所选题目：C：“智慧政务”中的文本挖掘应用

综合评定成绩：_____

评委评语：

评委签名：

“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。

针对问题一，群众留言分类问题，我们这次会使用 LSTM（Long Short-Term Memory），也就是长短期记忆模型来进行分类。具体步骤如下：首先通过数据清洗得到无空值的数据并统计各一级标签下的数据量，用 jieba 中文分词工具来把留言详情中的文本作合理划分，通过调用 Keras 工具，将前面分词后的留言信息数据进行向量化处理，然后建立关于留言内容的一级标签分类模型，并绘制损失函数的趋势图和准确率趋势图。最后通过画混淆矩阵热力图和计算 F-Score 对分类方法进行评价。

针对问题二，热点问题挖掘问题，首先排除掉无用的如用户编号信息，提取出留言的文本信息作为一个文档。再使用 jieba 专业工具进行分词、删掉无意义的停用词，对文本进行提取关键字操作，生成一个文档。最后使用 k 均值（k-Means）聚类，通过计算轮廓系数找到最优的 K 值为 10，挖掘出整个文档的前十类热点问题，以及每个问题中的十个关键词，包括地点、时间、对象等，并以点赞数、相关主题出现次数为指标，可以计算热度指数。

针对问题三，答复意见的评价问题，我们结合了情感词典对抽取的关键词通过句子算法建立句子倾向和程度向量，还要使用支持向量机（SVM）和 word2vec 划分句子分类，再通过标准化、降维等操作进行精确性、完整性、情感分析。

关键词：文档分类　混淆矩阵　　K-means 聚类　　情感分析

Text Mining Application of “智慧政务”

Abstract: In recent years, with online questioning platforms such as WeChat, Weibo, Mayor’s Mailbox, and Sunshine Hotline have gradually become an important channel for the government to see public opinion, gather citizens’ wisdom, and collect people’s popularity. The amount of text data related to various social conditions and public opinions has continued to rise,bringing great challenges to the relevant departments, which mainly rely on manpower for message division and hot-spot sorting . At the same time, with the development of big data, cloud computing, AI and other technologies, the establishment of a smart government system based on natural language processing technology has become a new trend of social administrate innovation and development, which has a great impact on improving the management level and efficiency of government.

Aiming at the problem of the first,we will apply LSTM model to classify.First is washing the data and counting the amount of data under each primary label,then use jieba--Chinese word segmentation tool--to divide words in message detail region.Later,launch a LSTM model and transform previous data which has been divided into vector through Keras.Then establish a first-level label classification model about the content of the messages and paint trend line of loss function as well as accuracy.Lastly give evaluation to this classification by drawing Confusion Matrix and count F-Score.

Aiming at the problem of the second,first eliminate useless information such as user ID and extract the text information of the message as a document.Then use jieba to do segmentation and remove stop words as keywords from text to generate a document similarly.Finally apply k-means to find top 10 hottest problems include ten keywords of each,such as location,time,object,etc.This procedure is benefit to later steps.And we can consider likes and dislikes,frequency of related topics as index to calculate hot index.

Aiming at the problem of the third,do word segmentation again,and combine emotion dictionary with the keywords extracted to establish sentence propensity and degree vector by sentence algorithm.Besides,we should use SVM and word2vec to divide the classification of sentences.Then analyse accuracy ， holistic and emotion through standardize and dimensionality reduction.

Key words: word segmentation,Confusion Matrix,F-score,K-means clustering,emotion analysis

目 录

1. 挖掘目标

本次建模的目标是根据网上收集的群众问政留言记录，以及相关管理人员给予的答复，采用数据挖掘技术，进行数据清洗、数据探索、LSTM 模型、利用 jieba 分词、k-means 聚类算法达到以下三个目标：

(1) 根据附件 2 给出的一、二、三级分类标准，建立关于留言内容的一级标签分类模型，使用 F-Score 对分类方法进行评价，对内容进行分类以便于后续对问题的处理。

(2) 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聚类，定义合理的热度评价指标，给予不同指标相应权重，并给出评价结果，发现高热度问题有利于反映群众最迫切的需求，便于相关措施的开展。

(3) 针对附件 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。

2. 总体流程

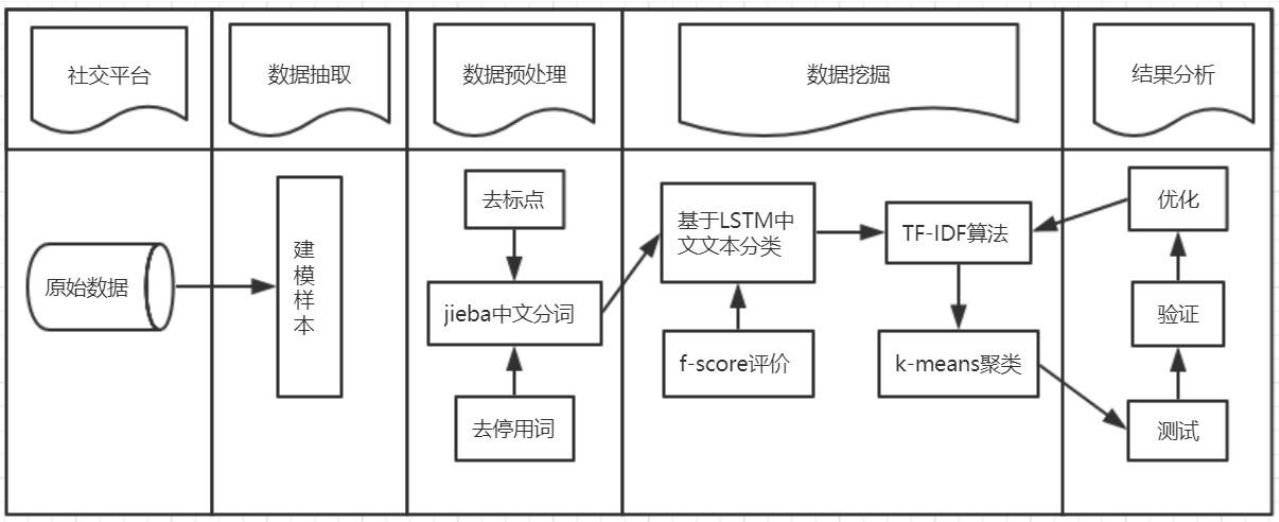

图 2-1 框架流程图

3. 问题 1

3.1. 分析方法与过程

3.1.1. 数据预处理

1. 数据第一步处理

题目给出的数据中，一共 9210 条，通过调用 pandas 函数对给出的数据进行筛查，查看各列是否有空值，有几个空值。如果有空值，对空值数据进行清洗。并根据一级标签统计一下各个类别的数据量，可以可视化的方式查看类别分布。然后将一级标签类转换成 id 号（0,1,2,3…），便于之后训练分类模型。为了避免留言详情中的标点符号和停用词增加计算的复杂度和系统内耗，我们需对留言详情信息进行预处理工作,包括删除前面所提到的文本中的标点符号、特殊符号、一些无意义的停用词（stopword），完成初步数据处理工作。

2. 对留言详情信息进行中文分词

在对留言信息数据进行挖掘分析前，先要将离散的文本信息转换为计算机可以辨别的结构化信息。附件 2 中，用中文文本的方式给出数据，为了便于转换，首先用 jieba 进行分词，jieba 分词是 Python 的一个中文分词组件，具有文本分类的常用功能，比如分词、词性标注、关键词抽取等功能，而且可以导入自定义词典。jieba 分词中，首先通过对照典生成句子的有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取。对于未登陆词（不在词典中的词）使用 HMM 进行新词发现。在此运用 jieba 分词同时，过滤掉停用词（中文停用词包含了一些日常使用频率很高的常用词，例如：吧，吗，呢，啥等一些感叹词等,这些高频常用词对文本的主要意思没有影响,所以要被过滤掉。）

3.1.2. LSTM 建模

1. 调用 Keras 模型

调用 Keras 工具，将前面分词后的留言信息数据进行向量化处理，将逐条数据转化为一个全为整数的向量。

步骤：

1) 设置最频繁使用的 10000 个词
2) 设置每条数据最大词语数为 250 个，溢出就截去，缺漏就补 0。
3) 设置 Embeddingceng 层的维度
4) 创建分词器 Tokenizer 对象
5) 填充 X,让 X 的各个列的长度统一

6) 多类标签的 onehot 展开

2. 拆分训练集和测试集

用 sklearn 将处理好的数据划分为训练(train)部分和测试(test)部分，为后面做准备。

训练集（Training set）的作用是拟合模型，提前调参，训练分类模型。在后面结合测试集检验效果时，会选出同一参数的不同取值，拟合出多个分类器，测试集是通过训练集得出最优模型后，使用测试集进行模型预测得到的。用来权衡该最优模型的性能和分类能力。即可以把测试集当做未存在的数据集，当确定好模型参数后，再使用测试集来对模型性能进行评估。

运用留出法（hold-out）将数据集划分为两个相互排斥而不相交的集合，其中一个为训练集 S，另一集合为测试集 T，在 S 上训练出模型后，用 T 来评估其测量误差，作为泛化误差预计。S/T 的划分要尽量保证数据分布的一致性，这样才能防止因为数据划分过程引入意外误差而对最终成果造成影响。常见做法是把大概 2/3~4/5 的样本作为 S,其余为 T。

3. 定义 LSTM 序列模型

通过 LSTM 模型，建立关于留言内容的一级标签分类模型。

定义 LSTM 模型步骤：

1) 模型的第一层叫作嵌入层（Embedding），使用长度为 100 的向量来表示一个词语；

2) SpatialDropout1D 层在训练中每次更新时，要将输入进单元的依比例随机设置为 0，可以避免过拟合；

3) LSTM 层包含 100 个记忆单元；

4) 输出层为包含 7 个分类的全连接层；

5) 由于是多分类，所以激活函数设置为’softmax’；

6) 由于是多分类，所以损失函数为分类交叉熵（categorical_crossentropy），当应用 categorical_crossentropy 损失函数时，标签为多类模式，比如若有 10 个类别，每一个样本的标签应该是一个 10 维的向量，该向量在对应有值的索引位置为 1 其余位置为 0；

定义好 LSTM 模型后，就要准备训练数据，其中设置 5 个训练周期且 batch_size

为 64。

4. 生成损失函数趋势图和准确趋势图

损失函数能优秀地反映模型与实际数据差距的工具，能够更好地对后面优化工具（梯度下降等）进行剖析和研究。生成损失函数趋向图和准确趋向图，直观地看到测试集和训练集随着训练周期损失和准确率的趋势。

3.1.3. LSTM 模型的评估

1. 生成混淆矩阵

混淆矩阵(confusion matrix)也叫作误差矩阵，是体现精度评价的一种标准的格式，它是用 n 行 n 列的矩阵。详细评价指标包括总体精度、制图精度、用户精度等，这些精度指标从不同的方面反映了图像分类的精准性。 在人工智能中，混淆矩阵是可视化编程的工具，特别用于监督学习，在无监督学习一般叫做匹配矩阵。在图像精度评价中，主要用于比拟分类结果和实际测量值，能把分类结果的精度显示在一个混淆矩阵中。混淆矩阵是通过将每个实测像元的坐标和分类与分类图像中的相应位置和分类相比较计算的。混淆矩阵的每一列代表了预测类别 ，列的总数表示预测为该类别的数据的数量，并且每一列中的数值表示真实数据被预测为该类的数量;每一行代表数据的真实归属类别，行的数据总量表示该类别的数据实例的数量。

混淆矩阵的主对角线上表示的是预测对的个数,除了主对角线之外其余都是预测有误的个数。

2. 计算 F1-score

F1 分数（F1 Score），简单来说是统计学中度量二分类模型的精准度的一种指标，它不仅可以计算出分类模型的精确率（precision）也能计算召回率（recall），可以看作是二者的一种调和平均，F1 分数的最大值是 1，最小值是 0。

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

其中 Pi 是第 i 类的准确率，Ri 是第 i 类召回率，调和平均。

3.2. 结果分析

3.2.1. 预处理结果

表 3-1 分类统计表

一级标签	数目
城乡建设	2009
劳动和社会保障	1969
教育文体	1589
商贸旅游	1215
环境保护	938
卫计卫生	877
交通运输	613

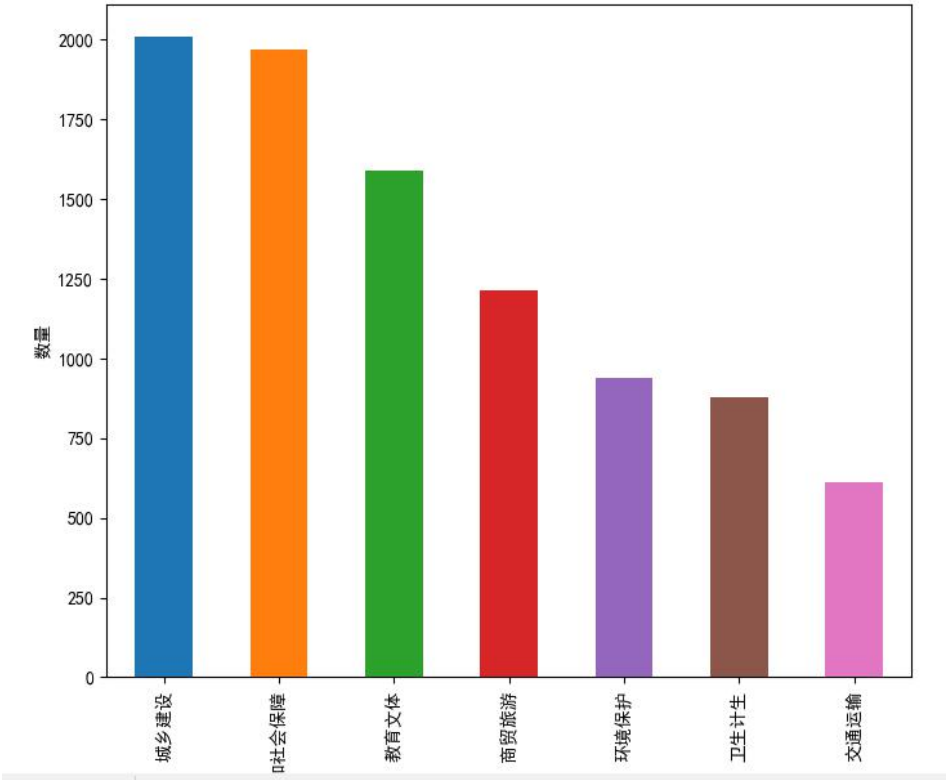

图 3-1 类目分布图

从上表 3-1 看到一共有 7 个类别，分别是：城乡建设、社会保障、教育文体、商贸旅游、环境保护、卫生计生、交通运输；各个类别的数量不一致，按序递减，分布不太均匀。表 3-1 给出了具体数值，城乡建设和社会保障与交通运输数量相差一倍以上，说明工作部门应该在城乡建设上多投入精力，而交通运输方面则做的比较完善，获得了人民群众的好评。

表 3-2：附件 1 定级分词

	一级标签	Cut_review
0	城乡建设	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆圈 西湖 建筑…
1	城乡建设	位于 书院 路 主干道 在水-方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏…
2	城乡建设	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…
3	城乡建设	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年不洗 现在 自来水 龙头…

表 3-2 所示，经过去空、转换 id、分词、过滤停用词等操作后，使每条留言内容提取到有用关键词 cut_review，完成数据预处理，为后面建立 LSTM 模型，对留言内容建立一级分类模型做准备。

3.2.2. LSTM 模型

表 3-3：LSTM 建模结果表

Layer(type)	Output Shape	Param #
embedding_1(Embedding)	(None,250,100)	1000000
spatial_dropout1d_1(Spatial)	(None,250,100)	0
lstm_1(LSTM)	(None,100)	80400
dense_1(Dense)	(None,7)	707

通过调用 Keras 工具，将附件 2 上的分词的留言信息数据进行向量化处理，把每条数据转换成相应的一个全为整数的向量，然后拆分训练集和测试集，由表 3-3 及处理结果可以知道，一共有 83642 个不相同的词语。拆分的训练集合和测试集合分别为（5710,250）（5710,7），（3500,250）（3500,7）；LSTM 层包含 100 个记忆单元，输出层为包含 7 个分类的全连接层。

3.2.3. 损失函数趋势图和准确趋势图

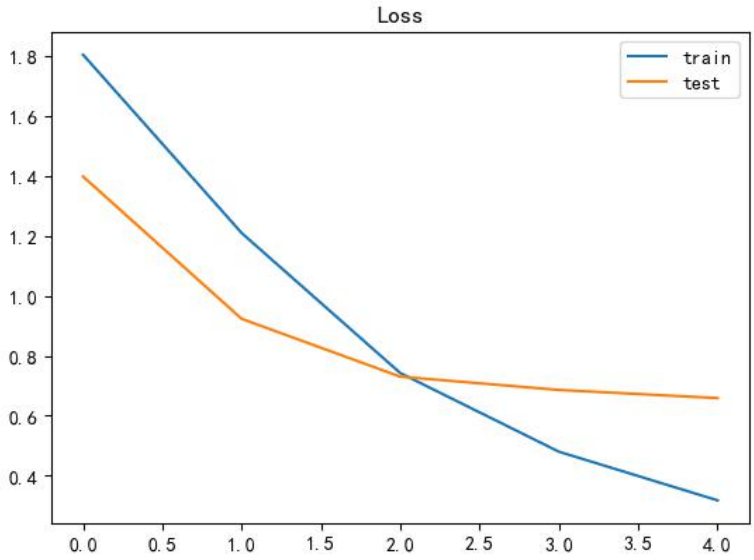

图 3-2：损失函数趋势图

从图 3-2 中可以看出,随着训练周期的增加,模型在训练集中损失逐渐减小,这是经典的过拟合现象,但在测试集中，损失随着训练周期的增加由一开始的从大逐步变小,再逐渐增大。

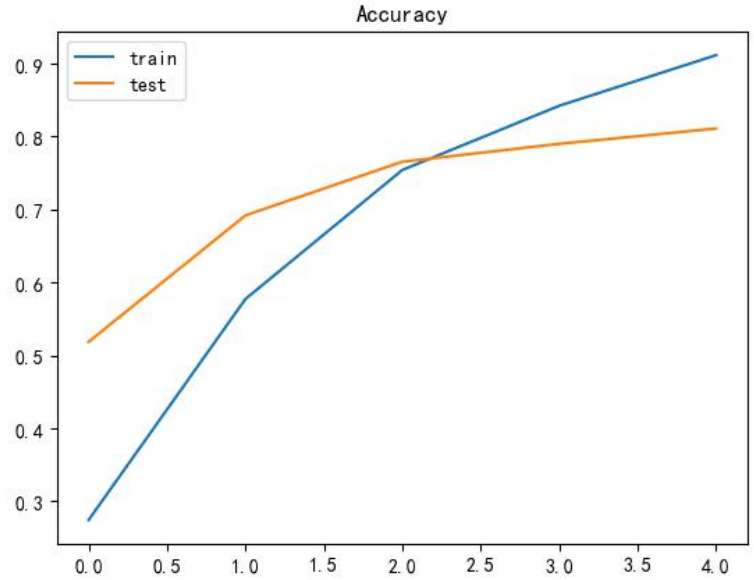

图 3-3：准确趋势图

从图 3-3 中我们可以看出,随着训练周期的增加,模型在训练集中准确率越来越高,这是典型的过度拟合现象,但在测试集中，准确率随着训练周期的增加由一开始的从小逐步变大，再逐渐减小。

3.2.4. 混淆矩阵和 F1 分数

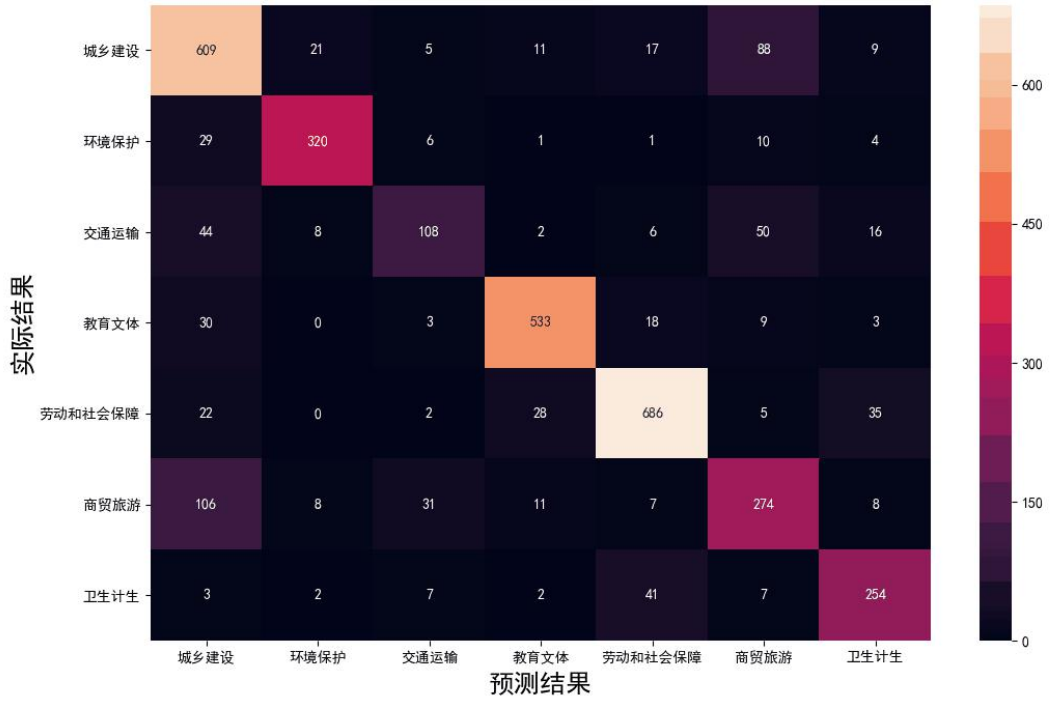

图 3-4 混淆矩阵

混淆矩阵的主对角线表示预测对的个数,除了主对角线外其他都是预测错误的数字，颜色越浅数量越大，反之越小。从上图混淆矩阵可以看出城乡建设、商贸旅游错误数量比较多，环境保护、教育文体错误数量较少。

多分类模型平时不使用准确率来评价模型的优劣,这是因为准确率不能很好地反应出每一个分类的准确性。而且当训练数据不平衡(比如有的类数据很多,有的类数据很少)时，准确率不能反映出模型的实际预测精度,此时可以借助借助于 F1 分数、ROC 等指标来进行评价模型的好坏。

表 3-4 F1-score 结果

	Precision	Recall	F1-score	support
城乡建设	0.72	0.80	0.76	760
环境保护	0.89	0.86	0.88	371
交通运输	0.67	0.46	0.55	234
教育文体	0.91	0.89	0.90	596
劳动社会保障	0.88	0.88	0.88	778

商贸旅游	0.62	0.62	0.62	445
卫生计生	0.77	0.80	0.79	316
Avg/total	0.80	0.80	0.79	3500

根据表 3-4 的运行结果来看，城乡建设，环境保护，交通体育，教育问题，劳动和社会保障，商贸旅游，卫生计生中，教育文体分类的准确率和召回率最高。从 F1 分数图像上看，教育文体的 F1 分数最大（90%），交通运输最小，只有 55%，可能是因为交通运输的本身数据集内训练数据太少，模型学习不够充分，招致预测出现的失误较大。而总的模型准确率为 79.5%，因此可以认为模型的分类效果可靠。

4. 问题 2

4.1. 分析方法与过程

4.1.1. 题目要求解析

附件 3 中提供了留言编号、留言用户、留言主题、留言时间、留言详情、反对数、点赞数等数据，因为只有某一时段内群众集中反映的某一问题可称为热点问题，所以要对各种特定问题的留言进行分类，对不同的热度分类指标加以权重，给出分类的准确度，留言编号、用户等不重要信息可以不予关注。

4.1.2. 基本流程图

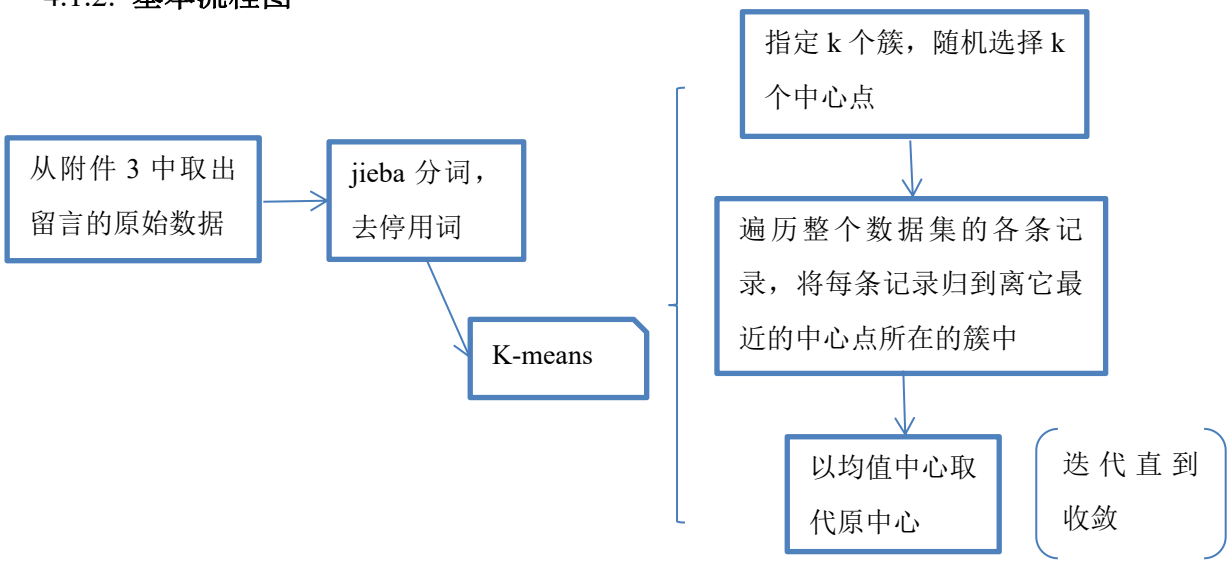

图 4-1 问题 2 流程图

4.1.3. 数据预处理

1. 取附件 3 中第 5 列数据存储到‘data.txt’文件中
2. 去除标点符号，导入 jieba 分词工具，导入停用词库‘stopwords.txt’
3. 分词生成‘seg.txt’，记录了留言的关键字词

4.1.4. LDA 模型

在留言详情的识别研究中，留言中的特征词是整个模型中唯一的可观察变量，LDA 模型在已知主题数目的情况下，调节特征词语在潜在主题上的概率分布完成留言的生成过程。在这个过程中，可以获得各个特征词语在潜在主题上的概率分布情况和每条留言在潜在主题上的概率分布情况。如果留言在某个潜在主题概率分布值越高，那它成为留言主题的可能性越大，如果潜在主题同时又是其他多篇留言的主题, 那么它就可能是整个留言集中的留言热点。如果某个特征词在某个主题上的概率分布值越高, 说明此特征词对该主题的贡献也就越大, 也就越有可能成为该主题的热点特征词。统计留言语料库中的主题分布情况，对每个主题出现的次数从高到低排序，设定阈值，选择排列在前若干位的主题作为留言集中的评论热点，再根据词汇表中的特征词在这些主题上的概率分布得到热点词。

4.1.5. K-means 聚类

1. 确定 k 值

k 均值算法的一个难点在于 k 值需要自己给定，常用的方法有手肘法和轮廓系数法，经过考量本题我们可以以 10 作为 k 值。

2. 优化目标函数

损失函数公式：

$$RSS_k = \sum_{x \in \omega_k} |x - u(\omega_k)|^2$$

$$RSS = \sum_{k=1}^{K} RSS_k$$

其中 ω_k 表示第 k 个簇，$u(\omega_k)$ 表示第 k 个簇的中心点，RSS_k 是第 k 个簇的损失函数，RSS 表示整体的损失函数。优化目的就是使用恰当的记录归属方案，使整体的损失函数最小。当损失函数的曲线出现较明显拐点或趋于平缓时，我们判断是收敛的。

3. 中心点的选择

k-meams 算法可以保证收敛，但却无法保证收敛于全局最优点，当初始中心点选取不好时，只能到达局部最优点，整个聚类的效果也会比较差。可以采用以下方法：

1) 选择彼此之间距离尽量远的那些点作为中心点；

2) 先使用层次进行初步聚类输出 k 个簇，以簇的中心点的作为 k-means 的中心点的输入；

3）多次随机选择中心点训练 k-means，选择效果最好的聚类结果。

4.2. 结果分析

表 4-1 问题 2 结果

Cluster 0:	公园 市委 市政府 具体 建设 国家 城市 周边 中心 规划
Cluster 1:	业主 物业 小区 电梯 物业公司 社区 问题 开发商 没有 多次
Cluster 2:	公交 地铁 路口 公交车 出行 车辆 建议 大道 方便 道路
Cluster 3:	请问 村民 什么 拆迁 是否 a7 户口 办理 政府 现在
Cluster 4:	公司 有限公司 没有 工资 员工 西地省 领导 社保 2018 工作
Cluster 5:	小区 幼儿园 居民 业主 生活 严重 影响 没有 部门 环境
Cluster 6:	开发商 车位 业主 购买 交房 销售 购房 合同 问题 政府
Cluster 7:	居民 严重 噪音 影响 施工 晚上 扰民 生活 部门 休息
Cluster 8:	学校 学生 家长 孩子 老师 小学 教育 中学 没有 教育局
Cluster 9:	没有 领导 一个 部门 现在 希望 问题 西地省 政府 解决

根据表 4-1 的结果分析，对于文本的 4338 行数据，我们分成了 10 个簇，得到了每个簇筛选了排名前十的关键字，用 TF-IDF 做文本特征选择，因为词频-逆文件频率比较能筛选出能代表整个文本的特征词。从运行结果可以看出，模型的准确率达到 88.14%，我们认为模型效果很好。

5. 问题 3

5.1. 分析方法与过程

5.1.1. 题目解析

经过对留言的基本分析之后，会有相关人员进行答复，但是回复的内容、针对对象、准确性难以考察，所以需要对其进行评价。

5.1.2. 数据预处理

先对附件 4 中的数据进行清洗，去除空值、无效值或没有参考价值的数据，和前两问类似，对答复进行 jieba 中文分词、去停用词。

5.1.3. 准确性评价

主要识别用户的称呼名是否正确，回复的相关内容与问题的主题是否有关联，对特定词加以权重计算准确性。

5.1.4. 情感分析评价

1) 导入知网的情感词典，用于对文本中一些情感性、评价性的词做色彩分类。比如正面情感：赞赏，快乐，感同身受，开心，鼓掌；负面情感：哀伤，疑惑，鄙视，不满意；正面评价：点赞，动听，正确 ；负面评价：丑，难，超标，不真实。另外，因为基本每条回复都带有“感谢您对我们工作的支持、理解与监督！”这句话，其中的正面词语可以忽略不计或降低权重。
2) 因为 TF-IDF 偏向于词频作为指标，有一些弊端，我们此时可以使用 Word2Vec 把数据中的字词组合转换成数值向量，在词语聚类、词性分析上比较合适，从 gensim 中导包后进行调参训练；
3) 构建朴素贝叶斯模型进行分类；
4) 此时可以使用 sklearn 中的 PCA（主成分分析）方法进行降维，支持向量机 SVM 进行分类，AUC 评价真阳率和伪阳率。

对答复的情感分析评价可以评判处理结果的好坏，处理方式的积极或消极。

比如对于盼望有电梯的评论答复是“...A3 区住建局高度重视，立即组织精干力量调查处理，现回复如下：为了完善住宅使用功能，提高我区既有多层住宅居民的宜居水平，2018 年 6 月 7 日，A 市 A3 区人民政府办公室下发了《关于 A 市 A3 区既有多层住宅增设电梯实施方案》的通知。该方案明确了增设电梯条件及流程。详情可咨询政务服务中心住建局受理申请老旧小区加装电梯窗口，咨询电话：0000-00000000...”，动作迅速且开展了具体措施，有关键词“高度重视”、“完善”、“提高”等词，是好的答复，有

利于民生；

而对于希望更改车站名的答复是“...公交站点的设置需要方便周边的市民出行，现有公交线路均使用该三处公交站站名，市民均已熟知，因此不宜变更...”，有“不宜”字眼，说明对留言的建议消极采纳。

6. 结论

本文提出了一个基于文本挖掘的针对平台用户留言信息分析、分类、评级、回复意见评价的模型，可以用于微信、微博等公共办事平台评论分析，或者类似于新闻评论、产品评论的文本情感分析。分摊了传统人工检索信息、分类文本的巨大工作量，方便工作人员更有针对性、更有效率地处理问题，还可以明确工作指标，有哪些问题可以解决并得到了解决，哪些问题本可以解决却得到了消极的反应，这都是可以明晰的。同时，如果想获得更客观精准符合民意的答复评价，也可以设置一个专门的板块让留言者或游客对答复内容进行打分，由此可以获得反馈。

7. 参考文献

[1] 马治涛. 文本分类停用词处理和特征选择技术研究. 硕士学位论文

[2] 凤丽洲. 文本分类关键技术及应用研究第 4 章基于主动学习和增量学习的垃圾邮件分类方法

[3] 延丰、杜腾飞等. 基于情感词典与 LDA 模型的股市文本情感分析. 上海大学信息与工程学院