基于文本挖掘的智慧政务系统构建

摘要：民生问题一直是政府部门高度重视的问题之一，微信、微博、市长信箱、阳光热线等网络平台逐步成为了市民反馈问题、政府了解民意的重要平台。但由于近年来互联网的飞速发展，各类社情民意相关的文本数据量不断增长。因此，建立基于自然语言处理技术的智慧平台、利用云计算、网络文本分析和数据挖掘等技术对群众的留言进行研究具有十分重大的意义。

针对问题 1，利用 pandas 读取文件中的”留言主题”、”留言详情”、”一级分类”，得到主要的留言信息及其对应的分类。利用 jieba 分词和哈工大停用词词库对留言详情进行分词、去停，得到描述留言的关键词，用得到的关键词作为特征构建词袋模型，然后再用基于词袋模型的朴素贝叶斯来进行文本分类。根据朴素贝叶斯分类器，将每个类贴上对应一级分类的类别。将数据按照一定比例分为训练集和测试集，对训练集进行训练，根据实际情况修改分类器，当训练集的 f-score 的值高达一定程度时，对测试集进行测试分类。

针对问题 2，利用问题 1 的方法对文件提取有用信息，利用 tf-idf 算法提取 tf-idf 特征，设置 5 个质心，采用 K-means 算法进行文本聚类，使得相似度较高的文本聚集，通过迭代的方法，得到最好的聚类结果。然后对聚类结果分析，得到留言中的热点问题及其点赞数等。

针对问题 3，仍采用分词、提取关键词、转词向量的方法，与前面不同的是，这里使用 Word2vec 词向量算法，再以短文本相似度作为相关性的度量指标，以回复时间差、反馈信息等作为可解释性、完整性的度量指标，综合权重较大的度量指标，形成一套综合评价系统。

关键词：数据挖掘，文本多分类，文本聚类，TF-IDF

Construction of intelligent government system based on Text Mining

Abstract: The issue of people's livelihood has always been one of the issues highly valued by government departments. Network platforms such as WeChat, Weibo, Mayor's Mailbox, and Sunshine Hotline have gradually become an important platform for citizens' feedback and the government's understanding of public opinion. However, due to the rapid development of the Internet in recent years, the amount of text data related to various social conditions and public opinion continues to increase. Therefore, the establishment of a smart platform based on natural language processing technology, the use of cloud computing, network text analysis and data mining and other technologies to study the message of the masses are of great significance.

For problem 1, use pandas to read the "message subject", "message details", and "level 1 classification" in the file to get the main message information and its corresponding classification. Use jieba word segmentation and Harbin University of Technology stop word lexicon to segment and stop the message details, get keywords describing the message, use the obtained keywords as features to build a bag of words model, and then use the simple Bayesian based on the bag of words model To classify text. According to the naive Bayes classifier, each class is pasted to the class corresponding to the first class classification. The data is divided into a training set and a test set according to a certain ratio, the training set is trained, and the classifier is modified according to the actual situation. When the f-score value of the training set is up to a certain level, the test set is tested and classified.

For problem 2, use the method of problem 1 to extract useful information from the file, use the tf-idf algorithm to extract tf-idf features, set 5 centroids, and use the K-means algorithm for text clustering, so that texts with higher similarity are clustered Through the iterative method, the best clustering results are obtained. Then analyze the clustering results to get the hot issues and the number of likes in the message.

For problem 3, the method of word segmentation, keyword extraction, and word transfer is still used. The difference is that the Word2vec word vector algorithm is used here, and the short text similarity is used as a measure of relevance to respond to time difference and feedback Etc. as a

measure of interpretability and completeness, and a measure with a relatively large weight, forming a comprehensive evaluation system.

Keywords: data mining, text multi classification, text clustering, TF-IDF

目录

1、挖掘目标

1.1 挖掘背景

随着我国的信息技术不断发展，网络平台的信息不断完善，政府了解民意的方式越来越多元化，手机、电脑、网络相辅相成，人们越来越倾向于依托于网络的速度传递信息。尤其是近年来，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，随着“智慧政务”建设步伐的快速推进，如何运用大数据、云计算、人工智能等技术，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中的重要课题之一，对提升政府的管理水平和施政效率具有极大的推动作用。

1.2 挖掘目标

本文将针对网络问政平台的群众留言进行分类，给出对应所属的分类；通过提取关键词、对文本进行分类以及文本聚类对某一时段内集中爆发地反映特定地点或特定人群问题的留言进行分类，分类好的留言将会依照特征关键词发送往与该事情有关的相关部门，有利于政府及时、着重得为百姓解决难题。最后，留言者是否对政府回复感到满意，则需要通过反馈系统，我们通过 CBOW 模型、训练词向量等方法深入研究分析了回复的相关性、可解释性、完整性，从而建立了反馈系统，更进一步的提高了政府的服务水平与办事效率。

1.3 研究现状

近年来，随着互联网的飞速发展和随着研究工作的深入，自然语言处理也随着快速发展，研究者们开始从传统机器学习转向深度学习。各种词表、语义语法词典、语料库等数据资源日益丰富，句法分析、词语切分等技术处于快速发展完善阶段。对于文本向量化技术，在基于神经网络的 word2vec 崭露头角时，谷歌工程师在 word2vec 的基础上进行拓展，提出了 doc2vec 技术，目前看来后者略优于前者。在预训练模型方面，CMU 与谷歌大脑提出新的 NLP 预训练模型 XLNet，在 SQuAD、GLUE 等 20 个任务全面超越 BERT[11]；复旦大学计算机科学技术学院副教授邱锡鹏介绍了研究组最新提出的 star-transformer 模型，通过预训练模型以及知识增强(比如 ELMo、BERT、GPT、ERNIE 等)提高模型泛化能力，在自然语言任务上获得了更好的性能。

2、 符号说明及分析方法与过程

2.1 符号说明

符号	说明
$p(a)$	随机变量 a 的分布概率
$p(a\|b)$	随机变量 a 和 b 的条件分布
$p(a,b)$	随机变量 a 与 b 的联合分布
d_i	第 i 个向量的特征
t_j	单词表的第 j 种单词
T	词袋向量 x 和 y 所构成的二元组合
$x^{(i)}$	数据集第 i 个样本的特征向量
$y^{(i)}$	数据集第 i 个样本的标准答案（非结构化预测）
$\boldsymbol{y}^{(i)}$	数据集第 i 个样本的标准答案（结构化预测）
$y\hat{}^{(i)}$	数据集第 i 个样本的预测答案（非结构化预测）

2.2 问题 1 分析方法与过程

2.2.1 流程图

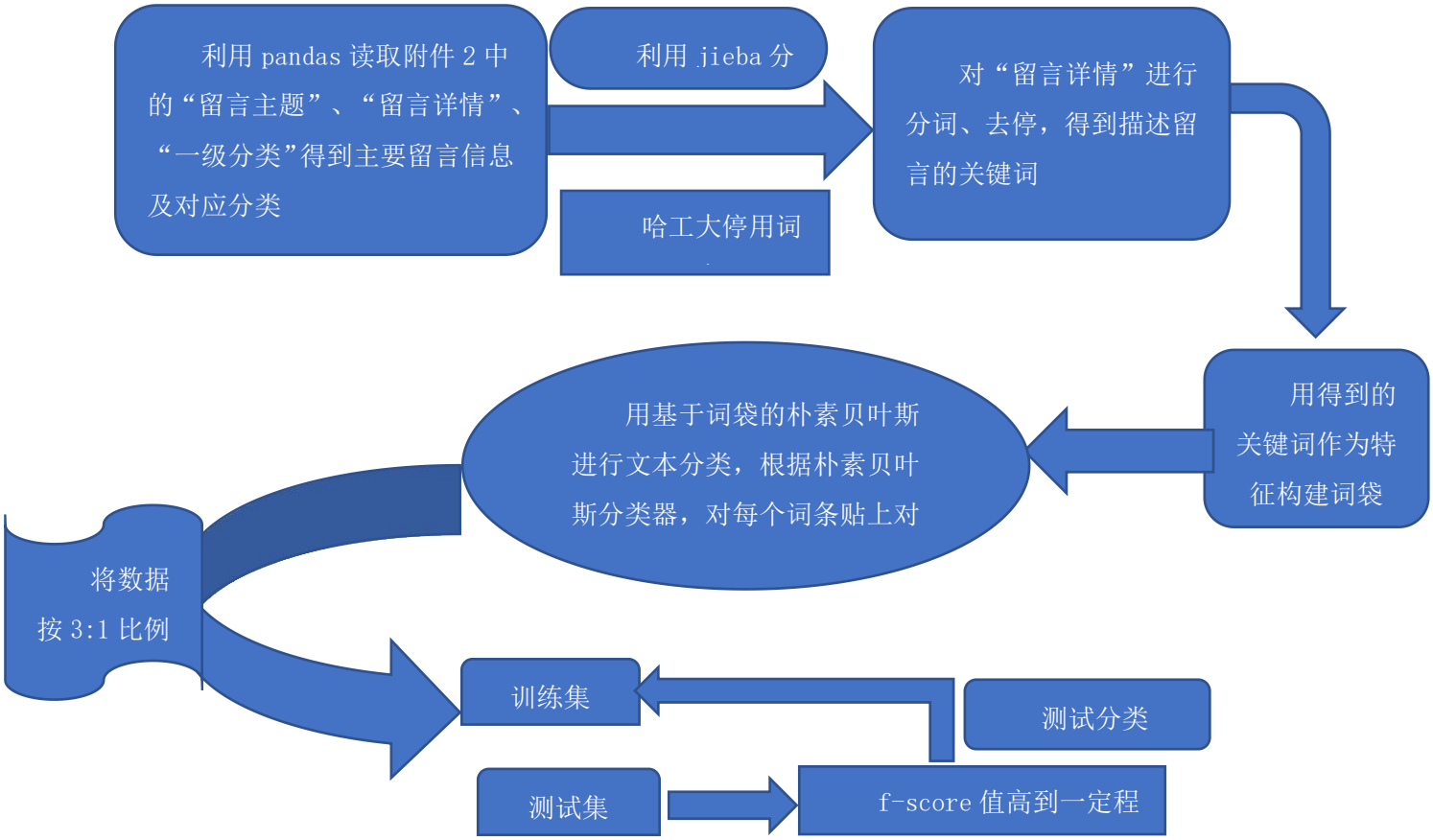

图 1：问题 1 流程图

2.2.2 问题 1 数据分析：

（1）本文中我们按 3:1 的比例将数据分为训练样本集和测试样本集作为实验数据，在对“留言详情”进行分词、去停之后得到关键词，并将关键词进行分类，对“留言详情”贴上对应的一级分类标签。通过对训练样本集和测试样本集的观察分析，本问总结了将留言进行分村类存在的技术难点如下：

文本语义带来的词语交叉

在处理文本信息中，经常会碰到同样的词语运用在不同的句子里面会产生不同的语义。因此，为了正确识别留言内容，怎样让计算机处理、理解自然语言中一个句子的意义成为本文中要解决的一个难点。

（2）长文本的无意义表达太多

在“留言详情”的文本中，存在了大量无关于问题本身的表达，如：问候的礼貌性用语、留言者表示自身愤怒、着急等心理状态的用语、各种互联网用语等。这些无意义表达，这使得存在以下难点：

①词频、词共现频率等信息不能充分利用，丢失掉词语语义间潜在的关联。

②文本较难规范，使得文本中出现不规则特征词和分词词典无法识别的未登陆词。

由此，大大增加了提取留言中有用特征词的难度。

2.2.3 文本多分类

2.2.3.1 数据预处理

（1）数据描述：通过对样本集的观察，可以发现附件 2 中主要信息的字段大多为文本格式，需要将其量化成数值形式才能对其进行分析。一级分类有 城乡建设（占 21.8%），环境保护（占 10.1%），交通运输（占 6.6%），教育文体（占 17.2%），劳动和社会保障（占 21.3%），商贸旅游（占 13.1%），卫生计生（占 9.5%）。对于文本中的信息，存在大量噪声特征与其他无用信息，所以为保证分类的准确率，必须对文本进行中文分词、去停用词等。于是本文先对数据进行预处理。

（2）文本预处理：

①去特殊符号：未经处理的原文本存在前后有许多空格的现象，在计算机输出即为 \n、\t、\xa0、\u3000，如图 3.1 。这将会对分析造成一定影响，可以使用 strip()函数将字符串前后的空格去除，或者使用正则表达式去空格。

②中文分词：中文句子并不像英文句子，词与词之间没有明确的分界，人们是根据日常生活经验和上下文关联进行分词理解，要将文本转化为文本向量，须将文本进行分词提取关键词。由于分词的算法较多，不管是基于规则的算法，还是基于 HMM、CRF 的算法，其分词效果在具体任务中，差距并不算太大，故本文采用较成熟的 Jieba[1]留言详情进行分词，Jieba 算法：

●基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)。

●动态规划查找最大概率路径, 找出基于词频的最大切分组合。

●对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法[2]。

Jieba 分词过程：

●加载字典, 生成 trie 树。

●给定待分词的句子, 使用正则获取连续的 中文字符和英文字符, 切分成 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, 也就是识别未登录词，如进行中国人名、外国人名、地名、机构名等未登录名词的识别。

● 使用 python 的 yield 语法生成一个词语生成器, 逐词语返回。

部分分词结果如图 2，由图得分词后的词列表中，还存在有许多标点符号与无用的停用词，所以下一步要进行过滤停用词。

['\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', 'A3', '区', '大道', '西行', '便', '道', ',', ',', '未管', '所', '路口', '至', '加油站', '路段', ',', ',', '人行道', '包括', '路灯', '杆', ',', ',', '被', '圈', '西湖', '建筑', '集团', '燕子', '山', '安置', '房', '项目', '施工', '围墙', '内', '。', '每天', '尤其', '上下班', '期间', '这', '条', '路上', '人流', '车流', '极', '多', ',', ',', '安全隐患', '非常', '大', '。', '强烈', '请求', '文明城市', 'A', '市', ',', ',', '尽快', '整改', '这个', '极', '不', '文明', '的', '路段', '。', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t']

['\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', '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', '。', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t']

['\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', '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', '!', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t']

['\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', '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', '!', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t', '\t', '\n', '\t', '\t', '\t', '\t']

图 2：部分分词结果

③过滤停用词

文本中如果大量使用停用词容易对页面中的有效信息造成噪音干扰，所以搜索引擎在运算之前都要对所索引的信息进行消除噪音的处理。这些无用的词称为 stopword，表现为使用十分广泛，甚至是过于频繁的一些词或字，文本中出现频率很高，但实际意义又不大的词。为去除停用词，需判断一个词是否为停用词，本文使用哈工大的停用词表，若该词出现在停用词表中，则要过滤掉，经过过滤后便得到文本的关键词，如图 3。

A3 大道 西行 未管 路口 加油站 路段 人行道 包括 路灯 西湖 建筑 集团 燕子 安置 项目 施工 围墙 上下班 期间 路上 人流 车流 安全隐患 请求 文明城市 整改 文明 路段
位于 书院 主干道 在水一方 大厦 一楼 四楼 人为 拆除 电等 设施 烂尾 多年 护栏 围着 占用 人行道 护栏 锈迹斑斑 倒塌 危机 过往行人 车辆 请求 部门 牵头
市政府 交警支队 安监局 环保局 A3 区政府 A3 杜鹃 文苑 小区 业主 涉及 网上 写信 方式 一件 引发 安全事故 杜鹃 雷峰 大道 交界处 杜鹃 文苑 小区 一家 汽车 检测站 紧密 锣鼓 装修 施工 开门 营业 疾呼 相关 部门 前来 调查 关闭 这家 汽车 检测站 道理 家门口 汽车 检测站 一件 事情 反对 网上 举报 形式 反对 这家 汽车 检测站 第一 这家 汽车 检测站 租用 场地 危房 符合 相关 法律 检测站 场地 国家 法律 明文规定 汽车 检测站 检测 宽敞 明亮 整洁 通风 排水 照明设备 工艺 布局合理 防护 设施 齐全 检测站 停车场 小于 检测 面积 这家 汽车 检测站 租赁 一栋 危房 汽车 检测站 改造 车主 生命 车辆 儿戏 更是 生产 制度 开玩笑 试问 危房 开设 汽车 检测站 事故 这家 汽车 检测站 尚未 办理 手续 咨询 交警 得知 机动车辆 检测站 部门 办理 相关 手续 工商行政管理局 营业执照 技术 监督局 组织 机构 代码证 仪器设备 检定 省级 水务局 税务 登记证 交通厅 检测 许可证 物价局 收费 许可证 环保局 环评 报告 保证 这家 汽车 检测站 办理 证件 透露 此家 汽车 检测站 老板 工作人员 夸下海口 出钱 寻找 公司 各个部门 公关 上马 办证 真假 不得而知 希望 交警 交通 工商 多个 部门 依法办事 违规 办事 第三 场地 进出口 道路 狭窄 具备 开设 检测站 条件 开设 汽车 检测站 百分之百 拥堵 安全隐患 周边 居民 何在 第四 汽车 检测站 污染 周边 居民 环境 影响 汽车 检测 包括 一项 尾气 检测 污染 破坏 环境 更会 毒害 居民 噪音 居民 环保局 信访 尊敬 市政府 及市 交警支队 安监局 环保局 领导 试问 汽车 检测站 开业 汽车 检测站 无证 违建 汽车 检测站 扼杀 事故 未发 非法 无证 非法 无证 制造 污染 安全隐患 汽车 检测站 施工 期望 市政府 及市 交警支队 投诉 积极行动 叫停 装修 检测站 依法 叫停 非法 无证 汽车 检测站
胡书记 您好 感谢您 百忙之中 查看 这份 留言 父亲 5.1 A6 金星 北路 明发 国际 工地 工作 5.7 工地 施工 发生 泥土 塌方 受伤 治疗 期间 工地 拒绝 支付 医疗 费用 态度恶劣 工地 生产 深究 父亲 维权 民工 维权 希望 民工 工地 工作 应有 保护 希望 政府 感谢
K8 丁字街 商户 摆摊 前段时间 丁字街 交通 几天 丁字街 做生意 商户 商品 摆到 影响 这条 交通 摩托车 城管局 领导 制定 措施 制止 形为
南门 前段时间 整改 劝阻 摆摊 占道 情况 改善 情况 几天 慢慢 有人 带头 慢慢 摆出来 商户 干脆 钩子 货物 门口 屋檐下 电线 上有政策 对策 城管 检查 稍微 好点 城管 一走 摆出来 样子 希望 部门 强硬 措施 每次 不痛不痒 整治 不到 效果 二小 门口 那条 马路 市场 小菜 卖鱼 水果 成堆
K8 县冷 江东 蓝波 酒店 外墙 装修 架子 无人 施工 路政 酒店 门口 三个 多月 影响 酒店 营业 酒店 施工队 人员 情况 时间 搞好 营业 施工人员 答复
酒店 每月 房租 三万多 工资 水电 费用 一个月 十几万 酒店 三个 营业 损失 酒店 承受 部门 弥补 损失 营业
九亿 广场 城区 休闲 娱乐 场所 景观 很漂亮 每到 晚上 人到 玩耍 两个 公厕 黑黑的 外面 大小便 影响 不好 如果说 不好 管理 景观灯 并网 开关 希望 解决 谢谢
石期 市镇 农贸市场 旁边 公厕 早厕 臭气熏天 老百姓 厕所 无从 下脚 公厕 长年 无人 管理 漏雨 已成 危房 这座 早厕 气味 难闻 夏天 蚊蝇 乱飞 安全 卫生 隐患 石期 市要 发展 公厕 革命 希望 领导 真抓实干 确切 回复 民众 干净 卫生 公厕 希望 网友 关注

图 3：过滤之后的文本关键词

2.2.3.2 文本特征提取

经过数据预处理后，虽然去除了停用词，但还是含有大量词语，不利于后面的文本向量化，所以还需对预处理后的词列表进行关键词提取，以此来降低向量空间维数，从而简化计算，大大提高文本处理的速度和效率。本文采用的方法是 TF-IDF 算法[3]。

TF-IDF（Term Frequency-Inventory Document Frequency，词频-倒排文档频次）

在 TF-IDF 计算方法中，一个词语的重要程度不光正比于它的在文档中的频次，还反比于有多少文档包含它。TF-IDF 的计算方法有许多变种，最基本的形式如下：

$$\mathrm{TF-IDF(t,d)} = \frac{\mathrm{TF(t,d)}}{\mathrm{DF(t)}} = \mathrm{TF(t,d)} \times \mathrm{IDF(t)} \qquad (2-1-1)$$

其中，t 代表单词（term），d 代表文档（document），TF(t,d)代表 t 在 d 中的出现频次，DF(t)代表有多少篇文档包含 t。DF 的倒数(inverse)称为 IDF。

$$\mathrm{TF(t,d)} = \frac{n_{td}}{\sum_k n_{kd}} \qquad (2-1-2)$$

实际应用时会对(2-1-1)式做一些拓展，如加一平滑、对 IDF 取对数防止浮点数下溢出等。IDF 的另一计算式(2-1-2)采用了拉普拉斯平滑，避免有部分新的词没有在语料库中出现过而导致分母为 0 的情况出现，增强了算法的健壮性。

$$\mathrm{IDF(t)} = \log\left(\frac{|D|}{\mathrm{DF(t)}+1}\right) \qquad (2-1-3)$$

|D|为文档总数。

TF-IDF 算法就是 TF 算法与 IDF 算法的综合使用，对于这两种方法如何组合，经过大量的理论推导和实验研究后，发现式(2-1-3)的计算方式之一。

$$\mathrm{tf} \times \mathrm{idf(t,d)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t)} = \left(\frac{n_{td}}{\sum_k n_{kd}}\right) \times \log\left(\frac{|D|}{\mathrm{DF(t)}+1}\right) \qquad (2-1-4)$$

传统的 TF-IDF 算法中，仅考虑了词的两个统计信息（出现频次，在多少个文档中出现），除了上面的信息外，还可以加上每个词的词性、出现的位置等，加上这些辅助信息，能对关键词提取起到很好的提高作用。

2.2.3.3 文本向量化

通过上面提取特征值后，这里我们采用词袋模型，将提取的特征值转化为词袋向量。

词袋模型[4]：(Bag of Words，简称 BoW)，即将所有词语装进一个袋子里，不考虑其词法和语序的问题，即每个词语都是独立的，把每一个单词都进行统计，同时计算每个单词出现的次数。当然，词袋模型有很大的局限性，因为它仅仅考虑了词频，没有考虑上下文的关系，因此会丢失一部分文本的语义。但是大多数时候，如果我们的目的是分类聚类，则词袋模型表现的很好。

定义由 n 个文档组成的集合为 S，定义其中第 i 个文档d_i的特征向量为d_i，其计算方式如下：

$$d_i = (TF(t_1,d_i),TF(t_2,d_i),\ldots\ldots,TF(t_i,d_i),\ldots\ldots TF(t_m,d_i))$$

（2-1-5 ）

其中t_j 表示词表中第 j 种单词，m 为词表大小。 $TF(t_j,d_i)$ 表示单词 t_j在文档d_i中的出现次数。为了处理长度不同的文档，通常将文档向量处理为单位向量，即缩放向量使得||d|| = 1 。

以特征的 id 为下标，频次作为数值，假设一共有 n 个特征，一篇文档就转化为 n 维的词袋向量。这里将词袋向量记作 x，向量的第 i 维记作x_i。将类别记作 y ∈Γ ={c1,c2,.....,ck}，其中 k 为类别总数。则语料库（训练数据集）T 可以表示为词袋向量 x 和类别 y 所构成的二元组的集合：

$$T = \{(x^{(1)},y_1),(x^{(2)},y_2),\ldots\ldots,(x^N,y_N)\} \quad (2\text{-}1\text{-}6)$$

至此，文本已转化为向量，下一步将进行文本分类。

2.2.3.4 文本分类

文本分类又称文档分类，指的是将一个文档归类到一个或多个类别的自然语言处理任务。文本分类是一个典型的监督学习任务，其流程总结为：人工标注文档的类别、利用语料库训练模型、利用模型预测文档的类别。上面我们已经准备好语料库和特征值，接下来将要进行分类器处理，本文采用的分类方法是朴素贝叶斯法[5]。

在各种各样的分类器中，朴素贝叶斯法可算是最简单常用的一种生成式模型。朴素贝叶斯法基于贝叶斯定理将联合概率转化为条件概率，然后利用特征条件独立假设简化条件概率的计算。

2.2.3.5 朴素贝叶斯法原理

朴素贝叶斯的目标是通过训练集学习联合概率分布 p（X，Y），由贝叶斯定理[6]可以将联合概率转换为先验概率分布与条件概率分布之积：

$$P(X = x,Y = ck) = p(Y = ck)\, p(X = x \mid Y = ck) \quad （2\text{-}1\text{-}7）$$

其中，类别的先验概率分布 p(Y=ck)很容易估计，通过统计每个类别下有多少个样本即可（极大似然），即：

$$P(Y = ck) = count(Y = ck)/N \quad （2\text{-}1\text{-}8）$$

而 p(X = x ｜ Y=ck) 则难以估计，因为 x 的量级非常大。这一点可以从式(2-1-9)

$$p(X = x|Y = c_k) = p(X_1 = x_1,\ldots\ldots,X_n = x_n|Y = c_k),k = 1,2,\ldots\ldots,k$$

（2-1-9）

假设第 i 维x_i 有m_i 种取值，那么组合起来 x 一共有$\prod_{i=1}^{n} m_i$种。该条件概率分布的参数数量数指数级的，特别是当特征数量达到十万量级时，参数估计实际不可行。

为此，朴素贝叶斯法“朴素”地假设了所有特征是条件独立的。该条件独立性假设为：

$$p(X=x|Y=C_k)=p(X_1=x_1,\ldots\ldots,X_n=x_n|Y=c_k)\prod_{i=o}^{n}p(X_I=x_i|Y=c_k)$$

（2-1-10）

于是，又可以利用极大似然来进行估计：

$$p=(X_i=x_i|Y=ck)=count(X_i=x_i,y_i=c_k)/count(y_i=ck)$$

（2-1-11）

也就是说，给定类别为 ck 的条件下，特征向量第 i 维为某个特定值x_i 的概率等于类别为c_k且第 i 维为x_i的样本数量除以类别c_k下的所有样本数量。有了 $p(Y=c_k)$和 $p(X_i=x_i|Y=c_k)$之后，朴素贝叶斯模型的参数估计即训练就结束了。

预测时，朴素贝叶斯法依然利用贝叶斯公式找出后验概率 $p(Y=c_k|X=x)$最大的类别c_k 作为输出 y。用公式描述即：

$$y=arg\max_{ck}P(Y=C_K|X=x) \qquad (2-1-12)$$

将贝叶斯公式带入上式得到：

$$y=arg\max_{c_k}[p=(X=x|Y=c_k)\times p(Y=c_k)]/p(X=x)$$

(2-1-13)

由于分母 p(X = x)与 c_k 无关，在求最大后验概率时可以忽略掉，即：

$$y=arg\max_{c_k}p(X=x|Y=c_k)\times p(Y=c_k) \qquad (2-1-14)$$

然后将独立性假设式(2-1-10)到最后的预测分类函数：

$$y=arg\max_{c_k}p(y=c_k)\prod_{i=0}^{n}p(X_i=x_i|Y=c_k) \qquad (2-1-15)$$

2.2.2 部分预测结果

杨书记：　　您好！　　我们是住望仙路187号凯丰.上景鑫苑D栋的住户。今年开始，在我小区东侧开建的“园梦康乐城”，距我们居住的D栋，实测间距14.8米，现在已建到10多层了。低层的住户白天都得开灯，阳台都已经没有一点阳光了。我们的采光权找谁呢？！！！　　房产证也遥遥无期，因开发商拖欠房产局维修基金等资金，导致我们的房产证只办理了100多户（询问房产局得知），证也无法拿到。其他多住户的都没有办理。盼杨书记百忙之中协调指导解决。　　上景鑫苑业主

['城乡建设']

西地省J市J1区增福街长冲三组有一家不锈钢门花厂，厂里污水不经过任何处理，黑心老板晚上把污水偷偷直排到附近河道，重要的是门花厂附近还有一大片居民区，污水主要成分有硫酸 硝酸 磷 对附近居民用地，用水造成了巨大影响 希望有关部门能进快解决，谢谢！

['环境保护']

以上为训练好的模型，通过自定义函数对文本进行分类预测结果，观察发现，对于不同类别，其分类效果和准确性还是比较高的。

2.2.3 模型评估效果

本文采用 F1-sorce[7]进行模型评价。对于每一个类别 c 的分类结果，有如下几个概念：

• True Positives(TP)：被正确地划分为该类的样本个数，即实际为该类且被分类器划分为该类的实例数（样本数）。

• False Positives(FP):被错误地划分为该类的样本个数，即实际为其他类却被分类器划分为该类的实例数（样本数）。

• False Negatives(FN):被错误地划分为其他类的样本个数，即实际为该类却被分类器划分为其他类的实例数（样本数）。

• True Negatives(TN):被正确地划分为其他类的样本个数，即实际为其他类且被分类器划分为其他类的实例数（样本数）。

而精确率、准确率、召回率、F1 的定义如下：

• 精确率（Precision）：$$P=\frac{TP}{TP+FP}$$

• 准确率（Accuracy）：$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

• 召回率（Recall）：$$R=\frac{TP}{TP+FN}$$

• P、R 调和平均：$$F1=\frac{2\times P\times R}{P+R}$$

我们需要衡量模型在所有类别上的整体性能，则可以利用将这些指标在文档级别进行平均：

$$\overline{P}=\frac{\sum_{c_i\in C}TP}{\sum_{c_i\in C}TP+\sum_{c_i\in C}FP}$$

$$\overline{R}=\frac{\sum_{c_i\in C}TP}{\sum_{c_i\in C}TP+\sum_{c_i\in C}FN}$$

$$\overline{F_1}=\frac{2\times\overline{P}\times\overline{R}}{\overline{P}+\overline{R}}$$

其中 C={c1,c2……,ck}。也就是将所有类别下的 TP、FP 和 FN 求和然后计算这些评测指标。这里平均是指微平均（micro-average），还可以使用（macro-average）。

本文使用模型评估训练集得到 F1-score 为 0.97,测试集得到的 F1-score 值为 0.89.

2.3 问题 2 分析方法与过程

2.3.1 数据分析

（1）附件 3 的数据集，和附件 2 类似，有效信息主要为“留言详情”“留言主题”，少了“一级分类”，多了“点赞数，反对数”。

（2）地点和人群的识别

对热点问题的挖掘，即是对一个对相似文本的聚类问题。所以要对文本中的特定地点与特定人群进行识别，但在识别中，存在以下几个难点：

地名、人名、专有名词不容易被 jieba 分词识别出来，且存在地名嵌套现象，构成规律十分复杂。

文本中的缩略词语表示存在多样性，如：“A 市魅力之城小区”可缩略为“魅力小区、魅小”等名称。因此，较难提取构成规则，增大了文本识别的难度。

中文的命名识别研究仍然处于不成熟的阶段，各类词的共享性和制约性都对命名识别带来一定的难度。

（3）相似度计算

1）在对文本提取关键词之后，我们需要对它进行相似度计算，但在这些数据所提取到的特征多，故特征矩阵较大。

2）对文本进行两两之间计算相似计算量大。

2.3.2 数据预处理

问题 2 的数据预处理过程与问题 1 的大致相同，只在提取特征值处加入了命名实体识别，在特征集中加入地名识别，加强聚类效果，由于数据集中人名出现极少，故不设置人名识别。

在中文分词前，需对文本数据进行命名实体识别，提取地名特征加入到之后的特征集中；对数据观察得，地名分为实体地名和英文地名（如:K3 区、A5 县等），后者识别较简单，可用正则表达式提取；而前者实体地名识别，嵌套情况复杂，一个地名经常和另外一个地名组合成新地名，也常常嵌套着人名，本文使用基于条件随机场的命名实体识别[8]。

条件随机场的定义为：设 X=(X1,X2,...,Xn)和 Y=(Y1,Y2,...,Ym)是联合随机变量，若随机变量 Y 构成一个无向图 G=(V,E)表示的马尔科夫模型，则其条件概率分布 P(Y|X)称为条件随机场(Conditional Random Field,CRF),即

$$P(Yv|X,Yw,w\neq v)=P(Yv|X,Yw,w\backsim v) \qquad (2\text{-}2\text{-}1)$$

其中 w∽v 表示图 G=(V,E)中与结点 v 有边连接的所有结点，w≠v 表示结点 v 以外的所有结点。

这里简单举例说明随机场的概念：现有若干个位置组成的整天，当给某一个位置按照某种分布随机赋予一个值后，该整体就被称为随机场。以地名识别为例，假设我们定义了如表 1 的规则。

标注	含义
B	当前词为地理命名实体的首部
M	当前词为地理命名实体的内部
E	当前词为地理命名实体的尾部
S	当前词单独构成地理命名实体
O	当前词不是地理命名实体或组成部分

表 1 地理命名实体标记

现有个由 n 个字符构成的 NER 的句子，每个字符的标签都在(B,M,E,S,O)中选择，当为每个字符选定标签后，就形成了一个随机场。在这个例子中，X 是字符，Y 是标签，P(X|Y)就是条件随机场。

在本文中我们假设 X,Y 结构相同，即结构如图所示。

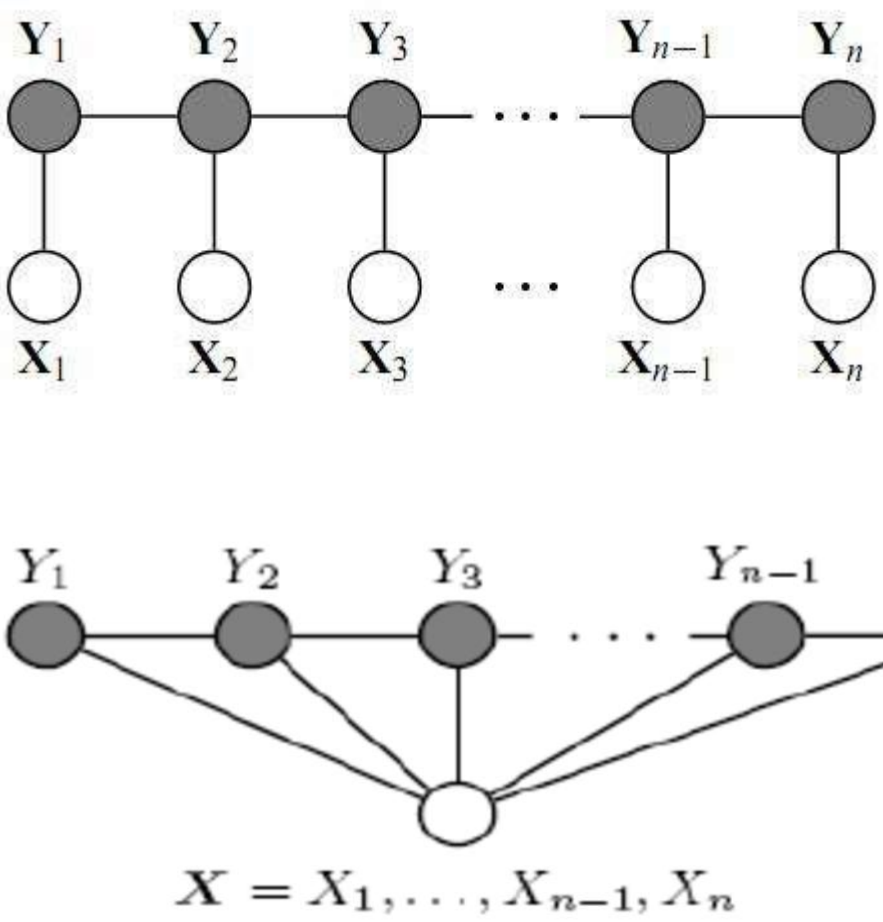

图 4：线性链条随机场结构示意图

一般称这种结构为线性链条件随机场(linear-chain Conditional Random Fields，linear-chain CRF)。其定义如下:

设 X=(X1,X2,...,Xn)和 Y=(Y1,Y2,....,Yn)均为线性链表示的随机变量序列，若在给点的随机变量序列 X 的条件下，随机变量序列 Y 的条件概率分布 P(X|Y)构成条件随机场，且满足马尔科夫性：

$$P(Y_i \mid X, Y_1, Y_2, \ldots, Y_n) = P(Y_i \mid X, Y_{i-1}, Y_{i+1}) \qquad (2\text{-}2\text{-}2)$$

则称 P(Y|X)为线性链的条件随机场。

在 CRF 中有两种特征函数，分别是转移函数 tk(yi-1,yi,i)和状态函数 sl(yi,X,i)。tk(yi-1,yi,i)依赖于当前和前一个位置，表示从标注序列中位置 i-1 的标记 yi-1 转移到位置 i 上的标记 yi 的概率。sl(yi,X,i)依赖当前位置，表示标记序列在位置 i 上为标记 yi 的概率。通常特征函数取值为 1 或 0，表示符不符合该条规则约束。完整的线性链 CRF 的参数化形式如下:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, i) + \sum_{i,l} \mu_l s_l(y_i, X, i)\right) \qquad (2\text{-}2\text{-}3)$$

其中

$$Z(x) = \sum_y \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, i) + \sum_{i,l} \mu_l s_l(y_i, X, i)\right) \qquad (2\text{-}2\text{-}4)$$

Z(x)是规范化因子，其求和操作是在所有可能的输出序列上做的；λk 和 μl 为转移函数和状态函数对应的权值。使用 CRF 做命名实体识别时，目标是求 $\arg\max_y P(y \mid x)$。

2.3.3 K 均值算法

k 均值算法(k-means)[9]基本原理：给定 n 个向量 d1,d2,.....,dn∈RL 以及一个整数 k，要求找出 k 个簇 S1,S2,....,Sk 以及各自的质心 c1,c2,...,ck∈RL，使得下式最小：

$$\text{minimize } \tau_{\text{Euclidean}} = \sum_{r=1}^{k} \sum_{d_i \in S_r} \| d_i - c_r \|^2 \qquad (2\text{-}2\text{-}5)$$

其中||di - cr||是向量与质心的欧拉距离，τEuclidean 称为聚类的准则函数(criterion function)。也就是说，k 均值以最小化每个向量到质心的欧拉距离的平方和为准则进行聚类，所以该准则函数有时也称作平方误差和(sum-of-squared-errors)函数。而质心的计算就是簇内数据点的几何平均:

$$s_i = \sum_{d_j \in S_i} d_j$$

$$c_i = \frac{s_i}{|S_i|} \qquad (2\text{-}2\text{-}6)$$

其中，s_i 是簇S_i 内所有向量之和，称作合成向量(composite vector)。

生成 k 个簇的 k 均值算法是一种迭代式的算法，每次迭代都在上一步的基础上优化聚类结果。其步骤如下：

（1）选取 k 个点作为 k 个簇的初始质心；

（2）将所有点分别分配给最近的质心所在的簇；

（3）重新计算每个簇的质心；

（4）重复步骤(2)和步骤(3)直到质心不再发生变化。

（5）输出结果

k 均值算法虽然无法保证收敛到全局最优，但能够有效地收敛到一个局部最优点。

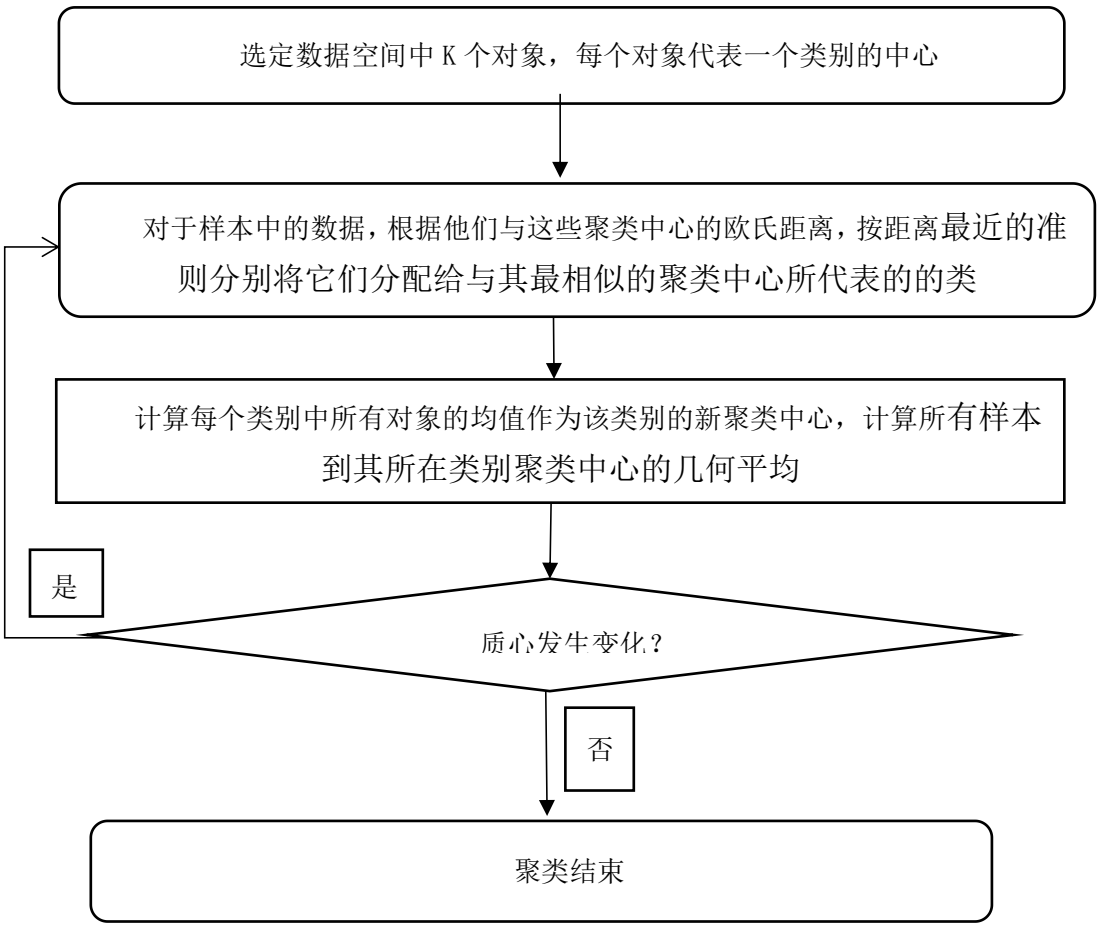

图 5：k-means 均值算法流程图

2.3.4 聚类结果

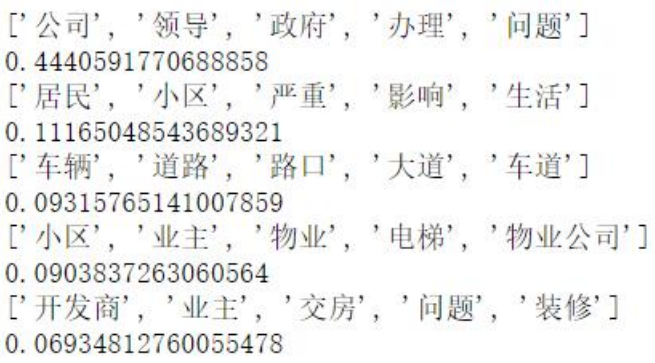

图 6：聚类结果

图 4 为聚类结果中占比最大的五个质心，即为热度最高的五个热点问题，将前五个热点问题的留言详情写入“热点问题明细表.xls ”,这里以问题的留言条数占总数的比作为热点指数。

2.4 问题三分析方法与过程

2.4.1 数据分析

从问题的角度出发，针对附件 4 中的答复意见，从相关性、完整性、可解释性等角度对答复意见给出评价方案，首先需要对有关部门给出的答复意见进行整理和分析，本题的难点大致分为两点：

相关性、完整性、可解释性应该如何进行度量。

这需要根据前面文本挖掘的一些信息去考虑怎么从前面的文本、词的频率、分类等信息来提炼出相关性、完整性、可解释性的度量指标。

对于题目给出的数据需要做出评价并建立一套可行的评价法案。

由于附件 4 的答复意见比较复杂，牵涉到多个方面，采用什么评价方法显得尤其关键。解决此问题，可以采用建立综合评价的方法，综合前面提炼出的比较合适的度量指标形成一套综合评价的体系。

2.4.2 答复相关性

本文采用词向量的相似度来评价文本之间的相关性，即将“留言详情”“答复意见”转化为词向量，从而根据其词向量的相似度来判断相关性，这里使用 word2vec 中的 CBOW 模型。

2.3.2.1 CBOW 模型

CBOW(Continuous Bag of Words Model)[10]是一种基于窗口的语言模型。一个窗口指的是句子中的一个固定长度片段，窗口中间的词语成为中心词，窗口中其他词语称为中心词的上下文。CBOW 模型通过三层神经网络接受上下文的特征向量，下面介绍三层神经网络模型。

神经网络的输入为 4 个上下文单词的独热向量，x(c-m),...,x(c-1),x(c+1),...,x(c+m)（其中 m 为窗口半径，x(c)为中心词，它不包含在内），输出为中心词的独热向量 y。神经网络模型从输入层到隐藏层的权重矩阵 V∈Rn×|V|，从隐藏层到输出层的权重矩阵记作 U∈Rn×|V|。其中，n 为词向量的维度，是一个由使用者自由定义大的超参数。V 和 U 都是词表中所有单词的词向量构成的矩阵，分别称作输入词向量矩阵和输出词向量矩阵。v 中第 i 个列向量为第 i 个单词的输入词向量，记作 vi。

CBOW 模型的任务为 6 步：

● 给定窗口半径 m，为窗口内除了中心词外所有单词分别为生成独热单词：x(c-m),...,x(c-1),x(c+1),...,x(c+m)∈R|V|。独热向量的生成可以通过对词语的词典序进行独热编码实现。

● 将输入权重矩阵乘以独热向量，得到每个单词的输入此向量：vc-m=Vx(c-m),vc-m+1=Vx(c-m+1),.....,vc+m=Vx(c+m)∈R|V|.

● 将这 2m 个上下文词语的词向量求平均，得到上下文词向量：$\hat{v} = \frac{v_{c-m} + v_{c-m+1} + ... + v_{c+m}}{2m} \in \mathrm{R}^{\mathrm{n}}$。

●利用输出词向量矩阵 u 乘以上下文词向量，得到一个分数向量：$\mathrm{z} = \mathrm{U}\hat{v} \in \mathrm{R}^{|\mathrm{V}}$

●利用 softmax 函数将分数向量转化为概率分布：$\hat{y} = soft\max(z) \in \mathrm{R}^{|\mathrm{V}|}$。其中 softmax 函数将向量第 j 维做如下转换：$\mathrm{softmax}_{\mathrm{j}}(z) = \frac{e^{z_j}}{\sum_k e^{z_k}}$。

●CBOW 模型希望自己的预测尽量精准，即希望 $\hat{y} = soft\max(z) \in \mathrm{R}^{|\mathrm{V}|}$ 和真实概率分布 y∈R|V|尽量相似，于是使用交叉熵[11]作为损失函数，利用随机梯度下降算法来优化两个参数矩阵 V 和 U。

其中交叉熵损失函数的定义为：

$$\mathrm{H}(\hat{y}, y) = -\sum_{j=1}^{|V|} y_j \log(\hat{y}_j) \qquad (2\text{-}3\text{-}1)$$

由于向量 y 为独热向量，其元素只有在下标 j 等于中心词的下标 i 时才不为 0，于是上式简化为：

$$\mathrm{H}(\hat{y}, y) = -y_i \log(\hat{y}_i) = -\log(\hat{y}_i) \quad (2\text{-}3\text{-}2)$$

CBOW 希望式(2-3-2)的值小，即希望预测出的分布里中心词的概率 p(wc|wc-m,...,wc-1,wc+1,...,wc+m)变大。对式(2-3-2)中的参数求偏导数，利用梯度下降法更新参数即可得到 V 和 U。

2.4.2.2 训练词向量

要训练词向量就必须有大量的语料库，由于中文的语料库较少，这里我们采用语料库来源有两部分：

（1）网络收集的词典合并；

（2）在公众号文章上做新词发现，可用性较高。

训练完成得到的模型，可用于计算短文本相似度，步骤为：

①短文本关键词提取；

②关键词向量化；

③相似度计算。

若模型计算得到关于“留言详情”“答复意见”的短文本相似度达到一定数值时，即判定为相关。

2.4.3 答复可解释性、完整性

可解释性：要求答复能够提供有效信息于留言者，准确回复留言者的问题，帮助留言者理解答复的含义。

答复的可解释性的度量指标有答复时间差、反馈评分等。根据留言时间和答复时间的距离判断，一般答复时间与留言时间离得越久，则答复内容对于留言者提供的有效信息便会减少，即可解释性减少。

根据反馈评分判断，由留言者对答复意见进行评分，将其所认为能否在其中获取有效信息以及能否看懂其中意思进行信息反馈，得到的反馈信息将占有最大的权重判断可解释性。

完整性：要求答复内容能与留言信息一一对应。

3、结论

对老百姓的留言进行分析分类研究，对于政府部门解决民生、对于我国的发展都具有重大意义。同时这也是文本分析、文本分类的一个课题、难题。传统的文本解读方法已经不能满足愈来愈庞大的数据量。

总结本次比赛，本文通过 pandas，Jieba 对数据与文本进行预处理。通过 TF-IDF 算法，将词袋模型转化为词袋向量，对预处理后的数据、文本进行特征词提取。实现了给不同的留言贴上对应的一级分类标签。通过 k-means 算法，对群众所反映的问题进行聚类，分析了在一段时间内集中爆发的问题，以及问题的地点、人群等。更有利于相管的部门更快、更准确地处理职责范围内的问题，早日给群众一个满意的答复。最后我们通过 CBOW 模型、训练词向量等方法深入分析了相关部门给出的完整性、相关性、可解释性。并给出了一套评价系统。

但是，我们最后的文本分类效果精确率不是特别高；第二问中的聚类效果不好，这是因为初始质心选取不当，可能收敛到一个较差的局部最优点，且只设置了 10 个质心；综合评价系统中可解释性和完整性的度量指标较少，故评价系统不全面。总结得到，上述不足高度取决于特征选取的好坏，这也涉及到当今文本挖掘的不足，提醒我们不应停留在传统的机器学习算法中，而应走进深度学习，比较两者间的优劣，我们也会继续对文本挖掘进行深入探讨与学习。

4、参考文献

[1]王庆福. 隐马尔可夫模型在中文文本分词中应用研究[J] 无线互联科技 2016 第 13 期

[2]赵庶旭，伍宏伟，刘昌荣. 基于隐马尔可夫模型的交通最优路径模型[D] 测试科学与仪器(英文版) 2020

[3]、[6] 吴军 . 数学之美[M] 北京市丰台区成寿寺路 11 号：人民邮电出版社 2014

[4]黄春梅，王松磊. 基于词袋模型和 TF-IDF 的短文本分类研究[J] 软件工程 2020

[5]、[9] 何晗 .自然语言处理入门[M] 北京市丰台区成寿寺路 11 号：人民邮电出版社 2019

[7]秦彩杰，管强. 一种基于 F-Score 的特征选择方法[J] 宜宾学院学报 2018

[8]涂铭，刘祥，刘树春. Python 自然语言处理实战核心技术与算法[M] 北京市西城区百万庄大街 22 号：机械工业出版社 2018 65-68

[10]安俊颖. 深度学习方法训练词向量[J] 中国新通信 2018 20 期

[11]张超然，裘杭萍，孙毅，王中伟 . 基于预训练模型的机器阅读理解研究综述[J].计算机工程与应用 2020-04-20