“智慧政务”中的文本挖掘应用

摘要

随着科技的不断进步，多种多样的交流方式，也在我们的生活中普遍，为了使政府更能了解人民群众的思想，提升人民群众幸福感，我们将对人民的意愿一一了解实现。近年来，群众都喜欢在微信，微博以及市长信箱中留言，以致导致很多大数据。随着留言的增多，我们需要靠人工来进行对这些留言的划分和整理，但是由于能够使问题能够更快得到解决，我们需要帮助相关部门将对这些工作进行简化，而对于我们来说，我们所做的就是希望靠人工把这些留言根据轻重缓急的重要性，分好类别群众反映的问题，能够更快的解决。

对于问题一也就是建立关于留言内容的一级标签分类模型，我们用题中所给出的分类方法进行评价，也就是说对于题中所给的数据，我们可以先提炼出关键词来，把它分别填到三级标签对应的体系里面，然后再根据我们所给的公式，来找出 Pi 和 Ri 评价我们的分类方法。

对于问题二，首先根据附件 3 提供的数据对人群的留言进行分类，做出合理的热度评价指标，通过利用 python 软件对文本预处理和 excel 表格对某一时间段、人群关注的热点问题、热点问题的留言和点赞数进行分类筛选，选出前五的热点问题以及热点问题的留言。

对于问题三的留言答复意见，我们可以根据他留言的时间来选择我们对留言意见回复的先后顺序，而根据他相应的分类标准，我们可以制定相关的统一回复的语言来减少我们工作量，还可以邀请他们对我们的回复进行评价，让我们从中吸取到更多的优点和找出缺点，步步改进，步步优化。

关键词：热度评价指标　　文本预处理

一．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，以往主要依靠人工来进行留言划分和热点整理的相关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。因此，利用自然语言处理和文本挖掘的方法解决相关问题具有重大意义。

二．模型假设

1. 假设题目给出的数据是准确的，没有存在个人主观因素太偏激的想法。
2. 人工处理数据时网络良好，所有数据均无差错。
3. 假设同一时间段内热点问题的特定地点是随机抽取的，保证某件事能反应社会关注的问题。

四．问题分析

问题一也就是建立关于留言内容的一级标签分类模型，我们用题中所给出的分类方法进行评价，也就是说对于题中所给的数据，我们可以先提炼出关键词来，把它分别填到三级标签对应的体系里面，然后再根据我们所给的公式，来找出 Pi 和 Ri 评价我们的分类方法。

问题二的分析：大数据发展的时代，对信息的处理我们能更快、更多地进行分析、整理，筛选出对我们有用的信息。在第二问中，我们利用 python 软件及数据挖掘算法对文本挖掘处理，筛选出排名前五的热点问题，以及相应热点问题对应的留言信息。基于数据挖掘技术与文本挖掘的差异性，应用数据挖掘之前我们先对文本数据进行预处理，实现附件 3 中对热点问题的词频统计分析。根据题目的要求，我们选择某一时段内特定人群的留言进行归类，从人民生活中的相关问题，城市建设，社会经济发展区，环境污染等四个方面对人群关心的热点问题进行评价。

问题三的留言答复意见，我们可以根据他留言的时间来选择我们对留言意见回复的先后顺序，而根据他相应的分类标准，我们可以制定相关的统一回复的语言来减少我们工作量，还可以邀请他们对我们的回复进行评价，让我们从中吸取到更多的优点和找出缺点，步步改进，步步优化。

五．模型的建立与求解

5.1 第一问的模型建立与求解

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

5.2 第二问的模型建立与求解

5.2.1 模型建立

利用附件 3 中群留言的特点，我们主要从人民生活中的相关问题，城市建设，社会经济发展区，环境污染等四个方面对文本进行归类，文本预处理过程是在文本中提取关键词表示文本的过程，通过词语或关键字在文档中出现次数的统计可分析出人群关心的热点问题倾向，对我们所要筛选出的问题有很大帮助。

运用文本分类特征提取方法对文本信息进行提取，针对本题我们看到人群留言信息中有很多热点问题是相似的，比如车贷诈骗问题，因此从我们观察的量中可以选择词频来作为我们的分类途径。建立模型，特征提取步骤如下：

用卡方检验的方法：

X 表示某个词，Y 表示某一类别，X_i 表示这个词的取值，而在本题中，假设只有两种情况，出现和不出现，Y_j 表示某一类别，类别有多样，支取其一就好，那么我们得知X_i 和 Y_j 在整个数据中出现的概率，X_i 在整个数据中出现的概率和 Y_j 在整个数据中出现的概率分别为：

$$P(X = x_i, Y = y_j)$$

$$P(X = x_i)$$

$$P(Y = y_j)$$

利用公式评价一个事件对另一个事件出现的信息为：

$$M(Y, X) = \sum_{y_j \in Y} \sum_{x_i \in X} P(X = x_i, Y = y_j) \log_2 \frac{P(X = x_i, Y = y_j)}{P(X = x_i) P(Y = y_j)}$$

经计算和分析，利用 python 软件和 excel，我们得出以下数据：

表 1-热点问题表

热度排名	问题 ID	热度指数	时间范围	人群	问题描述
1	A00077171	2097	2019/8/19 11:34	A 市 A5 区汇金路五矿万境 K9 县小区业主	群租房的出现给别墅小区带来的困扰

2	A00087522	1762	2019/4/11 21:02:44	梅溪湖金毛湾的业主	A 市教育局未将金毛湾楼盘纳入配套入学的问题
3	A00031682	821	2019/2/21 18:45:14	A4 区 p2p 公司 58 车贷受害人	A4 区 p2p 公司 58 车贷非法经，营损害顾客权利
4	A00056543	790	2019/2/25 9:58:37	严惩 A 市 58 车贷特大集资诈骗案保护伞	58 车贷爆雷，公司董事和负责人逃离，
5	A000106161	733	2019/3/1 22:12:30	承办 A 市 58 车贷案警官应跟进关注留言	A 市 A4 区经侦没有对 58 车贷案发的投诉情况跟进市领导的留言

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	点赞数	反对数
1	208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019/8/19 11:34	我是 A 市 A5 区汇金路五矿万境 K9 县 24 栋的一名业主，我们小区一开始的定位是一个高端别墅小区，……	2097	0
2	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！我是梅溪湖金毛湾的一名业主，和其他业主一样因为当初金毛湾的承若学校都是金毛建的，……	1762	5
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，非法经营近四年。……	821	0
4	217032	A00056543	严惩 A 市 58 车贷特大集资诈骗案保护伞	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 https://baidu.com/ 亿。……	790	0
5	194343	A000106161	承办 A 市 58 车贷案警官应跟进关注留言	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，也引起市领导的重视，公布了受害人的留言，……	733	0

从上面两个表格数据可知，人民最近的热点问题大多数是关于车贷，居住地得不到环境保证，以及关心孩子上学等问题。

八．参考文献

[1]曹鲁慧,邓玉香,陈通,李钊.一种基于深度学习的中文文本特征提取与分类方法[J].山东科学,2019,32(06):106-111.