第八届“泰迪杯”
全国数据挖掘挑战赛

1.挖掘目标

本次建模目标是利用网络信息平台系统发布的群众问政留言数据，利用python，R和excel等工具对留言数据进行不同类别的分类、jieba分词的方法及朴素贝叶斯算法，达到以下三个目标：

1） 利用文本分词和文本聚类的方法对非结构化的数据进行文本挖掘，根据聚类结果，结合群众留言的内容和性质进行分类；并利用F-score对分类方法进行评价。

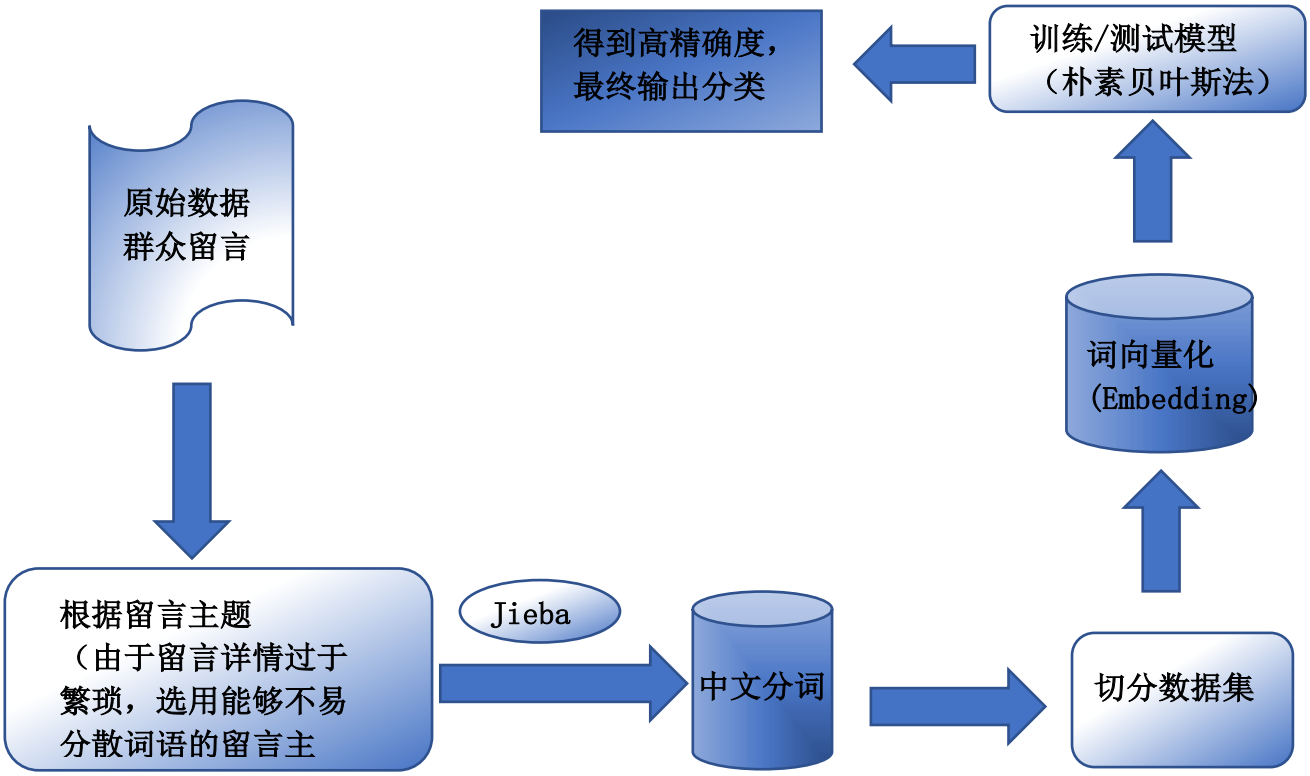
2） 根据群众留言数据，定义热度评价指标，挖掘出一段时间，集中爆发，多人反映的热点问题，在进行留言分类。

3） 根据政府对留言的答复情况，对答复从多方面角度评价，给出评价方案。

1.问题分析方法与过程

2.1 问题1分析方法与过程

2.1.1流程图



2.1.2数据预处理和jieba中文分词

在题目给出的数据中，出现了部分重复的群众留言数据。例如留言编号303与留言编号319只字不差地描述了一种现象，表达了重复的留言信息。考虑到政务人员在审核群众留言信息时，要查看最新的留言，即每天都会对留言信息进行更新，因此在去重的时候应该取更新时间最晚的记录，去掉历史记录。
在对群众留言信息进行挖掘分析之前，必须先把非结构化的文本信息转换为计算机能够识别的结构化信息。在原始附件2中，以中文文本的方式给出了留言数据，包括时间，留言主题和留言详情，为了便于转换，先要对这些留言信息进行中文分词。我们选用jieba 分词是因为其能够基于前缀词典来高效快速的进行词图扫描，并且生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)，同时采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的HMM 模型，使得能更好的实现中文分词效果。此外，我们引入了pandas，numpy和jieba数据包。考虑到留言详情字数过多，过于繁琐，若是使用留言详情，则会出现分词不精确并且繁杂的现象，扰乱分类，所以我们将留言主题的数据保存在2.csv文件中，并且进行jieba空格分词。

2.1.3划分数据集

得到jieba分词后，我们首先进行数据切分。分类问题需要x（特征），和y（label），这里分词后x=留言主题，y=分类评级，随后按8:2的比例切分为训练集和测试集。并且为下一步词向量化做铺垫。

2.1.4词向量化

词向量将我们的整个留言主题所出现的词语一一排列，然后每行数据去映射到这些列上，出现的就是 1，没出现就是 0，这样，文本数据就转换成了 01 稀疏矩阵。我们选用 sklearn 的方法进行词向量化，其中我们选用CountVectorizer 用到的函数有如下：
max_df：在超过这一比例的文档中出现的关键词（过于平凡）eg.投诉，举报，请问等词语，我们将此类词语去除掉。
min_df：在低于这一数量的文档中出现的关键词（过于独特）eg.有限公司等，我们将此类词语去除掉。
token_pattern：主要是通过正则处理掉数字和标点符号。
stop_words：设置停用词表，这样的词我们就不会统计出来（多半是虚拟词，冠词等等），需要列表结构，所以代码中定义了一个函数来处理停用词表。停用词表经过我们专门化设置，我们想去除专用名词例如：小区，希望等词语。

2.1.5测试和训练数据集

我们按 8:2 的比例切分为训练集和测试集。通过训练集和机器学习，形成训练模型，我们选用 sklearn 来实现朴素贝叶斯模型进行机器训练。
步骤如下：

第一步：朴素贝叶斯是一种概率分类器，这意味着对于文档 d，在所有类

别　中，分类器返回具有给定文档的最大后验概率的类别

$$c = argmax P(c|d) \quad 1.1$$

$c \epsilon C$

第二步：贝叶斯分类的直觉是利用贝叶斯规则将公式 1.1 转化为其他具有一些有用性质的概率，贝叶斯规则如式 1.2 所示。

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad 1.2$$

第三步：将 1.2 代入 1.1，得到

$$c = argmax P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad 1.3$$

训练集得到的精确度结果为 0.9089。

最后，我们进行数据测试来验证精确度，得到 0.8116，得到较高的精确度。

2.2 问题 2 分析方法与过程

2.2.1 数据筛选和问题识别

由于题目要求找出某一时间段内反映特定地点或人群的 5 大热点问题，所以我们首先进行时段排序。由于原始附件 3 的数据主要集中在 2020 年初和 2019 年，所以我们根据留言主题的时间段，将所有的留言主题拆分至 4 个不同的时间段，每个时间段间隔约为 2 个月，并将数据储存在 1.txt；2.txt；3.txt；4.txt 中。

2.2.2 问题归类和定义热度

我们利用R来进行jieba分词，我们选用采用混合形式分词（mixseg），并且利用worker函数进行构架分词器和dict(DICTPATH)函数构建系统词典，其中我们选用了stop_word和user(USERPATH)两个参数，分别是停用字词典和用户词典，并且将停用字典输出，保存为stopwords.txt文件，包括“一直”“一起”“不仅”等词语。接着，我们使用result函数进行分词。此外，我们为了让选出的词语更为精确，我们去除字符长度小于2的词语，例如，的，请等词语。最后，我们使用reslut函数让词语按照频数排序，并将频数结果写入excel文件，包括result1，result2，result3，result4四个文件，显示频数最高的前100个词语，并且加载词云包，画出词云图，以更直观的形式展示热点词语。

我们对热度进行定义，其中点赞数和反对数的总和占30%，同一时间段出现的次数占70%，设置第一名为100分。

我们将热点词语与附件3中的留言主题进行比对，选出5大热点问题。

针对result1文档，可以看出扰民问题非常严重，我们将所有扰民问题进行归类，输出至“扰民问题.xsl”文档的第一个sheet中。再根据词云图，找出可能扰民的原因所在，包括“地铁”和“搅拌机”，最终得到“A市丽发新城小区附近搅拌站噪音扰民和污染环境”是热点问题。A市立发共出现12次，点赞和反对数工52票。

针对result2文档，根据词云图，最终得到“A市A5区汇金路五矿万境K9县存在一系列问题”

针对result3文档，可以看出新城问题比较严重，我们将所有新城问题进行归类。再根据词云图，找出可能扰民的原因所在，最终得到“A市富绿物业丽发新城强行断业主家水”。

针对result4文档，可以看出投诉问题和欠贷保护伞问题非常严重，我们将拖欠问题进行归类，输出至“投诉问题.xsl”文档的第一个sheet中。再根据词云图，找出可能投诉的原因所在，包括“高压”和“车贷”，最终得到“关于A6区月亮岛路沿线架设110kv高压线杆的投诉”和“严惩A市58车贷特大集资诈骗案保护伞”。

2.2.3输出五大热点问题和问题明细

五大热点问题

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	217032	100	2019/2-2019/3	A市58车贷	严惩A市58车贷特大集资诈骗案保护伞
2	208636	88.714	2019.7-2019.8	A市A5区汇金路	A市A5区汇金路五矿万境K9县存在一系列问题
3	262052	5.941	2019/2-2019/3	A6区月亮岛路沿线架	关于A6区月亮岛路沿线架设110kv高压线杆的投诉
4	214282	3.405	2020.2-2019.12	A市丽发新城小区	A市丽发新城小区附近搅拌站噪音扰民和污染环境
5	193091	2.613	2019/6-2019/7	A市富绿物业丽发新城	A市富绿物业丽发新城强行断业主家水

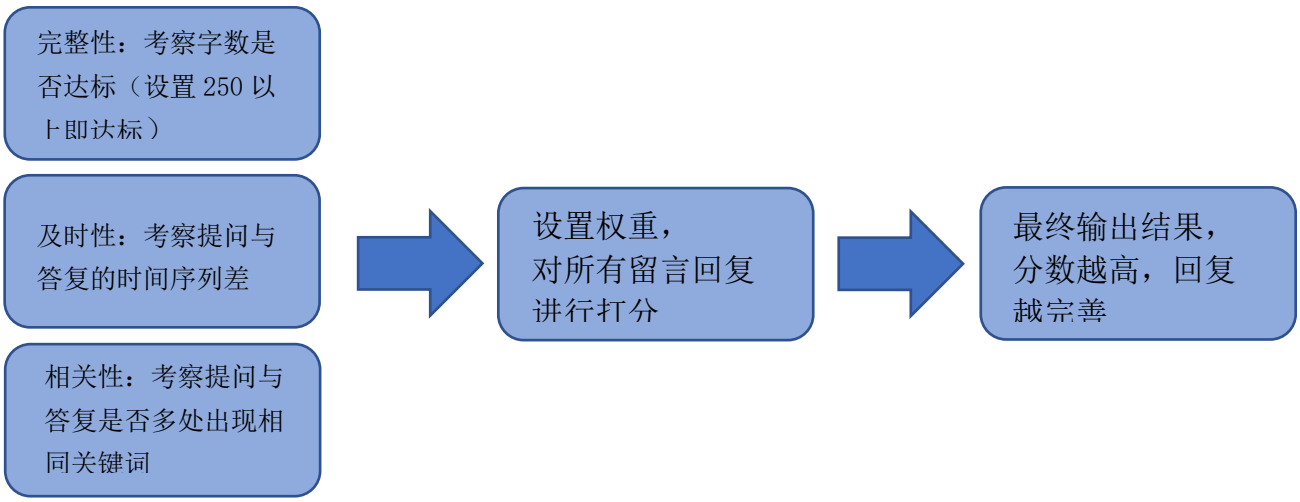
问题明细

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
1	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	息总是失望，	0	821
1	217032	A00056543	市58车贷特大集资诈骗案件	2019/2/25 9:58:37	、苏纳弟弟为	0	790
1	194343	A000106161	市58车贷案警官应跟进关注	2019/3/1 22:12:30	没有跟进市领	0	733
1	268251	A000106090	立案近半年毫无进展，单位	2019/2/2 15:03:05	和资产，这种	0	25
1	240554	A00029163	车贷老板跑路美国，经侦拖	2019/2/10 20:58:40	是涉嫌保护伞	0	6
1	226265	A000106448	王办理58车贷案件，还我们	2019/5/28 15:08:51	这样的经侦我	0	3
1	254532	A000106062	性退出立案近半年没有发过	2019/1/14 22:08:20	在58官网上关	0	3
1	272413	A000106062	贷恶性退出，A4区立案已近	2019/1/14 20:23:57	我们出借人又	0	2
1	234320	A000106592	让A市因为58车贷案件而臭名	2019/7/8 17:16:57	还是一如既往	0	0
1	218132	A000106090	请求过问A市58车贷案件进展	2019/1/29 19:15:49	办案警官毛浚	0	0
1	264119	A00084445	个月过去，A4区公安分局未	2019/1/19 9:47:23	，未查封资产	0	0
1	272858	A00061787	性退出案件为什么不发布案	2019/1/16 23:21:21	立案近半年，	0	0
1	223787	A00034861	案件创造全国典型诈骗案，	2019/1/11 21:12:34	任由犯罪嫌	0	0
2	208636	A00077171	汇金路五矿万境K9县存在一	2019/8/19 11:34:04	生过狗咬人，	0	2097
2	220716	A00089049	与汇金路之间路段）人行经	2019/1/23 10:32:59	烂泥巴到处都	0	0
3	262052	A00072424	亮岛路沿线架设110kv高压	2019/3/26 14:33:47	5kv以上电力	0	78
3	272089	A00061602	6区月亮岛路110kv高压线	2019/4/9 17:10:01	体操学校、西	2	55
3	218442	A00099016	路架设高压电线环评造假，	2019/4/8 21:19:40	害。次工程基	0	22
3	268250	A00072424	亮岛路沿线架设110KV高压	2019/3/26 10:17:31	在市中心地区	0	10
3	254865	A00099016	亮岛路沿线架设110kv高压	2019/4/3 17:36:58	在市中心地区	0	5
3	234885	A00060375	岛路11万伏高压线没用地埋	2019/4/5 13:01:17	而且破坏市容	0	2
3	231773	A00010141	岛路架设高压电线，强烈要	2019/4/12 14:59:14	使用的环评报	0	1
4	219174	A00081998	发新城小区内垃圾站散发严	2019/7/3 23:27:02	饭睡觉都能闻	0	3
4	203393	A00053065	侧面建设混泥土搅拌站，粉	019/11/19 14:51:52	尘，严重影响	0	2
4	243692	A909201	城小区附近的搅拌站噪音严	2019/11/15	尘极大，都飘	0	2
4	272224	A909224	城小区噪音大粉尘大,求撤走	2020/1/9	太大无法呼吸	0	1
4	190108	A909240	丽发新城小区旁边建搅拌站	2019/12/21	几千名学生的	0	1
4	217700	A909239	市城小区旁的搅拌站严重影响	2019/12/21	毛到丽发新城	0	1
4	244512	A00094706	丽发新城小区粉尘大的孩子	2019/12/5	人都无法正常	0	1
4	272361	A909242	发新城小区旁建搅拌厂严重	2019/12/4	常上班不在家	0	1
4	281546	A00051470	城小区附近搅拌站粉尘大,无	2019/11/29	很大,无法正	0	1
4	284576	A00063717	丽发新城小区云塘路还没有2	2019/7/26 17:41:39	一身灰，家里	0	0
4	215842	A909210	2区丽发新城小区附近太吵	2020/1/26	厂是怎么回事	0	0
4	214282	A909209	城小区附近搅拌站噪音扰民	2020/1/25	，烦死了不仅	0	0
4	239648	A909211	新城小区附近搅拌站明目张	2020/1/6	立！都不敢开	0	0
4	235362	A909215	新城小区附近水泥搅拌站非	2020/1/6	严重危害居民	0	0
4	234327	A909212	噪声不断的丽发新城小区	2019/12/26	还产生大量粉	0	0
4	216824	A909214	工砂石料噪音污水影响丽发	2019/12/25	严重扰民的噪	0	0
4	284485	A909222	搅拌厂建在A市丽发新城小区	2019/12/18	拌厂，危害广	0	0
4	238212	A909203	新城小区附近建搅拌站合理	2019/12/12	居民应是一个	0	0
4	239336	A909213	区丽发新城小区遭搅拌站严	2019/12/11	作，离居民区	0	0
4	255276	A909219	望领导“拯救”丽发新城小	2019/12/11	共受尘土影响	0	0
4	213464	A909233	发新城小区附近违建搅拌站	2019/12/10	站。该搅拌站	0	0
4	283482	A909232	新城小区附近搅拌站的一些	2019/12/7	生产产生了灰	0	0
4	215563	A909231	新城小区旁边的搅拌厂是否	2019/12/6	噪音和灰尘。	0	0
4	268109	A909230	2区丽发新城小区开发商违	2019/12/5	音污染小区居	0	0
4	233158	A909242	新城小区旁建搅拌厂严重扰	2019/12/5	家我还能忍忍	0	0
4	260979	A909229	春云街道丽发新城小区环境	2019/12/4	得了疾病住院	0	0
4	281348	A909219	领导“拯救”丽发新城小区	2019/11/24	，更有多名	0	0
4	281943	A909216	丽发新城小区附近仍存在非	2019/11/15	身心健康！请	0	0
5	193091	A00097965	绿物业丽发新城强行断业主	2019/6/19	共地摊上买的	0	242

以上两表均保存在“留言输出.xsl”中，方便查看。

2.3 问题 3 分析方法与过程

2.3.1 定义评价指标



2.3.2 操作评价指标

我们定义：答复及时性，完整性，相关性，都达标的回复为 100 分，即每个回复的初始值为 100 分，若答复在上述三项指标中有不达标者，进行减分。

完整性和及时性的考察可以直接使用 excel 进行操作。

及时性：

直接通过 excel 的函数功能，将“答复时间”和“留言时间”相减，形成时间序列差，得到的差值越小，说明政府工作人员的回复越及时，反之，差值越大，则说明该问题的回复不及时。此外，我们将所有时间差取均值，查看平均答复时间，输出在“问题三.xsl”中。

我们定义，时间差超过均值的，即为答复较慢，且超过均值时间越长，回复越慢。超过一天，倒扣 5 分，以此类推。

在“问题三.xsl”，我们定义“时间差值”：“答复时间均值”—“及时性”，若差值为正数，则说明回复时间达标，且正差值越大，回复的越快；反之，差值为负数，则说明回复时间未达标，且负差值越大，回复的越慢。

完整性：

由于政府工作人员工作繁忙，每日需回复大量问题，所以他们通常都采用精简式回答，有句式清晰，回答逻辑完整等特点。但通常此类回答的字数不会过多，经过统计，得到答复均值在 360 字。考虑到每个问题的特殊性，我们定义达到 250 字以上，即完整性指标达标。我们在 excel 中使用 len 函数，得到所有留言回复的字数。

我们定义，字数少于250字的，即为答复不完整，且低于250字越多，完整性越差。对于未超过的回复，以10字为一单位，每少10字倒扣5分，以此类推，未超过10字，则不扣分。

在“问题三.xsl”，我们定义“完整性差值”：“完整性”—250（上述设置的标准线）若差值为正数，则说明完整性指标达标，且正差值越大，回复越全面，越完整；反之，差值为负数，则说明完整性未达标，且负差值越大，回复越笼统，越没有针对性。

相关性：

根据“留言主题”与“答复意见”的关键词进行比配。

将留言主题进行关键字提取，主要提取地点和事件，在“答复意见”是否有相同的关键词出现。

上述所有内容，均已保存在“问题3.xsl”中，方便查看。

3.结合研究结果，给政府留言答复情况提建议

从工作人员角度出发，政府工作人员每日都会收到大量的群众留言，导致留言回复不及时或者不够完整的情况出现。此外，群众留言分类较困难，无法将留言数据进行统一类别划分，若是能够进行类别划分，则工作人员可以快速定位，或者转至相关部门，更有针对性的解决。

一．分类留言信息

可以采取分级办理的模式。首先，对网民的留言反映进行分类，层层分类，设置1级 2级3级分类，将留言层层筛选，分类至最匹配的层块中。例如，1级分类为“卫生计生”，2级分类则可以是“医患纠纷”或者“医政监管”，针对“医患纠纷”的3级分类可以是“医疗事故争议”、“医护人员权益“等。并且较为具体的民生问题，例如小区物业事件、涉及本地区政府自身建设、详询本地政策情况等方面的留言，按照属地化解的原则转至本市的相关部门直接处理，不仅节约时间，更能有针对性，较专业地回答网民留言，并且要求该部门工作人员直接限时回复网民。

此外，对网民留言反映关于中央、省委、省政府较大决策的问题情况，例如，地铁建设，财政拨款等群众群众关心的热点、难点问题，可以将此类问题推送至省政府督查室，由他们统一处理，可以转交专项省直有关政府交办，也可以由他们汇总审核后在省级留言平台公开回复网民。

二．是实行限时办结

设置 “谁承办、谁把关、谁回复、谁负责”的原则，按照上述分类，将留言分为咨询建议类，和调查处理类，针对咨询建议类，转交的有关相关单位应在收到转办之日起 10 个工作日内给出答复意见，因为此类问题并不复杂，通常政府工作人员或是查询政府内部工作网络即可解决；而针对较复杂的调查处理类留言，原则上 20 个工作日内给出答复意见，但也并非绝对，若是碰到棘手问题（超过 20 个工作日），政府工作人员应该主动联系网民，说明情况，适当延迟解决时间，但不得超过 40 个工作日。

处理网民所反映的问题，应该秉承着“一定解决问题，适当化解问题，明确解决时限，对内容重要、时间紧急的留言，应该特殊处理，特事特办，为民解忧，积极开辟绿色通道”的原则，加快办理进程。

此外，各地各部门收到的留言转交可能存在分类错误的情况，应该及时表示无法按时处理，并且退回留言处理，或者对难以解决的问题，及时在专栏中提出延办申请，并且告知留言提出者，说明延办理由和延办时限，未按要求提交延办申请或者未向留言提出者进行说明，一律视为迟办漏办，轻者将在全市范围内进行通报，重者将在全省范围内通报，并且实施一定的处罚措施。

三．留言应附联系方式

为确保网民的留言都能够有着落，被解决，政府部门应该及时掌握 d 网民留言的办理情况，在留言平台上，需要随时跟踪办理、回复情况，对网民留言办理情况还将进行量化分值考核，可以从及时性，完整性，相关性三处进行考核，省政府应设立督查室，由他们“不定期，不告诉”的原则，抽查办理留言处理情况，或者不定期选择有针对性的多次提及，但未解决的热点问题进行回访督查，确保群众反映的问题事事有落实。

此外，还应设置相关的处罚模式，并且要杜绝恶意回复，例如，杜绝回复套话，无针对性，字数过少的情况，真真切切，实干地帮助群众解决实际困难和问题。但同时也需要明确，网民的留言不包含详细地点，详细人物，事项模糊的情况，一律不予处理。

从群众角度出发，群众的部分留言杂乱，话语繁琐，没有主题和终点，难以让工作人员短时间内了解问题详情。群众应该在留言详情页中，主动提交自己的姓名和电话，并且确保留言信息内容真实，杜绝恶意多次留言，同一个问题间隔应大于 15 个工作日。

4 结论

对政府问政留言平台进行分析研究，了解政府群众留言和政府工作人员回复的特点和改进趋势。目前，随着社会水平的发展，群众感知到的问题越来越多，群众也越来越敢与提出自己看法，并且群众越来越关心自己提出的问题是否得到重视，所以政府问政平台的完善对政府满意度有着重大意义。但，与此同时，针对问政留言的文本分析和文本分类的一个难题。传统的文本解读已经不能满足数据量如此庞大的政府留言体系。

本文采用根据 jieba 分词方法和朴素贝叶斯法进行数据分析，划分目前问政留言平台中出现较多的留言分类，并对此进行三级分类，最终利用 F-score 公示，检验分类成果。

在有了留言分类的基础上，留言问政处理的难点有所降低，但依旧难以取定义热点问题，政府无法统一快速的定义重大热点问题，由分析结果可以看出，热点问题需要多方面考虑，某一时间段内，大量的，不同用户的同时反映，难点在于如何对该特定人群或者特定地点的留言进行归类，并且定义合理的热度评价指标，其中需要识别恶意点赞，恶意反对，恶意多次重复留言的情况。

最后，需要对每个留言回复进行评价，关键在于如何从三个指标，相关性，完整性和可解释性量化评价，每个留言事项都有其特殊性，不能一概而论。所以我们又将问题定义为咨询建议类和调查处理类，两大分类的留言时限处理不一样，可以缓解该类矛盾。

同时，我们也应该看出，群众线上留言问政是一大趋势，越来越多的群众乐于在线上提出自己的看法，统计线上留言问政情况是目前政府面临的一大困难，如何让群众高效，真实的留言，如何让政府工作人员快速并且让群众满意地回复处理留言等问题都有待于解决。

5. 参考文献

[1] 罗罗攀. Python 有趣|中文文本情感分析(2019-03-31).http：//www.

tipdm.org/tzjingsai/1628.jhtml

[2] 浪大大. 朴素贝叶斯和情感分析(2019-03-26).
https://zhuanlan.zhihu.com/p/60346098

[3] 触摸壹缕阳光.训练集、验证集和测试集（2018-11-4）

https://zhuanlan.zhihu.com/p/48976706