智慧政务中的文本挖掘

摘要

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。

对第一题的数据，我们采用 TF-IDF 的方法去筛选每一类标签对应的关键词，得知该类问题的主要信息并将其作为筛选的标准，对剩余问题的关键词筛选，将筛选出不必要的关键词，只留下有筛选意义的关键词，将这些关键词与反应的问题文本进行匹配，本着能可抓错不放过一个的原则将所有问题进行分类，再根据题目给出的公式计算对应的值。对第二题，查看数据后，我们发现有几个市民反应的问题获得了较其他问题多出很多的赞数，因为赞数悬殊，所以我们默认最热点的五个问题为赞数最高的前五类问题（不是前五个，是前五类，因为前五个问题中有几类问题重复了多次），对这前五类问题提取关键词，对全部反应的问题进行关键词匹配，将匹配上的问题归类为其中的某一类，从而得到表 2。

一、 模型建立

第一题，采用 TF-IDF 的方法去筛选关键词：

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

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

可以得到一个全新的参数 TF-IDF

$$TF - IDF = TF * IDF$$

并以此为参考权重去判断一个词语在这个内容中占比的大小，对每一一级标题均采取这个举措，就可以得到各个一级标题的关键词，但这并不是最终关键词，我们先将 TF-IDF 得到的关键词进行筛选后，得到最终的关键词，然后将含有关键词的意见反馈收集起来就将其分类完毕。

第二题，我们发现赞数最高的前五类问题的赞数明显较其他问题悬殊，于是我们认为这五类问题即为最热的问题。

同样采用 TF-IDF 的算法去适配关键词，但我们发现关键词并不能适配到很好的意见反馈，于是我们将这五类问题的地点提取，筛选地点的关键词，然后对此进行适配得到了非常良好的结果。从而去筛选出了热点问题表。

二、 结果展示

在 TF-IDF 算法下，我们对一类标题，各自筛选出了 50-100 个关键词（各类标题不等，且相关意见越多，关键词越少）（仅展示部分）。

	words
1	报销
2	补交
3	补缴
4	参保
5	产假
6	城乡居民
7	城镇职工
8	低保
9	电力局
10	改制
11	岗位
12	工龄
13	工勤
14	工伤
15	工伤保险
16	公积金
17	供销社

	words
1	保健品
2	产品
3	超市
4	车位
5	充绒
6	传销
7	春运
8	导游
9	地产
10	儿童票
11	狗
12	广告
13	混凝土
14	活禽
15	价格
16	检疫
17	江湖

	words
1	办学
2	笔试
3	毕业证
4	补课
5	补课费
6	补贴
7	补助
8	代课
9	代课老师
10	档案
11	放假
12	高中
13	公办
14	寒假
15	机构
16	家长
17	监考

我们对此进行筛选时发现，关键词匹配出的结果并不能很好的适配原结果，查阅资料发现，查全率和查准率在常规情况下，似乎并没有非常好的结果，于是我们继续了模型的实现。

对 7 类一级标题计算查全率和查准率得到

P_1=0.583　P_2=0.441　P_3=0.233　P_4=0.427　P_5=0.490

P_6=0.389　P_7=0.368

R_1=0.778　R_2=0.794　R_3=0.863　R_4=0.926　R_5=0.877　R_6=0.756　R_7=0.946

其中 P_i 为查准率，R_i 为查全率，从而可以计算出 F-score

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

得到 **F=0.551**

对第二题，由于我们默认赞数前五为热点问题，所以可以简单得到热点问题表为

问题ID	热度指数	时间范围	地点/人群	留言详情
1	2097	2019/1/15	A市A5区汇金路五矿万境	A市A5区汇金路五矿万境K9县存在一系列问题
2	1767	########	A市金毛湾	反映A市金毛湾配套入学的问题
3	821	2019/1/8 3	A市A4区广大人民	A市A4区58车贷诈骗案
4	669	2019/1/30	A4区绿地海外滩小区	A4区绿地海外滩小区距长赣高铁太近
5	242	2019/4/2 3	A市富绿物业丽发新城	A市富绿物业丽发新城业主大规模反映物业问题

对这五个问题进行关键词的挑选，我们发现若按上式的 TF-IDF 法筛选关键词的话，我们并不能得到准确的关于此类问题的反馈意见，并且会得到许多没有意义的意见反馈，于是我们采取了另一种关键词筛选方式，我们发现这些问题类型总是关

于某一个地点发生的，于是我们提取这五类问题的地点文本，对他们进行了关键词的筛选，然后去匹配附件 3 中的含有此类地点的反馈意见，得到了非常良好的效果，于是我们便采用了这类方式。然后得到了留言明细表（仅展示部分数据）。

		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	1	283732	A0002149	A市五矿万	2019/1/15 10:29	我们是A市五矿万境水岸三	0	0
3	1	275491	A0006133	A市五矿万	2019/9/10 9:10	关于五矿万境K9县负一楼面	0	0
4	1	262599	A0001004	A市五矿万	2019/9/19 17:14	我是西地省A市五	0	0
5	1	252650	A0001053	A市五矿万	2019/9/11 15:16	尊敬的相关部门，本人家庭于2018年	0	0
6	1	212349	A0001007	请A市教育	2019/5/24 16:32	尊敬的A市教育局领导：我是一名幼升小孩子的家长	1	0
7	1	215507	A0001032	A市五矿万	2019/9/12 14:48	预交房23栋没有通往负一楼	0	1
	1	271488	A0002356	A5区教育局	2019/6/4 17:04	目前A5区砂子塘万境水岸小学有3612	0	2

最终我们得到了一份详细准确的明细表。

三、 模型优缺点

优点：1、第一题我们得到了较好的查全率和查准率

2、第一题我们对关键词进行了二度筛选，筛选掉了类似“你”，“我”，“他”此类无实际意义但 TF-IDF 权重占比很高的词汇

3、第二题我们采用了地点作为关键词进行筛选，得到了良好的结果

缺点：1、第一题我们并没有做一些其他的分析去协助关键词筛选。

2、第二题我们没有去筛选热点问题而是直接决定了热点问题

3、我们并没有办法解决第三题

四、 参考文献

1. https://blog.csdn.net/kMD8d5R/article/details/89629284 -CSDN

2. https://mp.weixin.qq.com/s?__biz=MzA3MTM3NTA5Ng==&mid=2651060726&idx=2&sn=45960ca6236b94c1f5a83f74597e789e&chksm=84d9d861b3ae51771a5a826a76f323a930fba218aa68c1f75d5b1e5c1169157353defd59ec21&scene=21#wechat_redirect 黄天元