基于朴素贝叶斯模型的“智慧政务”中的文本分类与处理

摘要：随着互联网信息时代的不断发展以及网民数量不断增长，网络信息量也呈现爆炸式地增长，我们要获取有效的信息，这必然给网络文本的处理带来了诸多困难。于是文本分类技术在提高文本处理效率上显得格外重要，本文利用 Python 语言，对互联网留言意见进行分级标题分类以及处理留言热点问题的挖掘问题。

问题 1，对全部留言信息数据进行分类。选取朴素贝叶斯模型，首先对表格留言数据中进行中文文本预处理；再对数据文本进行假设并结合 TF-IDF 进行特征提取；最后引入朴素贝叶斯模型的发哦一个合理的对于留言内容的一级标签分类模型。

问题 2，热点问题的挖掘，基于 LDA 主题模型,对内容进行分类聚类，即把相似主题的留言放在一起：首先对留言数据预处理，将处理后的留言数据导入；然后通过代码计算确定五个主题-词语“A 市噪音扰民”、“A 市街道问题”、“A 市咨询问题”、“A 市地铁问题”、“A 市街道问题”--这就是五个热点问题主题；最后基于 LDA 模型进行问题热度指数计算，将每个主题对应的热点问题留言信息给出表格。

关键词： 文本分类、朴素贝叶斯、TF-IDF、LDA 主题模型

目录

1 问题重述

1.1 问题背景

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、聚集民气的重要渠道，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市政效率具有极大推动作用。

附件中给出了收集自互联网来源的群众问政留言记录，及相关部门对群众留言的答复意见。请利用自然语言处理和文本挖掘的方法解决一系列问题。

1.2 解决问题

1.2.1 问题一 群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照附件 1 中的三级标签划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，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，且差错率比较高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。并且通过 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

1.2.2 问题二

某一时段群众集中反映的某一问题可称为热点问题，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。请根据附件 3 将某一时段内反映特定地点或特定人群问题地留言进行归类，定义合理的热度评价指标，并给出评价结果，按照表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

表 1-热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	…	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店有眼噪音扰民
2	2	…	2017/06/04 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
…	…	…	…	…	…

表 2-热点问题留言明细表

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	360104	A012417	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/08/18 14:44:00	A 市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天到凌晨还在营业……	0	0
1	360105	A120356	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，噪声严重扰民，有很大的油烟味往楼上窜，没办法居住……	1	0
1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
…	…	…	…	…	…	…	…
			魅力之城小		您好：我是魅力之城小区的业主，小		

2 群众留言分类

2.1 文本分类背景介绍

随着互联网信息时代的不断发展以及网民数量不断增长，产生的网络信息量也呈指数级增长趋势[1]。根据 2020 年 4 月中国互联网信息中心（CNNIC）发布第 45 次《中国互联网发展状况的统计报告》的数据显示，截止 2020 年 3 月，我国网民规模已达 9.04 亿，手机上网比例达 99.3%。日常获取各类信息的渠道不再只有传统报纸、杂志，这些信息来源渠道变得多种多样，例如网民手机端的各类广告短信、各微信公众号等自媒体发布平台、各微博官方发布等。处在高速发展

的网络时代，巨大的信息容量、丰富多样的信息来源渠道、查询速度以及更新速度快等一系列特点使文本信息达到了空前的规模，文本形式的信息使得用户获得消息变得更加便利、可以看到来自世界各地各色各样的新闻娱乐科技信息[1]。与此同时，面对巨大规模的网络信息，快速精准找到对我们有用的信息内容也显得较为困难。

近年来，随着微信、微博、市长信箱、阳关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同样不断攀升，给政府信息处理人员带来了诸多操作困难。因此，文本分类技术的出现及发展显得格外重要，在处理大量的群众意见留言的分类问题，就是建立一种文本分类模型。文本分类技术是一项实用价值极大、领域较为广泛的技术[3]。文本分类技术在国内外的出现也有较长一段时间了，其也得到了深入的研究发展和较为广泛的应用，在众多研究专家的努力下，文本分类已经形成了比较成熟的体系。

文本分类就是依照特征或内容，根据拟定的分类系统，将需要划分的文本分配到一个或者多个的先前已经定义好的分类中的一种分类方法[4]。而分类方法大致可分为两个类别：基于规则的分类方法以及基于机器学习的分类方法。

前者的代表分类方法就是决策树方法[4]，决策树算法的主要目的是构造精度较高、且规模较小的决策树。后者基于机器学习的分类方法则是通过学习给出的训练集，进而总结归纳除各分类的模板，从而使用这些模板对文本进行分类。这类方法的优点就是简单易懂可行性较高，且通常精度较高。

2.2 基于朴素贝叶斯分类法文本分类

2.2.1 朴素贝叶斯模型

发展到 20 世纪 90 年代，基于机器学习文本分类方法逐渐显示出其优越性，其不需要先验知识、精度达到专家级别、极高的工作效率等特性使其在领域内得

到更多的应用。基于机器学习的分类方法有 KNN 分类方法[9]、神经网算法[2][10]、支持向量机（SVM）方法[10]、朴素贝叶斯分类器[1]等，在这些方法中，朴素贝叶斯又具有许多优势，主要包含有独到的不确定性知识的形式表达、结合先验知识的学习特性、丰富的概率表达能力、以及其有效性和简洁性[1]。本题的留言文本分类也是在基于朴素贝叶斯模型上完成。

朴素贝叶斯算法是一类均以贝叶斯公式为基础的一类分类算法的总称，统称为贝叶斯算法。朴素贝叶斯的原理极其简单且容易实现，常用于垃圾邮件的分类问题，1000 封垃圾邮件能够被过滤掉 995 封，且无一漏判[5] 。首先，条件概率贝叶斯公式：

$$P(Y \mid X)=\frac{P(Y)P(X \mid Y)}{P(X)} \cdots\cdots\cdots\cdots\cdots(1)$$

其中引入 Y 的先验概率 $P(\mathrm{Y})$；

Y 的后验概率 $P(\mathrm{Y}|\mathrm{X})$；

联合概率为 $P(\mathrm{Y},\mathrm{X})$。且推出另一公式：

$$P(Y,X)=P(Y \mid X)P(X)=P(X \mid Y)P(Y) \cdots\cdots\cdots\cdots（2）$$

贝叶斯算法的核心思想是在假设各个特征相互独立的前提下，首先计算每一个类别的先验概率 $P(Y_i);i=1,2,3...$；再通过贝叶斯公式计算求出各个特征属于某一个类别的后验概率；最后选出具有最大后验概率估计值的类别即为最终类别。

这里我们将 X 理解成“具有某种特征”，把 Y 理解成“分类的标签”或者说“属于一类”，则在分类问题中就有：

$$P(\text{“属于某一类”}|\text{“具有某种特征”})=\frac{P(\text{“具有某种特征”}|\text{“属于某一类”})P(\text{“属于某一类”})}{P(\text{“具有某种特征”})}$$

……(3)

结合题设把模型放到留言中，例如，

$$P(\text{“属于城乡建设”}|\text{“A市西湖建筑集团占道施工有安全隐患”})=\frac{P(\text{“A市西湖?”}|\text{“属于城乡建设”})P(\text{“属于城乡建设”})}{P(\text{“A市西湖建筑集团”})}$$

……(4)

即先验概率就成了：

$P(\text{“属于城乡建设”})$；

后验概率就成了：

$P(\text{“属于城乡建设”}|\text{“A市西湖建筑集团占道施工有安全隐患”})$；

贝叶斯方法就是把计算具有“某种特征的条件下属于某类”的条件概率转换成“属于某类的条件下具有某种特征”的条件概率，计算后者的方法就简单许多，提供含有已知特征标签的样本作为训练样本即可。

2.2.2 TF-IDF 与文本预处理简介

进行数据分析时，文本预处理占据了一大半时间。主要步骤是首先把一句话拆分成更细粒度的词语表示[5] ；再去除留言文本的非中文部分、标点符号、数字、停用词等无关成分，接着进行 TF-IDF 特征处理。

TF-IDF 是一种用来评估某个字词对于一个文件集或一个语料库的其中一份文件的重要程度的一种统计方法[6]。其主要思想为：某个单词再一篇文章出现的频率较高，且在其他文章中很少出现，则认为这个单词具有比较好的类别区分能力，适合用于文本分类。其中 TF 是词频（Term Frequency），表示词条或者说关键字出现在一段文本中的频率，既有表达式：

$$TF_w = \frac{\text{在某一类别中词条w出现的次数}}{\text{该类别中所有的词条数目}} \cdots\cdots\cdots\cdots(4)$$

IDF 是逆向文本频率（Inverse Document Frequency），某特定词语的 IDF 是将总文件数目除以含有该特定词语的总文件数所得的商取对数得到，即含有该特定词条的文档越少，IDF 越大，则该词条的类别区分能力就越好。为了避免出现词条不在语料库导致的分母为 0 的情况，分母为文档数+1，既有表达式：

$$IDF_w = \log\left(\frac{\text{语料库的文档总数}}{\text{含有词条w的文档数+1}}\right) \cdots\cdots\cdots\cdots(5)$$

某个词语在某一特定文件中为高词频，且在整个文件集合中为低文件词频，则可以产生高权重的 TF-IDF，即通过 TF-IDF 来过滤常用词，保留有区别作用的关键词。公式有：$TF-IDF = TF * IDF$; TF-IDF 算法简单易懂、容易实现，但同时也没有完美的算法，其过于简单的结构并没有考虑词语的语义信息，无法处理一词多义与一义多词的情况。

2.2.3 文本预处理具体操作:

由原本已知的测试数据，现根据测试数据及以上模型进行问题解决。

（1）中文分词：由于中文文本存在着一些特别点，所以在处理上也会有别于英文分词。分词也就是将一整句话拆分成更精细粒度的词语，例如给出附件 2 的一条留言 ：

“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。”

而通过中文分词，我们就可以得到：

（“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”、“极”、“多”、……、“文明”、“的”、“路段”）

我们可以将其理解成一个向量，其中每一个维度都表示特征词在文本中的特定位置的存在，通过将长句拆分成更小的单位在自然语言处理上时非常常见的。

（2）这里我们采用最常见的 jieba 分词：

```
>>data_cut = data.apply(jieba.lcut)
```

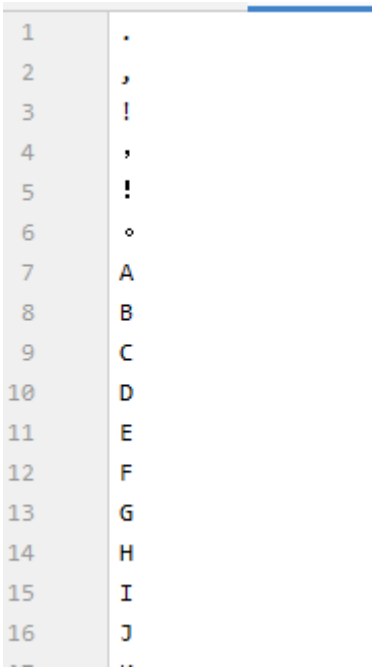
以下为部分分词后的代码：

图 1 分词后的部分文本数据

```
0              [A3, 区, 一米阳光, 婚纱, 艺术摄影, 是否, 合法, 纳税, 了, ？]
1       [咨询, A6, 区, 道路, 命名, 规划, 初步, 成果, 公示, 和, 城乡, 门牌,...
2          [反映, A7, 县, 春华, 镇金鼎村, 水泥路, 、, 自来水, 到户, 的, 问题]
3           [A2, 区, 黄兴路, 步行街, 大, 古道, 巷, 住户, 卫生间, 粪便, 外排]
4       [A, 市, A3, 区, 中海, 国际, 社区, 三期, 与, 四期, 中间, 空地, 夜...
```

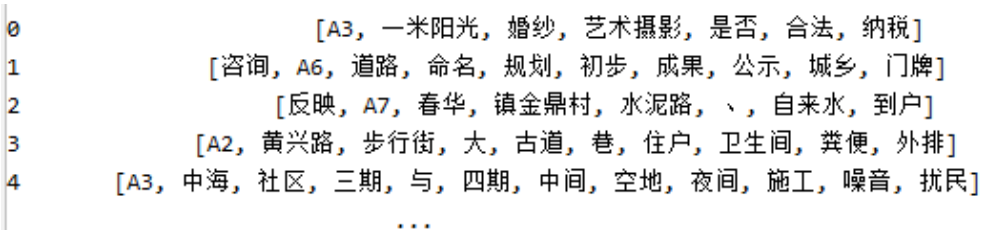
（3） 去留停用词 ：显然上面处理过后的文本中还存在许多无效的词，比如“？ ”、“A7”等，这些对于处理数据效率有影响的词就是停用词。这里，我们将停用词库 stopwords 直接导入到 py 函数中进行文本处理，图 2 为一部分的停用词表：

图 2 停用词示例



然后再对文本进行去留停用词处理,以下是上图 1 数据的去停用词之后的数据：

图 3 去停用词后部分文本数据

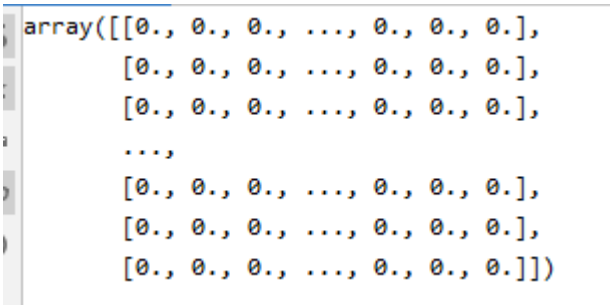


很明显，和只进行分词相比，在排除无用的语气词或者连接词之后，数据冗余大大降低，数据维度也降低了不少，减小了我们的计算难度，并且提高了计算效率。

（4） 特征提取与特征权重计算：在文本挖掘时，需要将文本表示出来，从复杂的非结构化形式转化成计算机能够识别、机器学习能够实现的结构化形式。文本的预处理已经在上文中有所叙述，提出用分词法将文本分词，然后又使用去停用词的方法留下大量对文本影响较大的词语，但若是直接使用该文本词语，则会造成特征向量的维数灾难，因此，这里，我们考虑用一些方法对现

有的去停用词的文本进行降维处理。而方法我们上文中已经提及，及是 TF-IDF 权值向量。图 4 为一部分代码结果：

图 4 原文本转换成的稀疏矩阵



```
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.]])
```

可见，原文本已经基于词袋转换成了稀疏矩阵。至此，我们的文本预处理过程结束，图 5 为以上文本预处理操作的代码实现。接下来，对预处理之后的文本进行模型预测与建立。

图 5 预处理代码实现

```
##导入库函数
from 文本向量化 import data_process_liuyan
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.naive_bayes import GaussianNB

#对要进行的数据进行预处理
adata,data_qutin,labels  = data_process_liuyan()
data_tr,data_te,labels_tr,labels_te = train_test_split(adata,labels,test_size=0.1)
data_tenew = data_te
countVectorizer = CountVectorizer()

data_tr = countVectorizer.fit_transform(data_tr)
X_tr = TfidfTransformer().fit_transform(data_tr.toarray()).toarray()

data_te = CountVectorizer(vocabulary= countVectorizer.vocabulary_).fit_transform(data_te)
X_te = TfidfTransformer().fit_transform(data_te.toarray()).toarray()
```

2.2.4 模型预测与建立

前面已经介绍了朴素贝叶斯模型的概念，这里进入文本分类的重点，采用朴素贝叶斯分类器，它会给留言分配使用特征词表示的类标签，且类标签取自附件

2 建立的资料库。

如图 6 代码对预处理过后的文本数据启用朴素贝叶斯模型所示

图 6 朴素贝叶斯模型实现代码

```
##启用朴素贝叶斯模型
model = GaussianNB()
model.fit(X_tr,labels_tr)
model.score(X_te,labels_te)
print('{}{}'.format(data_tenew,labels_te))
```

得到如图 7 所示的部分代码实现结果：

图 7 处理后的部分数据

```
4612                          恳请 西地省 教育厅 责成 西地省 工业 职业 技术...
5832                          以前 是 L9 某 单位 聘用制 员工 （ 临时工 ...
7279    垄断 势力 与 黑势力 悄悄地 走进 西地省 B7 老百姓 身边 \r\n B7 红砖 协会...
5958                          胡 厅长 ： 你好 1 我原 是 国企 一名 工程 ...
3898    是 一个 失独 农村 老人 失去 独生子 没有 赡养人 没有 监护人 住院 没有 护理 人 ...
                                ...
1822    G1 盛唐 四月 天 天天 停电 业主 正当 维权 既 没 阻碍交通 也 没 闯 市政府 房...
4307                          您好 作为 一个 高三 即将 参加 高考 学生 家长...
34      易 书记 \r\n 您好 \r\n 冒昧 打扰 烦请 见谅 敝 人 作为 您 子民 作为 7...
3880                          张 主任 您好 妈妈 是 农村 父亲 2003 年 ...
5348                          根据 西地省 行政 程序 规定 有关 “ 规范性 文...
Name: message, Length: 906, dtype: object4612       教育文体
5832    劳动和社会保障
7279       商贸旅游
5958    劳动和社会保障
3898       卫生计生
```

由此可见分类效果比较理想，并且我们可以得到具体的分类效果值如图 8：

图 8 分类效果值

```
>>> model.score(X_te,labels_te)
0.6986754966887417
```

可见，分配的准确值达到了 70%。

2.3 结果评价与分析

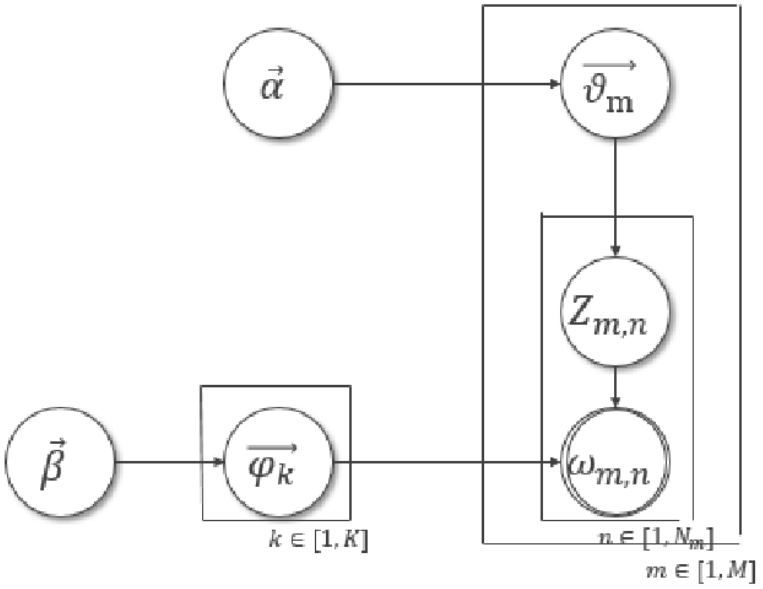
使用朴素贝叶斯模型得到的结果具有较强的依耐性，依赖于各个变量之间相互独立，及视为每个词语之间不会存在交叉含义，这是很难实现的。但是由于本文中是将文本内容作为分类标准，因此，各个变量之间独立的概率会增大，因此，这个地方可以使用高斯朴素贝叶斯分类器。

3 热点问题挖掘

3.1 LDA 主题模型

LDA 主题模型是一个概率增长模型，它可以用来模拟离散数据。LDA 主题模型的基本思想是，该模型采用的是包字法，它以每个文档为词频向量，文本信息生成为数字信息，不考虑单词之间的原始顺序。每个文档可以被看作是一些主题的多项式分布，每个主题可以被看作是一些单词的多项式分布。LDA 包含文档层、主题层以及词汇层三层拓扑结构，故其又被称为三层贝叶斯概率模型。同时该模型层次结构之间的逻辑非常清晰，且每一层均由相应的随机变量或参数控制。LDA 主题模型如图 9 所示：

图 9 LDA 主题模型



参数 $\vec{\alpha},\vec{\beta}$ 定义了 LDA 模型中的文档层集，其中向量$\vec{\alpha}$ 描述了整个文本数据集中主题信息的分布情况，刻画了文本集中潜在主题间的相对强弱，矩阵$\vec{\beta}$ 描述了主题信息集合关于特征词分布的情况，表征的是潜在主题自身在文本集中的概率分布，β_{ij} 表示的是第 j 个特征词属于第 i 个潜在主题的概率。$\vec{\alpha},\vec{\beta}$ 为先验概率分布参数，一般事先给定。假如文档集中特征词的规模 N，潜在主题数为 K，则参数 α 为 K 维向量，参数 β 为 K*N 阶矩阵。

$\overrightarrow{\vartheta_m},\overrightarrow{\varphi_k}$ 定义了 LDA 模型的主题层，$\overrightarrow{\vartheta_m}$ 表示每篇文档的主题分布，是一个 K 维的向量，即向量$\overrightarrow{\vartheta_m}$ 表示第 m 篇文档分布，$\Theta=\{\vartheta_m\}_{m=1}^{M}$ 表示有 M 个文档主题分布。表示第 k 个主题的词分布，是一个长度为 V 的向量，$\overrightarrow{\varphi_k}$ 即表示第 k 个潜在主题各特征词的比例分布，$\Phi=\{\overrightarrow{\varphi_k}\}_{k=1}^{K}$ 表示有 K 个主题特征词分布。参数 W，Z 定义了 LDA 模型的特征词层，反映了文档集层特征词的概率分布。向量 W 表示每篇文档的特征词分布，Wm,n 是可观测变量，表示第 n 个词在第 m 篇文档中所占的比重。参数 Z 表示潜在主题的特征词分布，Zm,n 表示第 n 个潜在主题在第 m 篇文档中所分配的比重。热点问题挖掘的计算具体步骤如下：

3.2 热点问题挖掘主要步骤

3.2.1 准备工作

对作品附件中的表格“附件 3”进行文本预处理，首先除去数据中的非中文部分；接着对数据基于 jieba 进行中文分词；然后通过在作品附件中的记事本文件“停用此词表”引入停用词。最后，建立语料库并定义 dis 余弦相似度函数，导入预处理之后的留言数据。这里，我们使用的是上述提出的 LDA 主题模型。图 10 为准备工作具体代码：

图 10 LDA 代码实现

```
#########LDA主题模型
import pandas as pd
from gensim import models
import jieba
import re
from gensim.corpora import Dictionary
from gensim.models import LdaModel
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
import xlwt

data = pd.read_csv('附件3处理.csv',usecols = [1])
data.columns = ['message']
data_str = data['message']
data = list(data['message'].str.split(' '))


dictionary = Dictionary(data)# +建立语料库
#def distance(x1,x2):#定义dis余弦相似度函数

 #  l1 = len(x1)
  # l2 = len(x2)
  # if l1 > l2: ##规定L1为小
  #     temp = l2
  ##     l2 = l1
  #     l1 = temp
  # for i in range(l1):




  # return
```

3.2.2 确定 5 个主题-词语模型

由上文提到的 LDA 主题模型的超参数 β，这里我们人为的定义为 5；图 11 为确定主题-词语模型的具体代码，

图 11 确定主题-词语代码实现

```
bow = [dictionary.doc2bow(message) for message in data]              ########打印出词频和词语 准备的语料库
x=5                                    ##确定主题数目
data_hot = LdaModel(corpus = bow, id2word = dictionary, num_topics = x)
data_hot_0 = data_hot.print_topic(0)          ########提取出了主题
print('主题-词语：\t')
i=0
topic  = {}
jieba.load_userdict('jieba.txt')
for i in range(x):#####循环得到5个主题-词语模型
    #print(data_hot.print_topic(i))
    topic[i] = data_hot.print_topic(i)#主题-词语模型赋值
    print(topic[i])
    #distance(topic[i],topic[i+1])
    i += 1
```

图 12 为结果输出代码，由此可见主题-词语的输出结果。

图 12 输出结果

```
主题-词语：
0.029*"A7" + 0.015*"建议" + 0.011*"建设" + 0.007*"规划" + 0.007*"对" + 0.007*"经济学院" + 0.006*"路" + 0.006*"A5" +
 0.005*"被" + 0.005*"希望"
0.011*"A7" + 0.010*"A3" + 0.008*"投诉" + 0.007*"苑" + 0.007*"滨河" + 0.007*"销售" + 0.007*"违规" + 0.007*"房屋" +
 0.006*"溪湖" + 0.006*"A2"
0.019*"A7" + 0.012*"A2" + 0.008*"A3" + 0.008*"附近" + 0.008*"扰民" + 0.007*"A9" + 0.007*"A1" + 0.006*"路" + 0.006*"反映
 " + 0.006*"幼儿园"
0.015*"A7" + 0.011*"噪音" + 0.011*"A3" + 0.009*"扰民" + 0.009*"不" + 0.007*"A2" + 0.007*"1" + 0.006*"还" + 0.006*"投诉"
  + 0.006*"建议"
0.017*"A3" + 0.013*"咨询" + 0.010*"A7" + 0.010*"扰民" + 0.007*"A1" + 0.007*"业主" + 0.007*"A4" + 0.007*"A6" + 0.006*"社
>>>
```

3.2.3 问题热度指数计算

LDA 主题模型的输出包括主题－词语的概率分布和文档－主题的概率分布，所以我们可以根据文档－主题的概率分布来估算微博话题的强度。此处我们假设所给文本的文档－主题概率最大的值为该条微博博文所归属的主题。由此我们可以使用公式来表示一个主题中的某个热点问题的强度 $Strength(k)$

$$Strength(k)=\sum_{i=1}^{N}\theta_i^k$$

式中 N --语料库中所包含的文本总数量；

θ_i^k ：--每个微博文本归属于主题 k 的概率。

每篇文本 θ_i^k 计算原则为：（1）若该文档一主题概率最大的值，为我们正在计算的主题，那么 θ_i^k 的值为该条微博博文的文档一主题概率最大的值；（2）否则博文对应的 $\theta_i^k = 0$ 。

且将第一个热度问题的热度指数默认 10，剩下的热度指数在第一的热度指数上进行线性改变。在计算完对应的热度之后，根据题设条件代入问题 ID，并同步到 excel 表中，完成所需要的的 2 个表 ，最终分类部分结果如图 13，图 14：

图 13 部分热点问题留言明细表格

言用户	留言主题	留言时间	留言详情	点赞数	反对
)028571	国际社区三期与四期中间空地夜间	2019/11/22 16:54:42	情况，而是直接打电话给投诉业主，态度强硬恶劣充满无奈，同时表明	0	0
)084085	语桐梓坡路与麓松路交汇处地铁	2019/9/17 4:25	老人已经被折磨的不行，周边邻居也是苦不堪言，请相关部门协调，按照	0	0
)053484	园小区乐果果零食炒货公共通道摆	2019/5/31 17:06:13	社区青青家园小区乐果果零食炒货公共通道摆放空调/冰柜外机，持续大	0	0
)097934	市利保壹号公馆项目夜间噪声扰民	2019/7/3 6:23:25	晚上开始，到目前30日凌晨2点还在施工中，且噪声极大，接近有80-100	0	0
)048792	门至万芙路段经常有改装车飙车，	2019/6/18 23:03:31	而且还投诉无门，多次拨打A2区交警大队的电话（0000-00000000通过11	0	0
)09139	家丽南路丽发新城居民区附近搅拌	2019/11/19 18:07:54	居民区，开发商在小区旁50米处建搅拌站，运渣车吵得人精神崩溃，灰尘	0	1
)09204	A2区丽发新城附近建搅拌站噪音	2019-11-13 11:20:21	下，在离小区不到百米的地方建搅拌站。可想而知，一个大型搅拌站每天	0	0
)060165	地省师大附中的校园喇叭太扰民	2019/10/11 10:31:08	的最高标准，有时候一开就开得早，早上6点就开始，能持续几个小时，	0	0
)072847	新城违建搅拌站，彻夜施工扰民	2019/12/26 13:55:15	民区，灰尘，噪音污染严重；3、搅拌站几百米外就是小学，扬尘严重影	0	0

图 14 部分热点问题表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	10	2019/1/10 8:29:17到2020/1/6 11:29:11	A市市民	A市噪音扰民
2	2	7.4	2019/1/6 20:36:34到2020/1/6 20:45:34	A市市民	A市街道问题
3	3	6.2	2019/1/10 7:46:31至2020/1/5 1:22:17	A市市民	A市咨询问题
4	4	5.3	2019/1/3 10:34:29到2020/1/7 22:50:33	A市市民	A市地铁问题
5	5	4.3	2019/1/6 20:36:34到2020/1/6 20:45:34	A市市民	A市街道问题

4 结语

4.1 算法优缺点总结

朴素贝叶斯分类模型虽然算法简单易懂以及有着准确稳定的分类效果，但是文本的数据独立性假设条件是朴素贝叶斯分类器的不足，在大多数实际情况，不

能保证数据的独立性，数据之间往往是有一定联系的。所以当数据间的相关性较大时会导致分类效果的大打折扣。

TF-IDF 中采用$TF-IDF=TF*IDF$值的大小作为依据提取特征词，算法快速简单。但是由于 IDF 逆向文件频率简单的结构，不能够最好地反映该词条的重要程度，因此 TF-IDF 算法的精度在某些情况下并不是特别高。

4.2 模型算法改进

在问题一上，可加入逻辑回归与朴素贝叶斯形成对比，并加入一定的评价标准例如通过精准率和召回率的对比来测试分类效果，模拟出最合适的分类模型。同时，结合应用段落标注等技术，对处于不同位置的词语额外附加不同位置的权重，对 TF-IDF 算法进行优化[11][12]。

5 参考文献

[1] 何伟，基于朴素贝叶斯的文本分类研究算法[D]，南京邮电大学，2018 年（02），4-4.

[2] 张波 黄晓芳，基于 TF-IDF 的卷积神经网络新闻文本的分类优化[J]，西南科技大学学报，35（1）：，2020.

[3] 董露露，基于特征选择及 LDA 模型的中文文本分类研究与实现[D]，合肥：安徽大学，2014.

[4] 刘冬瑶 刘世杰 陈宇星 张文波 周振，新闻文本自动分类技术概述[J]，电脑知识与技术，13（35）:，2017.

[5] 龙心尘 寒小阳，NLP 系列(2)_用朴素贝叶斯进行文本分类(上)，

https://blog.csdn.net/longxinchen_ml/article/details/50597149，2020.4.27.

[6] Asia-Lee，TF-IDF 算法介绍及实现，

https://blog.csdn.net/asialee_bird/article/details/81486700（本文为博主原创文章，遵循 CC 4.0 BY-SA 版权协议），2020.4.28.

[7] 刘建平 Pinard，中文文本挖掘预处理流程总结，

https://www.cnblogs.com/pinard/p/6744056.html，2020.4.27

[8] 张萌，微博热点话题发现方法的研究和实现，北京交通大学，2018 年(06)

[9] 李宏志 李苡兰 赵生慧，基于 Spark 的大规模文本 KNN 并行分类算法，湖南科技大学学报(自然科学版)，2020 年(01):95-102

[10] 丛成;吕哲;高翔;王敏，基于支持向量机的钢板缺陷分类问题的研究，物联网技术，2020 年(04):39-41+46

[11] 隗中杰.文本分类中 TF-IDF 权重计算方法改进.软件导刊.2018 年(12):43-46

[12] 张瑾.基于改进 TF-IDF 算法的情报关键词提取方法.情报杂志.2014 年(04):157-159