第八届“泰迪杯”全国数据挖掘挑战赛论文报告

所选题目：C 题 ：“智慧政务”中的文本挖掘应用

基于 skip-gram 模型和密度的聚类算法的文本挖掘系统

综合评定成绩：_____

评委评语：

评委签名：

基于 skip-gram 模型和密度的聚类算法的文本挖掘系统

摘　要

在当今这个信息技术飞速发展的时代，互联网与我们的生活息息相关。互联网对于政务的帮助也是日渐加深，“智慧政务”充分利用物联网、云计算、移动互联网、人工智能、数据挖掘等新一代信息技术为基础，构建一个政府、市场和社会多方协同的公共价值塑造，实现政府管理与公共服务的精细化、智能化、社会化，实现政府和公民的双向互动，让市民充分享受信息时代给公民生活带来的便捷，是“互联网+”在民生服务领域的落地。

随着各类网络问政平台的使用率不断攀升，其中各类社情民意相关的文本数据量也在不断增涨，若仍旧使用传统的人工方式来进行留言划分和热点整理，效率将会相当低下。此时基于自然语言处理技术的智慧政务系统成为了这类问题的必然趋势。

所以我们设计了一套算法流程，大量的留言数据进行数据清洗、词向量转化、通过 word2vec 中 SKIP-GRAM 模型原理提取出留言数据关键词，进行留言分类，通过基于密度的聚类算法对于社会热点问题进行提取，最后设计了一套对于答复意见的评价流程，监督政府相关部门工作人员的答复情况，以及回收市民对于其答复内容的满意程度，根据其反馈的意见进行政务工作的再优化。

在留言主题和留言内容关键词的文本挖掘当中，由于需要将留言内容分布到七个一级主题，会用到较高的维度，所以我们使用了两个循环，即先通过指向性明确的留言主题将其分类到相应的一级主题，再将剩下的留言内容通过循环系统进行分类，将高维度问题降成低维度问题。

这套完整的算法流程，经过测试，我们的系统对于政务留言的分类、热点问题提取以及回复意见的评价有着一定的效果，相信可以对于数据挖掘、文本提取等方面的政务工作有所帮助。

关键词：中文文本清洗分类、SKIP-GRAM 模型、基于密度的聚类算法

Abstract

In the age of rapid development of information technology, the Internet is closely related to our life. Internet for government assistance is also deepening, "smart government" to make full use of the Internet of Things, cloud computing, mobile Internet, artificial intelligence, data mining and other new generation of information technology as the basis for building a government, market and social coordination of public value shaping, to achieve government management and public services, intelligent, social, to achieve the two-way interaction between the government and citizens, so that citizens fully enjoy the information age to the convenience of citizens' lives, is "Internet" in the field of people's livelihood services.

With the use of various types of network political platform surging, including all kinds of social sentiment public opinion-related text data volume is also increasing, if the traditional manual way to divide messages and hot-spot finishing, efficiency will be quite low. At this time, the intelligent government system based on natural language processing technology has become the inevitable trend of this kind of problem.

So we designed a set of algorithmic process, a large number of message data for data cleaning, word vector conversion, through word2vec SKIP-GRAM model principle extracted message data, message classification, through density-based clustering algorithm for social hot issues, and finally designed a set of evaluation process for the response to comments, supervision of the government related departments staff response, as well as the recovery of public satisfaction with its response content, Re-optimization of government work based on the feedback from them.

In the text mining of message theme and message content keywords, because of the need to distribute the message content to seven first-level topics, will use a higher dimension, so we used two loops, that is, first through the point of clear message theme to classify it into the corresponding level of the topic, and then the remaining message content through the loop system classification, the high dimension problem reduced to a low dimension problem.

This complete algorithm process, after testing, our system for the classification of government messages, hot issues and comments to the evaluation has a certain effect, I believe that can be for data mining, text extraction and other aspects of government affairs work is helpful.

Keywords: Chinese text cleaning classification, SKIP-GRAM model, density-based clustering algorithm

目　　录

1 研究背景

随着物联网、大数据、数据挖掘、云计算等现代信息技术的不断发展，“互联网”在人们的生活中占比逐渐增大，使用计算机代替传统人工对数据和文本进行处理是必要的倾向，使用计算机处理不但能提高效率，还能在一定程度上提高操作的准确性，使用计算机处理还可以将无用的、重复的信息，使有限的人工可以快速处理关键问题。

互联网以及数据挖掘对于“智慧政务”的帮助是可以想见的，随着互联网技术和时代文化的日新月异，文本和词汇呈现出多元化、更新快的特点，加之中文的语序顺序以及词义表达特殊，在文本的分析及文本句意的表达分析方面给文本分类工作带来了巨大的挑战。于是我们需要一个可以辨明句意，同时为热点问题的收集提供帮助的文本挖掘应用的出现。

2 分析方法

2.1 问题分析

2.1.1 问题一

一、检索关键词

1、预处理：文本清洗，去除无用的信息（即文本提取）

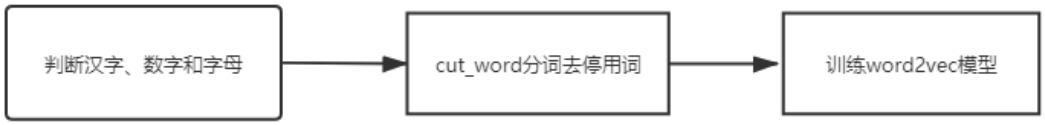


图 1 预处理流程

因为“智慧政务”的信息收集渠道较为正规，所以收集上来的文本大多经过市民的斟酌，对于其中的文本清洗，目的是去掉多余的部分，为接下来的文本分类做准备。由于中文文字的特殊性，我们需要对于文本进行预处理，去掉文本中冗杂多余的部分，如标点、介词等。我们使用相应函数对于字符串进行筛选的处理，输入待处理的字符串，返回清洗后的字符串，将标点、介词等无用信息删除。然后对于 word2vec 进行

词向量训练，计算两个词的相似度，找出与某个词最相近关系的词，找出一段文本中不同类的词，并保留模型，方便接下来的使用。

2、文本关键词提取算法：基于词的关联信息的特征量化

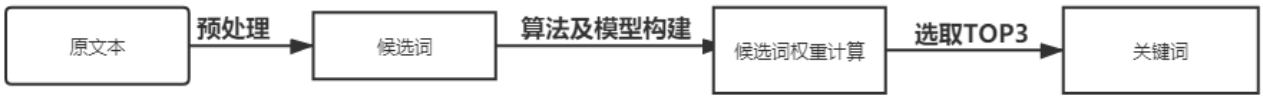


图 2 文本关键词提取流程

词的关联信息是指词语之间、词与文档之间的关联程度信息，包括互信息、TF-IDF 值。

首先针对词性、词频和位置信息，一般在一段文本中，词性为动词的的词的附近的词更加等代表这段文本的中心含义，所以我们将针对动词附近的词进行筛选；其次词频高以及位置靠前的词语，也是文本撰写人想要重点表达和引人注意的词，针对这个特性，我们设计文本关键词提取程序。

其次我们针对于词跨度，即一个词或者短语在文本中首次出现和末次出现的距离进行文本关键词提取程序优化，一个词跨度越大说明这个词对于这个文本越重要，可以反映文本主题。词跨度计算公式如下（图 3），通过这个方法，可以减少文本中的噪声，提高关键词选取的准确性。

$$span_i = \frac{last_i - first_i + 1}{sum}$$

图 3 词跨度计算公式

最后我们针对一个词的 TF-IDF 值对他赋予权重，然后选取候选词权重中的 top3 来成为这段文本的关键词。TF 是指这个词在文档中出现的频率，假设一个词 w 在文本中出现了 m 次，而文本中词的总数为 n，那么一个词的 IDF 是根据语料库得出的，表示这个词在整个语料库中出现的频率。假设整个语料库中，包含词 w 的文本一共有 M 篇，语料库中的文本一共有 N 篇，则

$$IDF_w = \log_2 \frac{N}{M}$$

由此可得词 w 的 TF-IDF 值为：

$$TFIDF_w = TF_w \times IDF_w$$

通过以上方法，我们设计文本关键词提取程序，将关键词提出，进行接下来的步骤.

二、SKIP-GRAM 模型（图 4）

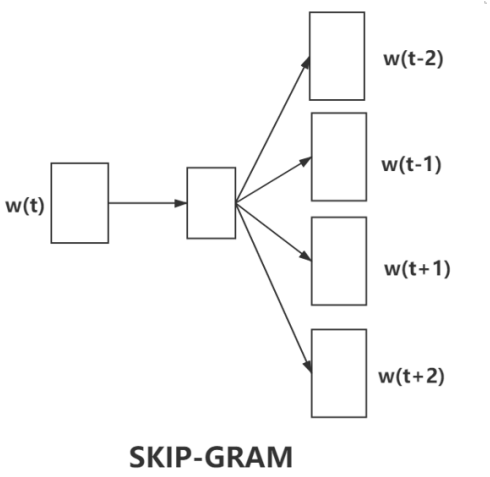


图 4 SKIP-GRAM 模型

进行一次以上的步骤后，我们会得到一个粗略的文本分类，即通过留言主题中的关键词对于留言进行分类进入一级标签，但是还是有些留言主题中会出现带有两种及以上的一级标签的性质，所以我们需要对于留言详情进行二次检索，程序同上。在留言详情检索后，大部分的留言都被分进了对应的一级标签中，还剩下少部分的留言包含了太多混淆的因素，需要运用传统人工方式进行区分。

三、F-score 评价分类方法

通过模型我们可以得到精确率(Precision)和召回率(Recall)评估指标，我们通过这两个指标可以使用 F-score 评价分类方法来对于此分类方法进行评价，从而更好地优化本模型。对于 Precision 和 Recall，虽然从计算公式来看，并没有什么必然的相关性关系，但是，在大规模数据集合中，这 2 个指标往往是相互制约的。理想情况下做到两个指标都高当然最好，但一般情况下，Precision 高，Recall 就低，Recall 高，Precision 就低。所以在实际中常常需要根据具体情况做出取舍，例如一般的搜索情况，在保证召回率的条件下，尽量提升精确率。而像癌症检测、地震检测、金融欺诈等，则在保

证精确率的条件下，尽量提升召回率。

所以，很多时候我们需要综合权衡这 2 个指标，这就引出了一个新的指标 F-score。这是综合考虑 Precision 和 Recall 的调和值。

$$F-Score=(1+\beta^2)\cdot\frac{Precision\cdot Recall}{\beta^2\cdot Precision+Recall}$$

当 β =1 时，称为 F1-score，这时，精确率和召回率都很重要，权重相同。当有些情况下，我们认为精确率更重要些，那就调整 β β 的值小于 1，如果我们认为召回率更重要些，那就调整 β 的值大于 1。

2.1.2 问题二

一、文本归纳及分类

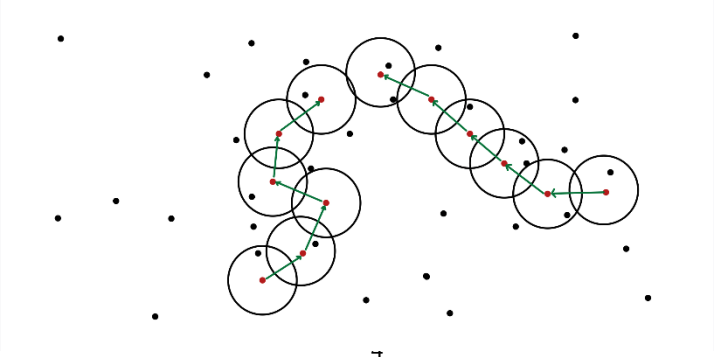
（1）.根据附件三的具体信息，我们对已有的文本（留言主题和留言详情）进行缩句，即提取句子主要成分。重点提取信息是特定的时间、特定地点、发生的问题。

（2）密度聚类算法

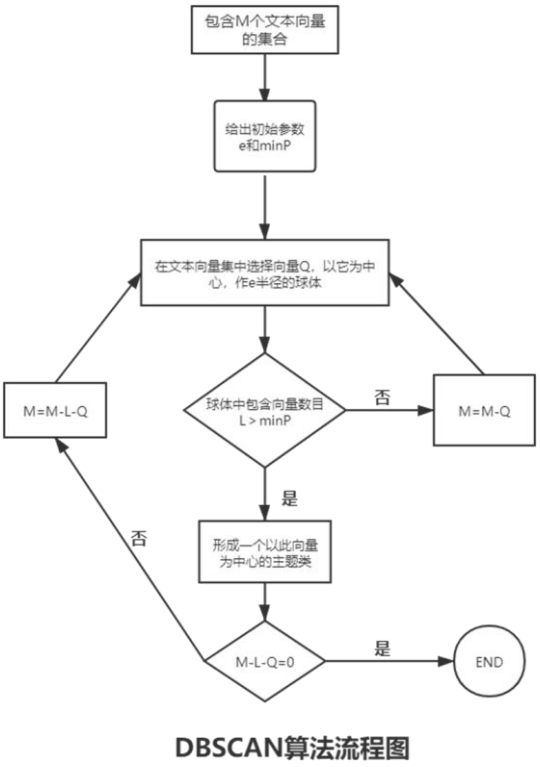
DBSCAN（Density-BasedSpatial Clustering of Applications with Noise）是一种基于密度的聚类方法（如图表 1）。DBSCAN 的聚类定义是：由密度可达关系导出的最大密度相连的样本集合，即为我们最终聚类的一个类别，或者说一个簇。

在每一个类别或主题中可以有一个对象或者多个重点对象。如果只有一个重点对象，那么在这一个类别中，其他非重点对象都属于这一个重点对象所在的领域中；如果有多个重点对象，那么类别中的任意一个重要对象所属的领域中一定有一个其他的重要对象，否则这两个重要对象无法密度可达。

DBSCAN 使用的方法如下，对于留言主题和留言详情，判断在可行范围内是否包含足够多的样本点，我们会先预估一个最小的样本点数目，然后对于有密度连接性的文本对象进行类别分类。具体操作如图表 2



图表 1



图表 2

二、热度评价指标

定义

热度评价指标是指通过多个方面多层次对某一信息或关键词出现次数、反映情况的评定。

具体内容

对于群众反映的热点问题，从同类具有关联度性文本个数、问题提出次数、关键词频率、点赞和反对数四个维度进行给分，四个维度分数占比是 40%、20%、20%和 20%（满分 100）。给分的规则是基于每一个维度里的次数，如在同类具有关联度性文本个数这一维度中，出现相关联文本最多的文本可得 40 分，出现相关联文本次数第二名的则按照出现次数占第一出现次数的比例，乘以这一维度的总分 40 分得到结果。另外，点赞是加分，而反对则是减分。最后，各项得分加在一起，总分最高的则是热点越高。

2.1.3 问题三

我们对本题基于答复的完整性、相关性、可解释性的顺序对答复意见的质量提出评价方案。我们根据内部检索以及市民评星级为标准对答复意见进行评分评价。具体方案如流程图（图 5）所示。

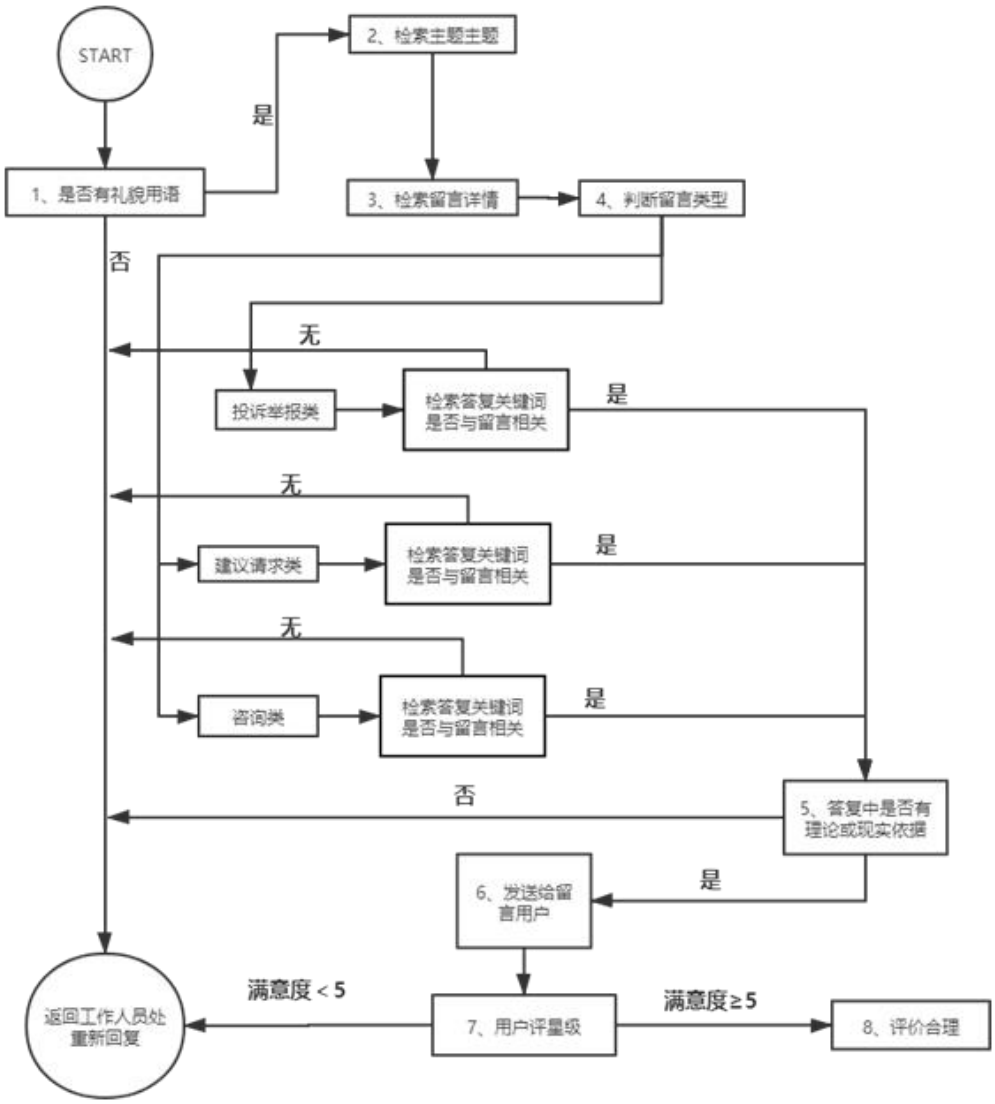


图 5 问题三流程图

一、内部检索测评

①从完整性来看，政府相关部门对于市民提出的留言问题进行答复，需要满足规范，按照标准流程进行答复。

首先是礼貌用语，中国作为礼仪之邦，待人处事懂礼仪、知进退，所以作为政府部门更要以身作则，对待市民的留言，在答复中，必不可缺少敬语称呼。根据政府提

高市民满意度各项资料数据显示，有敬语的回复使市民满意度更高，所以答复的开头使用礼貌用语，能够使市民获得重视感、尊重感，一个好的开头，留下好的印象，提高市民的满意度。根据以上说明，我们认为答复是否有礼貌用语为评价的第一点准则，也是一个答复完整性的开头。

其次是判断类型进行分类，市民的留言数量庞大、类型多种多言，所以政府部门必须对留言进行分类，以便系统地进行答复，同时减轻工作量。按照留言时间先后顺序进行列表排序，利用文本特征选择方法，使用文档频率的数学方法针对留言主题选取最据分类信息特征，在留言主题的特征中，我们对留言详情进行检索，使用 word2vec 对留言详情进行分类列表。从而我们获得投诉举报、建议请求、咨询这三大类留言。确立留言类型，根据不同类型分配任务给不同的工作人员，各司其职，能够更好地针对市民的留言进行明确分析、答复。

然后对答复内容进行检索，判断其相关性以及内容的完整性。检索每一条答复意见中的关键词，判断它与留言主题关键词是否匹配。通过量化分析处理，对答复内容是否完整正确进行判断。以上两点皆为是则进行下一步，若为否，系统则自动返还工作部门，工作部门需进行查漏补缺，重新进行内部测评。对答复内容进行检索分析，可以减少答复结果的错误率，减少市民不满意而重新工作的时间，提高政府管理水平和施政效率以及市民满意度。

最后将答复发送给留言用户，并附与评星表。在收到留言到答复意见发送给留言用户之间所需要的时间也是重要的评价标准之一，确保半个月内进行回复。在数字参考咨询环境下，用户对于即使回复和立即获得答案的期待更加强烈。

故答复中含有礼貌用语、回复时间短且答复具有高度的完整性，用户的满意度将会更高

②从相关性看，政府答复意见的内容必须与留言问题相关，不相关则为无效答复。政府必须针对用户留言问题进行系统分析、数据挖掘，正确掌握问题的核心。

根据用户的具体问题，相关部门应该对工作人员进行分类，提高工作人员对于某一类型问题答复的专业性。具有专业性的答复人员，对于用户留言也能进行更专业化速度化的答复，并且相关工作人员可以通过云计算，在程序中输入准确的特定的关键词从而提高个人工作效、提高用户满意度。

对于答复意见的相关性，内部评价系统需要不断优化升级，才能针对各种各样的庞大的留言问题进行数据挖掘，保证答复真实有效。

③从可解释性的角度来看，政府相关部门给出的答复是需要有理论支撑的，并且是能够让用户理解、实操的。

政府给出的答复要有相关的法律法规、政策支持，按照实际情况答复，不能无根据答复，答复必须严谨。即使是无法解决或是尚且未能解决的问题也要给出对应的解释，不能敷衍了事。所以政府给出的答复意见必须要根据实际情况回答，有相关的解释和理论的支撑。

政府给出的答复必须要使市民可理解的。留言用户的文化层次参差不齐、年龄也尽不相同，所以政府给出的答复不能是晦涩难懂的，而应该简单直白，让用户一看便知。简单易懂的答复便于用户迅速进行操作，解决问题，客户对政府的满意度、信赖度也会提高。

二、市民星级评价

市民对政府相关部门给出的答复意见进行满意度星级评价，这是对答复意见进行评价的最直接方式。

根据评价指标规定，最高星级为十星，超过五星为较满意该回复，该回复较好，该部门的工作顺利完成；若是低于五星则为不太满意自己所收到的回复，系统将自动把留言内容返回相关部门工作人员处，工作人员需重新回复。

市民进行星级评价能够提高市民的政治参与度，对政府答复规范和效率有了更高的要求。也能够直观地看出政府相关部门答复意见的效果。

内部检索测评和市民星级评价相结合是我们针对政府相关部门的答复意见所设计的评价方案，并利用文本特征选择进行数学建模，建立智能评价系统。

3 模型说明及流程图

3.1 实验平台

本文的实验在 word2vec 中进行。简单来说，word2vec 是一群用来产生词向量的相关模型。这些模型的作用在于表示词与词之间的关系。

3.2 总体流程图

3.2.1 问题一流程

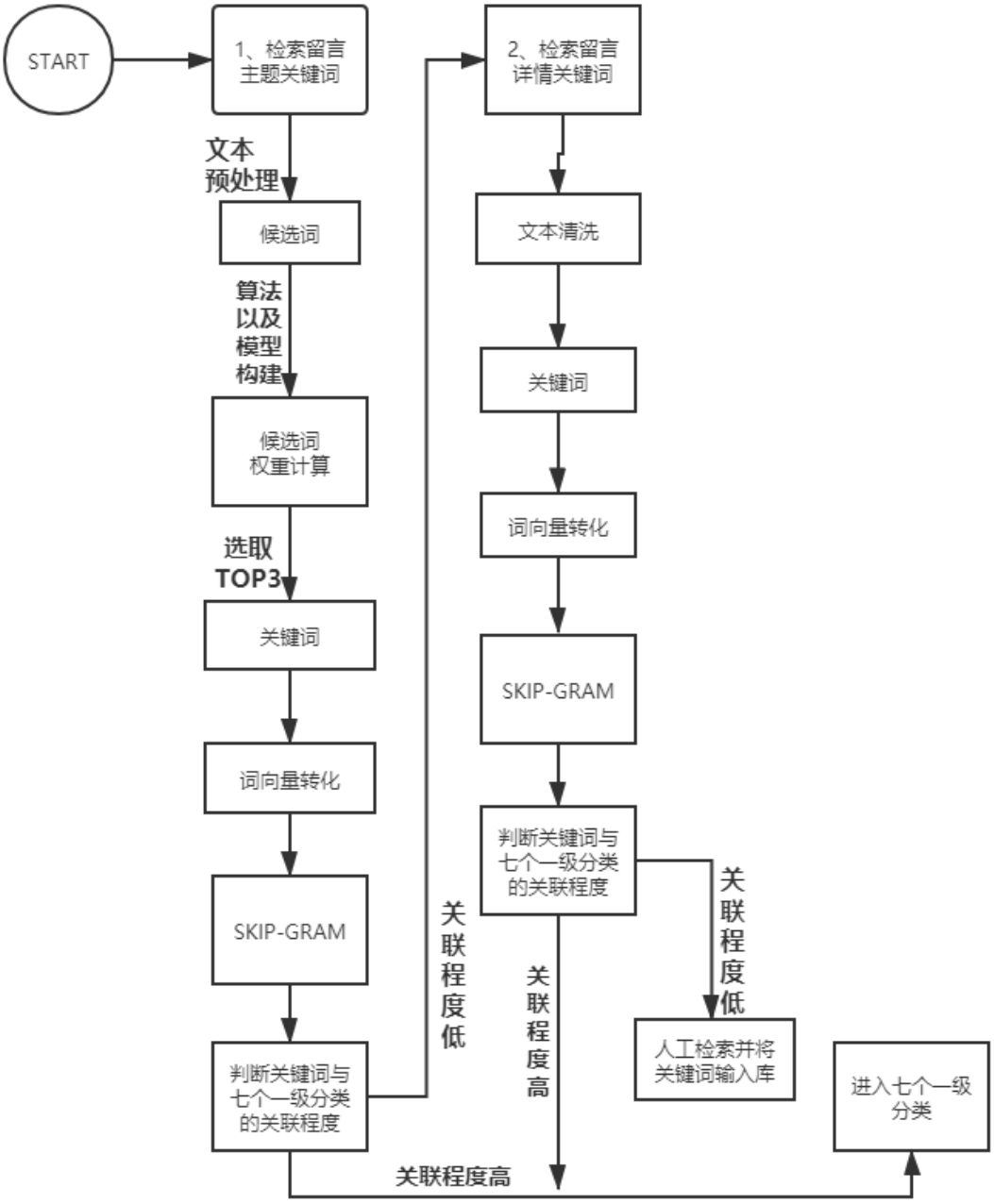


图 6 问题一流程图

3.2.2 问题二流程

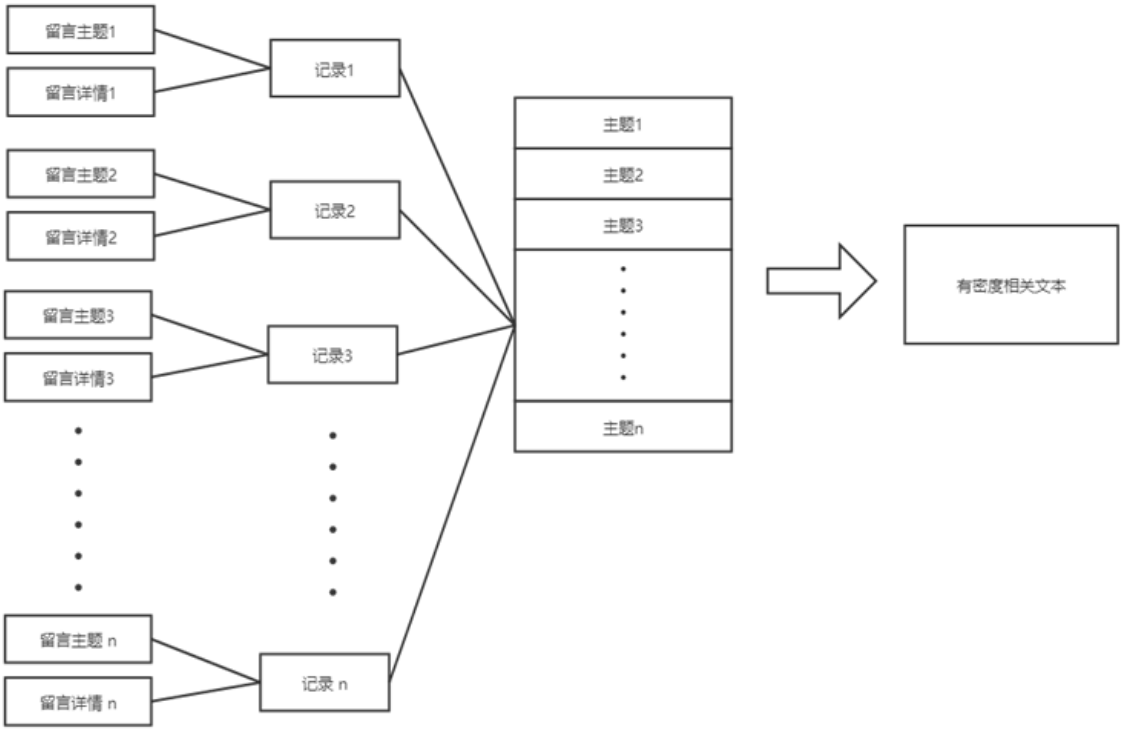


图 7 问题二流程图

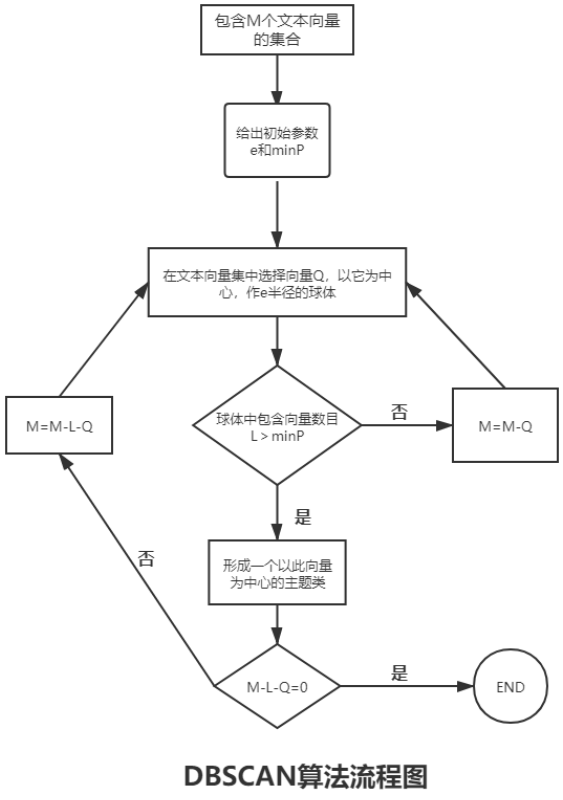


图 8 DBSCAN 算法流程图

3.2.3 问题三流程

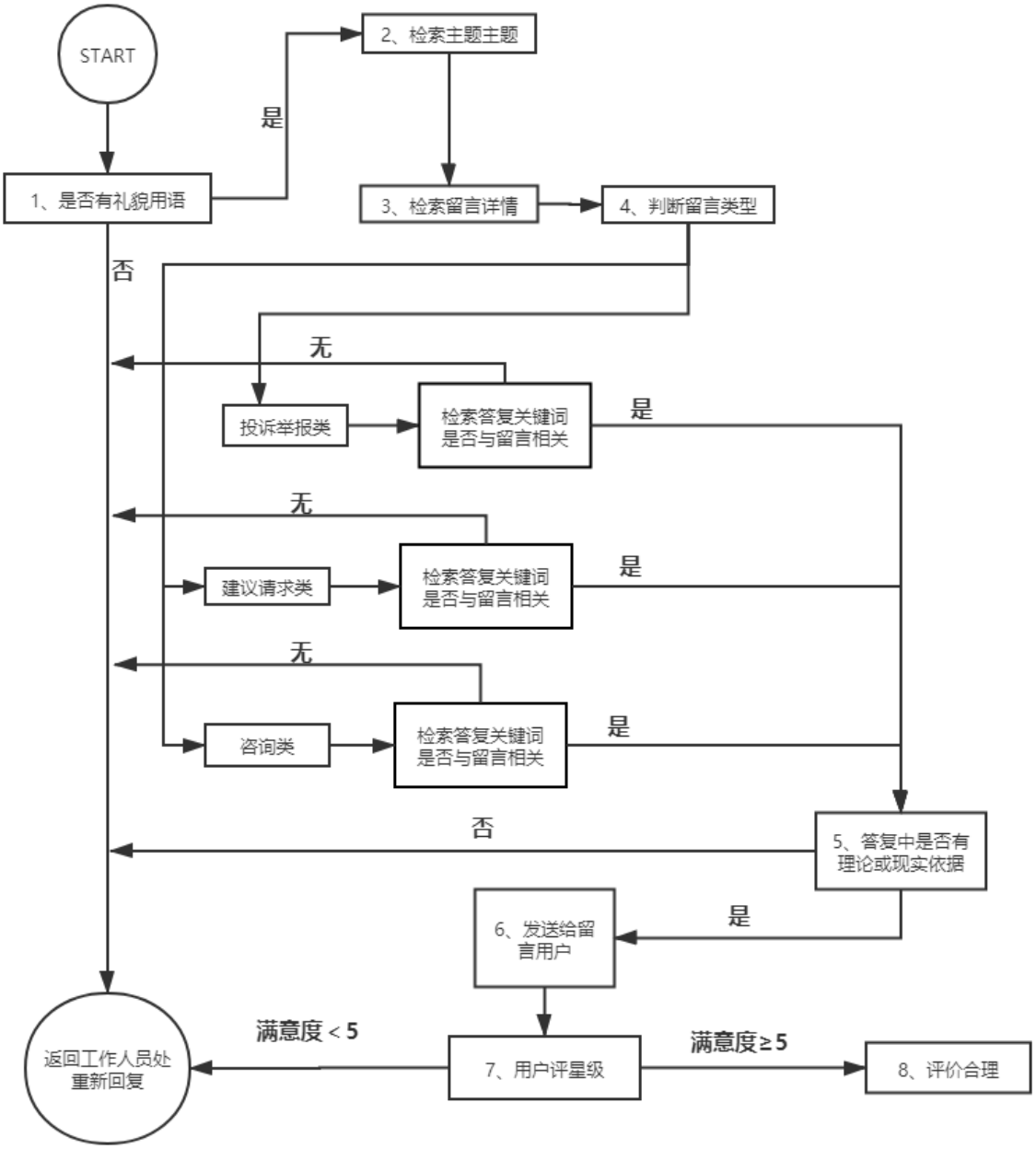


图 9 问题三流程图

4 优缺点分析

4.1 优缺点分析

4.1.1 模型优点

通过模型的分析我们可以知道，此文本清洗及留言分类模型的优点是可以辨别中文文本中部分词义表达模糊的问题，去除一定量的噪声以达到准确的分类，同时可以获得较快速的一级标题的分类。

针对于热点问题的分析，我们采用了多维度分析以及基于密度的聚类算法，使对于热点问题的分析有依据的同时可以随着留言不断叠加看到热点问题变化的倾向，使工作人员能尽快的针对于社会上的热点问题进行解决方案的策划和出台。

关于相关部门对于留言答复意见的评价，我们从答复的相关性、完整性、可解释性方面有一定的程序评价，但是由于反馈意见的市民有自己的思想和想法，我们设计了反馈意见的市民对于留言答复满意程度的意见反馈调查，通过真实可靠的数据去达到对于留言答复意见评价系统的更加完善。

4.1.2 模型缺点

对于文本分类，为了表达准确，我们设置了两个循环，这导致时间会比较长，无法一击即中的达成留言的分类工作，同时对于有两个及以上留言关键词特性的留言，我们仍需要传统的人工方式进行分类，这导致我们的程序在正式使用当中可能出现还是需要较多人工的帮助问题。此外，这个系统需要深度学习，所以无法及时快速的供给“智慧政务”系统进行正式使用。

在第二个问题的文本相关性算法中，有一个比较大的不足是计算较为复杂繁琐，需要文本与文本之间一一对应，比较耗时。同时，算法过程也需要优化。我们认为，一个良好的模型应该是通过精炼而又不复杂的算法就可以得出结论，因此我们小组在模型设计上仍需改进，争取更加优化模型。

针对留言答复意见的评价，由于考虑到尊重市民个人的喜好倾向，我们将市民的意见反馈也输入进评价系统，即市民的评价会对于答复的工作人员的工作质量有所评价和影响，所以如果有人恶意的大量给工作人员恶评，那么数据将很不好看，还是缺失了一些合理性和人性化。

对于内部检索检测

①若是系统出现故障，则将会堆积大量的未检测答复，所以需要优化程序，添加备用方案

②用户留言若是涉及多个关键词和不同类型，我们的程序便无法精准检索，可能会出现偏差。

③内部环节还是需要一定的人工支持，并没有达到最优化效果

5 结论

5.1 结果反思

对于“智慧政务”系统的具体使用，其实是有很多超越我们想象的方面的问题去考虑的，不单单只是我们在这个系统中考虑到的问题，还需要让系统有学习功能，在深度学习后才能应用到真实的“智慧政务”系统的具体使用当中，所以我们还需结合更多的专业知识以及结合更多“智慧政务”方面的传统人工处理中遇到的问题，来更好的优化系统，去改进出更加人性化更加符合真实生活中可以合理使用的文本挖掘应用。

6 参考文献

【1】 CSDN 博主「saltriver」，分类模型的评估方法-F 分数(F-Score)，https://blog.csdn.net/saltriver/java/article/details/74012163 ，2020/5/6

【2】 简书博主「华小锐」，中文文本清洗、分词、训练词向量，https://www.jianshu.com/p/2a75954cda13，2020/5/6

【3】 Cnblogs 博主「刘建平 Pinard」，word2vec 原理(一) CBOW 与 Skip-Gram 模型基础，https://www.cnblogs.com/pinard/p/7160330.html，2020/5/6

【4】 郝立丽. 汉语文本数据挖掘[D].吉林大学,2009.

【5】 时志芳. 移动投诉信息中热点问题的自动发现与分析[D].北京邮电大学,2013.

【6】 刘宁,陈凌云,熊文涛.基于文本挖掘的网络热点舆情分析——以问题疫苗事件为例[J].湖北工程学院学报,2019,39(06):60-64.

【7】 马小龙.网络留言分类中贝叶斯复合算法的应用研究[J].佛山科学技术学院学报(自然科学版),2013,31(02):43-47+68.

【8】 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(05):1-8.

【9】 薛金成,姜迪,吴建德.基于 word2vec 的专利文本自动分类研究[J].信息技术,2020,44(02):73-77.