“智慧政务”中的文本挖掘应用

摘要

随着社会的发展进步，近年来，人民生活水平逐渐提高，但是还是不可避免地存在一些问题，尤其是我国还是一个入口大国，解决人与人之间的问题是我国不得不面临的一个重大问题。这些问题不可避免地繁杂，在计算机技术如此发达的今天，利用计算机来处理居民反应的情况是非常有利的。

对于问题 1，第一个问题是要求我们对居民提供的数据进行一级标签分类，一级标签分类，并使用 F－Score 对分类方法进行评价。由于居民反应的数据情况内容特别多，在本次中我们需要对这些文字数据进行片段化处理，在本问中，采用提取关键词的方法，将关键词与类别的联系紧密的归为一类。

对于问题 2，本问是进行热点挖掘类问题，居民反应的各种问题，其中不可避免的会重复同一种类型的问题，而热点评价中，出现次数也是可以当做热点评价的一种评价指标。但在此热点问题之中并没有用明确的数字表示关注程度，因此这里对排名的处理就需要进行一下间接性的转换。

对于问题 3，本问是一个开放性问题，是针对答复质量的分析，由于问题回答的数据较大，对前面的反映问题的归类如果没有准确地归类，那么在答复中就会出现错误，问题 3 也是对前面模型的一种检验和评价。

关键词：文本分词　jieba 工具　热度评价

Application of Text Mining in "Intelligent Government"

Abstract

With the development and progress of society, the living standard of the people has gradually improved in recent years, but there are still some problems that are unavoidable. In particular, China is still a big import country, solving the problems between people is a major problem that our country has to face. These problems are inevitably complex, in the computer technology so developed today, the use of computers to deal with the reaction of the residents is very beneficial.

For Question 1, the first question asks us to classify the data provided by the residents by a first-level label, a first-level label, and to evaluate the classification method using f-Score. Because the residents respond to the data content is very much, in this we need to carry on the fragment processing to these characters data, in this question, uses the method of extracting the key word, the key word and the category connection close grouping.

For Question 2, this question is a hot spot mining type of question, residents respond to various questions, which will inevitably repeat the same type of question, and hot spot evaluation, frequency of occurrence can also be used as a hot evaluation of the Evaluation Index. But there are no clear numbers for how much you care about this hot topic, so dealing with rankings here requires an indirect shift.

As for Question 3, this question is an open-ended question and an analysis of the quality of the responses. Since the data on the responses to the questions are relatively large, the categorization of the previous response questions, if not accurately categorized, will result in errors in the responses, question 3 is also a test and evaluation of the previous model.

Keywords: Text segmentation JIEBA tools Thermal evaluation

一、挖掘目标

在本次建模中，利用系统平台发布的数据，使用 jieba 中文分词工具对居民反应的热点问题进行了分词，提取重要词汇，且在此过程中出去垃圾信息，使用聚类算法，并在 F－Score 方法下对分类进行评价，目的是达到以下目标：

1 利用分词工具对反应问题进行处理，对这些词汇进行归类，并使用评价标准进行评价。

2 对居民反应的问题进行热点处理。

3 对答复的评价。

二、分析方法与过程

1、对问题 1 的处理

1.1 数据预处理

题目中给出的数据，量大且内容多，不可能一字一句地去读出来，因此提取关键词就成为了必须要进行的一个步骤。

1.2 对反应的问题进行中文分词

在对问题分类之前，需要对数据进行非结构的文本信息处理，转换为计算机可以识别的语言，这里使用的工具是 jieba 进行分词。之后就是对这些分词进行分类，一级标签是一个范围大的分类标准，其中肯定存在与之相关的常见词汇，如交通问题中不可避免会出现“交通”等关系程度大的词汇。因此比较分词与类别的热门词汇就为我们对这些问题进行分类提供了一种方法。

1.3 TF-IDF 算法

在对居民反映的问题进行分词之后，需要将这些分词转换为向量，以便于分析使用。在本问题中我们使用的是 TF-IDF 算法。

在此算法中 TF 表示的是权重，即词汇在本反应问题中出现的次数，也可以称之为词频。

TF＝某个词的个数/文本处理后的总词数。

IDF 指的是逆文档频率，越大则表明该词与该文本的关系程度越大，也就是代表性越大

TF-IDF＝TF×IDF

TF-IDF 表示的是权重向量，此法得出的值即代表本词的重要性。

1.4 反应问题的归类

从上面的已经处理的词汇向量中，选取三个 TF-IDF 值最大的，将这些词与将要分类问题的词进行比较

在这里须再引入一种表示关系，假设为 FL

FL＝TF-IDF/预分类类别的关键词汇。

通过比较 FL 值，选出最大的值所对应的类别即可将此条文本归于相对应的类别。

2、对问题 2 的处理

在所给的信息中存着一个明显的数据，即赞成和反对数，这两个数可以反应该问题的受关注程度，在这里可以采用 Excel 中的 sun 公式计算出和，通过排序

得出热度大小。按照要求做出表格即可

结果：

表 1 热点问题表

热度排名	热度指数	时间范围	地点/人群	问题描述
1	2097	2019/8/19 11:34:04	A市A5区的业主	群租房的安全问题
2	1767	2019/4/11 21:02:44	梅溪湖金毛湾的一名业主	配套入学的问题
3	821	2019/2/21 18:45:14	A4区p2p公司	车贷案
4	790	2019/2/25 9:58:37	西地省展星投资有限公司	车贷特大集资诈骗案
5	733	2019/3/1 22:12:30	A市A4区	贷案警官应跟进关注留言

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	反对数	点赞数
2	208636	A00077171	汇金路五矿万境K9县存在一系	2019/8/19 11:34:04	我是A市A5区汇金路五矿万境K9县24栋的一名业主，我	0	2097
1	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	书记先生：您好！我是梅溪湖金毛湾	5	1762
3	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	尊敬的胡书记：您好!A4区p2p公司58车贷，非法经营近	0	821
3	217032	A00056543	A市58车贷特大集资诈骗案保	2019/2/25 9:58:37	胡市长：您好！西地省展星投资有限公司设立58车贷	0	790
3	194343	A000106161	A市58车贷案警官应跟进关注	2019/3/1 22:12:30	胡书记：您好！58车贷案发，引发受害人举报投诉，也	0	733

3、对问题 3 的处理

问题 3 是一个开放性问题，针对答复的相关性，即回答问题的符合程度；完整性，即答复的全面程度；可解释性，即答复的规范程度，给出评价方案。

该问题中又需要对答复进行分词处理，再结合现实中的情况得出一般答复的重要词汇，继续进行比较，以此来得出对答复意见的评价。