基于数据挖掘的“智慧政务”文本挖掘应用研究

摘要：随着网络问政平台逐步发展，各类社情民意相关的文本数据量不断攀升，给人工处理信息的工作人员带来了巨大的挑战。为了帮助相关部门的工作人员解决困难，我们采用自然语言处理和文本挖掘等方法解决问题。在第一问中，我们首先针对群众留言内容对其进行分词、过滤停用词等数据处理，然后根据其对应的一级标签基于深度学习中的 LSTM 模型和 CNNtext 模型建立一级标签分类模型，最终准确率达到 90%左右。在第二问中，我们采用 TF-IDF 方法，将数据预处理后的群众留言进行相似度计算，通过热度指数公式对热点进行排行，将排在前五的热点和热点留言明细进行展示并储存。在第三问中，我们利用 python 中 gensim 库构建词典，通过 doc2bow 稀疏向量生成语料库，利用 TF-IDF 模型算法求出回复内容与留言详情的相似度，并利用相关性、完整性进行解释说明。从最终结果来看，虽然存在一定误差，但具有一定的实际意义，有较大的推广性。

关键词：自然语言处理；深度学习； LSTM；Gensim；TF-IDF；CNNtext

1. 引言

自然语言指的是日常使用的语言，如汉语、英语等。而自然语言处理旨在研制可对人类口头或书面语言做出非预设反应，可对外部信息进行有效通信交流的计算机系统或软件系统。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分支，随着其技术应用范围不断扩大，在数据处理领域占有越来越重要的地位。

如今，我们正身处一个数据爆炸的时代，越来越多的群众利用微信、微博、市长信箱、阳光热线网络问政平台向有关部门提供意见。然而，目前留言划分和热点整理仍然以人工逐一阅读的方式来获取信息。这种人工方式不仅工作量大、效率低，而且在面对大规模的非结构化自由文本时，人工方式将显得无能为力。因此，为了能有效地对民众意见进行重要性划分，精准提取热点问题，使有关部门能及时回复并采取相关措施，从而提升政府的管理水平和施政效率，我们采用了自然语言处理的相关技术，将其在文本挖掘、情感分析等领域的一些有效方法运用于标签划分、热点问题提取以及文本评价信息研究之中，弥补人工分析的不足，大大减少人工分析的工作量。

2. 相关技术及工具

2.1 Pkuseg 分词

Pkuseg 是一个多领域中文分词工具包。选择该分词工具包是因为其三大优点：支持多领域分词，支持用户自训练模型，分词准确率较高。Pkuseg 目前支持新闻领域，网络领域，医药领域，旅游领域，以及混合领域的分词预训练模型。Pkuseg 是基于经典的 CRF 模型，辅以 ADF 训练方法和精调的特征，实现更快的训练速度、更高的测试效果和更好的泛化能力。

通过与 jieba 中文分词工具包相比较明显可以看出 jieba 的分词效果会将一些地名小区名拆分开来，并且一些固定的词组也被分开，例如：占道被分为占和道；烧烤摊被分为烧烤和摊，而这个情况在 Pkuseg 分词中得到了改善。

pkuseg 具有如下几个特点：

1) 多领域分词。不同于以往的通用中文分词工具，此工具包同时致力于为不同领域的数据提供个性化的预训练模型。根据待分词文本的领域特点，用户可以自由地选择不同的模型。
2) 更高的分词准确率。相比于其他的分词工具包，当使用相同的训练数据和测试数据，pkuseg 可以取得更高的分词准确率。
3) 支持用户自训练模型。支持用户使用全新的标注数据进行训练。

2.2 LSTM

2.2.1 LSTM 介绍

长短期记忆(LSMT)是一种特殊的 RNN，主要为了解决长序列训练过程中的梯度消失和梯度爆炸问题，相比普通的 RNN，LSTM 能够在更长的序列中有更好的表现。LSTM 有两个传输状态，一个c^t(cell state)和一个h^t(hidden state)。其主要输入输出结构图如下图 1 所示：

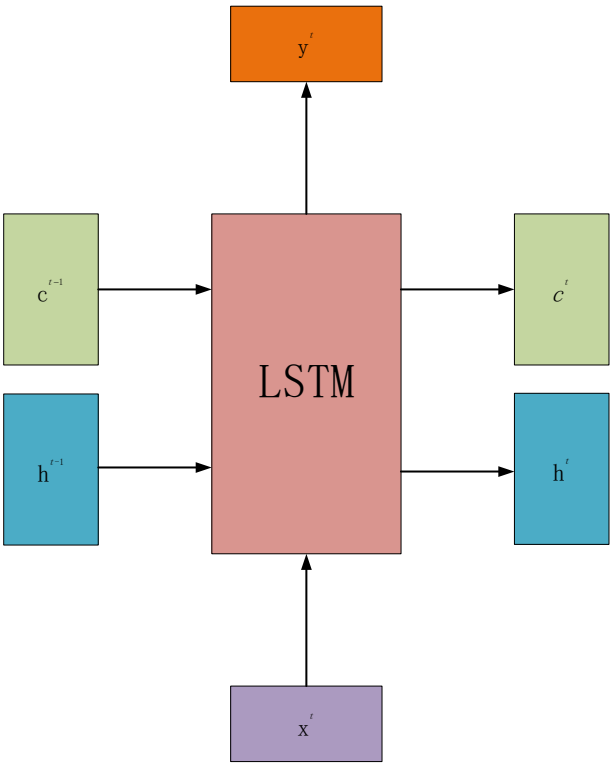

图 1 LSTM 输入输出结构图

2.2.2 深入 LSTM 结构

使用 LSTM 的当前输入x^t和上一个状态传递下来的h^{t-1}拼接训练得到四个状态，其中z^f, z^i, z^o是由拼接向量乘以权重矩阵后，再通过一个 sigmoid 激活函数转换成 0 到 1 之间的数值来作为一种门控状态。而 z 则是将结果通过一个 tanh 激活函数转换成-1 到 1 之间的值。z, z^f, z^i, z^o这四个状态在 LSTM 内部主要由三个阶段：

1) 忘记阶段。这个阶段主要是对上一个节点传进来的输入进行选择性忘记。简单来说就是会 “忘记不重要的，记住重要的”。具体来说是通过计算得到的z^f(f 表示 forget)来作为忘记门控，来控制上一个状态的c^{t-1}哪些需要留哪些需要忘；
2) 选择记忆阶段。这个阶段将这个阶段的输入有选择性地进行“记忆”。主要是会对输入x^t进行选择记忆。当前的输入内容由前面计算得到的 z 表示。而选择的门控信号则是由z^i(i 代表 information)来进行控制；
3) 输出阶段。这个阶段将决定哪些将会被当成当前状态的输出。主要是通过z^o来进行控制的。并且还对上一阶段得到的c^o进行了放缩。

2.3 CNNtext 模型结构

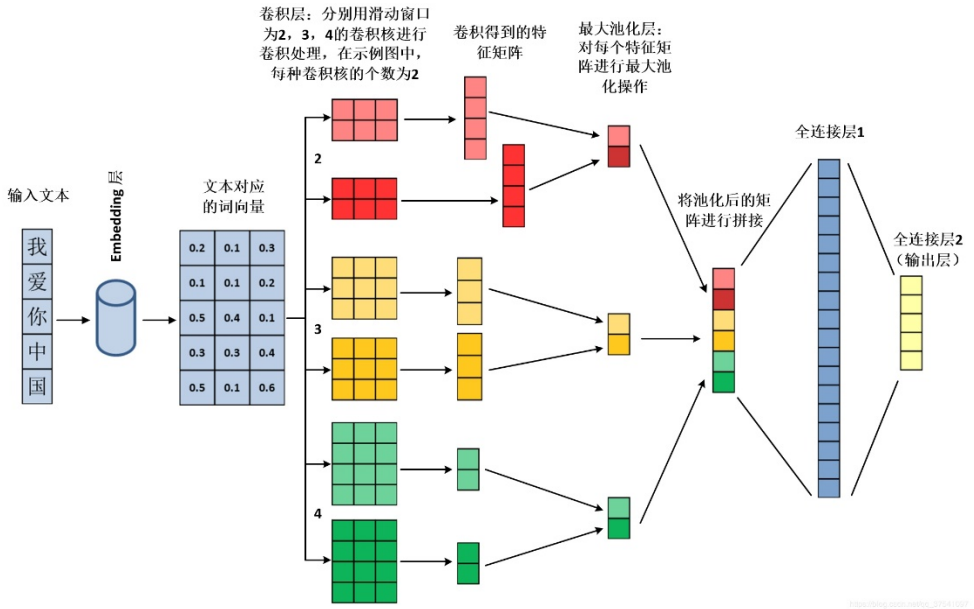

图 2 CNNText 模型

CNNtext 模型结构如图 2，主要流程分为以下几步：

1) 将中文文本通过 embedding 层转换为词向量，图中词向量以三维为例（例如“我”对应的是[0.2, 0.1, 0.3]词向量），在本文代码中采用的是 64 维。（本图中省略了将中文文本转换为对应词 id 的过程）

2) 通过不同的滑动窗口进行卷积处理，图中以滑动窗口分别为 2、3、4 为例，并且每种滑动窗口的卷积核个数为 2，实际使用过程中每种滑动窗口的卷积核个数可自己设定，本文代码中采用的是 64。

3) 对卷积操作生成的特征矩阵使用最大池化处理。

4) 将池化后的特征矩阵进行拼接。

5) 将特征矩阵进行扁平化或压缩维度，图中未绘制。

6) 连接全连接层 1

7) 连接全连接层 2，通过 activation=softmax 输出每个类别的概率

2.4 Gensim

Gensim 是开源的第三方 Python 工具包，用于从原始的非结构化的文本中，无监督地学习到文本隐层的主题向量表达。核心概念为文件、语料库、向量和模型。

训练语料的预处理是将文档中原始的字符文本转换成 Gensim 模型所能理解的稀疏向量

的过程。创建 Gensim corpora.Dictionary 对象，为语料库中出现的所有单词分配了唯一的整数 id。corpora.Dictionary 对象中 doc2bow()方法仅计算每个不同单词的出现次数，将该单词转换为其整数单词 id，然后将结果作为稀疏向量返回。similarities 相似度接口用于计算文档对之间的相似度或者一篇特定文档和其他文档之间的相似度。

2.5 术语频率*反向文档频率，Tf-Idf

术语频率*反向文档频率(Tf-Idf)是一种流行的向量空间模型。它是一种用于资讯检索与资讯探勘的常用加权技术，由 TF 和 IDF 组成。

TF 表示词频，统计文本中每个词出现的频率。其表达公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

其中分子表示该词在文件中的出现次数，分母为文件中所有字词出现的次数总和。

IDF 表示逆文本频率，返回一个词在语料库中所有文本中出现的频率，反映词语在文本中的重要性。如果一个词在多个文本中出现，它的 IDF 值反而越低；反过来，若一个词在较少的文本中出现，它的 IDF 值反而越高。其表达公式如下：

$$IDF(x) = log\frac{N}{N(x)} \tag{3}$$

其中 N 表示语料库中文本数量，N(x)表示含有词 x 的文本数。如果一个词在语料库中不出现，那么 N(x)则为 0，而分母不能为 0，出现计算错误。为了防止出现这种分母为 0 的现象，最常用的方法是使用拉普拉斯平滑对上述公式进行处理，进行平滑处理后的公式如下：

$$DF(x) = log\frac{N+1}{N(x)+1} + 1 \tag{4}$$

根据公式(2)与公式(4)，我们就得到 TF-IDF 的计算公式：

$$TF - IDF = TF(x) \times IDF(x) \tag{5}$$

3. “智慧政务”文本挖掘具体应用

3.1 群众留言分类

3.1.1 数据预处理

3.1.1.1 ID 变换及字符过滤

原始数据集中所包含的信息有留言编号、留言用户、留言主题、留言时间、留言详情以及一级标签六个字段，所示数据如下图 3：

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
24	A00074011	A市西湖建筑集团占道施工有安全隐患	2020/1/6 12:09	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。	城乡建设

图 3 附件 2 原始数据

由于原始数据中的一级标签字段为文字类型，因此首先要将其转换成数字形式的 id，最终得到的 ID 形式如下图 4 所示：

	一级标签	label_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

图 4 标签 ID 数据

紧接着我们可以发现，由于我们的留言内容都是中文，而数据中存在许多“,”、“。”、“\n”、“\t”等非必要字符，因此需要将其利用正则表达式过滤掉。而由于此题我们针对的是留言详情建模，因此在过滤字符时，只针对留言详情进行处理，最终得到的结果如下图 5 所示：

clean_review
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...

图 5 字符过滤后的数据展示图

3.1.1.2 分词及过滤停用词

如上图 3 所示，去掉无意义的字符之后，所有中文文字都连成了一个句子，无法提炼关键信息，因此，我们采用了 Python 中文分词组件包结巴“jieba”分词将其切分成多个词语。由于分词之后发现其中包含了很多日常使用频率很高的常用词，如“吧”、“吗”、“呢”、“的”等一系列无法反映出文本的主要意思的词，因此为了减少计算复杂度，我们利用网络上下载的中文停用词库将这些无意义的词语过滤掉，最终得到的结果如下图 6 所示：

clean_review	cut_review
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...	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...
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...	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏...
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...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公...
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...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉...
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...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉...

图 6 分词及去停后数据展示图

3.1.2 LSTM 建模

数据预处理完成以后，接下来便是进行 LSTM 建模工作。

1) 第一步我们将分词去停之后的数据(cut_review)进行向量化处理，将每条数据转换成一个整数序列的向量；

2) 第二步设置最频繁使用的词的个数为 50000；

3) 第三步设置每条数据最大的词语数为 250 个，若超过此设置量则将会被截去，不足将会被补 0；

4) 第四步将数据进行训练集和测试集的拆分；

5) 第五步定义一个 LSTM 的序列模型。模型的第一层是嵌入层(Embedding)，第二层是 SpatialDropout1D，第三层为 LSMT 层，最后是输出层。由于是多分类，因此将激活函数设置为“softmax”，顺势函数设为分类交叉熵“categorical_crossentropy”；

6) 第六步利用定义好的 LSTM 模型训练数据，设置训练周期数为 5，batch_size 为 64。

经过上述 6 个步骤之后，训练得到的训练集准确率达到 97%，测试集准确率达到 85%。

紧接着，我们通过画混淆矩阵来评估模型的表现，混淆矩阵如下图 7 所示：

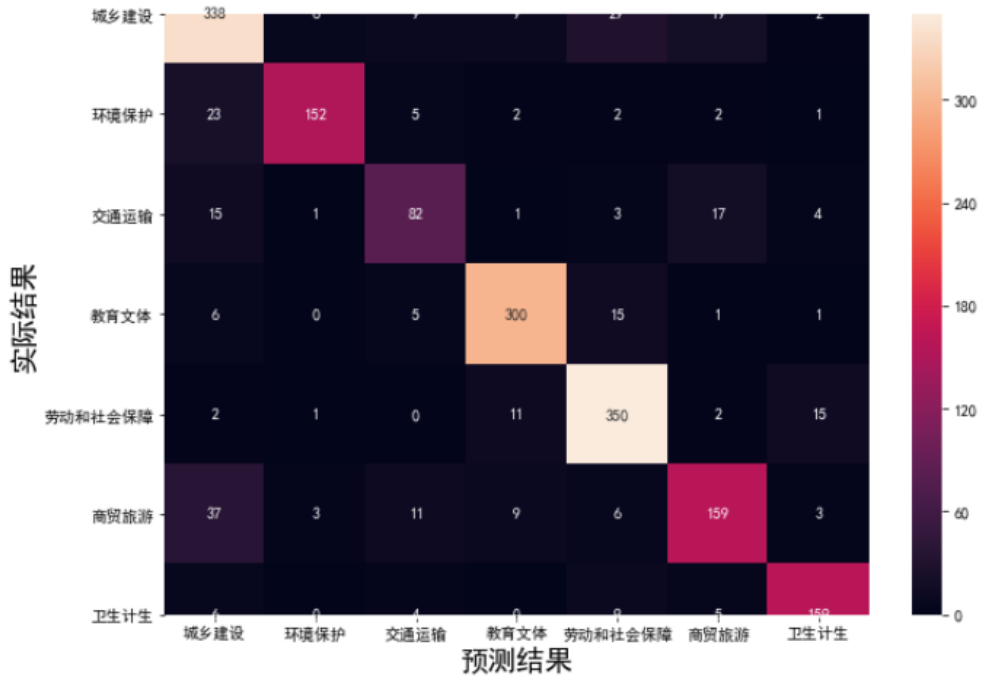

图 7 混淆矩阵展示图

由图 5 所示的混淆矩阵可以发现“环境保护”类的预测准确率较好，而“城乡建设”类预测的错误数量相对较多。

3.1.3 CNNtext 建模

数据预处理方式与前面类似，不再赘述，接下来构建 CNNC 模型

1) 定义输入层，输入向量维度与文字序列长度相同（seq_lenght = 600）
2) 嵌入层，将词汇的 one-hot 编码转为词向量
3) 构建长度为 3，4，5 的词窗口的卷积层，进行最大池化处理
4) 合并三个经过卷积和池化后的输出向量，再进行扁平化
5) 构建全连接层 1 和 2，在之间添加 dropout 减少训练过程中的过拟合，其中全连接层 2 为输出层
6) 编译模型，其中优化器选择为‘adam’，loss 计算方法为’categorical_crossentropy’
7) 模型搭建成功后，对模型进行训练，其中训练时传入的训练集 batch_size 为 64，验证集 batch_size 为 200，每步的训练次数为 1000，训练次数为 2。

通过训练结果如下：

```
Epoch 00001: LearningRateScheduler reducing learning rate to 0.001.
Epoch 1/2
1000/1000 [==============================] - 243s 243ms/step - loss: 0.3401 - accuracy:
0.8853 - val_loss: 0.3226 - val_accuracy: 0.8985

Epoch 00002: LearningRateScheduler reducing learning rate to 0.0005000000237487257.
Epoch 2/2
1000/1000 [==============================] - 287s 287ms/step - loss: 0.0119 - accuracy:
0.9985 - val_loss: 0.4079 - val_accuracy: 0.8945
```

图 8 CNNtext 模型训练结果

分析：

训练集准确率达到 99%，而测试集准确率接近 90%。通过 CNNtext 模型与 LSMT 模型训练结果比较，无论是训练集准确率还是测试集准确率，CNNtext 模型都要优于 LSMT 模型，而训练率准确率明显高于测试集准确率，主要问题在于分词方法存在缺陷以及数据量较小，对结果造成了偏差。总的来说，无论是训练集准确率还是测试集准确率，结果还是比较令人满意。

3.2 热点问题挖掘

3.2.1 数据预处理

3.2.1.1 ID 变换及字符过滤

我们的留言内容都是中文，而数据中存在许多“，”、“。”、“\n”、“\t”等非必要字符，因此需要将其利用正则表达式过滤掉。而由于此题我们针对的是留言详情建模，因此在过滤字符时，只针对留言详情进行处理，最终得到的结果如下图 9 所示：

clean_review
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...

图 9 字符过滤后的数据展示图

3.2.1.2 分词及过滤停用词

如上图 3 所示，去掉无意义的字符之后，所有中文文字都连成了一个句子，无法提炼关键信息，因此，我们采用了 Python 中文分词组件包分词 pkuseg 将其切分成多个词语。由于分词之后发现其中包含了很多日常使用频率很高的常用词，如“吧”、“吗”、“呢”、“的”等

一系列无法反映出文本的主要意思的词，因此为了减少计算复杂度，我们利用网络上下载的中文停用词库将这些无意义的词语过滤掉，最终得到的结果如下图 10 所示：

	留言主题	clean_review
2830	投诉A3区坪塘镇白泉村润泉山庄破产案分配不公	投诉A3区坪塘镇白泉村润泉山庄破产案分配不公
19	投诉A市温斯顿英语培训学校拖延退费！	投诉A市温斯顿英语培训学校拖延退费
208	咨询A市公交409线路问题	咨询A市公交409线路问题
787	A市园博家园何时开工建设？	A市园博家园何时开工建设
1033	A市何时开通地铁环线？	A市何时开通地铁环线

图 10 分词及去停后数据展示图

3.2.2 构建语料库

文本挖掘最重要的是关键词的提取，在本题使用的数据是留言主题，由于主题已经是精简过后的文字，因此，经过分词并去除停用词之后即可认为是关键词。利用 Gensim 库创建 corpora.Dictionary 对象，为语料库中出现的所有单词分配了唯一的整数 id。利用 corpora.Dictionary 对象的 doc2bow 将每个文本的关键词转化成矢量。

	留言主题	clean_review	cut_review	vec
1049	建议将A7县泉塘招商局物流公司和中国铁建重工外迁	建议将A7县泉塘招商局物流公司和中国铁建重工外迁	[建议, A7, 县泉塘, 招商局, 物流, 公司, 铁建, 重工, 外迁]	[(32, 1), (549, 1), (938, 1), (1237, 1), (1657...
4266	咨询A市办理社保卡问题	咨询A市办理社保卡问题	[咨询, 市, 办理, 社保卡]	[(681, 1), (1046, 1), (1482, 1), (2662, 1)]
3348	A市810路公交高峰运力严重不足	A市810路公交高峰运力严重不足	[市, 810, 路, 公交, 高峰, 运力, 严重不足]	[(535, 1), (1482, 1), (3098, 1), (3473, 1)]
2354	A市长城物业多次停水停电，影响居民正常生活	A市长城物业多次停水停电影响居民正常生活	[市长, 城, 物业, 停水, 停电, 影响, 居民, 生活]	[(485, 1), (487, 1), (1168, 1), (1425, 1), (15...
2381	万科魅力之城小区底层门店深夜经营，各种噪音扰民	万科魅力之城小区底层门店深夜经营各种噪音扰民	[万科, 魅力, 城, 底层, 门店, 深夜, 经营, 噪音, 扰民]	[(107, 1), (1085, 1), (1168, 1), (1628, 1), (1...

图 11 关键词转换成矢量后的数据展示图

3.2.3 计算文本相似度

利用 Gensim 的 models.TfidfModel 进行主题向量转换，通过挖掘语料中隐藏的语义结构特征，最终可以变换出一个简洁高效的文本向量。最后利用 Gensim 提高的文本相似度查询 API：similarities。利用 similarities 这个 API 接口中的 SparseMatrixSimilarity()方法计算并返回与各个文本的相似度的稀疏矩阵。

	1	2	3	4	5	6	7	8	9	10	...	4317	4318	4319	4320	4321	4322	
2276	0.088054	0.021264	0.0	0.01853	0.057839	0.032345	0.0	0.187946	0.000000	0.047190	...	0.017831	0.000000	0.016383	0.014807	0.0	0.000000	0.0
730	0.000000	0.000000	0.0	0.00000	0.061020	0.044382	0.0	0.011984	0.007519	0.000000	...	0.000000	0.006292	0.000000	0.000000	0.0	0.005477	0.0
4251	0.000000	0.000000	0.0	0.00000	0.000000	0.000000	0.0	0.170085	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0
3453	0.000000	0.000000	0.0	0.00000	0.007195	0.000000	0.0	0.013454	0.067745	0.000000	...	0.000000	0.007064	0.000000	0.000000	0.0	0.006149	0.0
2913	0.017015	0.076823	0.0	0.01148	0.069234	0.000000	0.0	0.000000	0.000000	0.009119	...	0.011048	0.000000	0.010150	0.009174	0.0	0.000000	0.0

5 rows × 4326 columns

图 12 各个文本的相似度的稀疏矩阵展示图

3.2.4 计算相似文本数

计算与各个文本的相似度之后，会发现有些文本的相似度很低。由于要计算热度指数就需要知道相似的留言数量的多少，而只有相似度达到一定的高度的留言才会反映同一类问题。因此设置阈值进行相似文本的筛选，并计算出各个文本的文本相似数量。一个字典存储文本 ID 和文本相似数量方便后续的热度指数计算。在此设置的阈值为 5，就是一个文本相似文本数量要大于 5 个才能分为一类问题。

考虑到同一类文本的多个 ID 的文本相似数量都较高，将会影响热度问题的反映，因此在将同类文本 ID 从字典中删除之后再进行下一个文本 ID 的寻找。

3.2.5 热度指数公式

热搜榜是广大用户原创内容平台（User Generated Content，简称“UGC”）如微博、贴吧、知乎等广泛使用的排序机制（Rank-order Mechanism）。该机制根据读者评级对 UGC 进行降序排列，评级越高者展示位置越靠前，评级最高则将置顶榜首。热搜的形成其根本还是在于用户在一段时间内对某一地点、某一事情、某一人物的讨论，当用户的发帖数形成一定数量时，系统根据算法自然而然形成热搜。而用户评级方式则是对特定内容的浏览、评论、转发与点赞，用户的浏览量、点击数、点赞数、反对数以及评论数将构成 UGC 排行的直接依据。在留言数据中包含与热搜榜帖子类似的指标如：留言内容、留言时间、点赞数、反对数等，根据指标生成以下的热度指数公式。

热度指数 = 文本相似率*0.5 + 点赞率*0.2 - 反对率*0.1 + 热点消失率*0.2

（0.5 为文本相似率影响因子，0.2 为点赞率影响因子，0.1 位反对率影响因子，0.2 为热点消失率因子）

其中：

文本相似率 = 热中总文本数/数据总文本数

点赞率 = 热点总点赞数/数据总点赞数

反对率 = 热点总反对数/数据总反对数

热点效率率 = Ln（1/热点持续时间）

话题持续时间 = 热点截止时间 - 热点开始时间

3.2.6 热点地点和问题描述提取

4017	1	283482	A909232	丽发新城小区附近搅拌站的一些问题	2019-12-07 09:21:32	我是A市A2区丽发新城小区的居民，我要反应我们小区的一些问题。1.小区违建部分到今未拆除，并...	0	0
1050	1	213464	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民	2019-12-10 12:34:21	我是暮云街道丽发新城小区的业主，我要投诉开发商在小区附近违建大型搅拌站。该搅拌站的设备太吵了...	0	0
1089	1	214282	A909209	A市丽发新城小区附近搅拌站噪音扰民和污染环境	2020-01-25 09:07:21	你们管不管A2区丽发新城小区啊！这个附近建了个搅拌厂啊，天天吵天天吵，烦死了不仅吵还臭！说好...	0	0
68	1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-11-13 11:20:21	我是A2区丽发新城小区的一名业主，我要投诉同发投资有限公司在未经小区业主同意的情况下，在离小...	0	0
1797	1	231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-12-02 11:20:21	尊敬的领导，我是A2区丽发新城小区的一名业主，再次投诉同发投资有限公司在未经小区业主同意的情...	0	0
2707	1	253040	A909202	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-12-04 12:10:21	投诉A2区丽发新城小区附近违建搅拌站！该站每天早6点一直到晚8点设备都在运行，每天耳边都是各...	0	0

图 13 附件 3 原始数据

通过查看同一热点的留言主题发现，地点名词几乎类似。用 pkuseg 分词构建同一热点留言主题的语料库，生成词频字典，将词频映射留言主题中，过滤掉词频较小的词语同时标注词性，选择词性为“ns”（地点名词）作为地点，其他词性合并作为问题描述。

3.2.7 结果分析

	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
0	1	1	27.877957	2019-11-13 至 2020-01-25	A市新城	丽发小区附近搅拌站噪音扰民和污染环境
1	2	2	24.267456	2019-08-06 至 2019/8/31	伊景园滨河苑	关于捆绑销售车位的投诉
2	3	3	11.930953	2018-11-15 至 2019/7/7	A市	反映人才补贴问题
3	4	4	9.886749	2019/8/12 至 2019/8/26	A市	建议经开区泉星公园项目规划进一步优化
4	5	5	9.311155	2019/1/21 至 2019/9/12	A3区西湖茶场村	街道五组什么时候能拆迁

图 14 热点问题表

2707	1	253040	A909202	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-12-04 12:10:21	投诉A2区丽发新城小区附近违建搅拌站！该站每天早6点一直到晚8点设备都在运行，每天耳边都是各...	0
2093	1	238212	A909203	丽发新城小区附近建搅拌站合理吗？	2019-12-12 10:23:11	请问在居民区附近百米的地方建搅拌站合理吗？丽发新城小区附近不到百米的地方建了搅拌站，每天设备...	0
3554	1	272224	A909224	丽发新城小区噪音大粉尘大,求撤走搅拌站	2020-01-09 19:46:10	我是暮云街道丽发新城小区的一名业主,这里简直无法生活了,因为一个搅拌站,白天粉尘太大无法呼吸...	0
33	1	188809	A909139	A市万家丽南路丽发新城居民区附近搅拌站扰民	2019-11-19 18:07:54	\n\t\t\t\t\t\t\n\t\t\t\t\t\t\tA市万家丽南路丽发新城居民区，开发商在小区旁...	0
832	1	208285	A909205	投诉小区附近搅拌站噪音扰民	2019-12-15 12:32:11	尊敬的领导，我是A市暮云街道丽发新城的一名业主，最近遇到了意见特别烦心的事情，我是做小区安保...	0
2808	1	255008	A909208	投诉小区附近搅拌站噪音扰民	2019-11-18 12:23:22	暮云街道丽发新城边上在建大型搅拌站，听说是从别的地方搬过来的，体会最深的就是噪音很大，扬尘污...	0
3075	1	261072	A909207	投诉小区附近搅拌站噪音扰民	2019-11-23 23:12:22	投诉A市暮云街道丽发新城附近大型搅拌站水泥厂噪音严重扰民，扬尘污染环境，希望有关部门回复，在...	2
3321	1	266665	A00096279	投诉小区附近搅拌站噪音扰民	2019-12-04 17:23:22	开发商把特大型搅拌站，水泥厂从绿心范围内搬迁到A市暮云街道丽发新城的居民区，产生的噪音和扬尘...	0

图 15 热点问题留言明细表

如上图的热点问题表和热点问题留言明细表可以发现，整体效果不错。其中 A 市丽发新城小区中噪音扰民和环境污染问题排名第一，其次是 A 市伊景园滨河苑项目捆绑销售车位问题、A 市人才补贴问题、A 市经开区泉星公园项目优化问题和 A3 区西湖街道茶场村街道五组拆迁问题。其中通过加入热点消失因子，发现热点的持续时间对热度指数影响较大，但这也更加符合生活实际情况。在本问题中仍有一些缺陷，如问题描述中出现的‘丽发’、‘经开区’、‘泉星公园’等词本该属于地点名词，从而可能造成问题描述有问题，经过反复调试发现问题在于标注词性时，这些词被标注成了‘n’（名词），而非理想中的地点名词。通过各项结果与原数据进行比较，发现在合理的热度指数公式下，与实际分类情况与热度指数情况相符合。

3.3 答复意见评价

针对相关部门对群众留言的答复，为了验证其答复意见与群众所提问题是否相关以及给出的建议是否合理，我们利用 python 中的 gensim 库中的 TF 模型等一些算法计算相关部门回复详情与群众留言主题以及详细内容的相似度，具体流程如下：

1) 第一步将去掉无意义字符后的群众“留言详情”文本数据通过 jieba 分词进行处理，形成一个二维维数组，并将二维数组生成词典；
2) 第二步将二维数组通过 doc2bow 稀疏向量，形成语料库，并使用 TF 模型算法，将语料库计算出 TFIDF 值；
3) 第三步获取词典 token2id 的特征数(字典里面键的个数)，计算稀疏矩阵相似度，建立一个索引；
4) 第四步将相关部门“答复意见”文本数据进行 jieba 分词处理，通过 doc2bow 计算测试数据的稀疏向量，最终求得“答复意见”文本数据与“留言详情”文本数据的相似度。

经过上述四步最终得到的相似度结果如下图 16(a)与图 16(b)所示：

留言主题	clean_review	clean_answer	score
A2区景蓉华苑物业管理有问题	2019年4月以来，位于A市A2区桂花坪街道的A2区	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”的调查核实情况向该网友答复如下：	0.775635
A3区潇楚南路洋湖段怎么还没修好?	潇楚南路从2018年开始修，到现在都快一年了，i	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区洋湖街道高度重视，立即组织精干力量调查	0.759239
请加快提高A市民营幼儿园老师的待遇	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：为了改善和提高民办幼儿园教师待	0.751056
在A市买公寓能享受人才新政购房补贴	尊敬的书记：您好！我研究生毕业后根据人才新	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映的问题交由市房屋交易管理中心办	0.656488
关于A市公交站点名称变更的建议	建议将“白竹坡路口”更名为“马坡岭小学”，原“马坡	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“白竹坡路口”更名为“马坡岭小学”，原“马坡	0.870181
A3区含浦镇马路卫生很差	欢迎领导来A市泥泞不堪的小含浦镇滚泥巴，这个	网友“A00077538”：您好！针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、含浦街道高度重视，现回复如下：您留言	0.62325
A3区教师村小区盼望早日安装电梯	尊敬的胡书记：您好！过去在小区买房是为了自	网友“A000100804”：您好！针对您反映A3区教师村小区盼望早日安装电梯的问题,A3区住建局高度重视，立即组织精干力量调查	0.751796
反映A5区东澜湾社区居民的集体民生问	我做为一东澜湾社区居民，我替社区几千户居民	网友“UU00812”您好！您的留言已收悉。现将有关情况回复如下：一、关于小区附近幼儿园的问题。经A5区教育局、黎托街道核	0.802389
反映A市美麓阳光住宅楼无故停工以及	我是美麓阳光a栋803业主，现工地已停工二个月	网友“UU008792”您好！您的留言已收悉。现将有关情况回复如下：据查，美麓阳光项目位于A3区枫林路，建筑面积25613.32平方	0.705222
反映A市洋湖新城和顺路洋湖壹号小区	胡书记好！根据规划，洋湖新城和顺路两侧都有	网友“UU008687”您好！您的留言已收悉。现将有关情况回复如下：您所反映的地点为洋湖新城片区和顺路两厢，北至先导路，南	0.75017

图 16 相似度结果(a)

建议用法制思路综合管理农村环境卫生	尊敬的易书记:　　农村环境卫生确实存在很大	网友：您好！留言已收悉	0.128565
希望解决清水塘三小学生出行安全问题	尊敬的易书记:　　您好！我是清水塘三小学生	网友：您好！留言已收悉	0.074966
希望A市其他市直部门的领导在平台上	书记你好！我在问政西地省名下给A市公安部	网友：您好！留言已收悉	0.167968
请求解决A市含浦锑都家园小区用水问	易书记，您好！我是含浦锑都家园小区的业	网友：您好！留言已收悉	0.096724
请求解决蒋垄家火车噪音问题	尊敬的易书记 及市委主管部门:　　最近以来，	网友：您好！留言已收悉	0.033352
对A市地图、西地省地图出版社及A市地	对A市地图、西地省地图出版社及A市地名委	网友：您好！留言已收悉	0.022178
举报A7县泉塘社区的"当代和坤"杨国建	举报信各级督查纪检部门:　　你们好！辛苦	网友：您好！留言已收悉	0.020308
反对拆除A市劳动广场中央绿化岛	劳动广场中央绿化岛已经有几十年的历史，经	网友：您好！留言已收悉	0
西地省长株潭两型社会建设的建议	易书记:　　西地省长株潭两型社会建设叫了20	网友：您好！留言已收悉	0.102329
建议取消绕城高速收费	A市绕城高速（三环线）系A市城市交通网络	网友：您好！留言已收悉	0

图 16 相似度结果(b)

如上图 16 (a)与图 16 (b)所示结果可以发现，回复内容与群众留言内容的相似度有两种情况，第一种如(a)图所示，其相似度较高，大约有 70%左右，而第二种如(b)图所示，其相似度特别低，大约只有 9%左右。将两张图进行对比可以发现：(1)从相关性的角度来看，几乎所有答复都与群众所提问题相关，只有一些答复除了写明“已收悉”以外，并未做出相关解释，并不具有相关性；(2)从完整性角度来看，相似度较高的文本中，相关部门的答复意见篇幅占据较大，其首先明确指明了所答问题是针对的哪位网友，接着明确表示具体回复信息，说明此问题的情况，并在末尾写明答复日期，具有完整性。而在相似度低的文本中可以发现，答复内容除了表明已收到以外，并未展示其他信息，因此回答相对来讲是不完整的，不具有完整性；(3) 从可解释性角度来讲，相似度较高的文本中，相关部门针对群众所提问题作出的回应内容中均存在解释说明，表明此问题应当如何解决或者实地考察过并且说明原因为何不能解决，具有可解释性。而在相似度低的文本中并未认真作出回应，没有根据群众留言进行调查解释，不具有可解释性。

4. 结束语

4.1 总结

至此，基于数据挖掘的“智慧政务”文本挖掘应用工作基本完成。针对第一问群众留言分类问题，我们通过对原始数据进行分词、过滤停用词等数据预处理，然后建立 LSTM 模型和 CNNtext 模型，最终训练得到预测准确率分别在 86%和 90%左右，得出结果 CNNtext 模型在文本处理方面由于 LSTM 模型；针对第二问的热点挖掘问题，同样对数据进行预处理，然后计算文本相似率，通过热度指数公式对热点进行排行，其中热点消失率的增加对热度排行起到了关键的作用，使情况更加接近现实。针对第三问的答复意见评价问题，我们对数据进行了预处理之后，利用 python 中的 gensim 库构建语料库，利用 TF 模型算法计算出 TF-IDF 值，然后计算稀疏矩阵相似度，建立一个索引，最后求出相关部门答复内容与群众留言内容

之间的相似度，并利用相似度从答复的相关性、完整性、可解释性三个角度对答复意见的质量进行了评价说明。

4.2 不足与展望

虽然 pkuseg 分词效果较好，但是其速度慢，导致花费时间较长。并且由于经验不足和能力稍有欠缺，参数调优过程不理想。虽然结果具有一定误差，但是此项目中所用方法都具有一定的实际意义，具有一定的推广性。

参考文献

[1] 王烟.自然语言处理技术在建筑使用后评价中的应用[J].南方建筑，2019(1):82-87.

[2] 方明之.自然语言处理技术发展与未来[J].信息科技搜索，2019.

[3] 曾小芹.基于 Python 的中文结巴分词技术实现[J].信息与电脑(理论版),2019,31(18):38-39+42.

[4] 张尚田，陈光，邱天.基于融合特征的 LSTM 评分预测[J].计算机与现代化,2020(3):49-53.

[5] 景永霞，苟和平，孙为.基于 TextRank 的 KNN 文本分类算法研究[J].洛阳理工学院学报(自然科学版)，2019(3):66-69+76

[6] 张建娥.基于 TFIDF 和词语关联度的中文关键词提取方法[J].情报科学,2012,30(10):1542-1544+1555.

[7] 叶建成.利用文本挖掘技术进行新闻热点关注问题分析[D].广州大学本科毕业论文,2018.