“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。

本文聚焦“智慧政务”中的实际问题。运用机器学习与深度学习相关知识，结合自然语言处理技术，对留言分类、热点挖掘以及答复质量评价三类问题提出了有效的解决方案。运用词嵌入技术表示文本，通过使用朴素贝叶斯与卷积神经网络训练分类模型。将命名体识别与 k-means 聚类相结合进行热点挖掘。使用 tf-idf 算法将文本转化为稀疏矩阵，从而计算文本相似度、并结合完整性、时效性等指标，建立了留言答复评级体系，对每一条留言进行了综合评定。通过多组对比试验，选取最优的模型参数，运用可视化技术呈现，为“智慧政务”的实现提供了技术支持。

关键词：word2vec、朴素贝叶斯、CNN、命名体识别、k-means、tf-idf

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.

This paper focuses on the practical problems in "intelligent government". By using machine learning and deep learning and combining with natural language processing technology, this paper proposes effective solutions to three problems: message classification, hot spot mining and response quality evaluation. The text is represented by word embedding technique and the classification model is trained by naive bayes and convolutional neural network. The named body recognition and k-means clustering were combined for hot spot mining. The tf-idf algorithm was used to transform the text into a sparse matrix, so as to calculate the similarity of the text, and combined with the indicators such as integrity and timeliness, a rating system for message replies was established, and each message was comprehensively evaluated. Through a series of comparative experiments, the optimal model parameters are selected and presented by visualization technology, which provides technical support for the realization of "intelligent government affairs".

Keywords:word2vec, NB, CNN, Namedbody recognition, k-means, tf-idf

目录

1.绪论

1.1 问题重述

问题 1：群众留言分类

根据标签体系表和用户留言数据集来建立一个关于留言内容的一级标签分类模型，并使用 F1-score 作为评价指标

问题 2：热点问题挖掘

根据附件-3 中的留言用户留言数据集，挖掘出某一时段内反应特定地点或者人群的热点问题，并建立合理的量化指标，对热点问题的热度进行量化并将排名前五的热点问题写入热点问题表并保存为 xls 文件。按表 2 的格式给出相应热点问题的留言信息并保存为 xls 文件。

问题 3：答复性意见评价

针对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。

1.2 问题分析

1.2.1 群众留言分类

本问题着眼于如何根据已标记的文本数据集训练模型，并对未来可能出现的留言进行合理的预测分类，属于有监督的学习。可以使用的算法有传统的朴素贝叶斯（NB）、SVM 等，以及比较前沿的深度学习算法，如卷积神经网络模型（CNN）、长短时记忆模型（LSTM）、胶囊网络（CapsNet）等。

我们首先训练一个朴素贝叶斯分类器对文本数据集进行分类，朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法，有着坚实的数学基础，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC 模型与其他分类方法相比具有最小的误差率。因此，我们将 NB 的分类效果作为我们的 Baseline 。

NB 模型假设每个句子包含词语出现的概率是相互独立的。因此属于一个词袋模型。该模型忽略掉了文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合。此外，传统的机器学习算法需要人工提取文本特征，而特征的提取因方式方法而异很难找到最优的提取方法，为了结合语法和语序信息，并且实现自动特征提取。我们使用深度学习中的卷积神经网络（CNN）模型来对文本进行分类。卷积神经网络是一类包含卷积计算且具有深度结构的前馈神经网络，具有表征学习能力，能够按其阶层结构对输入信息进行平移不变分类。深度学

习往往需要大量的数据，对于中等规模的数据，为了提高泛化效果，我们将尝试使用 CNN 与动态路由算法结合，来降低较少数据的影响。

1.2.2 热点问题挖掘

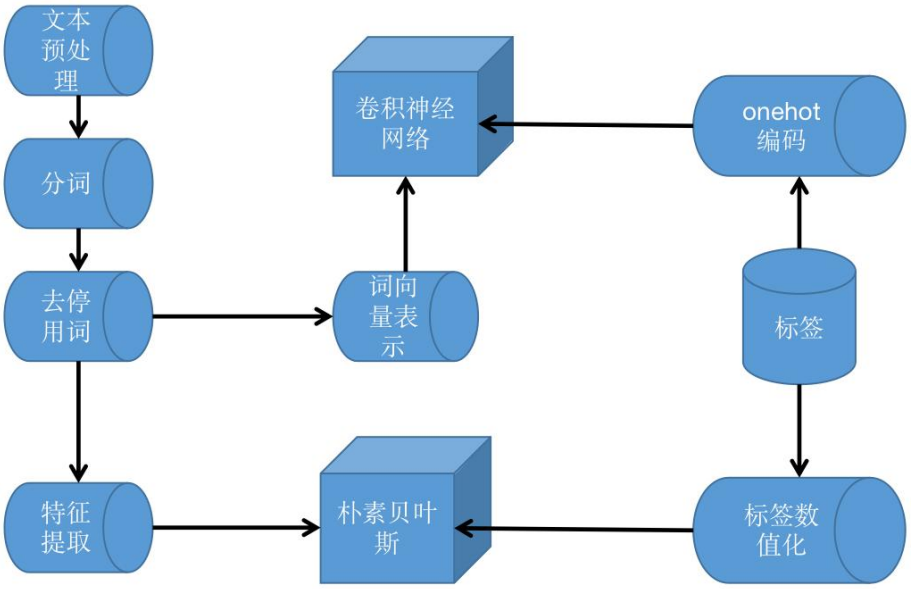
该问题难点在于地点或人群表达形式多样化以及特征多造成的两者之间相似度计算量大。对于特定地点或者人群发现，可通过命名体识别技术识别出留言中的地名、机构名等词性的词。对于热点问题的发现，采用聚类等方法对留言主题做聚类分析，根据问题的点赞数、反对数、反应频率等指标定义热度评价指标，对聚类结果的热度进行量化评价后发现热点问题。

1.2.3 答复性意见评价

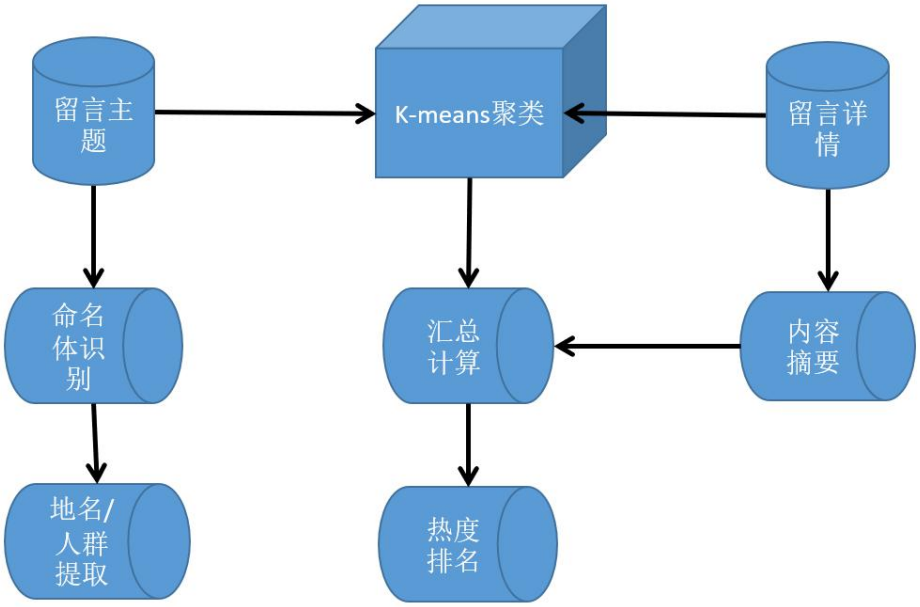
该问题属于开放性问题，针对相关部门对留言的答复意见从多个角度对答复意见的质量指定评价方案，可以使用文本相似度评价答复的相关性、从句法分析的角度衡量回复的完整性、运用文本匹配技术来评价答复的可解释性。

2.解题思路

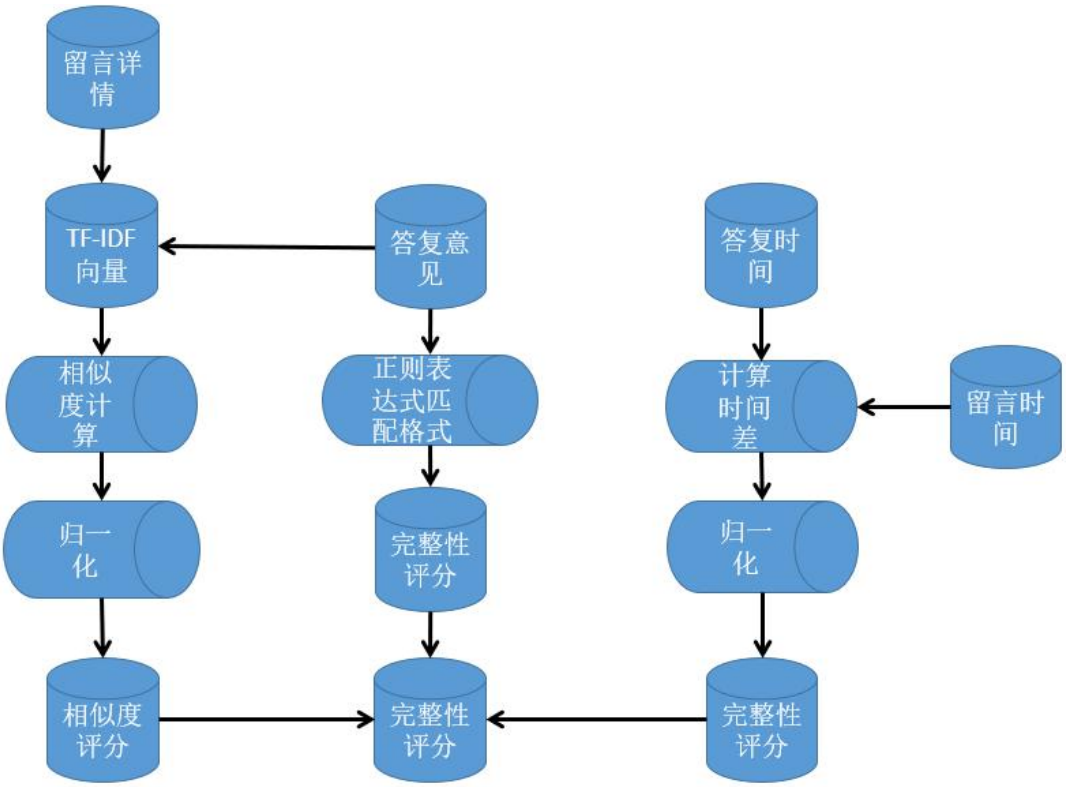
2.1 群众留言分类：



2.2 热点问题挖掘：



2.3 答复性意见评价：



3.自然语言处理

3.1 自然语言的特点

自然语言文本的字、词、句等各个层次上广泛存在歧义性或多义性。一段文本从形式上看是由汉字和标点组成的字符串。但是，形式上一样的字词，在不同的场景或不同的语境下，有着不同的含义。人之所以能够消除歧义，是依靠知识的积累，结合语法和上下文语境来获得正确的理解。一句话或一个词可能有多个含义，一个相同或相近的语义同样有多种表达方式，自然语言的形式与其意义之间是一种多对多的关系。

3.2 数据集分析

3.2.1 附件 2 数据集分析

```
sample.head()
```

	留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
0	24	A00074011	A市西湖建筑集团占道施工有安全隐患	2020/1/6 12:09:38	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...	城乡建设
1	37	U0008473	A市在水一方大厦人为烂尾多年，安全隐患严重	2020/1/4 11:17:46	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...	城乡建设
2	83	A00063999	投诉A市A1区苑物业违规收停车费	2019/12/30 17:06:14	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...	城乡建设
3	303	U0007137	A1区蔡锷南路A2区华庭楼顶水箱长年不洗	2019/12/6 14:40:14	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...	城乡建设
4	319	U0007137	A1区A2区华庭自来水好大一股霉味	2019/12/5 11:17:22	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...	城乡建设

图 3-1 数据预览

```
#检查是否有重复和缺失值
print('是否重复',sample.duplicated().any())
pd.DataFrame(sample.info())
```

```
是否重复 False
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9210 entries, 0 to 9209
Data columns (total 6 columns):
留言编号    9210 non-null int64
留言用户    9210 non-null object
留言主题    9210 non-null object
留言时间    9210 non-null object
留言详情    9210 non-null object
一级标签    9210 non-null object
dtypes: int64(1), object(5)
memory usage: 431.8+ KB
```

图 3-2 数据集信息

由图 3-1 可知，附件二共包含 9210 条记录，共有六个变量：['留言编号', '留言用户', '留言主题', '留言时间', '留言详情', '一级标签']，且每列均不包含空值，每行无重复记录。

我们主要使用留言主题、留言详情、和一级标签进行建模。在附件 1 数据集中，“一级标签”覆盖七类主题：

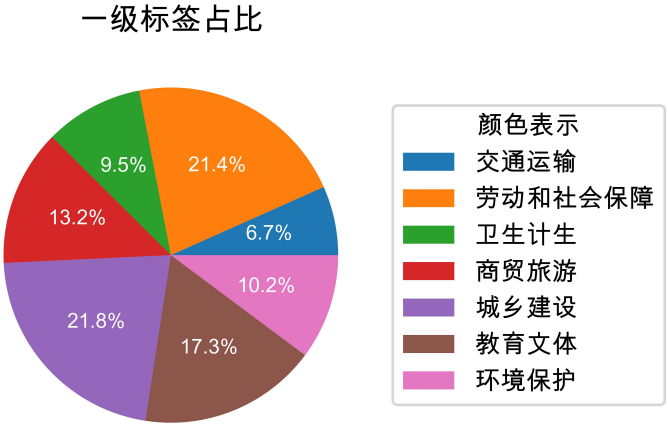


图 3-3 一级标签占比

由图 3-3 可以比较直观的观察出各类留言的数量占比

3.2.2 附件 3 数据集分析

```
[164]: print(f'附件三共有{sample.shape[0]}行,{sample.shape[1]}列')
       sample.head()
```

附件三共有4326行,7列

[164]:

	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
0	188006	A000102948	A3区一米阳光婚纱艺术摄影是否合法纳税了?	2019/2/28 11:25:05	\n\t\t\t\t\t\n\t\t\t\t\t座落在A市A3区联丰路米兰春天G2栋320，...	0	0
1	188007	A00074795	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	\n\t\t\t\t\t\n\t\t\t\t\tA市A6区道路命名规划已经初步成果公示文件，...	0	1
2	188031	A00040066	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	\n\t\t\t\t\t\n\t\t\t\t\t本人系春华镇金鼎村七里组村民，不知是否有相关...	0	1
3	188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	\n\t\t\t\t\t\n\t\t\t\t\t靠近黄兴路步行街，城南路街道、大古道巷、一步...	0	1
4	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	\n\t\t\t\t\t\n\t\t\t\t\tA市A3区中海国际社区三期四期中间，即蓝天璞...	0	0

```
[ ]:
```

图 3-4 附件三数据预览

3.2.3 附件 4 数据集分析

```
[167]: print(f'附件四共有{sample.shape[0]}行,{sample.shape[1]}列')
       sample.head()
```

附件四共有2816行,7列

[167]:

	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
0	2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53
1	2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10
2	2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14
3	2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42
4	2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30

[ ]:

图 3-5 附件四数据预览

3.3 文本预处理

3.3.1 词袋模型

3.3.1.1 模型原理

计算机只能处理数字信息，所以我们首先需要将人类的自然语言转化成计算机可以理解的数字形式。

在词袋模型中，将每一个句子转化为一个句向量，句向量的每一个维度都表示这一个词的词频信息。我们使用在这里 tf-idf 算法来实现词袋模型。对于一个词语数量为 n 的句子，首先计算每个词语在全部文本中的词频，然后计算每个词语的 tf-idf 值作为句向量的一个值。

词频（term frequency, TF）：指的是某一个给定的词语在该文件中出现的次数。

$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

公式 3-1

逆向文件频率（inverse document frequency, IDF）：IDF 的主要思想是：如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

$$IDF = log(\frac{\text{语料库的文档总数}}{\text{包含词条}w\text{的文档数} + 1}), \text{分母之所以要加}1\text{，是为了避免分母为}0$$

公式 3-2

$$TF - IDF = TF * IDF$$

公式 3-3

如果一个词语拥有较高的 tf-idf 值，则表示该词语对该句子的“特点”形成贡献率更高。在实际应用中，往往将 tf-idf 值较低的词语删去，保留 tf-idf 较高的词语，我们称之为特征提取。

3.3.1.2 实验步骤

1.分词

我们首先使用正则表达式去除标点符号以及数字和字母。然后使用 jieba 分词，对留言详情进行进行分词处理。每一条留言转化成一个词语列表。

2.去停用词

根据已有的停用词列表，去除一些无用的高频词语，比如“的”、“了”、“常常”等，这些词语也许有很高的词频，但并不会对模型的效果做出很多贡献。

3.标签数值化

对于七类一级标签，将每一个标签用 0-6 中的一个数值表示

4.特征提取

将每条留言转化为一个 tf-idf 向量表示。

3.3.2 Word2vec 词向量模型

3.3.2.1 模型原理

词向量，也被称为分布式词表示，指的是将词表示成一个定长的连续的实值向量。词向量计算是通过训练的方法，将语言词表中的词映射成一个长度固定的向量。词表中所有的词向量构成一个向量空间，每一个词都是这个词向 量空间中的一个点。这种向量有许多优点：

(1) 相似的词对应的词向量在特征空间中的距离也相近

(2) 词向量包含了更多的语义信息，每一维都有特定的含义。

词向量是训练语言模型时得到的副产物。因此，近年来词 向量引起了许多研究者的广泛关注。为了得到词向量，许多学者提出了多种语言模型，并在许多自然语言处理（NLP）任务中取得了很好的成果。

3.3.2.2 实践步骤

由于附件中的数据太少，难以训练出好的 word2vec 模型，我们从维基百科中下载了一个包含 80 多万个词语的语料库，并使用该语料库训练 word2vec。

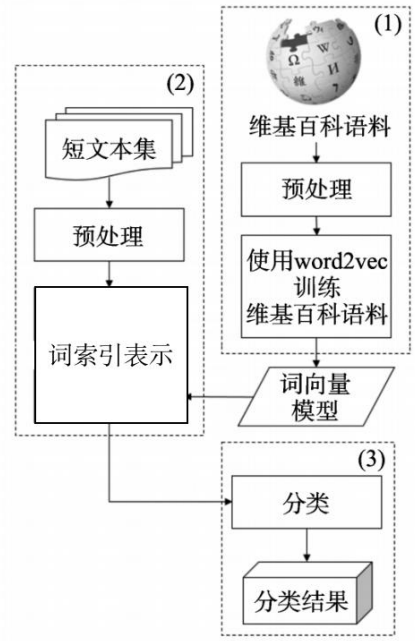


图 3-6 Word2vec 训练与应用

1.从维基百科中下载语料库，并使用 opncc 将繁体语料转换为简体

2.对转换为简体的语料库进行去标点、英文数字等文本预处理操作

3.使用 jieba 对语料进行分词

4.使用 gensim 库对分词结果进行词向量的训练

3.4.2.3 模型参数

参数	参数说明	取值
size	词向量维度	100
min_counts	忽略词频低于 min_counts 的词	5
window	窗口大小	5

4.群众留言分类

4.1 朴素贝叶斯原理

设有样本数据集$D=\{d_1,d_2,...,d_n\}$,

样本数据有m个特征($feature$)，对应集合$X=\{f_1,f_2,...,f_m\}$

样本可分为k个类别($class$)，对应集合$Y=\{c_1,c_2,...,c_k\}$,

假设各特征间相互独立，Y的先验概率$P_{prior}=P(Y)$,Y的后验概率$P_{post}=P(Y|X)$

条件概率$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)}=\frac{P(Y)\prod_{i=1}^{m}P(x_i|Y)}{P(X)}$

即$p(class|features)=p(c_j|f_1,f_2,...,f_m)=\frac{p(c_j)\prod_{i=1}^{m}p(f_i|c_j)}{p(f_1,f_2,...,f_m)}$

4.2 模型参数

参数	参数说明	取值
test_size	测试集大小	0.3

alpha	*拉普拉斯平滑参数*	*1.0*
cv	*k 折交叉运算折数*	*1,2...10*
ranom_state	*分割随机数种子*	*42*

4.3 结果可视化

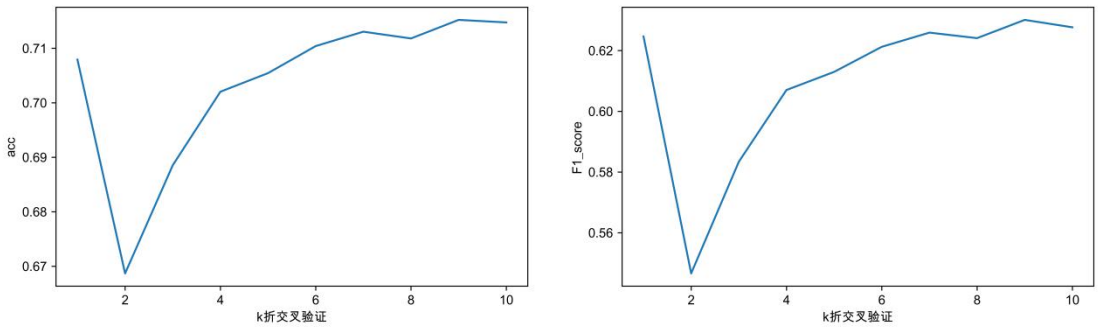


图 4-1 准确率与 F1_score 折线图

由图 4-1 可以看出，使用 9 折交叉运算分类器的准确率和 F1_score 都达到了最高，但不使用交叉运算的准确率和 F1_score 和使用 9 折交叉运算的效果的相差只有 0.008 和 0.006，为了减少计算量，我们并不采用交叉验证的方法。

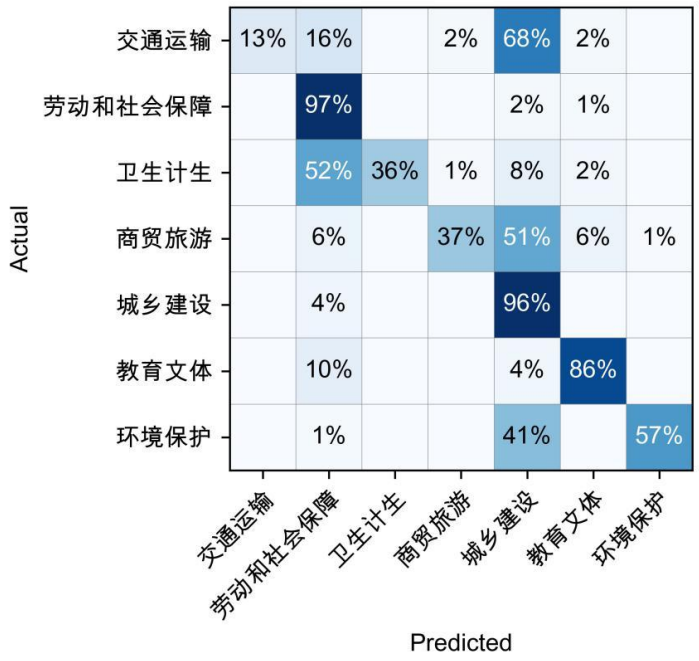


图 4-2 混淆矩阵热力图

我们使用在测试集上的预测结果与真实类别数据绘制了混淆矩阵热力图。在图 4-2 中，我们可以看到，分类效果最好的是卫生计生和城乡教育类，基本上有 95%以上的留言都被正确的分类。分类效果最差的是交通运输类，只有 13%的数据被正确预测。因此，想要提高模型的准确率，需要提高交通运输类的分类准确率，可以使用上采样的方法来增加交通运输类的训练数据量来减少泛化误差，由于时间有限，加之我们只希望吧朴素贝叶斯模型作为一个 Baseline，我们不做进一步的实验，转而使用新的模型——卷积神经网络模型。

4.4 卷积神经网络（CNN）原理

卷积神经网络最开始用于计算机视觉领域，是深度学习在该领域的基石。随着耶鲁大学的 kim 提出 TextCnn 算法，卷积神经网络开始逐渐应用于自然语言处理领域。

与前馈神经网络不同的是，卷积神经网络具有稀疏交互、参数共享、等变表示的特点。

稀疏交互是指卷积核的尺度远小于输入的尺度，每个卷积核仅与前一层局部区域的神经元依靠权重相连，我们可以通过多个卷积核捕捉文本中各个部分的局部特征，再将这些局部特征进一步的抽象为具体的特征。

我们规定每个卷积核的参数并不会随着位置的变化而变化，这叫做参数共享。参数共享使得卷积操作具有平移不变性，同时大大简化了计算量。

4.5 卷积神经网络结构：

一个卷积神经网络通常包含以下几种层：

卷积层：卷积神经网路中每层卷积层由若干卷积单元组成，每个卷积单元的参数都是通过反向传播算法优化得到的。卷积运算的目的是提取输入的不同特征，第一层卷积层可能只能提取一些低级的特征，更多层的网络能从低级特征中迭代提取更复杂的特征。

线性整流层：这一层神经的激活函数（Activation function）使用线性整流（ReLU）。

池化层：通常在卷积层之后会得到维度很大的特征，将特征切成几个区域，取其最大值或平均值，得到新的、维度较小的特征。

全连接层： 把所有局部特征结合变成全局特征，用来计算最后每一类的得分。

4.6CNN 模型训练

4.6.1 模型参数

表 4-1 word2vec 参数

参数	*参数说明*	*取值*
size	*词向量维度*	*100,300*
vocab	*包含词汇量*	*682505*
Sequence_length	*padding 定长*	*500*

经过多次对比试验，在 CNN 网络 embedding 层中使用预训练的 word2vec 准确率和 f1_score 反而更低，经过分析，可能是因为维基语料库与留言内容有着较大的差异，因此我们选择使用 embedding 层自训练的词向量模型。

表 4-2 CNN 模型 A(Conv1D)

参数	参数说明	取值
kernel_size	卷积核大小	3
filters	卷积核数量	128
stride	步长	1
dropout	随机丢弃比率	0.2
optimizer	优化器	Adam
Maxpool	最大池化窗口	2
batch_size	批处理大小	128
epoch	迭代轮数	10
validation	验证集大小	0.1

表 4-3 CNN 模型 B(Conv1D)

参数	参数说明	取值
kernel_size	卷积核大小	5
filters	卷积核数量	64
stride	步长	1
dropout	随机丢弃比率	0.2
optimizer	优化器	Adam
Maxpool	最大池化窗口	2
batch_size	批处理大小	128
epoch	迭代轮数	10
validation	验证集大小	0.1

表 4-4 CNN 模型 C(Conv2D)

参数	参数说明	取值
kernel_size	卷积核大小	(2,100),(3,100)
filters	卷积核数量	128
stride	步长	1
dropout	随机丢弃比率	0.2
optimizer	优化器	Adam
Maxpool	最大池化窗口	(sequence_length-kernel_size+1, 1)
padding	填充	valid
batch_size	批处理大小	128
epoch	迭代轮数	10
validation	验证集大小	0.1

表 4-4 CNN 模型 D(Conv2D)

参数	参数说明	取值
kernel_size	卷积核大小	(2,100),(3,100),(4,100)
filters	卷积核数量	128
stride	步长	1
dropout	随机丢弃比率	0.2
optimizer	优化器	Adam
Maxpool	最大池化窗口	(sequence_length-kernel_size+1, 1)
padding	填充	valid
batch_size	批处理大小	128
epoch	迭代轮数	10
validation	验证集大小	0.1

表 4-5 CNN 模型 E(Conv2D)

参数	参数说明	取值
kernel_size	卷积核大小	(3,100),(5,100),(7,100)
filters	卷积核数量	128
stride	步长	1
dropout	随机丢弃比率	0.2
optimizer	优化器	Adam
Maxpool	最大池化窗口	sequence_length-kernel_size+1, 1
padding	填充	valid
batch_size	批处理大小	128
epoch	迭代轮数	10
validation	验证集大小	0.1

4.6.2 数据预处理

1.分层抽样。

由图 3-3 可知，不同类型的文本分布不均，因此需要进行分层抽样，确保每类都能抽到数量相对均衡的数据。

2.标签 one-hot 表示。

3.句子末尾 padding 处理

4.词向量表示（可选）。

4.6.3 结果可视化

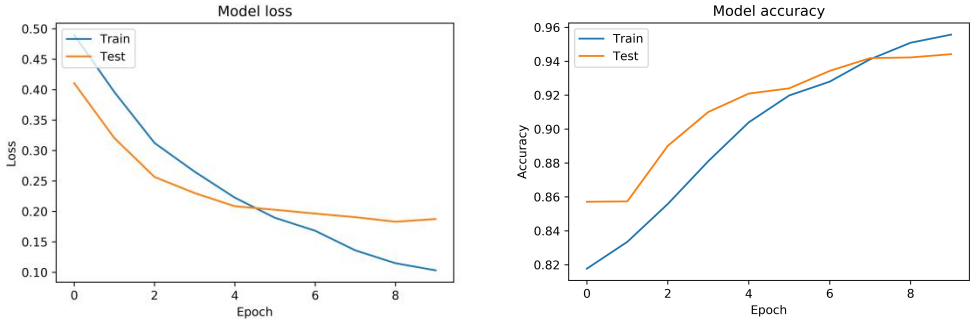


图 4-4 模型 A 损失值与准确率

由图 4-4 可以看出，在第八轮迭代开始，测试集的损失值开始升高，准确率上升变得缓慢，模型开始出现过拟合。

对于一维卷积神经网络，我们调整了不同的卷积核数量和迭代轮数，并从中选取了准确率和 F1_score 最高的组合即模型 A。限于篇幅原因不在文中呈现对比试验。

接着我们希望将我们预训练的词向量应用到 CNN 中，于是我们进行了二维 CNN 探索。考虑到每条留言内容的不是很长，我们选取的卷积核尺寸和数量也都都选取的比较小，层数也比较单一，以避免过拟合。

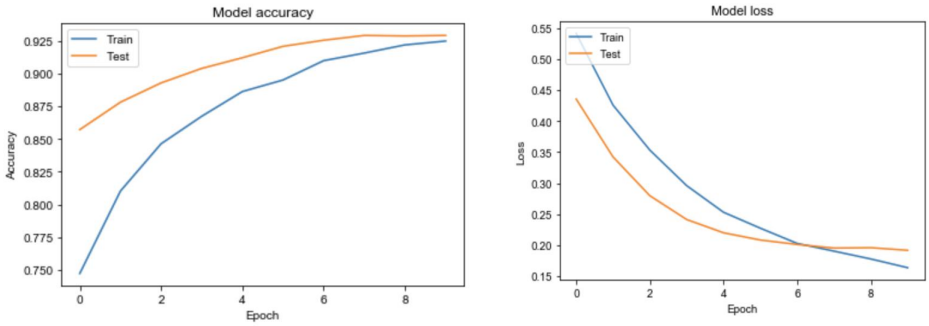


图 4-4 模型 D 损失值与准确率

根据模型的分类结果，我们选择模型 A 作为结果。该模型比 NB 分类器的 F1——score 值增加了 0.2919，体现出比较好的分类效果。

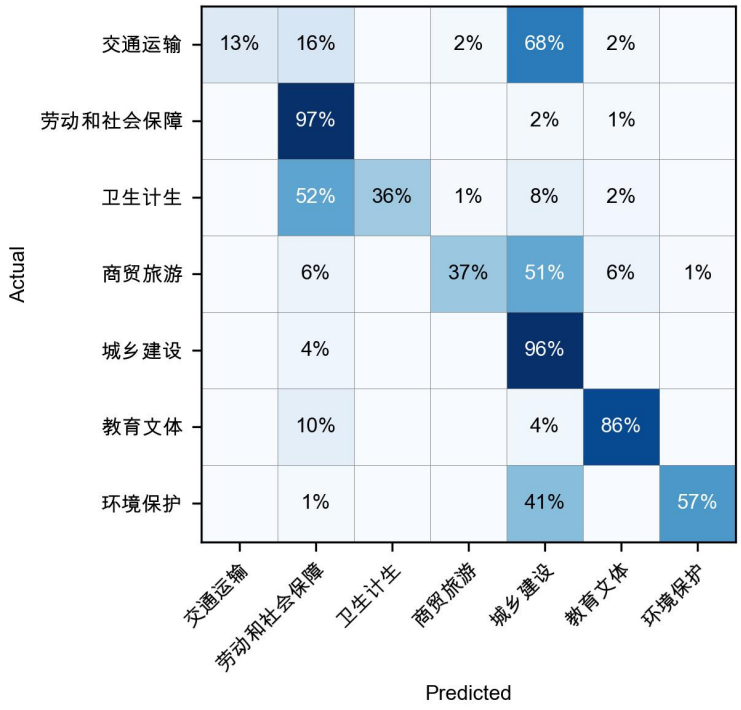
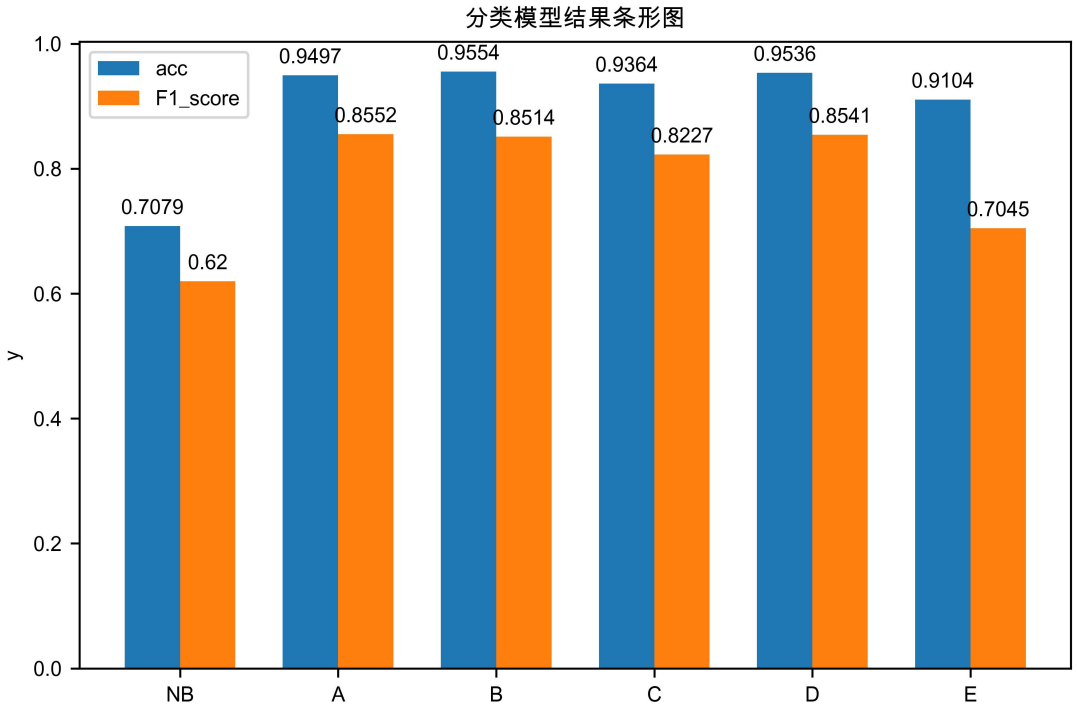


图 4-5 模型 A 混淆矩阵

有图 4-5 可以看出，交通运输类的分类效果比较差，大部分的交通运输容易被分到城乡建设类中去，原因有二，一方面是交通运输类留言内容数据量较少，另一方面交通运输类的留言与城乡建设类的留言相似度较高，比较难区分。出现相同问题的还有卫生计生与劳动和社会保障类。

表 4-6 模型分类结果

卷积模型	测试集准确率	F1_score
A	0.9497	0.8552
B	0.9554	0.8514
C	0.9364	0.8227
D	0.9536	0.8541
E	0.9104	0.7045



5.热点问题挖掘

5.1 命名体识别

5.1.1 pyhanlp 分词工具简介

Pyhanlp 是 HanLP 的 python 接口，可以实现自定义词典、极速词典分析、关键字提取、自动摘要、文本分类等多重功能，内部算法经过工业界和学术界考验。

5.1.2 地点/人群的识别

通过对附件 3 数据的初步的观察，我们发现市、区、县等地名被进行了脱敏处理，另外还含有一些该市特有的街道名、村镇名、小区名。所以首先提取出这些特殊的地名，存入 txt 文件中。接着，向分

词词典中加入本案例中出现的特殊地名，对每一条留言的留言主题进行分词并标注词性。提取出词性为‘nr’(人名)、‘ns’(地名)、‘nt’(机构名)、‘nz’(其他专名)的词作为每一条留言对应的地点/人群。

5.2 热点问题发现

5.2.1 热力评价指标的定义

热力评价指标需要全面客观评价某一问题的热度，在本案例中问题热度f受点赞数、反对数、频数三个因素影响，函数表达式定义如下：

$$f = \text{点赞数}+\text{反对数}+\text{频数}$$

5.2.2 k-means 聚类

K-means 算法原理简述如下：

假设有n个点$d_1,...,d_n$，需要将其分成k个簇$S_1,...,S_k$，簇的质心为$C_1,...,C_k$，
对于传统$K-means$算法，满足条件：最小化所有点到质心的距离的平方和

$$\min\ I_{Euclidean} = \sum_{r=1}^{k}\sum_{d_i \in S_r} \|d_i - c_r\|^2$$

其中$\|d_i - c_r\|$是点与质心的欧式距离
质心的计算为簇内点的几何平均：

$$s_i = \sum_{d_i \in S_r} d_j$$

$$c_i = \frac{s_i}{|S_i|}$$

其中，s_i是簇S_i内所有向量之和，称为合成向量
在Hanlp中，$K-means$的准则函数发生变化，
满足条件：最大化同一簇内的点与质心的相似度

$$\max\ I_{\cos} = \sum_{r=1}^{k}\sum_{d_i \in S_r} \cos(d_i, c_r) = \sum_{r=1}^{k}\|s_r\|$$

聚类的步骤如下：

1、在数据集中选取 k 个点作为 k 个簇的初始聚类质心。

2、计算质心外所有点到每个质心的距离，将这些点分配给最近的质心所在的簇。

3、对每个点，计算将其移入另一个簇时 $I_{\cos}$ 的增大量，找出最大增大量，并完成移动。

4、重复步骤 3 直到达到最大迭代次数，或簇的划分不再变化。

通过不断调整超参数 k，k 的取值最终确定为 900。根据热力评价指标进行排序后，排名前 5 的热点问题如图所示。

热度指数	时间范围	地点/人群	问题描述
2394	2019/01/29至2019/07/08	A市A4区58车贷	6000受害人分布于全国
2142	2019/01/15至2019/09/19	A市A5区汇金路五矿K9县	小区的保洁也是一直跟物业投诉然后也是一直无果
2114	2019/01/08至2019/12/31	A市金毛湾	明确于2018年7月31日前将金毛湾纳入周南梅溪湖中学或西雅中学配套入学范畴
685	2019/08/23至2019/09/05	A4区绿地海外滩高铁	你们要在这小区旁边建高铁铁路是何其扰民
299	2019/04/23至2019/12/31	A市丽发新城	物业未提供相关合同服务标准

6.答复性意见的评价

6.1 制定评分指标

由图 3-4 可以看到，附件四一共包含着 2816 条数据，7 列变量。为了建立一个答复意见质量的合理评价体系，我们从相似度、完整性、时效性三个方面对答复意见进行量化分析，并综合三个指标的评分对每一条答复意见进行评级。这三个指标的量化规则制定如下：

相似度：答复的内容是否与群众所提问题相匹配是衡量答复意见质量重要指标，如果答非所问，或虚与委蛇都不是一条让人满意的答复。

因此我们将文本相似度作为评价指标之一。分别计算留言详情与答复意见的 tf-idf 向量。并计算每一条留言详情与对应答复意见句向量的欧氏距离，对欧氏距离进行归一化处理，并把归一化的结果乘以 100 作为相似度评分。

完整性：固定的格式是必要的，比如敬语、对留言的感谢以及落款。这体现了对留言重视与尊重。我们从回复中发现了一些特定的格式，如果一条答复意见具备其中之一，我们就提高该回复的完整性评分。我们使用正则表达式，匹配答复意见中三个内容，包括引语，答谢，以及日期落款。匹配到三项中的任意一项，增加 33.3333 分。满分约等于 100　此外，过短的回复被认为是无内容的，我们设置了一个数字阈值（字数的下十分位数）如果一条回复少于该字数阈值，我们认为该条回复没有合理的答复问题，完整性得分直接被判定为 0。

时效性：留言的回复时间也是一个很重要的指标，这体现了对留言问题的重视程度以及答复效率。我们把留言时间与答复时间的时间差，作为时效性度量指标。对时间差作归一化处理，由于时间差越大说明时效性越低，因此我们的归一化公式有所不同

$$y = \frac{x_{max} - x}{x_{max} - x_{min}}$$

最后对三个指标求平均值，得到答复意见的综合评分并规定评级方式：高于综合评分上四分位数的评为 A，高于中位数的评为 B,高于下四分位数的评为 C，低于下四分位数为 D。

6.2 结果展示

```
df.head()
```

	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	综合评分	答复质量等级
0	2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	93.317767	A
1	2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	87.453209	A
2	2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	67.929135	D
3	2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	81.189047	B
4	2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30	58.122700	D

图 6-1 评级展示

7.总结与展望

总体来说，我们的模型的优势在于模型简单实用、代码可复现性强，效率较高。当然，由于时间原因，模型无法进一步优化，我们提出一些模型优化的具体方法如下：

对于文本分类问题，使用 CNN 模型的 F1-score 并不很高，我们认为可以在四个方面进行改进，一方面，由于训练集数据较少，我们可以使用过采样的方式增加数据量，另一方面，对于比较少的数据，可以结合动态路由算法，实践表明，CNN 结合动态路由算法使用较少的数据也可以得到比较好的结果，最后也可以使用与留言回复相似度较高的语料训练词向量。当然，也可以对 NB 分类器进行改进，比

如使用平均词向量替代 tf—idf 句向量作为输入，与多个分类器进行组合等，遗憾的是由于时间原因，我们并未能实现这四种技术提升。

热点问题挖掘中使用的文本聚类属于无监督学习，所以在聚类的准确率上难以保障，常出现同一类别的留言主题被区分到两个簇或同一个簇中包含不同主题的留言的情况。k-means 聚类的一大缺点是超参数 k 值难以确定，需要人工判断聚类效果，据此对 k 值不断修正，不符合便捷、高效需求。Pyhanlp 接口除了提供 k-means 聚类外，还提供了另一种聚类方法：repeated bisection。repeated bisection 算法相较于 k-means 算法可通过设定准则函数阈值 β 作为停机准则，避免超参数 k 难以确定的问题，并且，在速度上远优于 k-means。在测试过程中 repeated bisection 聚类效果不及 k-means，考虑到数据量较小，所以最终选用 k-means 进行聚类。在后续研究中当数据集较大时，或许 repeated bisection 算法能够取得更好地效果。

留言评价体系的不足在于主观性太强，三类指标的评分机制过于死板，没有体现出三者的联系，如果时间允许，可以通过调查问卷、专家访问等方式制定三个指标之间的权重，达到更合理的标准。

8.参考文献

[1]崔懿心.基于机器学习算法的社交数据挖掘与用户偏好的建模[J].电子技术与软件工程,2019(14):174-175.
[2]张弛,张贯虹.基于词向量和多特征语义距离的文本聚类算法[J].重庆科技学院学报(自然科学版),2019,21(03):69-72+77.
[3]何养明. 基于卷积神经网络结合词向量的中文短文本分类研究[D].重庆理工大学,2019.
[4]何愉,卫陈泉,陆钰华.基于深度神经网络与主题模型的文本情感分析——以上海迪斯尼景区游客满意度调查为例[J].统计科学与实践,2016(12):17-21.
[5]蔡慧苹. 基于卷积神经网络的短文本分类方法研究[D].西南大学,2016.
[6]魏德华. 微博热点话题发现问题的研究及实现[D].福州大学,2017.
[7]韩春燕,刘玉娇,琚生根,李若晨,苏翀.中文微博命名体识别[J].四川大学学报(自然科学版),2015,52(03):511-516.
[8]朱君,程雅梦.电力工单文本数据分析挖掘模型研究[J].电力需求侧管理,2017,19(S1):87-89.
[9]吴柳,程恺,胡琪.基于文本挖掘的论坛热点问题时变分析[J].软件,2017,38(04):47-51.