基于机器学习的“智慧政务”文本挖掘应用

摘　要

随着互联网日新月异的发展，人们越来越离不开网络。尤其在当下人人都能在互联网虚拟世界中畅所欲言，非结构化数据不断增加，因此，文本挖掘的价值定位和流行度也不断上升。越来越多的机构意识到利用文本挖掘从他们的文本资源库中提取知识的重要性。

本文在针对比赛给出的来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系列数据集做出了以下三个问题的探讨。

问题 1：群众留言分类。本文采用了基于 SVM 一类对多类的方法对群众留言进行多分类建模。对数据集进行正则化清洗、jieba 分词、TF-IDF 词向量化后，在 9210 条数据按照 7:3 比例划分训练集和测试集的情况下得到了 0.9 的 F1 值。

问题 2：热点问题挖掘。本文首先定义了留言热度 Hot；其次用余弦相似度将“留言主题”分成不同的相似类；最后使用“留言主题”匹配相同字符串和正则表达式的方式来提取留言地点或人群。

问题 3：答复意见的评价。本文对相关部门答复意见的评价主要有三个指标：相关性、完整性以及时效性。相关性的评价利用余弦相似度以及词频向量计算留言详情以及答复意见的相似度来表达；完整性的评价则是采用 LDA 主题模型和余弦相似度进行主题覆盖度表达；时效性的评价则是利用留言时间到答复时间的时间跨度来表达。其创新在于利用主题覆盖度评价回复的完整性以及利用时间数据得出回复的时效性，最后将文本相似度、主题覆盖度以及时效性三者按权重结合得出答复意见的最终评价得分。

【关键词】SVM 多分类　余弦相似度　TF-IDF　主题模型

Abstract

With the rapid development of Internet, people are increasingly inseparable from the Internet. Especially at present, everyone can speak freely in the virtual world of the Internet, unstructured data is increasing, so the value orientation and popularity of text mining are also rising. More and more organizations realize the importance of using text mining to extract knowledge from their text databases.

This paper discusses the following three problems about the records of the public political messages from the Internet and the response data set of some public messages from relevant departments.

Problem 1: classification of mass message. In this paper, one-against-all method based on SVM is used to model the mass message. After regular cleaning, Jieba word segmentation and TF-IDF word vectorization, the F1 value of 0.9 is obtained when 9210 pieces of data are divided into training set and test set according to 7:3 ratio.

Problem 2: hot issues mining. In this paper, we first define the hot degree of message; secondly, we use cosine similarity to divide "message subject" into different similar classes; finally, we use "message subject" to match the same string and regular expression to extract the message location or crowd.

Problem 3: evaluation of the responses. In this paper, there are three main indicators to evaluate the response of relevant departments: relevance, integrity and timeliness. Correlation evaluation uses cosine similarity and word frequency vector to calculate message details and reply opinion similarity to express; integrity evaluation uses LDA topic model and cosine similarity to express topic coverage; timeliness evaluation uses the time span from message time to reply time to express. Its innovation lies in the use of topic coverage to evaluate the integrity of the reply and the use of time data to get the timeliness of the reply. Finally, text similarity, topic coverage and timeliness are combined according to the weight to get the final evaluation score of the reply opinion.

【Key words】 SVM multi classification; cosine similarity; TF-IDF; LDA

目 录

摘 要 I

Abstract II

1 引言 1

2 问题 1：群众留言分类 1

2.1 问题背景 1
2.2 问题描述 1
2.3 模型分析 2
2.3.1 多分类问题描述 2
2.3.2 one-against-all 2
2.4 问题分析求解 3
2.4.1 数据预处理 3
2.4.2 jieba 分词 3
2.4.3 词向量化 4
2.4.4 建模 5

3 问题 2：热点问题挖掘 6

3.1 问题背景 6
3.2 问题描述 6
3.3 问题分析求解 7
3.2.1 热点问题定义 7
3.2.2 留言相似性计算 7
3.2.3 提取留言地点或人群 8
3.2.4 结果分析 8

4 问题 3：答复意见评价 9

4.1 问题背景 9
4.2 问题描述 9
4.3 问题分析求解 10
4.3.1 数据预处理 10
4.3.2 问题求解及工具 10
4.3.3 数据融合 14

5 不足与展望 15

参考文献 16

基于机器学习的“智慧政务”文本挖掘应用

1 引言

随着互联网日新月异的发展，人们越来越离不开网络。尤其在当下人人都能在互联网虚拟世界中畅所欲言，非结构化数据不断增加，因此，文本挖掘的价值定位和流行度也不断上升。越来越多的机构意识到利用文本挖掘从他们的文本资源库中提取知识的重要性。

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，对提升政府的管理水平和施政效率具有极大的推动作用。

本文针对比赛给出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等系列数据集进行文本挖掘。主要处理以下三个问题：

1.群众留言分类。将群众的留言进行大类区分以便将其划分至对应的部分进行处理；

2.热点问题挖掘。网络舆情一触即发，整理群众的留言内容便于有关部分及时地掌握舆情走向，早预防、早处理；

3.答复意见的评价。对有关部分的答复进行量化评分，从数据角度得到相关部分工作的效率和态度，让政务得以向前发展。

接下来将依次对每个问题进行详细分析。

2 问题 1：群众留言分类

2.1 问题背景

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对留言进行分类，以便后续将群众留言分派至对应的职能部门处理。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。

2.2 问题描述

为解决繁重的手工作业以及提升留言分类的准确率，我们基于机器学习对数据集《用户留言详情表》建立关于留言内容的一级标签分类模型。

该问题涉及两张表，其中《用户留言详情表》涉及“留言编号”、“留言用户”、“留言主题”、“留言时间”、“留言详情”、“一级分类”共 6 个字段，其结构如图 2.2-1 所示。表含有相应热点问题对应的留言信息以及对应留言详情的一级分类

标签。

留言编号	留言用户	留言主题	留言时间	留言详情	一级分类
744	A089211	建议增加 A 小区快递柜	2019/10/18 14:44	我们是 A 小区居民...	交通运输

图 2.2-1 《用户留言详情表》结构

《内容分类三级标签体系》涉及“一级分类”、“二级分类”、“三级分类”、共 3 个字段，其结构如图 2.2-2 所示。表中含有留言信息可能涉及的分类标签。此次数据集共有 15 个一级分类。

一级分类	二级分类	三级分类
城乡建设	安全生产	事故处理
城乡建设	安全生产	安全生产管理
城乡建设	安全生产	安全隐患
...	...	...

图 2.2-2 《内容分类三级标签体系》结构

该问题要解决的问题是：对《用户留言详情表》中的“留言详情”进行分析后，参照《内容分类三级标签体系》表中的“一级分类”标签，最后给出该条留言的一级分类标签。最后选择 F1-score 对模型进行评价。

2.3 模型分析

该问题是一个纯文本多分类的问题。在传统的机器学习中，我们选取了支持向量机 SVM 进行多分类建模。

2.3.1 多分类问题描述

多类分类问题描述如下：

给定含 N 个样本的训练集$\mathrm{X}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 K 维特征向量$x_n\in R^K$，类标签$y_n\in\{1,2,\cdots,M\},n=1,2,\cdots,N$。训练集数据共 M 个类。任务是找到决策函数$\mathrm{y}=f(x)$或者规则用于预测新数据的类别。

2.3.2 one-against-all

多分类的方法有“成对分类方法”和“一类对余类”两种方法。本文中采用的是“一类对余类”方法，也称为 one-against-all 或 one-against-the-rest。

成对分类方法是基于 binary SVM 的，最早实现 SVM 对多类别进行分类就是这种方法。one-against-one 适合实际应用，也是 LIBSVM 库采用的方法。其含义为：对于每一个类，将其作为正类（标签为 1），而其余 M-1 个类的所有样本作为负类（标签为 0），构造一个 binary SVM，训练数据找出这两大类的决策边界。在训练时，对于 k 个类别的样本数据，需要训练 k 个 SVM 二类分类器，在构造第 i 个 SVM 子分类的样本数据标记为正类，其他不属于 i 类别的样本数据标记为负类。依次对剩下的类进行这样的二分类操作。直至每个类别都作为正类

训练出模型。每个模型都有一个类别预测值，最后选择最大值所属类作为该类别的预测类别输出。

该方法仅训练 k 个分类器，个数较少，其分类速度相对较快。但每个分类器的训练都是将全部的样本作为训练样本，这样在求解二次规划问题时，训练速度会随着训练样本的数量的增加而急剧减慢；同时由于负类样本的数据要远远大于正类样本的数据，从而出现了样本不对称的情况，且这种情况随着训练数据的增加而趋向严重。解决不对称的问题可以引入不同的惩罚因子，对样本点来说较少的正类采用较大的惩罚因子 C。当有新的类别加进来时，需要对所有的模型进行重新训练。

2.4 问题分析求解

本文对问题 1 群众留言分类在研究框架如图 2.4-1 所示。

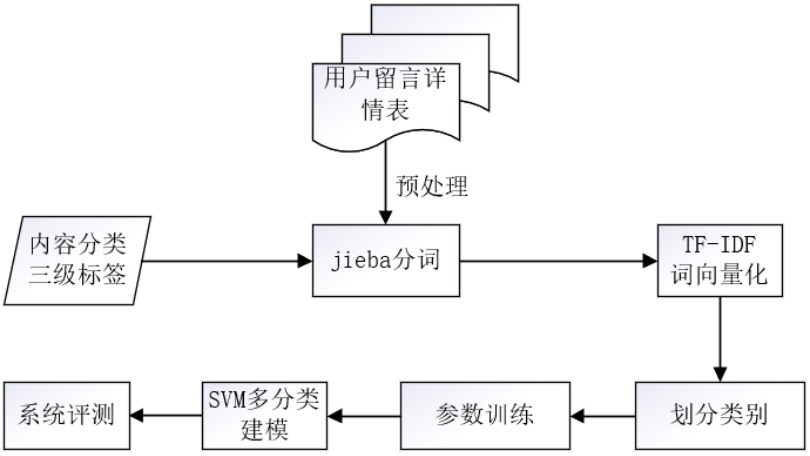


图 2.4-1 群众留言分类研究框架

2.4.1 数据预处理

此次实验需要《用户留言详情表》中的“留言详情”文本进行数据处理。采常用的文本预处理方法，我们需要采用正则化对文本中的特殊无意义符号，如“\n”、“\t”、“\r”等进行删除处理。

2.4.2 jieba 分词

数据清洗完成之后，对文本采用 jieba 分词进行分词操作。本文中采用精确模式分词。

一、jieba 的三种分词模式

jieba 是一个 python 实现的中文分词组件，在中文分词界非常出名，支持简、繁体中文，还可以加入自定义词典以提高分词的准确率。Jieba 分词结合了基于规则和基于统计这两类方法。其提供了三种分词模式：

精确模式，试图将句子最精确地切开，适合文本分析；

搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合

用于搜索引擎分词；

全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义。

二、参数简介

为实现以上三种分词效果，我们可使用 jieba.cut 和 jieba.cut_for_search 两个方法进行分词，两者所返回的结构都是一个可迭代的 generator，可使用 for 循环来获得分词后得到的每一个词语，或者直接使用 jieba.lcut 以及 jieba.lcut_for_search 直接返回 list。其中：

jieba.cut 和 jieba.lcut 接受 3 个参数：

1.需要分词的字符串（unicode 或 UTF-8 字符串、GBK 字符串），但在实际运用中，尽量不要使用 GBK 字符串，因为其可能错误解码成 UTF-8 而导致分词结果出错；

2.cut_all 参数：是否使用全模式，默认值为 False；

3.HMM 参数：用来控制是否使用 HMM 模型，默认值为 True。

而 jieba.cut_for_search 和 jieba.lcut_for_search 仅需要前 2 个参数即可。

HMM 模型，即隐马尔可夫模型（Hidden Markov Model, HMM），是一种基于概率的统计分析模型，用来描述一个系统隐性状态的转移和隐性状态的表现概率。在 jieba 中，对于未登录到词库的词，使用了基于汉字成词能力的 HMM 模型和 Viterbi 算法，其大致原理是：

采用四个隐含状态，分别表示为单字成词，词组的开头，词组的中间，词组的结尾。通过标注好的分词训练集，可以得到 HMM 的各个参数，然后使用 Viterbi 算法来解释测试集，得到分词结果。

2.4.3 词向量化

本文采用 TF-IDF 进行词向量化，得到每个词的向量化表示，以便后期将其送入模型中进行训练。在进行此向量化之前采用哈工大词典进行去停用词处理，减少停用词的干扰。

TF-IDF 算法，其英文全称为 term frequency–inverse document frequency，是一种信息检索和数据挖掘科目中较为常见的加权技术。该技术采用一种统计方法，根据字词在文本中出现的次数和在整个语料中出现的文档频率来计算一个字词在整个语料中的重要程度。它的优点是能过滤掉一些常见的却无关紧要本的词语，同时保留影响整个文本的重要字词。计算方法如公式（2.4.3-1）所示。

$$\mathrm{TF-IDF}_{i,j} = TF_{i,j} \times IDF_{i,j} \qquad (2.4.3\text{-}1)$$

由英文单词 Term Frequency 可知 TF 即代表词频，表示一个词条 t 出现在某一文档中的次数，出现次数越多代表着该词条的重要性越高，但需要特别关注的是，次数的提高并不能正比推导出相关度的提高，因此大多数时候 TF 都是需要

做平滑处理的。本文中采用最小支持度（词频数）10，以便帮助过滤掉出现太多的无意义词语。TF 词频的计算方式如公式（2.4.3-2）所示。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad (2.4.3\text{-}2)$$

其中，$n_{i,j}$为特征词t_i在文本d_j中出现的次数，$\sum_k n_{i,j}$是文本d_j中所有特征词的个数。计算的结果即为某个特征词的词频。

同理，IDF 在学术上即被称为逆文本频率指数，其与包含词条 t 的文档数目有关。文本频率是指某个关键词在整个语料所有文章中出现的次数。逆文档频率又称为倒文档频率，它是文档频率的倒数，主要用于降低所有文档中一些常见却对文档影响不大的词语的作用。IDF 计算方式如公式（2.4.3-3）所示。

$$IDF_{i,j} = \log_{10} \frac{|D|}{1 + |D_{t_i}|} \qquad (2.4.3\text{-}3)$$

其中，$|D|$表示语料中文本的总数，表示文本中包含特征词t_i的数量。为防止该词语在语料库中不存在，即分母为 0，则使用$1 + |D_{t_i}|$作为分母。

2.4.4 建模

本文采用径向基函数（RBF）做 SVM 的核函数，其中涉及的参数为 gamma 和 C。gamma 和 C 参数值的热力图（来自网络）如图 2.4.4-1 所示。

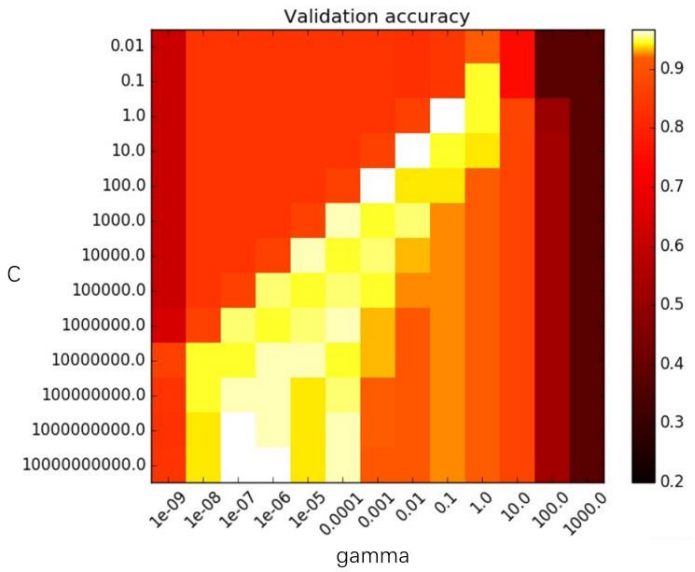


图 2.4.4-1 gamma 和 C 的热力图

由图 2.4.4-1 可见，参数 gamma 定义了单个训练样本的影响大小。参数 gamma 可以看作被模型选中作为支持向量的样本的影响半径的倒数。模型的行为对于参数 gamma 十分敏感。如果参数 gamma 过大，支持向量的影响半径将小到只能影响到它自己，这时无论如何调整参数 C 也不能避免过拟合。当参数 gamma 非常小时，模型会过于拘束不能捕捉到数据的复杂性或“形状”。任何选中的支持向

量的影响区域将包含整个训练集。模型的结果将表现得像是用一组超平面分割两类或多类的高密度中心的线性模型。

参数 C 在误分类样本和分界面简单性之间进行权衡。低的 C 值使分界面平滑，而高的 C 值通过增加模型自由度以选择更多支持向量来确保所有样本都被正确分类。

在此次实验中，我们选择了 GridSearchCV 来训练参数。GridSearchCV 可以保证在指定的参数范围内找到精度最高的参数，但这也是网格搜索的缺陷所在，其遍历所有可能参数的组合，在面对大数据集和多参数的情况下，非常耗时。

本文的参数训练范围为 C：[1e-3, 1e-2, 1e-1, 1, 10, 100, 1000]；gamma：[0.001, 0.0001]。对于每一个 one-against-all 的 SVM 选取最佳参数进行建模，然后再应用至训练集中。

将全部数据集按照 7:3 的比例划分训练集和测试集，训练模型，得到如表 2.4.4-1 所示的结果。表中给出了每条留言属于某一类别的概率，以及该条留言最后的类别预测 predict 值。尽管全部数据集已经 9000 多条，目前模型的 F1 值也能达到 0.9，但是对于机器学习而言，其数据量仍然不够，更大的数据集将得到更好的效果。

表 2.4.4-1 热点问题表

留言编号	留言用户	留言主题	留言时间	留言详情	一级分类	label	城乡建设	predict	环境保护	交通运输	商贸旅游	卫生计生	教育文体	劳动和社会保
67969	U0005844	西地省为	2015/3/2		劳动和社会保障		-0.95229	劳动和社	-1.18778	-1.55122	-1.58125	-1.66726	-2.04012	1.832707
150184	U000968	请问过去	2016/3/2		卫生计生		-1.96903	卫生计生	-1.07495	-1.5958	-1.36102	0.741261	-0.22438	-1.12296
24445	U0006799	A市汽车南	2013/10/	我 姑娘	交通运输		-0.92257	交通运输	-0.75227	-0.10919	-1.1719	-1.61818	-1.05236	-1.55725
179646	U000154	C4市潭市	2016/9/1		教育文体		-1.61657	教育文体	-1.67027	-1.2705	-1.12991	-1.21356	0.992787	-1.32636
30877	U0004274	请您帮忙	2012/6/1		卫生计生		-1.87211	卫生计生	-1.31313	-1.5397	-0.83459	0.308973	-0.97703	-0.96663
24876	U0003177	如何申办	2012/3/3		教育文体		-0.90242	卫生计生	-0.83312	-1.0844	-1.08468	-0.2094	-0.60021	-0.80784
228847	U0004497	C2区葩金	2017/7/2		教育文体		-1.15859	教育文体	-1.62662	-1.50641	-2.41747	-1.5252	1.900348	-1.27033
229752	U0005438	G市下岗买	2018/1/8		劳动和社会保障		-1.84919	劳动和社	-1.77982	-1.5863	-1.62745	-1.7561	-1.69075	2.222518
129260	U0007574	咨询L9县	2018/12/	昨天 有事	商贸旅游		-0.67896	商贸旅游	-1.22417	-0.90481	-0.1372	-0.72866	-0.89253	-1.29611
119288	U0008139	请E4县教	2015/5/1		教育文体		-1.58044	教育文体	-1.59224	-0.99118	-1.51441	-1.38625	1.607313	-1.50134
116279	U0003703	咨询关于	2015/5/7		劳动和社会保障		-1.82383	劳动和社	-1.46158	-1.54353	-1.48173	-1.26818	-1.84064	1.720583
38146	U0008381	南方机动	2012/8/3		劳动和社会保障		-1.93194	劳动和社	-1.44588	-1.67992	-1.76159	-2.01459	-1.27497	2.420942
85323	U0007365	A市融程花	2014/5/1		劳动和社会保障		-1.73799	劳动和社	-1.38222	-1.16264	-1.24627	-1.69889	-1.63946	1.662684

3 问题 2：热点问题挖掘

3.1 问题背景

随着互联网及移动设备的普及，越来越多的服务机构，包括一些政府机构开始在互联网上提供留言通道，并将这些留言进行整理和针对性处理。服务人员将某一时段内群众集中反映的某一问题可称为热点问题，如“智能政务”[5]中“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。然而，现有的热点问题挖掘方式，仅仅是人为进行统计与发现，消耗大量的服务成本，且服务效率低下，为此，运用自然语言处理进行热点问题挖掘势在必行[6]。

3.2 问题描述

经过小组成员的仔细分析，热点问题挖掘主要存在以下三点问题：

1.如何定义热点问题，即：热点问题的计算方式。

2.如何利用留言主题对所有留言信息进行相似性计算，提取出热点问题。

3.如何提取留言信息的地点或相关人员信息。

对于问题 2 中存在的问题，小组查阅了相关的资料和文献，并提出了解决这些问题的可行性解决方法，通过对附表三中所有的数据测试，我们发现我们提出的解决方案具有较高的应用价值。以下将针对问题 2 存在的问题来描述我们具体的解决方法。

3.3 问题分析求解

3.2.1 热点问题定义

经过对附表三中的信息进行仔细分析，我们发现附表三中的“留言主题”、“留言详情”、“点赞数”、“反对数”与我们需要定义的热点问题有关，而其他信息则具有较低的相关性（时间信息会因时间的流逝而失去热度，即：假设相关人员即使处理留言，那么现有未处理留言的留言时间都较新，所处地位相同）。在相关的留言信息中，“留言详情”虽然与热点问题具有一定的相关性，但是考虑到“留言详情”结构复杂、长度不一、且描述水平也不同，且“留言主题”是“留言详情”的高度概括，为此，我们未将“留言详情”信息考虑在热点问题的计算在内。定义一给出了留言热度定义：

定义一：给出一个阈值 α，相似留言数 Similar_Num，所有相似留言点赞数之和 agree_num，所有相似留言反对数之和 disagree_num，留言热度 Hot 计算如公式（3.2.1-1）所示：

$$Hot = (1-\alpha) * Similar\_num + \alpha * (agree\_num - disagree\_num) \quad (3.2.1\text{-}1)$$

在附件三的数据环境下，我们初步将 α 的值设为 0.5，即：我们考虑相似留言数与点赞数与反对数的差，具有等价关系。根据 C 题题目要求，我们的目标是筛选出问题热度 Hot 排名前 5 的留言。

3.2.2 留言相似性计算

留言热度 Hot 的计算涉及到的参数中，相似留言数 Similar_Num 并未直接给出，为此我们需要利用留言相似性来计算相似留言数。我们利用余弦相似度的方式来计算“留言主题”的相似性：首先，对于附件三中的每一个“留言主题”进行分词和去停用词处理，去停用词可以防止无效词对相似性计算产生干扰。然后，我们将“留言主题”中剩余的词转化成词向量。最后，我们依次以某个“留言主题”为模板，利用余弦相似度计算其他“留言主题”与其相似性，给定两个“留言主题”词向量 a 和 b，其余弦相似性 θ 由点积和向量长度给出，如公式（3.2.1-2）所示：

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2}\sqrt{\sum_{i=1}^{n}(b_i)^2}} \qquad (3.2.1\text{-}2)$$

并将相似性大于 Φ（similarity>=Φ）的“留言主题”归为相似类，在附件三的数据环境下，我们设置 Φ 为 0.5。统计每个相似类中的“留言主题”数为相似留言数 Similar_Num。

值得注意的是，我们只在当两个“留言主题”中具有相同词数大于 2 时，才会计算它们之间的相似度，这是为了防止两个过短的不相干“留言主题”词向量相似性计算时出像相似性过高的问题。

3.2.3 提取留言地点或人群

生成热点问题表还需要从“留言主题”或“留言详情”中提取相关的留言地点以及人群信息。对于相关的留言地点，我们使用字符串匹配的方法提取留言中的地址信息。对于相似类中的所有“留言主题”，进行两两匹配，匹配两个“留言主题”中的相同字符串，并筛选出最长的、以地址关键字结尾的字符串作为我们的地址信息，对于人群信息我们使用正则表达式对“留言主题”或“留言详情”中的内容进行匹配。在热点问题表中，我们依次选择相关人员、留言地址，“留言主题”+“相关人员”作为热点问题表中地址/人群的一列。

3.2.4 结果分析

对于问题 2 的解决，我们首先使用了定义一来定义留言热度 Hot；其次，我们使用余弦相似度来将“留言主题”分词不同的相似；最后，我们使用“留言主题”匹配相同字符串和正则表达式的方式来提取留言地点或人群。

总的来说，我们使用的方案对于问题 2 的解决具有较好的效果，训练模型，应用于全部数据集，得到如表 3.2.4-1 所示的“热点问题表”，以及如表 3.2.4-2 所示的“热点问题留言明细表”。

表 3.2.4-1 热点问题表

热点排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	22.5	2019/8/1至2019/8/1	A9市	咨询A9市高铁站选址的问题
2	2	15	2017-06-08至2019/11/27	来自西地省经济学院体育学院的一名即将大四的学生	A市涉外经济学院强制学生实习
3	3	14.5	2019-11-15至2020-01-25	暮云街道丽发新城小区的一名业主	丽发新城小区旁边建搅拌站
4	4	12.5	2019-08-01至2019/8/7	广铁集团铁路职工	A市伊景园滨河苑协商要求购房同时必须购买车位
5	5	10	2019/1/21至2019/9/12	西湖街道茶场村五组的村民	请问A3区西湖街道茶场村五组是如何规划的

表 3.2.4-2 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	203187.0	A0002471(	咨询A9市高铁站选址的问题	2019/8/1	尊敬的A市	53.0	10.0
2	233759.0	A909118	A市涉外经济学院强制学生实习	2019/04/	各位领导	0.0	0.0
2	360112.0	A220235	A市经济学院强制学生实习	2019-04-	各位领导	0.0	0.0
2	360113.0	A3352352	A市经济学院强制学生外出实习	2018-05-	A市经济学	3.0	0.0
2	195917.0	A909119	A市涉外经济学院组织学生外出打工合理吗？	2019/11/(	一名中职	0.0	1.0
2	242062.0	A0002888(	西地省涉外经济学院变相强制学生“社会实践”	2019/11/	请制止和	0.0	0.0
2	360114.0	A0182491	A市经济学院体育学院变相强制实习	2017-06-(	书记您好	9.0	0.0
2	266368.0	A0003892(	A市涉外经济学院寒假过年期间组织学生去工厂工作	2019/11/	关于西地	0.0	0.0
2	360111.0	A1204455	A市经济学院组织学生外出打工合理吗？	2019-11-(	一名中职	1.0	0.0
2	235521.0	A0006920	A3区枫林三路涉外经济学院外街理发店扰民	2019/10/	A市A3区，	0.0	0.0
3	190108.0	A909240	丽发新城小区旁边建搅拌站	2019-12-	丽发新城	0.0	1.0
3	238212.0	A909203	丽发新城小区附近建搅拌站合理吗？	2019-12-	请问在居	0.0	0.0
3	283482.0	A909232	丽发新城小区附近搅拌站的一些问题	2019-12-(	我是A市A	0.0	0.0
3	243692.0	A909201	丽发新城小区附近的搅拌站噪音严重扰民	2019-11-	领导您好	0.0	2.0
3	281546.0	A0005147(	丽发新城小区附近搅拌站粉尘大,无法呼吸	2019-11-	我是暮云	0.0	1.0
3	247160.0	A0001041(	A市丽发小区建搅拌站，噪音污染严重	2019/11/	发同投资	0.0	0.0
3	247160.0	A0001041(	A市丽发小区建搅拌站，噪音污染严重	2019/11/	发同投资	0.0	0.0
3	244512.0	A0009470(	搅拌站丽发新城小区粉尘大的孩子生活不了	2019-12-(	我是暮云	0.0	1.0
3	213464.0	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民	2019-12-	我是暮云	0.0	0.0
3	281943.0	A909216	举报A2区丽发新城小区附近仍存在非法搅拌站	2019-11-	A2区辖区	0.0	0.0
3	217700.0	A909239	丽发新城小区旁的搅拌站严重影响生活	2019-12-	开发商把	0.0	1.0
3	272224.0	A909224	丽发新城小区噪音大粉尘大,求撤走搅拌站	2020-01-(	我是暮云	0.0	1.0
3	239336.0	A909213	A市A2区丽发新城小区遭搅拌站严重污染	2019-12-	敬爱的领	0.0	0.0
3	235362.0	A909215	暮云街道丽发新城小区附近水泥搅拌站非法经营何时休	2020-01-(	暮云街道	0.0	0.0
3	214282.0	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-	你们管不	0.0	0.0
3	216824.0	A909214	搅拌站大量加工砂石料噪音污水影响丽发新城小区环境	2019-12-	最近一段	0.0	0.0
3	268109.0	A909230	我要举报A市A2区丽发新城小区开发商违规建设搅拌站	2019-12-(	我要举报	0.0	0.0
3	203393.0	A0005306(	A市丽发新城小区侧面建设混泥土搅拌站，粉尘和噪音污染严重	2019/11/	发同投资	0.0	2.0
3	239648.0	A909211	A市A2区丽发新城小区附近搅拌站明目张胆污染环境	2020-01-(	丽发新城	0.0	0.0
4	191001.0	A909171	A市伊景园滨河苑协商要求购房同时必须购买车位	2019-08-	商品房伊	1.0	12.0
4	268626.0	A909186	A市伊景园滨河苑坑害购房者	2019-08-	A市伊景园	0.0	0.0
4	230554.0	A909174	投诉A市伊景园滨河苑捆绑车位销售	2019-08-	投诉A市伊	0.0	0.0
4	230554.0	A909174	投诉A市伊景园滨河苑捆绑车位销售	2019-08-	投诉A市伊	0.0	0.0

值得一提的是，在实践过程中我们将相似性大于 Φ（similarity>=Φ）的“留言主题”归为相似类，而在分类效果相差不大的情况下，我们会尽量选择 Φ 更小的值（$\Phi \in [-1,1]$），这么做是因为 Φ 越小，相似类中的项就越多，这项可能都反映了同一个地方中类似的问题，虽然可能存在细微差别，到大致的问题方向相同。而选择更小的 Φ 值将他们分在一起，有助于调查统计，或解决这些问题的人更加清楚明了的观察到这些问题，这可以大大提高相关人员解决留言问题的效率。

最后，必须承认的是，虽然我们的方案具有较好的效果，但面对结构和表达数据留言，仍然会存在一些问题，如分类错误，地址或人群信息提取失败等。

4 问题 3：答复意见评价

4.1 问题背景

近年来，随着微博、微信、市长信箱、阳光热线等网络问政平台逐渐成为政府了解民意、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。随着大数据、云计算、人工智能等技术的发展，建立自然语言处理技术的智慧政务系统已经是社会治理创新的新趋势。

4.2 问题描述

针对以上背景，本部分主要解决对相关部门答复意见的评价，从相关性、完整性以及时效性三个方面展开。

4.3 问题分析求解

本文对答复的评价主要有三部分，然后将三部分得到的分数按一定的权重得出最后的评分。

4.3.1 数据预处理

1.时间数据处理

时间这一类数据对于评价相关部门的答复具有重要意义，体现了部门的行动效率。本文对时间数据的处理主要有三部分。文本中时间数据是文本形式，首先需要将数据清洗，化为年、月、日三个信息，去除具体时间信息；其次，利用化出的时间信息，计算具体的回复天数，并根据天数给予评分；最后，对时间评分作归一化处理。整体流程如图 4.3.1-1 所示。

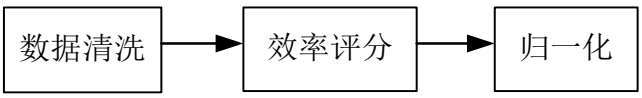


图 4.3.1-1 时间数据处理图

2.文本数据处理

本文对文本数据的处理主要包括：首先，对文本中的标点符号等符号替换；其次，对文本进行分词以及词性标注；最后，去除文本停用词。整体流程如图 4.3.1-2 所示。

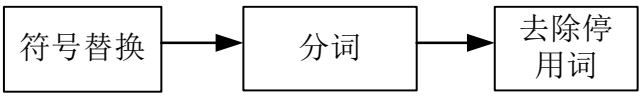


图 4.3.1-2 文本数据处理图

4.3.2 问题求解及工具

1.余弦相似度

向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值越接近 1，向量间的夹角越小，相似度越高；反之，表明越不相似。使用余弦相似度计算两段文本的相似性需要将分词结果编码，形成词频向量，最后利用余弦函数公式计算相似度。如公式（4.3.2-1）所示。

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2}\sqrt{\sum_{i=1}^{n}(b_i)^2}} \qquad (4.3.2\text{-}1)$$

其中，a、b 是两个文本向量，a_i，b_i是向量的分量。

2.主题模型

LDA(latent Dirichlet allocation)主题模型是一种三层贝叶斯主题模型，通过无监督学习发现文档隐含的主题信息。在该模型中，每个文档表示为多个主题的概率分布，每个主题表示为多个单词的概率分布。本文主要把每个文档提取出一个

主题以及多个单词的概率分布，得到每个文档的一个主题分布，最后得到一组单词的概率分布。利用这些单词的概率分布结果来计算本文所需要的主题覆盖度。LDA 主题模型如 4.3.2-1 所示。

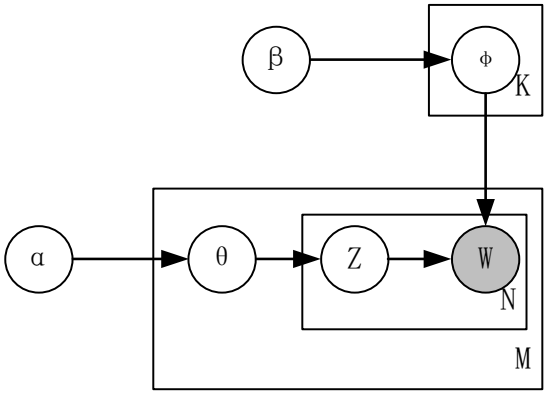


图 4.3.2-1 主题模型图

其中，β和α为语料级别先验参数，α代表每个文档下主题多项分布的 Dirichlet 先验参数，β代表每个主题下单词多项分布的 Dirichlet 先验参数；θ和 φ 是隐含变量，θ代表每个文档与主题之间的多项分布，φ 代表每个主题与语料单词之间的多项分布；K 代表主题数，N 代表每个文档的单词数，M 代表中文档数；$Z_{m,n}$是第 m 个文档中第 n 个词的主题，$W_{m,n}$是第 m 个文档中第 n 个词。

3.相关性

在得出文本相关性这部分，本文主要利用 TF-IDF 实现。它是一种统计方法，认为一个词在一个文档中的重要性随着词出现的次数增加而增加。因此，通过统计词频可以得出文本的词向量，通过这些词频可以很好地表征文本，有利于计算文本相似性。

留言与答复的相关性通过数据中的留言详情和答复意见文本的相似度得出。相似度求解过程如下：

(1)文本数据预处理：符号替换、分词、去停用词；

(2)处理完的文本利用文本特征提取，获取词频，表征为词向量；

(3)利用余弦相似度计算词向量的相似度，得出文本相似度。

将全部答复意见数据集通过以上步骤训练模型，得到全部的答复相关性分布图，如图 4.3.2-2 所示。

由图 4.3.2-2 可知，大部分的答复意见与留言的相关性处于 0.5-0.7 分，说明大部分答复都处于中立的角度，浅显的回答了群众的留言意见。

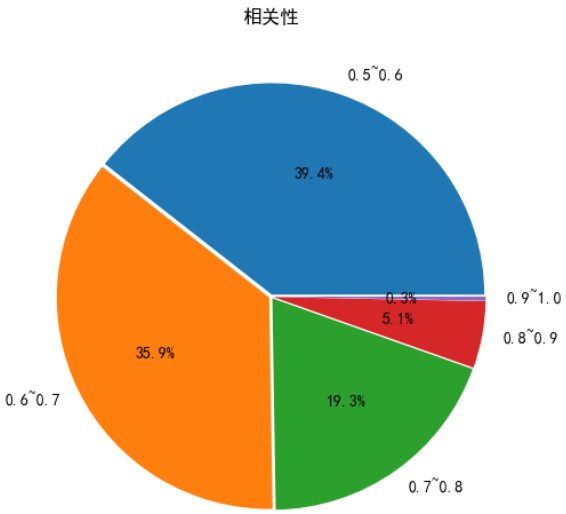


图 4.3.2-2 全部数据的回复相关性占比图

4.完整性

相关部门对于网友留言回复的完整性评估主要从主题入手。每个文档都隐含主题，主题则反映了文档的核心思想；主题则是由一组主题词的概率分布组成，主题词能充分反映主题的内涵。因此，本文计算主题覆盖度主要利用主题词来实现。

相关部门的回复完整性评价利用留言与答复的主题覆盖度来计算。主题覆盖度求解过程如下：

(1)文本数据预处理：符号替换、分词、去停用词。

(2)已处理的文本输入，创建语料的词典，通过词典将语料转换为矩阵。

(3)利用矩阵以及词典训练 LDA 模型，每个文本求出一个主题以及六个主题词。

(4)将主题词清理，利用余弦相似度以及主题词的词向量表征计算主题覆盖度。

将全部答复意见数据集通过以上步骤训练模型，得到全部的答复完整性分布图，如图 4.3.2-3 所示。

由图 4.3.2-3 可知，大部分的答复意见的完整性处于 0.5-0.6 分，依旧是一个偏中间的分数。说明大部分答复都能较为完整的回复群众的留言，答复与留言主题差不多匹配。

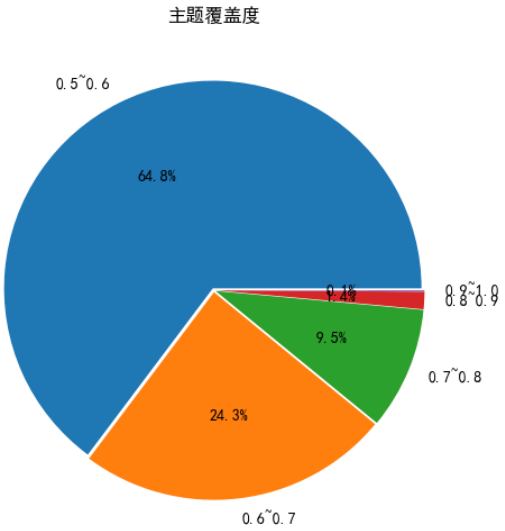


图 4.3.2-3 全部数据的回复主题覆盖度占比图

5.时效性

回复时间的长短可以体现相关部门的工作效率，将时间作为评估相关部门回复的指标之一可以更好地从多维度评价回复工作。时间越长，相关部门在回复留言工作方面的效率越低；反之，越高。根据网络问政相关部门回复时间规定，对相关部门回复效率定级。时效性求解过程如下：

(1)时间数据格式清理，形成固定格式：年、月、日。

(2)根据以上清理数据，求解出答复时间，并给出工作效率定级。

(3)对工作效率级别做归一化处理，得出效率分数。

对于以上处理的时间数据清理完毕后，计算出回复天数，根据天数进行工作效率定级。规则如表 4.3.2-1 所示。等级越高则回复效率越高。

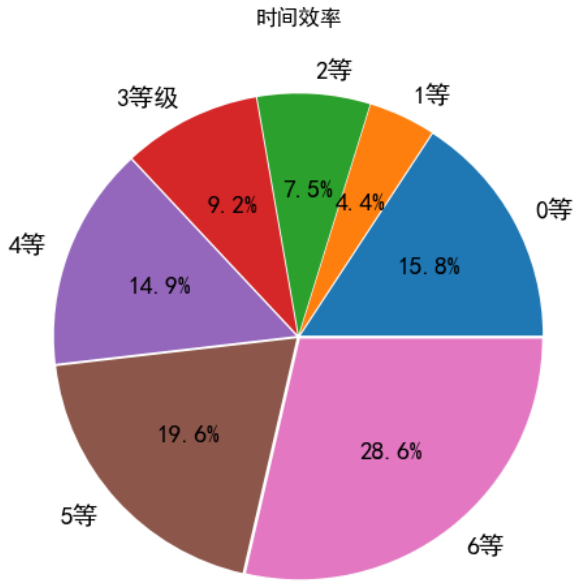
表 4.3.2-1 效率定级表

效率等级(级)	回复时长 T(天)
6	T<=5
5	5<T<=10
4	10<T<=15
3	15<T<=20
2	20<T<=25
1	25<T<=30
0	T>30

将全部答复意见数据集通过以上步骤训练模型，得到全部的答复时效性分布图，如图 4.3.2-4 所示。

由图 4.3.2-4 可知，63.1%的答复意见处于 4 等及以上时效，即大部分的留言在 15 天之内得到了答复。由此可见大部分行政部门的工作效率是可观的，但高

效率的部门数目仅为及格。反观剩余部分留言，超过 30 天的答复占据 15.8%，可见懈怠部门仍是大多数。这种要么及时回复，要么搁置的工作作风在智慧政务上亟待解决。



4. 3. 2-4 全部数据的回复时效性占比图

4. 3. 3 数据融合

根据以上步骤得到的相关性、主题覆盖度以及时效性来得出最后对相关部门的评分。将以上数据按一定权重，加权求和给定评分。本文权重设置为 0.5、0.3、0.2。如公式（4.3.2-2）所示。

$$score = 0.5 * similarity + 0.3 * theme + 0.2 * time \quad (4.3.2\text{-}2)$$

其中，similarity 代表相似性，theme 代表主题覆盖度，time 代表时间效率。根据公式得出的 score 来评分。规则如表 4.3.2-2 所示。

表 4.3.2-2 效率定级表

评分(分)	Score 值(分)
1	score<0.6
2	0.6<=score<0.7
3	0.7<=score<0.8
4	0.8<=score<0.9
5	0.9<=score<=1.0

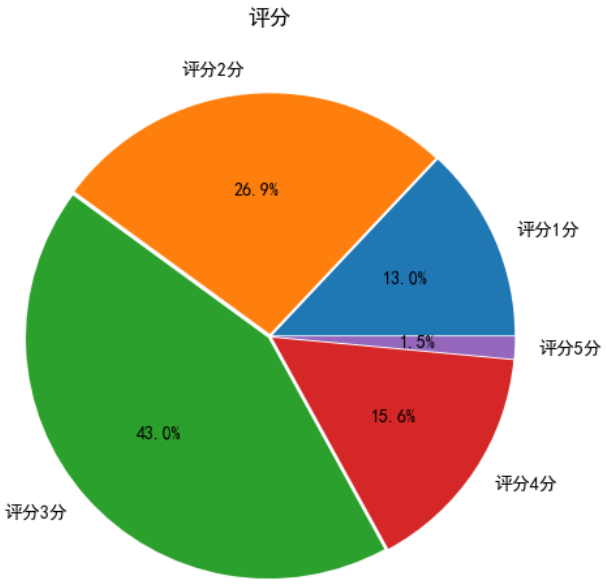
根据表 4.3.2-2 规则，假设相关性得分为 0.7 分，完整性得分 0.6 分，时效性得分 0.6 分，则该条回复最终评分为：

Score=0.5*0.7+0.3*0.6+0.2*0.6=0.65(分)

即在 5 分制下，该条留言最终评分为 2 分。

将上述规则应用于全部数据集，我们可以得到如图 4.3.2-5 所示的回复得分

的分布图。



4.3.2-5 全部数据的回复评分占比图

由图 4.3.2-5 可见，大部分的答复意见处于 2-3 分之间，这对于满分值为 5 的评分标准来说是较为中立的分数，说明大部分答复意见并不能很好的解答群众的留言问题，这类答复意见涉及的相关部门还需要改进和完善。

5 不足与展望

本文基于机器学习的“智慧政务”文本挖掘涉及多个学科的理论方法和科学技术，在基本完成赛题要求的基础上，结合更深层次的理论研究和在实践应用中的积累完善，我们还可以在以下几个方面进一步的进行优化和研究：

1.在群众留言分类问题上，由于采用的机器学习 SVM，因此需要大量的数据进行学习，但是赛题中全部数据也仅有 9010 条留言数据可用于分析。在仅有示例数据时，模型的 F1 值仅能达到 0.4，在得到全部数据时模型后模型 F1 值达到了 0.9。

2.在热点问题挖掘上，面对结构及表达相对复杂的留言数据，仍然会存在一些问题，如分类错误，地址或人群信息提取失败等。在未来，我们在改进方案的同时，将会进一步的研究多条件下协同分类及词性标注等策略来对留言进行分类，从而解决热点问题挖掘等相关问题。

3.在答复意见评价上，本模型从相关性、完整性以及时效性入手。相关性计算时主要利用词频以及词向量来计算相似度，这样表达的文本特征信息量有些不足，未能考虑语义信息，后期或许可以结合深度学习对文本特征表达这一方面优化。另外，本文最后未能从答复的可解释性给出评价，若能加上可解释性的评价，那么对答复的评价质量会更好。

参考文献

[1] 邓乃扬，田英杰，2004:《数据挖掘中的新方法——支持向量机》，科学出版社。

[2] SVM核函数RBF的参数https://blog.csdn.net/wn314/article/details/79972988

[3] 《支持向量机原理详解(八):多类分类》知乎分享，参考网址为，SVMhttps://zhuanlan.zhihu.com/p/66933242

[4] 廖国琼，姜珊，周志恒，万常选，2017:《基于位置社会网络的双重细粒度兴趣点推荐》，计算机研究与发展，第54卷11期：P2600-2610。

[5] 权军,易萍,2019:《浅议人工智能在人社系统电子政务公共服务领域应用的意义》，劳动保障世界，第34期：P56-57。

[6] 翟云，2019:《人工智能+政务：开启智慧治理新征程》，学习时报，第3期。