“智慧政务”中的文本挖掘应用

摘要

本文旨在基于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，通过文本特征工程、文本向量表达化与 BTM 模型，建立关于留言内容的文本分类模型、热点问题表、热点问题明细留言表，结合量化分析，建立关于答复意见的质量一套评价方案。

针对任务一，在数据清洗完成的基础下，通过文本数据向量表达化的方式，进行标签识别分类，从而得到一级标签分类模型。

针对任务二，通过建立 BTM 模型，忽略用户个性化问题，提出基于词对主题模型的话题特征提取方法，首先，引入用户因素进行主题建模，引入用户的吉布斯采样对模型参数推导，最后使用 JS 和余弦相似度联合判断话题是否为同一类，从而得到同一问题的留言，接着使用 K-means 算法进行聚类分析，输出包含留言数最多的 5 个聚类到“热点问题表”，即为前五个热点问题，同时将这五个问题包含的留言整理到“热点问题留言明细表”。

针对任务三，构建出词条和释义语言量化关系，并对问题详情与答复意见进行数值化映射到[0,1],同时结合生活经验，做出答复意见的质量评价方案。

关键词：智慧政务；文本挖掘； BTM 模型；指标评价方案；python；Excel

Abstract

The purpose of this paper is to establish a text classification model, a list of hot issues and a detailed list of hot issues through text feature engineering, text vector representation and BTM model based on the records of public political messages from the Internet and the response opinions of relevant departments to some of the public messages. Combined with quantitative analysis, a set of evaluators for the quality of the response opinions is established Case.

For task one, on the basis of the completion of data cleaning, through the way of text data vector expression, label recognition and classification are carried out, so as to get the first level label classification model.

According to task 2, by building BTM model and ignoring user personalization, a topic feature extraction method based on word to topic model is proposed. Firstly, user factor is introduced to model the topic, Gibbs sampling is introduced to deduce the model parameters, and finally JS and cosine similarity are used to jointly judge whether the topic is the same type, so as to get the message of the same problem. Then K-means algorithm is used for clustering analysis, and the output of 5 questions with the largest number of comments are clustered into the "hot issues table", which is the first five hot issues. At the same time, the comments contained in these five questions are sorted into the "hot issues message list".

According to task three, the quantitative relationship between entry and paraphrase language is constructed, and the question details and reply opinions are numerically mapped to [0,1], at the same time, combined with life experience, the quality evaluation scheme of reply opinions is made.

Key words: smart government; text mining; BTM model; index evaluation scheme; python; Excel

目录

1 问题分析

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、汇聚民意、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的挑战。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，对提升政府管理水平和施政效率具有极大的推动作用。

该题目给出四个附件，分别是分类三级标签体系（附件 1），附件 2-4 的数据来源于互联网公开渠道。

问题所给的任务一要求根据附件 2，建立关于留言内容的一级标签分类模型。并使用 F-Score 对分类方法进行评价。该问题需要克服的困难有文本语义带来的词语交叉（比如交通局的亲属拖欠我们工资）、多分类问题带来的难度、数据不平衡带来的影响以及长文本无意义的表达太多等，对于该问题，我们将运用 SVM 算法进行解决进行解决。

任务二可以拆分为三个子任务，分别是

①. 子任务 1：问题识别，即如何从众多留言中识别出相似的留言

②. 子任务 2:问题归类,把特定地点或人群的数据归并,即把相似的留言归 为同一问题，结果对应表 2

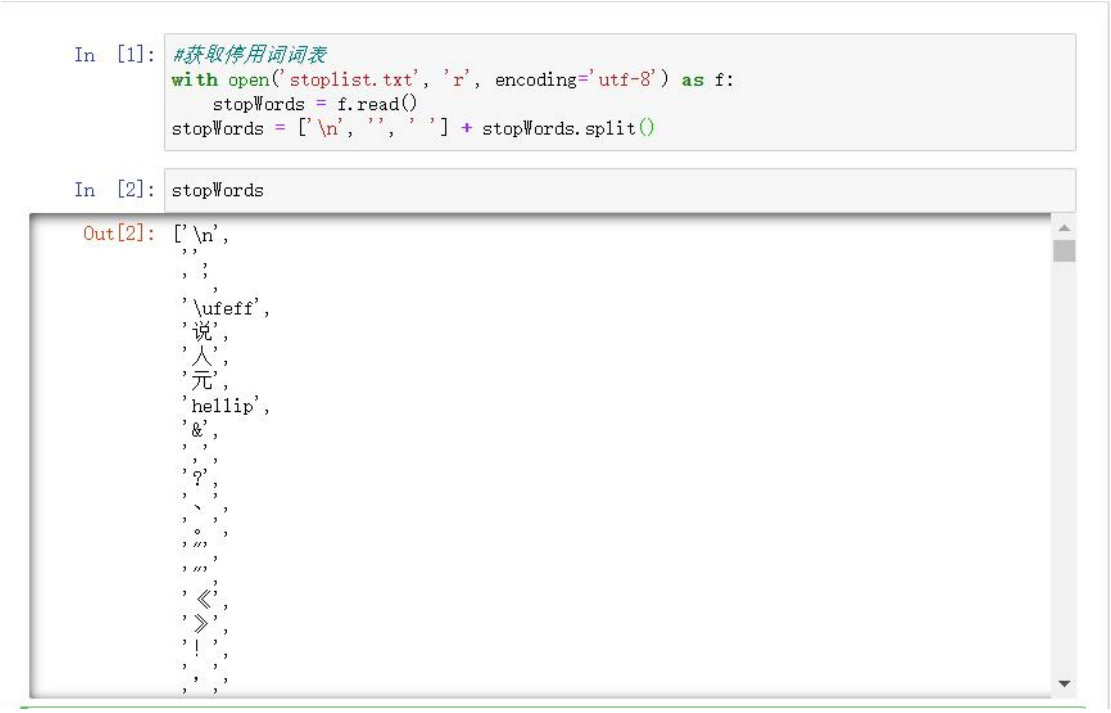
③. 子任务 3：热度评价，热度评价指标的定义和计算方法，对指标排名之后得出对应表 1

任务三要求针对附件 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，并尝试实现。该任务我们主要结合实际生活经验，从客观角度对答复意见做出评价方案。

2 数据准备

2.1 停用词词表的建立与导入

通俗地来说，停用词是对文本分析无作用的词，需要将它“停用”，类似一些符号、“的”、“一”等词，删除停用词之后会相对地节省空间和提高效率。我们通过百度下载了停用词词表，并将其保存为 txt 文件，命名为“stoplist”。



```
In [1]: #获取停用词词表
        with open('stoplist.txt', 'r', encoding='utf-8') as f:
            stopWords = f.read()
        stopWords = ['\n', '', ' '] + stopWords.split()

In [2]: stopWords

Out[2]: ['\n',
         '',
         ' ',
         '  ',
         '\ufeff',
         '说',
         '人',
         '元',
         'hellip',
         '&',
         ',',
         '?',
         '  ',
         '、',
         '。',
         '“',
         '”',
         '《',
         '》',
         '！',
         '，',
         '：',
```

图 1停用词词表的导入

2.2 建立地方名词典

本文先通过停用词词库对留言进行清洗，接着主要采用 jieba 分词的方法对留言进行分词，最后提取出词性为“ns”（地名）和“nt”（机构团体）的即为地方名、词性为“nr”（人名）的即为人群。

Jieba 允许开发者指定自己自定义的词典，以便包含 jieba 词库里没有的词。虽然 jieba 有新词识别能力，但是自行添加新词可以保证更高的正确率。用法是 jieba.load_userdict(自定义词典的路径)，词典格式和 dict.txt 一样，一个词占一行；每一行分三部分，一部分为词语，另一部分为词频，最后为词性，用空格隔开[1]。我们通过以下步骤建立起我们的词典：

- 参考附件 3，把常见的地方名写进去，并且把词性设为“ns”；比如 A 市。
- 参考附件 3，把常见的群体称呼写进去，并且把词性设为“nr”；比如学生。

将获得的词典存放于 txt 文件，并命名为“dict”，效果图如下图所示：

dict - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
邮差 6 nr
老人 6 nr
年轻人 6 nr
北京大学东门 6 ns
鼓楼大街 6 ns
广阳城 6 ns
```

图 2建立地方名词典

3 任务一：文本分类模型的建立

3.1 文本分类（NLP）

文本分类是自然语言处理中研究最为广泛的任务之一，通过构建模型实现对文本内容进行自动分类。文本分类的大致流程：文本预处理、抽取文本特征、构造分类器。其中研究最多的就是文本特征抽取，更广义上说是文本表示。关于文本表示，研究者从不同的角度出发，提出大量的文本表示模型。本文将使用基于向量空间模型的方法进行研究。[2]

3.2 基于向量空间模型（VSM）的方法

向量空间模型是将文本表示成实数值分量所构成的向量，一般而言，每个分量对应一个词项，相当于将文本表示成空间中的一个点。向量不仅可以用来训练分类器，而且计算向量之间的相似度可以度量文本之间的相似度。

采用向量空间模型进行文本表示时，需要经过分词、特征词选择、模型表示和分类算法模型的建立。[3]

3.2.1 文本特征选择

在文本处理过程中,将群众留言中的每个文本实行分词处理后,常是统计出每个文本出现的词以及相应的词频,然后将每个文本出现的词合并成一个词空间,所以词空间中出现的不同词相当多。表示一篇文本的时候,无论文本用向量空间模型还是概率统计模型来表示,文本的特征的维度都和词空间的维度一样。而每个文本中出现的词只占词空间中很少一部分,文本特征表示非常稀疏。使分类算法的时间复杂度和空间复杂度增加,而且对文本特征表示的不准确性严重影响了文本分类性能。因此,需要对文本特征进行筛选,选出最能代表文本类别的特征,这个过程就是特征选择。

特征选择的一般步骤是：

- 从文本集中取得所有的特征项,构成文本特征集合 F；
- 对集合中的每一项用下面将要介绍的特征评估函数进行打分,然后按照分值由高到低排序,得到有序集合；
- 假设需要选取 N 个分类需要的特征项,则从集合中选取前 N 个特征项,构成最终的分类特征项，将用于训练分类器和分类测试。

3.2.1.1 文本特征选择方法—互信息

互信息（MI）法的基本思想是：互信息越大，特征t_i和类别C_j共线的程度越大。如果 A、B、C、N 的含义和卡方统计中的约定相同，那么t_i和C_j的互信息可由下式计算：

$$I(t_i, c_j) = \log\frac{P(t_i,C_j)}{P(t_i)P(C_j)}$$

$$= \log\frac{P(t_i|C_j)}{P(t_i)}$$

$$\approx \log\frac{A \times N}{(A+C)\times(A+B)}$$

其中，A 为在类别 C 中特征词 t 出现的文档数； B 为在除了类别 C 的其他类别中特征词 t 出现的文档数；C 为在类别 C 中特征词 t 未出现的文档数；N 为所有类别中的文档数的总和。如果共有 m 个类别，那么每个特征词将得到 m 个相关度值，取这 m 个值的平均值作为每个特征词的权值，权值大的特征词被保留的可能性大。如果特征t_i和C_j类别无关，则 $P(t_i,C_j) = P(t_i)\times P(C_j)$，那么 $I(t_i, c_j) = 0$。为了选出对多类文档识别有用的特征，有最大值方法和平均值方法两种方法：

$$I_{Max}(t_i) = Max_{j=1}^{M}\left[P(C_j)\times I(t_i, c_j)\right]$$

$$I_{AVG}(t_i) = \sum_{j=1}^{M} P(C_j)I(t_i, c_j)$$

3.2.1.2 特征权重算法—TF-IDF 权重法

不同的特征项对文本的重要程度和区分度是不同的，所以在对文本分类模型进行形式化的时候，需要对所有特征项进行赋权重处理，根据群众留言的特点，本文将采用 TF-IDF 权重法。

TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

(1) TF 是词频(Term Frequency)[4]

这个数字通常会被归一化(一般是词频除以文章总词数)， 以防止它偏向长的文件。公式：

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

即

$$TF_w = \frac{\text{在某一类中词条 w 出现的次数}}{\text{该类中所有的词条数目}}$$

其中 $n_{i,j}$ 是该词在文件中出现的次数，分母则是文件中所有词汇出现的次数总和；

IDF 是逆向文件频率(Inverse Document Frequency)

某一特定词语的 IDF 可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含词条的文档越少， IDF 越大，则说明词条具有很

好的类别区分能力。公式：

$$\mathrm{idf_i} = \log\frac{|D|}{|\{j:t_i \in d_j\}|}$$

即

$$\mathrm{IDF} = \log\frac{\text{语料库的文档总数}}{\text{包含词条 w 的文档数}+1}$$

分母之所以加 1 是为了避免分母为 0。其中，$|D|$是语料库中的文件总数，$|\{j:t_i \in d_j\}|$表示包含词语t_i的文件数目(即 $n_{i,j} \neq 0$ 的文件数目）。

知道了“词频”（TF）和“逆向文件频率”（IDF）之后，将这两个值相乘，就得到了一个词的 TF-IDF 值，

即 TF-IDF=词频(TF)×逆文档频率(IDF)

可以看到，TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。所以，自动提取关键词的算法就很清楚了，就是计算出附件 2 中已经经过预处理的分词的 TF-IDF 值，然后按降序排列，取排在最前面的几个词。某个词对文章的重要性越高，它的 TF-IDF 值就越大。所以，排在最前面的几个词，就是这篇文章的关键词。

结巴分词可以实现 TF-IDF 算法

```
import jieba.analyse

text='关键词是能够表达文档中心内容的词语，常用于计算机系统标引论文内容特征、
信息检索、系统汇集以供读者检阅。关键词提取是文本挖掘领域的一个分支，是文本检索、
文档比较、摘要生成、文档分类和聚类等文本挖掘研究的基础性工作'

keywords=jieba.analyse.extract_tags(text, topK=5, withWeight=False, allowPOS=())
print(keywords)
```

运行结果：

```
['文档', '文本', '关键词', '挖掘', '文本检索']

Process finished with exit code 0
```

图 3结巴分词实现 TF-IDF

3.2.2 文本表示

空间向量模型需要一个“字典”：文本的样本集中特征词集合，这个字典可以在样本集中产生，也可以从外部导入。

有了字典后便可以表示出某个文本。先定义一个与字典长度相同的向量，向量中的每个位置对应字典中的相应位置的单词，,对应文本中的出现某个单词,在向量中的对应位置,填入“某个值”，实际上填入的“某个值”,就是特征词的权重(Tem Weight)。例如：[1,1,1,1,1,0,0]。

3.2.3 分类器-贝叶斯分类器[5]

贝叶斯分类器是各种分类器中分类错误概率最小或者在预先给定代价的情况下平均风险最小的分类器。它的设计方法是一种最基本的统计分类方法。其分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

贝叶斯定理：贝叶斯公式是建立在条件概率的基础上寻找事件发生的原因即大事件 A 已经发生的条件下，分割中的小事件B_i的概率。设B_1，B_2…是样本空间Ω的一个划分，则对任一事件 A（P(A)>0），有贝叶斯定理：

$$P(B_i|A)=\frac{P(A|B_i)}{P(A)}=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)}$$

对于每个特征 x，我们想要知道样本在这个特性 x 下属于哪个类别 c，即求后验概率 P(c|x)最大的类标记。这样基于贝叶斯公式，可以得到：

$$P(c|x)=\frac{P(x,c)}{P(x)}=\frac{P(c)P(x|c)}{P(x)}$$

即

$$P(\text{类别}|\text{特征})=\frac{P(\text{特征}|\text{类别})}{P(\text{特征})}.$$

3.3 实验结果

3.3.1 统计词频

3.3.2 生成词云

我们通过 jieba 分词与词频统计得到附件 2 中关于“留言详情”的词云，并通过附件 1 和附件 2 文本匹配，将关于一级分类的词云分别词频出来，具体结果如下图：

```
In [18]: def my_word_cloud(data=None, stopWords=None, img=None):
             dataCut = data.apply(jieba.lcut)  # 分词
             dataAfter = dataCut.apply(lambda x: [i for i in x if i not in stopWords])  # 去除
             wc = WordCloud(font_path=r'C:\Windows\Fonts\HYGangYiTi-65W',background_color='white'
             wordFre = pd.Series(_flatten(list(dataAfter))).value_counts()  # 统计词频
             mask = plt.imread(img)
             wc.fit_words(wordFre)
             plt.imshow(wc)  # 词云
             plt.axis('off') #关闭坐标

In [19]: d_cloud(data=df['留言详情'], stopWords=stopWords, img='aixin.jpg')  # 群众问政留言的词z
```

图 4留言详情词云

以下分别是关于留言详情中所有一级分类便签、以及各个一级分类便签的词云：

图 5留言详情

图 6城乡建设

图 7环境保护

图 8卫生计生

图 9交通运输图

图 10教育文体

图 11劳动和社会保障

图 12商贸旅游

4 任务二：热点问题挖掘

4.1 文本表示模型：

话题发现阶段的前提是如何让计算机处理中文文本，即将群众问政留言的问题和内容转化为计算机可以处理的形式。因此，为了方便操作处理，需要选取合适的结构化表示模型。还有一个关键原因就是便于处理和识别的模型可以

提高话题发现算法的效率，因此模型的选取是一个很重要的环节。目前常见的文本分类模型有基于规则的模型、基于概率的模型、基于几何的模型以及基于统计的模型，其中概率主题模型是最常用的文本表示模型，常用模型如下：潜在语义分析模型、概率隐形语义分析模型、LDA 模型、词对主题模型、Word2vec 模型。本文主要采用词对主题模型解决。[6]

4.1.1 分词

导入了数据准备中建立的“停用词词库”和“地方名词典”后，就开始对留言进行分词，最后把结果写入到表 1 的 E 列（地点/人群）

4.1.2 词语权重模型

词语权重使用差异系数来表示，计算公式为：

$$v(w)=\frac{\sqrt{\frac{\sum_{d=1}^{D}(f_d-\overline{f})^2}{Num}}}{\overline{f}}$$

使用 python 编写的差异系数算法进行权重的计算，最终得到留言词语的权重。

4.2 主题提取

计算留言中每个词语的权重之后，就可以进行特征的提取，开始挖掘留言的本题。本文采用改进后的 BTM-Master 程序，在其原来的基础上加上每个词语的权重。

用 python 编写程序，程序的参数设置如下：

参数	说明
K	留言集主题个数，根据实际情况确定，通常根据经验确定，可以进行实验确定。
W	留言集中词语的数量
alpha	即 α，P(z)的狄利克雷先验参数，通常 α=50/K
beta	即 β，P(w|z)的狄利克雷先验参数，通常 β=0.01
niter	吉布斯抽样的迭代次数，niter=1000
weigh_dir	留言集中词语的差异系数，即词语的权重
save_step	保存结果的阶段数，save_step=100
docs_pt	训练集所在的路径

model_dir	输出结果所在的目录

表 1

为了使实验取得更好的效果，需要确定合适的主题个数。先将主题个数 K 设为 9-23，间隔为 1。通过文本 F1 值来选定最合适的主题个数。综合评价 F1 值将分类后的召回率和准确率综合起来，弥补了召回率和准确率的单一性。F1 值得计算公式如下：

$$recall = \frac{TP}{TP + FN}$$

$$percision = \frac{TP}{TP + FP}$$

$$MicroF1 = \frac{2 \times precision \times recall}{precision + recell}$$

$$MicroF1 = \frac{\sum_{i=1}^{|c|} MicroF1}{|c|}$$

其中，TP(true positive)表示为实际为某个类别也被判断为该类别的样本个数，FP(false positive)表示实际不是某一类别但被错误的判断为该类别的样本个数， FN(false negative)表示实际为该类别但被错误的判断为某一类别的样本个数。recall 是召回率，代表正确分类样本数与该类别样本总数的比值，反映了模型分类结果的完备程度。precision 是准确率，代表正确分为该类的样本数与分为该类的样本数的比值，体现了模型分类结果的准确性。MacroF1 的值越大，说明该参数设定下分类的效果越好。不同主题个数下的 MacroF1 值如下图所示。从图中可以看出，随着主题个数的增加，MacroF1 值逐步增加，当主题个数为 15 时，MacroF1 值最大，之后随着主题个数的继续增大，MacroF1 值小范围波动后逐步下降。因此，最终选定主题个数 K=15，之后的实验对比等 W-BTM 模型均以此为条件。

图 13

参数设置完成后，程序开始自动运算。运算结果输出 p(z|d)，p(z|d)包含

一个 K N 的矩阵，即文档-主题矩阵，为每个留言的主题概率，选取概率最大的为对应留言的主题

整个主题挖掘过程中，输入文档格式如图 4-2 所示，其中每一行那个代表一个文本，也就是一条语料。输出文档格式如图 4-3 所示[7]。

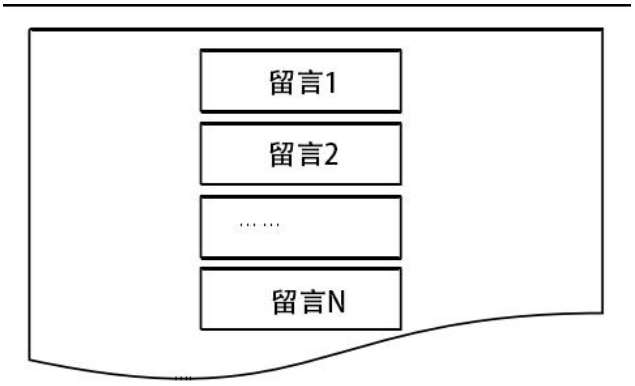


图 1 4 W-BTM 模型输入文档模式

	主题1	主题2	主题3	……
留言1	D1TI	D1T2	D1T3	……
留言2	D2TI	D2T2	D2T3	……
……	D3TI	D3T2	D3T3	……
留言N	DNTI	DNT2	DNT3	……

图 1 5 W-BTM 模型输出文档模式

4.3 热点问题挖掘

聚类算法是研究分类问题的一种统计分析方法，近年来，聚类算法在自然语言处理文本分析领域的研究越发广泛。学者将传统聚类算法进行改进，以使相关聚类算法能在自然语言处理领域取得更为显著的效果。所谓聚类，它是按照某个规定的标准把一个数据集分割成不同的类簇，达到簇内相似，簇间分离的目的。本文采用基于划分的聚类对留言的主题进行归类，并挖掘出热点问题。该类聚类算法的关键主要是 K 分组的构造，聚类在反复迭代中根据预先给定的 K 值将文本数据分成不同的分类，最后得到较好的聚类中心和聚类结果[6]。下面采用 K-means 算法对留言进行处理：

①. 随机选择 K 个对象，每个对象初始的代表了一个簇的中心；

②. 对除这 K 个对象以外的剩余其它对象，计算其与各个簇心的距离，将它分到距离最近的类簇中；

③. 再次计算每个簇的平均值，更新聚类中心；

④. 不断重复步骤 2、3，当目标函数收敛或达到最大迭代次数输出聚类结果，即可完成对留言主题的划分。

⑤. 最后统计每个聚类包含的留言数，输出包含留言数最多的 5 个聚类到表 1，即为前五个热点问题

4.4 操作结果：

4.4.1 热点问题的挖掘

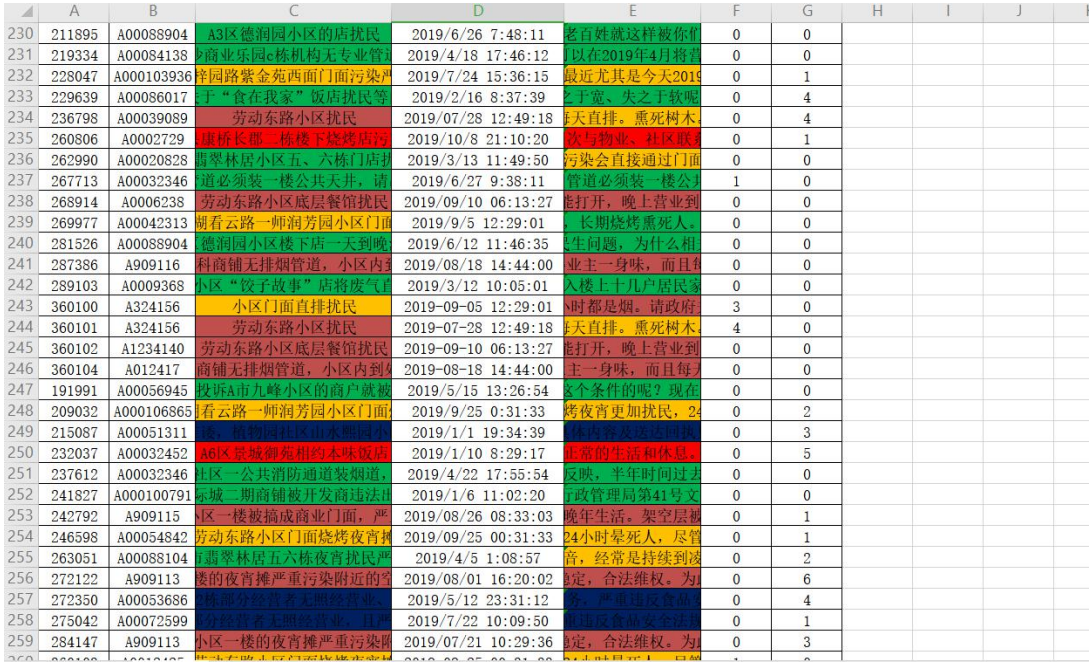
排名前五的热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	896	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	730	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
3	3	552	2019/07/11 至 2018/09/01	A 市伊景园滨河苑	广铁集团在伊景园滨河苑项目中捆绑销售车位
4	4	446	2019/01/06 至 2019/12/19	A 市	社保缴纳方式繁琐
5	5	310	2019/05/24 至 2019/11/02	A2 区	交通拥挤，道路不合理

表 3 热点问题表

4.4.2 同一问题的筛选

对于热点问题 1——小区临街餐饮店油烟噪音扰民，通过词频出现概率的大小我们发现，该问题中，A 市、油烟、扰民、A5 区、魅力之城、餐饮等词的同时出现的频率较大，接下来就借用这些词进行同一问题留言的筛选，此处我们运用了 Excel 对数据进行查找标记，找到了 307 条相关的留言，再通过多个词同时出现来判别与主题关联最大的留言，与此同时删除了部分重复的数据，最终得到了 5 条关于该热点问题的留言，部分数据截图如下图所示：



	A	B	C	D	E	F	G	H	I	J
230	211895	A00088904	A3区德润园小区的店扰民	2019/6/26 7:48:11	老百姓就这样被你们	0	0			
231	219334	A00084138	商业乐园c栋机构无专业管理	2019/4/18 17:46:12	以在2019年4月将营	0	0			
232	228047	A000103936	锋园路紫金苑西面门面污染严	2019/7/24 15:36:15	最近尤其是今天2019	0	1			
233	229639	A00086017	于“食在我家”饭店扰民等	2019/2/16 8:37:39	之于宽、失之于软呢	0	4			
234	236798	A00039089	劳动东路小区扰民	2019/07/28 12:49:18	天直排。熏死树木	0	4			
235	260806	A0002729	康桥长廊一栋楼下烧烤店污	2019/10/8 21:10:20	次与物业、社区联	0	1			
236	262990	A00020828	翡翠林居小区五、六栋门店扰	2019/3/13 11:49:50	污染会直接通过门面	0	0			
237	267713	A00032346	道必须装一楼公共天井，请	2019/6/27 9:38:11	管道必须装一楼公共	1	0			
238	268914	A0006238	劳动东路小区底层餐馆扰民	2019/09/10 06:13:27	能打开，晚上营业到	0	0			
239	269977	A00042313	湖看云路一师润芳园小区门面	2019/9/5 12:29:01	长期烧烤熏死人。	0	0			
240	281526	A00088904	德润园小区楼下店一天到晚	2019/6/12 11:46:35	生问题，为什么相	0	0			
241	287386	A909116	科商铺无排烟管道，小区内到	2019/08/18 14:44:00	业主一身味，而且每	0	0			
242	289103	A0009368	小区“饺子故事”店将废气直	2019/3/12 10:05:01	入楼上十几户居民家	0	0			
243	360100	A324156	小区门面直排扰民	2019-09-05 12:29:01	时都是烟。请政府	3	0			
244	360101	A324156	劳动东路小区扰民	2019-07-28 12:49:18	天直排。熏死树木	4	0			
245	360102	A1234140	劳动东路小区底层餐馆扰民	2019-09-10 06:13:27	能打开，晚上营业到	0	0			
246	360104	A012417	商铺无排烟管道，小区内到处	2019-08-18 14:44:00	主一身味，而且每天	0	0			
247	191991	A00056945	投诉A市九峰小区的商户就被	2019/5/15 13:26:54	这个条件的呢？现在	0	0			
248	209032	A000106865	看云路一师润芳园小区门面	2019/9/25 0:31:33	烤夜宵更加扰民，24	0	2			
249	215087	A00051311	楼、植物园社区山水畔园小	2019/1/1 19:34:39	体内容及送达回执	0	3			
250	232037	A00032452	A6区景城御苑柑约本味饭店	2019/1/10 8:29:17	正常的生活和休息。	0	5			
251	237612	A00032346	社区一公共消防通道装烟道，	2019/4/22 17:55:54	反映，半年时间过去	0	0			
252	241827	A000100791	东城二期商铺被开发商违法出	2019/1/6 11:02:20	行政管理局第41号文	0	0			
253	242792	A909115	A区一楼被搞成商业门面，严	2019/08/26 08:33:03	晚年生活。架空层被	0	1			
254	246598	A00054842	劳动东路小区门面烧烤夜宵扰	2019/09/25 00:31:33	24小时晕死人，尽管	0	1			
255	263051	A00088104	翡翠林居五六栋夜宵扰民严	2019/4/5 1:08:57	音，经常是持续到凌	0	2			
256	272122	A909113	楼的夜宵摊严重污染附近的空	2019/08/01 16:20:02	定，合法维权。为	0	6			
257	272350	A00053686	栋部分经营者无照经营业，	2019/5/12 23:31:12	务，严重违反食品	0	4			
258	275042	A00072599	分经营者无照经营业，且再	2019/7/22 10:09:50	重违反食品安全法	0	1			
259	284147	A909113	小区一楼的夜宵摊严重污染附	2019/07/21 10:29:36	定，合法维权。为	0	3			

图 16 处理过程

问题ID为1的留言明细表

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
360104	A012417	A市魅力之城商铺无排烟管道，小区内到处油烟味	2019/8/18 14:44	导致大量排入小区内，每天进出都搞得业主	0	0
360105	A120356	区魅力之城小区一楼被搞成商业门面，噪音扰民严	2019-08-26 08:33:03	北大门两侧的楼栋下面一楼，本来应是架空	1	0
360106	A235367	魅力之城小区底层商铺营业到凌晨，各种噪音好烦	2019-08-26 01:50:38	商铺越发嚣张，不仅营业到凌晨不休息，各	0	0
360107	A0283523	动东路魅力之城小区一楼的夜宵摊严重污染附近	2019-07-21 10:29:36	各小区的一楼，开了几家夜宵快餐店，厨房	3	0
360109	A0080252	科魅力之城小区底层门店深夜经营，各种噪音扰	2019-09-04 21:00:18	您好：我是万科小区的业主，小区的一楼	0	0

图 17 问题 ID 为 1 的留言明细

最终得到的问题 ID 为 1 的留言明细表如下：

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	360104	A012417	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/8/18 14:44	A 市小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量排入小区内，每天进出都搞得业主一身味，而且每天到凌晨还在营业，烧烤店烧烤吆喝……	0	0
1	360105	A120356	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019-08-26 08:33:03	我们是小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面……	1	0
1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种	2019-08-26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音……	0	0

			噪音好痛苦				
1	360107	A0283523	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	2019-07-21 10:29:36	局长：　　你好，劳动东路小区的一楼，开了几家夜宵快餐店，厨房内多个灶台随意排放，……	3	0
1	360109	A0080252	万科魅力之城小区底层门店深夜经营，各种噪音扰民	2019-09-04 21:00:18	您好：我是万科小区的业主，小区的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业……	0	0

表 4--热点问题 1 的留言明细表

对热点问题 2——学校强制学生去定点企业实习，在该问题中，出现频率最多的词语为“A 市”、“经济学院”、“学校”、“实习”、“强制”等，我们采用与上个问题相同的方法进行挖掘，最终得到关于该问题的留言，过程部分截图如下：

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
211395	A00050903	西地省财政校园宽带被垄断	2019/9/14 17:57:34	卡都有免费赠送学	0	0
233759	A909118	A市涉外学生	2019/04/28 17:32:51	须去学校安排的几个	0	0
235521	A0006920	区枫林三路涉外外街理发店扌	2019/10/15 18:59:08	上8.30左右至晚上	0	0
240721	A00050903	西地省财政涉嫌宽带垄断	2019/9/14 17:56:17	卡都有免费赠送学	0	0
242062	A00028889	地省涉外变相学生“社会实践	2019/11/27 23:14:33	是从学生提升专业	0	0
264084	A00074365	以报名名额已满拒绝让学生	2019/3/19 23:11:44	加考试的人员估计超	0	0
266368	A00038920	寒假过年期间组织学生去工	2019/11/22 14:42:14	虽说不是性的，但	0	0
360110	A110021	寒假过年期间组织学生去工厂	2019-11-22 14:42:14	虽说不是性的，但不	0	0
360112	A220235	A市学生	2019-04-28 17:32:51	去学校安排的几个点	0	0
360113	A3352352	A市学生外出	2018-05-17 08:32:04	校很小但是这几年	3	0
360114	A0182491	A市体育学院变相	2017-06-08 17:31:20	并且公司也要和我	9	0
195917	A909119	涉外组织学生外出打工合理吗	2019/11/05 10:31:38	个小时以上，（晚班	0	1
211800	A00046925	堂只开放十余个窗口，有些	2019/2/25 23:24:54	也因为卖完了而吃不	0	0
360111	A1204455	市组织学生外出打工合理吗?	2019-11-05 10:31:38	个小时以上，（晚班	1	0

图 18 处理过程

问题ID为2的留言明细表

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
360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42	关于西地省A市经济学院寒假过年期间组	0	0
360111	A1204455	A市经济学院组织学生外出打工合理吗?	2019-11-05 10:31:38	一名中职院校的学生,学校组织我们学生	1	0
360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	各位领导干部大家好，我是A市经济学院	0	0
360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	A市经济学院强制16届电子商务跟企业物	3	0
360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	书记您好，我是来自西地省经济学院体育	9	0

图 19 问题 ID 为 2 的留言明细

最终得到的问题 ID 为 1 的留言明细表如下：

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
2	360110	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42	关于西地省 A 市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面……	0	0
2	360111	A1204455	A 市经济学院组织学生外出打工合理吗？	2019-11-05 10:31:38	一名中职院校的学生,学校组织我们学生在外边打工,在外省做流水线工作，还要倒白夜班。本来都在学校好好上课，十月底突然说组织外出外省……	1	0
2	360112	A220235	A 市经济学院强制学生实习	2019-04-28 17:32:51	各位领导干部大家好，我是 A 市经济学院的一名学生，临近毕业，学校开始组织学生参加实习，……	0	0
2	360113	A3352352	A 市经济学院强制学生外出实习	2018-05-17 08:32:04	A 市经济学院强制 16 届电子商务跟企业物流专业实习。其中我们企业物流专业实习 6 个月，去江苏。暑假……	3	0
2	360114	A0182491	A 市经济学院体育学院变相强制实习	2017-06-08 17:31:20	书记您好，我是来自西地省经济学院体育学院的一名即将大四的学生，系里要求我们在实习前分别去指定……	9	0

表 5--热点问题 ID 为 2 的留言明细

通过此类方法，进而得到了热点问题排名前五的热点问题明细表，并存放于“表 2-热点问题明细表”中。

5 任务三：答复意见质量评价方案

5.1 答复意见质量总则

答复意见质量的总体评价原则，是以用户为导向，做到让用户满意。

5.2 答复意见质量细则[8]

答复意见质量必须做到：从提交者、提交部门看来，答复结论是及时的、可用的。及时是指相关部门答复群众的时间快速，而可用是指答复内容足够清

晰且指导性强，该答复能够为群众解决问题、满足群众的要求。为满足答复结论的及时、可用，给出评价细则如下：

5.2.1 及时性：

衡量答复意见及时性，在制度上，通常定义为确定的时间段要求，如八个工作日；而在业务上，则是重点满足群众答复时间要求。

5.2.2 相关性：

答复意见质量的关键点在于是否接纳。为了保证答复意见的相关性，答复责任主体应严格执行下列原则：

①. 强化团队运作机制

②. 规范执行问题分析活动，确保意见答复是跟问题是紧密相关的，并进行科学的优先级排序；

衡量答复意见的相关性，可以从两方面来审视：从过程来看，意见是否经过严格的分析和集体决策，所做出的结论是否有充分的理由说服群众；从结果来看，结论是否实现了群众利益最大化，提高了社区的管理水平和条件保障。

5.2.3 完整性

答复的结论除了接纳与否外，还包括更多的要素，才能构成完整的答复。例如对需求优先级的排序，此外，拒绝的需求还应当说明相关的决策依据，而对于接纳的需求，则需要重点说明给出的解决方案。对于拒绝的需求，如果未能给出清晰的理由，则可能导致提交人的争议，影响社区和谐。对于接纳的需求，如果未能描述解决方案，则可能导致居民群众的不满。

衡量答复意见的完整性，就在于除接纳与否外，是否给出了相关的所有应当给出的支持信息，包括相关要素，以及上述的解决方案。

5.2.4 可解释性

答复意见的可解释性，表现在结论的发布渠道上，不同的渠道应保持一致性，也包括结论不能随着时间随意发生变化。（结论的变更必须经过变更流程后才能进行更改）。具体表现在：答复意见的正式发布渠道是通过特定平台回复用户的问题需求，对于热点问题的答复意见，还需要在问政平台向群众公布。相关部门的不同工作人员，在回复答复意见时，也应该保持答复格式和结论一致，而不能随意修改或者给出其他不同的结论。答复意见在发送、实现的过程中，所有人都不能对结论随意进行修改，只能通过严格的变更流程来进行更改。

衡量结论的可解释性，主要通过反向追溯方式，即审视需求的结论是否在不同渠道、不同时间发生了未经受控的更改。需求结论的可解释性，体现了问政平台及相关部门管理体系运作的规范程度。

5.3 答复意见质量评价方案

通过对附件 4 内容的量化分析，我们从答复意见的相关性、完整性、可解释性等角度建立了对答复意见的质量的四类评价指标，分别为服务能力、服务质量、管理能力、条件保障，每一项各 25 分，满分为 100 分，具体评价方案如下表：

指标	分值	评价内容
服务能力	25	1.深入了解群众的需求，竭尽所能为群众“办好事” 2.为社会公众、企业、政府及非盈利组织提供服务 3.评估标准基本固定，评估结果较准确 4.各部门积极配合，逐步提高办事效率，提高群众对电子政务的认知度和利用率
服务态度	25	1.始终遵循“以公民为中心”的服务理念，加强互动沟通，开展社会监督，规范服务投诉 2.认真对待群众所反馈的问题，及时给出处理意见、结果 3.群众对电子政务的满意度有所提高
管理能力	25	1.部门间信息共享水平、协同办公能力较高，能在最快时间内解决问题并给予答复意见 2.不断完善管理体系，对政府工作流程进行优化和改造，以标准化服务的方式实现各类跨部门的联动业务，提高政府办事效率 3.立法完善，有明确的规章制度并且严格执行
条件保障	25	1. 信息系统功能完整，使用方便、完全 2.初步建成信息共享和数据交换平台，实现政务智慧化 3.信息基础设施基本完善，软硬件投资逐步增加 4.通过开放 API 接口，完成各系统之间的数据交换调用，达到实现一站式服务的目的
总计	100	

表 6 答复意见质量评价方案

参考文献

[1] 陶瑞同学.Python 中文文本分析实战：jieba 分词+自定义词典补充+停用词词库补充+词频统计[EB/OL]. 2018.

[2] 徐国海.文本表示简介.研究方向:自然语言处理和知识图谱.2018 https://blog.csdn.net/sigai_csdn/java/article/details/81870381

[3] zsffuture. NLP ---文本分类(向量空间模型(Vector Space Model)VSM)).2018. https://blog.csdn.net/weixin_42398658/java/article/details/85063004

[4] Shendu.CC.文本分类学习（三） 特征权重（TF/IDF）和特征提取. https://www.cnblogs.com/dacc123/p/8707638.html

[5] 陈小虾.贝叶斯分类器详解.2019 https://blog.csdn.net/ch18328071580/java/article/details/94407134

[6] 董静. 基于主题模型和聚类算法的网络热点话题发现[DB/OL]. 河北大学，2019.

[7] 张雅君. 基于 W-BTM 的短文本主题挖掘及文本分类应用[DB/OL]. 山西财经大学，2017.

[8] 需求答复质量评判标准 http://www.szgcl.com.cn/sv.aspx?TypeId=426&FId=t8:426: