“智慧政务”中的文本挖掘应用

摘要

电子政务是从政府的角度出发，服务于社会、企业和个人的电子商务应用之一。作为一种基于网络，符合 Internet 标准，面向政府机关、企业以及社会公众的信息服务和信息处理系统，信息的获取、利用和开发是必须解决的问题。目前的网络技术不具备信息自主开发能力。网络提供给用户的只是信息素材或粗加工过的信息，不能立即应用于实际，而且为了得到这类原始信息或数据，通常要经过一连串的网上操作，查询效率低，即信息的利用率低【1】。更重要的是我们想通过对政务文本数据的挖掘，发现一些有用的信息反馈给相关部门，以改善以后的工作。

群众留言分类文本挖掘的过程是：文本的预处理，进行分词处理与停用词处理并绘制词云，展示高频词汇；然后文本向量分类，需要用到 TF-IDF 权重矩阵与朴素贝叶斯；最后是模型的训练与评价，需要使用 F-Score 评价模型，展示分类的模型好坏。

热点问题挖掘的过程是：数据的探索与预处理，需要进行文本去重、机械压缩、短句删除，用到的方法是自动化或半自动化处理文本；然后是建模与诊断，情感倾向分析、语义网络分析、LDA 主题分析，根据诊断结果进行模型优化与重构。

答复意见的评价过程是：文本的预处理，构建模型，展现答复意见的相关性、完整性、可解释性。

关键词：TF-IDF 权重矩阵、朴素贝叶斯、F-Score、自动化或半自动化处理文本、LDA 模型

Abstract

E-government is one of the e-commerce applications that serve the society, enterprises and individuals from the point of view of the government. As an information service and information processing system based on network, in line with Internet standards, and facing government agencies, enterprises and the public, the acquisition, utilization and development of information must be solved. The current network technology does not have the ability of independent information development. The network only provides users with information material or rough information, which can not be applied to practice immediately, and in order to get this kind of original information or data, it usually has to go through a series of online operations, and the query efficiency is low. That is, the utilization rate of information is low [1]. More importantly, we want to find some useful information to feed back to the relevant departments through the mining of government text data, in order to improve the future work.

The process of text mining in mass message classification is as follows: text preprocessing, word segmentation and stop word processing, drawing word cloud, showing high-frequency vocabulary; then text vector classification, need to use TF-IDF weight matrix and naive Bayes; finally, model training and evaluation, need to use F-Score evaluation model to show the classification model is good or bad.

The process of hot issue mining is: data exploration and preprocessing, text de-duplication, mechanical compression, short sentence deletion, automatic or semi-automatic text processing, and then modeling and diagnosis, emotional tendency analysis, semantic network analysis, LDA topic analysis, model optimization and reconstruction according to the diagnosis results.

The evaluation process of response comments is: pre-processing of the text, building a model to show the relevance, completeness and interpretability of responses.

Key Words : TF-IDF weight matrix, naive Bayes, F-Score, automatic or semi-automatic text processing, LDA model.

目录

1.挖掘的背景与目标

1.1 挖掘的背景

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。附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

1.2 挖掘目标

（1）群众留言分类，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，进行不同问题的归类。
（2）热点问题挖掘，根据附件 3 进行问题分类，定义合理的热度评价指标，给出排名前 5 的热点问题。
（3）答复意见的评价，展现相关答复意见的相关性、完整性、可解释性。

2.问题分析

2.1 群众留言分类分析

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。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。 给出的数据很多，需要进行文本处理，进行相关问题的分类，该问题的重点就是怎么很好的评价问题之间的相关性，需要运用 F-Score 进行相关模型的评价。

2.2 热点问题挖机分析

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。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，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题 对应的留言信息，并保存为“热点问题留言明细表.xls”。这个问题主要是进行热点问题的分析与排名，需要进行同一类问题的分组，在进行留言的热度评价排名。

2.3 答复意见的评价

针对附件 4 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，展示回复的有用性与实在性。

3.分析方法与过程

3.1 群众留言分类的方法与过程

3.1.1 导入数据

导入相关数据，进行数据抽样：

data - DataFrame

索引	1	2	3	4	5
留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
24	A0007401l	A市西港建筑集团占道施工有安全隐患	2020/1/6 12:09:38	A3区大道西行…	城乡建设
37	U0008473	A市在水一方大厦人为烂尾多年，安全隐患严重	2020/1/4 11:17:46	位于书院路王…	城乡建设
83	A0006399l	投诉A市A1区亮物业违规收停车费	2019/12/30 17:06:14	尊敬的领导：…	城乡建设
303	U0007137	A1区蒂裡南路A2区华庭榠项水箱长年不洗	2019/12/6 14:40:14	A1区A2区华庭…	城乡建设
319	U0007137	A1区A2区华庭自来水好大一股霉味	2019/12/5 11:17:22	A1区A2区华庭…	城乡建设
379	A0001677l	投诉A市盛世耀凯小区物业无故停水	2019/11/28 9:08:38	我在2015年购…	城乡建设
382	U0005806	咨调A市楼盘集中供暖一事	2019/11/27 17:14:11	由于西地省地…	城乡建设
445	A0001920l	A3区桐梓坡西路可可小城长期停水得不到解决	2019/11/19 22:39:36	尊敬的胡书记…	城乡建设
476	U0003167	反映C4市收取城市垃圾处理费不平等的问题	2019/11/15 11:44:12	我们是梅家田…	城乡建设
530	U0008488	A3区魏家坡小区脏乱差	2019/11/10 18:59:24	尊敬的A市政…	城乡建设
532	U0008488	A市魏家坡小区脏乱差	2019/11/10 12:30:27	尊敬的A市政…	城乡建设
673	A0008864l	A2区泰华一村小区第四届非法业委会涉嫌侵占小区业主公共资金	2019/10/24 11:29:02	请求依法监督…	城乡建设
994	U0005196	A3区梅溪湖壹号御湾业主用水难	2019/9/18 22:43:26	我住在梅溪湖…	城乡建设
1005	U0006509	A4区鸿涛嘉荣湾强行对入住的业主关水限电	2019/9/18 13:36:14	尊敬的领导：…	城乡建设
1110	A0009977l	地铁5号线施工导致A市林楚国际星城小区三期一个月停电10来次	2019/9/9 11:07:48	地铁5号线施…	城乡建设
1309	U0005083	A6区河和泉郡用电的问题能不能解决	2019/8/21 15:12:20	尊敬的领导：…	城乡建设
1440	A0003288	A市林楚国际新城从6月份开始停电好多次了	2019/8/6 10:28:55	A市A5区朝晖…	城乡建设
1775	U0002150	始A9市城区南西片区城铁站设立的建议	2019/7/4 18:52:39	肯定是选择在…	城乡建设
1783	U0004763	请A6区政府加大对清水新城的绿化建设	2019/7/4 14:25:30	尊敬的领导：…	城乡建设
1827	U000613	A5区楚府线几个小区经常停电	2019/7/1 20:14:52	A5区楚府线，…	城乡建设
2603	A0009965l	请调查西地省建望集团及西地省辉东安建工程有限公司的违法行为	2019/4/20 16:50:49	我叫冷疱，是…	城乡建设
3607	A0004652l	A2区山水嘉园1栋三单元群租房扰民	2019/1/8 10:08:32	A市A2区黄谷…	城乡建设
3742	A0001388l	A3区杜鹃文苑小区外的非法汽车检测站要开业了！	2018/12/26 10:13:37	A市政府、市…	城乡建设
3800	U0001518	建议A市、B市、C市联合修建中速磁悬浮（最高时速150km以上）西部快线	2018/12/20 1:23:52	A市、B市、C…	城乡建设

Index	nd	np	ns	message	label
167060	U0001726	EMS坑人的服务与速度，让人无语	2014/1/15 10:55:05	…	交通运输
172346	U0007225	I4县的出租车天燃气价格为什么…	2018/11/3 19:48:52	…	交通运输
132467	U0006516	L3县借母溪乡至凤滩的路不通	2017/5/28 8:12:44	…	交通运输
185648	U000789	嘉雨路预备役师段何时能修通？	2013/8/16 10:08:09	…	交通运输
167717	U0001924	关于G市网约车新政的个人建议	2017/11/9 10:21:38	…	交通运输
6260990	U0003141	D2区黄茶路55号渣土随意运输，…	2018/9/20 19:05:03	…	交通运输
15041	A00053551	A市新增500台出租车经营权竞标…	2013/4/13 16:19:22	…	交通运输
151309	U0005722	M14县出租车2011年的燃油补贴为…	2012/11/5 11:25:10	…	交通运输
174487	U0007344	I6市要发展，交通会是个大问题	2011/10/20 17:26:16	…	交通运输
39899	U0001208	强烈要求在B5县渌水大桥上安装…	2016/3/17 13:18:11	…	交通运输
109961	U0002714	K8县出租车挂纸牌交警队作担保	2016/11/4 16:57:44	…	交通运输

3.1.2 数据的预处理：

去除留言中的 X 序列，并进行了结巴分词，去停用词处理，对处理数据进行函数封装，进行词频统计，制作了词云图。

3.1.3 文本向量表示：

获取训练样本的 TF-IDF 权限向量，并获取了测试样本的 TF-IDF 权限向量。

3.1.4 模型训练与评价：

构建 F-Score 评价模型，进行模型的测试修改，获取最终的评价数据。

3.2 热点问题挖机分析的方法与过程

数据的探索与预处理，需要进行文本去重、机械压缩、短句删除，用到的方法是自动化或半自动化处理文本；然后是建模与诊断，情感倾向分析、语义网络分析、LDA 主题分析、根据诊断结果进行模型优化与重构。

3.3 答复意见的评价的方法与过程

3.3.1 导入数据

我们先导入数据，并进行了留言主题的统计：

	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
0	2549	A00045581	A2区景春华苑物业管理有问题	2019/4/25 9:32:09	\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...	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区景春花苑物业管理有问题”的调查核...	2019/5/10 14:56:53
1	2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	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
2	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\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...	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
3	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\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...	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
4	2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\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...	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

K7县人民反对拆除五一桥来建商业廊桥 3
投诉J5县国城国际开发商严重违反合同 3
反对在A7县江背镇乌川湖村兴建涵洞 3
投诉K6县和苑新城的多项问题 2
投诉J1区下眉桥五牛城烂尾楼的问题 2
希望加快推进B9市醴宫公路拓宽建设 2
投诉A7县星沙金科时代中心房屋存在严重安全隐患 2
咨询J11市婚假期政策 2
咨询J9县的油茶种植政策 2
对K10县关于“三类人员”考试一事回复的质疑 2
反映A6区大汉汉园网签合同的问题 2
咨询关于A7县安沙镇户口迁移政策 2
请求合理设置L市城区的28路公交车终点站 2
投诉K3县二中强制高三学生补课 2
投诉J7县安监局监员特殊岗位津贴的问题 2
K市潇楚佳苑小区对面工地施工噪音大，业主多次投诉仍未果 2
反映A市潇楚北路三汉矶大桥到普瑞大道段路面减速带的问题 2
咨询A7县体育公园建设情况 2
咨询K市科目三道路驾驶技能考试等相关问题 2
A7县恒大翡翠华庭的房子问题多多，投诉维权无门 2
反映A8县市楼盘质量问题 2
举报K7县锦江村无证养猪场严重污染水源 2
我想咨询我这情况能转业到M市吗？ 2
希望解决清水塘三小学生出行安全问题 2
反映A8县木水清华二期临时用电与住房物业问题 2
投诉L市城东移动营业厅发票金额与实付金额不一样 2
请求解决B6县农民工尘肺病群体免费医疗的问题 2
K10县邝胡社区邝家村恶性涉黑采石场投诉 2
B市自闭症补助咨询 2
投诉K1区207国道隔离带中间的人行道设置不合理 2

3.3.2 去重处理

去除重新词汇，去除冗余操作如下：

```
def condense_1(str):
    for i in [1, 2]:
        j=0
        while j < len(str)-2*i:
            if str[j: j+i]==str[j+i: j+2*i] and str[j: j+i]==str[j+2*i: j+3*i]:
                    k = j+2*i
                    while k+i<len(str) and str[j: j+i]==str[k+i: k+2*i]:
                            k += i
                    str = str[: j+i] + str[k+i:]
            j += 1
        i += 1
    for i in [3, 4, 5]:
        j = 0
        while j < len(str)-2*i:
            if str[j: j+i]==str[j+i: j+2*i]:
                    k = j+i
                    while k+i<len(str) and str[j: j+i]==str[k+i: k+2*i]:
                            k += i
                    str = str[: j+i] + str[k+i:]
            j += 1
        i += 1
    return str
```

3.3.3，随机取值

随机取其中一组答复意见如下：

```
data.iloc[14]
```

'网友“UU0081227”您好！您的留言已收悉。现将有关情况回复如下：261路公交车全程24公里，配车20台，高峰期发车间距为7-8分钟/趟，平峰为10-15分钟/趟，经查看近期发车时刻表，其发车间隔正常。由于驾驶员工作时间长，劳动强度大，造成车队驾驶员短缺，公司正在积极组织调配人员充实该线路运力，公司人事部正在积极进行驾驶员招募工作，条件具备后将增加该线路配车。感谢您对我们工作的支持、理解与监督！2019年1月8日'

3.3.4 字符串化

将答复意见字符串化：

```
data.astype('str').apply(lambda x: len(x)).sum()
```

```
1008233
```

```
data1 = data.astype('str').apply(lambda x: condense_1(x))
```

```
data1.apply(lambda x: len(x)).sum()
```

```
992156
```

```
data2 = data1.apply(lambda x: len(x))
```

```
data3 = pd.concat((data1,data2),axis = 1)
```

```
data3.columns = ['答复意见','长度']
```

```
data3.head()
```

3.3.5 文本去重后结果

进行了相关文本去重处理，筛选出答复意见：

	答复意见	长度
0	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...	454
1	网友“A023583”：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区洋湖街...	303
2	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...	357
3	网友“A0110735”：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映的...	298
4	网友“A09233”，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“白竹坡路口...	159

3.3.6 去除结巴分词

去除结巴分词，将下列语句进行机器学习：

```
list(jieba.cut('我爱北京天安门，天安门前国旗升'))
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\lenovo\AppData\Local\Temp\jieba.cache
Loading model cost 0.980 seconds.
Prefix dict has been built succesfully.
```

```
['我', '爱', '北京', '天安门', '，', '天安门', '前', '国旗', '升']
```

```
help(jieba.cut)
```

```
Help on method cut in module jieba:

cut(sentence, cut_all=False, HMM=True) method of jieba.Tokenizer instance
    The main function that segments an entire sentence that contains
    Chinese characters into seperated words.

    Parameter:
        - sentence: The str(unicode) to be segmented.
        - cut_all: Model type. True for full pattern, False for accurate pattern.
        - HMM: Whether to use the Hidden Markov Model.
```

```
list(jieba.cut('我爱北京天安门，天安门前国旗升',cut_all = True))
```

```
['我', '爱', '北京', '天安', '天安门', '', '', '天安', '天安门', '门前', '国旗', '升']
```

```
list(jieba.cut('我爱北京天安门，天安门前国旗升',HMM = False))
```

```
['我', '爱', '北京', '天安门', '，', '天安门', '前', '国旗', '升']
```

```
list(jieba.cut_for_search('我爱北京天安门，天安门前国旗升'))
```

```
['我', '爱', '北京', '天安', '天安门', '，', '天安', '天安门', '前', '国旗', '升']
```

```
data5 = data4.apply(lambda x: list(jieba.cut(x)))
```

```
data5.head()
```

3.3.7 分词处理

分词处理，自定义函数进行机械压缩：

```
0    [现将，网友，在，平台，《，问政，西地省，》，栏目，向，胡华衡，书记,...
1    [网友，“，A023583，”，：，您好，！，针对，您，反映，A3，区...
2    [市民，同志，：，你好，！，您，反映，的，“，请，加快，提高，民营,...
3    [网友，“，A0110735，”，：，您好，！，您，在，平台，《，问政...
4    [网友，“，A09233，”，，，您好，，，您，的，留言，已，收悉，...
Name: 答复意见，dtype: object
```

3.3.8 构建模型

最后构建模型将回复进行量化处理展现答复意见的相关性、完整性、可解释性：

```
feel = list(feeling['word'])
def classfi(list1):
    SumScore = 0
    for i in  list1:
        if i in feel:
            SumScore += feeling['score'][feel.index(i)]
    return SumScore
```

```
date7 = data6.apply(lambda x:classfi(x))
```

```
negfile = 'liuyan_jd_neg_delStop.txt'
posfile = 'liuyan_jd_pos_delStop.txt'

neg = pd.read_csv(negfile, encoding = 'utf-8', header = None) #读入数据
pos = pd.read_csv(posfile, encoding = 'utf-8', header = None)

neg[1] = neg[0].apply(lambda s: s.split(' ')) #定义一个分割函数，然后用 apply 广播
pos[1] = pos[0].apply(lambda s: s.split(' '))
```

```
from gensim import corpora, models

#负面主题分析
neg_dict = corpora.Dictionary(neg[1]) #建立词典
neg_corpus = [neg_dict.doc2bow(i) for i in neg[1]] #建立语料库
neg_lda = models.LdaModel(neg_corpus, num_topics = 3, id2word = neg_dict) #LDA 模型训练
print("\n负面评价")
for i in range(3):
    print("主题%d : " %i)
    print(neg_lda.print_topic(i) ) #输出每个主题
```

```
#正面主题分析
pos_dict = corpora.Dictionary(pos[1])
pos_corpus = [pos_dict.doc2bow(i) for i in pos[1]]
pos_lda = models.LdaModel(pos_corpus, num_topics = 3, id2word = pos_dict)
print("\n正面评价")
for i in range(3):
    print("主题%d : " %i)
    print(pos_lda.print_topic(i) ) #输出每个主题
```

4.总结

本文的主要目的是利用数据挖掘，文本处理，可视化和数学建模技术实现智慧政务中的文本挖掘应用。首先，通过引入文本，进入垃圾文本处理操作，对原有文本实现垃圾分类，采用了数据探索，数据读取，数据抽取，去除 x 序列，去除结巴分词。去除停用词，函数分装，词频统计，词云图绘制，获取样本的 tf-idf 权值向量，然后建立模型。其次，对附件 3 中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。最后，针对附件 4 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进行评价和建模。

参考文献

【1】版权声明：本文为 CSDN 博主「3851391」的原创文章，遵循 CC 4.0 BY-SA 版权协议，转载请附上原文出处链接及本声明。
原文链接：https://blog.csdn.net/recls/java/article/details/531108