摘要

随着大数据、 云计算、人工智能等技术的发展，网络问政平台的群众问政留言记录，及相关部门对群众留言的答复意见，具有大量值得挖掘的信息。并且以往的人工进行留言划分和热点整理在大数据时代效率较为低下。

因此，本文根据题目中的具体问题，构建了文本分类模型，热点挖掘模型与答复意见质量评价模型。帮助政府直观地了解民意，提升政府的管理水平和施政效率。

针对问题一，我们构建了融合 LDA 与 Skip-gram 的 TextCNN-XGBoost 文本分类模型。我们首先对附件 2 进行数据预处理，将数据转换为纯文本数据，并去除停用词的干扰。然后通过 Skip-gram 与 LDA 构建词向量，并通过 TextCNN 提取特征，降低维数，避免数据的高维稀疏。最后通过 XGBoost 集成学习进行分类。

针对问题二，我们构建了热点挖掘模型。我们首先对附件 3 的文本数据进行预处理，通过 TexSmart 识别文本数据中命名实体，对分词模型添加了命名实体等自定义词汇，然后通过基于 Skip-gram 的 Word2Vec 模型构建词向量，并利用了 Tf-Idf 值与额外权重对词向量矩阵加权平均得文本数据的向量，最后再与命名实体的词向量拼接得最终的输入向量。接着通过基于余弦相似度的聚类算法将留言归类，然后构建并计算热点指数的指标，对热点问题筛选并排序。最后通过 Seq2Seq 自动生成热点问题的描述。

针对问题三，我们构建了答复意见质量评价模型。我们首先对附件 4 的文本数据进行预处理。然后通过对相关政务网站及资料的浏览，我们构建了四个评价指标，分别是：可解释性、时效性、相关性与完整性，通过 AHP 层次分析法确定指标的权重。然后通过命名实体识别模型与 TF-IDF 关键词提取等技术，求出各个指标的得分，最终得到每一条答复意见的质量评分。

关键词： TextCNN-XGBoost；TexSmart；Seq2Seq;AHP

Abstract

With the development of big data, cloud computing, artificial intelligence and other technologies, there are a lot of information worthy of mining in the records of mass political messages on the network political platform, and the response opinions of relevant departments to the mass messages. And the past manual message division and hot spot sorting are inefficient in the era of big data.

Therefore, according to the specific problems in the topic, this paper constructs text classification model, hot spot mining model and reply quality evaluation model. Help the government understand the public opinion intuitively and improve the management level and governance efficiency of the government.

To solve the first problem, we build a text CNN xgboost text classification model which integrates LDA and skip gram. First, we preprocess the data in attachment 2, transform the data into plain text data, and remove the interference of stop words. Then, the word vector is constructed by skip gram and LDA, and the feature is extracted by textcnn to reduce the dimension and avoid the high-dimensional sparse data. Finally, it classifies them by xgboost integrated learning.

For the second problem, we build a hot spot mining model. First, we preprocess the text data in Annex 3, recognize the named entity in the text data through texsmart, add the named entity and other customized words to the segmentation model, then construct the word vector through word2vec model based on skip gram, and use TF IDF value and extra weight to weigh and average the word vector matrix to get the vector of the text data, and finally, with the word of the named

entity The final input vector is the result of vector splicing. Then, the message is classified by cosine similarity based clustering algorithm, and then the index of hotspot index is constructed and calculated to filter and sort the hotspot issues. Finally, the description of hot issues is automatically generated by seq2seq.

In view of the third question, we build the quality evaluation model of the reply. First, we preprocess the text data in Annex 4. Then, by browsing related government websites and materials, we build four evaluation indexes, which are interpretability, timeliness, relevance and integrity, and determine the weight of the indexes by AHP. Then through named entity recognition model and TF-IDF keyword extraction technology, we can get the scores of each index, and finally get the quality scores of each reply.

Keywords: TextCNN-XGBoost; TexSmart; Seq2Seq;AHP

目录

1 引言

1.1 问题描述

随着互联网时代的发展，微信、微博、市长信箱和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，大数据、人工智能等技术的迅速发展，建立基于自然语言处理技术的智慧政务系统取代人工劳动已经成为了新趋势。而我们的任务，就是根据相关部门对部分群众留言及答复意见。利用自然语言处理和文本挖掘的方法，构建文本分类模型，热点挖掘模型与答复意见质量评价模型。

问题一中，在处理网络问政平台的群众留言时，需要将留言进行分类，以便后续将群众留言分派至相应的职能部门处理。现根据附件 2 的数据，建立关于留言内容的分类模型。

问题二中，某一时段内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。现根据附件 3 的数据，将某一时段内反映特定地点或特定人群问题的留言进行归类，并定义合理的热度评价指标，并给出评价结果。

问题三中，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、可解释性和时效性的角度对答复意见的质量给出一套评价方案，并实现。

1.2 篇章结构

本文共分为五个部分，各章内容安排如下：

第一章，对论文需要解决的问题进行描述，并简单介绍整篇论文的结构安排。

第二章，对附件中数据进行探查，以对数据进行预处理。

第三章，对附件中的群众留言进行分词，并在语料库中检索分词，爬取语料训练词向量。进而提取群众留言的特征，构建文本分类模型。

第四章，运用热点挖掘模型识别群众留言及答复意见的地点、人物等信息进行归类，挖掘特定地点特定人群的问题。

第五章，综合考虑答复意见的可解释性、完整性、相关性与时效性，建立多方面、立体的群众留言答复质量评价体系。

2 数据探查

在对文本信息进行分析时，我们需要提取文本的一些特征，例如：时间、点

赞数和反对数等。本节主要对数据进行一个初步的探索，观察有无缺失值、异常值。

我们使用 SPSS 分别对附件 2、附件 3 与附件 4 进行统计，结果如下所示：

统计

		留言用户	留言主题	留言时间	留言详情	一级标签	留言编号
个案数	有效	9210	9210	9210	9210	9210	9210
	缺失	0	0	0	0	0	0

图 2-1 附件 2 描述统计图

统计

		留言用户	留言主题	留言时间	留言详情	点赞数	反对数
个案数	有效	4326	4326	4326	4326	4326	4326
	缺失	0	0	0	0	0	0
平均值						2.89	.12
中位数						.00	.00
标准 偏差						47.833	1.028
范围						2097	53
最小值						0	0
最大值						2097	53

图 2-2 附件 3 描述统计图

统计

		留言用户	留言时间	留言主题	留言详情	答复时间	答复意见
个案数	有效	2816	2816	2816	2816	2816	2816
	缺失	0	0	0	0	0	0

图 2-3 附件 4 描述统计图：

根据统计结果，我们确定在附件 2、附件 3、附件 4 中无缺失值，数据正常。

3 结合主题模型词向量的 CNN-XGBoost 短文本分类

本章针对问题一，首先分别建立基于 LDA 的文本表示模型和基于 Word2Vec 的文本表示模型，然后处理数据并将群众留言内容转换为词向量文本矩阵，再通过 text-CNN 提取特征，最后通过 XGBoost 算法进行分类。

3.1 数据的选取与预处理

本节主要详细介绍了数据选取的理由以及对原始数据进行预处理的步骤。

3.1.1 数据的选取

附件 2 中反映群众留言内容的有两部分，一是留言主题，二是留言详情。根据观察，留言主题内容精炼，是留言详情的集中反映；而留言详情内容较多，长度较长，且留言详情有较多的冗余词语，对文本主题的干扰比较大。因此，我们选择留言主题作为对留言文本进行分类的数据。

3.1.2 数据预处理

中文文本处理复杂，与英文相比，词与词之间缺乏明显的间隔符。预处理包括：文本去噪、中文分词与去除停用词。

文本预处理需要对文本数据进行清洗，使用正则表达式去除非中文符号。使文本内容转化为纯文本数据。然后对纯文本数据进行分词，引入融合了百度停用词表、哈工大停用词表、四川大学及其智能实验室停用词库、知网停用词表等常用的停用词词典，去除无实际意义的词，降低数据维数，提高精确度。

最后，对附件二的一级分类标签作哑变量处理。用数字表示所属一级分类。

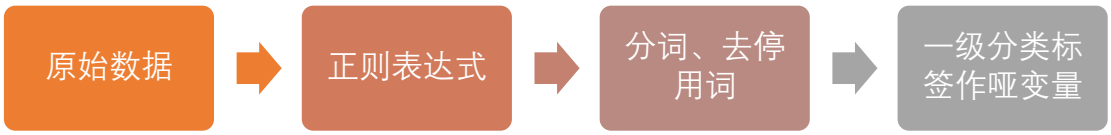

图 3-1 数据预处理的流程

3.2 文本表示模型[1]

3.2.1 基于 Word2vec 的文本表示

计算机只能识别机械语言，为了使计算机能够识别文本语言，需要将文本语言转换为数据矩阵。

Word2Vec 是一种基于神经网络的词向量生成模型。该模型是用深度学习网络对语料数据的词语及其上下文的语义关系进行建模，以求得到低维度的词向量，对文本数据进行数字表示。

Word2Vec 主要有 CBOW 和 Skip-gram 两种模型。Skip-gram 模型是在给定当前词预测上下文，如下图所示。

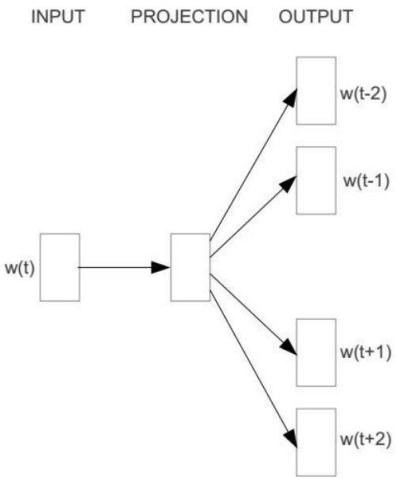

图 3-2 Skip-gram 模型

本文使用 Skip-gram 模型，将数据预处理后的文本数据进行词向量表示。生成形如$w = (d_1, \cdots, d_n)$的词向量。其中w表示特征词，d_i表示特征词w的第i个维度。本文训练的词向量维度是 200 维。

由于附件中的文本数据有限，仅使用附件的文本数据训练词向量效果有限。因此，我们先对附件的文本数据进行分词，再在 CCL 北京大学语料库中检索分词后词语的语料，爬取语料进行训练，并且训练时设定词共现窗口大小为 5。

图 3-3 语料库截图

3.2.2 基于 LDA 的文本表示

隐含狄利克雷分布主题模型是 2003 年提出的在文档-主题和主题-词语 2 个层次上建模的三层贝叶斯模型。

基于 LDA 的主题词向量是在 LDA 生成的主题-词语分布φ基础上进行的。$\varphi_{i,w}$表示某个主题Z_i出现词语W的概率情况，

$$\varphi_{i,w} = p(w|Z_i)$$

本文要产生的词向量是由词语W出现在每个主题上的概率组成。v_w表示词语W的主题词向量，$v_w(i)$表示主题词向量在第i个维度上的取值。

$$v_w(i) = p(Z_i|w)$$

因为 LDA 模型中的主题是隐含的，并不一定对应人们认为的某个主题，因此假定每个主题出现概率相同，则$v_w(i)$可表示为：

$$v_w(i) = \frac{p(w|Z_i)}{\sum_{i=1}^{k} p(w|Z_i)} = \frac{\varphi_{i,w}}{\sum_{i=1}^{k} \varphi_{i,w}}$$

本文基于 LDA 生成 200 维的词向量。

3.3 基于 TextCNN-XGBoost 的文本分类模型[2]

3.3.1 TextCNN

本文提出的 TextCNN-XGBoost 文本分类模型主要包括卷积层、池化层和构造分类器。若一个句子的长度为 n。通过 1.2 训练好的 400 维的词向量进行词嵌入，即词语$w = (v_1, v_2)$，其中v_1是 Skip-gram 训练的词向量，v_2是基于 LDA 的主题词向量。即每个词表示维200×2维的向量。将句子转化为$n \times 200 \times 2$的数据矩阵作为输入。

卷积层通过使用$2 \times 200, 3 \times 200, 4 \times 200, 5 \times 200, 6 \times 200$尺寸的卷积核进行特征提取，对提取后的特征进行池化处理，然后将不同尺寸的卷积核提取的特征进行全连接，最后通过 XGBoost 分类器完成文本分类。

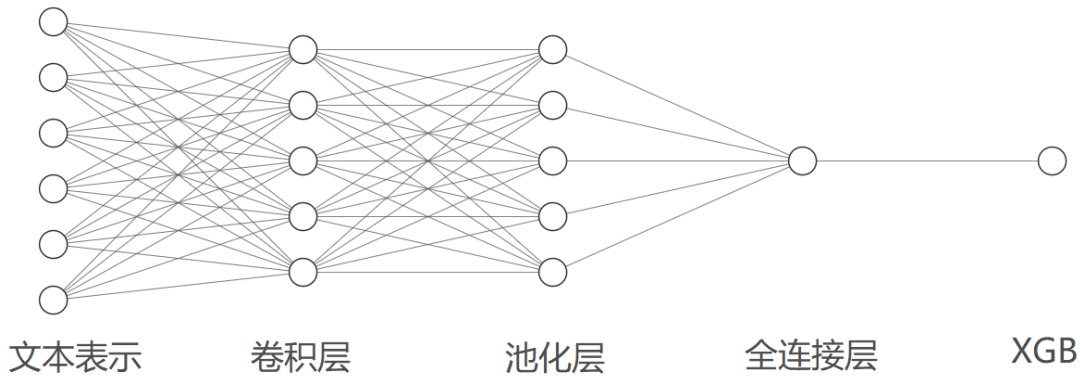

图 3-4 基于 TextCNN-XGBoost 的文本分类模型

3.3.2 卷积层与池化层

句子表示为$n \times 200 \times 2$维的数据矩阵，通过高度为$h = (2,3,4,5,6)$，宽度为 200 的卷积核进行卷积，提取新的特征向量。然后采用 1-max-pooling 最大池化，减少维数，获取最大特征值。再通过全连接层，拼接组成新的特征向量作为

XGBoost 分类器的输入。

3.3.3 XGBoost

XGBoost 是对梯度提升算法的改进，求解损失函数极值时使用了牛顿法，将损失函数泰勒展开到二阶，另外损失函数中加入了正则化项，避免过拟合。我们将 TextCNN 全连接层提取的特征向量，作为 XGBoost 分类器的输入，进行文本分类。

3.4 模型求解

3.4.1 模型算法流程

Step1 文本数据预处理，将留言转换为纯文本数据，分词、去除停用词并保存分词结果。再对类别标签作哑变量处理。

Step2 在 CCL 北京大学语料库中检索分词的语料并爬取，采用 Skip-gram 模型训练词向量

Step3 采用 LDA 主题模型对附件的文本数据生成词向量

Step4 根据训练好的词向量，将文本映射为数据矩阵

Step5 将数据矩阵作为 TextCNN 的输入，提取特征

Step6 提取的特征作为 XGBoost 的输入，并进行分类。

Step7 对 TextCNN-XGBoost 采用网格搜索，寻找最优的超参数组合

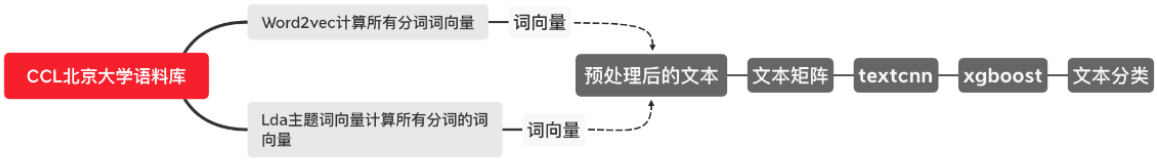

图 3-5 模型算法流程

3.4.2 求解结果

经过上述处理，并采用十折交叉验证，模型求解结果为：

	precision	recall	f1-score	support
0	0.91	0.96	0.94	589
1	0.96	0.93	0.94	296
2	0.96	0.92	0.94	188
3	0.96	0.94	0.95	493
4	0.94	0.96	0.95	601
5	0.94	0.90	0.92	354
6	0.95	0.93	0.94	242
accuracy			0.94	2763
macro avg	0.94	0.93	0.94	2763
weighted avg	0.94	0.94	0.94	2763

图 3-6 模型求解结果

可见，融合了 LDA 与 Word2Vec 词向量的 TextCNN-XGBoost 分类器取得了很好的结果，f1-score、recall 与 precision 均达到 90%以上。

3.5 本章小结

本章综合 LDA 与 Word2Vec 构建词向量，并通过 TextCNN 提取特征、降维，最终通过 XGBoost 进行分类，取得了很好的效果。模型分类精度高，极大程度地解决了特征维数高、数据稀疏的问题。

4 热点问题挖掘

一个热点问题中往往有关键词，于是关键词相似的留言可以视为同一类问题。所以，首要任务就是将同一类留言聚在一起，再通过一些指标来衡量它的热度。先将留言中的命名实体识别出来，再将文本向量化，根据文本向量的余弦相似度进行聚类。

4.1 命名实体识别

4.1.1 TexSmart 的命名实体识别

TexSmart 是由腾讯 AI Lab 构建的一个文本理解系统，提供分词、词性标注、命名实体识别(NER)、语义联想等基础自然语言理解功能。本文主要采用 TexSmart 的 NER 模块，对文本数据进行命名实体的识别，对比起现有的开源工具，其支持上千种实体类型，对于本文中的文本数据识别表现更加出色。

本文将采用粗粒度与细粒度结合的方式，对文本数据进行命名实体识别，下面将介绍两种方法。

4.1.1.1 粗粒度 NER

粗粒度 NER 采用有监督学习方法，分别提供了两种模型：CRF 和 DNN，用于识别人名、地名、机构名等命名实体。在测试两种模型的识别效果时，我们发现对于短文本 DNN 模型的识别效果优 CRF 模型，而对于长文本 CRF 模型的识别效果要优于 DNN 模型。因此，本文主要采用 CRF 模型对“留言详情”进行命名实体识别，DNN 模型对“留言主题”进行命名实体识别。

4.1.1.2 细粒度 NER

细粒度 NER 采取一种混合方法，它融合了无监督细粒度实体识别方法、有监督的序列标注模型以及腾讯 AI Lab 在 2017 年国际知识图谱构建大赛夺冠的实体链接方法。其中有监督模型只识别出其中 12 类粗粒度的实体，无监督模型结合腾讯 AI Lab 所维护的知识图谱 TopBase 可以识别出一千多种细粒度的实体类别。

4.1.1.3 NER 的实体类型

下面表格将列出本文所使用的实体类型。

表 4-1 本文所用命名实体类型

标签	描述	文本数据中的例子
loc.generic	地点，包括国家、城市、桥梁等	“景蓉花苑”、“黄兴路步行街”、“魅力之城小区”、“卫生间”
org.generic	机构、包括政府、公司、学校等	“一米阳光婚纱艺术摄影”、“景蓉华苑业委会”、“A 市经济学院”、“新明园林建设集团有限公司”
work.generic	作品，包括书籍、电影、手机应用程序等	“问政西地省”
event.generic	事件，包括历史事件、国际会议、体育赛事等	
activity.generic	行为	“实习”，“涉黄”，“违法行为”、“扰民”
org.division	部门	“区住房和城乡建设局”
structure.route	路名	“劳动东路”
basic.job_profession	职业	“学生”
basic.position_of_person	职位	“书记”、“局长”

4.1.2 NER 应用

4.1.2.1 留言主题识别

本文首先采用粗粒度 NER 中的 DNN 模型对“留言主题”进行命名实体识别，接着获取在第一轮识别中，识别结果为空的数据，并整合成新的数据集，对该数据集采用细粒度 NER 对“留言主题”进行第二轮识别，以降低主题识别为空的数据量，最后将识别结果存入表格。

4.1.2.2 留言详情识别

首先采用粗粒度 NER 中的 CRF 模型对“留言详情”进行命名实体识别，接着获取在第一轮识别中，识别结果为空的数据，并整合成新的数据集，对该数据集采用细粒度 NER 对“留言详情”进行第二轮识别，以降低详情识别为空的数据量，最后将识别结果存入表格。

4.1.2.3 人群识别与答复意见识别

对于人群的识别，直接采用细粒度 NER 对留言详情进行命名实体识别，并存入表格。

针对附件 4 中的答复意见，采用细粒度 NER 对答复意见进行命名实体识别，并存入表格。

4.1.2.4 识别留言所属行政区域

通过观察附件 3 中的数据，可以得出该数据所属行政区域都由 26 个英文字母与数字 1-9“市”、“区”、“县”、“县城”、“乡”所组成，因此无法使用 NER 识别留言所属行政区。

针对此，我们首先将所有可能的行政区域用字典的形式存储下来，再用正则表达式获取每条留言详情中包含英文字母，最后将获取的字母作为索引条件在行政区域字典中获取与留言详情中的行政区域相同的值，并存入表格；接着获取在第一轮识别中为空的数据，组合成新的数据集，针对其留言主题再进行第二轮的行政区域识别，最后存入表格。行政区域识别效果如下图所示：

	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	涉及行政区域
0	188006	A000102948	A3区—米阳光婚纱艺术摄影是否合法纳税了?	2019/2/28 11:25:05	座落在A市A3区联丰路米兰春天G2栋320，一家名叫一米阳光婚纱艺术摄影的影楼，据说年单这一...	0	0	A市A3区
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	A市A6区道路命名规划已经初步成果公示文件，什么时候能转化成为正式的成果，希望能加快完成的路...	0	1	A市A6区
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	本人系春华镇金鼎村七里组村民，不知是否有相关水泥路到户政策和自来水到户政策，如政府主导投资村...	0	1	A7县
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	靠近黄兴路步行街，城南路街道、大古道巷、一步两搭桥小区（停车场东面围墙外），第一单元一住户卫...	0	1	A市
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	A市A3区中海国际社区三期四期中间，即蓝天璞和洲幼儿园旁边那块空地一直处于三不管状态，物业不...	0	0	A市A3区
5	188073	A909164	A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质	2019/3/11 11:40:42	作为麓泉社区麓谷明珠小区6栋居民，我们近期感觉到很震惊也很伤心。购房、签合同、办产权时从来都...	0	0	A3区
6	188074	A909092	A2区富绿新村房产的性质是什么?	2019/1/31 20:17:32	“二高一部”发出关于针对非法集资的打击的通知中是针对的金融犯罪方面，然而通知中说明的4点特征...	0	0	A市
7	188119	A00035029	对A市地铁违规用工问题的质疑	2019/5/27 16:04:44	我是一名在A市某地铁站上班的安检员，我是由中介公司介绍来上班的，安检员岗位分两个班次，一个班...	0	0	A市
8	188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	12月21日下午17时52分许，6路公交车（司机座位旁边的汽车编号3283）在A3区大道金星...	0	0	A3区
9	188249	A00084085	A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民	2019/9/17 4:25:00	保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工，噪音严重扰民，家里老人已经被折磨的不行，...	0	0	A3区

图 4-1 行政区域识别效果

全部的识别结果请看：附件/过程数据/附件 3_命名实体识别.xlsx 与附件 4_命名实体识别.xlsx

4.2 基于 Seq2Seq-Attention 的文档自动摘要

4.2.1 Seq2Seq

Seq2Seq 模型是 Cho 于 2014 年[4]提出的，是 RNN 最重要的一个变种。

RNN 是一类用于处理序列数据的神经网络，其网络架构如下图所示 。

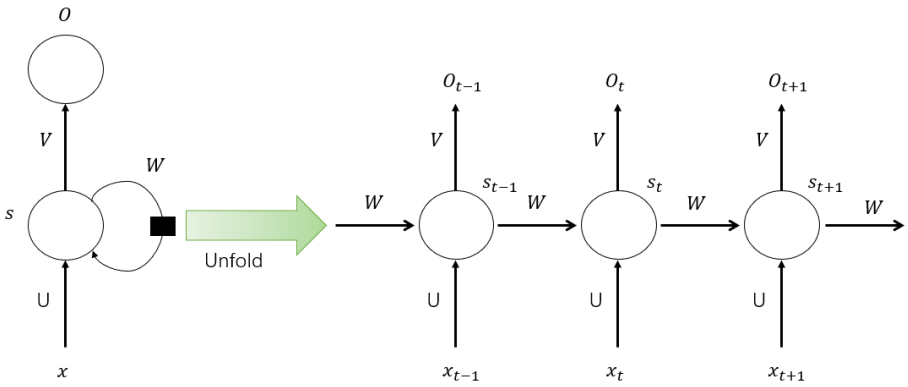

图 4-2 RNN 的网络架构

Seq2Seq 属于 encoder-decoder 结构的一种，其基本思想是，利用两个 RNN，一个 RNN 作为 encoder，另一个 RNN 作为 decoder。encoder 模块负责将输入序列压缩成指定长度的向量，这个向量就可以看成是这个序列的语义，该过程称为编码，如下图所示，而获取语义向量的方式是直接将最后一个输入的隐状态作为语义向量C。

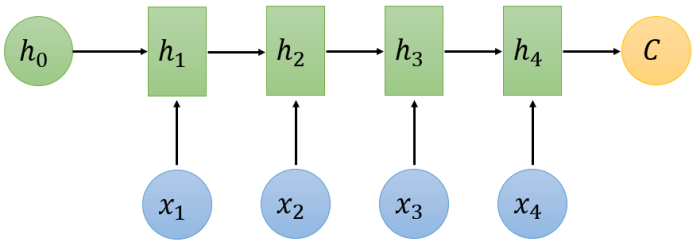

图 4-3 encoder 模块

decoder 模块负责根据语义向量C生成指定的序列，这个过程称为解码，如下图所示。

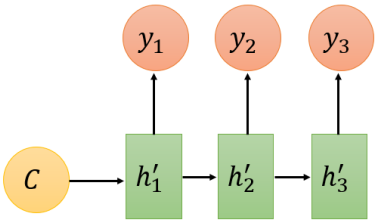

图 4-4 decoder 模块

完整的 Seq2Seq 模型构架如下图所示。

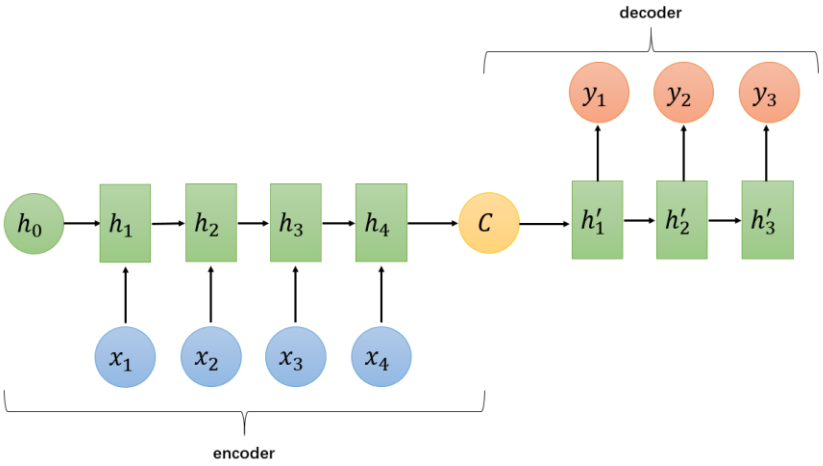

图 4-5 Seq2Seq 模型

4.2.2 Attention 机制

Attention 机制，又称为注意力机制，最早于 Neural Machine Translation by Jointly Learning to Align and Translate[5]中应用于自然语言处理领域。Attention 机制模仿于人类的注意力机制，在生成每个词的时候，对不同的输入词语给予不同的关注权重。

4.2.3 Seq2Seq-Attention

Seq2Seq 模型尽管非常灵活，但是其 encoder 模块给出的都是一个固定维数的向量，这就存在信息损失，最终在生成文本时，生成每个词所用到的语义向量都是一样的，这就导致该模型对于文本摘要提取存在极大的误差。

针对上述问题，本文使用 Seq2Seq-Attention 模型对文本摘要进行提取。该模型由 Seq2Seq 模型与 Attention 机制结合而成，在 encoder 模块模仿人类的注意力机制，提高某些词语的权重，再传入 decoder 模块，其网络架构如下图所示。该模型极大的增强了对文本摘要提取的能力。

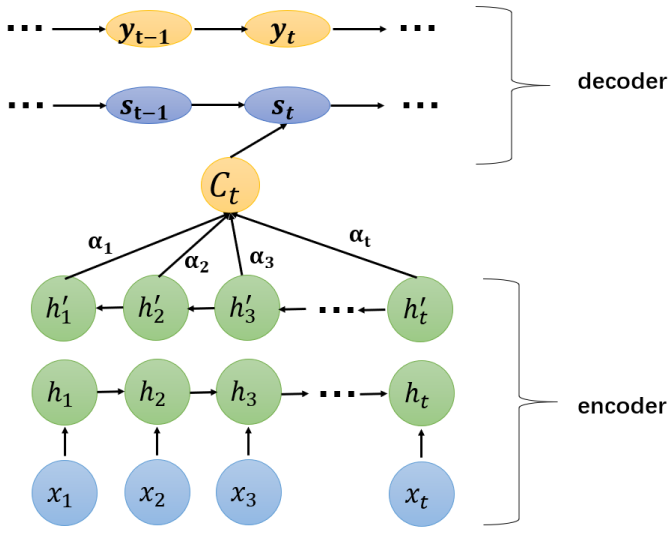

图 4-6 Seq2Seq-Attention

4.3 文本向量化及特殊处理

4.3.1 基于命名实体识别的分词

由于文本数据中有大量的类似于“A 市”、“一米阳光婚纱艺术摄影”的命名实体，所以普通的分词模型的效果十分不好。于是，将上一节提取到的命名实体与行政区域等词汇汇总成自定义词表，喂入分词模型，使得分词模型更好。如：

表 4-2 两种分词结果的对比

分词	分词结果
jieba	'A3', '区', '一米阳光', '婚纱', '艺术摄影', '是否', '合法', '纳税', '了'
自定义词库的 jieba	'A3 区', '一米阳光婚纱艺术摄影', '是否', '合法', '纳税', '了'

按上述方法将附件 3 中留言主题的文本数据分词，留以进一步加工。

4.3.2 基于 Word2Vec 的文本表示

与 1.2.1 思路相同，将上一小节分词后的词汇总，再在 CCL 北京大学语料库中检索相应语料，爬取语料。然后将附件 3 中所有文本与爬取获得的语料进行分词，再喂入词向量模型进行训练。由此得到每个词的词向量，维度为 200 维，以键值对的形式储存在.npy 文件中以备用。

4.3.3 TF-IDF 值的计算

TF-IDF（term frequency-inverse document frequency，词频-逆文本频率）是一种用于信息检索与文本挖掘的常用加权技术。它是一种统计方法，用以评估某词条对于一个文档集火一个语料库中的其中一份文档的重要程度。某词条的重要性随着它在文档中出现的次数成正比增加，但同时会随着它在语料库中 出现的频率成反比下降。

$$TF = \frac{在某文档中某词条出现的次数}{在该文档中所有词条的数目}$$

$$IDF = log\left(\frac{语料库的文档总数}{包含该词条的文档数 + 1}\right)$$

$$TF\text{-}IDF = TF \times IDF$$

由此利用 Python 的第三方库 Gensim 计算附件 3 中每一条留言主题的每一个词对应的 TF-IDF 值，并将其组成若干个一维向量，大小为$N_i \times 1$，其中N_i是第i条留言主题的分词后的词数。利用这个来调整句子中某个词重要性的比重，以突出更能体现主题的词语。

4.3.4 额外的加权

我们经过试验发现，只有 TF-IDF 值的加权，效果不是特别明显，而针对问题二，突出那些命名实体十分必要，于是对于每个词按类别额外加权：

$$命名实体:行政区域:其他词汇 = 9:3:1$$

4.3.5 留言主题的向量化

通过词向量模型训练可以得到每个词的词向量，每个文本可以用向量矩阵表示。而向量矩阵无法直接进行聚类运算或相似度运算，普遍的做法是对词向量矩阵取平均得到文本向量表示。这里我们使用加权平均[3]。

$$\boldsymbol{x}_i = \sum_{j=1}^{N_i} \boldsymbol{v}_{ij} t_{ij} w_{ij}$$

表 4-3 本小节的符号表示

符号	意义	大小	备注
$\boldsymbol{x_i}$	第i条留言主题的向量	200×1 的向量	
$\boldsymbol{v_{ij}}$	第i条留言主题的第j个词的词向量	200×1 的向量	$j = 1 \ldots N_i$
$\boldsymbol{t_{ij}}$	第i条留言主题的第j个词的 TF-IDF 值	标量	
$\boldsymbol{w_{ij}}$	第i条留言主题的第j个词的额外权重	标量	
$\boldsymbol{N_i}$	第i条留言主题的词数	标量	

4.3.6 特殊处理

为进一步突出命名实体的影响，将每条留言详情中命名实体识别出的关键词也进行向量化，并对词向量矩阵取平均，也获得了 200 维的向量。

然后，将 2.2.4 获得的向量与命名实体向量拼接为 400 维的向量，作为最终的输入数据。

4.4 热点挖掘

为了挖掘热点问题，应将内容相似的留言分在一起，首先想到的就是聚类算法。据前人研究，文本聚类大多用的是 K-Means 算法。

4.4.1 K-Means

K-Means 聚类的目的是将观测值划分为k个类，使每个类中的观测值距离该类的中心（类均值）比距离其他类中心都近。

但使用该算法，必须事先确定k值；而在问题二的热点挖掘中，我们很难确

定k的范围。经过试验，我们得出k取 3900-4050 能得出若干个效果较好的簇，但大多数的簇中的留言内容不相似。并且当k较大时，簇中样本数最大值也降低，一些留言很多的类容易被拆分。所以我们放弃这个算法。

4.4.2 基于向量余弦相似度的算法

先根据上节得到的向量计算余弦相似度矩阵$M_{N\times N}$，其中m_{ij}为第i条留言的向量与第j条留言的向量的余弦相似度:

$$m_{ij} = \cos(\theta) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|}$$

设置阈值$\varphi = 0.95$，当$m_{ij} \geq \varphi$时则认为第i条留言的向量与第j条留言的向量相似，记为 True；反之，记为 False。从而得到布尔矩阵，大小与余弦相似度矩阵一致。

聚类算法流程如下：

Step1　记集合H储存着所有列数。按行遍历整个布尔矩阵，寻找 True 最多的一行，储存该行中值为 True 的列数，即记下若干个j值，储存在集合h，并将h储存。

Step2　更新H，令$H = H - h$，即取其余集。

Step3　循环 Step1-Step2，直到H为空集。

储存的若干个h就是最终被聚成的若干个类。

4.5 热点指数的计算

4.5.1 热点指数的定义

原数据中每条留言都有留言用户、留言时间、反对数、赞同数等数据，这些可以协助衡量热点指数。

4.5.1.1 留言个数

注意到有相同用户在短时间内多次重复留言，为避免其引起的不良影响，记该类的留言个数为该类留言中用户的个数。第i类的留言个数记为m_i。

4.5.1.2 时间集中程度

若一个类中的留言的留言时间较为集中，则代表该问题反映得交频繁，应视为热点问题。留言时间的集中程度可以用时间的标准差来体现。具体公式为

$$t_i = \sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}\left(d_j - \bar{d}\right)^2}\ ,(j = 1 \ldots N_i)$$

其中t_i为第i类留言的时间集中程度，N_i为第i类中留言的个数，d_j为第i类中

第j条留言的留言时间与该类中留言时间最早的相差天数，$\bar{d}$为第i类中与该类中留言时间最早的平均相差天数。

若一个类中只有一条留言，则将其时间集中程度设为其余类的时间集中程度的中值。

4.5.1.3 点赞分数

注意到每一类的总点赞数与总反对数情况如下：

表 4-4 每一类总点赞数与总反对数的情况对比

	平均值	标准差	最小值	最大值
总点赞数	3.205244	51.176013	0	2097
总反对数	0.131356	1.093854	0	53

即总点赞数普遍要比总反对数要高，为了缩小两者之差，均取其开平方。最终的点赞分数定义为开平方后的总点赞数与开平方后的总反对数作离差标准化后之差，具体公式为

$$s_i = \frac{\sqrt{y_i} - a_{min}}{a_{max} - a_{min}} - \frac{\sqrt{n_i} - a_{min}}{a_{max} - a_{min}}$$

其中s_i为第i类留言的点赞分数，y_i为第i类留言的总点赞数，n_i为第i类留言的总反对数，a_{max}为总点赞数与总反对数中的最大值，a_{min}为总点赞数与总反对数中的最小值。

4.5.1.4 热度指数

由热度的意义可以得出热度与上述三个指标有关，初步将热度定义为

$$r_i = \frac{s_i^a m_i^b}{t_i^c}$$

其中r_i为第i类的热度，a, b, c均为参数。

根据数据的实际情况，热度高的类大致分为两种，一种是总点赞数比较高，但留言个数较少；另一种是总点赞数较少，但留言个数较多。为了平衡这两种问题成为热点的概率，经过一系列调试后，确定$a = 1, b = 1.2, c = \frac{2}{3}$。最终确定热度指数的定义为

$$r_i = \frac{s_i m_i^{\frac{6}{5}}}{t_i^{\frac{2}{3}}}$$

4.6 根据热度指数排序

计算出每一类的热度指数后，对热度指数进行离差标准化，并放缩到[0,100]

之间，然后对热度指数进行降序，经过简单筛选得出热度前五的热点问题：

表 4-5 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	100	2019-11-02 至 2020-01-09	A2 区丽发新城小区居民	A 市 A2 区丽发新城小区遭搅拌站明目张胆污染环境
2	2	87.0263	2019-07-18 至 2019-09-01	A 市伊景园滨河苑居民	A 市伊景园滨河苑捆绑销售车位
3	3	67.82851	2019-10-29 至 2019-10-30	A4 区教育局教师	未落实发放文明单位奖
4	4	35.32986	2019-03-26 至 2019-04-09	A6 区月亮岛路业主	违法架设高压线
5	5	27.48664	2019-03-06 至 2019-03-06	A 市地铁	地铁换乘辛苦

其中地点/人群由 4.1.1 的命名实体识别得出，问题描述由 4.3.1 的 Seq2Seq 模型得出。

4.7 本章小结

本章通过针对命名实体识别的文本向量构造，用基于余弦相似度的聚类算法挖掘出了热点问题，并定义与计算了热点指数，并通过 seq2seq 自动生成问题描述。即本章的模型可以一键式挖掘热点并生成热点描述，利于政府工作人员直观地了解市民留言的重点，对提升政府的管理水平和施政效率具有极大的推动作用。

5 答复意见质量评价模型

图 5-1 青岛卫生监督对于政务答复意见的质量评价

根据青岛卫生监督对于政务答复意见的质量评价:“在答复工作中，注重工作质量，切实解决群众诉求。对群众反映的问题，能够解决的及时帮助解决，暂时解决不了的向群众说明情况并答复阶段性处理意见，确实无法解决的及时向群众解释清楚原因，并在办理期限内将办理结果答复来话人。“及查阅相关政务答复意见的网友意见。

我们认为一个好的答复意见应该具备以下特点：

（一）可解释，即答复涉及到的内容或论据是有依据的，明确指出问题涉及到的相关的规章制度、政府部门与民间组织，不是凭空捏造和踢皮球。

（二）时效性，即在国家规定的时间内完成政务答询。

（三）完整性，答复意见应该对问题完整回答。

（四）相关性，答复意见应与问题高度相关，而不是作假大空的回复。

因此，我们构建答复意见质量评价模型，从这四个评价指标，对答复意见的质量给出一套评价方案。

图 5-2 答复意见质量指标图

5.1 指标权重

接着利用层次分析法计算出每个指标的权重。最后，对答复意见的质量给出了一套评价方案，建立答复意见质量评价模型。

5.1.1 层次分析法

层次分析法，简称 AHP，再 20 世纪 70 年代中期由美国运筹学家托马斯·塞蒂正式提出。它是一种定性和定量相结合的、系统化、层次化的分析方法。该方法用决策者的经验判断各衡量目标之间能否实现的标准之间的相对重要程度，并合理地给出每个决策方法的每个标准的权重，利用权重求出各个方案的优劣次序，在本文中，其用于求出可解释性、时效性、完整性、相关性四个指标的权重，以建立起答复意见质量评价模型。

5.1.1.1 AHP 基本步骤

首先，**建立层次结构模型。**将决策目标、考虑因素和决策对象按照相互关系分为：最高层、中间层和最低层。最高层为决策目的；最低层为决策的备选方法；中间层为决策的准则。

然后，**构造判断矩阵。**计算出方案对应的一致矩阵。下表将列出判断矩阵a_{ij}的标度方法：

表 5-1 判断矩阵a_{ij}的标度方法

标度	含义
1	表示两个因素相比，具有同样重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素稍微重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要

2, 4, 6, 8	上述两相邻判断的中值
倒数	因素i与j比较的判断a_{ij}，则因素j与i比较的判断$a_{ji}=1/a_{ij}$

接着，**层次单排序及其一致性检验。**首先进行层次单排序，对应于判断矩阵最大特征根λ的特征向量，经过归一化后记为W，W的元素该该层元素对于上一层因素中的某因素相对重要性的排序权值。接着进行一致性检验，对此定义一致性指标$CI=\frac{\lambda-n}{n-1}$，CI越接近 0，其一致性越强。为了衡量CI的大小，引入随机一次性指标RI，其具体取值如表 2 所示：

表 5-2 随机一致性指标RI

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

依此定义**一致性比率**：$CR=\frac{CI}{RI}$，并认为一致性比率$CR<0.1$时，认为判断矩阵通过一致性检验，可用其归一化特征向量作为权向量，否则需要重新构造判断矩阵。

最后，**层次总排序及其一致性检验。**首先计算最下层对目标的组合权向量，在检验其一致性，若检验通过则按照该组合权向量表示的结果进行决策。

5.1.2 模型算法流程

Step1　输入四个指标的预设权重比

Step2　计算其一致矩阵，进行层次单排序并检验其一致性

Step3　进行层次总排序以及一致性检验

Step4　输出权向量

5.1.3 计算结果

我们预设权重比为：

$$可解释性:时效性:完整性:相关性=3:2:2:3$$

最终计算出可解释性、时效性、完整性、相关性的权重如下表所示：

指标	权值
可解释性	0.3
时效性	0.2
完整性	0.2
相关性	0.3

表 5-3 指标权重

5.2 可解释性

我们认为，可解释性即答复的内容是有依据的，不是凭空捏造的。这样，市民明确地可以知道有哪些相关文件、栏目或部门是与问题相对口的，则可以更好地帮助市民解决问题。

因此，我们采用 2.中的命名实体识别模型对答复意见进行组织、部门、作品的识别，部分识别结果如下所示：（全部结果请看/附件/过程数据/附件 4_score.xlsx）

表 5-4 命名实体识别示意表

可解释性-作品	可解释性-组织	可解释性-部门
问政西地省	景蓉华苑业委会	区住房和城乡建设局,区住房和城乡建设局
关于 A 市 A3 区既有多层住宅增设电梯实施方案	A3 区住建局,政务服务中心住建局	区人民政府,办公室
A 市 A5 区城镇小区配套幼儿园专项整治工作方案	东澜湾幼儿园,A5 区教育局,江河水利置业投资发展有限公司,A5 区教育局,A5 区卫计局,花桥门诊部,沙湾门诊部,医疗机构	区教育局,区教育局,卫生,区政府

我们认为一个高质量的答复意见，应该同时具备作品、组织和部门，因此记可解释性得分为：

$$Score_1 = \begin{cases} 100\% * 30\% * 100，\textit{三个都有} \\ 80\% * 30\% * 100，\textit{有其中两个} \\ 60\% * 30\% * 100，\textit{仅有一个} \end{cases}$$

若三个都没有，则$Score_1 = 0$

5.2.1 可解释性得分算法流程

Step1　采用命名实体识别模型识别出答复意见中的组织、部门、作品

Step2　计算

$$Score_1 = \begin{cases} 100\% * 30\% * 100，\textit{三个都有} \\ 80\% * 30\% * 100，\textit{有其中两个} \\ 60\% * 30\% * 100，\textit{仅有一个} \end{cases}$$

若三个都没有，则$Score_1 = 0$

5.3 完整性

完整性，即答复意见能完整地回答市民留言的所有问题。在留言文本中，与

市民遇到的问题相关的词语在留言文本中占据重要地位。类似地，答复市民问题的相关词语也在答复意见中占据重要地位。

因此我们对留言文本提取关键词$U=(u_1,\cdots,u_i)$，对答复意见提取关键词$V=(v_1,\cdots,v_j)$，若答复意见完整地回复了留言文本，则$\forall u_k \epsilon U, 1\le k\le i,\ \exists v_t$，使$f(u_k,v_t)$ >常数m。

即对于市民留言中的任意一个关键词，在答复意见总存在一个关键词与之高度相关。

因此，我们运用 TF-IDF 模型分别对留言详情与答复意见提取关键词。分别记为

$$U=(u_1,\cdots,u_i),\ V=(v_1,\cdots,v_j)$$

其中有$(u_1,\cdots,u_t), t\le i$的关键词在答复意见中存在与之高度相关的关键词。接着计算在市民留言详情的关键词U中，在答复意见中存在高度相关的关键词的比率k

$$k=\frac{t}{i}。$$

则

$$Score_2=k*20\%*100$$

5.3.1 完整性得分算法流程

Step1　运用 TF-IDF 模型提取留言文本与答复意见的关键词，分别记为

$$U=(u_1,\cdots,u_i),\ V=(v_1,\cdots,v_j)$$

Step2　用余弦相似度度量词之间的相关性，记为$f(u_k,v_t)$，若$f(u_k,v_t)>0.6$则认为高度相关,则$u_k\in U'$

Step3　统计U中在答复意见的关键词中存在与之高度相关的词的关键词，记为：

$$U'=(u_1,\cdots,u_t), t\le i$$

Step4　计算比率$k=\frac{t}{i}$

Step5　计算

$$Score_2=k*20\%*100$$

5.4 相关性

相关性，即答复意见与市民留言相关。这里，我们运用余弦相似度度量答复意见与市民留言的相关性。

$$\theta=\frac{A\cdot B}{\|A\|\cdot\|B\|}$$

接着计算相关性得分：

$$Score_3 = \theta * 30\% * 100$$

5.4.1 相关性得分算法流程

Step1　计算相关性系数θ

Step2　计算相关性得分

$$Score_3 = \theta * 30\% * 100$$

5.5 时效性

根据 2005 年 1 月 10 发布的《信访条例》第四章信访事项的受理中的第二十一条以及 2015 年 10 月 26 日发布的《信访事项网上办理工作规程（试行）》第三章信访工作机构的受理办理第十一条，得到自收到信访人的来信起 15 日内决定是否受理并书面告知信访人，并按要求通报信访工作机构。

因此，本文将 15 天t_0作为回复延迟的最后期限，求出每一条文本数据的留言时间与回复时间之间的间隔Δt，利用下面的公式得到文本数据的时效性得分系数S_t：

$$S_t = \frac{(t_0 - \Delta t)}{t_0}$$

S_t的取值范围为$(-\infty, 1)$，当S_t的值越接近 1 的时候，说明回复的速度越快，该文本数据的时效性越强。

同时，时效性越强，说明政府部门对人民的事物越上心。因此，我们对时效性越强的答复给予一定奖励，时效性强，会有一定额度的加分。

计算系数$(1 + 10\% * S_t)$

计算时效性得分：

$$Score_4 = \begin{cases} (1 + 10\% * S_t) * 100 * 20\%, S_t \in (-10,1) \\ 0, \quad 其他 \end{cases}$$

5.5.1 时效性得分算法流程

Step1　计算时效性得分系数

$$S_t = \frac{(t_0 - \Delta t)}{t_0}$$

其中$t_0 = 15$，Δt为回复的间隔时间。

Step2　计算时效性得分

$$Score_4 = \begin{cases} (1 + 10\% * S_t) * 100 * 20\%, S_t \in (-10,1) \\ 0, \quad 其他 \end{cases}$$

5.6 模型求解

5.6.1 模型算法流程

Step1　文本数据预处理，转换为纯文本数据

Step2　计算可解释性、完整性、相关性和时效性的得分$Score_1, Score_2, Score_3, Score_4$

Step3　计算总得分$Score = Score_1 + Score_2 + Score_3 + Score_4$

5.6.2 求解结果

部分求解结果如下所示：（完整结果见附件中/过程数据/附件 4_score.xlsx）

表 5-5 答复意见质量评价模型求解结果示意图

留言编号	相关性	时效性	可解释性-作品	可解释性-组织	可解释性-部门	完整性	score
17619	0.9725	0.6710	换届政府工作报告	A7 县委，新高地	县政府，县委，	1	100.5194
87360	0.9190	-2.2716	关于签订 2017 年度保证民生用气责任书的通知	0	住房和城乡建设局	0.75	82.0272
78506	0.8583	-0.8068	0	0	教体局	0.3333	68.8009
12451	0.4883	-2.1974	0	0	0	0	30.2541

统计

score

个案数	有效	2816
	缺失	0
平均值		81.17510068
中位数		85.46290956
标准 偏差		14.33161787
范围		84.18389499
最小值		16.99158488
最大值		101.1754799

图 5-3 答复意见质量得分统计图

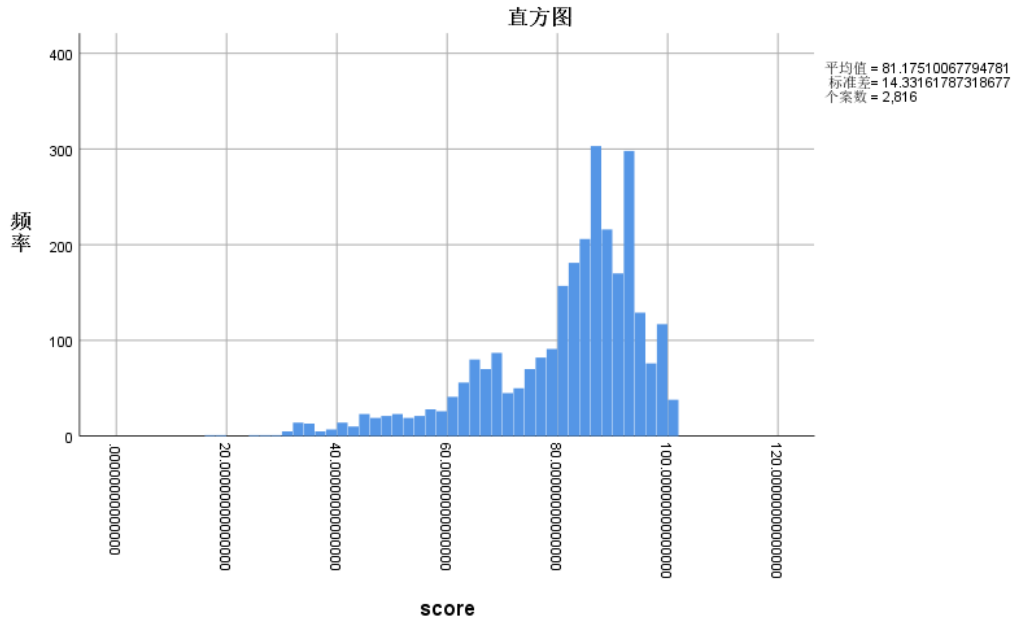

图 5-4 答复意见质量得分直方统计图

由上述图表得知，模型能很好的识别出我们认为的高质量答复与低质量答复，并且分布较为符合正态分布。

其中得分集中在区间[80,100]，说明答复意见的质量普遍较高，能很好地回答人民群众的问题。

5.7 本章小结

本章在可解释性、时效性、完整性和相关性四个维度综合评价答复意见的质量。构建了答复意见的质量评价模型，构建了一个答复意见质量评价方案，能多方面、立体化地对答复意见做出质量评估。政府能直截了当地了解到答复意见哪个方面有欠缺，需加以改进；哪个方面表现地比较好，应继续保持。同时，我们的模型根据现实情况，对时效性特别高的答复意见给予一定加分奖励，充分贴合实际，符合市民的期望。并且，模型对高质量与低质量的答复意见识别能力较强，能客观地做出符合人们期望的质量评价，对提升政府的管理水平和施政效率具有极大的推动作用。

6 总结

本文主要运用了 textcnn、xgboost 与 seq2seq 等深度学习的框架与算法，构建了 textcnn-xgboost 文本分类模型，热点挖掘模型与答复意见质量评价模型来解决三个问题。。

首先我们对附件中的市民留言及答复意见进行分词，并在语料库中检索分词，爬取语料训练词向量。进而提取群众留言的特征，构建文本分类模型。

其次，运用热点挖掘模型识别群众留言及答复意见的地点、人物等信息进行归类，挖掘特定地点特定人群的问题，再通过 seq2seq 生成问题描述。方便政府精确高效地处理问题。

最后，我们综合考虑答复意见的可解释性、完整性、相关性与时效性，建立多方面、立体的质量评价体系，有助于政府部门清楚地发现答复意见的优劣。对提升政府的管理水平和施政效率具有极大的推动作用。

参考文献

[1] 牛雪莹.结合主题模型词向量的CNN文本分类[J].计算机与现代化,2019(10):7-10.
[2] 龚维印,王力.基于卷积神经网络和XGBoost的文本分类[J].通信技术,2018,51(10):2337-2342.
[3] 杨波,杨文忠,殷亚博,何雪琴,袁婷婷,刘泽洋.基于词向量和增量聚类的短文本聚类算法[J].计算机工程与设计,2019,40(10):2985-2990+3055.
[4] Cho K , Van Merrienboer B , Gulcehre C , et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer ence, 2014.
[5] Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. Iclr 2015 1–15 (2014).