“智慧政务”中的文本挖掘应用

摘要：

随着社会的发展，网络问政平台逐渐成为了政府倾听民意、响应民声、集中群力的重要渠道，各类社情民意相关的文本数据量不断增加，给相关部门的整理工作带来了极大挑战。同时随着大数据、人工智能等科学技术的发展，智慧政务系统已经是社会管理的新趋势。因此，运用文本分析和数据挖掘对提升政府管理和施政效率具有极大的促进作用。

对于问题一，通过对附件二的留言信息进行 excel 到 txt 的格式转换，使用 jieba 中文分词工具，对留言信息进行分词，再采用 TF-IDF 算法得到各个留言信息的 TF-IDF 权重向量，利用朴素贝叶斯分类器进行分类，建立关于留言内容的一级标签分类模型。

对于问题二，通过对附件 3 的留言信息运用 jieba 中文分词工具 LDA 模型的主题分析进行某一时段特定地区的留言分类以及对不同类型的问题细化分类，利用 TextRank 算法计算主题权重，获得合理的热度评价指标。

对于问题三，根据数据挖掘与分析的结合思想，通过使用 excel 分析留言回复的时效性，采用文本分类的方法，先用 jieba 对数据进行预处理，后使用 Bunch 和 TF-IDF 创建向量空间并进行特征提取，采用朴素贝叶斯分类器对数据进行分类。综合两方面的信息对平台上的留言回复进行评定。通过此结果对便民服务平台和留言回复系统进行升级改进，为人民提供更高效的服务。

本文基于 Python 语言，结合第三方库 jieba 和 sklearn 库，利用 TF-IDF 算法与朴素贝叶斯分类器，实现中文文本分类，并利用 TextRank 算法实现热度评价指标的定义，反馈评价方案。

关键词：jieba 中文分词、TF-IDF 算法、朴素贝叶斯分类器、LDA 模型、创建向量空间。

目录

1.问题重述

1.1 问题背景

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。

1.2 待解决的问题

1.2.1 群众留言分类

根据网络问政平台的群众留言，建立关于留言内容的一级标签分类模型，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。

1.2.2 热点问题挖掘

针对网络问政平台的群众留言，将某一时段内反映特定地点或特定人群问题的留言进行归类，并定义合理的热度评价指标，给出评价结果。

1.2.3 答复意见的评价

通过网络问政平台上相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

2.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

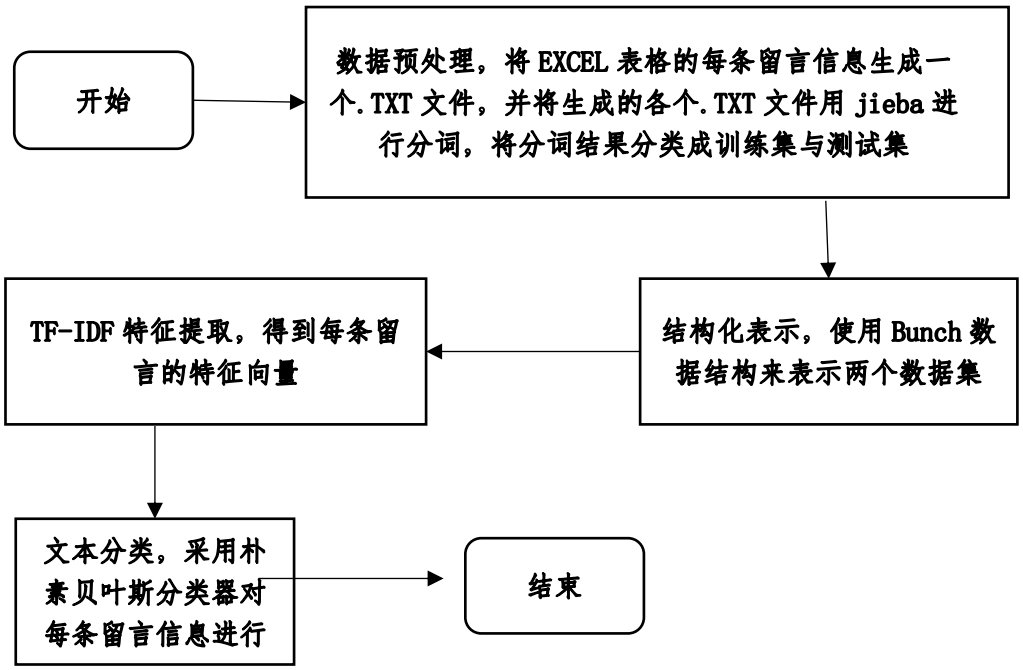

图 1　群众留言分类流程图

2.1.2 具体步骤

2.1.2.1 数据预处理

进行留言分类之前，首先要对留言信息进行预处理，主要是生成 TXT 文件、文本分词。

首先，我们使用 xlrd 模块读出 excel 文件中的数据，然后使用传统的 write()方法循环地将每行的数据写入到.txt 文件中,使得每行的留言信息位于同一个.txt 文件中。同时将数据分为训练集与测试集。

文本分词可以使用 jieba 进行分词，jieba.cut()方法可以实现文本的分词，该方法接受三个输入参数，第一个参数为待分词的字符串，第二个参数为采用的分词模式，第三个参数为是否使用 HMM 模型。其中，分词模式分为三种：精确模式、全模式和搜索引擎模式，

本文采用精确模式，此模式适合进行文本分析。

2.1.2.2 结构化表示

我们采用的是 Scikit-Learn 库中的 Bunch 数据结构来表示训练集和测试集。我们在 Bunch 对象里面创建了 4 个成员：target_name 存放整个数据集的类别集合；label 存放所有文本的标签；filenames 存放所有文本文件的名字；contents 存放分词后的文本文件。

2.1.2.3 TF-IDF 特征提取

TF-IDF 值与一个词在留言信息文本中出现的次数成正比，某个词文本重要性越高，TF-IDF 值越大。

我们首先去除训练集与测试集中每个文本中的停用词。接着，读取并写入 bunch 对象后，利用 bunch 对象构建 TF-IDF 词向量空间对象，并使用 TfidVectorizer 初始化向量空间模型，使用 TfidfTransformer()统计每个词语的 TF-IDF 权值。最后，将数据集中所有文本文件映射到同一个 TF-IDF 词向量空间中，将文本转为词频矩阵，单独保存字典文件。

2.1.2.4 文本分类

我们采用的是朴素贝叶斯分类器，这个分类器有我们需要的封装好了的函数——MultinomialNB，该函数能够获取训练集的权重矩阵和标签，进行训练，然后获取测试集的权重矩阵，进行预测，并给出预测标签。

2.2 问题 2 分析方法与过程

2.2.1 数据筛选

（1）根据附件 3 对某一时段内反应特定地点问题的留言进行归类，得到 A1 区、A2 区、A3 区、A4 区、A5 区、A6 区、A7 县、A8 县、

A9 市等 15 个地区。

（2）根据附件 3 对不同问题进行分类，得到 148 种不同问题。

2.2.2 群众反映热点问题分析流程图

根据附件 3 的数据作为模型的基础，对其进行如下图三个阶段的分析处理，定义出合理的热度评价指标，并给出最终结果。

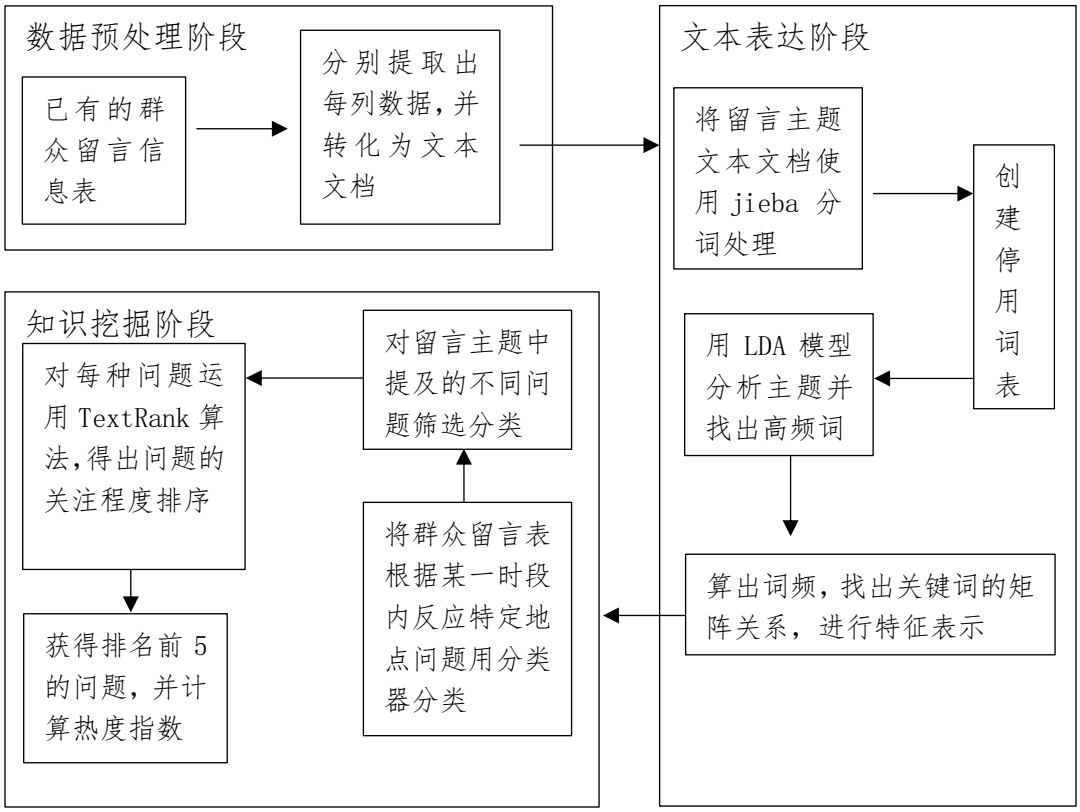

图 2　群众反映热点分析流程图

2.2.3 具体步骤

2.2.3.1 数据预处理

在对某一时段内反应特定地点或人群问题的留言进行归类前，首先需要确定分类的类别，由于附件 3 所给的留言中都提到了地点，因此选取特定地点分为 A1 区、A2 区、A3 区、A4 区、A5 区、A6 区、A7 县、A8 县、A9 市、C5 市、E7 县、K3 县、K9 县、高新区、A 市其他 15 类。（详情见“按照特定地点归类.xlsx”）

其次，对 Excel 表格进行处理，由于求的是热点问题，我们未对文本去重处理，用 Python 对 Excel 的每列数据分别进行提取并使用 write()方法保存为 txt 文件，其中主要用到的文本是留言主题的文本文档，并对其进行整理。

2.2.3.2 文本表达

在使用文本数据前，需要对留言主题进行数据清洗，进行 jieba 分词处理及词性筛选,创建停用词表,并过滤停用词,提取出关键词。为了提高数据的精确性，问题 2 采用 LDA 主题模型，从文档中抽取出主题，找出高频词，并找出关键词间的矩阵关系，进行特征表示。

2.2.3.3 知识挖掘

在得出关键词之间的关系后,运用分类器对留言主题根据关键词进行特定地点分类，接着再对留言主题中提及的问题进行筛选分类，将噪音、污染、户口等问题分成 148 类，可得到热点问题留言表。

为了寻找出最受人们关注的热点问题，使用 TextRank 算法迭代传播各节点的权重，直至收敛，对关键词进行排序，得出问题的关注热度。TextRank 算法的公式如下：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

最后，运用热度指数公式计算得出各种问题的热度指数，对其排序，得出热度排前 5 名的热点问题表。热度指数公式如下：

$$\text{热度指数} = \frac{\text{该问题留言条数+该问题点赞数+该问题反对数}}{\text{总留言条数+总点赞数+总反对数}}$$

2.3 问题 3 分析方法与过程

2.3.1 问题分析及评价方案介绍

问题 3 要求对于答复意见的质量给出一套完整的评价方案，并尝试实现。对于网络问政平台，从人民的角度来讲其留言的答复最重要的一方面是及时，一方面是有效。及时，则从时效性，（答复时间与留言时间的间隔时间表示），时间间隔越短，时效性越好。有效，则从可执行性来表达，可执行性即告诉人们一个该如何去解决，或已解决的结果，如此才能满足人们对于问政平台的需求。制定的评价方案：条件 1 为于提问后 10 天之内的回复为优秀，反之为差；条件 2 为回复内容为可执行性内容则记为 1，反之为 0。

2.3.2 答复评价问题流程图

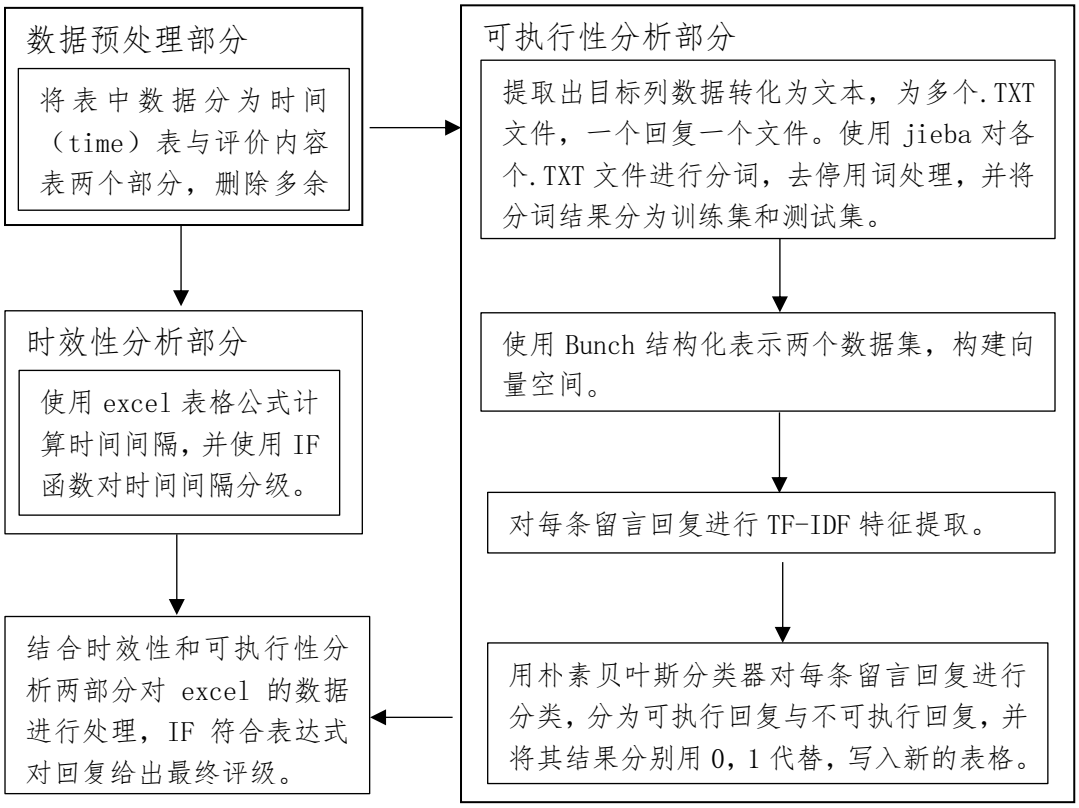

图 3　答复评价问题流程图

2.3.3 具体步骤

2.3.3.1 数据预处理

在分析回复前，先对已知数据进行预处理，将 excel 表格中的信息分为时间数据和回复内容数据，删除冗杂信息，将“留言时间”和“回复时间”置于一个表中，回复信息及留言 ID 置于一表中，以便以下的数据提取与采用。

2.3.3.2 可执行性分析

此处对于留言内容可执行性的分析，主要采用的文本分类方法，将留言内容分为可执行与不可执行两部分。

先使用 xlrd 模块读取 excel 文件中的数据，然后使用传统的 write()方法将 excel 中留言回复列的数据转化为 txt 形式,使得一行的留言信息位于一个.txt 文件中。同时将数据分为训练集与测试集。之后使用 jieba 进行分词处理，并导入停用词库对回复信息进行去停用词处理。

2.3.3.3 结构化及 TF-IDF 特征提取

分词处理后，由于回复文本只有一行，则得到一个个词袋，后采用 Scikit-Learn 库中的 Bunch 数据结构来表示训练集和测试集，分别保存在一个文件夹中。其次采用了 1/4 的数据进入训练集数据使用 TfidVectorizer 初始化向量空间模型，使用 TfidfTransformer()统计每个词语的 TF-IDF 权值。接着将所有的文本数据映射到一个词向量空间中就成功创建一个 TF-IDF 词向量空间实例，将文本转为权重矩阵，并单独保存字典文件。这样就完成了将训练集数据转换为 TF-IDF 词向量空间中的实例。

2.3.3.4 设计分类器

首先要把测试数据也映射到上面这个 TF-IDF 词向量空间中，我们采用的是朴素贝叶斯分类器，使用其封装好的函数---MultinomialNB，读取 bunch 对象，导入训练集、测试集，训练分类器，输入词袋向量和分类标签，即获取训练集的权重矩阵和标签，进行训练。接着获取测试集的权重矩阵，进行预测，并用 0，1 分别表示不可执行性与可执行性写入新的 excel 表格。

2.3.3.5 时效性分析

时效性分析只是在 excel 中操作，首先利用 excel 中的求差值的方法在 2.3.3.4 形成的新表格中，计算出留言时间与答复时间的时间差，记住设置单元格格式，标明单元格属性。然后使用 IF 函数以计算出的时间间隔为标准，10 天之内回复则为优，反之为差。

2.3.3.6 最终评级

结合 2.3.3.2 的可执行性分析结果及 2.3.3.5 的时效性分析结果，再次使用 excel 表格的 IF 复合函数对最终等级进行划分：“优，1”为优秀回复；“优，0”为良好回复；“差，1”为合格回复；“差，0”为不合格回复。

3.结果分析

3.1 问题 1 结果分析

本实验采用五分之一的数据作为训练集，训练后，得到的结果如下图所示。实验列出了留言信息实际类别与预测类别的不相同时的情况，并计算了朴素贝叶斯分类器错误比率、精度、召回度，其中错误比率约为 18.024%，精度与召回度均达到 0.82，说明该分类器具有较好的分类性能。

```
./test_corpus_seg/环境保护/2845.txt : 实际类别: 环境保护  -->预测类别: 城乡建设
./test_corpus_seg/环境保护/2848.txt : 实际类别: 环境保护  -->预测类别: 劳动和社会保障
./test_corpus_seg/环境保护/2862.txt : 实际类别: 环境保护  -->预测类别: 教育文体
./test_corpus_seg/环境保护/2864.txt : 实际类别: 环境保护  -->预测类别: 劳动和社会保障
./test_corpus_seg/环境保护/2865.txt : 实际类别: 环境保护  -->预测类别: 商贸旅游
./test_corpus_seg/环境保护/2867.txt : 实际类别: 环境保护  -->预测类别: 城乡建设
./test_corpus_seg/环境保护/2871.txt : 实际类别: 环境保护  -->预测类别: 劳动和社会保障
./test_corpus_seg/环境保护/2896.txt : 实际类别: 环境保护  -->预测类别: 城乡建设
./test_corpus_seg/环境保护/2906.txt : 实际类别: 环境保护  -->预测类别: 城乡建设
./test_corpus_seg/环境保护/2911.txt : 实际类别: 环境保护  -->预测类别: 劳动和社会保障
./test_corpus_seg/环境保护/2923.txt : 实际类别: 环境保护  -->预测类别: 劳动和社会保障
./test_corpus_seg/环境保护/2932.txt : 实际类别: 环境保护  -->预测类别: 劳动和社会保障
预测完毕!
error_rate: 18.023887079261673 %
精度：0.823
召回：0.820
f1-score：0.814
```

图 4　部分数据分类器训练后结果

3.2 问题 2 结果分析

该问题将附件 3 所给的所有数据进行两次按照不同类型的分类，分出不同的问题 ID，并使用 LDA 主题模型获取了热度问题权重，计算出热度指数。问题 2 获得的部分结果如表 1、表 2 所示。（具体详情见“热点问题表.xlsx”和“热点问题留言明细表.xlsx”。）

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	102	0.1454	2019/1/14 至 2019/7/8	A 市 A4 区 58 车贷案受害人	58 车贷诈骗未合理解决
2	137	0.1292	2019/5/5 至 2019/9/19	A 市 A5 区五矿万境 K9 县居民	房屋交房后仍有许多问题
3	8	0.1085	2019/4/11 至 2019/4/12	A 市金毛湾	学区房配套入学的问题
4	15	0.0423	2019/8/23 至 2019/9/6	A4 区绿地海外滩小区	居民生活受小区与高铁站过近影响
5	6	0.0149	2019/6/19 至 2019/6/20	A 市宝绿物业丽发新城	物业强行断业主家水

表 1　热点问题

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	191153	A909097	泉塘小区麻将馆	9/12/15 18:43	并没有什么	0	0
1	198493	A00011876	汇小区一住改	19/7/14 7:36:	。2、该麻	0	0
1	211453	A00011876	的非法住改商	19/9/16 17:29:	。社区，11	0	0
1	218587	A00055713	市场小区47栋商	019/9/6 0:15:	干扰到附近的	1	0
1	218590	A00075700	松雅安置小区	019/4/2 0:01:	有甚者直接	0	0
1	225880	A00011876	非法住改商，	19/7/8 13:28:	。2、该麻	0	0
1	232854	A00097784	置小区夜宵店噪	9/5/25 11:42:	是下午五点	0	0
1	233377	A00086041	园小区a7栋夜	19/7/11 16:06:	居民无法休	5	0
1	237329	A00097784	小区宵夜店噪音	9/5/29 22:53:	问是下午五	0	0
1	237409	A00050277	城三期小区麻	19/2/15 23:40:	也说把问题	4	1
1	247523	A00022530	乐小区业主反映	9/12/31 9:03:	害极大，现	0	0
1	247859	A00011774	动摊贩占道经	9/10/16 14:45:	酒，大声中	1	0
1	249586	A00099441	路晶华美地小	19/3/22 19:53:	规范》、《	0	0
1	254873	A00040338	小区同一首歌	19/11/8 1:16:	孩。每晚既	0	0
1	256887	A00094487	安置小区KTV	019/5/9 0:09:	果不好，不	0	0
1	259405	A909098	城西小区ktv深	9/10/30 12:58	还唱歌不停	1	0
1	287540	A909098	小区橘红酒店ktv	19/12/6 10:22:	依然没有得	0	0
1	225191	A00017755	第三小学大放	19/9/4 15:03:	乐，中午十	0	0
1	269825	A00061707	雅湖小学噪音扰	19/11/5 14:45:	音越来越大	4	0
1	197619	A909107	路和灰埠路上	9/10/21 22:48	息。车辆经	5	0

表 2　热点问题留言明细表

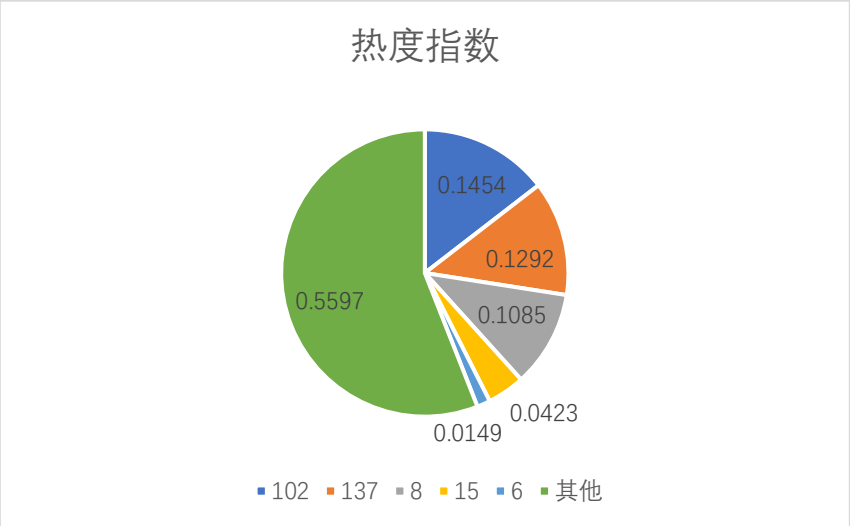

图 5　热度指数统计饼图

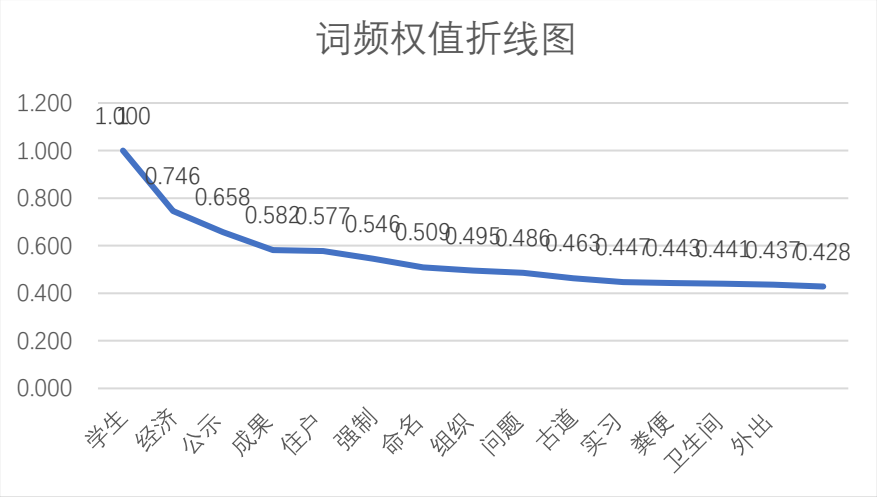

图 6　部分词频权值折线图

3.3 问题 3 结果分析

该实验采用了 excel 与 python 相结合的方法，对时效性分析采用传统的 excel 分析计算，对可执行性分析通过文本分类，用 1/4 的数据为训练集进行训练，采用朴素贝叶斯分类器对测试数据进行预测，并将其分类写入表格，最后回到 excel 结合两方面的数据对回复进行综合评定。问题三获得的部分结果见表 3（答复意见评价表）。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	时间间隔	时效性	可执行性	评定结果
2549	A00045581	蓉华苑物业管理	2019/4/25 9:32:09	却以交20万保证	费，在业主大会	2019/5/10 14:56:53	365	差	1	合格
2554	A00023583	海路洋湖段怎么这	2019/4/24 16:03:40	生意带来很大影响	且换填后还有三	2019/5/9 9:49:10	353	优	1	优秀
2555	A00031618	A市民营幼儿园	2019/4/24 15:40:04	更是加大了教师的	教职工要依法签订	2019/5/9 9:49:14	354	优	1	优秀
2557	A000110735	能享受人才新政	2019/4/24 15:07:30	市，想买套公寓，	下（含），首次	2019/5/9 9:49:42	354	优	1	优秀
2574	A0009233	公交站点名称变	2019/4/23 17:03:19	岭小学”，原“	”的问题。公交站	2019/5/9 9:51:30	376	差	1	合格
2759	A00077538	含浦镇马路卫生	2019/4/8 8:37	巴冲到右边，越	有说明卫生较差	2019/5/9 10:02:08	745	差	1	合格
2849	A000100804	村小区盼望早日	2019/3/29 11:53:23	社区惠民装电梯的	政府办公室下发了	2019/5/9 10:18:58	982	差	0	不合格
3681	UU00812	澜湾社区居民的集	2018/12/31 22:21:59	，天寒地冻的跑好	及设施设备采购等	2019/1/29 10:53:00	684	差	0	不合格
3683	UU008792	光住宅楼无故停	2018/12/31 9:55:00	到相关准确开工信	检查后，西地省楚	2019/1/16 15:29:43	389	差	1	合格
3684	UU008687	路洋湖壹号小区	2018/12/31 9:45:59	等地方做立体绿化	要求完成了建设	2019/1/16 15:31:05	389	差	0	不合格
3685	UU0082204	大托街道大托新村	2018/12/30 22:30:30	局审批通过《温室	地征收补偿款给原	2019/3/11 16:06:33	1697	差	0	不合格
3692	UU008829	D区安置房人防	2018/12/29 23:27:51	置房地下室近两万	发[2014]7号文件	2019/1/29 10:52:01	731	差	1	合格
3700	UU00877	请求修建一座人行	2018/12/29 11:55:34	量从小区开车出去	具体选址，招标	2019/1/14 14:34:58	386	差	1	合格
3704	UU0081480	芒果金融平台涉	2018/12/28 17:18:45	关政府部门的大力	，已由银盆岭派出	2019/1/3 14:03:07	140	优	1	优秀
3713	UU0081227	增开A市261路公	2018/12/28 7:53:25	上！天寒地冻，其	员工作时间长，	2019/1/14 14:33:17	414	差	1	合格
3720	UU008444	披塘路交叉路口	2018/12/27 15:18:07	https://baidu.com	路口两端各拆除20	2019/3/6 10:26:14	1651	差	1	合格
3727	UU0081194	梓坡路益丰大药	2018/12/27 1:55:21	各种理由拒绝退货	的信息进行投诉信	2019/1/3 14:02:47	180	优	1	优秀
3733	UU008706	市梅溪湖开办一	2018/12/26 16:51:40	建议在艺术中心先	期二期金菊路与雪	2019/1/14 14:32:40	453	差	1	合格

表 3　答复意见评价表

4. 结论

如今，科技发展之迅速让人惊叹，在各领域高速运转的当今社会，智慧政务系统显得尤为重要。与此同时，信息时代的到来使得大数据的出现，我们生活中产生的数据越来越多。为了方便人们的日常生活，人们对大数据进行了处理应用。

网络问政平台中群众留言板块的文本挖掘与处理，不仅能够加快政府部门处理相关热点问题的效率，减轻了政府工作人员的负担，还可以让公职人员了解民意，更加具体地落实“为人民服务”这一宗旨。

“智慧政务”中的文本挖掘应用通过对群众留言分类、热点问题挖掘和答复意见的评价三个问题锻炼了当今大学生的动手实践能力与思维能力，同时也提高了人们对政务工作的了解程度。

参考文献

[1] 张良均,王路,谭立云,苏剑林等.Python 数据分析与挖掘实战.机械工业出版社.2016:310-335

[2] 廖一星,严素蓉.基于 Python 的中文文本分类的实现.浙江财经大学.2016

[3] 夏海峰，陈军华.基于文本挖掘的投诉热点智能分类.上海师范大学.2013

[4] 美团算法团队.美团机器学习实践.人民邮电出版社.2018