
“智慧政务”中的文本挖掘应用

摘要

本文探究的是基于自然语言处理技术的智慧政务系统。通过机器学习及自然语言处理技术迎合社会治理创新发展的新趋势，提升政府的管理水平和施政效率。

对于问题一，使用朴素贝叶斯分类方法，对自然语言进行处理，对其做向量化使机器能尽量理解和表示人类的语言。利用中文停用词剔除无用词语并将自然语言文本的中文词语分割，作为模型特征，将特征向量化。分割数据集，使其分为训练数据集和测试数据集，数据集比例为 3:1，使用训练数据集进行分类模型训练，并使用测试数据集进行模型准确率测试。最后验证本文建立的分

类模型的正确性，并对模型做出客观评价及推广。

对于问题二，通过分离热度相关变量，建立热度变量表。通过数学模型对变量相关度权值拟合并建立热度评价模型，分割时间段，将时间按月分为不同的时间块，按建立的热度评价模型将一时间区块内反映特定地点或特定人群问题的留言进行归类，并给出按热度排名前五的热点问题及相应热点问题的留言信息。

对于问题三，通过对附件中问题回复的数据，分析情感及思考模型，建立一个回复必须与事项相关，要有依据的支持，并使得提问者便于理解的评价方案。

关键字：朴素贝叶斯 特征分类模型 热度评价模型

一、问题重述

1.1 问题的背景

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 问题的提出

1、群众留言分类

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。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

通常使用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

2、热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发

现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据附件 3 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，定义合理的热度评价指标，并给出评价结果，给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

3、答复意见的评价

针对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。

二、问题分析

2.1 问题一分析

利用 Python 机器学习框架 scikit-learn，使用朴素贝叶斯分类方法，对自然语言进行处理，对其做向量化使机器能尽量理解和表示人类的语言。利用中文停用词剔除无用词语并将自然语言文本的中文词语分割，作为模型特征，将特征向量化。分割数据集，使其分为训练数据集和测试数据集，数据集比例为 3:1，使用训练数据集进行分类模型训练，并使用测试数据集进行模型准确率测试。

2.2 问题二分析

问题二，通过分离热度相关变量，建立热度变量表。通过数学模型对变量相关度权值拟合并建立热度评价模型，分割时间段，将时间按月分为不同的时间块，按建立的热度评价模型将一时间区块内反映特定地点或特定人群问题的留言进行归类，并给出按热度排名前五的热点问题及相应热点问题的留言信息。

图 2 数据噪点

利用 python 内置函数进行数据清洗并进行中文分词，清洗后的示例数据如图 3 所示，分词后的示例数据如图 4 所示

```
0    A3区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项...
1    位于书院路主干道的在水一方大厦一楼至四楼人为拆除水、电等设施后，烂尾多年，用护栏围着，不但占...
2    尊敬的领导：A1区苑小区位于A1区火炬路，小区物业A市程明物业管理有限公司，未经小区业主同意...
3    A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道...
4    A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道...
Name: 留言详情, dtype: object
```

图 3 清洗后数据

```
0    A3 区 大道 西行 便道 ， 未管 所 路口 至 加油站 路段 ， 人行道 包括 路灯 ...
1    位于 书院 路 主干道 的 在水一方 大厦 一楼 至 四楼 人为 拆除 水 、 电等 设施 ...
2    尊敬 的 领导 ： A1 区苑 小区 位于 A1 区 火炬 路 ， 小区 物业 A 市程明 ...
3    A1 区 A2 区华庭 小区 高层 为 二次 供水 ， 楼顶 水箱 长年 不洗 ， 现在 自...
4    A1 区 A2 区华庭 小区 高层 为 二次 供水 ， 楼顶 水箱 长年 不洗 ， 现在 自...
Name: 留言详情, dtype: object
```

图 4 分词后数据

通过 scikit-learn 将数据集划分为训练集和测试集，划分比例为 3:1，数量分布如图 5 所示

```
总数量 (9211,) 训练集数量 (6907,) 测试集数量 (2303,)
```

图 5 训练集测试集数量分布

利用 CountVectorizer 向量化工具，依据词语出现频率转化向量得到示例数据如图 6 所示

```
—— 一下 一个多月 一个月 一中 一事 一人 一件 一份 ... 黄金 黑 黑心 黑烟 黑社会 黑色 鼓励 齐全 龙
0    0    0    0    0    0    0    0    0    0    ...    0    0    0    0    0    0    0    0    0    0
1    0    0    0    0    0    0    0    0    0    0    ...    0    0    0    0    0    0    0    0    0    0
2    0    0    0    0    0    0    0    0    0    0    ...    0    0    0    0    0    0    0    0    0    0
3    0    0    0    0    0    0    0    0    0    0    ...    0    0    0    0    0    0    0    0    0    0
4    0    0    0    0    0    0    0    0    0    0    ...    0    0    0    0    0    0    0    0    0    0
[5 rows x 3727 columns]
```

图 6 特征向量化

采用朴素贝叶斯（Multinomial naive bayes）分类模型进行朴素贝叶斯分

类，未经特征向量化的训练集内容输入，做交叉验证，计算得出模型分类准确率的均值，如图 7

```
'stop_words.' % sorted(inconsistent))
D:\Study\Python\virtualEnvironment\TDB\lib\
'stop_words.' % sorted(inconsistent))
0.8274210365818515
```

图 7 分类准确率均值

使用之前分离出的测试数据集进行模型准确率验证，得到平均准确率结果均值约为 0.84 符合预期，如图 8 所示

```
D:\Study\Python\virtualEnvironment\TDB\lib\site-pa
'stop_words.' % sorted(inconsistent))
0.8363004776378636
```

图 8 测试数据集准确率均值

80 左右的准确率虽然勉强及格，但是对于真正使用起来并不算高，所以对模型进行优化，鉴于数据集数量并不多，所以目前的最高准确率均值为 0.87 左右，如图 9 所示，参数：

```
max_df = 0.13 #去除掉超过这一比例的文档中出现的关键词（过于平凡）
min_df = 3 #去除掉低于这一数量的文档中出现的关键词（过于独特）
```

```
original 0.8617326734229585
test 0.8732088580112897
```

图 9 优化后的模型平均准确率

3.1.2 问题一模型的评价

通过 F-Score 对分类方法进行评价，其公式为

$$F - Score = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

其中 Precision 为精确率，Recall 召回率，当 β 为 1 时，称为 F1-Score，其公式为

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

利用 sklearn 内置的 f1_score 方法可以求得每一类的 F1-Score 评价，结果如图 10 所示，分别对应城乡建设，环境保护，交通运输，教育文体，劳动和社会保障，商贸旅游，卫生计生七个一级标签。

```
0.7278688524590166
0.9148325358851673
0.88
0.8218694885361552
0.8544698544698545
0.9057527539779683
0.9098039215686274
```

图 10 每一类的 F1-Score 评价结果

通过求平均数，即通过下述公式可计算得到模型的整体评价，结果如图 11 所示。

$$F - Score = \frac{1}{n} \sum_{i=1}^n \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i}$$

```
0.8592282009852558
```

图 11 模型整体评价

3.2 问题二

3.2.1 问题二数据处理

导入./data/附件 3.xlsx，使用问题一中建立的分类模型对附件 3 进行一级标签的填充，填充结果数据示例如图 12 所示

	留言编号	留言用户	留言主题	...	反对数	点赞数	一级标签
0	188006	A000102948	A3区一米阳光婚纱摄影是否合法纳税了?	...	0	0	商贸旅游
1	188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌问题	...	0	1	城乡建设
2	188031	A00040066	反映A7县春华镇金鼎村水泥路、自来水到户的问题	...	0	1	城乡建设
3	188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	...	0	1	城乡建设
4	188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	...	0	0	环境保护

图 12 填充结果数据示例

对数据粗略观察后发现时间数据列存在噪点，不属于标准时间格式，如图 13 所示，进行修复后的数据如图 14 所示

```
2019/1/16 10:08:33 <class 'str'>
2018-11-15 16:07:12 <class 'datetime.datetime'>
2019/3/27 23:21:36 <class 'str'>
2019/8/22 13:42:06 <class 'str'>
2019-08-22 00:00:00 <class 'datetime.datetime'>
2019/3/5 10:49:54 <class 'str'>
2019/8/16 13:33:59 <class 'str'>
2019-08-21 21:00:21 <class 'datetime.datetime'>
2019/7/16 22:49:58 <class 'str'>
```

图 13 不标准的数据格式

```
2019-01-28 10:20:07 <class 'datetime.datetime'>
2019-07-11 17:52:36 <class 'datetime.datetime'>
2019-05-21 11:27:54 <class 'datetime.datetime'>
2019-02-11 14:09:40 <class 'datetime.datetime'>
2019-01-07 14:27:13 <class 'datetime.datetime'>
2019-04-19 10:01:56 <class 'datetime.datetime'>
2019-12-10 11:57:56 <class 'datetime.datetime'>
2019-07-04 14:25:30 <class 'datetime.datetime'>
2019-05-08 09:21:26 <class 'datetime.datetime'>
2019-05-13 11:28:08 <class 'datetime.datetime'>
2019-01-03 10:34:29 <class 'datetime.datetime'>
2019-11-14 16:03:48 <class 'datetime.datetime'>
2019-05-16 11:46:27 <class 'datetime.datetime'>
2019-06-19 16:50:17 <class 'datetime.datetime'>
2019-11-15 13:59:35 <class 'datetime.datetime'>
2019-10-23 10:36:32 <class 'datetime.datetime'>
2019-02-18 15:43:01 <class 'datetime.datetime'>
2019-07-05 20:56:10 <class 'datetime.datetime'>
```

图 14 处理后的数据格式

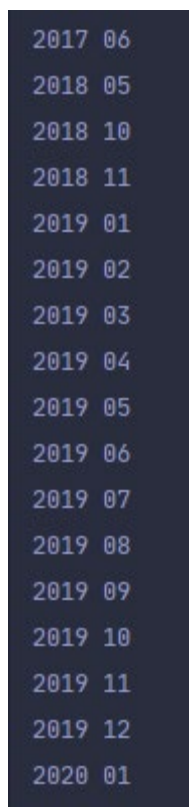
3.2.2 热度评价体系

热度相关变量见表 1

表 1 热度相关变量

time	时间
approval	赞成
disapproval	不赞成
issue	相关问题

先将时间进行排序并划分区块，以月为基准，可将所有问题分为 17 个区块，具体如图 15 所示



```
2017 06
2018 05
2018 10
2018 11
2019 01
2019 02
2019 03
2019 04
2019 05
2019 06
2019 07
2019 08
2019 09
2019 10
2019 11
2019 12
2020 01
```

图 15 时间范围区块

进行信息处理，可以获得每个时间区块内的一级标签数量，如图 16 所示

2017 06	{ '城乡建设': 0, '环境保护': 0, '交通运输': 0, '教育文体': 2, '劳动保障': 0, '商贸旅游': 0, '卫生计生': 0 }
2018 05	{ '城乡建设': 1, '环境保护': 0, '交通运输': 0, '教育文体': 1, '劳动保障': 0, '商贸旅游': 0, '卫生计生': 0 }
2018 10	{ '城乡建设': 1, '环境保护': 0, '交通运输': 0, '教育文体': 0, '劳动保障': 1, '商贸旅游': 0, '卫生计生': 0 }
2018 11	{ '城乡建设': 1, '环境保护': 0, '交通运输': 0, '教育文体': 0, '劳动保障': 1, '商贸旅游': 0, '卫生计生': 0 }
2019 01	{ '城乡建设': 175, '环境保护': 27, '交通运输': 34, '教育文体': 24, '劳动保障': 42, '商贸旅游': 43, '卫生计生': 15 }
2019 02	{ '城乡建设': 132, '环境保护': 7, '交通运输': 29, '教育文体': 28, '劳动保障': 27, '商贸旅游': 18, '卫生计生': 16 }
2019 03	{ '城乡建设': 193, '环境保护': 24, '交通运输': 31, '教育文体': 45, '劳动保障': 26, '商贸旅游': 43, '卫生计生': 7 }
2019 04	{ '城乡建设': 204, '环境保护': 49, '交通运输': 34, '教育文体': 40, '劳动保障': 29, '商贸旅游': 35, '卫生计生': 7 }
2019 05	{ '城乡建设': 203, '环境保护': 41, '交通运输': 37, '教育文体': 36, '劳动保障': 33, '商贸旅游': 32, '卫生计生': 6 }
2019 06	{ '城乡建设': 176, '环境保护': 34, '交通运输': 23, '教育文体': 26, '劳动保障': 28, '商贸旅游': 39, '卫生计生': 13 }
2019 07	{ '城乡建设': 235, '环境保护': 41, '交通运输': 34, '教育文体': 29, '劳动保障': 34, '商贸旅游': 39, '卫生计生': 6 }
2019 08	{ '城乡建设': 225, '环境保护': 32, '交通运输': 26, '教育文体': 30, '劳动保障': 27, '商贸旅游': 44, '卫生计生': 8 }
2019 09	{ '城乡建设': 200, '环境保护': 41, '交通运输': 23, '教育文体': 49, '劳动保障': 17, '商贸旅游': 50, '卫生计生': 8 }
2019 10	{ '城乡建设': 146, '环境保护': 32, '交通运输': 37, '教育文体': 23, '劳动保障': 21, '商贸旅游': 35, '卫生计生': 3 }
2019 11	{ '城乡建设': 136, '环境保护': 47, '交通运输': 27, '教育文体': 21, '劳动保障': 26, '商贸旅游': 23, '卫生计生': 1 }
2019 12	{ '城乡建设': 167, '环境保护': 50, '交通运输': 30, '教育文体': 35, '劳动保障': 29, '商贸旅游': 47, '卫生计生': 9 }
2020 01	{ '城乡建设': 39, '环境保护': 14, '交通运输': 5, '教育文体': 2, '劳动保障': 7, '商贸旅游': 11, '卫生计生': 2 }

图 16 时间区块内一级标签数量

引入热度计算模型，可计算出每一时间块中热度最高的问题，公式如下，

$$F = \frac{issuemount * \log(approval - disapproval)}{totalissuemount}$$

表 2 公式变量说明

issuemount	时间块内相关问题数量
approval	时间块内赞成数量
disapproval	时间块内不赞成数量
totalissuemount	时间块内总问题数量

排名前 5 的问题详情如图 17 所示

{ '2019-08': 7.64826383890192, '城乡建设': 225, 'F': 4.401174378396245, 'ID': 288636, 'question_info': '我是A市A5区汇金路五矿万境M9县24栋的一名业主，我们小区一开始的定位是
{ '2019-09': 6.505784060128229, '城乡建设': 200, 'F': 3.362162304975829, 'ID': 263672, 'question_info': '您好，近日看到了渝长厦高铁最新的红线征地范围以及走向经过，其经过北
{ '2019-06': 5.488937726156687, '城乡建设': 175, 'F': 2.841985627447989, 'ID': 193891, 'question_info': '位于A市A2区碧云街道丽发新城是2014年首次交房以来，到目前已入住几万
{ '2019-05': 3.7376696182833684, '城乡建设': 203, 'F': 1.960586388918666, 'ID': 267630, 'question_info': '沈书记：您好！请求您百忙中关注下居住在地铁3号线松雅西地省站西北方
{ '2019-07': 3.295836866004329, '城乡建设': 234, 'F': 1.8494624140168179, 'ID': 231690, 'question_info': 'A市及A4区各相关领导：A市N年前，就根据A市城市发展的需求，做出了洪

图 17 热度排名前五的问题

建立表格“热点问题表.xls”和“热点问题留言明细表.xls”，见附件

3.3 问题三

经过数据筛查后发现，留言主题、留言详情与答复意见为关键数据，且数据之间存在关联。以留言主题为核心，分析答复意见与留言主题和留言主题的相关性，并结合留言详情，分析答复意见与留言详情之间的对应性，以保证答

复的完整。利用 pandas 将数据导入 python 中之后，我们对数据进行了总体的分析，答复意见应围绕留言主题展开，留言主题中的关键字在答复意见中出现的次数越多，说明答复意见与留言之间的相关性越高，我们对数据进行切片处理后发现，答复意见中出现的相同关键词主要有：“感谢”、“理解”、“支持”等表明歉意的词语。另外，我们将答复意见中对留言详情进行重述的意见进行了特殊标记。发现存在重述的意见占比并不高。

综上，我们给出了三套较为标准的答复意见模板：

您好！经核查，您所反映的事项与编号×××的投诉件内容一致，我局正在对您反映的问题进行调查处理，调查处理结果将在规定时限内向您反馈。如有疑问，欢迎拨打我局（单位名称）电话×××咨询。感谢您对我们工作的支持、理解与监督！ ×年×月×日

您好！您于×月×日咨询的××问题，现答复如下： 按照《××细则》（文号）第×条规定（需指明相关文件依据），办理××业务（或申请××）需提交以下资料（或办理手续如下）：

一、…… 二、…… 三、…… 感谢您对我们工作的支持、理解与监督！
×年×月×日

您主要反映（梳理投诉事项或投诉诉求）： 1.…… 2.…… 二、调查核实情况

针对您反映的问题，我局进行了深入调查：

1.……

2.……（具体做了哪些调查工作，并针对投诉事项逐一调查回应）

三、处理情况 情况一：

根据调查情况，您反映的事项属实（基本属实或部分属实），我局后续将对××问题进行督促整改，整改措施如下：

1.……

2.……（说明整改计划和时间安排）

感谢您对××工作的关注和支持。感谢您对我们工作的支持、理解与监督！ ×年×月×日