政府A市是一座历史名城是一座具有幸福感的城市幸福感的体现就在于市委市政府想民之所...	尊敬 市委 市政府 A 市是 一座 历史 名城 一座 具有 幸福感 城市 幸福感 体现 在于...
14	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA市A5区朝晖路锦楚国际新城三区从6月份开始...	0	A市A5区朝晖路锦楚国际新城三区从6月份开始到现在一共停电不少于9次每次都不说什么原因停电而...	A 市 A5 区 朝晖路 锦楚 国际 新城 三区 6 月份 现在 一共 停电 不少 9 次 ...
21	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t你们好，上周提交了请求迎丰公园从人性关怀的角...	0	你们好上周提交了请求迎丰公园从人性关怀的角度考虑延后清晨的路灯熄灯时间半小时貌似没有得到任何...	你们好 上周 提交 请求 迎丰 公园 人性 关怀 角度 考虑 延后 清晨 路灯 熄灯 时间 ...
30	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...	0	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 A 市程明 物业管理 有...
9	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 李书记您好，感谢您的阅读。十二五期间，非...	0	李书记您好感谢您的阅读十二五期间非省会地级市的轨道交通规划与建设已经席卷而来截止到2016年...	李 书记 您好 感谢您 阅读 十二五 期间 非 省会 地级市 轨道交通 规划 建设 已经 席...
4	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t K8县丁字街的商户乱摆摊，前段时间丁字街...	0	K8县丁字街的商户乱摆摊前段时间丁字街的交通好了几天最近那些在丁字街做生意的商户又开始把商品...	K8 县 丁字街 商户 乱 摆摊 前段时间 丁字街 交通 好 几天 最近 丁字街 做生意 商...
18	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tK9县城更新了公交线路，新的公交车也在试运行...	0	K9县城更新了公交线路新的公交车也在试运行中市民出行是一项重大民生工程省市城内公交都是1元出...	K9 县城 更新 公交线路 新 公交车 试运行 中 市民 出行 一项 重大 民生 工程 省市 ...

图 3.1.2-1 分词和去停用词处理后的数据

(2) 文本权重计算

文本权重计算使用的是 TF-IDF，是一种加权统计法。它通过统计的方法来计算和表达某个关键词在文本中的重要程度。TFIDF 是由两部分组成，一部分是 TF(Token Frequency)，表示一个词在文档中出现的次数，即词频。另一部分是 IDF(Inverse Document Frequency)，表示某个词出现在多少个文本中(或者解释为有多少个文本包含了这个词)，即逆向文档频率，通常由公式 $IDF_t = \log((1+|D|)/|D_t|)$，其中$|D|$表示文档总数，$|D_t|$表示包含关键词 t 的文档数量。TFIDF 的值就是由这两部分相乘得到的。

最终得到词向量及权重分配结果(部分)如 **图 3.1.2-2**

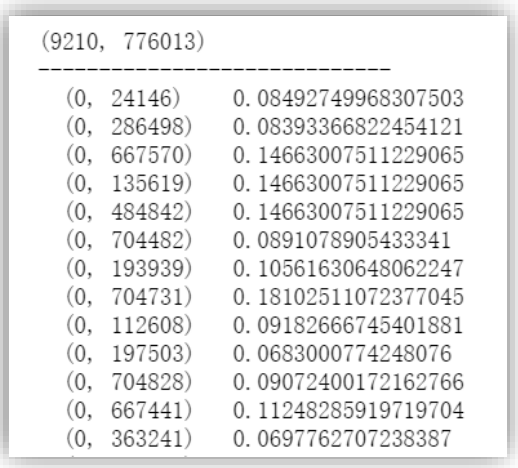


图 3.1.2-2

3.1.3 初级分类

通过对数据中分词的权重及其对应的标签进行统计分类，并列出每个标签对应最大频率的四个词语和四个词组 如 **图 3.1.3**（部分）。由此我们可以看出每个标签对应的词语和词组都与该标签有很大的相关性，所以可以初步判断分类结果较为满意。

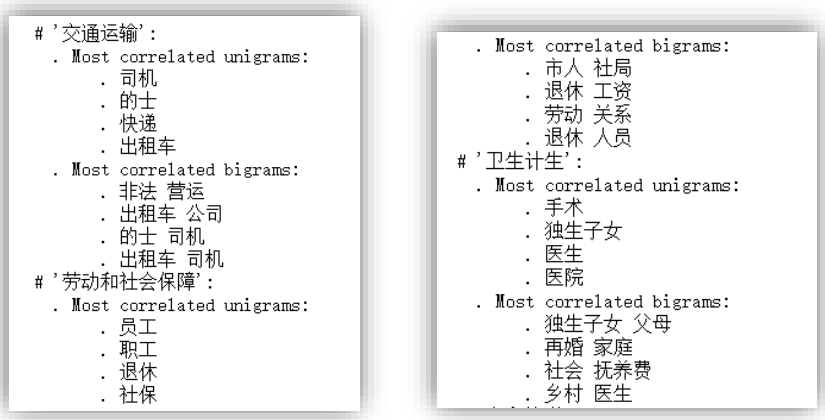


图 3.1.3

3.1.4 朴素贝叶斯分类

1) 原理：根据贝叶斯定理

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

x=(x1,x2,...xD)表示含有 D 维属性的数据对象。训练集 S 含有 K 个类别，表示为 y=(y1,y2,...yK)。

$$P(\mathbf{x}|y_k) = \prod_{d=1}^{D} P(x_d|y_k) = P(x_1|y_k)P(x_2|y_k)...P(x_D|y_k)$$

2) 实现过程：其中令 x=分类标签，y_k 为每个留言中词语 tfidf 权重则，以此构建分类器，可以得出每个留言能被分类在标签 x_i 的概率，取最高的 P 值对应的 x_i 即分类标签。并将分类结果存入

3.1.5 分类效果检验

利用 F-Score 对分类方法进行评价

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

最终结果如**图 3.1.5**:

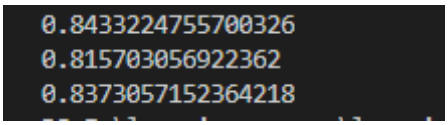


图 3.1.5

对所有数据进行利用分类器进行分类，得到宏 F1=0.8433，微 F1=0.8157，加权 F1=0.8373

3.2 热点问题挖掘模型

3.2.1 基本思路：流程图如**图 3.2.1**

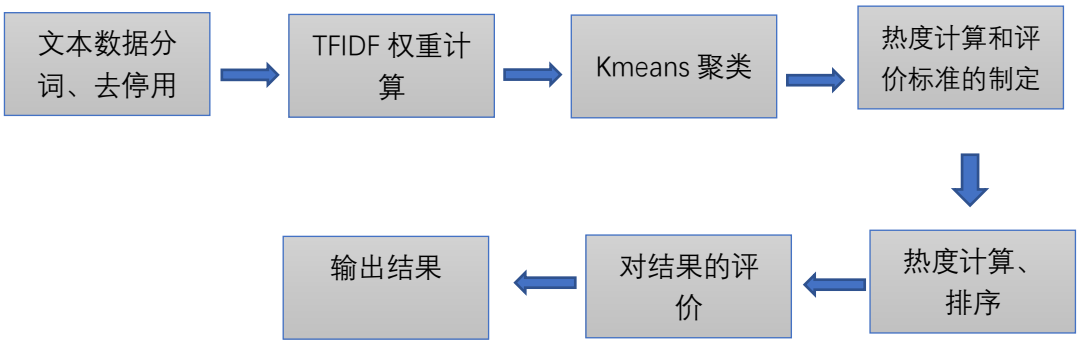


图 3.2.1

3.2.2 文本处理

1.对留言主题进行分词和去停用词处理，方法同 3.1.2。
2.文本 TFIDF 权重计算，方法同 3.1.3。

3.2.3 K-Means 聚类算法和 K-Means++

<1>定义:

聚类是一个将数据集中在某些方面相似的数据成员进行分类组织的过程，聚类就是一种发现这种内在结构的技术，聚类技术经常被称为无监督学习。

k 均值聚类是最著名的划分聚类算法，由于简洁和效率使得他成为所有聚类算法中最广泛使用的。给定一个数据点集合和需要的聚类数目 k，k 由用户指定，k 均值算法根据某个距离函数反复把数据分入 k 个聚类中。

<2>传统 K-Means 算法原理:

输入是样本集 D={x1,x2,...xm}D={x1,x2,...xm},聚类的簇树 k,最大迭代次数 N

输出是簇划分 C={C1,C2,...Ck}C={C1,C2,...Ck}

1) 从数据集 D 中随机选择 k 个样本作为初始的 k 个质心向量： {μ1,μ2,...,μk}{μ1,μ2,...,μk}

2）对于 n=1,2,...,N

a) 将簇划分 C 初始化为 Ct=∅t=1,2...kCt=∅t=1,2...k

b) 对于 i=1,2...m,计算样本 xixi 和各个质心向量 μj(j=1,2,...k)μj(j=1,2,...k)的距离：

dij=||xi−μj||22dij=||xi−μj||22，将 xixi 标记最小的为 dijdij 所对应的类别 λiλi。此时更新 Cλi=Cλi∪{xi}Cλi=Cλi∪{xi}

c) 对于 j=1,2,...,k,对 CjCj 中所有的样本点重新计算 μj=1|Cj|∑x∈Cjxμj=1|Cj|∑x∈Cjx

d) 如果所有的 k 个质心向量都没有发生变化，则转到步骤 3）

3） 输出簇划分 C={C1,C2,...Ck}

<3>初始化优化的 K-Means++算法原理：

K-Means++的对于初始化质心的优化策略也很简单，如下：

a) 从输入的数据点集合中随机选择一个点作为第一个聚类中心 μ1μ1

b) 对于数据集中的每一个点 xixi，计算它与已选择的聚类中心中最近聚类中心的距离

D(xi)=argmin||xi−μr||22r=1,2,...kselectedD(xi)=argmin||xi−μr||22r=1,2,...kselected

c) 选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)D(x)较大的点，被选取作为聚类中心的概率较大

d) 重复 b 和 c 直到选择出 k 个聚类质心

e) 利用这 k 个质心来作为初始化质心去运行标准的 K-Means 算法

k 个初始化的质心的位置选择对最后的聚类结果和运行时间都有很大的影响，因此需要选择合适的 k 个质心。如果仅仅是完全随机的选择，有可能导致算法收敛很慢。K-Means++算法就是对 K-Means 随机初始化质心的方法的优化。

3.2.4 基于 K-Means++的文本聚类

实现过程：

1) 文本处理，即分词、计算 TF-IDF 权重等
2) 确定聚类 K 值
3) 输出聚类标签

3.2.5 定义热度指标

热度指标用于衡量群众集中反映问题的热度情况。我们从热度二字着手，深入分析附件 3 所给数据，提取出能反映热度的几大特征：同一问题的留言条数、点赞数、反对数的多寡等。

通过对这些数值的权重分配得出该问题的热度指标，在这里我们简单的将留言数加上点赞数与反对数的算术平均，即

$$\boldsymbol{hot}_{\boldsymbol{热度}} = \boldsymbol{C}_{\boldsymbol{留言}} + \frac{\boldsymbol{D}_{\boldsymbol{点赞}} + \boldsymbol{D}_{\boldsymbol{反对}}}{\boldsymbol{2}}$$

3.2.6 热度表以及明细表

1） 热度表

将附件 3 中的数据通过 Python 中 Pandas 的 DataFrame 操作，将聚类标签与每一条数据相挂钩，然后筛选出每一类问题的其中一个作为热度表中的代表问题。接着将热度用列表排序并返回索引，再用这个索引来排序其他数据，则将热度表按照热度排名给罗列出来。

1	1	1063.5	2019-01-15	至	2019-11-11	反映A5区圭塘路五矿万境水岸路段拥堵及执法不为问题
2	2	906	2019-01-10	至	2019-12-31	A7县橄榄城小区的孩子到泉塘小学上学不方便
3	3	784.5	2019-01-20	至	2019-08-08	承办A市58车贷案警官应跟进关注留言
4	4	425.5	2019-01-11	至	2019-07-08	严惩A市58车贷特大集资诈骗案保护伞
5	5	354.5	2019-01-30	至	2019-09-06	A4区绿地海外滩小区距渝长厦高铁太近了
6	6	224	2019-01-07	至	2020-01-08	咨询A6区道路命名规划初步成果公示和城乡门牌问题
7	7	138	2019-06-19	至	2020-01-09	A市万家丽南路丽发新城居民区附近搅拌站扰民
8	8	113	2019-01-03	至	2019-05-22	关于A市金星北片110kv及以上高压线的现状和规划的几个问题
9	9	106.5	2019-01-22	至	2020-01-07	咨询A市公交409线路问题
10	10	96	2019-01-02	至	2019-11-08	问问A市经开区东六线以西泉塘昌和商业中心以南的有关规划

完整数据详见附件中的“热点问题表.xlsx”。

2） 明细表

明细表类似热度表，但要求给出每一个热点问题的具体留言情况。前面我们是每一类问题取出一条来作为代表，每条便对应一个热度指标，已然排序。在此基础上，我们便可轻松地通过几个基本语句将明细表给出。

	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	198961	A00010395	反映A5区圭塘路五矿万境水岸路段拥堵及执法	2019-11-11 16:30:39	尊敬的领导：你好！现反映：	0	3
1	208069	A0009443(	A5区五矿万境K9县的开发商与施工方建房存在	2019-05-05 13:52:50	本人是A5区涧井街道汇金路五	0	2
1	208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019-08-19 11:34:04	我是A市A5区汇金路五矿万境	0	2097
1	215507	A00010323	A市五矿万境K9县存在严重的消防安全隐患	2019-09-12 14:48:07	预交房23栋没有通往负一楼的	0	1
1	234086	A0009986(	A市五矿万境K9县房子的墙壁又开裂了	2019-06-20 09:30:44	五矿万境K9县的房子又出问题	0	6
1	252650	A00010531	A市五矿万境K9县交房后仍存在诸多问题	2019-09-11 15:16:02	尊敬的相关部门，本人家庭于	0	0
1	262599	A00010042	A市五矿万境K9县房屋出现质量问题	2019-09-19 17:14:49	我是西地省A市五矿万境K9县	0	0
1	275491	A0006133(	A市五矿万境K9县负一楼面积缩水	2019-09-10 09:10:22	关于五矿万境·K9县负一楼面	0	0
2	283732	A0002149(	A市五矿万境水岸三期违规建设使用垃圾站	2019-01-15 10:29:32	我们是A市五矿万境水岸三期	0	0
2	188799	A00010734	A7县橄榄城小区的孩子到泉塘小学上学不方便	2019-05-22 12:24:16	我是橄榄城的一名业主，小孩	0	1
2	208650	A0007478(	咨询A市金科山水洲房产证问题	2019-09-23 20:39:05	胡书记，您好！我于2013年按	0	0
2	210292	A0003770(	指出A3区潇湾镇通程商业广场处红绿灯问题	2019-03-26 09:10:27	潇湾镇通往一桥中间(通程商	0	0
2	216587	A0008411(	A市山水湾小区经常有粪池车停在里面	2019-07-23 12:42:54	A市万家丽北路山水湾小区成	0	0

完整结果详见“热点问题留言明细表.xlsx”

3.3 答复意见的评价模型

3.3.1 基本思路

通过计算答复内容与留言内容的文本相似度来衡量相关性；通过答复意见的文本长度和是否包含某些特定词语来衡量完整性和可解释性。最后对数据分配不同权重计算综合指标。

3.3.2 相关性检验

首先对数据分词和去停用词处理，方法与第一问相同；再使用自然语言 Python 中的 Difflib 模块进行自动差异性和相似度分析，记为**相关性（R）**，并添加到 **Data.xlsx**。得到结果如 **图 3.3.2**（部分）

留言详情	答复意见	答复时间	相关性（S）
	收取停车管理费，在业主大会结束后业委	2019/5/10 14:56:53	0.562827225
面的生意带来很大影响，里	需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	0.33490566
同时更是加大了教师的工作压	办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	0.41074856
	年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	0.247126437
“马坡岭小学”，原“马坡岭	留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	0.415300546
再把泥巴冲到右边，越是上下	您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	0.30625
为老社区惠民装电梯的规范	市A3区人民政府办公室下发了《关于A市A	2019/5/9 10:18:58	0.46350365
好远，天寒地冻的跑好远，	修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	0.318644068
也没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	0.404371585
交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	0.462608696
规划局审批通过《温室养殖	司支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	0.326530612
区安置房地下室近两万平方	米，按长人防发[2014]7号文件要求，鄱阳	2019/1/29 10:52:01	0.174796748
峰，大量从小区开车出去的业	分局配合进行具体选址，招标（邀标）进	2019/1/14 14:34:58	0.290748899
省相关政府部门的大力支持	决的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	0.138747885
小时以上！天寒地冻，其他公	正常。由于驾驶员工作时间长，劳动强度大	2019/1/14 14:33:17	0.380645161
址：https://baidu.com/。	出的“披塘路路口两端各拆除20米中间花坛	2019/3/6 10:26:14	0.531400966

图 3.3.2

3.3.3 完整性和可解释性检验模型

1） 完整性和可解释性的一个重要指标是每条“答复意见”的文本长度，如果文本长度不够，答复内容就少，完整性和可解释性相应地也就越低；相反地，如果文本长度较长，答复涉及内容越多，文本很大概率较为完整，对问题的解释也较为深入详细。考虑到高效和便捷性，采用 Excel 中文本长度计算工具，得到文本长度数据，记为**文本长度（L）**，同样添加到附件 4.xlsx。

2）

A) “答复意见”文本中不乏有不完整但是文本长度较大的，对此，可以建立模板来检验完整性。我们的方法是建立有关留言开头、中间、结束语中高频词语的**词表 A**。

例如：“因此”,“感谢”,“支持”,“监督”,“疑问”,“欢迎”,“致电”,“明确”……

通过检测文本中是否包含这些词，对文本进行打分，每包含**词表 A** 中的一个词语**得分**（记为 **S**）增加 1。

B) 对于可解释性，建立有关专业用词的**词表 B**,

例如：“根据”,“规定”,“经查”,“编制”,“调查”,“程序”,“据查”,“通知”,“明确”,“依法”……

这些词的存在使得答复内容有理有据，是可解释性高的表现。所以每包**词表 B** 中的一个词语**得分**（**S**）增加 2。

C) 对于没有参考意义的答复采用惩罚机制，建立有词语的**词表 C。**

例如：“转”,“转交”,“移至”,“转至”,“等候”,“等待”,“咨询”,“后续”,“交”,“已交”,“已转”

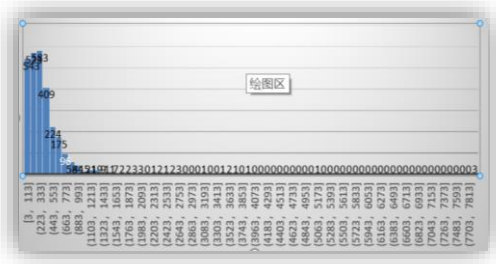
答复内容每包含一个**词表 C** 中的词，**得分**（**S**）减少 4。

3） 综合上述**得分**（S），并将其保存到 **Data.xlsx**

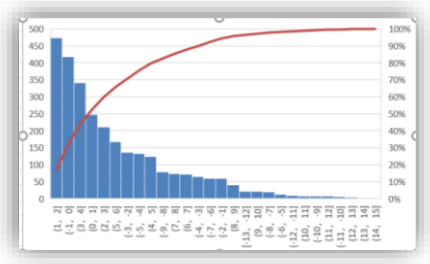
3.3.4 建立综合评价模型

1）数值分析

对**文本长度**（**L**）和**得分**（**S**）做图表如下



文本长度（**L**）分布直方图



得分（**S**）分布直方图

数据初步处理通过对数据的观察，“答复意见”**文本长度**（**L**）范围大，为（3,7883），因此对其进行对数处理；完整性和可解释性**得分**（S）跨度较小，分布较为均匀，为保数值足够小，取得分与极差的比值，即 $\frac{Si}{Max(S)-Min(S)}$ ，(i=1,2,⋯,2816)

2） 模型建立

考虑到对于答复内容的评判参考价值不同，所以使用**相似性**（**R**）、**文本长度**（**L**）和完整性与可解释性**得分**（S）三个变量建立模型时需要分配不同的权重。权重先后依次为 **R，L，S** 对“答复意见”最终的**综合评价**用 **CA** 表示，建立如下模型:

$$\mathbf{CA_i = R_i + 0.4 \times (\lg L_i - 2) + \frac{0.2 \times S_i}{Max(S) - Min(S)}}\ ,\quad (\ i=1,2,\cdots,2816\ )$$

（CA 值越大表明答复内容相关性、完整性和可解释性越高。）

3.3.5 模型效果展示

利用上面建立的模型，计算所有答复意见的综合评价并保存至 **Data.xlsx**。

H	I	J	K	L	M
相关性（R）	答复内容文本长度（L）	得分（S）	lg（L）-2	S/（Max（S）-Min（S））	综合评价（CA）
0.562827225	454	9	0.657055853	0.321428571	0.889935281
0.33490566	305	4	0.484299839	0.142857143	0.557197025
0.41074856	357	5	0.552668216	0.178571429	0.667530133
0.247126437	310	4	0.491361694	0.142857143	0.472242543
0.415300546	161	3	0.206825876	0.107142857	0.519459468
0.30625	232	3	0.365487985	0.107142857	0.473873765
0.46350365	245	5	0.389166084	0.178571429	0.654884369
0.318644068	624	6	0.79518459	0.214285714	0.679575047
0.404371585	505	4	0.703291378	0.142857143	0.714259565
0.462608696	224	2	0.350248018	0.071428571	0.616993617
0.326530612	489	4	0.689308859	0.142857143	0.630825584
0.174796748	427	4	0.630427875	0.142857143	0.455539327

做出**综合评价（CA）**的频率分布直方图得

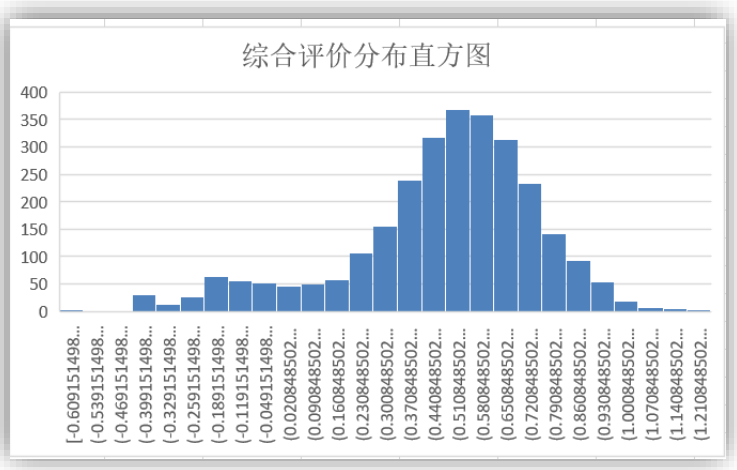


图 3.3.5

通过直方图 **3.3.5** 可以观察到综合评价大体上呈正态分布，符合一般的统计规律，因此可以作为“答复意见”评价的参考。

四、总结

本文主要运用数学模型和数据挖掘技术，使用自然语言 Python 以及算法模块等对留言进行多分类，文本聚类和文本相似度等处理和计算。首先，文本数据的清理对文本分类、多分类模型的建立和使用起重要作用，结巴分词和去停用词大提上已经满足基本需求，但对于语义和词语在不同语句中的含义识别方面有很大不足，这对聚类也有一定的影响。其次，问题二中使用 K-Means 是聚类最常用的算法，聚类效果明显，但是会出现陷入局部最优的情况，我们使用 K-Means++在一定程度上减少了这种影响，但也出现新的问题。最后在对文本的是否人性化分析中使用三个变量，我们做的仅仅据此建立一般的非线性模型，如果在此基础上进行文本情感分析等更深层次的处理，得到的结果将会进一步精确。希望以后可以进一步改进 Kmeans 算法和评价模型。

参考文献

[1] 刘慧.基于 KNN 的中文文本分类算法研究.西南交通大学，2010

[2] 弭晓月 毕冠群 黄成梓基于双重注意力机制与 Bi-LSTM 的智能阅读系统.山东大学，2018

[3] 浅析文本挖掘（jieba 模块的应用）

[EB/OL].https://www.cnblogs.com/wj-1314/p/8034023.html

[4] 二分类模型评估指标的计算方法与代码实现

[EB/OL].https://blog.csdn.net/GeForce_GTX1080Ti/article/details/78877318?depth_1-utm_source=distribute.pc_relevant.none-task&utm_source=distribute.pc_relevant.none-task

[5] 基于 python 中 jieba 包的中文分词中详细使用

[EB/OL].https://blog.csdn.net/meiqi0538/article/details/80218870?depth_1-

utm_source=distribute.pc_relevant.none-task&utm_source=distribute.pc_relevant.none-task

[6] 文本挖掘深度学习之 word2vec 的 R 语言实现

[EB/OL].https://blog.csdn.net/u011955252/article/details/70495921

[7] 邹 海，李 梅， 一种用于文本聚类的改进二分 K-均值算法，技术与方法，2010，1674-7720(2010)12-0064-04

[8] 鲁明羽 ,姚晓娜，魏善岭，基于模糊聚类的网络论坛热点话题挖掘，大连海事大学学报，2008, 1006-7736(2008)04-0052-03