“智慧政务”中的文本挖掘应用

摘要

本文主要研究了根据给出的市民留言数据来进行群众留言分类，热点信息的挖掘以及答复意见的评价。对于现在“智慧政务”的实施有着极大的意义，方便更快捷更便利的解决市民的生活问题。

对于问题一，建立关于留言内容的一级标签分类模型，用深度学习中的 LSTM 长短期记忆网络对中文文本进行多分类，先进行数据预处理，然后进行 LSTM 的建模工作，定义好 LSTM 模型以后，我们要开始训练数据，接下来我们通过画混淆矩阵和求 F1 分数来评估我们模型的表现。不仅如此我们还用了另一种算法朴素贝叶斯。

对于问题二，将某一时间段内反应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进行评价。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。建立了留言问题归类模型，在归类用到了两种聚类方法 k-means 和 DBSCAN，k-means 是基于划分的聚类算法，DBSCAN 是基于密度的算法,两者相比较发现 DBSCAN 更为优秀，但针对未到分类留言又进行了 k-means 算法，最终将两者的分类进行合并。热度评价指标则通过分析每条留言的信息，根据需要及时的热点问题，得到热点计算公式，对每类留言进行了评价。

关键词：LSTM 长短期记忆网络，朴素贝叶斯，k-means，DBSCAN，热点评价

目录

1
绪论..... 1
1.1
研究背景和研究意义..... 1
1.1.1
研究背景..... 1
1.1.2
研究意义.....3
1.2
研究内容..... 4
1.3
论文结构..... 5
2
建立文本一级分类标签模型以及评价..... 7
2.1
LSTM 长短期记忆网络..... 7
2.2
朴素贝叶斯算法..... 12
2.3
其他分类算法..... 14
2.3.1
Svm 算法..... 14
2.3.2
GaussianNB 算法..... 14
2.3.3
MultinomialNB..... 15
2.3.4
KNeighborsClassifier.....15
3
文本归类..... 17
3.1
k-means 算法..... 17
3.2
DBSCAN 算法..... 19
3.3
其他聚类算法 21
3.3.1
基于层次的方法 17
3.3.2
基于网络的方法..... 22

4
热度排序 22
4.1
合理热度指标建立..... 22
4.2
留言问题的关键词提取..... 23
4.3
答复留言的评价..... 23
5
建模前对文本进行的处理..... 24
5.1
向量归一化处理..... 25
5.2
文本向量化处理..... 25
5.3
中文分词..... 25
6
总结与展望..... 26
6.1
总结..... 26
6.2
展望..... 27
参考文献.....28

1.绪论

1.1 研究背景和研究意义

1.1.1 研究背景

互联网的快速发展推动着人类迅速发展，他加入人们生活的方方面面，人们从原来的互通书信到现在的手机短信，各种社交软件。从原来的跋山涉水去见朋友家人，到现在方便快捷的视频通话。人们原来想买东西还需要货币并且前往相应的商铺，再到现在在网上就可以轻松购物并且可以买到世界各地你想要的东西。在网络上大家可以畅所欲言，许多人通过相同的兴趣在网络上相识，结交好友，互联网的诞生改变了人们的生活方式，加快了各方面的发展。中国互联网络信息中心（CNNIC）发布第 45 次《中国互联网络发展状况统计报告》，网民规模突破 9 亿 为数字经济发展打下坚实用户基础。《报告》显示，截至 2020 年 3 月1日，我国网民规模为 9.04 亿，互联网普及率达 64.5%，庞大的网民构成了中国蓬勃发展的消费市场，也为数字经济发展打下了坚实的用户基础。CNNIC 主任曾宇指出，当前，数字经济已成为经济增长的新动能，新业态、新模式层出不穷。在此次疫情中，数字经济在保障消费和就业、推动复工复产等方面发挥了重要作用，展现出了强大的增长潜力。由此可见互联网已经深深嵌入我们的平常生活中，社会人民已经离不开它。

在这个大数据时代的今天，我们的生活发生了翻天覆地的变化，正如知乎上的用户黑格尔说的，数据与我们日常生活的联系从未如此紧密过，从没有像今天如此活跃，具体的记录着人类与世界。从最初的计算机，摄像头到家用计算机，智能手机，再到大数据和人工智能，我们不断升级采集和利用数据的方式。而现在，从一辆车的每日碳排放量统计到全球气温的检测，从预测个人在网上喜好分析到总统选举时投票趋势的预测，我们都可以做到。数据将人与人，人与世界连接起来，构成一张繁密的网络，每个人都在影响世界，又在被他人影响着。传统的统计方法已经无法处理这种相互影响的数据，这么办？答案是让机器自己来处理数据，从数据中习得知识。这便是当代人工智能的本质。与传统的数据记录定义不同，这种数据是有“生命”的。它更像是我们身体的一种自然延伸：聆听我们的声音，拓宽我们的视野，加深我们的记忆，甚至组成一个以数据形式存在的“我”。然而我们也在为处理这些数据而努力着，面对着每天庞大数量的数据，我们还需要思考如何从中获得高效的信息，从里面挖掘出对我们有用的数据，这是人工智能诞生帮助我们解决这一问题，它出现在我们生活中，帮助我们解决了许多人工解决起来麻烦的问题。在你的手机阅读软件中，书籍被分好了类这样你就可以根据你的需求找到你想看的那本书，这么庞大的书记量如果让人工来处理不仅费时可能还会发生差错。毋庸置疑人工智能已经对我们的生活提供了很大的便利，省去了我们不少麻烦，让我们的一些愿望从原来的不敢想到现在可以轻松实现。实现这种技术离不开

的是自然语言，自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。它所能实现的自动分类，自动聚类对文本数据的处理有巨大影响，它可以有效的管理文本方面的信息。

生活中社交媒体的热点新闻，娱乐圈大事不仅需要分类，更需要对他们的热度进行排序，从关注度，事件对社会产生的影响等等方面进行考量，让他们呈现在大众的视野中，对于他们的关注和讨论，是亿万网友们通过网络来实现的。由于每天发生的新闻众多，普通市民们无法从众多事件中发现出重大新闻，这时就需要对这些新闻进行处理，对这些文本进行划分，这样网民们可以根据自己对哪些方面感兴趣来进行相应搜索。对于政府部门而言，每天收到大量的留言，需要解决群众的问题，更需要对问题的各方面进行考虑，由重到轻地按照一定顺序的帮大家排忧解难，保证社会的平稳。

1.1.2 研究意义

本文需要解决对给出数据的一级分类建立好模型后用 F-score 对建立好的模型进行评价以及对留言进行归类排序，最后对部门给出的回复进行评价，运用到的分类算法以及聚类算法，还有对文本数据处理的方法相结合完成将市民留言等一系列问题进行解决，这些放到现实生活中政府进行对市民留言的解决也是有很大意义的。

本文运用到的 k-means 算法和 DBSCAN 算法进行比较选择出对于聚类较优的算法来进行对留言信息的聚类，但在实验过程中由于得到

的结果不太理想，于是将两种算法结合得到了最终的归类数据，具有创新意义。

公众给出的建议问题中，需要将问题的热度以及它的排序进行实现，这样才能快速准确的进行分配，并且让大众知道问题所在，还可以收集到他们对问题提出的看法，供处理问题时进行参考。将问题的热度量化，更加方便对它们进行排序，相关部门就可以依次来解决这些问题，市民们也可以发现哪些问题需要被重视，引发讨论。结合多因素来决定排序的规则，从给出的问题角度来搜寻有用的信息，最终对这些问题进行排序。

1.2 研究内容

本文主要进行对文本分类，对文本聚类，在整理出一个合理的热度评价方式，对归类后的留言问题进行排序，最后对给出的答复意见从各个角度进行评价。其中涉及了许多技术，中文分词，文本向量化，LSTM 长短期记忆网络分类，k-means 聚类，DBSCAN 聚类，热点评价以及他的排序。建立好适合的模型是解决这些留言问题的关键，所以要选出最佳的模型来对他们进行分类，聚类，评价。

本文对第一题采取了两种建模方式，一种用了 LSTM 长短期记忆网络，一种用了朴素贝叶斯算法，主要研究了他们分类的流程以及他们对文本进行分类后用 F-score 对他们进行评价，其中 LSTM 长短期记忆网络还运用了混淆矩阵。LSTM 会对训练集进行测试，要防止他过度拟合，通过损失函数趋势图和准确率趋势图来判断。朴素贝

叶斯算法获得的模型，其中还涉及到词云图以及 TF-IDF 算法给每个词加权，最后通过朴素贝叶斯创建了模型。

本文第二题是对文本进行聚类，聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集，这样让在同一个子集中的成员对象都有相似的一些属性。对于文本聚类本文对两种聚类算法进行了比较，k-means 算法和 DBSCAN 算法，对他们进行比较选出最优算法并且对数据进行归类。K-means 算法最重要的是要取到 k 值，可以用手肘法。DBSCAN 需要确定 eps， min_samples。在最后的归类模型创造时对单一的算法进行了改进。

在进行一系列建模操作之前，我们获得的文本数据无法直接使用，需要对获得的文本进行预处理，包括对他们进行去空去重，替换或去掉里面某些不用的文字，最后将他们进行中文分词，这些操作都是为之后的建模做铺垫，他们完成的准确与否会大大影响之后建模后的结果，本文采用的是 jieba 分词。

本文第二题后面还需要对热度进行评价并且将他们按照热度进行排序，考虑综合因素，以及题目给出的要求，建立了热度计算公式，以多方面对留言的影响，结合相应理论知识将符合要求的计算公式得到，为了方便排序以及计算和呈现在大众视野后更方便看懂，需要将热度量化，变成方便的数字，此时便需要将文本量化，这又涉及到另外的知识了。

1.3 论文结构

本文共分为章，各章节分布以及每章节具体内容如下；

第一章是绪论，主要针对需要解决的问题介绍了现代互联网技术研究背景，对分类聚类以及热度评价的研究意义以及本文对于为了解决这些问题需要研究的内容。

第二章是对文本分类的分析以及对解决这个问题所用到的两种技术 LSTM 长短期记忆网络和朴素贝叶斯算法，建立朴素贝叶斯模型还会用到 TF-IDF 算法，在此基础之上还会介绍几种其他的分类方法，其中对文本的预处理也是解决问题的关键。最后对建立的模型进行 F-score 的评价

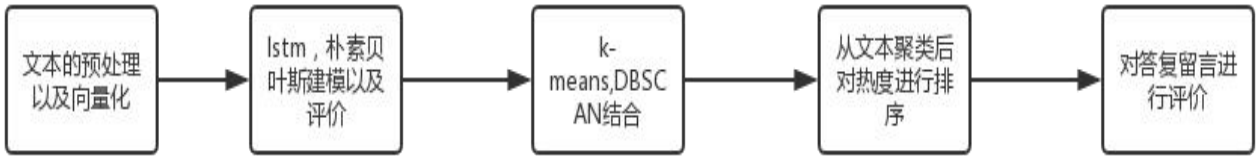
第三章是对文本进行归类的分析以及制作热度问题表和热度问题留言明细表，其中需要将两种算法进行比较，选出对于留言问题归类较优的算法，在得到结果后对模型在进行改进，对聚类问题进行分析。最后得出了最终的聚类结果。

第四章是对热度问题的分析，如何执行一个合理的热度评价指标，其中涉及到文本量化，将热度转化为方便排序的数字，让大家更直观的看见哪些是当下及时的热度问题，从多个方面去制定这个指标。最后实现了对热度指标的判定以及表格的制作。

第五章是对处理问题的前提文本预处理进行分析，还要关键字的提取，文本预处理是对之后建模做铺垫，是关键所在，中文分词用到的 jieba 分词，对他进行详细的分析，在朴素贝叶斯模型还会制作直观的词云图。还有对答复留言进行多角度的评价。

第六章是对本文研究结果的总结以及对工作的分析，还有待提高的地方，另外对今后人工智能学习做出展望。

论文总体工作的流程图如图一所示



2.建立文本一级分类标签模型以及评价

对于第一问用到了 LSTM 和朴素贝叶斯来对文本进行分类，同时还介绍了其他几种算法，这些分类算法都是当代自然语言研究的方向之一，众多分类方法值得人们去研究探讨，同时 F-score 也是对模型建立后进行评价，本章就会研究主要的两种算法以及介绍其他多种算法。

2.1 LSTM 长短期记忆网络

LSTM 在 RNN 的基础上扩展出来的，整体上看，LSTM 和 RNN 是类似的，既向前传递信息，又处理当前信息。但是，LSTM 允许保留或者忘记信息，如图，表示 LSTM 的一个 cell，可以看到，其核心为 cell 的 各种状态 & 控制门，可以把有用的信息保留在 cell 中，并移除无用信息。LSTM 一共拥有三个门，这三个们需要同一个函数进行激活，那就是 Sigmoid 激活函数，每个门都包含 Sigmoid 激活，这个函数与 Tanh 的区别在于，Sigmoid 将取值约束在[0,1]。这个特点可以帮助保留和移除信息。如果是 0，则移除；是 1，保留；中间数，能体现数据的重要性。

就留言文本分类来看，我们得到了大量留言数据，并且告诉了我们通过附件一建立一级标签分类，我们可以从附件一得到这些留言会被分成多少类，且这些留言只能被分到一级分类里的类中去。

对于具体的 LSTM 建模：

要建立一级标签分类模型，要先对文本进行处理，

统计数据总数、空值、各个类别的数据量

```
all objects: 9210.
在 cat 列中总共有 0 个空值.
在 review 列中总共有 0 个空值.
       cat  count
0     城乡建设   2009
1  劳动和社会保障   1969
2     教育文体   1589
3     商贸旅游   1215
4     环境保护    938
5     卫生计生    877
6     交通运输    613
```

全部数据量+空值统计+各类数据量

我们的数据有两个字段,其中 cat 字段表示类别,review 表示用户的评价信息,数据总量为 9210，并且没有空值

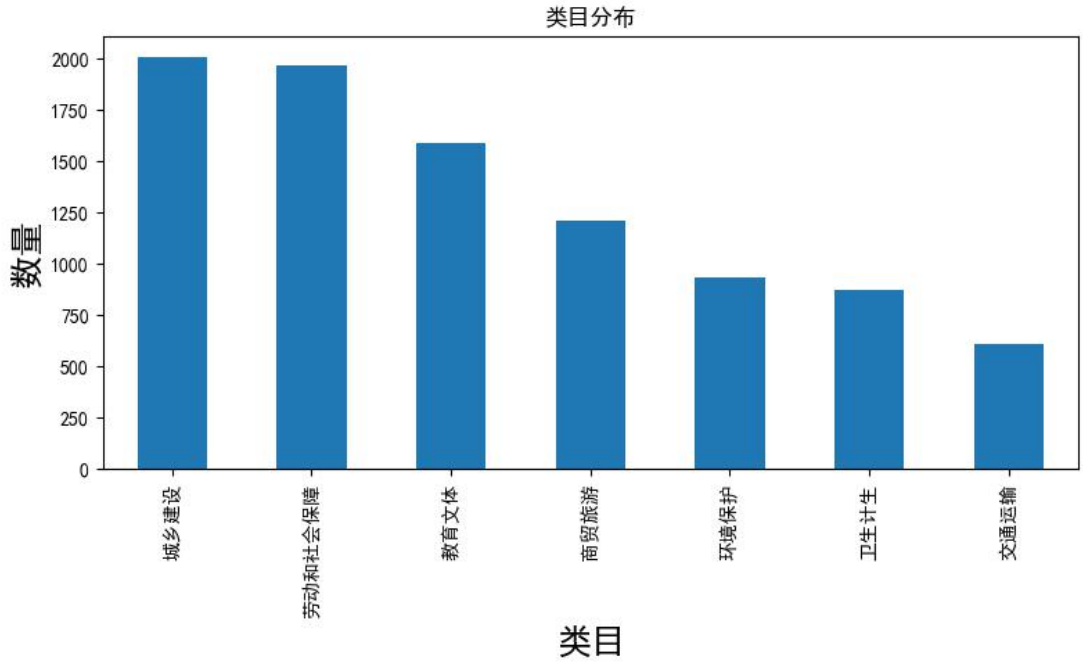
删除字母、数字、汉字以外的符号，去停用词

中文停用词包含了很多日常使用频率很高的常用词,如 吧,吗,呢,啥等一些感叹词等,这些高频常用词无法反应出文本的主要意思,所以要被过滤掉。我们过滤掉了 review 中的标点符号和一些特殊符号，并生成了一个新的字段。接下来我们要在新的字段的基础上进行分词,把每个评论内容分成由空格隔开的一个一个单独的词语。为了能

够提出更有用的数据。

	cat	...	cut_review
3391	交通运输	...	作为 一个 老百姓 我 觉得 滴滴 平台 是 个 好 平台 既 解决 了 大部分 失业 人员...
3265	交通运输	...	请问 市长 先生 M4 市 出租车 不 打表不去 乡镇 这 问题 要 什么 时候 才 解决 ...
3448	交通运输	...	公交车 所有 的 交通工具 车乱 停乱 靠 行人 没有 保障 所有 的 交通车 想 怎么 开...
7270	商贸旅游	...	目前 我市 主要 大 医院 医生 开出 的 药都 只有 该 医院 有 这种 垄断 式 的 销...
2093	环境保护	...	2018 年 12 月 2 日晚 10 点 30 分 B9 市 欧洲 城 小区 工地 施工 ...
2987	交通运输	...	嘉华 年 出租车 公司 发包 出租车 每台 18 万多 的 高价 而 其它 出租车 公司 只...
108	城乡建设	...	A1 区 北路 街道 10 号 地 副 10 号 地 7 号 地 的 安置 小区 的 卫生 ...
5744	劳动和社会保障	...	我 是 一名 下岗职工 现年 52 岁 2007 年 8 月 招聘 到 M 市 铁通 分公司...
6174	劳动和社会保障	...	贺 厅长 我们 是 临 G5 县 交通运输 局 局属 事业单位 县 运管处 县 县乡 公路 ...
3953	教育文体	...	尊敬 的 陈 局长 您好 我 是 1999 年 解聘 的 幼师 请问 您 教育局 对 199...

用图形化的方式再查看一下各个类别的分布。



Tokenizer 文本向量化

将文本或数据向量化，文本数据是非结构化的，自然语言无法处理他们，需要将文本转化为结构化数据，然后计算机才能使用，这样才可以继续后面的建模。

统计所有文本中不相同的词语个数

设置最频繁使用的词个数为 50000 每条处理后的数据最大长度为 250 Embeddingceng 层的维度为 100

```
[10 rows x 5 columns]
共有 84197 个不相同的词语.
(9210, 250)
(9210, 7)
(8289, 250) (8289, 7)
(921, 250) (921, 7)
```

多类标签 onehot 展开

one-hot 是比较常用的文本特征特征提取的方法。one-hot 编码，又称“独热编码”。其实就是用 N 位状态寄存器编码 N 个状态，每个状态都有独立的寄存器位，且这些寄存器位中只有一位有效。

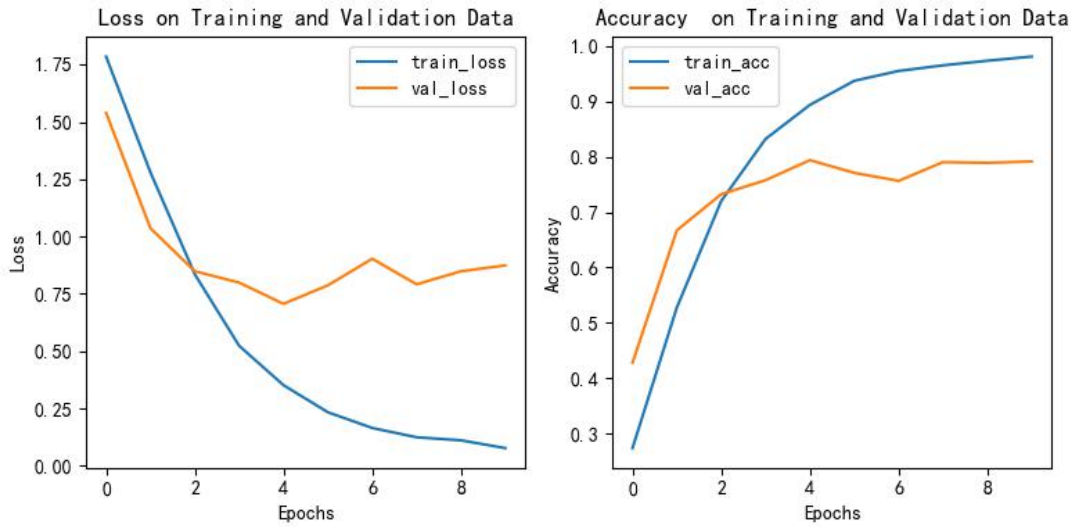
查看 model_summary

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 250, 100)          5000000
_________________________________________________________________
spatial_dropout1d_1 (Spatial (None, 250, 100)          0
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_1 (Dense)              (None, 7)                 707
=================================================================
Total params: 5,081,107
Trainable params: 5,081,107
Non-trainable params: 0
_________________________________________________________________
None
```

定义 LSTM 模型

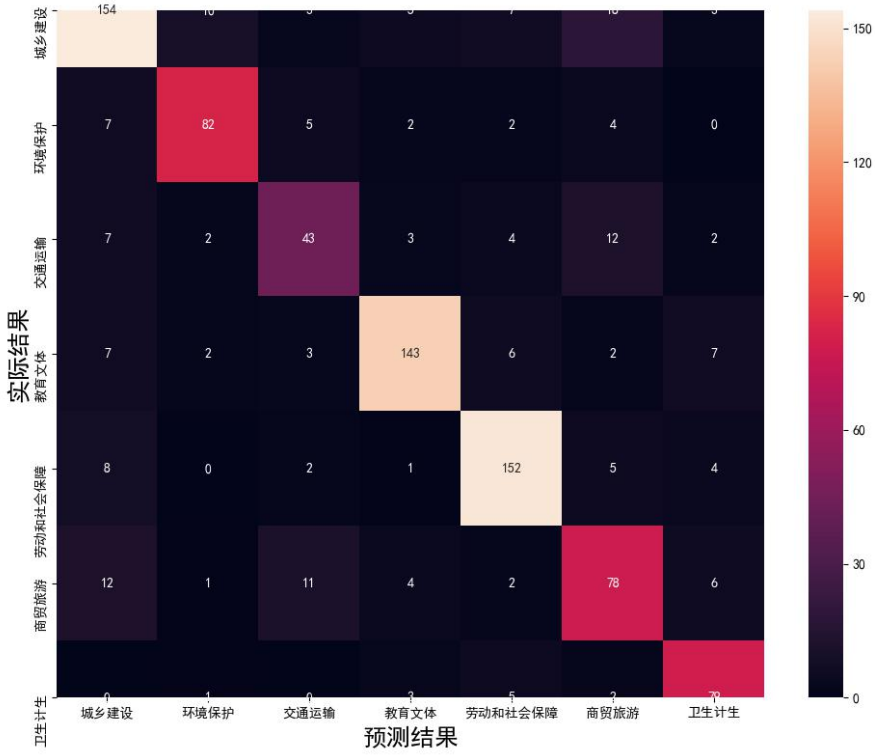
设置训练周期为 100 一次训练所选取的样本数为 64，然后开始训练数据

损失函数趋势图和准确率趋势图



通过画混淆矩阵和求 F1 分数来评估建立的分类模型。

混淆矩阵是对分类问题的预测结果的总结。使用计数值汇总正确和不正确预测的数量，并按每个类进行细分，这是混淆矩阵的关键所在。混淆矩阵显示了分类模型的在进行预测时会对哪一部分产生混淆。它不仅可以让您了解分类模型所犯的错误，更重要的是可以了解哪些错误类型正在发生。正是这种对结果的分解克服了仅使用分类准确率所带来的局限性。



混淆矩阵

	precision	recall	f1-score	support
城乡建设	0.79	0.77	0.78	200
环境保护	0.84	0.80	0.82	102
交通运输	0.64	0.59	0.61	73
教育文体	0.89	0.84	0.86	170
劳动和社会保障	0.85	0.88	0.87	172
商贸旅游	0.64	0.68	0.66	114
卫生计生	0.78	0.88	0.83	90
accuracy			0.79	921
macro avg	0.78	0.78	0.78	921
weighted avg	0.79	0.79	0.79	921

pre_recall_f1_support

2.2 朴素贝叶斯算法

首先留言信息的去重去空

在本次提供的留言信息中存在一些重复的留言数据，考虑到部门分配提高部门解决效率，需要取出重复的留言内容，因此利用函数将重复内容删除并将去重后的数据保存到 data_dup 里。同时对于得到的留言信息中有为空的记录，为了不影响之后的分析也需要去除。

然后对留言信息进行中文分词

在对留言进行一级分类之前要把非结构化文本信息转换成计算机能够识别的结构化信息，这里用到 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，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），同时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，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。之后将数据预处理函数进行封装，对词频进行统计后绘制词云图。词云图可以一目了然的看出哪些词语出现的频率高。

词云图见附件词云图。

TF-IDF 算法

在对民众留言信息进行分次后，需要把这些词转化成向量，以供挖掘分析使用，这才用 TF-IDF 算法，给每个词加权，通过统计词频增加权重，词频高的权重高。具体算法如下：TF-IDF）是一种用于信息检索与文本挖掘的常用加权技术。

TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

$$TF-IDF=TF*IDF$$

$$IDF(X)=\log\frac{N}{N(X)}$$

朴素贝叶斯建立模型

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，

$$P(Y \mid X) = \frac{P(Y) * P(X \mid Y)}{P(X)}$$

从给出的 496 条数据，七大类中在去重去空操作后每类抽取 33 条样本，利用这些样本进行分词，求 TF-IDF 算法以及朴素贝叶斯分类把样本分成了七类。

2.3 其他分类算法

2.3.1 Svm 算法

Svm 又称为支持向量机，是一种二分类的模型。当然如果进行修改之后也是可以
用于多类别问题的分类。支持向量机可以分为线性核非线性两大类。其主要思想
为找到空间中的一个更够将所有数据样本划开的超平面，并且使得本本集中所有
数据到这个超平面的距离最短。
优点：
分类思想很简单，就是将样本与决策面的间隔最大化
分类效果较好
缺点：
对大规模数据训练比较困难
在评价时得到平均数起伏过大（0.27）标准差起伏过大（两个数量级）

2.3.2GaussianNB 算法

高斯贝叶斯 ：适用于特征值符合正态分布的数据，不需要知道具体每个样本的
数值，只需知道样本符合什么样的正态分布（均值、方差）即可计算；
优点：即使条件独立性假设很难成立，但是朴素贝叶斯算法在实践中表现出乎意
料的好。该算法很容易实现并能随数据集的更新而扩展。在数据较少的情况下仍

然有效，可以处理多类别问题。
缺点：对于输入数据的准备方式较为敏感。
在评价时得到标准差起伏过大（两个数量级） 平均数起伏（0.16）

2.3.3 MultinomialNB 算法

多项式贝叶斯：不知道特征值符合哪种分布的时候，使用多项式贝叶斯算法计算
每个特征的概率，所以需要知道每个特征值的数值大小（**最常用于文本分类**）。
优点：

方法简单，分类准确率高

在接受大数据量训练和查询时速度快

缺点：

由于贝叶斯定理假设一个属性值对给定类的影响独立于其它属性的值，而此假设在实际情况中经常是不成立的，因此其分类准确率可能会下降，即无法处理基于特征组合所产生的变化结果

在评价时得到平均数起伏（0.28）

2.3.4 KNeighborsClassifier 算法

存在一个样本数据集合，也称作训练样本集，并且样本集中每个数据都存在标签，
即我们知道样本集中每一数据 与所属分类的对应关系。输人没有标签的新数据
后，将新数据的每个特征与样本集中数据对应的 特征进行比较，然后算法提取
样本集中特征最相似数据（最近邻）的分类标签。一般来说，我们 只选择样本
数据集中前 K 个最相似的数据，这就是 K-近邻算法中 K 的出处,通常 K 是不大于
20 的整数。 最后 ，选择 K 个最相似数据中出现次数最多的分类，作为新数据
的分类。
优点：

简单，易于理解，易于实现，无需估计参数。
训练时间为零。它没有显示的训练，不像其它有监督的算法会用训练集 train
一个模型（也就是拟合一个函数），然后验证集或测试集用该模型分类。KNN 只
是把样本保存起来，收到测试数据时再处理，所以 KNN 训练时间为零。
KNN 可以处理分类问题，同时天然可以处理多分类问题，适合对稀有事件进行分
类。
特别适合于多分类问题(multi-modal,对象具有多个类别标签)， KNN 比 SVM 的
表现要好。
KNN 还可以处理回归问题，也就是预测。
和朴素贝叶斯之类的算法比，对数据没有假设，准确度高，对异常点不敏感。
缺点：
计算量太大，尤其是特征数非常多的时候。每一个待分类文本都要计算它到全体
已知样本的距离，才能得到它的第 K 个最近邻点。
可理解性差，无法给出像决策树那样的规则。
是慵懒散学习方法，基本上不学习，导致预测时速度比起逻辑回归之类的算法慢。
样本不平衡的时候，对稀有类别的预测准确率低。当样本不平衡时，如一个类的
样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该
样本的 K 个邻居中大容量类的样本占多数。
对训练数据依赖度特别大，对训练数据的容错性太差。如果训练数据集中，有一

两个数据是错误的，刚刚好又在需要分类的数值的旁边，这样就会直接导致预测
的数据的不准确。
在评价时得到平均数起伏(0.18)　标准差起伏（一个数量级）

2.3.5 tree.DecisionTreeClassifier

分类决策树模型是一种描述对实例进行分类的树形结构. 决策树由结点和有向
边组成. 结点有两种类型: 内部结点和叶节点. 内部节点表示一个特征或属性,
叶节点表示一个类.
决策树(Decision Tree),又称为判定树, 是一种以树结构(包括二叉树和多树)
形式表达的预测分析模型.
优点:
速度快: 计算量相对较小, 且容易转化成分类规则. 只要沿着树根向下一直走到叶, 沿途的分裂条件就能够唯一确定一条分类的谓词.
准确性高: 挖掘出来的分类规则准确性高, 便于理解, 决策树可以清晰的显示
哪些字段比较重要, 即可以生成可以理解的规则.
可以处理连续和种类字段
不需要任何领域知识和参数假设
适合高维数据
缺点:
对于各类别样本数量不一致的数据, 信息增益偏向于那些更多数值的特征
容易过拟合
忽略属性之间的相关性
在评价时得到标准过大 与上方几个分类器的标准差不是一个数量级

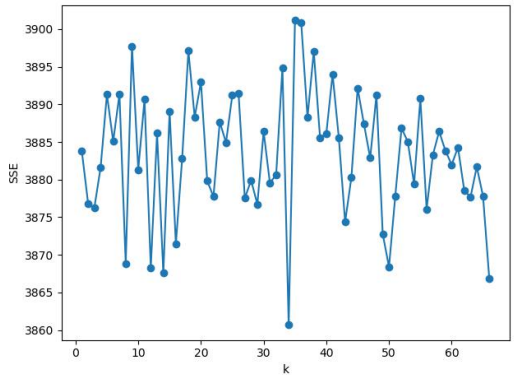
3.文本归类

3.1 k-means 算法

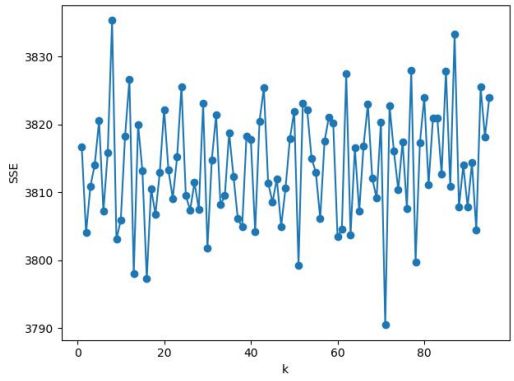
算法思想：

- 随机选取 k 个点，作为聚类中心；
- 计算每个点分别到 k 个聚类中心的聚类，然后将该点分到最近的聚类中心，这样就行成了 k 个簇；
- 再重新计算每个簇的质心（均值）；
- 重复以上 2~4 步，直到质心的位置不再发生变化或者达到设定的迭代次数

要对留言进行分类用到了 k-means 算法和 DBSCAN 算法，首先用到了 jieba 分词，把文本转化为 tf-idf 的特征矩阵。分别调用 sklearn 包的 kmeans 模型，DBSCAN 模型。对于 k-means 经过计算图片表示很难通过手肘法获取 k 值(质心数的取值根据经验法 2<=k<= 根号 n 由此确定 k 的范围　n 为对象的数量 即数据总数量,再根据经验法 k= （对象的数量除以 2 ）开平方　)根据给出的 4000 多条留言信息，由此设置质心数 67 和 96。



kmeans(SSE)67 类



kmeans(SSE)96 类

手肘法拐点的选取随着分类的类别数增加，SSE 的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓，手肘法即为选取那个拐点。上图过于跌宕起伏无法使用手肘法获得 k 值。所以这里不适用。

3.2 DBSCAN 算法

DBSCAN 是一种著名的密度聚类算法，它使用一组关于“邻域”的参数来描述样本分布的紧密程度。它的概念包括：

1.邻域：对于任意样本 i 和给定距离 e，样本 i 的 e 邻域是指所有与样本 i 距离不大于 e 的样本集合；

2.核心对象：若样本 i 的 e 邻域中至少包含 MinPts 个样本，则 i 是一个核心对象；

3.密度直达：若样本 j 在样本 i 的 e 邻域中，且 i 是核心对象，则称样本 j 由样本 i 密度直达；

4.密度可达：对于样本 i 和样本 j，如果存在样本序列 p1，p2，…，pn，其中 p1=i，pn=j，并且 pm 由 pm-1 密度直达，则称样本 i 与样本 j 密度可达；

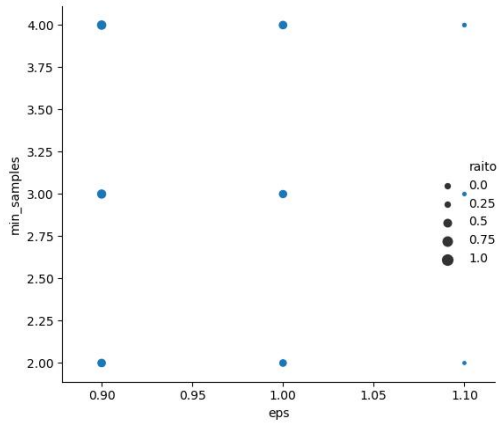
5.密度相连：对于样本 i 和样本 j，若存在样本 k 使得 i 与 j 均由 k 密度可达，则称 i 与 j 密度相连。

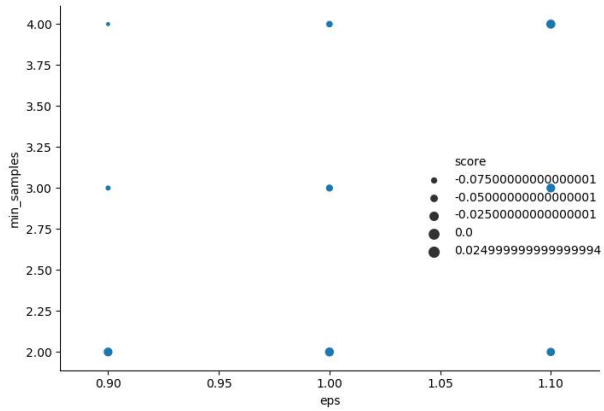
DBSCAN 通过轮廓系数，噪声比，分簇的数目进行评价。其中轮廓系数公式为：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

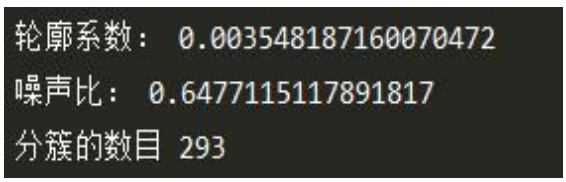
$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, a(i)<b(i)\\ 0, a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, a(i)>b(i)\end{cases}$$

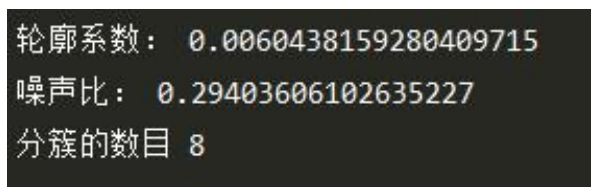
当我们将 4000 多条数据带入后得到，通过下面图片找到了 eps=1 min_samples=2 和 eps=1.1 min_samples=4





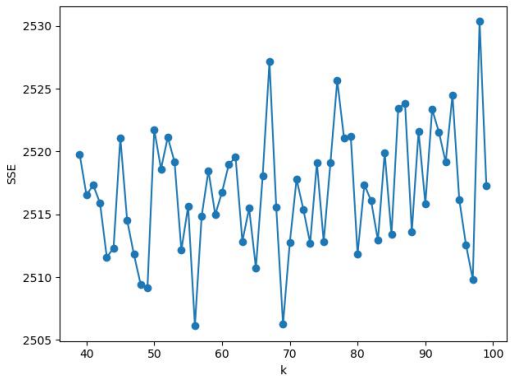
两种搭配比较二者的轮廓系数 噪声比 分簇的数目()选择了 eps=1 min_samples=2 导出数据





根据DBSCAN算法将4000多条留言数据分为了293类,其中有2000多条留言不属于这293类，他们都各自成为一类，总共分的3000多类，此次分类不太理想，进行了改进。

保留 293 类（每类至少包含两条以上留言），将剩余的 2000 多类只有再次进行归类，DBSCAN 结果不太理想，于是将他们进行了 k-means 聚类算法，得到了以下图



根据图示选择了下降幅度最大的 69，于是 K 值附为 69，将这次的分类与 DBSCAN 分得的 293 类进行合并得到了最终留言的分类。得到了热点问题留言明细表。

3.3 其他聚类方法

3.3.1 基于层次的方法

层次聚类主要有两种类型：合并的层次聚类和分裂的层次聚类。前者是一种自底向上的层次聚类算法，从最底层开始，每一次通过合并最相似的聚类来形成上一层次中的聚类，整个当全部数据点都合并到一个聚类的时候停止或者达到某个终止条件而结束，大部分层次聚类都是采用这种方法处理。后者是采用自顶向下的方法，从一个包含全部数据点的聚类开始，然后把根节点分裂为一些子聚类，每个子聚类再递归地继续往下分裂，直到出现只包含一个数据点的单节点聚类出现，即每个聚类中仅包含一个数据点。

它的流程

1.将每个对象看作一类，计算两两之间的最小距离；

2. 将距离最小的两个类合并成一个新类；

3. 重新计算新类与所有类之间的距离；

4. 重复 2、3，直到所有类最后合并成一类

3.3.2 基于网络的方法

基于网络的方法：这类方法的原理就是将数据空间划分为网格单元，将数据对象集映射到网格单元中，并计算每个单元的密度。根据预设的阈值判断每个网格单元是否为高密度单元，由邻近的稠密单元组形成”类“。

它的流程：

1、 划分网格

2、 使用网格单元内数据的统计信息对数据进行压缩表达

3、 基于这些统计信息判断高密度网格单元

4、 最后将相连的高密度网格单元识别为簇

4.留言热度问题

4.1 合理热度指标建立

对于热度评价，根据数据，文本分为两种类型

1.突发性 短时间内文本数激增的问题 （持续时间短 数量多）

2 持续型 短时间内解决不了的问题 （数量多 持续时间长）

题目中要求及时发现热点问题，所以本文认为突发型问题排序应更加靠前。本文认为，对问题观点具有的歧义越大，对热度影响也越大，对于给出的留言问题数据中能够体现歧义程度的就是点赞数和反对数，将点赞数与反对数相加得到歧义程度，当话题中意见倾向不统一的歧义越大，则此时对话题热度影响也越大。突发性要求持续时间短，数量多，所以公式里包含了最新一条数据距离当前时间的时间差和评论数量，由于时间差和评论数量差距过大，进行热度排序前已进行过对参数的归一化（将所有数据缩小到 0 到 1 之间）下面是热度计算公式：

热度计算公式 = 0.4 * 一类中总的赞成与反对的和+ 0.6 * ((评论数量+1) / (最新一条数据距离当前时间的时间差+1))

（公式中+1 是为了避免等于 0 的情况出现，干扰排序）

（4：6 的权重比例分配有利于突出突发型数据)

对每类文本分词，计算权值，排序，得出关键词，结合关键词和文本，人工整理出地点/人群以及问题描述两列，对每类问题进行热度计算，根据热度指数进行热度排序。结合起来得到热点问题表。

4.2 留言问题关键字提取

基于于统计特征的关键词抽取算法的思想是利用文档中词语的统计信息抽取文档的关键词。通常将文本经过预处理得到候选词语的集合，然后采用特征值量化的方式从候选集合中得到关键词。基于统计特征的关键词抽取方法的关键是采用什么样的特征值量化指标的方式，目前常用的有三类：

1 基于词权重的特征量化

2 基于词的文档位置的特征量化

3 基于词的关联信息的特征量化

对热点问题的留言详细进行的关键字提取是基于词权重的特征量化，基于词权重的特征量化主要包括词性、词频、逆向文档频率、相对词频、词长等。

4.3 答复留言的评价

相关性：选取的标准必须百与事项时相关的，相关的标准有助于得出结论，便于预期使用者作出决策。

完整性 ：答复的内容是否语句完整，对问题是否有完整的回答。

可解释性：指的是标准的可以获取和使预期使用者理解的。

答复意见无外乎分为几种

对解决问题的留言：

问题已经解决 解决结果是什么

问题正在解决 正在如何解决

问题没有解决 为什么没有解决

对咨询问题的留言：

是否所答为所问（就是说有没有对问题回答到点上）

答复的内容是否与留言问题沾边，就看答复问题的内容里重要词汇与留言详细里的词汇是否一致，完整性就是答复的全不全面，有没有解决留言里的所有问题，以及解决的程度，

5.建模前对文本进行处理

5.1 向量归一化处理

当不同的特征成列在一起的时候，由于特征本身表达方式的原因而导致在绝对数值上的小数据被大数据“吃掉”的情况，这个时候我们需要做的就是对抽取出来的 features vector 进行归一化处理，以保证每个特征被分类器平等对待。

在计算热度指标时，里面的两个变量时间和留言条数，由于他们之间差距太大，会对算出来的指标造成影响，所以将向量归一化，让他们数值都在 0-1 之间，计算出来的结果会减小误差。

5.2 文本向量化处理

文本向量化就是将文本表示成一系列能够表达文本语义的向量，是文本表示的一种重要方式。目前对文本向量化大部分的研究都是通过词向量化实现的，也有一部分研究者将句子作为文本处理的基本单元，于是产生了 doc2vec 和 str2vec 技术。

利用 world2vec 计算词语间的相似度有非常好的效果，word2vec 技术也可以用于计算句子或者其他长文本间的相似度，其一般做法是对文本分词后，提取其关键词，用词向量表示这些关键词，接着对关键词向量求平均或者将其拼接，最后利用词向量计算文本间的相似度。

在文本分类的 lstm 长短期记忆网络中就运用了 Tokenizer 文本向量化。

5.3 中文分词

基于词典的方法、基于统计的方法、基于规则的方法

本文两个模型都用了中文分词，用的是 jieba 分词中的精确模式，试图将句子最精确地切开，这是对数据的预处理，将大篇幅的文本按照一定逻辑分开，方便筛选出关键字对词进行加权等等操作。

6. 总结与展望

6.1 总结

在市民们平常的生活中对于自己的所看所听所想，他们想了解的需要政府帮忙解决的都可以通过网络平台让政府部门知晓。这种方便快捷的手段让政府部门每天都会收到成千上万的市民留言。原来的人工处理市民投来的反馈意见耗费人力资源多，处理起来也很慢，并且很难对口处理相关问题，于是运用计算机网络来处理这类问题是大势所趋。自然语言处理人工智能与计算机网络科学中值得研究的重要课题。

通过浅层传统的机器学习分析能够基本解决大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题，相较于人工来分类，速度快了许多并且减少了人工成本，能够将问题迅速分配到各个部门并进行相应的解决方案，它分类的准确度高,并行分布处理能力强,分布存储及学习能力强，能充分逼近复杂的非线性关系，具备联想记忆的功能等。LSTM 有预测功能，与序列和列表类型的数据密切相关，它拥有忘记门、输入门和输出门，可以更新信息，也可以忘记旧信息，如果更改了分类信息也可以用这个模型继续来对留言进行分类。

本文通过解决文本聚类分类以及热度排序，对答复进行回复的评价等问题，使用了 lstm 和朴素贝叶斯算法进行建模，并且还尝试了其他几种算法，测试后都能得出结果，但做的是比较传统的算法，并未进行创新，对于文本预处理方面进行了几项基本的处理，聚类方面将两种算法进行了对比，并且找到了最终聚类较好的方法，将两种算法结合，在热度排序上根据题目要求给出了相对合理的热度计算公式，最后根据之前所归类进行了最终的排序，根据关键字找出了留言的地点/人群，以及问题详细，最后一问进行了思路规划，但没找出方法来将它实现。

6.2 展望

在确定算法以及建模方面还比较生疏，还没有学习透彻，只学习到了皮毛，所以模型建立出来后精确度还有些偏差，用到的算法都还是传统的，没有进行太大改进，所以对于这种文本处理起来会有些误差，中文分词上随着词汇的更新，更多新奇词汇的诞生，需要根据实际情况像词库里添加一些词语，增加精确度，希望在基础知识学习透彻之后，能对人工智能以及这些自然语言算法有个更多自己的见解。

参考文献

[1] CSDN 基于 LSTM 的中文文本多分类实战 -派神-

[2] CSDN LSTM 特点及适用性 shincling

[3] CSDN 【个人整理】长短是记忆网络 LSTM 的原理以及缺点 LoveMIss-Y

[4] CSDN 文本聚类算法总结 小拳头

[5] CSDN 短文本聚类【DBSCAN】算法原理+Python 代码实现+聚类结果展示 jessie_weiqing

[6] 简书 集成学习聚类算法 DBSCAN 密度聚类算法详解和可视化调参 statr

[7] 在文章聚类中话题热度排序的研究与实现_张国锋

[8] 高校网络舆情话题热度趋势预测研究_李青

[9] 融合观点倾向的话题热度趋势建模研究

[10] CSDN Jieba 分词简介 韩明宇

[11] CSDN 中文分词算法总结 mandagod

[12] CSDN 文本向量化 refresh&grow

[13] CSDN 归一化处理方法 Danker01

[14] CSDN 各种聚类算法的系统介绍和比较 abc200941410128