“智慧政务”中的文本挖掘应用

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。

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that used to rely mainly on human to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of

smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government.

目录

1. 挖掘目标

本次建模的目标是针对“智慧政务”中的留言文本评论数据，在对文本进行基本的机器预处理、消息的相关性以及用户的点赞数反对数的统计，通过建立多种数据挖掘模型，实现对文本数据的深度分析挖掘，并形成一套回复信息的留言体系，以期望得到有价值的内在内容。

2. 问题一的分析方法与过程

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（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。使用 F-Score 对分类

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

方法进行评价：

， 其中 Pi 为第 i 类的查准率， Ri 为第 i 类的查全率。

F 方法

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，机器学习中分类模型的精确率(Precision)和召回率(Recall)评估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。

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。

通过对数据预处理，建立关于一级分类标签模型，通过对多个数据标签的分析以及 F-score 算法的运用，从而能够实现对“智慧政务”中复杂留言的分类，使工作大大简化，提高行政效率。

3. 问题一的结果

	原来分类	模型分类输出	比较\n",
"0	5	5	True\n",
"1	5	5	True\n",
"2	5	5	True\n",
"3	5	5	True\n",
"4	5	5	True\n",
"..	...	...	...\n",
"490	2	2	True\n",
"491	2	2	True\n",
"492	2	2	True\n",
"493	2	2	True\n",
"494	2	2	True\n",

建立关于一级分类数据标签的模型，通过对多个标签进行分类分析从而快速的实现对数据的预处理并挖掘数据的价值。

4. 问题二的分析方法与过程

MD5 算法原理

1、数据填充对消息进行数据填充，使消息的长度对 512 取模得 448，设消息长度为 X，即满足 X mod 512=448。根据此公式得出需要填充的数据长度。填充方法：在消息后面进行填充，填充第一位为 1，其余为 0。

2、添加消息长度在第一步结果之后再填充上原消息的长度，可用来进行的存储长度为 64 位。如果消息长度大于 264，则只使用其低 64 位的值，即（消息长度 对 264 取模）。在此步骤进行完毕后，最终消息长度就是 512 的整数倍。

3、数据处理准备需要用到的数据：4 个常数： A = 0x67452301, B = 0x0EFCDAB89, C = 0x98BADCFE, D = 0x10325476;

4 个函数：F(X,Y,Z)=(X & Y) | ((~X) & Z); G(X,Y,Z)=(X & Z) | (Y & (~Z)); H(X,Y,Z)=X ^ Y ^ Z; I(X,Y,Z)=Y ^ (X | (~Z));把消息分以 512 位为一分组进行处理，每一个分组进行 4 轮变换，以上面所说 4 个常数为起始变量进行计算，重新输出 4 个变量，以这 4 个变量再进行下一分组的运算，如果已经是最后一个分组，则这 4 个变量为最后的结果，即 MD5 值。

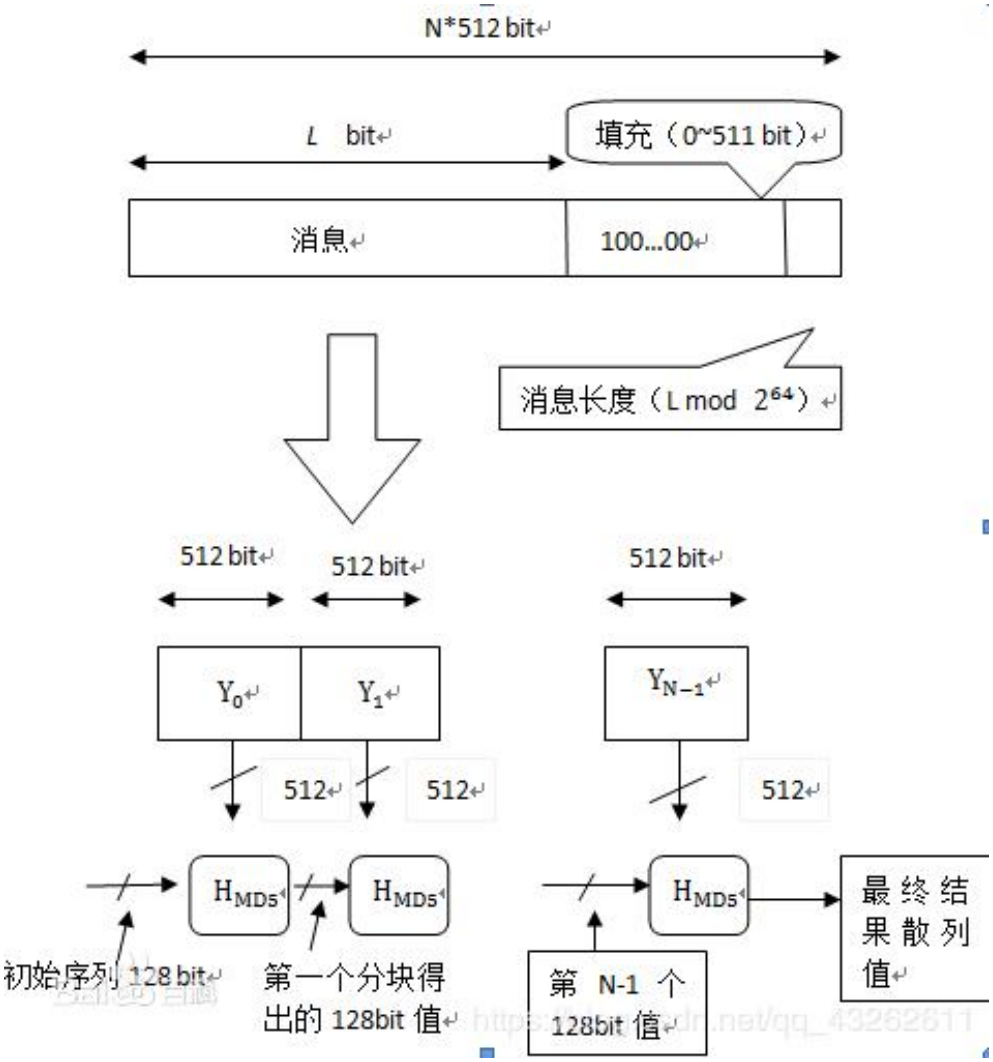
MD5 算法的基本流程

附加填充位

初始化链接变量

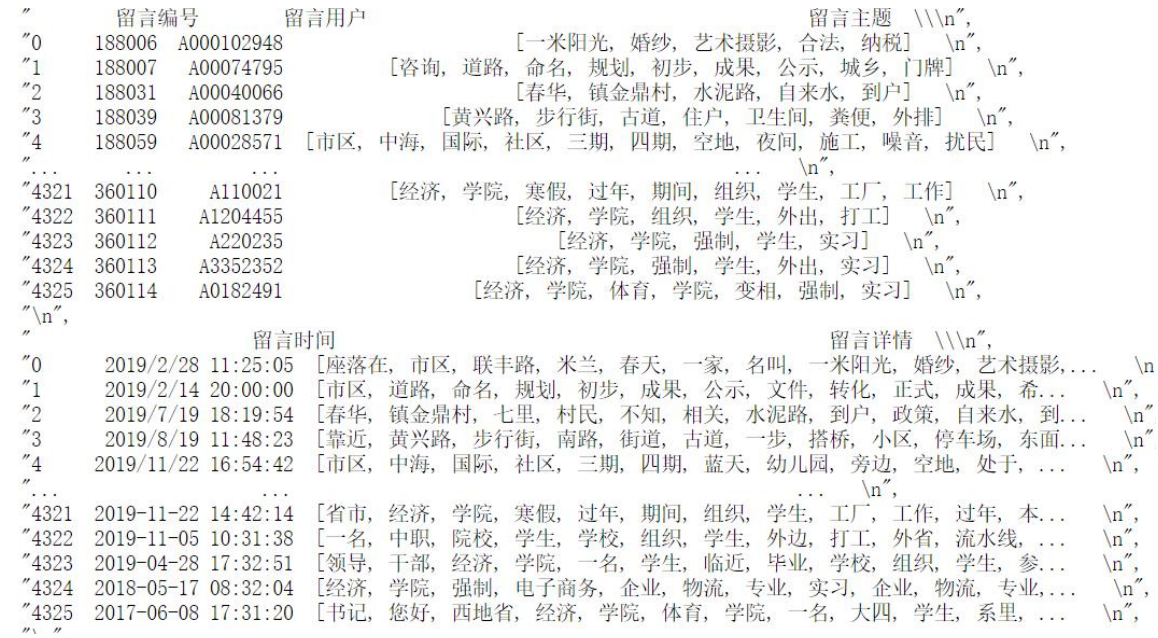
分组处理

步函数



运用 MD5 算法能够在问题一的基础之上实现对留言更为深层次的挖掘与分析，不仅仅在根据文本数据进行预处理之后，提炼出用户关注度较高的问题，根据其点赞数、反对数以及留言时间等更多的数据进行分析，从而能够分析出热点问题，为下一步信息回复做好准备，也大大提高了行政效率，更为快捷的进行回复。

5. 问题二的结果



	留言编号	留言用户	留言主题
0	188006	A000102948	[一米阳光，婚纱，艺术摄影，合法，纳税]
1	188007	A00074795	[咨询，道路，命名，规划，初步，成果，公示，城乡，门牌]
2	188031	A00040066	[春华，镇金鼎村，水泥路，自来水，到户]
3	188039	A00081379	[黄兴路，步行街，古道，住户，卫生间，粪便，外排]
4	188059	A00028571	[市区，中海，国际，社区，三期，四期，空地，夜间，施工，噪音，扰民]
...	...	...	...
4321	360110	A110021	[经济，学院，寒假，过年，期间，组织，学生，工厂，工作]
4322	360111	A1204455	[经济，学院，组织，学生，外出，打工]
4323	360112	A220235	[经济，学院，强制，学生，实习]
4324	360113	A3352352	[经济，学院，强制，学生，外出，实习]
4325	360114	A0182491	[经济，学院，体育，学院，变相，强制，实习]

	留言时间	留言详情
0	2019/2/28 11:25:05	[座落在，市区，联丰路，米兰，春天，一家，名叫，一米阳光，婚纱，艺术摄影，...
1	2019/2/14 20:00:00	[市区，道路，命名，规划，初步，成果，公示，文件，转化，正式，成果，希...
2	2019/7/19 18:19:54	[春华，镇金鼎村，七里，村民，不知，相关，水泥路，到户，政策，自来水，到...
3	2019/8/19 11:48:23	[靠近，黄兴路，步行街，南路，街道，古道，一步，搭桥，小区，停车场，东面...
4	2019/11/22 16:54:42	[市区，中海，国际，社区，三期，四期，蓝天，幼儿园，旁边，空地，处于，...
...	...	...
4321	2019-11-22 14:42:14	[省市，经济，学院，寒假，过年，期间，组织，学生，工厂，工作，过年，本...
4322	2019-11-05 10:31:38	[一名，中职，院校，学生，学校，组织，学生，外边，打工，外省，流水线，...
4323	2019-04-28 17:32:51	[领导，干部，经济，学院，一名，学生，临近，毕业，学校，组织，学生，参...
4324	2018-05-17 08:32:04	[经济，学院，强制，电子商务，企业，物流，专业，实习，企业，物流，专业，...
4325	2017-06-08 17:31:20	[书记，您好，西地省，经济，学院，体育，学院，一名，大四，学生，系里，...

	反对数	点赞数
0	0	0
1	0	1
2	0	1
3	0	1
4	0	0
...	...	...
4321	0	0
4322	1	0
4323	0	0
4324	3	0
4325	9	0

6. 问题三的分析方法与过程

BM25 算法，通常用来作搜索相关性平分。一句话概况其主要思想：对 Query 进行语素解析，生成语素 qi；然后，对于每个搜索结果 D，计算每个语素 qi 与 D 的相关性得分，最后，将 qi 相对于 D 的相关性得分进行加权求和，从而得到 Query 与 D 的相关性得分。

BM25 算法的一般性公式如下：

$$Score(Q,d)=\sum_{i}^{n}W_i\cdot R(q_i,d)$$

其中，Q 表示 Query，qi 表示 Q 解析之后的一个语素（对中文而言，我们可以把对 Query 的分词作为语素分析，每个词看成语素 qi。）；d 表示一个搜索结果文档；Wi 表示语素 qi 的权重；R(qi，d)表示语素 qi 与文档 d 的相关性得分。

下面我们来看如何定义 Wi。判断一个词与一个文档的相关性的权重，方法有多种，较常用的是 IDF。这里以 IDF 为例，公式如下：$$\mathrm{IDF}(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$$，其中，N 为索引中的全部文档数，n(qi)为包含了 qi 的文档数。

根据 IDF 的定义可以看出，对于给定的文档集合，包含了 qi 的文档数越多，qi 的权重则越低。也就是说，当很多文档都包含了 qi 时，qi 的区分度就不高，因此使用 qi 来判断相关性时的重要度就较低。

我们再来看语素 qi 与文档 d 的相关性得分 R（qi，d）。首先来看 BM25 中相关性得分的一般形式：

$$R(q_i,d)=\frac{f_i\cdot(k_1+1)}{f_i+K}\cdot\frac{qf_i\cdot(k_2+1)}{qf_i+k_2}$$

$$K=k_1\cdot(1-b+b\cdot\frac{dl}{avgdl})$$

其中，k1，k2，b 为调节因子，通常根据经验设置，一般 k1=2，b=0.75；fi 为 qi 在 d 中的出现频率，qfi 为 qi 在 Query 中的出现频率。dl 为文档 d 的长度，avgdl 为所有文档的平均长度。由于绝大部分情况下，qi 在 Query 中只会出现一次，即 qfi=1，因此公式可以简化为：

$$R(q_i,d)=\frac{f_i\cdot(k_1+1)}{f_i+K}$$

从 K 的定义中可以看到，参数 b 的作用是调整文档长度对相关性影响的大小。b 越大，文档长度的对相关性得分的影响越大，反之越小。而文档的相对长度越长，K 值将越大，则相关性得分会越小。这可以理解为，当文档较长时，包含 qi 的机会越大，因此，同等 fi 的情况下，长文档与 qi 的相关性应该比短文档与 qi 的相关性弱。

综上，BM25 算法的相关性得分公式可总结为：

$$Score(Q,d)=\sum_{i}^{n}IDF(q_i)\cdot\frac{f_i\cdot(k_1+1)}{f_i+k_1\cdot(1-b+b\cdot\frac{dl}{avgdl})}$$

从 BM25 的公式可以看到，通过使用不同的语素分析方法、语素权重判定方法，以及语素与文档的相关性判定方法，我们可以衍生出不同的搜索相关性得分计算方法，这就为我们设计算法提供了较大的灵活性。

运用 BM25 算法，通过使用不同的语素分析方法、语素权重判定方法，以及语素与文档的相关性判定方法，寻找到留言信息之间的相关性，形成一套对留言数据分析回复的成熟体系，进而能够帮助用户更为便捷的回复。

7. 问题三的结果

留言编号	留言用户	留言主题	留言时间	\\\n″,
2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n″,
2554	A00023583	A3区潇楚南路洋湖段怎么还没修好？	2019/4/24 16:03:40	\n″,
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n″,
2557	A000110735	在A市买公寓能享受人才新政购房补贴吗？	2019/4/24 15:07:30	\n″,
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n″,
...	...	...	...	\n″,
181267	UU008766	汽车北站进站口附近居民强烈反对建设I市平康肾病医院！	2018/12/12 15:20:46	\n″,
181603	UU008194	强烈反对I市9路公交车改线路	2018/6/12 8:51:03	\n″,
184423	UU0082115	对G7县文盛小学特色班的一点质疑	2018/10/11 20:02:52	\n″,
185799	UU008785	燃油税费改革政策的咨询	2012/9/4 23:14:44	\n″,
185986	UU008363	强烈呼吁宁朱公路拓宽提质改造	2011/10/3 21:52:37	\n″,

留言详情 \\\n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t2019年4月以来，位于A市A2区桂花坪街道... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t潇楚南路从2018年开始修，到现在都快一年了... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t地处省会A市民营幼儿园众多，小孩是祖国的未来... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t尊敬的书记：您好！我研究生毕业后根据人才新政... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t建议将“白竹坡路口”更名为“马坡岭小学”，原... \n″,
... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t我们是I市汽车北站进站口的周围居民。在这里的... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t强烈反对I市9路公交车改线路获悉从7月1日起... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\tG7县文盛小学引入特色班，每个学生必须参加，... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t贺厅长： 您 好！自燃油税费改革... \n″,
\\n\\t\\t\\t\\t\\t\\n\\t\\t\\t\\t\\t A8县朱良桥乡可以说是A8县最破烂的乡了... \n″,

答复意见	答复时间 \n″,
现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53 \n″,
网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10 \n″,
市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14 \n″,
网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42 \n″,
网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30 \n″,
... ...	\n″,
您的留言已收悉。关于您反映的问题，已转I1区委、区人民政府调查处理。	2019/1/8 16:54:53 \n″,
“UU008194”您的留言已收悉。关于您反映的问题，已转市交通运输局调查处理。	2018/7/4 16:55:53 \n″,
“UU0082115”您好！获悉关于“对G7县文盛小学特色班的质疑”的网帖后，我局领导高度重...	2018/10/24 9:22:07 \n″,
西地省平台《问政西地省》栏目组： 网民在贵栏目留言，咨询中央转移支付我省燃油税资金情况，...	2013/1/6 15:41:02 \n″,
“UU008363”： 您好！您的留言，我厅领导高度重视，要求相关部门进行了认真的调查和...	2012/2/28 10:19:55 \n″,

8. 结论

本文通过对“智慧政务”中的各类文本利用 F 算法、MD5 算法、BM25 算法等算法进行挖掘分析，得到了一定价值的结果，实现了对“智慧政务”中文本留言的分析以及一定程度上的对包括用户点赞数、反对数还有留言时间，答复意见、答复时间等更为细节的文本信息的挖掘与认识。而随着微信、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。

但是从我们的分析结果来看总体的效果还有待进步，这也会是我们在后来的进一步的对中文文本数据的研究过程中可以继续深入探讨的地方。