“智慧政务”中的文本挖掘应用

摘要

本文我们将利用自然语言处理和文本挖掘技术解决“智慧政务”中的文本挖掘应用问题，对群众问政留言记录进行标签分类并深入挖掘热点问题且对相关部门对部分群众留言的答复意见进行评价。

对于问题 1，先对数据进行清洗通过使用正则表达式删除留言详情中的累赘信息，从而简化数据。利用 jieba 中文分词工具对留言详情进行分词并通过自定义词典来增强歧义纠错能力解决文本歧义带来的词语交叉，再利用 TF-IDF 算法得到每个留言详情的 TF-IDF 权重向量，采用 K-eras 对 TF-IDF 权重进行聚类，得到每段详情留言的关键字向量。通过深度学习模型（CNN 卷积神经网络）对留言数据进行训练，进行一级标签分类，再通过 F-Score 对分类结果进行评价。

对于问题 2，先对去重后的留言数据其相似或者相同的留言主题内容出现次序、留言时间距当前时间的长短、点赞数和反对数这四个方面进行筛选，挖掘当前前五个热点问题。

对于问题 3，重新对留言内容进行清洗和去重操作，再清洗数据的过程中用 jieba 中文分词工具通过自定义词典来解决文本歧义带来的词语交叉，计算每段去重后留言数据的留言主题、答复意见关键字的重复率，以判别其两者的相关性，利用 SVM(Support Vector Machine)以及 word2vec 算法对留言与回复内容的相似性及时效性比较，从而得出其对应方案中的评价。

关键词：自然语言处理；文本挖掘技术；文本分类；CNN 卷积神经网络；K-eras 多标签分类；SVM(Support Vector Machine)；word2vec；

Abstract

In recent years, the online political platform has gradually become an important channel for the government to understand public opinions. However, it is extremely difficult to classify and sort out the comments and hot spots manually based on the text data of various social situations and public opinions that are constantly increasing. With the development of big data, 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. Therefore, we will use natural language processing and text mining techniques to establish the tag classification of the message content, dig into the hot issues and evaluate the replies.

For question 1, the data is first cleaned to simplify the data by using regular expressions to remove the verbatim information in the message details. Jieba Chinese word segmentation tool is used to segment message details and a custom dictionary is used to enhance the ambiguity correction ability to solve the word crossing caused by text ambiguity. Tf-idf weight vector of each message details is obtained by using tf-idf algorithm, and k-eras is used to cluster tf-idf weight to obtain the keyword vector of each message details. The deep learning model (CNN convolutional neural network) was used to train the message data, conduct first-level label classification, and evaluate the classification results by f-score.

For question 2, the four aspects of similar or identical message topic content occurrence order, message time distance from the current time, thumb up number and opposition number of the message data after reweighting are firstly screened, and the current top five hot issues are mined.

For question 3, cleaning and the content of the message to retry again, then the process of data cleaning in jieba Chinese word segmentation tools through the custom dictionary to solve text ambiguity of the words crossed, calculate each theme, and reply to leave a message after heavy data message opinions keyword repetition rate, to

identify the relevance of the two, by using the SVM (Support Vector Machine) and the contents of the message and reply word2vec algorithm similarity and timeliness, so as to obtain the corresponding scheme of evaluation.

Key words: natural language processing; Text mining technology; Text classification; CNN convolutional neural network; K-eras multi-label classification；SVM(Support Vector Machine)；word2vec；

目录

第一章 引言

1.1 研究背景与意义

1.1.1 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、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，且对提升政府的管理水平和施政效率具有极大的推动作用。但同时，在智慧政务建设过程中，也面临着地区间发展不均衡、业务协同不足、数字鸿沟障碍、信息安全威胁等问题。推动智慧政务的发展，将会是一个长期的过程。应在政府主导下，以促进电子政务与信息技术的深度融合为手段，以满足公众需求的智慧服务供给为目标，积极推动政府治理模式和政府职能的转换，实现智慧政务的透明化、高效化与智慧化发展[1]。

1.1.2 意义

智慧城市建设的核心是智慧政府，国家信息中心信息化研究部于施洋等认为，在智慧政府建设中，“智慧”代表着对事物能迅速、灵活、正确地理解和处理的能力[2]，智慧政务是智慧政府的基础，没有实现智慧政务的智慧政府只是空中楼阁。”智慧政务“文本挖掘应用有利于实现“政府迅速、灵活、正确地理解和处理”公众相关事项，实现多部门协同服务。自然语言处理和文本挖掘技术对“智慧政务”中的文本进行挖掘应用，有利于政务信息协同的分级分步建设，促进实现政务信息高效协同推动政务服务智慧化[3]。

1.2 论文研究内容与章节安排

本文的主要研究内容为：对收集自互联网公开来源的群众问政留言记录及相关部门对部分群众留言的答复意见，通过自然语言处理和文本挖掘技术解决“智慧政务”中的文本挖掘应用问题，对群众问政留言记录进行一级标签分类并挖掘热点问题，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进行评价。

章节安排为：第一章主要讲述“智慧政务”文本挖掘应用的研究背景以及研

究意义；第二章为通过自然语言处理和文本挖掘技术进行“智慧政务”文本挖掘应用的理论依据；第三章对研究问题进行重述以及具体分析；第四章是对第一题的详细分析，主要包括对数据进行清洗、文本分词、去停用词、文本特征选择、深度学习模型训练、模型验证、调优以及结果展示；第五章是对第二题所述问题的详细解答，在第一题对数据进行清理的基础下对相似主题进行归类且排序，再结合多因子约束选取前 5 个热点问题；第六章是对第三题的详细分析，计算去重后的留言主题、答复意见关键字的重复率判别两者的相关性，利用 SVM(Support Vector Machine)以及 word2vec 算法对留言与回复内容的相似性及时效性比较，得出评价；第七章是对最终所得结果以及模型表现进行阐述，表明所选模型的优点与不足以及对于该项工作的未来展望。

第二章 基础理论依据

2.1 自然语言处理技术

自然语言处理 (Natural Language Processing, NLP) 技术是研究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[4]。按照不同的分类标准、基于不同的分类原则，自然语言处理相关技术的分类结果也有所不同。

2.2 文本分类

文本分类(Text Classification 或 Text Categorization，TC)，或者称为自动文本分类(Automatic Text Categorization)，是指计算机将载有信息的一篇文本映射到预先给定的某一类别或某几类别主题的过程。文本分类另外也属于自然语言处理领域。本文中文本(Text)和文档(Document)不加区分，具有相同的意义。

2.3 TF-IDF 算法

TF-IDF 是一种统计方法，用以评估单词对于一个文件集或一个语料库中的其中一份文件的重要程度。如果某个单词在一篇文档中出现频率高，并且在其他文章中出现的频率低，则认为这个单词有很好的类别区分能力[5]。频词（Term-Frequency,TF)，衡量一个 term 在一篇文档中出现的频繁次数。逆文档频率（InverseDocumentFrequency,IDF)，是 一个词语普遍重要性的度量[6]。

2.4 CNN 卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）是一类包含卷积计算且具有深度结构的前馈神经网络（Feedforward Neural Networks)，是深度学习（deep learning）的代表算法之一[7-8]。卷积神经网络具有表征学习（representation learning）能力，能够按其阶层结构对输入信息进行平移不变分类（shift-invariant classification)，因此也被称为“平移不变人工神经网络（Shift-Invariant Artificial Neural Networks, SIANN)”[9]。

2.5 Word2vec

Word2vec 模型 Word2vec 模型是由 Google 的 TomasMikolov 团队提出并实现的分布式词向量表示模型，普遍应用于自然语言处理 (NLP)。该模型可以在较短的时间内，从大规模的语料库中学习到高质量、多角度表达的词向量[10]。一篇文档可

以通过这种模型得到该文档中每个词的低维度（100-500）向量表达，从而可以方便的计算词与词之间的语义相似度[6]。

2.6 SVM

支持向量机（Support Vector Machine, SVM）是一类按监督学习（supervised learning）方式对数据进行二元分类的广义线性分类器（generalized linear classifier),其决策边界是对学习样本求解的最大边距超平面(maximum-margin hyperplane）SVM 使用铰链损失函数（hinge loss）计算经验风险（empirical risk）并在求解系统中加入了正则化项以优化结构风险（structural risk)，是一个具有稀疏性和稳健性的分类器。SVM 可以通过核方法（kernel method）进行非线性分类，是常见的核学习（kernel learning）方法之一

第三章 问题重述与具体分析

3.1 问题重述

1.1.1 群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（三级标签体系）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 差错率高等问题。本文我们将依据一定的划分体系对收集的公开的问政留言建立关于留言内容的一级标签分类模型，且使用 F-Score 对分类方法进行评价：

$$\mathrm{F_1}=\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\frac{2P_iR_i}{P_1+R_i} \tag{1}$$

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率

1.1.2 热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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。本文将对留言内容中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，定义合理的热度评价指标，并给出评价结果，按表 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	小区临街餐饮店油烟噪音扰民
2	2	…	2017/06/08 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，嗓声严重扰民，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
…	…	…	…	…	…	…	…

1.1.3 答复意见的评价

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、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

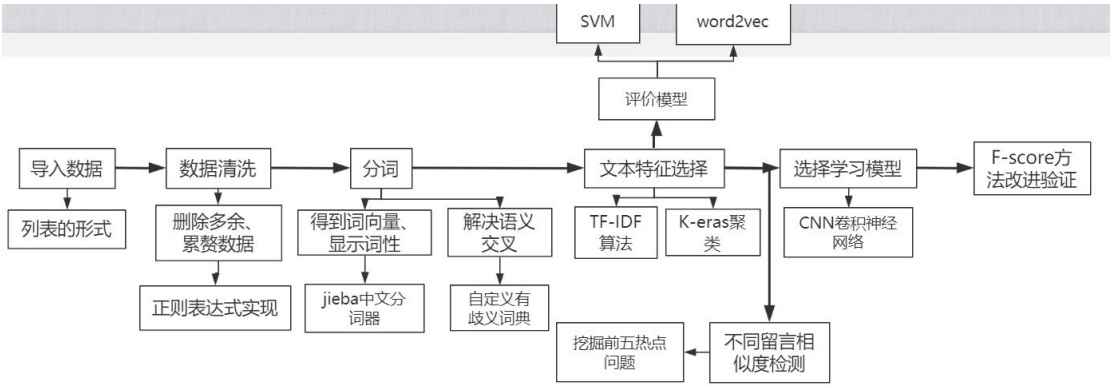
3.2 总体研究思路与框架

将数据以列表的形式导入，得到包含全部数据的列表，用正则剔除数据中重复、无用的词进行数据清洗，再用 jieba 中文分词器进行分词得到各种词向量并且给予词性，自定义词典来增强歧义纠错能力解决文本歧义带来的词语交叉（例如各种地名或者小区的各种简称等同于其全称），再利用 TF-IDF 算法得到每个留言详情的 TF-IDF 权重向量，采用 K-eras 对 TF-IDF 权重进行聚类，得到每段详情留言的关键字向量。再通过深度学习模型（CNN 卷积神经网络）对留言数据进行训练，进行一级标签分类，最后对所求模型通过 F-Score 方法进行改进和评估。

对去重后的留言内容，通过问题 ID 对比不同留言主题的关键词向量，筛其出相似度较高的留言并记录不同留言内容相似流言的个数，通过留言条数的比较得到 5 个当前热点问题，在此基础上再根据结合留言时间距当前时间的长短、点赞数和反对数这三个方面进行筛选，得到较为全面且合理的当前前五个热点问题。对于问题 3，则需重新对留言内容进行清洗和去重操作，再清洗数据的过程中用 jieba 中文分词工具通过自定义词典来解决文本歧义带来的词语交叉，计算

每段去重后留言数据的留言主题、答复意见关键字的重复率，以判别其两者的相关性，利用 SVM(Support Vector Machine)以及 word2vec 算法对留言与回复内容的相似性及时效性比较，从而得出其对应方案中的评价。

总体研究思路思路框架图示列如下：



第四章 第一小题详细解答

4.1 数据导入

将附件 2 的内容都以列表的形似导入，得到完整的数据，同理将附件 1 的一级分类标签以列表的形式导入

4.2 数据清洗

用正则表达式提出文本中重复、累赘词语（例如，尊敬的**，您好、以及空格空行）从而达到简化数据的目的。接着通过自定义词典来增强歧义纠错能力解决文本歧义带来的词语交叉（例如各种地方、地区的简称造成的歧义）

4.3 jieba 分词

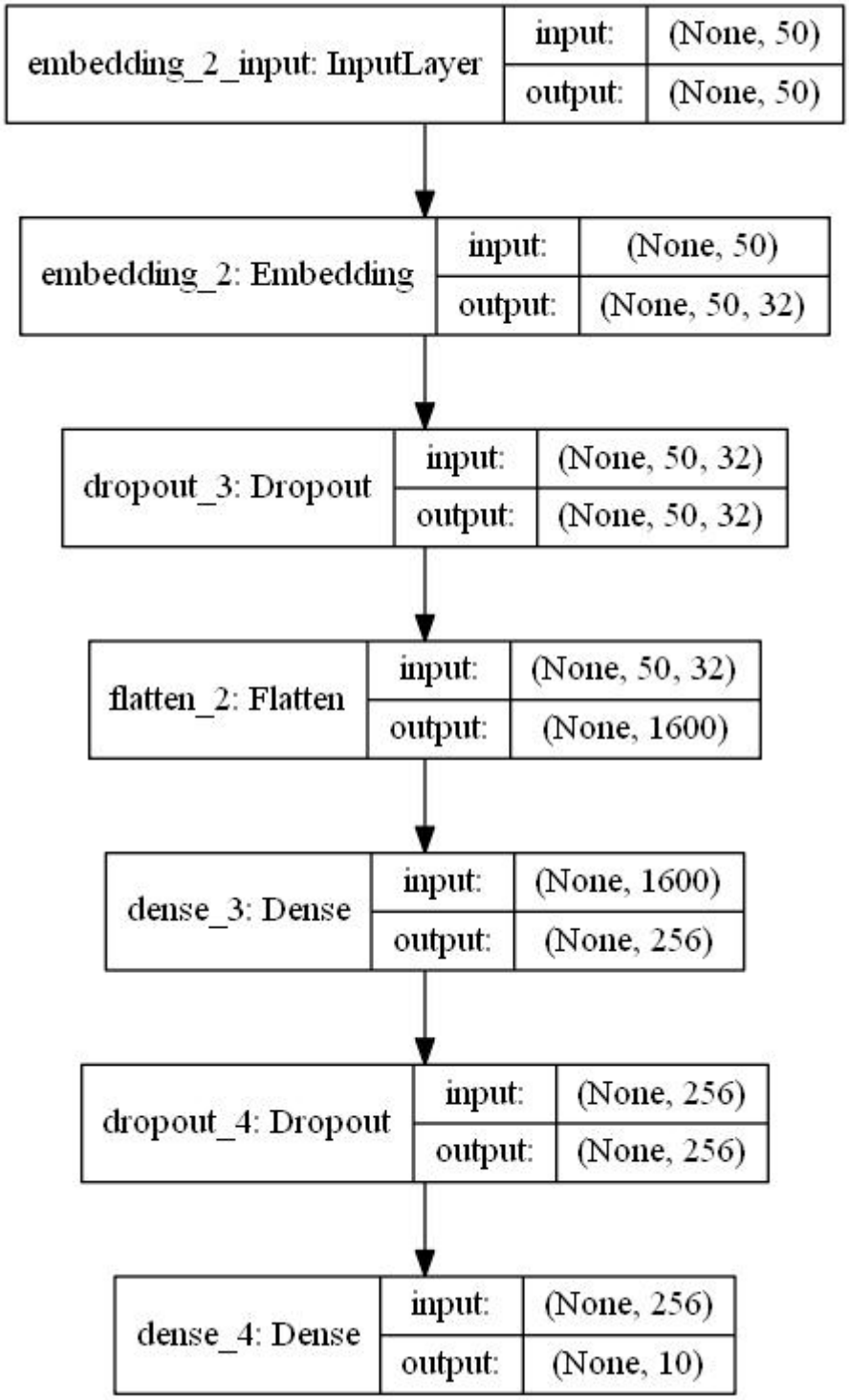
利用 jieba 中文分词工具对留言详情进行分词

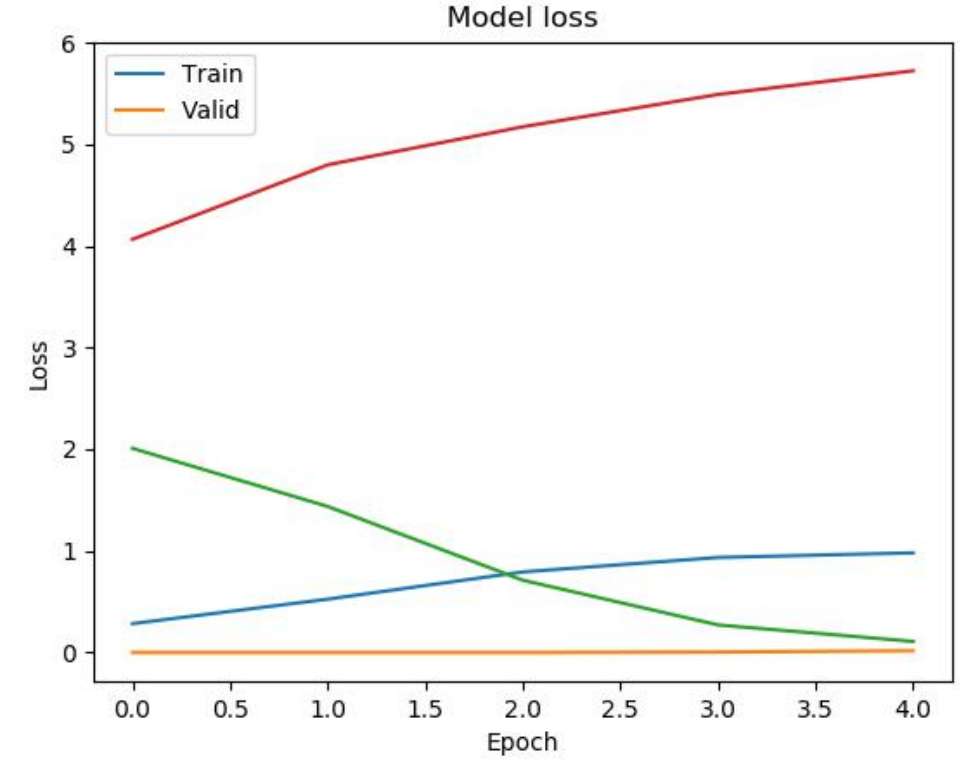
4.4 提取文本特征

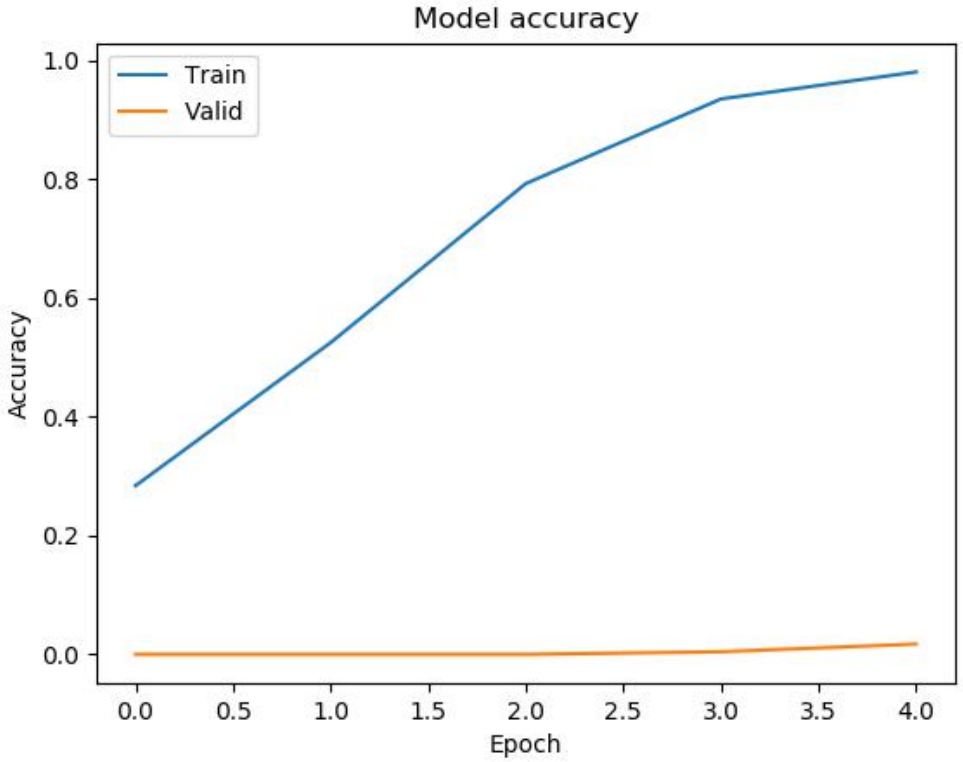
利用 TF-IDF 算法得到每个留言详情的 TF-IDF 权重向量，采用 K-eras 对 TF-IDF 权重进行聚类，得到每段详情留言的关键字向量。

4.5 模型学习 标签分裂

通过深度学习模型（CNN 卷积神经网络）对留言数据进行训练，进行一级标签分类，再通过 F-Score 对分类结果进行评价。




Model loss
Train
Valid
Loss
Epoch
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Model accuracy
Train
Valid
Accuracy
Epoch
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

第五章 第二小题详细解答

5.1 热点问题挖掘思路概述

基于对第一题对数据进行清理的基础下对相似主题进行归类且排序，我们给出通过简单的列表、字典、文本操作选取前 5 个热点问题。

5.2 热点评价指标

①主要标准依据计算相似主题出现次数排序，如果主要（第一）指标相同；②次要指标出现时间最早，更次要指标是点赞数和反对数

5.3 具体步骤

先提取留言主题放入列表中，再用集合剔除重复的列表中留言主题，再循环计算列表重复出现的元素的次数；按字典集合中，每一个元组的第二个元素排列，遍历出来的一个元组，便能找出排名前十的留言主题，为了进一步确认前 5 个热点问题，还要将排名前十的留言主题作为指标从附件三中找出相似留言主题，基于排名前十的留言主题用计算数器分别计算排名前十的留言主题出现的次数；如果第一指标相同，次要指标就比较出现时间最早；前两个指标相同，更次要指标就比较是点赞数和反对数。

第六章 第三小题详细解答

针对相关部门对留言的回复进行评价方案：

这里我们会从其回复与留言的相关性，完整性和时效性（及时性）对这相关部门的回复进行评价。

首先从相关性这方面进行比较，我们在如何评判留言与回复的相关性时是采用了比较留言与回复两者之间的文字重复率，也就是关键词的重复率，从而比较出两者之间的相关性，在此会分为四个等级：“与用户留言相关性高”，“与用户留言相关性一般”，“与用户留言相关性不大”，“与用户留言没有相关性”例如附件 4 中的留言编号为 2549 留言用户为 A00045581 的关于“A2 区景蓉华苑物业管理有问题”的留言，从给出的材料中可以看出该留言与回复之间的关键词重复率不高，所以该留言的回复可能与该留言的相关性不高，我们就会给出一个为“与留言相关性不大”的评价。

其次，我们会从留言的完整性进行评价，如何去评判留言回复的完整性，这是个

难题，在这里我们是从回复对留言问题中有没有提及留言者，有没有对问题有一个简单的复述，有没有在留言中说出如何去解决问题以及从回复中涉及的流程进行评价，在此我们同样会分为四个等级：“该回复完整性高”，“该回复完整性一般”，“该回复完整性较低”，“该回复完整性低”，同样是看附件 4 中的留言编号为 2549 留言用户为 A00045581 的关于“A2 区景蓉华苑物业管理有问题”的留言，从该留言的回复中可以看出该留言是有对留言问题进行一个简单的复述的，并且该留言是有提及到留言者的，而且是对其在解决时的流程有一定的描述的，还有就是该回复是有给出解决方法和解决结果的，所以在完整度这方面我们是会给出“该回复完整性高”的评价。

然后就是我们会从回复的时效性进行评价，时效性在这里也就是相关部门在对留言进行回复时是否及时回复进行评价，而且回复是否及时这是对相关部门回复的评价所占比较高的，我们在此会将其分为几个等级，首先是回复对留言的时长不超过 5 天的为“回复问题速度快”，对回复对留言的时长超过 5 天，不超过 15 天的为“回复问题速度一般”，对回复对留言的时长超过 15 天，不超过 30 天的为“回复问题的速度较慢”，对回复对留言的时长超过 30 天的将会被评为“回复问题的速度非常慢”，这里举与上面同样一个例子，留言编号为 2549 的留言，留言时间为：2019/4/25 9:32:09，而相关部门的回复时间为：2019/5/10 14:56:53，可以看出该留言是属于超过 15 天但不超过 30 天的，所以在此会得出该留言回复在时效性这方面的评价为“回复问题的速度较慢”。

最后我们会综合以上三种评价方式的结果，再次分成四个等级：“该回复用户非常满意”，“该回复用户较为满意”，“该回复用户不太满意”，“该回复用户很不满意”，具体的分级标准为：若为“该回复用户非常满意”：则需相关性高，完整性高，回复问题速度快（其中有两个评判标准为 1 级且有一个标准为 2 级的亦可归为此类）；若为“该回复用户较为满意”，则需有两个评判标准为 2 级，另外一个标准为 3 级及以上；若为“该回复用户不太满意”，则需有两个评判标准为 3 级，另外一个标准为 4 级及以上；若为“该回复用户很不满意”，则是 3 个评判标准均为 4 级。

具体实现我们会用到 SVM(Support Vector Machine)这种判别方法，主要使用多类 SVM，多类 SVM 旨在通过使用支持向量机为实例分配标签，其中标签从有限的

几个元素集中绘制。这样做的主要方法是将单个多类问题减少为多个二进制分类问题。这种减少的常见方法包括:构建二进制分类器，区分(i)标签之间的一个和其余的(一对全部)或(ii)每对类之间(一对一)。一对一的情况的新实例的分类是通过获胜者采取所有策略完成的，其中具有最高输出函数的分类器分配类(重要的是输出函数被校准以产生可比较的分数）。对于一对一的方法，通过最大胜利投票策略进行分类，其中每个分类器将实例分配给两个类中的一个，然后对所分配的类的投票增加一个投票，最后是最多选票的课程决定了实例分类。

此外还会用到 word2vec 算法，基于词语相似度和文本长度，构造文本间相似度。需要引入函数 WORDSIM (wi, wpj)，对留言与回复进行比较关键词的相似度。

参考文献

[1] 颜培霞.我国政府治理现代化进程中的智慧政务建设研究[J].经济动态与评论,2018(01):175-184+194-195.

[2] 于施洋,杨道玲,王璟璇,张勇进,王建冬.基于大数据的智慧政府门户:从理念到实践[J].电子政务,2013(05):65-74.

[3] 马捷,蒲泓宇,张云开.基于复杂网络分析的智慧政务信息协同结构及特征研究——以深圳市为例[J].情报理论与实践,2020,43(01):24-32

[4] 李静,罗文华,林鸿飞.自然语言处理技术在网络案情分析系统中的应用[J].计算机工程与应用,2012,48(03):216-220.

[5] 白杨.大数据环境下的文本挖掘教学内容探讨[J].无线互联科技,2018,15(09):86-87.

[6] 李光明,潘以锋,周宗萍.基于自然语言处理技术的学生管理论坛文本挖掘与分析[J].智库时代,2019(29):122+127.

[7] Goodfellow, I., Bengio, Y., Courville, A.. Deep learning (Vol. 1). Cambridge：MIT press，2016

[8]Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G. and Cai, J., 2015. Recent advances in convolutional neural networks. arXiv preprint arXiv:1512.07108.

Goodfellow, I., Bengio, Y., Courville, A.. Deep learning (Vol. 1). Cambridge：MIT press，2016
[9]Zhang, W., 1988. Shift-invariant pattern recognition neural network and its optical architecture. In Proceedings of annual conference of the Japan Society of Applied Physics.
[10]陈慧. 基于特征细分的中文情感分析研究[D].上海师范大学,2018.