基于文本挖掘的“智慧政务”研究

摘要

本文为解决分类留言、热点挖掘和答复评估等“智慧政务”的文本挖掘问题，使用高斯朴素贝叶斯分类、K-Means 文本聚类和 LDA 主题模型进行文本数据挖掘，并评估文本挖掘应用的效果。

问题一，根据对附件 2 的留言详情和类别，首先利用 Excel 筛选得到 7 个类别，利用 Python 进行去重、jieba 中文分词、去停用词和 TF-IDF 模型向量化，完成文本数据预处理，并利用分词制作出所有类别的词云图，直观看出每个类别的区别；接下来，进行构建训练集与测试集，建立高斯朴素贝叶斯一级分类模型，对所有类进行分类，将 1 个七分类转换成 7 个二分类问题，利用模型分类七次，得到模型精度、F-Score、ROC 曲线和 AUC 数值进行评估，再进行 10 次实验评估分类的稳定性、作图分析，最终得出分类效果较为稳定，故取第一次 F-Score 为 F1=0.8587，模型的效率高、差错率较低。

问题二没有类别与留言信息对应，需要进行归类，利用 Python 对附件 3 留言主题进行去重、jieba 中文分词、去停用词和 TF-IDF 模型向量化，构建 K-Means 文本聚类模型，并使用 Calinski-Haeabaz Index(CHI)评价方法检验聚类效果，得到 15 个类别，利用 Excel 筛选得到 8 个类，定义影响热度的指标为：该类事件出现次数、该类问题维持事件、该类问题赞成数和反对数，分别给予 0.4、0.3、0.1、0.2 的权重，构建热度计算模型，得到热度排名前五的热点问题。

问题三从答复的完整性、相关性、可解释性和重要性四个角度出发，利用 Python 对附件 4 的答复内容进行机械压缩/去重、中文分词、去停用词，得到具有实际意义的分词，导入情感评分表，并使用 Python 进行情感分析评分，得到正向(民生、社会等问题)和负向(较严重问题)的分数和权重，构建 LDA 主题模型，分别得到 3 个答复内容的正向和负向的主题，并从正向和负向的角度对答复进行评价，结论是：答复内容充分和有感情且有交代解决方式和解决问题的部门。

综上，本文基于 Python 和 Excel 构建数学模型、数据模型完成文本挖掘，通过对模型的评估，可以得到结论有一定合理性，相信可以解决“智慧政务”中的文本挖掘问题。

关键词：TF-IDF　高斯朴素贝叶斯　K-Means 文本聚类　LDA 主题模型　情感分析

Research on "Smart Government" based on Text Mining

Abstract

In order to solve the text mining problems of "Smart Government", such as classified message, hot spot mining and reply evaluation, this paper uses Gaussian Naive Bayes classification, K-Means text clustering and LDA topic model to mine text data, and evaluates the effect of text mining application.

In Task 1, based on the message details and categories for Annex 2, seven categories are screened by Excel. Python is used for deduplication, Chinese word segmentation based on jieba library, stopwords removing, and TF-IDF model vectorization. Complete text data preprocessing, and use word segmentation to make word cloud diagrams of all categories, intuitively see the difference of each category. Next, build training set and test set, establish Gaussian Naive Bayes first level classification model, classify all categories, convert one seven-classification into seven two-classification problems, and use the model classification seven times, then get model accuracy, F-Score, ROC curve Line and AUC values to evaluate. After 10 experiments to evaluate the stability of classification and analyse drawing, it is concluded that the classification effect is relatively stable, so taking the first F-score as F1 = 0.8587, the model has high efficiency and low error rate.

In Task 2, there is no category corresponding to the message information, so it needs to be classified. We use Python to deduplicate the message topic in Annex 3, Chinese word segmentation based on jieba library, remove stopwords and TF-IDF model vectorization. Build K-means text clustering model, and use Calinski-Haeabaz Index (CHI) evaluation method to test the clustering effect, and then 15 categories are obtained. Eight categories are selected by Excel. The indexes that affect the heat degree are defined as: the number of such events, the maintenance events of such problems, the number of pros and cons of such problems. The weights of 0.4, 0.3, 0.1 and 0.2 are given respectively, and the heat degree calculation model is constructed to get the top five hot issues of heat degree.

In Task 3, from the perspective of completeness, relevance, explainability and importance of the reply, Python is used to mechanically compress / deduplicate the reply content of Annex 4, Chinese word segmentation and stopwords removing, so as to get the word segmentation with practical significance, import the emotional scoring table, and use Python to score the emotional analysis, so as to get the positive (livelihood, social and other issues) and negative (more serious questions) According to the score and weight of question), the LDA topic model is constructed, and three positive and negative themes of the reply content are obtained respectively. The reply is evaluated from the positive and negative perspectives. The conclusion is: the reply content is full and emotional, and the Department that has explained the solution and the problem-solving department.

To sum up, based on Python and Excel, this paper constructs mathematical model and data model to complete text mining. Through the evaluation of the model, we can get a conclusion that is reasonable, and we believe that it can solve the problem of text mining in "Smart Government".

Keywords: TF-IDF; Gaussian Naive Bayes; K-means Text Clustering; LDA Topic Model; Sentiment Analysis

目录

一、挖掘目标

1.1 挖掘背景

当今，网络信息迅猛发达，微信、微博、市长信箱、阳光热线等信息化产物为政府和公众带来了不少的实惠。政府通过这些方式了解民意、汇聚民智、凝聚民气，促进政府和民间的沟通成为常态，通过听取、吸纳建议，去寻求新的改变。

而与此同时，带来的便是数据量庞大，操作繁琐，效率低下，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。在人工智能、大数据、云计算等技术的发展下，代替以往的人工划分及整理，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，这对提升政府的管理水平和施政效率具有极大的推动作用。

然而，现如今大部分电子政务系统还是依靠人工根据经验处理，这无疑存在着工作量大、效率低，且差错率高等问题。通过收集互联网公开的群众问政留言和相关部门的答复意见，利用文本挖掘及自然语言处理技术对群众留言进行分类、对热点问题进行挖掘并且对答复意见进行评价，构建类似的智慧政务系统对提高政府相关部门的施政效率及管理水平有着重要意义。

1.2 挖掘目标

参照附件 1 中工作人员提供的内容分类三级标签体系，根据附件 2 的数据，建立关于留言内容的一级标签分类模型并通过 F-Score 对分类方法进行评价。

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，给出排名前 5 的热点问题和相应热点问题对应的留言信息。

最后，根据附件 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尝试实现。

二、总体流程与步骤

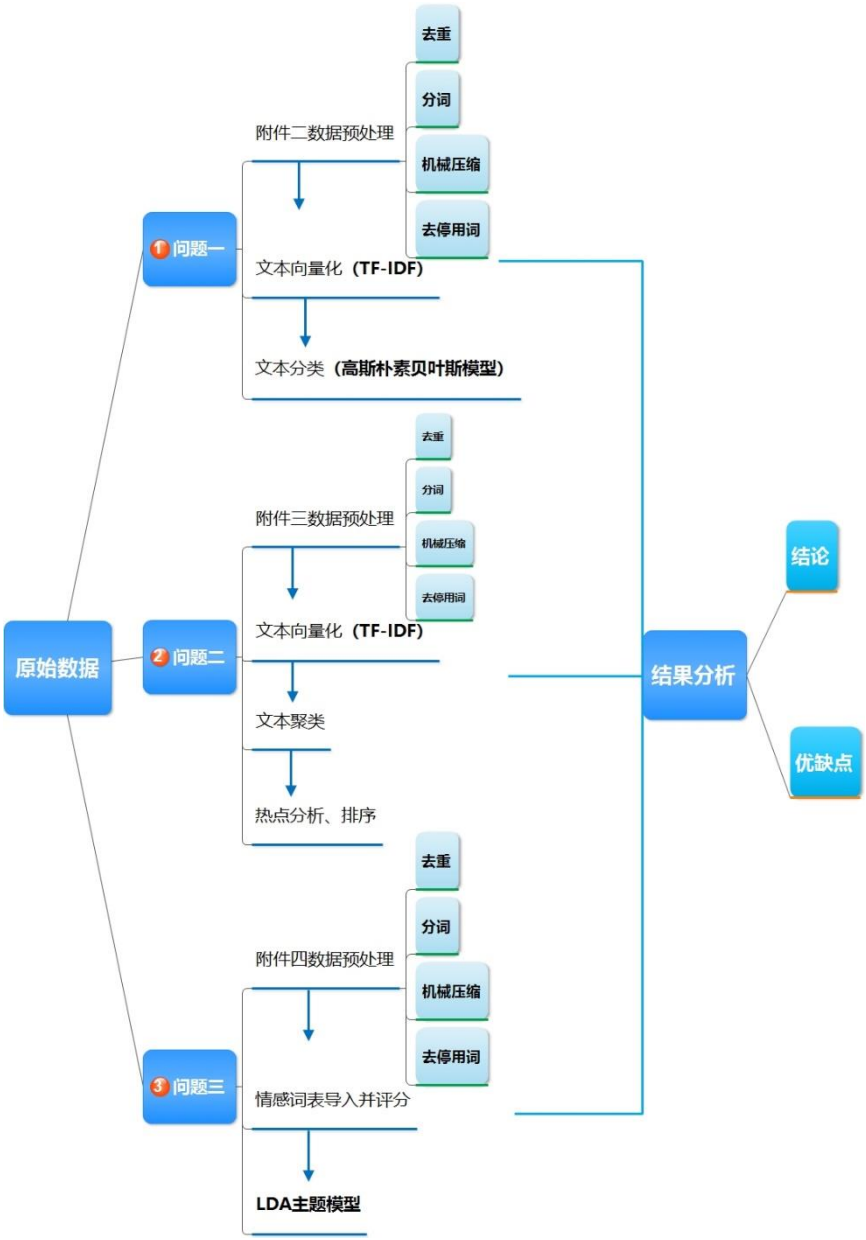


图1　总体流程图

三、文本数据的分析

通过观察附件 1、附件 2、附件 3 和附件 4 的数据，可以发现其数据是文本数据并且数据量较大，一共近 8 万左右的文本数据。根据对题目的分析，我们把附件 1 作为对分类参考的文本数据，而附件 2、附件 3 和附件 4 作为建立模型、解决问题和研究分析的为文本数据，所以必须进行数据预处理和量化成数值，不然会对后续分词、分类和聚类造成较大的影响和误差。对题目进行分析后，挑选出三个附件中较重要的数据：附件

2：留言详情、附件 3：留言主题、附件 4：答复详情。下面我们对三个附件中的重要数据进行去重/机械压缩、去停用词、中文分词以及 TF-IDF 向量化处理并建立数学模型，解决挖掘目标中的问题。

四、留言文本分类模型

4.1 问题一的基本流程

对于附件 2，由于在分类中留言详情可以作为信息量较大的信息，利用此文本进行构建文本分类模型可以使得分类效果更好，所以选择留言详情这一列文本数据并根据附件 2 的类别列，进行数据挖掘分析；并且由于附件 2 已有一级标签分类，所以利用 Excel 筛选统计并对应附件 1 的 15 个一级分类标签，我们得到附件 2 一共有 7 类一级分类标签，所以我分别设每个一级标签为 0，而其他六个标签为 1(例如第 1 次分类：城乡建设为 0，其他为 1；第 2 次分类：交通运输为 0，其他为 1，以此类推，进行 7 次)，每次取 100%文本数据进行文本分类，将 1 次多分类问题转变为 7 次二分类问题，并且判断 7 次文本分类的效果，进行数据挖掘和建立数学模型，具体流程如下所示：

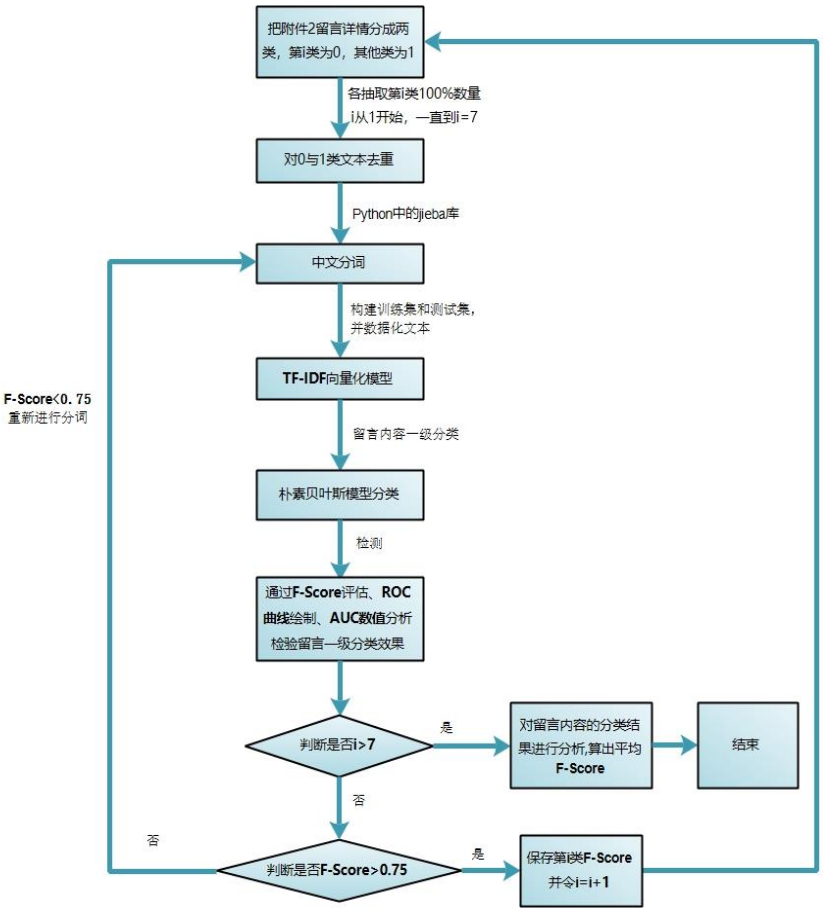


图2　问题一流程图

4.2 数据预处理

4.2.1 文本去重

首先我们对抽取的留言文本数据进行去重，利用 Python[1]对文本进行消去重复词语和清洗数据，部分去重结果如下：

108270	\n	\n	省十二届人大常委会第十三次会议分组审议了《西地省人口与计划生...
168028	\n\t\t\t\t\t\n\t\t\t\t\t		尊敬的环保局领导： 你们好！ 我是I市...
342799	\n	\n	我们在外务工人员现在有大部都买社保了，购买社保的还需要缴纳农村合...
345410	\n	\n	我是M3县燎原机械厂的双职工，俩口子在该企业工作三十多年设分文退...
117555	\n	\n	尊敬的胡伯俊厅长您好：\n\n 我名王微，女，44岁，M2县人...
117513	\n\t\t\t\t\t\n\t\t\t\t\t\t		K市K9县新圩镇株木水村世盛工贸公司，工厂排...
8074	\n\t\t\t\t\t\n\t\t\t\t\t		胡书记，您好！ 我是一位平民老百姓，请...
38579	\n	\n	独生子女大了.父母老了.还等的到发的那一天吗 \n ...
229363	\n	\n	杨书记好，我是L4县第一中学的一名普通的高三学生，我们之前已...
8230	\n\t\t\t\t\t\n\t\t\t\t\t		易书记您好！ 我是今年7月1号被朋友骗...
291151	\n	\n	近期K8县各乡镇都在交新农合，我在家乡一工厂务工，工厂已交社保，...
181666	\n	\n	我父亲在G7县天源机械有限责任公司工作27年左右，现年63岁...
98142	\n\t\t\t\t\t\n\t\t\t\t\t		居民拿自己用过的煤炭瓶去充气煤气站不充说过期...
303803	\n	\n	我是西地省一家民营企业的员工，公司给我交了五险一金，交的是西地省...
117927	\n	\n	厅长： 2015城镇居民收入西地省7840，湖北7720，西地...
153999	\n\t\t\t\t\t\n\t\t\t\t\t		我们是A市A2区老虎塘路65号银杏嘉园9栋全...
457	\n\t\t\t\t\t\n\t\t\t\t\t		您好！本人是19年7月份在A市A3区莱茵城购...
77889	\n\t\t\t\t\t\n\t\t\t\t\t		08年，F市城建局把两湖派出所前面的鱼塘...
15847	\n	\n	在县、乡两级实实在在做事的，都不是公务员，一般是机关职工和事...
44427	\n\n王局长：		您好！我是F7县一名新的乡村教师，请问我有想法向F市教育局发点自己写的小..
185648	\n\t\t\t\t\t\n\t\t\t\t\t		刘厅长： 您好，我想了解嘉雨路预备役师段什...
105839	\n	\n	在楚雅住院，才听病友说居民医保可以二次报销，（他是M7市的，...
138030	\n	\n	国务院都已经鼓励各企事业单位放2.5天小长假，但昭山示范区依...
334485	\n\n请求尊重我们的教育选择权还教育一方净土尊敬的政府、教育局领导：我们地处临近M2县的D..		
130577	\n	\n	请问一下领导：我是一家公立医院的正式职工，是集体户口，老公是...
129239	\n	\n	市、区教育局今年适儿单入学统计，符合在市区就读的小学一年级新...
11544	\n	\n	我们是今年高一新生的家长，希望局长关心一下B市二中的学校管理...
117517	\n\t\t\t\t\t\n\t\t\t\t\t		红太阳的职工宿舍上有个大烟囱，每天排放废气，...
36988	\n	\n	我是B市B6县三中一名学生的家长！我孩子他们学校利用暑假偷偷...
33083	\n	\n	今年，我邕初三毕业，想知道C市（县）一些高中学校的高考上线人...

图3 利用 Python 对文本去重及数据清理后的部分结果

4.2.2 文本分词

可以看到上图为去除停用词后的效果，但还未分词与去除停用词，为了让词语与词语之间有较明显的差别，下面我们先对抽取的文本数据进行中文分词处理，本文我们利用 Python 的第三方库--jieba 库[2](包括问题二和问题三)进行中文分词。

(一)jieba 库分词原理[2]：jieba 分词依靠中文词库：

- 利用一个中文词库，确定中文之间的关联概率；
- 中文间概率大的组成词组，得到分词的结果；
- 除了分词，可以加入和添加自定义的词组。

(二)jieba 库使用的数学和计算机算法[2]：

1.基于 Trie 树结构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)；

2.采用了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；

3.对于未登录词，采用了基于汉字成词能力的隐式马科夫(HMM)模型，使用了 Viterbi 算法(搜索最佳路径)。

(三)jieba 支持的三种分词模式[2]：

1.精确模式：试图将句子最精确地分开，适合文本分析；

2.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；

3.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。

通过 jieba 分词得到部分分词结果展示：

```
154040    ['\n', '\n', '西地省', '农机局', '流通', '协会', '在', 'C', '市', '九华', '召开', '《',...
94729     ['\n', '\n', '尊敬', '的', '黄', '县长', '：', '，', '，', '您好', '！', '请', '您', '帮'...
46224     ['\n', '\n', '裘', '书记', '：', '，', '，', 'A8', '县', '虎山', '特殊教育', '学校', ...
130791    ['\n', '，', '，', '，', '，', '，', '\n', '，', '，', '，', '，', '，', '，', 'L'...
298526    ['\n', '\n', '希望', 'C', '市', '十一', '中非', '驳人', '的', '场所', '2019', '年',...
22312     ['\n', '\n', '尊敬', '的', '周辉', '县长', '：', '，', '，', '，', '我名', '秦安', '加',...
153955    ['\n', '\n', 'A6', '区', '行政', '中心', '超市', '商品', '普遍', '比', '市场', '价格',...
147086    ['\n', '，', '，', '，', '，', '，', '\n', '，', '，', '，', '，', '，', '，', '，', '为'...
49811     ['\n', '\n', '，', '，', '2016', '年', '3', '月', '7', '日', '下午', '14', '点',...
72808     ['\n', '\n', '，', '，', '从', 'M3', '县', '，', '，', 'M4', '市至', 'B5', '县', '，', '，', ...
80132     ['\n', '\n', '陈', '局长', '：', '，', '，', '您好', '！', '打挠', '您', '繁忙', '的', ...
233244    ['\n', '，', '，', '，', '，', '，', '\n', '，', '，', '，', '，', '，', '，', '，', 'K'...
39451     ['\n', '\n', 'B', '市堤', '香蓝岸', '六', '、', '七栋', '因', '一楼', '靠', '路边', '门'...
273512    ['\n', '，', '，', '，', '，', '，', '\n', '，', '，', '，', '，', '，', '，', 'C', '市', '碱'...
64779     ['\n', '，', '，', '，', '，', '，', '\n', '，', '，', '，', '，', '，', '，', '尊敬', '的', ...
229305    ['\n', '，', '，', '，', '，', '，', '\n', '，', '，', '，', '，', '，', '，', '，', '2'...
52897     ['\n', '\n', '，', '，', 'D9', '县新', '车站', '搬迁', '后', '，', '，', 'D9', '县到', '大'...
352075    ['\n', '，', '，', '，', '，', '，', '\n', '，', '，', '，', '，', '，', '今年', 'D6',...
```

图4　通过 jieba 分词得到部分分词结果展示

4.2.3 使用停用词

停用词也叫功能词，一般指在文本内容中出现的次数和频率较高或者较低的虚词、介词、代词和一些字符，与其他类别的词语相比，停用词通常没有实际含义。根据部分分词结果的案例，可以发现，分词后的结果还是杂乱无章的，许多行中的一些词语的分词结果是不需要或是没有意义的，所以必须去除一些文本内的没有意义的词语，即去除留言详情中产生作用较小的分词。

◆ 使用停用词的具体原因：

1.单独分开就无意义的虚词和助词，在中文的表达中最常用的是功能性词语是限定词，如“个”、“的”、“那”、“是”、“得”等等。出现的频率较高，只有组成一个句子时，才会有实际意义。

2.在互联网上频繁出现但是没有实际意义的词语，比如“如果”、“http”，对文本分析作用不大，为了提高分词的效果和搜索的效率，也需要对这类词语停用。

所以通过导入停用词表（详情见附件），对中文分词进行去除停用词表的词语，优化分词效果，提高后续建模效率，利用 Python 处理分词，得到：

```
0
118751    [尊敬, 彭, 书记, , , , , , , , , , , , , ...
83716     [小学教师, 声讨, F, 市, 人民政府, 樹, 上午, 分, F, 市, 教育局, 素质...
228847    [外来, 户口, C, 市, 工作, 年, 小孩, C, 市, C2, 区鹿, 金, 小学,...
5631      [厅长, 同志, G8, 县自, 2011, 年, 元月, 全县, 实行, 药品, 差价, ...
109284    [19, 年, 29, K8, 县, 冷水, 镇政府, 内部, 群发, 休年, 假, 通知,...
277811    [请, D, 市, 纪委, 严查, 高新, 开发区, 教育局, 200, 小学, 学位, 秋...
117414         [K9, 县, 县城, 乡里, 中巴车, 一岁, 以内, 小毛, 仔都, 收钱, 买票]
238492    [尊敬, 陈, 主任, 西地省, 结核病, 防治所, 西地省, 胸科, 医院, 周边, 居民...
312805    [尊敬, 领导, 好多, 市, 各区, 教育, 新闻, 建, 学校, 升学, 唯独, A6,...
154089    [前两天, 朋友, 小区, 里, 碰到, 电梯, 故障, 楼层, 电梯门, 开, 电源, 吓...
316012    [急待, 抢救, 中国, 民间, 瑶族, 博物馆, 占, 全国, 总人口, 0.2%, 瑶族...
264773    [G, 市, 教育局, 领导, 2018, 年, 受, 刺激, 全市, 中小学, 师生, 暑...
273285    [六间, A3, 区, 一小, 李校长, 一问, 李校长, 通气会, 第一, 时间, 通知,...
161541    [尊敬, 刘厅长, 新, 农村, 环境, 领导, 关心, 领导, 质, 变化, 垃圾, 回收...
48777     [C3, 县, 分水, 坳, 村, 境内, 国道, 路面, 坑坑洼洼, 车子, 时, 尘土飞...
70401     [患者, 李雪梅, 1999, 年, 日因, 受伤, 送入, B6, 县, 医院, 治疗, ...
98936     [尊敬, 张, 厅长, K, 市, K2, 区岚角, 山镇, 中心医院, 医务人员, 屈, ...
124332    [近期, 云, 汇聚, 英在, L, 市, L1, 区, 登陆, 打着, 网络平台, 幌子,...
51551     [工伤, 管理中心, 民营, 康复, 医院, 借调, 人员, 年, 工伤, 管理中心, 发工...
170280    [G5, 县, 出租车, 秩序, 混乱, 重来, 规矩, 办事, G5, 县, 出租车, 重...
323014    [九华, 经开区, 小学, 就读, 划分, 议论, 事情, 听到, 最多, 公平, 开发商,...
160275              [文田镇, 石羊, 山寨, 申报, 旅游景点, 不错, 圣地, 请, 多多关照]
305038    [严, 书记, 钧鉴, 何义荣, K7, 县, 图书馆, 退休干部, 做县, 文物保护, 工...
210313    [45, 岁, 时, 女友, 2013, 生有, 儿子, 出生, 医学, 证明, 回家, 上...
117731    [K, 市, K1, 区萍洲, 西路, 阳光, 药厂, 每天晚上, 12, 点至, 点, 之...
77899     [兹, F, 市, F7, 县, 南江, 镇, 黄, 裴村, 老屋, 组与, 杨家, 组, ...
192011    [医保, 每口人交, 150, 财政, 资助, 50, 傻子, 几亿, 老百姓, 一共, 交...
103267    [尊敬, 詹, 书记, 老婆, 前, 日子, 肚子疼, 历害, 县, 医院, 一查, 竟是,...
167040    [L12, 市, 安江, 镇, 亿鑫, 旁边, 邮政, 报刊亭, 超出, 经营范围, 违反,...
282301    [农民工, 孩子, 跟随, 父母, 转学, 原, 学校, 盖章, 教导, 主任, 孩子, 档...
```

图5 利用 Python 处理分词后部分结果展示

可以看到去停用词后的中文分词的效果比起去停用词前的效果好了很多，那么对文本数据的预处理已经完成，下面对文本进行向量化成数值矩阵，利用建立文本分类模型并评估效果。

4.3 文本特征提取以及 TF-IDF 向量化

在上面我们已经利用 Python 对留言详情去重、分词以及去停用词得到有实际意义的文本。但是得到的新分词需要进行特征提取并且向量化，即给予分词中能表达出这个留言的特征的词语的权重大小和重要程度，才能继续建模分析。下面我们引入 TF-IDF 算法[3]。

1.TF-IDF 是一种适用于信息检索和文本数据挖掘的技术，也是一种统计方法，通过文档分词后的词语来评估一段文本的权重，优点是简单快速；

2.TF-IDF 由两个词语的组合，TF(Term Frequency,即词频)表示一个词语在文档中出现的次数和频率，IDF(Inverse Document Frequency)表示逆向文件频率，TF 和 IDF 是成反比关系的；

3.TF-IDF 算法的原理是，如果文档中的一个词语出现的频率(词频)很高，但这个词语在其他文档的词频是很小的，那么我们可以认为，该词语有较好的区分能力，适合用来分类；

4.TF-IDF 算法同样使用在问题二与问题三中的数据向量化中。

下面我们进行建立 TF-IDF 向量化模型：

(1)词频(TF)计算公式:

$$\mathbf{TF_{i,j}} = \frac{\text{该词语在文档出现的个数}}{\text{文档总词语的个数}} \tag{4.1}$$

转换成数学算式，即：

$$\mathbf{TF_{i,j}} = \frac{\mathbf{n_{i,j}}}{\sum_k \mathbf{n_{k,j}}} \tag{4.2}$$

其中，n_{ij}表示该词语在第 j 个文档出现的个数，$\sum_k n_{k,j}$表示该文档总词语的个数。

(2)逆向文档频率是指如果一个文档的词语高频出现但在其他多数文档低频出现，那么这个词语的权重会较高。逆向文档频率(IDF)计算公式：

$$\mathbf{IDF_i} = \log\left(\frac{\text{语料库的文档总数量}}{\text{包含该词语的文档数量+1}}\right) \tag{4.3}$$

(3)TF-IDF 模型计算词语权重：

$$\mathbf{w_{ij}} = \mathbf{TF_{i,j}} \times \mathbf{IDF_i} \tag{4.4}$$

我们抽取 80%的留言详情文本数据为训练集,而 20%的留言详情文本数据作为测试集，利用 Python 的 sklearn 库[1]将训练集和测试集进行 TF-IDF 向量化，使用 Python 的 CountVectorizer 使得向量长度相同，分别得到附件 2 训练集 X_tr 和测试集 X_te 的留言详情的词语-文档矩阵。

```
X_tr

array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

```
X_te

array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.02590445, 0.02590445, ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

图6　部分留言详情-文档矩阵输出结果

4.4 词云分析

根据分词和类别，我们对7个一级分类留言进行词云分析，使用Python的WordCloud库进行制作词云图[1]，得到城乡建设、劳动与社会保障、教育文体、商贸旅游、环境保护、卫生计生和交通运输的词云图：

图7　7个一级分类词云图

词云图可以直观得到7个类别的分词出现频率的大小，词语越大，表示出现的次数高，可以直观根据词语大小对不同类别进行区分和分析。

城乡建设：词语较大的是“业主”、“小区”、“居民”，可以得到词语出现较多的是社区类问题，那么可以利用词语进行区分。

劳动与社会保障：词语较大的是“工资”、“人员”、“劳动”，可以得到该类留言内容可能大多数是就业、工作类问题。

商贸旅游：词语较大的是“电梯”、“工资”、“价格”，可以得到该类留言内容可能比较多是旅游价格、旅游设施等问题。

教育文体：词语较大的是“学校”、“老师”、“教育”，可以得到该类留言内容

可能是学校类问题。

环境保护：词语较大的是“污染”、“环境”、“环保”，得到该类留言内容可能是环境污染等问题。

交通运输：词语较大的是“出租车”、“司机”、“收费”，可以得到该类留言内容可能是出行问题和出行费用等问题。

卫生计生：词语较大的是“医院”、“医生”、“生育”，可以得到该类留言内容可能是在医院的一些问题，如生育情况。

4.5 建立高斯朴素贝叶斯一级分类模型

通过 TF-IDF 向量化算法得到了训练集 X_tr 和测试集 X_te 的一个词语-文档矩阵，那为了建立一级分类标签模型，我们需要利用留言详情的词语-文档矩阵进行分类，并且评估。这里我们引入高斯朴素贝叶斯模型。

当样本变量$\mathbf{x_n}$是连续变量时，考虑使用高斯朴素贝叶斯公式[4]进行样本分类，其经典假设是：和每一个类相关的连续变量的分布是基于高斯分布[5]的。所以我们得到高斯朴素贝叶斯模型[5]：

$$\mathbf{P(x_i = u|y_k)} = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi\sigma_i^2}}}\mathbf{exp}\left(-\frac{\mathbf{u-\mu_{y_k}}}{\mathbf{2\pi\sigma_{y_k}^2}}\right) \tag{4.5}$$

其中y_k表示第 k 类，k=0 或 1；x_i表示样本变量；μ_{y_k}和$\sigma_{y_k}^2$表示y_k类的数学期望与方差。

利用 Python 中的 GaussianNB 建立高斯贝叶斯模型进行对留言详情进行分类，通过 7 个一级标签的类别分类，使用训练集合测试集的数据，从而得到 7 次留言文本分类模型的实验精度是：

表1 7 次留言文本分类实验精度表

分类	精度
城乡建设	0.7910
劳动和社会保障	0.8464
教育文体	0.8742
商贸旅游	0.8189
环境保护	0.9069
卫生计生	0.8575
交通运输	0.8740

平均	0.8527

再进行多次模型精度实验，得到 10 次实验精度折线图：

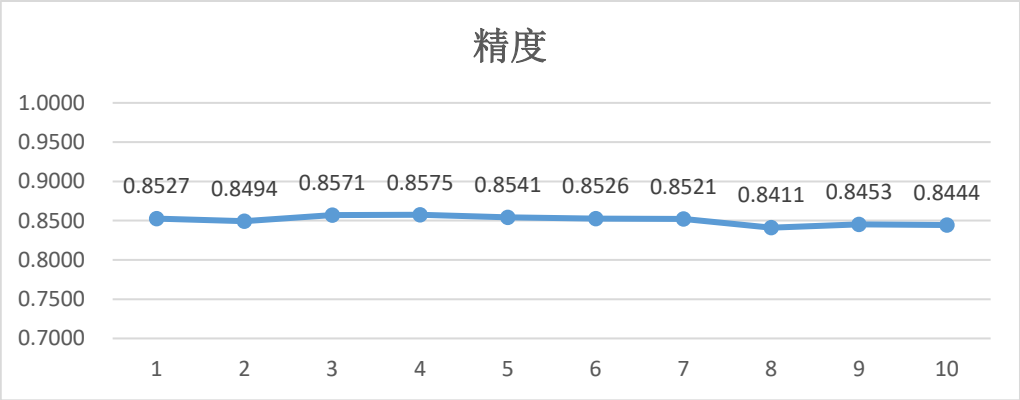


图8　基于 10 次实验后的平均精度变化折线图

可以看到精度基本趋于平稳，说明模型稳定性良好，且精度高，模型的实际意义大，下面对分类进行评价。

4.6 分类方法评价

由于我们将 1 个留言类别的多分类问题转化成 7 个留言二分类问题，那么我们先引入二分类时的 F-Score 评价分类方法[6]。

我们定义，根据附件 2，0 为留言的正样本，1 为留言的负样本。一般使用四个符号表示预测的所有情况[6]：

1. TP(真阳性)：留言的正样本被正确预测为留言的正样本；

2. FP(假阳性)：留言的负样本被错误预测为留言的正样本；

3. TN(真阴性)：留言的负样本被正确预测为留言的负样本；

4. FN(假阴性)：留言的正样本被错误预测为留言的负样本。

在 F-Score 评价分类方法中，有精确率(Precision)和召回率(Recall)的两个重要的概念。设附件 2 有第 i 个一级标签，i=1,2,…,7。并假定第 i 类为 0，其他类为 1(举例：城乡建设为 0，其他为 1)，那么可以得到 7 个 F-Score。

精确率(Precision)是在利用计算机编程预测的正样本中，真实的正样本的百分比，精确率的计算公式为：

$$\mathbf{Precision_i} = \frac{\mathbf{TP_i}}{\mathbf{TP_i + FP_i}} \tag{4.6}$$

召回率(Recall)是真实的正样本中，计算机编程预测的正样本的百分比，召回率的计算公式为：

$$\mathbf{Recall_i} = \frac{\mathbf{TP_i}}{\mathbf{TP_i + FN_i}} \tag{4.7}$$

从而得到 F-Score 的数学模型：

$$\mathbf{F-Score_i} = (\mathbf{1+\beta^2})\frac{\mathbf{Precision_i \times Recall_i}}{\mathbf{\beta^2(Precision_i + Recall_i)}} \tag{4.8}$$

β是衡量精确率(Precision)和召回率(Recall) 在 F-Score 模型中的重要程度，在本文我们取$\beta = 1$，即 Precision 和 Recall 同等重要。

那么就有：

$$\mathbf{F-Score_i} = \frac{\mathbf{2 \times Precision_i \times Recall_i}}{\mathbf{(Precision_i + Recall_i)}} \tag{4.9}$$

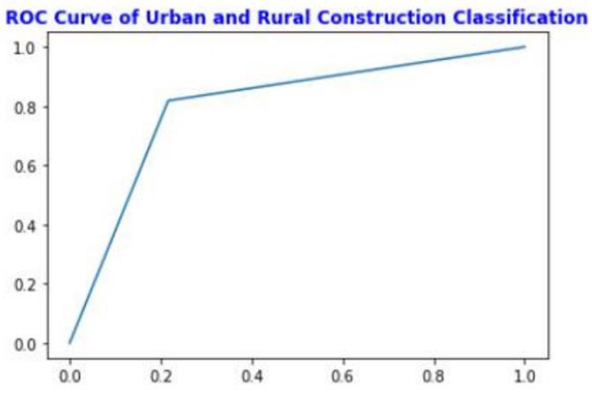
利用 Python 基于高斯朴素贝叶斯模型分类，得到 n 个分类的评价$\mathrm{F-Score}$的实验数值如下：

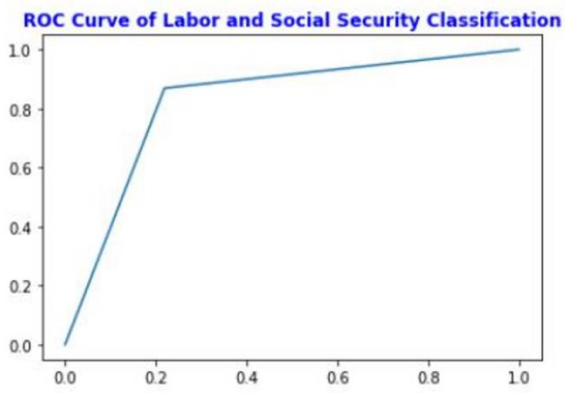
表2 基于朴素贝叶斯模型分类得到的 n 个分类的召回率、准确率、F-Score 数值表

分类	召回率	准确率	F-Score
城乡建设	0.8037	0.8037	0.8037
劳动和社会保障	0.8685	0.8373	0.8526
教育文体	0.8237	0.9375	0.8769
商贸旅游	0.8313	0.8112	0.8211
环境保护	0.8547	0.9444	0.8974
卫生计生	0.8649	0.8649	0.8649
交通运输	0.9270	0.8581	0.8912

可以看到 F-Score 数值基本大于 0.75，分类效果较好，模型评价较好，可以使用此模型。

并利用 Python 制作出每个分类的 ROC 曲线验证：





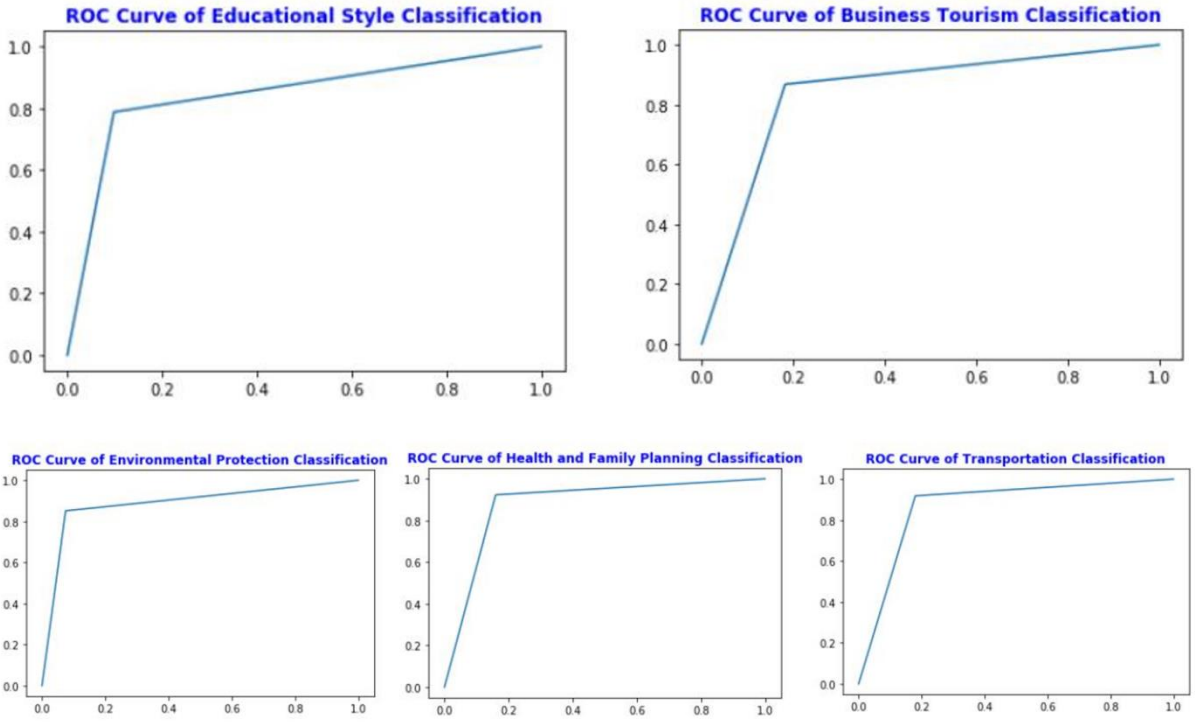


图9　7 个分类的 ROC 曲线图

同时得到 AUC 值：

表3　7 个分类的 AUC 值

分类	AUC
城乡建设	0.7902
劳动和社会保障	0.8459
教育文体	0.8791
商贸旅游	0.8189
环境保护	0.9045
卫生计生	0.8571
交通运输	0.8672
平均	**0.8519**

因为分成 7 次分类，那么得到总$\mathrm{F-Score}$为：

$$\mathbf{F_1} = \frac{1}{n}\sum_{i=1}^{7}\mathbf{F-Score_i} = \frac{1}{n}\sum_{i=1}^{7}\frac{2\times \mathbf{Precision_i}\times \mathbf{Recall_i}}{(\mathbf{Precision_i}+\mathbf{Recall_i})} \tag{4.10}$$

得到考虑了 n 分类的评价效果总$\mathrm{F-Score}$的平均值为$F_1 = 0.8587$。可以看到F_1达到了 85%以上，说明分类效果较好，高斯朴素贝叶斯一级分类模型模型的可行性较高。并再进行 9 次实验(共 10 次)，验证分类模型的效果是否平稳有效，并得到十次F_1和平均

AUC 的折线图：

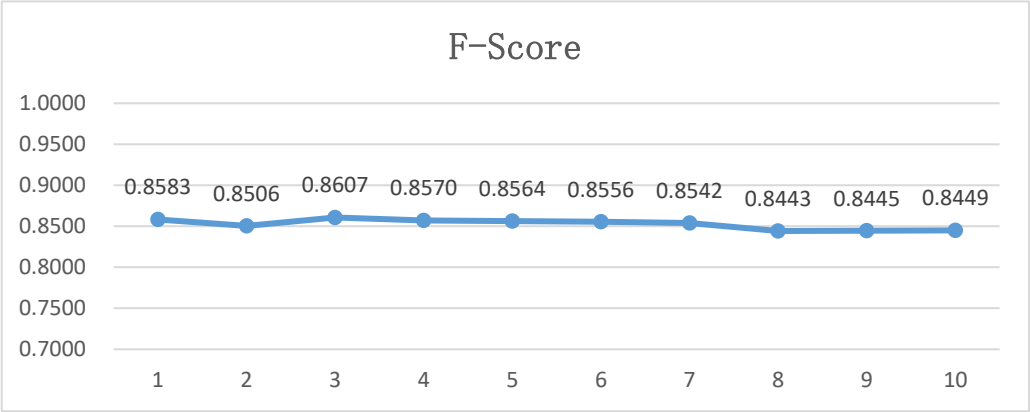


图10　基于 10 次实验后的 F1 变化折线图

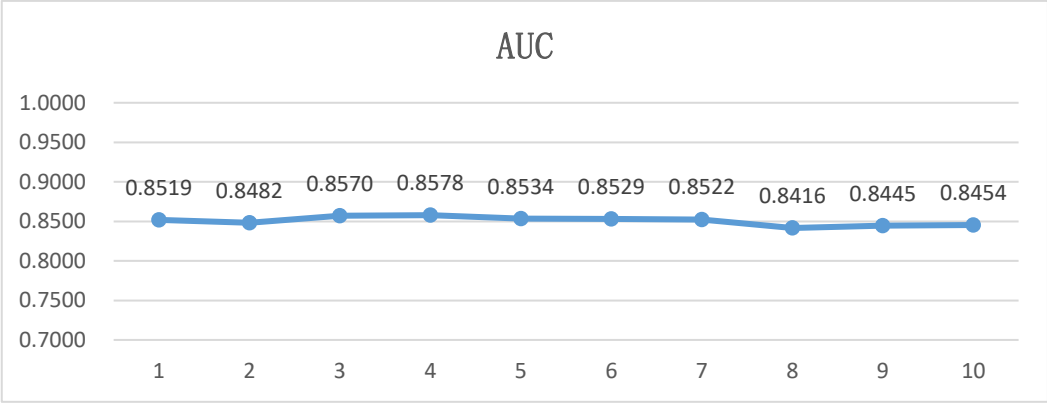


图11　基于 10 次实验后的平均 AUC 值变化折线图

可以看到利用不同的测试集与训练集进行了 10 次实验(详细的实验数据与准确率、召回率的 10 次实验折线图见附件中的“分类模型实验数据.xlsx”)，F-Score 平均值和 AUC 值趋于平稳，并且基本处于 0.84-0.86 之间，误差不超过 2%，并且数值较大，一级分类标签模型分类效果好，高斯朴素贝叶斯分类模型应用性高。由于多次试验结果误差较小，所以做结论时，采用第 1 次实验的数值，即$F_1 = 0.8587$、模型精度=0.8527、AUC 数值=0.8519 等结果数值进行结果研究分析。

五、热点问题挖掘

5.1 问题二的基本流程

此题我们使用附件 3 中的留言主题文本进行分析、归类；同时使用附件 3 留言的数量、时间和反对数、赞成数进行热点问题挖掘。

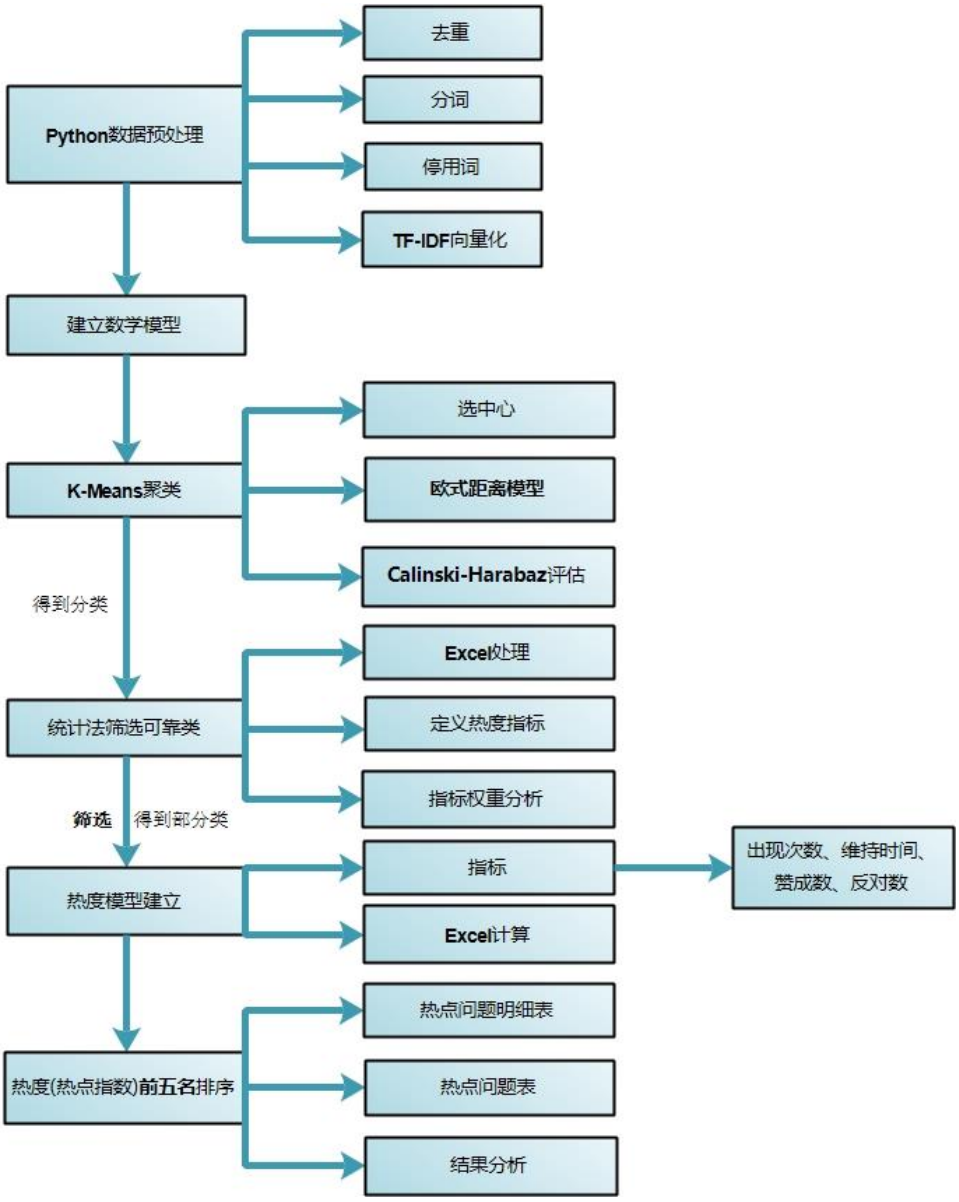


图12　问题二流程图

5.2 K-Means 文本聚类归类

5.2.1 留言主题文本预处理

对于附件 3，利用留言主题的较短的文本进行预处理，为建模奠定数据基础；所以对于下面的文本数据处理，对挑选的附件 3 的留言主题这列数据利用 Python 进行分析处理，即进行文本去重、中文分词、分词去停用词和分词 TF-IDF 向量化。效果展示在上文已经展示，所以本题只给出文本步骤，步骤如下：

（一）主要处理工具：Python 软件、Excel 软件。

（二）将附件 3 中的留言主题进行 Python 文本去重，去除一些重复出现的留言主题

文本数据，尽可能缩小信息量。

（三）调用 Python 的第三方库——jieba 库[2]，对去重后的留言主题进行中文分词。

（四）调用第一问中的停用词表，并对中文分词后的留言主题文本进行去停用词，得到需要向量化的分词向量，并保存成“data.txt”和“adata.txt”。

（五）利用 TF-IDF 向量化模型得到留言主题的 TF-IDF 留言主题向量，使用 Python 使得向量长度相同和维度相同，得到留言主题-文档矩阵。

5.2.2 K-Means 聚类模型

得到 TF-IDF 向量化后的留言主题的词语-文档矩阵，根据每个编号的留言主题的 TF-IDF 词语向量，进行归类，根据一级分类标签，利用 K-Means 聚类模型把留言主题归类成 15 类，根据定义衡量热度的指标进行热度排序。

建立 K-Means 聚类模型[7](算法步骤)

1.简要介绍:K-Means 聚类算法是一种无监督学习,将不知道类别的数据进行归类，主要通过数据与类中心距离的大小进行归类，用他们之间的距离大小衡量数据之间的相似度。

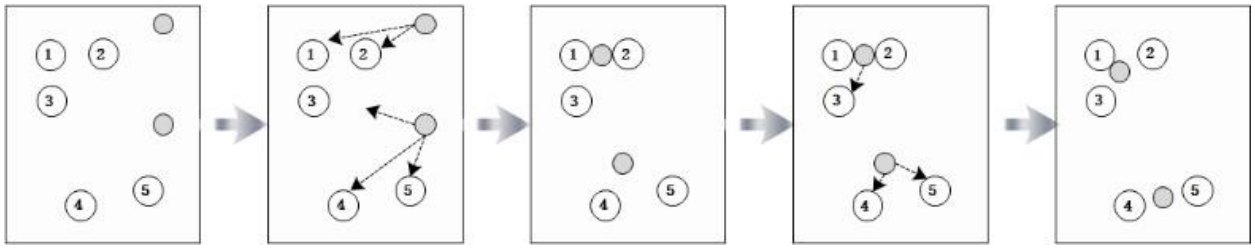


图13 K-Means 聚类模型演示

上图为 k=2(聚类中心数量为 2) 的情况，我们讨论当 k=15，即类中心为一级分类的数量。

2.首先初始化类中心，我们需要提前选择类中心的数量 k，即类别的数量，我们选择 k=15。

3.为了衡量留言主题之间的相似度，利用 TF-IDF 词语向量到聚类中的距离，通过大小来划分类别，将所有留言主题(样本)分成 k 个簇。

相似度(距离)计算公式[7]：

$$\mathbf{dist(i,j)} = \sqrt{\sum_{\mathbf{i=1}} \sum_{\mathbf{j=1}}^{\mathbf{k}} \left(\mathbf{X_i} - \mathbf{center_j}\right)^2} \qquad (5.1)$$

其中X_i为第 i 个留言主题，$center_j$表示第 j 个聚类中心。

4.通过 k 个簇的新样本得到 k 个均值作为新的类中心，计算出类中心的位置。

5.利用 Python 迭代计算直到达到最大迭代次数，满足命令则停止，反之则迭代计算更优的聚类中心点。

即得到满足此式[8]：

$$\mathbf{distbetter} = \arg\min \sum_{j=1}^{k} \sum_{x \in S_j} \left\| X_i - \mathbf{center}_j \right\|^2 \quad (5.2)$$

聚类$\mathbf{center}_j$是类S_j所有样本点的 means(均值)。

6.确定最终的聚类中心点，并得到每个留言主题的类别集合。

K-Means 聚类算法步骤如下：

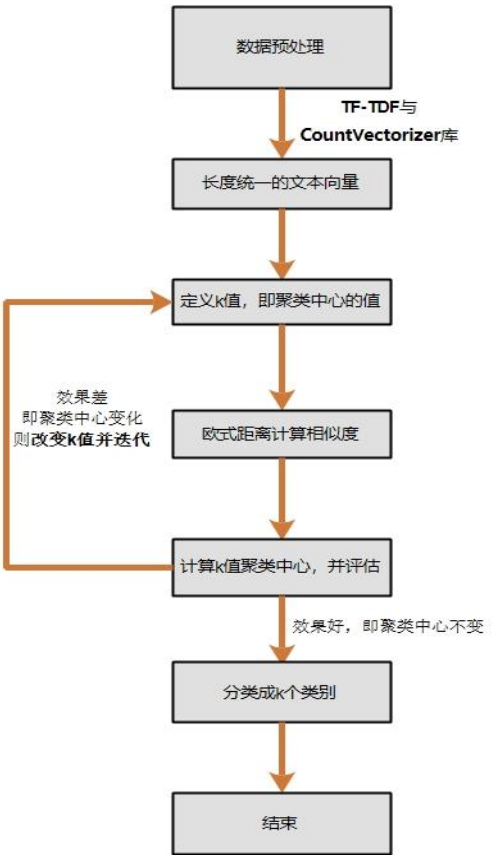


图14　K-Means 聚类算法流程图

利用 Python 进行计算，得到归类并保存在“聚类分析归类.xlsx”中，分为 0、1、2、3、4、5、6、7、8、9、10、11、12、13、14 一共 15 类。下面对归类的 15 个归类进行热点问题分析，并得到热度前五名的热点问题。

引入 Calinski-Harabaz Index 的评价方法[9]，来评价聚类效果 Calinski-Harabaz 的分数越大，那么聚类效果就越好，那么得到 Calinski-Harabaz 评价聚类的分数值的数学模型如下：

$$s(k)=\frac{tr(B_k)m-k}{tr(W_k)k-1} \tag{5.3}$$

在本文m是训练及样本数量，也是留言中主题总数，k是类别数量，B_k是每个聚类类别的协方差阵，W_k 是每个类的样本之间的协方差阵，tr为矩阵的迹。

利用 Python 建立 Calinski-Harabaz Index 的评价方法进行 CH 聚类效果评估，得到$s(15)\approx 13.5$，并继续利用 Excel 在“聚类分析归类.xlsx”中根据留言的主题进行热点问题分析。

5.3 热点问题排序

5.3.1 定义热点问题指标

热点（hot spot）[10]指的是比较受广大群众关注，或者欢迎的新闻或者信息，或指某时期引人注目的地方或问题。通过附件 3 我们得到一些文本数据和时间数据，我们定义 4 个指标作为评价热度的标准：

1.同类留言问题出现的次数。出现的次数可以衡量该问题的重要程度，多次被提出来表示这类问题被影响的范围广、被影响的人群广或是影响程度大，出现次数可以纵向说明热度大小，给予 0.4 评价热度的权重；

2.同类留言维持时间。问题维持的时间可以表明该类问题影响的时间跨度有多大，一定程度上表明该问题是较长存在或是较容易存在，并且需要尽量解决的问题，横向说明热度大小，给予 0.3 评价热度的权重；

3.同类留言问题赞成数。可以一定程度看出人们对此留言问题的赞成程度，在现实生活中，通过微博、微信、阳光热线等途径留言，信息或者留言被点赞是比较常有的事，所以给予较小的 0.1 的权重。

4.同类留言问题反对数。可以一定程度看出人们对此问题的反对程度，根据现实情况，一般不会去给留言或是信息反对数，如果存在反对很可能是一种能引起热度的一些话题，所以给予较大的 0.2 的权重。

5.3.2 热度模型的建立

(1)Excel 类别数据筛选

利用 Excel 对 15 个类进行高级筛选，对可靠的分类(分类正确、分类有效等因素)进行选择，最终得到 7 个类，分别是第 3、4、6、9、12、13、14 类，下面计算这几个类别的热度大小。

用 Excel 计算出每个类别留言问题出现的次数、维持时间、总赞成数和总反对数，

并保存成“聚类最终归类.xlsx”，并得到指标柱状图。

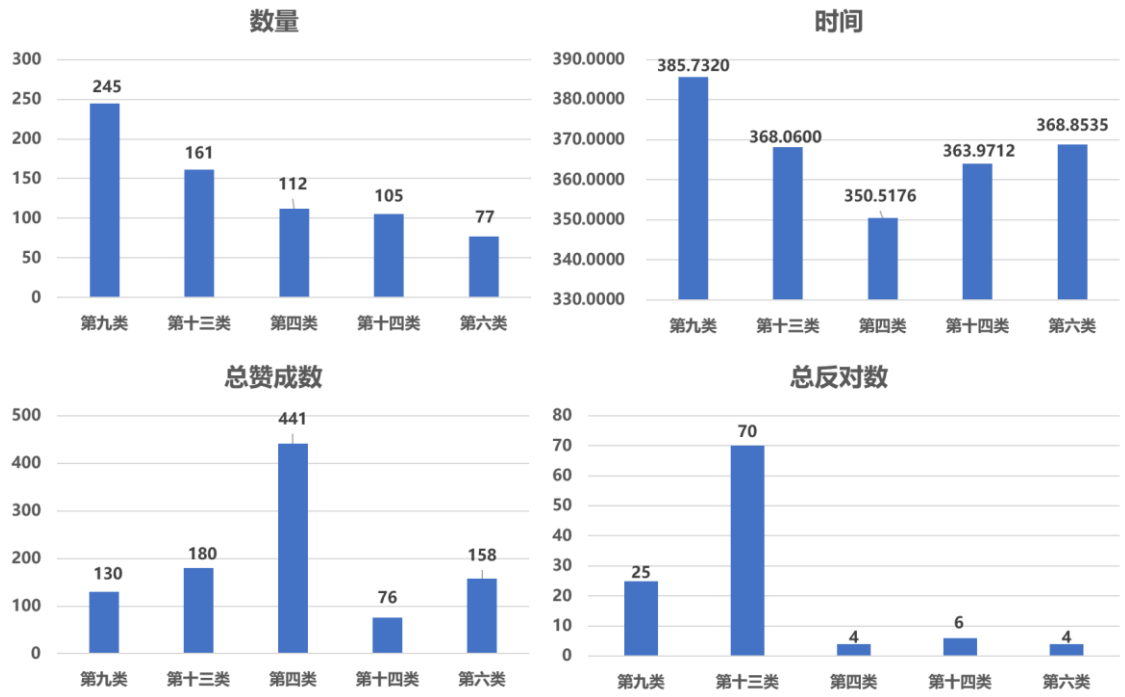


图15　热度前五名的数量、时间、总赞成数、总反对数柱状图

(2) 热度模型

根据给予每个指标的权重，建立热度模型[11]：

$$\textbf{hot spot} = \alpha \times \textbf{number} + \tau \times \textbf{time} + \gamma \times \textbf{approval} + \rho \times \textbf{objection} \quad (5.4)$$

根据权重大小，令$\alpha = 0.4, \tau = 0.3, \gamma = 0.1, \rho = 0.2$，得到：

$$\textbf{hot spot} = 0.4 \times \textbf{number} + 0.3 \times \textbf{time} + 0.1 \times \textbf{approval} + 0.2 \times \textbf{objection} \quad (5.5)$$

其中，hot spot为热度得分，number为同类留言问题出现的次数，time为同类留言维持时间，approval同类留言维持赞成数，objection同类留言维持反对数。

根据热度模型通过 Excel 计算最终得到前五名的热度得分：

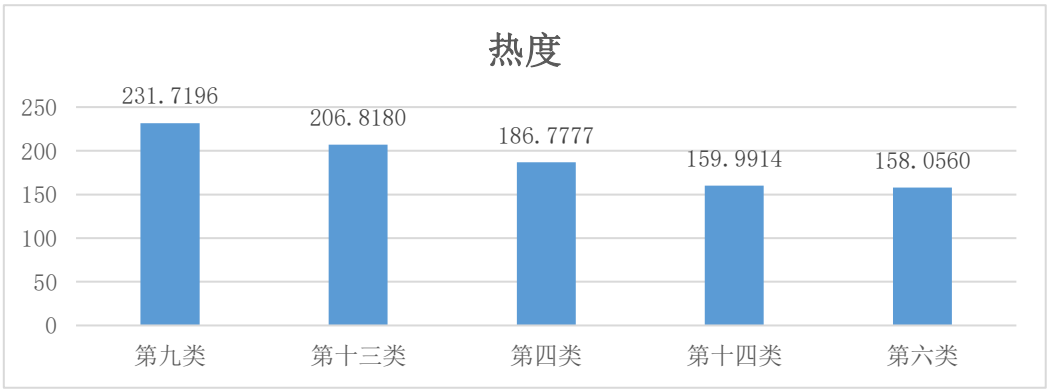


图16　热度排名前五的热度柱状图

表4　热度排名前五的七大指标汇总表

指标 \ 类别	九	十三	四	十四	六
数量/个	245	161	112	105	77
时间/天	385.73	368.06	350.52	363.97	368.85
总赞成数/个	130	180	441	76	158
总反对数/个	25	70	4	6	4
得分	231.7196	206.8180	186.7777	159.9914	158.0560
最初留言时间	2019/1/4 15:33:14	2019/1/4 15:35:56	2019/1/2 20:27:07	2019/1/1 21:10:17	2019/1/2 17:25:52
最后留言时间	2020/1/25 9:07	2020/1/7 17:02:22	2019/12/19 8:52:27	2019/12/31 20:28	2020/1/6 13:54:54

5.4 热点排序结果综合评价

我们得到热点问题前五名的热度得分为231.7196、206.8180、186.7777、159.9914、158.0560，并且问题描述如下所示：

表5　前五名热点排序表

热度排名	热度指数	时间范围	问题描述
1	231.7196	2019/1/4 15:33:14 至 2020/1/25 9:07:21	各地点噪音扰民、噪音污染
2	206.8180	2019/1/4 15:35:56 至 2020/1/7 17:02:22	咨询生活、工作等问题时遇到的问题
3	186.7777	2019/1/2 20:27:07 至 2019/12/19 8:52:27	各地点设施规划的时间和地点等问题
4	159.9914	2019/1/1 21:10:17 至 2019/12/31 20:28:51	各种违规行为，如违规建设、学校违规、小区违规等
5	158.0560	2019/1/2 17:25:52 至 2020/1/6 13:54:54	加快A市的建设、战略、交通等的力度

利用Excel，得到部分热点问题留言明细表如下(详细的热点问题表和热点问题留言明细表保存在作品附件中的“热点问题表.xlsx”和“热点问题留言明细表.xlsx”)。

表6　热度问题留言明细表

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	188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	A市A3区中海国际社区三期四期中间…	0	0
1	188249	A0008	A3区保利麓谷林语桐梓	2019/9/17	保利麓谷林语桐梓	0	0

		4085	坡路与麓松路交汇处地铁凌晨 2 点施工扰民	4:25:00	坡路与麓松路交汇处…		
1	188399	A0009 7934	A 市利保壹号公馆项目夜间噪声扰民	2019/7/3 6:23:25	您好，我想举报 A 市利保壹号公馆项目…	0	0
	…	…	…	…	…	…	…
1	360109	A0080 252	万科魅力之城小区底层门店深夜经营,各种噪音扰民	2019-09-04 21:00:18	您好：我是万科魅力之城小区的业主…	0	0
2	188007	A0007 4795	咨询 A6 区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	A 市 A6 区道路命名规划已经初步成果…	1	0
2	188455	A0003 5902	咨询异地办理出国签证的问题	2019/5/16 15:20:43	书记您好!我是外地人,2018 年毕业后…	0	0
2	188876	A0001 3435	咨询 A7 县榔梨龙华安置区外围马路修复问题	2019/2/26 11:29:49	尊敬的领导您好，榔梨龙华安置区…	0	0
…	…	…	…	…	…	…	…
2	342119	A0909 00	咨询移动通信业务问题	2019-10-18 23:52:58	尊敬的领导：你好，本人前期向西地省…	0	0
3	189739	A00051 608	请问 A3 区西湖街道茶场村五组是如何规划的	2019/9/12 8:30:47	请问领导，政府对于 A3 区…	0	0
3	190033	A00010 6961	A7 县泉塘泉星社区力都大厦旁新规划的菜市场什么时候启用？	2019/11/3 20:19:18	尊敬的领导，关于泉塘街道…	5	0
3	190717	A00011 2982	请问 A2 区西牌楼小区有无拆迁计划	2019/8/27 7:05:15	请问 A2 区西牌楼小区…	0	0
…	…	…	…	…	…	…	…
4	188119	A0003 5029	对 A 市地铁违规用工问题的质疑	2019/5/27 16:04:44	我是一名在 A 市某地铁站上班的安检员…	0	0
4	190377	A0001 06749	A 市 A3 区兰亭湾畔小区违规开餐饮	2019/8/18 11:39:56	A3 区兰亭湾畔小区违规开餐饮…	0	0
4	190408	A0007 467	F 市建工违规施工使我们无家可归还我房子	2019/4/6 14:36:11	2019.3.18F 市建工在我们房屋周边…	0	0
…	…	…	…	…	…	…	…
4	289606	A0009 3113	A 市楚郡未来实验学校针对四年级学生上调学	2019/7/16 22:49:58	各位领导：之前反应	0	0

			费是否违规？		的关于楚郡…		
5	190213	A0003 1618	请 A 市加快自来水深度净化改造力度	2019/1/16 12:53:26	地处时代倾城小区的自来水烧过开水后…	0	0
5	191872	A0003 1618	请 A 市加快轨道交通建设力度	2019/3/1 15:19:28	地处中部中心城市的 A 市在高铁和地铁…	9	2
5	193514	A0003 1618	请加快 A 市月亮岛片区公共服务力度	2019/3/20 16:39:22	地处月亮岛片区近年人口迅猛增长…	4	0
…	…	…	…	…	…	…	…
5	289729	A0003 1618	A 市加快国家中心城市建设刻不容缓	2019/4/30 15:35:00	地处中部大省 A 市常年人口流失严重…	3	0

六、答复意见主题评价

6.1 问题三的基本流程

问题三需要从答复的相关性、完整性、可解释性和重要性四个角度对附件 4 中的答复意见的质量给出一套评价方案。下面我们对评价方案进行建立模型并实现，步骤流程图如下：

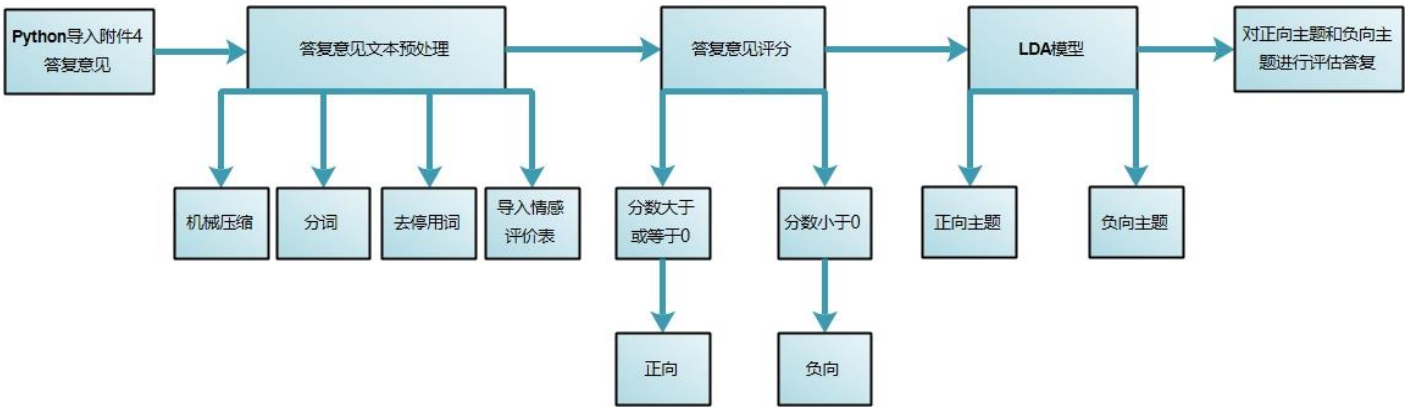


图17 问题三流程图

下面对答复的相关性、完整性、可解释性和重要性进行一定程度的定义：

1. 答复的相关性和完整性：对答复内容进行情感分析；

2. 答复的重要性：答复内容的去重和机械压缩；

3.答复的可解释性：答复内容 jieba 库[2]中文分词。

6.2 文本数据筛选与处理

（一）同前两题一致，将附件 4 中的答复内容进行 Python 文本去重，因为需要保证每一条答复的内容的重要程度尽可能接近，不会在向量化时，分词的权重变小，所以需要去除一些重复出现的答复内容；并建立自定义函数进行机械压缩(如“好啊好啊好啊”压缩成“好啊”)，进行机械压缩后字符串长度从 1015450 变为 1000783。去重和机械压缩，保证每条答复文本的重要性。

（二）调用 Python 的第三方库——jieba 库[1],对去重后的答复内容进行中文分词，保证每条答复内容数据的可解释性。调用前两问中的停用词表，并对中文分词后的答复内文本进行去停用词，得到需要向量化的分词向量。

（三）导入情感评价表[12](详情见附件 txt 表），保证每条答复内容数据的完整性和相关性。情感评价表是通过经验和相关题目得到的带有情感的词语信息和各个词语的权重的表格，利用 Python 建立评分函数并封装，得到答复内容的评分(权重)，并将答复内容词语划分成正向和负向，如果分数大于 0 则为正向词语，并给出正向评分(权重)，保存成 “pos.txt”；如果分数小于 0 则为负向词语，负向主题在本文的定义：问题较严重且较难解决的词语主题，给出负向评分(权重)并保存成 “neg.txt”。

6.3 建立 LDA 主体模型

LDA[13]（Latent Dirichlet Allocation，隐含狄利克雷分布）是文档主题生成模型，也称为三层贝叶斯贝叶斯概率模型，包含文档(d)、主题（z）、和词(w)三层结构，能对附件 4 的答复内容文本进行建模，通过 LDA 主体模型能够挖掘答复内容数据集的潜在主题，从而集中分析数据集中的关键特征词。LDA 是一种无监督机器学习技术，它采用了词袋模型（bag of words，简单记为 BOW）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。

下面引入生成模型的概念：每一篇文章的每一个词语都通过“以一定程度的概率选择了某个主题，并且从这个主题以一定程度的概率选择了某一个词语”这就是 LDA 模型中主题和文档的关系。而 LDA 模型简单来说就是：根据给定文档，推出文档的主题分布。

LDA 模型文档的生成[14]如下：

1.生成一篇文档时，从狄利克雷分布中取样生成文档 i 的主题分布θ_i；

2.从主题的多项式分布θ_i中取样生成文档 i 第 j 个词的主题$z_{i,j}$；

3.从狄利克雷分布中取样生成主题中取样生成主题$z_{i,j}$对应的词语分布对应的词语

分布$\emptyset_{z_{i,j}}$。

4.从词语的多项式分布$\emptyset_{z_{i,j}}$中采样最终生成词语$\omega_{i,j}$

隐含狄利克雷分布，即 LDA 模型的实现[14]如图所示:

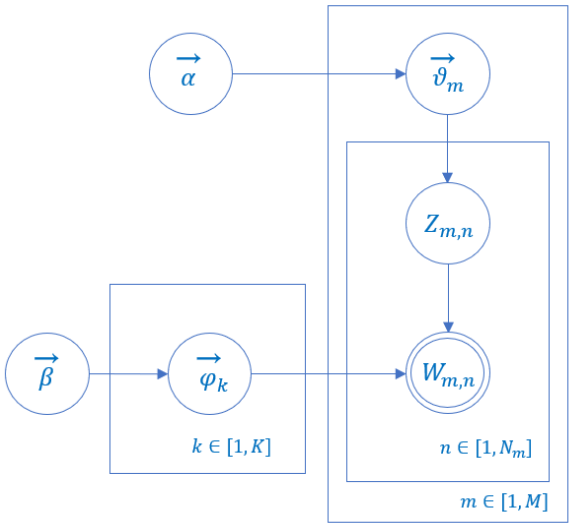


图18 LDA 模型实现演示

为了实现 LDA 主体模型，我们使用 Python 的 Gensim 库进行主题建模和主题分析，下面进行介绍 Python 的免费库之一——Gensim。

1.能够从文档中自动抽取语义主题；

2. Gensim 库[15]的核心概念：

A. 语料库(corpus)：文档的集合，也可以称为训练语料，可以推出文档主题；

B. 向量(vector):将一个文本或文档表示成一个特征数组(列表)；

C. 稀疏矩阵(Sparse vector)： 将答复-文档矩阵(A Sparse vector)的 0 值省略，以节省空间(因为空间较大，数据多)；

D. 模型(model):定义两个向量空间的变换，即从文本的一种向量表达变换成另一种。

文档中的某一个词语出现的概率有如下公式可以获得：

$$\mathbf{p}(\text{词语}|\text{文档}) = \sum_{\text{主题}} \mathbf{p}(\text{词语}|\text{主题}) \times \mathbf{p}(\text{主题}|\text{文档}) \qquad (6.1)$$

用简单的图像表示就是：

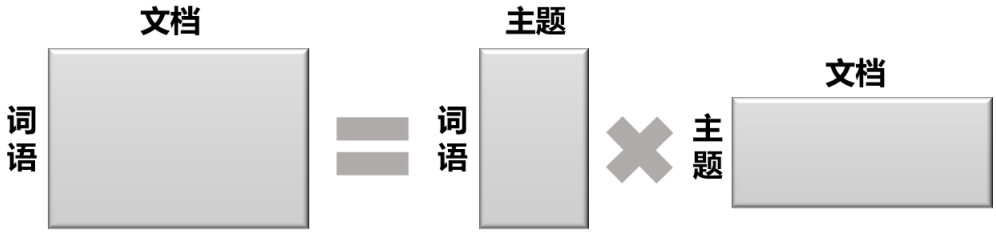


图19　文档中词语出现概论公式演示

利用 Python 进行对答复内容的 LDA 主题分析，设主题为 3 个，得到 3 个主题中每个特征词语的权重，部分结构如图所示：


```
[(0,
  '0.002*"[\'留言\', \'收悉\', \'转\', \'相关\', \'部门\', \'敬请\', \'关注\', \'后续\', \'回复\', \'谢谢\']" + 0.001*"[\'留言\', \'收悉
\', \'已转\', \'P9\', \'市\', \'调查\']" + 0.001*"[\'您好\', \'转交\', \'相关\', \'单位\', \'调查\', \'处置\']" + 0.001*"[\'留言\', \'收悉
\', \'转市\', \'教育\', \'体育局\', \'调查\']" + 0.001*"[\'网友\', \'提出\', \'J11\', \'市\', \'灵活\', \'就业\', \'社保\', \'补贴\', \'补
贴\', \'年\', \'补贴\', \'年\', \'收悉\', \'现\', \'答复\', \'感谢您\', \'关心\', \'关注\', \'我局\', \'工作\', \'西地省\', \'就业\', \'专
项资金\', \'管理\', \'办法\', \'楚财社\', \'2018\', \'25\', \'号\', \'文件\', \'第九条\', \'就业\', \'困难\', \'人员\', \'灵活\', \'就业
\', \'缴纳\', \'社会保险费\', \'给予\', \'数额\', \'社会保险\', \'补贴\', \'补贴\', \'标准\', \'4050\', \'女满\', \'40\', \'周岁\', \'男满
\', \'50\', \'周岁\', \'人员\', \'缴费\', \'60%\', \'非\', \'4050\', \'人员\', \'缴费\', \'40%\', \'就业\', \'困难\', \'人员\', \'社会保险
\', \'补贴\', \'期限\', \'法定\', \'退休年龄\', \'年\', \'就业\', \'困难\', \'人员\', \'可延长\', \'退休\', \'外\', \'人员\', \'最长\',
\'超过\', \'年\', \'初次\', \'核定\', \'享受\', \'社会保险\', \'补贴\', \'时\', \'年龄\', \'为准\', \'补贴\', \'方式\', \'先缴\', \'补\',
\'郴人\', \'社发\', \'2018\', \'83\', \'号\', \'文件精神\', \'2019\', \'年\', \'灵活\', \'就业\', \'人员\', \'社会保险\', \'补贴\', \'享受
\', \'对象\', \'享受\', \'养老保险\', \'补贴\', \'未满\', \'年限\', \'灵活\', \'就业\', \'困难\', \'人员\', \'全省\', \'年度\', \'岗\',
\'职工\', \'月\', \'平均工资\', \'100%\', \'60%\', \'缴费\', \'基数\', \'缴纳\', \'养老保险费\', \'享受\', \'养老保险\', \'补贴\', \'4050
\', \'人员\', \'补贴\', \'全省\', \'年度\', \'岗\', \'职工\', \'月\', \'平均工资\', \'60%\', \'缴费\', \'基数\', \'缴纳\', \'养老保险费\',
\'60%\', \'非\', \'4050\', \'人员\', \'补贴\', \'全省\', \'年度\', \'岗\', \'职工\', \'月\', \'平均工资\', \'60%\', \'缴费\', \'基数\',
\'缴纳\', \'养老保险费\', \'40%\', \'补贴\', \'方式\', \'仍为\', \'先缴\', \'补\', \'感谢您\', \'我局\', \'工作\', \'关心\', \'关注\', \'2
018\', \'年\', \'11\', \'月\', \'27\', \'日\']" + 0.001*"[\'网友\', \'UU0082316\', \'\\u3000\', \'\\u3000\', \'您好\', \'来信\', \'收悉\',
\'现\', \'回复\', \'\\u3000\', \'\\u3000\', \'我局\', \'建议\', \'如实\', \'上报\', \'上级\', \'交通\', \'主管部门\', \'出台\', \'相关\',
```


图20　部分正向词语结构图

表7　部分权重较大的正向主题表

主题一	主题二	主题三
相关	收悉	相关
收悉	调查	关注
后续	教育	已转
部门	法院	规划
谢谢	城乡规划	立案查处
就业	体育局	规范性
社保	严肃处理	行政许可
补贴	高度重视	转交
养老保险	签订协议	解决

处理程序	采取措施	协议

根据对留言的答复内容进行特征提取 3 个正向主题，正向的定义：反映答复的是否重视留言中的各种问题和问题的解决方法，我们得到了上表中 3 个主题中，部分权重较大的正向词语(全部词语保存成“正向主题.txt”以及在附件 Python 代码输出)，下面逐个主题分析：

主题 1 中的高频出现即高权重的特征词有：收悉、处理程序、就业、后续、养老程序、谢谢和社保等，“收悉”、“处理程序”和“后续”主要反映了答复内容已经收到答复，表明留言的问题已经收到，后续处理会加快；“谢谢”表明答复内容的态度是很好的，有答复的结尾，保证答复内容的完整性；“就业”、“养老程序”和“就业”表示了该主要偏向的问题是民生问题，同时弥补了答复的相关性。

主题 2 中的高频出现的特征词有：高度重视、收悉、签订协议、采取措施、严肃处理、城乡规划和教育等词语，“高度重视”、“签订协议”、“采取措施”、“严肃处理”表示答复内容对留言内容会进行处理，保证答复的可解释性；“城乡规划”和“教育”表示了该主要偏向的问题是社会问题，同时保证了答复的相关性。正向的定义：反映答复的是否重视留言中的各种问题和问题的解决方法

主题 3 中的高频出现的特征词有：立案查处、相关、协议、关注、规范性、行政许可和解决等词语，可以看到第三个主题主要反映了答复内容中对问题的做法和解决方法，保证了答复内容对问题的重要性。

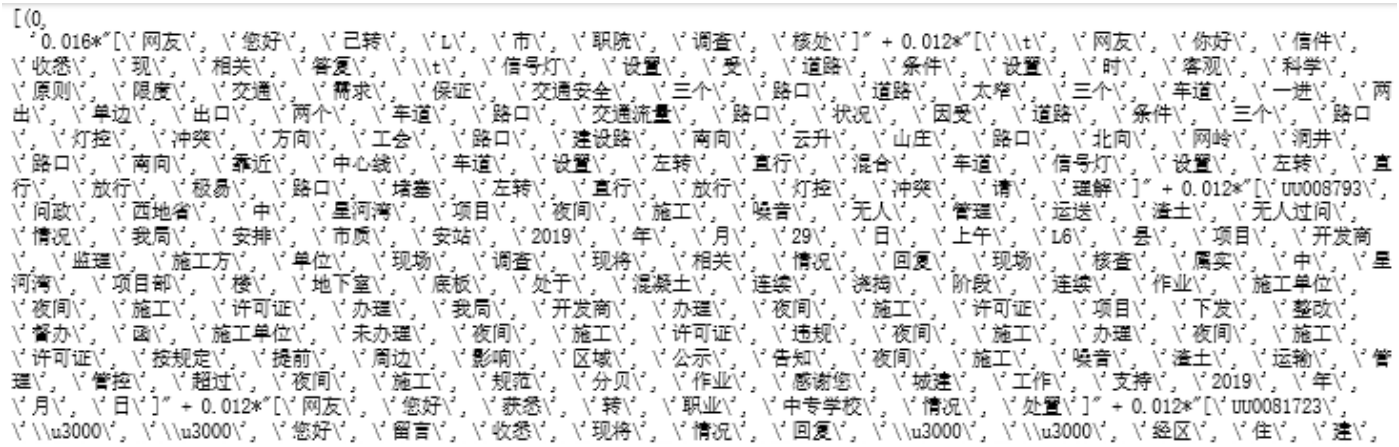


图21　部分负向词语结构图

表8　部分权重较大的负向主题表

主题一	主题二	主题三
交通安全	整治	施业单位
车道	飙车	秩序

噪音	物业	交通事故
教育	调查	物业
留校察看	尽量减少	依法

同样根据对留言的答复内容进行特征提取 3 个负向主题，负向主题在本文的定义是问题较严重且较难解决的词语主题，我们得到了上表中 3 个主题中，部分权重较大的负向词语(全部词语保存成“负向主题.txt”以及在附件 Python 代码输出)，下面逐个主题分析：

主题 1 中的高频出现即高权重的特征词有：车道、教育、噪音、交通安全、留校察看等，可以看到主题 1 反映了答复内容中比较严重的问题，

主题 2 中的高频出现的特征词有：整治、飙车、尽量减少、物业、调查等词语，“整治”、“尽量减少”、“调查”、“严肃处理”表示答复内容对负向的问题会进行处理改进、整治并且减少，保证答复的可解释性；“飙车”和“物业”表示了该主要偏向的问题，保证了答复的相关性。

主题 3 中的高频出现的特征词有：依法、交通事故、秩序等词语，可以看到第三个主题有一个词语是依法，说明答复中也会采取一些方法进行问题的解决，侧面表明了答复人对留言问题的关注，会尽量解决。

LDA 模型评估答复是：既能反映答复的是否重视留言中的各种问题和问题的解决方法(正向和负向)，也可以得到比较严重的问题所在(负向)。可以挖掘出留言的情感，并评估。对答复内容的评估有合理性。

七、总结

7.1 优点与缺点

表9 优缺点分析表

优点	1.F-Score 值较高，分类效果好，并多次实验检测评估； 2.Calinski-Harabaz Index 高，聚类效果好； 3.考虑多个影响热度的指标进行挖掘热点问题； 4.利用情感评论表情感分析，赋予了答复内容情感评分； 5.LDA 模型挖掘了答复内容的主题，从正反面、正负方向对答复内容进行主题分析和模型评估。

缺点	1.挖掘过程多步进行，较不灵活； 2.LDA 模型主题数量取值比较主观； 3.代码运行效率不高。

7.2 结论

问题一中，对留言内容进行了文本预处理，建立了高斯朴素贝叶斯分类模型对留言内容进行 7 次二分类，构建训练集和测试集进行实验，得到分类模型七次的平均精度为 0.8527，平均 F-Score 为 0.8587，并作图得到实验数据的 ROC 曲线，得到平均 AUC 值为 0.8519，三个通过试验后的数值均比较大，都大于 85%，说明分类模型可靠，再进行多次分类评估实验，得到多个评价指标，发现平均 F-Score 的 F1 指数稳定性高，说明建立的一级分类标签模型合适。

问题二中，同样对留言主题处理后，使用 K-Means 聚类算法进行归类，通过 Calinski-Harabaz 检验，得到 CH 指数约为 13，CH 数值越高，聚类效果越好。再利用 Excel 筛选归类，定义 4 个影响热度指标，建立热度模型，算出热度的得分，最终得到热点问题前五名(简略)分别是：

1.城市(或地区)的噪音问题；

2.咨询类问题；

3.规划问题；

4.违规问题；

5.加快地区各领域推动力度的问题。

问题三中，对答复内容进行构造评估模型，建立 LDA 主题模型，并对答复内容情感分析，得到较严重问题答复和解决方法(负向)和民生、社会类等问题答复和解决方法(正向)的主题词，最终通过 LDA 主题分析模型，从概率和权重的角度，评估得到答复内容较好，保证了答复的完整性、可解释性、相关性和重要性四个角度，并有较大权重提到问题的解决方案。

八、参考文献

[1] 董付国，Python 程序设计基础（第 2 版），北京：清华大学出版社，2018.

[2] https://blog.csdn.net/qq_44455289/article/details/104117576

[3] https://blog.csdn.net/zhaomengszu/article/details/81452907

[4] https://blog.csdn.net/fisherming/article/details/79509025

[5] 阿曼. 朴素贝叶斯分类算法的研究与应用[D].大连理工大学,2014.

[6] https://www.jianshu.com/p/f7ea71f2344f

[7] https://blog.csdn.net/PyRookie/article/details/81915078

[8] https://www.cnblogs.com/baiboy/p/pybnc6.html

[9] https://blog.csdn.net/u010159842/article/details/78624135

[10] https://m.baidu.com/sf_bk/item/%E7%83%AD%E7%82%B9/3388605?ms=1&rid=11333372823761842868

[11] https://m.doc88.com/p-9962760955356.html

[12] 雷小惠. 基于情感分析的评论挖掘技术研究[D].东南大学,2017.

[13] https://m.baidu.com/sf_bk/item/LDA/13489644?ms=1&rid=11448355206811247737

[14] https://www.cnblogs.com/mantch/p/11259347.html

[15] https://blog.csdn.net/u013230189/article/details/82753499