基于自然语言处理技术的智慧政务系统

摘要

随着网络问政平台的兴起，网络问政逐步成为了政府了解民意的重要渠道，各类社情民意相关的文本数据不断攀升，对相关的政府部门带来了极大的挑战。同时，随着大数据技术的发展，基于自然语言处理技术的智慧政务系统已经是社会创新发展的新趋势。在此，本文基于自然语言处理和文本挖掘的方法在 “智慧政务”的实际需要出发，研究了以下几个方面的内容：

将群众留言进行划分，以便于分派到相应职能部门处理。在经过数据的预处理之后，根据文本分类原理，设置测试文本和训练文本，使用算法训练分类器与线性的支持向量机。采用基于二分类的一对多的决策树策略导出关于留言内容的一级标签分类模型。再通过训练学习得到的模型对群众留言文本进行分类。关于分类效果的评价检测，我们采用F-Score分类评价模型对分类效果进行评价打分。

热点问题整理。群众在某一时段集中反应的某一问题的集中投诉，称为热点问题，这类问题产生速度快、数量大，如不及时处理势必会产生严重的负面影响。关于在海量文本数据挖掘出热点问题的方法，若直接采用文本聚类，会导致文本向量空间的维数过高导致聚类很差，使得反应同一主题的信息汇聚在一起。因此本文将热点问题转化为先求热点词，后求热点问题。在提取文本特征词库后，先对通过全局加权平均法进行平滑处理，然后通过提取热点词进行变量聚类。最后，将热点词类群用TF-IDF的文本相似度检测出留言内容中的热点问题。

对答复意见的质量评价。如何评价答复意见的质量评价，我们从及时性、可解释性、相关性、完整性的角度对答复意见值建立一套评价模型。我们先从不同角度研究其对应的不同的指标，通过指标对总体质量的正负影响程度，采用熵权法，将指标信息熵化，然后确定各指标的熵权重，再根据权重评分答复意见的质量评价。

关键字：基于二分类的一对多策略 ； TF-IDF ； 信息熵化 ； 全局加权平均法 ； 熵权法 ；

Abstract

With the rise of the platform of network politics, network politics has gradually become an important channel for the government to understand public opinion. The text data related to all kinds of social situations and public opinion is rising, which brings great challenges to the relevant government departments. At the same time, with the development of big data technology, an intelligent government system based on natural language processing technology has become a new trend of social innovation and development. Based on the actual needs of natural language processing and text mining in "smart government", this paper studies the following aspects:

Divide the message of the masses so that it can be handled by the corresponding functional departments. After data preprocessing, according to the text classification principle, test text and training text are set up, and an algorithm is used to train classifiers and linear support vector machines. A one to many decision tree strategy based on two classification is used to derive the first level label classification model of message content. Then through the model of training and learning, we classify the message text of the masses. For the evaluation and detection of classification effect, we use the F-score classification evaluation model to evaluate and grade the classification effect.

Sorting out hot issues. The concentrated complaints of a certain problem that the masses react in a certain period are called hot issues. These problems have a fast speed and a large number. If they are not handled in time, they will have serious negative effects. As for the method of mining hot issues in massive text data, if text clustering is used directly, the dimension of text vector space will be too high and the clustering will be very poor, so that the information reflecting the same topic will be gathered together. Therefore, this paper transforms hot issues into hot words first, then hot issues. After extracting the feature words of the text, the global weighted average method is used for smoothing, and then the hot words are extracted for variable clustering. At last, we use TF-IDF text similarity to detect the hot issues in

the message content.

The quality evaluation of the reply. How to evaluate the quality of the reply? We set up a set of evaluation models for the reply value from the perspective of timeliness, interpretability, relevance, and integrity. First, we study the corresponding different indicators from different perspectives. Through the positive and negative impact of indicators on the overall quality, we use the entropy weight method to entropy the information of indicators, then determine the entropy weight of each indicator, and then grade the quality evaluation of the reply opinions according to the weight.

Key words:one too many strategy based on two classification; TF-IDF; information entropy; global weighted average method; entropy weight method;

目录

1.绪论

1.1 研究背景及意义

随着计算机技术的发展以及信息网络时代的到来，每天网络上都会产生并且积累着大量的数据，海量的数据影响了人们对于数据的正确且有效的利用。例如政府需要了解民意，需要从民众中听取意见，可是大量的留言形成了大量的数据海，给以往主要依靠人工力量来进行留言的分类和热点整理及回复意见的相关部门的工作带来了极大挑战，使得相关部门的处理效率跟不上群众意见的上传速度，没有及时将问题分类，且整理热点问题回复群众，会使得政府失去民心，引起民众不满。因此，基于自然语言处理技术的政务系统是政府极为需要的，而这就需要我们使用数据挖掘之中的文本挖掘技术来处理数据。

文本挖掘技术中最为基础且重要的技术便是分类和聚类。对于类似于群众的意见处理等问题，我们都需要进行文本分类，在文本分类的基础上我们便可以更好的针对信息进行针对化处理。因此，建立一个有效的文本分类模型是我们处理信息的重要前提。

文本聚类的实现使得更容易发现新的问题。在政府的网络平台上有海量的数据，观察其中的数据，不难发现其中有这样的一类问题，在某段时间内有相当一部分数量的群众集中反映的问题，我们把这些问题统一归为热点问题。如果政府不能及时了解哪些问题是热点问题，不能及时挖掘出热点问题并作出回应，那么会使该热点问题迅速发酵，势必会产生严重的负面影响，对于这类热点问题必须迅速的挖掘出来，集中彻底解决，避免事态进一步的扩大。而热点问题的有效提取，能让政府快速了解民众遇到的困难，从根本上提高政府人员的工作效率。面对如何快速寻找热点问题，这就需要运用到文本聚类技术，因此建立一个有效的文本聚类模型是我们解决挖掘热点问题的重要方法。

1.2 问题重述

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。

附件给出了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。请利用自然语言处理和文本挖掘的方法解决下面的问题。

（1）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（参考附件 1 提供的内容分类三级标签体系）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。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。

（2）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。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，做一个表给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。做一个表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”。

（3）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。

2.问题分析

2.1 问题 1 的分析

本题需要根据附件 1 提供的内容分类三级标签体系对附件 2 的数据建立一个

关于留言内容的一级标签分类模型。而建立标签模型，实质上也就是进行文本分类任务，文本分类任务可以拆分为特征工程和分类器，那么我们大致分为 4 步建立模型并评价模型结果。

（1）获取训练文本集。文本集由预处理过后的文本数据组成，对于每个训练文本集建立一个分类标签。

（2）选择分类方法并训练分类模型。使用线性的支持向量机方法进行分类，再进行训练分类模型。

（3）用训练后的模型对附件 2 中的文本数据进行分类。

（4）最后通过使用 F - score 对这次建立的分类模型进行评价。

2.2 问题 2 的分析

针对在某短暂的时间内对某一问题的集中投诉，即热点问题，这类问题一般具有产生速度快、数量大的特点，如不及时处理必会产生严重的负面影响。我们需要从海量的文本数据中挖掘出什么是热点问题，本文不直接采用文本聚类的方法，因为传统的聚类可能会导致向量维度过高。因此本文采用 2 个步骤来解决热点挖掘问题。

（1）寻找热点词。通过全局加权平均法、肘部法确定最佳的 k 值即确定聚类的 k 值，以此来寻找热点词。进行 K - means 聚类，用以发现同一主题的热点词。

（2）用文本相似度检测文本数据中的热点问题。

2.3 问题 3 的分析

针对相关部门对留言的答复意见，我们从答复的相关性、完整性、可解释性以及及时性的角度做评价模型。

（1）从答复的相关性以及完整性考虑。用侧重点于比较文本之间内容相关的余弦相似度，来检验留言与回复的词向量相似程度。

（2）从答复的可解释性考虑。从句子的词性出发，考虑名词、形容词、动词，对可解释性的影响。

（3）从答复的及时性考虑。从留言与回复的时间间隔考虑政务处理的时间效率。

3. 总体流程图

基于自热语言处理留言文本的评价模型总流程图如图 3-1 所示：

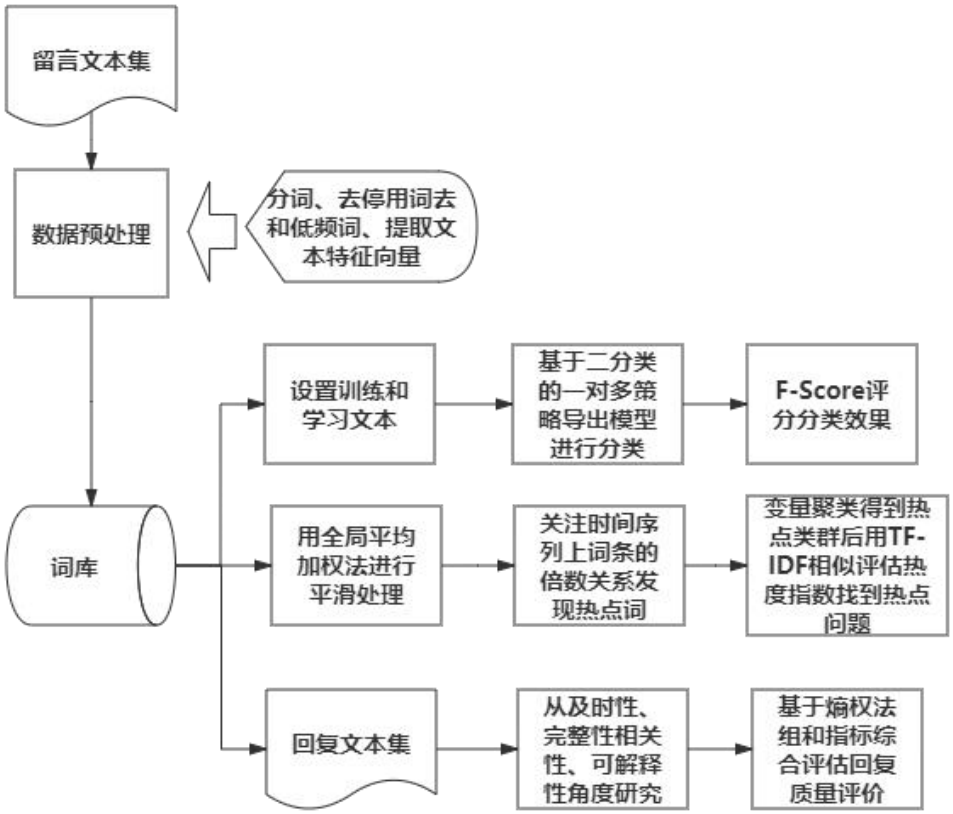

图 3-1 总体流程图

4. 实验环境

本文的实验在 Windows 10＋Python3 的环境下完成，使用的编程软件为 Anaconda 下的 Jupyter Notebook 。其中，使用了下列几种库：

sklearn 机器学习库	Pandas 库	Numpy 库	Jieba 库	wordcloud 库
matplotlib 库	seaborn 库	pkuseg 库	collections 库	PIL 库

5. 符号说明

符号	意义
w_i	词条
$F_i(x)$	决策函数
TF	正文本词频率
IDF	逆文本频率
TP	正样本被正确预测为正样本
FP	负样本被错误预测为正样本
TN	负样本被正确预测为负样本
FN	正样本被错误预测为负样本
P_i	第i类的查准率
R_i	第i类的查全率
$\mathrm{x}^{\mathrm{avg}}$	加权平均数
$f_1, f_2, ..., f_k$	词条的权
s(i)	轮廓系数
$\cos(\theta)$	余弦相似度
E_j	信息熵
x_i	指标
W_i	信息熵各指标的权重

6. 数据预处理

对于原来提供的文本数据中会存在着大量冗余信息和有异常数据，所以我们需要进行文本预处理。

文本预处理是指将多样化的原始数据文本去除原来可能有的标记或者其他无关的信息，使文本仅包含独立的词，且去除对文本分类没有意义的词的过程。关于文本预处理，需要有选择的进行去停用词和分词。最后将文本处理成计算机可以识别的数据——词向量即文本的向量化表示，便于后续展开研究。预处理的主要过程如图 6-1 所示：

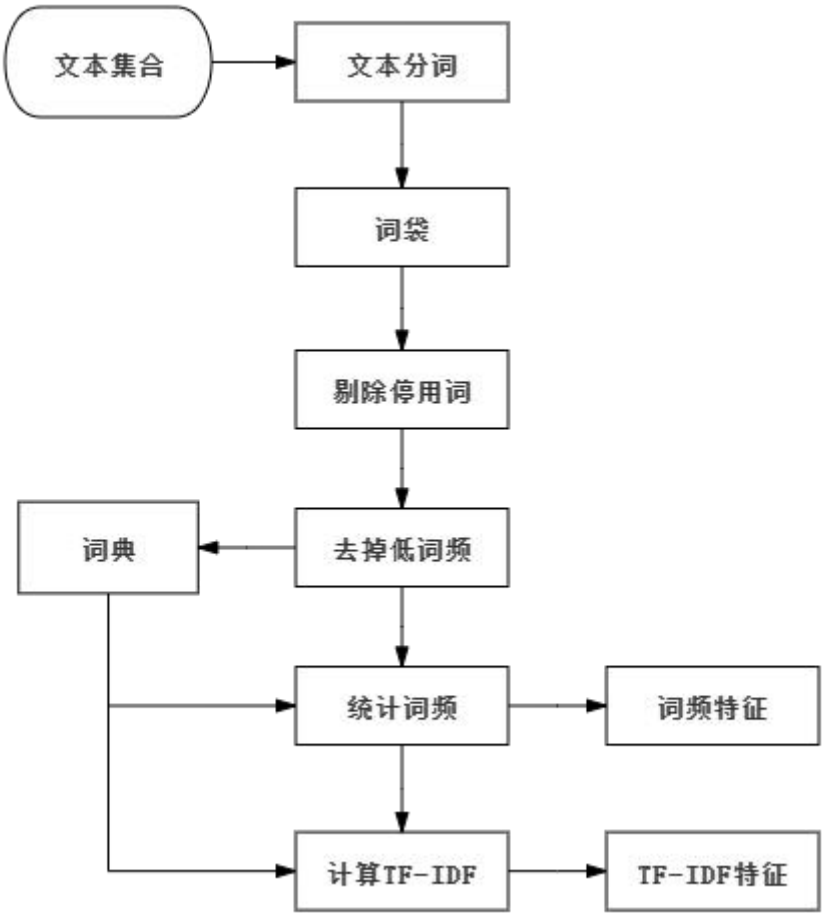

图 6-1 文本预处理流程图

6.1 中文文本分词

对文本集合进行词切分操作，选择 pkuseg 进行分词。 pkuseg 是由北京大学

语言计算与机器学习研究组研制推出的一套全新的中文分词工具包。LTP 是哈工大出品的自然语言处理箱，用户可以通过使用工具进行分词、词性标注、句法分析等等。选择 pkuseg 的理由：

1.高分词准确率。分割精度高。与其他的分词工具相比，在不同的数据区域，分词的准确率有了很大的提高。经过北京大学研究组测试结果表明，相同条件下，pkuseg 在 MsrA 和 ctb8 两个样本数据集上的分词错误率比其他分词工具分别降低了 79.33%和 63.67%。

2.多领域分词。相比于其他的分词工具包，pkuseg 工具包在不同领域的数据上都大幅提高了分词的准确度。

3.支持用户自我培训模式。支持用户使用新的注释数据进行培训。根据分类语料库手动添加更多的语料比如哈工大 LTP 里的分词表，使得语料清洗的更加干净。

最后总体分词的效果如下图 6-2 所示：

	topic_one	label
0	[大道, 西行, 便道, 未, 路口, 加油站, 路段, 人行道, 包括, 路, 灯杆, 圈...	城乡建设
1	[位于, 书院, 路主, 干道, 水, 一方, 大厦, 楼, 楼, 人为, 拆除, 水, 电...	城乡建设
2	[小区, 地面, 停车位, 程明, 物业, 征, 业主, 意见, 强制, 违规, 收费, 管...	城乡建设
3	[华庭, 小区, 高层, 次, 供水, 楼顶, 水箱, 长年, 洗, 自来水, 龙头, 水,...	城乡建设
4	[华庭, 小区, 高层, 次, 供水, 楼顶, 水箱, 长年, 洗, 自来水, 龙头, 水,...	城乡建设
...	...	...
9205	[夫妻, 农村, 户口, 女, 岁, 儿2, 岁, 斤, 治疗, 一级, 脑瘫', 纯女户,...	卫生计生
9206	[中心, 医院, 做, 无痛, 人流, 手术, 手术, 怀孕, 症状, 复查, 活胚芽, 宫...	卫生计生
9207	[再婚, 想, 小孩, 不知, 我省, 胎, 新, 政策, 先, 怀孕, 做, 西地省, 胎...	卫生计生
9208	[明白, 国家, 政策, 简化, 手续, 群众, 开, 奇葩, 证明, 生育, 证明, 实际...	卫生计生
9209	[领导, 你好, 未婚, 生子, 接受, 处罚, 小孩, 户, 小孩, 外地, 上学, 需,...	卫生计生

图 6-2 总分词效果图

6.2 剔除停用词

下载停用词表，对文本进行预处理，去除对文本特征没用的词，比如“啊”、“我”、“着”、“和”，还有空格、标点、介词等等。剔除这些停用词有助于提高文本分析的效率。停用词表如下图 6-3 所示：

	stopwords
0	去年
1	NaN
2	安全
3	食品
4	一个
...	...
2757	['
2758	']
2759	请问
2760	这几年
2761	两

图 6-3 停用词表示例图

6.3 去除低频词

中文文本中有一些低频词词，虽然它们有一定的意义，但只有少数出现过，相对于整个文本集来说，这些词不仅没有什么意义，反而可能在后续文本分类算法中出现偶然误差或者极端现象，因此需要将其进行剔除。在此选择使用 Counter 等库获取所有句子中所有词的词频，通过筛选频率为 2 以下低频词进行删除。得出词频图如下图 6-4 所示：

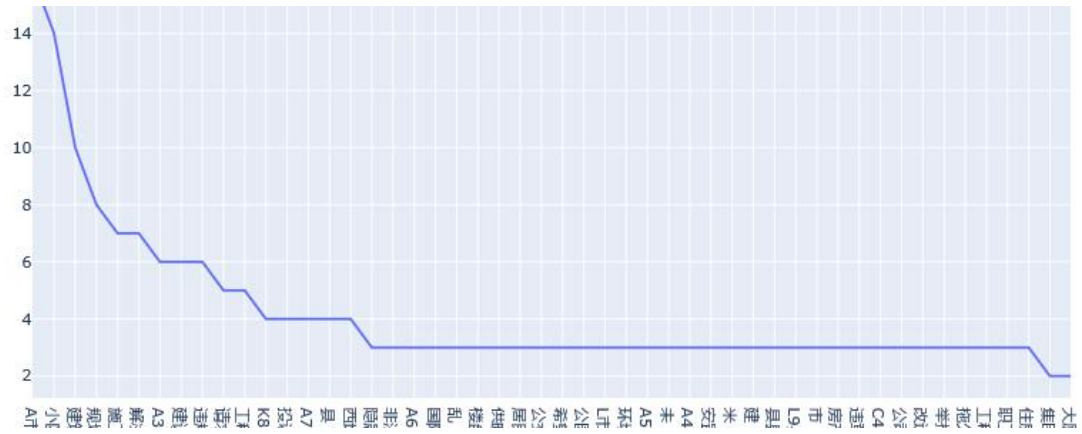

图 6-4 词频图

6.4 文本的特征提取和特征选择

得到文本分词结果后，选择 $TF-IDF$ 模型文本特征提取方法，并对特征进

行选择，构建词向量空间，约简特征，精良降低维度，减少后续计算量。

TF – IDF 模型是一类应用广泛的加权技术，TF(Term Frequency)是指正文本词频率，可以理解为文本内词汇出现的频率，即指给定单词在文件中出现的次数。此数值是对词数归一化后的处理结果，即对向量实施缩放处理，全部元素的合计值等于 1，由此避免它偏向长文档。在某一文档内的词条 w 而言，体现其重要性的表达式如下：*TF* 计算公式：

$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

IDF (Inverse Document Frequency)则是指逆文本频率，表示一个词语普遍关键性的度量。*IDF* 的主要思想是：如果包含词条 w 的文档较少，即该类中所有的词条数目 *n* 越小，IDF 越大（见下式），则词条 w 具有很好的分类能力。如果一种类型的文档 C 中包含 w 项的文档数为 *m*，而其他类型中包含 w 项的文档总数为 *k*，显然，包含 w 的所有文档数为 n = m + k。当 *m* 较大时，*n* 也较大，根据 *IDF* 公式，*IDF* 值很小，说明区分 w 的能力不强。相反的文档频率意味着包含 w 的文档越少，就 *IDF* 越大，这意味着词条具有很好的区分类别的能力。

IDF 计算公式：

$$IDF = \log(\frac{\text{语料库的文档总数}}{\text{包含词条}w\text{的文档数}+1})$$

分母之所以要加 1，是为了避免分母为 0。最后，*TF – IDF* 的值是这两个值的乘积：

$$TF - IDF = TF * IDF$$

预处理后提取特征向量示例图如 6-5 所示：

```
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.]])
```

图 6-5：特征向量示意图

关键词，如图 6-6 所示：

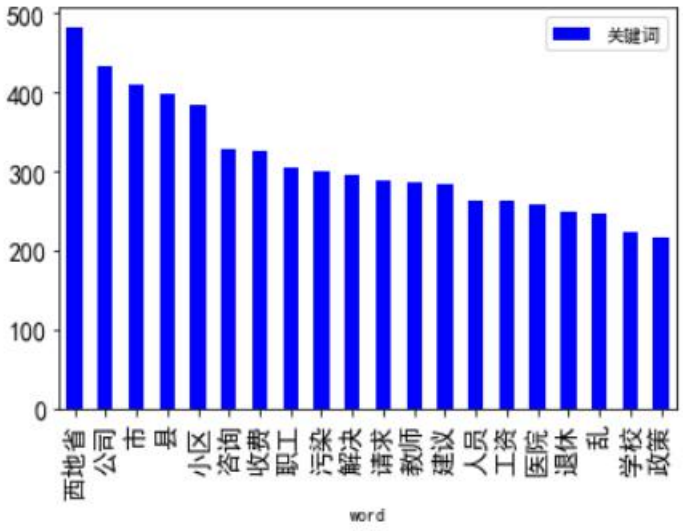

图 6-6 关键词图

由此可见，$TF-IDF$ 可以有效评估特定字词对于文本集或一个语料库的重要程度，通过某特定文本集中的高频率，结合此语言于文本集合内的低文本频率，生成高权重的$TF-IDF$。故而，将文本的重要词语保留下来。

7. 问题解答及过程

7.1 问题 1 的解答过程

对于问题 1，完整的解题流程图如图 7-1 所示：

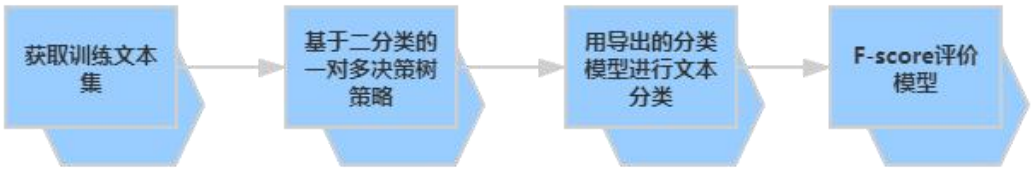

图 7-1 问题 1 解题流程图

7.1.1 获取训练文本集

首先根据文本分类原理，设置测试与训练文本比例为 3:7，文本的 30%被用于分类规则的准确性，文本的 70%用于训练学习。训练文本集由一组经过预处理的文本特征向量组成，每个训练文本有一个类别标签。

7.1.2 选择分类方法并训练分类模型

关于文本分类方法在此采用了基于二分类的一对多 (One - Against - All) 的决策树策略，分类器，使用算法训练分类器与线性的支持向量机。因为样本数据不均衡，采用多个二分类模型组合成一个多分类模型，每个类别训练一个模型。一个样本为正样本，其他的则为负样本。

即用一个分类器对应一个类别，每个分类器都把其他全部的类别作为相反类看待，然后训练得到对应的决策函数 $F_i(x)$ 。如此，便可利用训练集构造 N_c 个决策函数，即 N_c 个二值分类器。

在具体分类时，先将分类样本 x 带入决策函数 $F_i(x)$ ，若 $F_i(x)>0$ ，并且 $F_j(x)<0(j\neq i, j=1,...,N_c)$ ，则将其归为第 i 类。但是，如果遇到 x 对应多个决策函数都大于 0，则将 x 归类为使得原最优平面函数最大值所对应的类别，即

$$label(x)=\arg_{i=1,...,N_C}\max F_i(x)$$

得出交叉验证的学习曲线图，如图 7-2 所示：

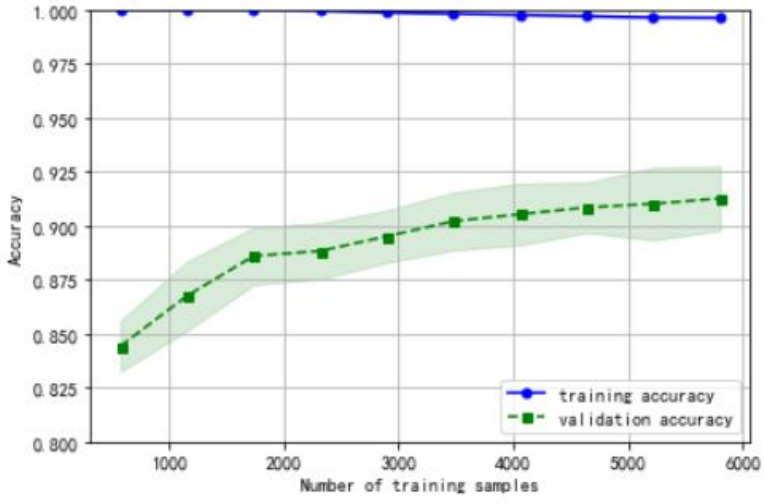

图 7-2 交叉验证学习曲线图

7.1.3 用导出的分类模型进行文本分类

根据图 7-3 一级分类标签图，用导出的分类模型进行文本分类。

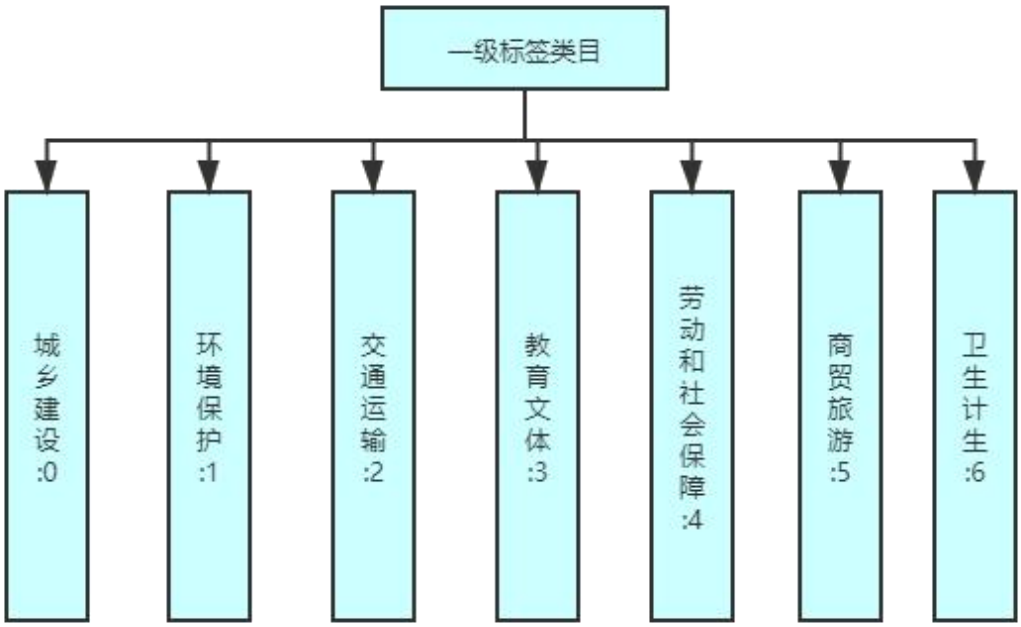

图 7-3 一级标签分类图

最终导出如图 7-4 所示的每个标签类别分类的结果：

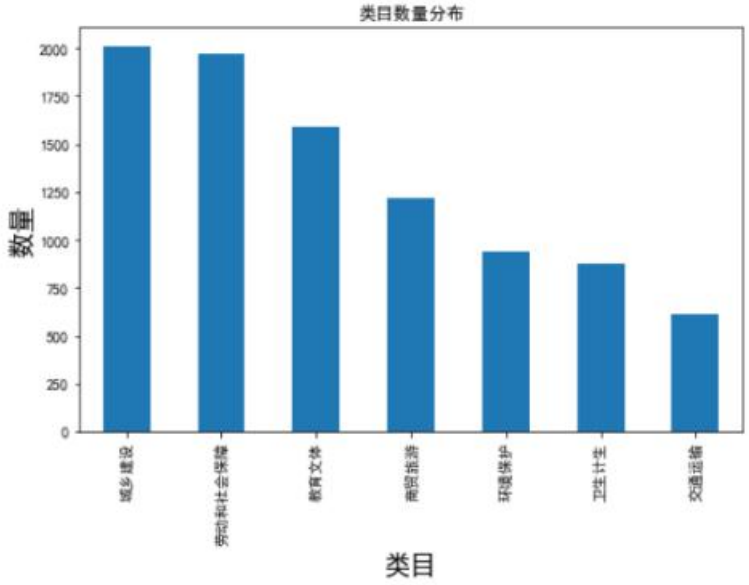

图 7-4 标签分类结果图

7.1.4 根据分类结果评估分类模型

分类评测模型通常为 F-score 模型，使用四个符号表示预测的所有情况:

TP (真阳性):正样本被正确预测为正样本	FP (假阳性):负样本被错误预测为正样本
TN (真阴性):负样本被正确预测为负样本	FN (假阴性):正样本被错误预测为负样本

评价方法介绍：

F-Score 的总计算公式:

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率

（1） P 查准率：

注意实际阳性样本在预测为阳性的数据中所占的比例（可能包括阴性样本）

计算公式：

$$P = \frac{TP}{TP + FP}$$

（2） R 查全率：

正确预测的数据集中真实正样本（不包括任何负样本）的比例

计算公式：

$$R = \frac{TP}{TP + FN}$$

（3） F-Score 评估模型评测结果如表 7-1 所示：

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

标签分类	Precision	Recall	F-score
0.城市建设	0.89	0.94	0.91
1.环境保护	0.94	0.94	0.94
2.交通运输	0.91	0.85	0.88
3.教育文体	0.94	0.95	0.95
4.劳动力和社会保障	0.95	0.94	0.95
5.商贸旅游	0.91	0.88	0.90
6.卫生计生	0.92	0.91	0.91

表 7-1 模型评估结果表

总体模型的评价 F-Score 为 0.9254433586681143。

对于问题 1 的标签分类正确用混淆矩阵表示如图 7-5：

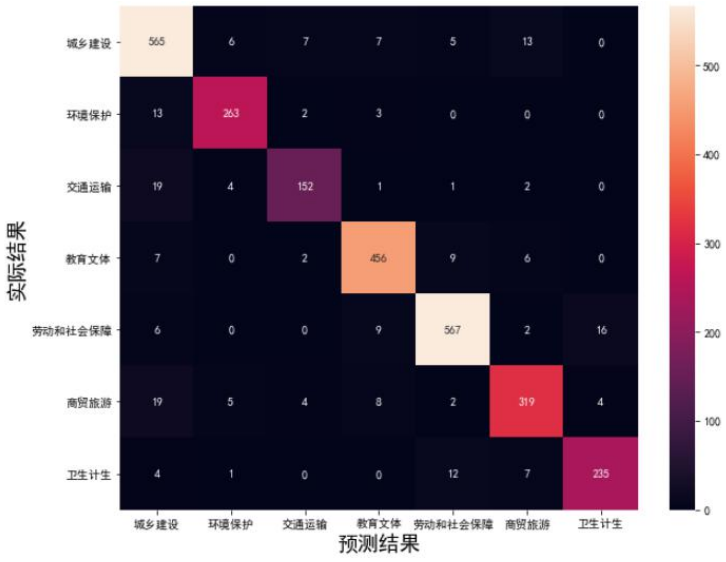

图 7-5 问题 1 标签分类结果图

7.2 问题 2 的解答过程

对于问题 2，完整的解题流程图如图 7-6 所示：

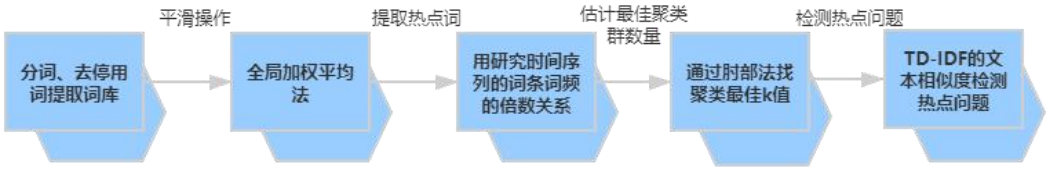

图 7-6 问题 2 解题流程图

对于文本数据，我们经过一系列的分词、去停用词以及提取词库，我们发现如图 7-7 所示问题都集中在 2019 年，所以我们选择 2019 年的所有数据进行处理。

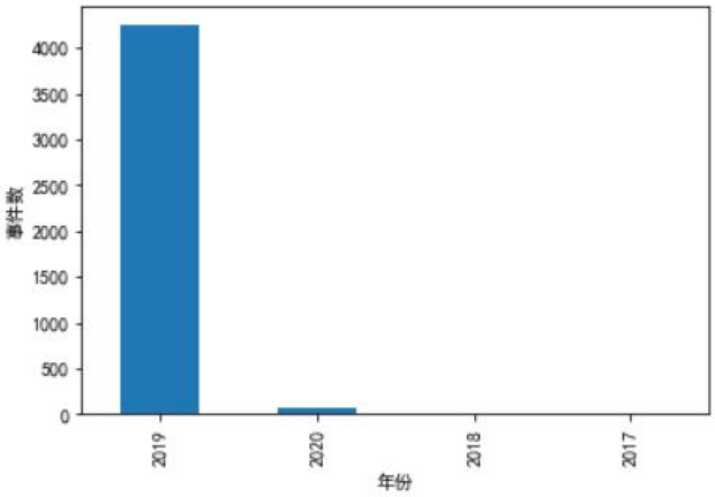

图 7-7 问题时间分布图

7.2.1 寻找热点词

如何从海量文本数据中挖掘出热点问题？若采用直接采用文本聚类方法，可能会导致向量的维度过高。因此本文通过先寻找热点词，那么热点词如何寻找呢？

（1）首先热点词与单位时间内词条词频之间的倍数关系有关，判断那些词条是增加速度快的，然而仅仅如此还不够，因为它不能防止样本太少带来的偶然误差，但我们也不可能忽视所有词频低的词条（而且很难定义），故为了检测热点词更准确，在此基础上我们需要先进行“平滑操作”——全局加权平均法，使得样本越大的词条越有能力增加评分，评分高的为热点词。

- 全局加权平均法。我们把每个词的得分都和全局平均分取一个加权平均！首先计算出这（已经经过数据预处理的词提炼的文本特征向量词库）每个词的

平均总频数。再计算出这四个词的平均得分 x^{avg}，假设已经有 n 个用户预先给每个词都打了平均分 x^{avg}，然后每个词都得到了平均分，但实际上，词数越多的词得到的评分越大。即样本越大的词，就越有能力把最终得分拉向属于自己本来应该的得分，样本太小的词，最终得分将会与全局平均分非常接近。实际运用中，这个 n 也可以由你自己来定，定得越高就表明你越在意样本过少带来的负面影响。这种与全局平均取加权平均的思想叫做 Bayesian average，从上面的若干式子里很容易看出，它实际上是最常见的平滑处理方法之一——分子分母都加上一个常数——的一种特殊形式。故此样本越大的词，就越有能力把最终得分拉向自己本来的得分，样本太小的词，最终得分将会与全局平均分非常接近。

- 加权平均法，即将各数值乘以对应的单位数，然后加总求和得到总体值，再除以全部的单位数。平均数的大小不仅取决于总体中每个单位的标志值的大小，而且取决于每个标志值出现的次数，因为每个标志值出现的次数对于它们在平均数中起着权衡轻重的作用的影响，所以叫做权数。加权平均数 $Xavg=\frac{x_1f_1+x_2f_2+\cdots x_nf_n}{n}$，其中 $f_1+f_2+\cdots+f_k=n$，$f_1,f_2,...,f_k$ 叫做权。通过数和权的乘积来计算，值得注意的是：算术平均实际上是一种特殊的加权平均，即权重相同的加权平均。比如 $f_1=f_2=f_3=\cdots=f_n$，那么加权平均数 $=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{n}$，实际上准确的说是一种算术平均。

（2）此时找出来热点词再和时间序列中观察单位时间内的倍数关系便可发现在某段时间或某段地点的热点词了。同时，对热点词进行聚类成热点类群，我们采用“轮廓系数”确定聚类的 k 值，估计热点词的最佳聚类数量，即热点词类群的数量，进行 K - means 聚类分簇热点词群。

- 轮廓系数（Silhouette Coefficient）结合了聚类的内聚性（Cohesion）和分离性（Separation），用于评估聚类的效果。该值处于 $-1\sim1$ 之间，值越高，聚类效果越好。具体计算方法如下：对于每个样本点 i，计算点 i 与同一个簇中的所有其他元素的平均距离值，记作 $a(i)$，用于量化簇中的内聚性。选取 i 外

的一个簇 b，计算 i 与 b 中所有点到簇内的平均距离，遍历所有其他簇，找到最近的平均距离,记作 b(i)，记录为 i 的邻居类，用于量化簇间的间隔。对于样本点 i，轮廓系数 $s(i)=\frac{[b(i)-a(i)]}{\max\{a(i),b(i)\}}$ 计算所有 i 的轮廓系数，平均值为当前聚类的整体轮廓系数，衡量数据聚类的紧凑性。由以上公式不难发现，如果若 s(i) 小于 0，说明 i 与其簇内元素的平均距离小于最近的另一个簇，说明聚类效果不好。如果 a(i) 趋于 0，或者 b(i) 足够大，即 a(i) < K 值确定：在实际应用中，此实验 Kmean 作为辅助聚类和标记。计算当前 k 的平均轮廓系数，最后选取轮廓系数最大的值对应的 k 作为最终的聚类数。

最终得出轮廓系数如图 7-8 所示：

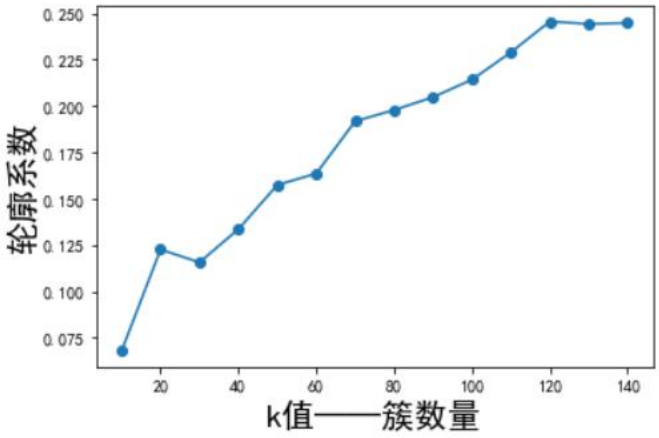

图 7-8 轮廓系数

又可以得出热点词的词云图，如图 7-9：

图 7-9 热点词词云图

7.2.2 检测留言中的热点问题

通过基于 TD - IDF 的文本相似度检测同一类群的热点词群在某段时间间隔的问题回复对象文本，发现留言中的热点问题。根据热度公式：

$$热度指数 = \frac{同类问题数量}{时间间隔}$$

最终得出表 7-2 热点问题表和表 7-3 热点问题留言明细表。

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	4.71157E-06	2019/6/10至2019/9/9	A7县星沙凉塘路旧城	旧城改造
2	2	3.56025E-06	2019/8/12至2019/10/12	泉星公园	泉星公园项目规划进一步优化的建议
3	3	1.87118E-06	2019/9/5至2019/9/25	魅力之城小区、3区梅溪湖看云路	烧烤夜宵摊油烟直排扰民
4	4	8.91158E-07	2019/2/15至2019/3/28	A市南山十里天池	A市南山十里天池虚假宣传及违规收取物业费
5	5	8.52989E-07	2019/3/26至2019/4/12	A6区月亮岛	A6区月亮岛路沿线架设110kv高压线杆的投诉

表 7-2 热点问题表

	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
2	A00035626	A7县星沙凉塘路旧城改造究竟要拖到何年何月才能开始?	2019-09-02 14:32:27	星沙的旧城改造今年就要全部结束了，而凉塘路的几排房子	6	0
3	A00035628	A7县星沙四区凉塘路旧城改造要拖到何年何月才能动工	2019-08-02 10:05:02	A7县星沙街道的旧城改造今年就要全部完工了，就偏偏留下	6	0
4	A00035631	A7县星沙街道凉塘路旧城改造什么时候可以进行	2019-06-10 10:24:59	2017年6月17日星沙街道凉塘路居民按旧城改造指挥部张贴	5	0
5	A00072477	A7县星沙四区凉塘路旧城改造要待何时	2019-07-04 14:10:30	在2017年6月份，星沙四区凉塘路的居民根据相关部门的通知	4	1
6	A00035629	A7县凉塘路的旧城改造要拖到何时才能动工?	2019-07-31 16:42:51	A7县星沙街道的旧城改造今年就要全部完工了，就偏偏留下	1	0
7	A00072486	A7县星沙街道凉塘路的旧城改造什么时候会启动?	2019-07-23 07:39:44	A7县星沙街道四区凉塘路群众在2017年6月时按照A7县旧城	1	0
8	A00098215	请问A7县星沙凉塘路的旧城改造要拖到何时何月何时才能再次启动?	2019-09-09 08:20:47	A7县星沙的旧城改造就快全部结束了，而凉塘路的几排房子	1	0
9	A00080342	建议A市经开区泉星公园项目规划进一步优化	2019-08-12 13:15:05	目前A市经济技术开发集团有限公司的泉星公园项目的规划拱	16	0
10	A00036841	给A市经开区泉星公园项目规划进一步优化的建议	2019-08-26 13:00:06	目前A市经济技术开发集团有限公司的泉星公园项目的规划拱	13	0
11	A00080343	请问A7县泉星公园何时开工，工期多长?	2019-10-16 13:03:13	泉星公园项目已经筹备了5年之久，人民群众非常期待，请	12	0
12	A00080342	A市经开区泉星公园项目规划需优化	2019-08-09 16:47:36	目前A市经济技术开发集团有限公司的泉星公园项目的规划拱	4	0
13	A00080343	建议公示A7县经开区泉星公园项目景观效果图	2019-09-29 13:25:08	泉星公园项目已经筹备了5年之久，人民群众非常期待，目	2	0
14	A00039089	魅力之城小区临街门面油烟直排扰民	2019-09-05 12:29:01	魅力之城小区楼下烧烤摊、快餐店无证经营，长期油烟烧烤	3	0
15	A324156	魅力之城小区临街门面油烟直排扰民	2019-09-05 12:29:01	魅力之城小区楼下烧烤摊、快餐店无证经营，长期油烟烧烤	0	3
16	A000106865	A3区梅溪湖看云路一师润芳园小区临街门面烧烤夜宵摊	2019-09-25 00:31:33	A3区梅溪湖看云路一师润芳园小区临街《山里人夜宵》《张	2	0
17	A00042313	A3区梅溪湖看云路一师润芳园小区临街门面油烟扰民	2019-09-05 12:23:38	《张傅师烧烤》《津市米粉菜馆》《山里人夜宵》《大众家	1	0
18	A00054842	A5区劳动东路魅力之城小区临街门面烧烤夜宵摊	2019-09-25 00:31:33	A5区劳动东路魅力之城小区临街夜宵摊、烧烤摊24小时经营	1	0
19	A0012425	A5区劳动东路魅力之城小区临街门面烧烤夜宵摊	2019-09-25 00:31:33	A5区劳动东路魅力之城小区临街夜宵摊、烧烤摊24小时经营	0	1
20	A00018255	投诉A市南山十里天池开发商	2019-02-15 15:56:31	领导，你好！我是十里天池的一名业主，我们楼盘本来合同	18	2
21	A00018927	A市南山十里天池虚假宣传及违规收取物业费	2019-02-15 16:52:20	1.南山十里天池小区在宣传及沙盘均表示交房时配套道路玉	6	0
22	A00021433	A市南山十里天池门前有土堆，业主出行难	2019-02-13 10:22:45	我是南山十里天池的业主，最近南山地产已经通知交付商品	5	0
23	A0005801	A市南山十里天池小区交房在即，但是小区正门却没有路	2019-02-15 15:55:29	我是南山十里天池的业主，我们目前小区前面是一潭土，出	5	0
24	A00018298	A3区南山十里天池（南蕃苑）楼盘交付存在问题	2019-03-05 08:19:16	尊敬的领导：您好！我是南山十里天池（南蕃苑）一期业主	3	0
25	A00021433	A市南山十里天池存在严重质量问题，开发商物业不闻不问	2019-03-28 17:32:04	领导好，我是A市南山十里天池小区的业主，这个月刚收房，	1	0
26	A00050125	A市南山十里天池玉佩路出行不便，开发商在踢皮球	2019-02-27 10:12:23	你好。对于南山十里天池的玉佩路，物业没权限，开发商踢	1	0
27	A00072424	关于A6区月亮岛路沿线架设110kv高压线杆的投诉	2019-03-26 14:33:47	联名信——坚决要求A市润和又一城、三润城、润和紫郡、滨	78	0
28	A00061602	关于A6区月亮岛路110kv高压线的建议	2019-04-09 17:10:01	尊敬的胡书记，您好！根据区政府和街道办及电力公司的回	55	2
29	A00099016	A6区月亮岛路架设高压电线环评造假，谁为民众做主	2019-04-08 21:19:40	A6区月亮岛路架设110kv高压电线，金星北多个小区呼吁：尽	22	0
30	A00072424	关于A6区月亮岛路沿线架设110kV高压电线杆的投诉	2019-03-26 10:17:21	联名信——坚决要求A市润和又一城、三润城、润和紫郡、滨	10	0

表 7-3 热点问题留言明细表

7.3 问题 3 的解答过程

对于问题 3，完整的解题流程图如图 7-10 所示：

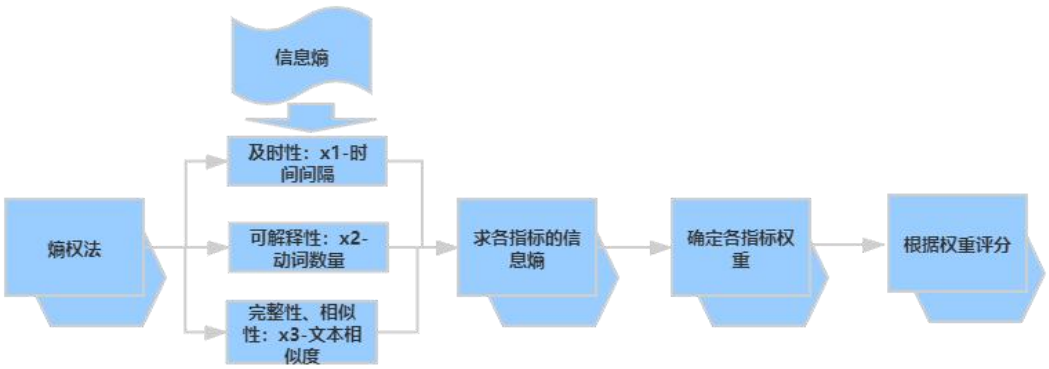

图 7-10 问题 3 解题流程图

7.3.1 各个评价的指标

（1）相关性及完整性。针对相关部门对留言的答复意见，从答复的相关性以及完整性考虑，我们可以采用文本的相似度即用回复的关键词和留言的关键词进行比较，如果相似度高即代表回复的相关性以及完整性高。对于文本相似度的算法，我们采用的是余弦相似度。余弦相似度就是用向量空间中两个向量夹角的 $\cos(\theta)$ 值作为衡量两个个体间差异的大小。$\cos(\theta)$ 越接近 1，就表明夹角越接近 0 度，即两个向量越相似，这就叫"余弦相似度"。

通过该原理，我们将文本进行，统计文本中的词频，写出词频向量，然后问题就变成了如何计算这两个向量的相似程度。我们可以把它们想象成空间中的两条线段，都是从原点([0,0,...])出发，指向不同的方向。两条线段之间形成一个夹角，如果夹角为 0 度，意味着方向相同、线段重合,这是表示两个向量代表的文本完全相等；如果夹角为 90 度，意味着形成直角，方向完全不相似；如果夹角为 180 度，意味着方向正好相反。因此，我们可以通过夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似，为正向指标。余弦相似度公式如下：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} * \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

（2）可解释性。我们通过人工检测发现寻找一些解释性比较强的回复内容作为样本文本集，通过统计分析其中形容词、动词、名词，我们发现动词数量与人工检测判定的回复文本可解释性有较强的为正相关。通过总样本的统计直方图，我们认为，动词属于谓语，是主语和宾语的中间部分，每个句子的组成也通常更需要动词，并且在政务处理中形容词不是解决问题的关键词，回复句子中动词比形容词更体现出如何解决，比如“归纳”、“召开”、“执行”、“责成”、“协商”、“拆迁”、“提高”、“沟通”、“确保”、“力争”等等。因此，我们采用提取“留言主题+留言内容”和“回复”的关键字文本特征，检测动词数量，由此反应文本可解释性。动词数量越多，表示回复内容对问题处理的描述越详细，为正向指标。如图 7-11 解释性词性统计：

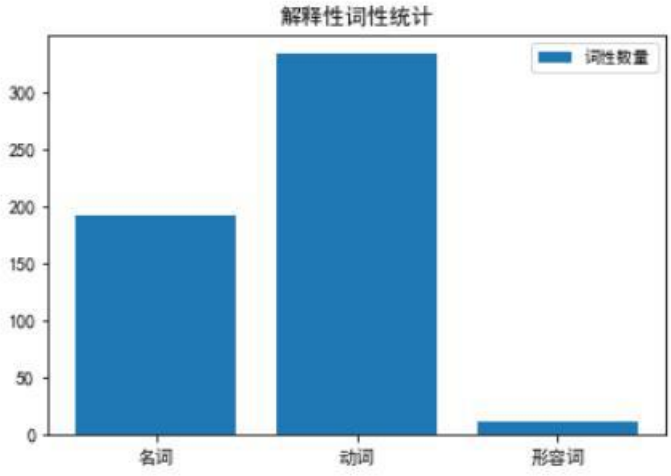

图 7-11 解释性词性统计

（3）及时性。“留言与回复”的时间间隔，它在一定程度上也反应了政务处理效率，及时处理群众问题也是重要指标之一，如果单纯解决办法好但不够及时，那么再好的解决方法也是会降低群众对回复质量的评价，所以我们将“留言与回复”的时间差加入政务留言回复质量的考察指标之一，“留言与回复”的时间差越长，表示在回复效率越低，为负向指标。

7.3.2 信息熵

对于各个评价指标我们采用的评价模型是熵权法。谈到熵权法，首先要了解信息熵。

（1）信息熵 E_j 常被用来当作系统的信息含量的量化指标，信息熵可以作为样品特性最佳分配的根据。对信息熵，一般来说，若某个指标的信息熵越小，表明指标值得变异程度越大，提供的信息量越多，在综合评价中所能起到的作用也越大，其权重也就越大，综合评估就越有用，其重要性就越大。相反，信息熵越大，所提供的信息量越小，总体评估的重要性就越小。

然而，在原有信息熵的定义中，若某个指标的信息熵越小，表明指标值得变异程度越大，提供的信息量越多，对质量评价是负向指标的时间差可以，但对判定总体质量的正向指标动词数量，却与之相反，故在使用熵权法之前，我们采用了正信息熵（即原有的概念）和负信息熵（在原有的公式增加负号）的划分，正向指标对应负信息熵，负向指标对应正信息熵，从而对正负指标归一化处理。

在处理政务文本信息时，通过判定指标对总体质量的影响是正相关还是负相关，修改对该信息熵的具体公式。

（2）熵权法赋权步骤有：

- 数据标准化：将各个指标的数据进行标准化处理。假设给定了 k 个指标 $X_1, X_2, ..., X_k$，其中 $X_i = \{x_1, x_2, ..., x_n\}$，假设对各指标数据标准化后的值为 $Y_1, Y_2, ..., Y_k$，那么：

$$Y_{ij} = \frac{x_{ij} - \min(x_i)}{\max(x_i) - \min(x_i)}$$

- 求各指标的信息熵：根据信息论中信息熵的定义，信息熵分为正信息熵为和负信息熵，正信息熵（即原有的概念）和负信息熵（在原有的公式增加负号），正向指标对应负信息熵，负向指标对应正信息熵，一组数据的信息熵 $E_j = -\ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij}$。其中 $p_{ij} = \frac{Y_{ij}}{\sum_{i=1}^{n} Y_i}$，如果 $p_{ij} = 0$，则定义 $\lim_{F_n \to 0} p_{ij} \ln p_{ij} = 0$。

- 确定各指标权重：根据信息熵的计算公式，计算出各个指标的信息熵为 $E_1, E_2, ..., E_k$，通过信息熵计算各指标的权重：$W_i = \frac{1 - E_i}{k - \sum E_i} (i = 1,2,...,k)$。

假设：

X1——用留言与回复的时间间隔体现其及时性

X2——用动词数量表示在回复内容中的详细的程度，体现其可解释性

X3——用文本相似度体现其完整性和相似性

根据指标权重的计算公式，可以得到

指标 X1、X2、X3 的权重表 7-4

X1	X2	X3
0.331136604343674	0.3456504615726636	0.32320731779929696

表 7-4 权重表

7.3.3 对评价模型评分

对每一个回复进行评分。根据计算出的权重，以及对每个回复 3 个指标的评分。设 Z_l 为第 l 个回复的的最终得分，则 $Z_l = \sum_{i=1}^{3} X_{li} W_i$，各个回复最终得分如下图 7-12：

	A	B	C	D	E	F
1		x1	x2	x3	score	rank
2	2038	0.194758	1596306	529	76.31403	1
3	1919	0.140474	176903	529	73.90854	2
4	2036	0.140474	924149	528	73.59173	3
5	2048	0.615844	706543	17	67.95758	4
6	770	0.473681	721982	103	65.6016	5
7	2084	0.064575	879407	358	59.23936	6
8	230	0.448632	287716	14	59.0109	7
9	388	0.318477	2745379	134	58.52252	8
10	1579	0.449436	416447	2	58.28086	9
11	420	0.391842	1996627	53	57.76836	10
12	1130	0.354185	3280569	80	56.96194	11
13	83	0.415486	1470371	14	56.83979	12
14	2636	0.087947	8779566	341	56.74471	13
15	664	0.391842	495801	11	55.73198	14
16	691	0.398047	1100655	8	55.67307	15
17	2329	0.354185	882841	36	55.11146	16
18	2215	0.137832	190186	215	54.69368	17
19	907	0.376053	339713	3	54.45847	18
20	4	0.358173	1356491	14	53.82551	19
21	2129	0.112343	5469564	248	53.5307	20
22	78	0.192397	1121161	149	53.27331	21
23	1134	0.252334	3014206	107	53.26696	22
24	627	0.346925	207358	6	53.13423	23
25	1940	0.221646	2980971	131	53.10101	24
26	1605	0.318477	1040538	33	52.97293	25
27	454	0.252334	2580445	97	52.8083	26
28	1085	0.318477	607280	25	52.63555	27
29	152	0.192397	2085233	141	52.45709	28
30	1029	0.318477	487545	21	52.43366	29
31	2479	0.318477	332876	14	52.06155	30
32	1737	0.020658	3280557	281	51.40055	31
33	568	0.284571	440085	33	51.37239	32

图 7-12 评价模型评分图

8. 结论

近年来，随着信息爆炸的大数据时代到来，以往传统的靠人工的政府管理水平和施政效率面临极大的挑战，因此如何使用大数据技术在自然语言处理和文本挖掘的方法应用在“智慧政务”中的文本挖掘，提升政务的效率，具有极其重要的研究意义。

本文首先对数据分析并进行预处理，经过分词操作、去停用词和低频词、提取文本特征向量后获得较干净的文本词向量。

解决问题 1 的群众留言分类处理。主要方法是设置训练和学习文本集，用基于二分类的一对多策略对文本进行分类，用 F-Score 评价分类效果。总体模型的评价 F-Score 为 0.9254433586681143，表现出较为良好的分类效果。具体得到的分类在**结果数据（文件夹）**下的**附件：Ti_1_Result.xls**

解决问题 2 的热点整理方面。采用先求热点词，后求热点问题，我们先对文本特征向量的词条进行平滑处理，过全局加权平均法进行平滑处理，然后研究在时间序列上同一词条的倍数关系，从而找到热点词，再通过提取热点词进行变量聚类。最后，将热点词类群用 TF-IDF 的文本相似度检测出热度指数留言内容中的热点问题。具体得到的热点问题表和热点问题详细表在**结果数据（文件夹）**下的**附件：热点问题表.xls、热点问题留言明细表.xls**

解决问题 3 的对答复意见的质量评价方面。如何评价答复意见的质量评价，我们从及时性、可解释性、相关性、完整性的角度对答复意见值建立一套评价模型。我们先从不同角度研究其对应的不同的指标，通过指标对总体质量的正负影响程度，其中的指标有动词数量、留言与回复的时间间隔、余弦相似度，采用熵权法，将指标信息熵化，然后确定各指标的熵权重，再根据权重评分答复意见的质量评价打分。具体关于答复意见的质量评分在**结果数据（文件夹）**下的**附件：回复排名.xls**

9. 参考文献

[1] 平源. 基于支持向量机的聚类及文本分类研究[D]. 北京邮电大学.

[2] 姜伦. 模糊聚类算法及其在中文文本聚类中的研究与实现[D]. 哈尔滨理工大学.

[3] 于游，付钰，吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报，2019(5).

[4] 郝立丽. 汉语文本数据挖掘[D]. 吉林大学.

[5] 黄春梅,王松磊.基于词袋模型和TF-IDF的短文本分类研究[J].软件工程,2020,23(03):1-3.

[6] 黄鑫沛,宋斐,李艳婷,夏唐斌.基于组合赋权法的海外基建环境综合评估研究[J/OL].工业工程与管理