“数据政务”中的文本挖掘应用

摘 要

近年来，网络问政平台逐渐成为政府了解民意、汇聚民智、凝聚民气的重要渠道。由于各类社情民意的数据量不断攀升，以往依靠人工来进行留言划分的方法已经无法及时的完成任务。随着大数据、人工智能、云计算等的迅速发展，建立处理自然语言处理技术的智慧政务系统已经成为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。

针对上述问题，本文主要运用自然语言处理和文本数据挖掘的方法来解决此类问题。

针对问题一，本文首先运用 Python 对文本数据进行预处理，提取附件 2 中留言详情的文本，删除文本中的停用词，再进行分词，最后绘出词云图，并生成高频词的直方图。其次建立 KNN 最邻近分类算法，对附件 2 中的数据进行标签分类，最后运用 F-Score 对 KNN 最邻近分类算法进行评价，得出此模型适合于对文本数据进行文本挖掘及分类并且分类效果较好的结论。

针对问题二，本文参考问题一的模型对附件 3 中的留言进行归类，然后运用因子分析的方法定义出某一问题的反对数、点赞数以及某一问题出现的次数作为热度评价指标对附件 3 中的数据进行评价，最后得出了人身安全、教育、纪检监察、政法和基建 5 个热点问题，并保存了相关明细表。

针对问题三，本文从答复的相关性、完整性和可解释性的角度对答复意见的质量给出了一套评价方案，并且给出了几条建议：

（1）希望政府部门能够将解决方案落实到实处，而不仅仅只是空谈。

（2）提供相应的咨询电话或服务平台，以供群众对问题的解决有深入的了解，方便群众跟进问题的解决进度。

（3）希望政府部门无论群众提出的事宜大小都能得到妥善解决。

（4）希望政府能够将问题解决的进度及时反馈给群众。

（5）希望政府能够针对群众提出的问题及时给出解决方案并付诸行动。

本文主要运用 Python 对文本数据进行数据挖掘处理，给出了对政府答复意见的质量的评价方案，还提出了一些建议，这些建议在实际运用中具有一定的价值。

关键词：Python KNN 预处理 归类

Abstract

In recent years, Wechat, micro-blog, mayor's mailbox and other online platforms have gradually become an important channel for the government to understand public opinion, pool people's wisdom and gather people's morale. Due to the increasing amount of data on social sentiment and public opinion, the former method of message division by hand can not finish the task in time. With the rapid development of big data, artificial intelligence and cloud computing, it has become a new trend of social governance to build intelligent government system that deals with natural language processing technology, it can promote the management level and administration efficiency of the government greatly. To solve these problems, this paper mainly uses the methods of natural language processing and text data mining to solve these problems. To solve the first problem, this paper first preprocesses the text data with Python, then establishes KNN nearest neighbor classification algorithm to label the data in Annex 2, and finally evaluates KNN nearest neighbor classification algorithm with F-Score, it is concluded that the model is suitable for text data mining and classification and the classification effect is good. In response to question 2, this paper classifies the comments in Annex 3 with reference to the model in question 1, then, the data in Annex 3 are evaluated by using the method of factor analysis to define the number of objections, the number of likes and the number of times a question appears, finally, the paper draws 5 hot issues of personal safety, education, discipline inspection, political law and infrastructure construction, and keeps the relevant detailed list. In response to question three, this paper gives a set of evaluation program for the quality of the responses from the point of view of relevance, completeness and Interpretability of the responses, and give a few suggestions: (1) hope that the government departments will be able to implement the solution, not just empty talk. (2) to provide consultation telephone or service platform for the masses to have in-depth understanding of the problem solving process, so as to facilitate the masses to follow up the progress of problem solving. (3) we hope that all matters raised by the public, irrespective of their size, will be properly resolved. (4) it is hoped that the government will be able to provide timely feedback to the public on the progress made in solving the problems. (5) it is hoped that the government will provide timely solutions to the problems raised by the masses and put them into action. This paper mainly uses Python to mine the text data, gives the evaluation scheme of the quality of the government's reply, and puts forward some suggestions, which have certain value in the practical application.

Key words : Python KNN Pretreatment Categorize

目 录

1.挖掘目标

1.1 挖掘背景

近年来，随着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 挖掘目标

根据附件给出的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。利用自然语言处理和文本挖掘的方法解决问题。

首先对数据进行处理，建立关于留言内容的以及标签分类模型对数据进行分类处理，并使用 F-Score 对分类方法进行评价。其次对群众反应的热点问题进行有针对性的处理，定义合理的热度评价指标，给出评价结果，并将结果保存。最后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。

2.分析方法与过程

本文首先对问题进行了分析，基于对问题的分析对数据进行预处理，对数据进行预处理之后建立了最适合的模型，最后运用 Python 对模型进行求解。

2.1 问题一的方法与过程

图 1 问题一的求解流程图

2.1.1 问题一的分析

问题一要求建立合适的一级标签分类模型对群众留言进行分类，之后运用 F-Score 对分类方法进行评价。针对问题一，本文首先对附件二中的数据运用 Python 进行数据预处理，之后建立了 KNN 最邻近分类算法对附件二中的数据进行标签分类，并在文章的最后给出模型求解的结果。

2.1.2 问题一的数据预处理

a) 读取 Excel 中的数据

```
import xlrd
def extract ( inpath ):
 data = xlrd.open_workbook ( inpath , encoding_override ='utf-8')
 tab1e = data .sheets ()[0]
 nrows =tab1e.nrows
 nco1s = tab1e.nco1s
 for i in range (1, nrows ):
  a11data=tab1e.col_va1ues( i )
 resu1t=a11data[5]
 print (result)
 inpath =('留言.xlsx')
 extract(inpath)
```

图 2 读取 Excel 数据代码分析

运用上述代码对附件 2 中的文本数据进行提取。

b) 文本处理

```
cleanedList = [x for x in user_labels_list if str(x) != 'nan']  # 去掉空值
mytext = ''.join(cleanedList)  # 把列表变成字符串
mytext = mytext.replace(";", " ")  # 把字符串中的分号;
stopwords=[line.strip()for line in open(r'C:\Users\admin\Documents\Tencent Files\216401003
```

图 3 文本处理代码分析

c) 删除停用词

```
data1=data_file.iat[i,4]
data2=jieba.cut_for_search(data1)
for wors not in stopwords:
    if wors not in stopwords:
        if len(wors)!=1:
            result.append(wors)
            filename='taidi.txt'
            with open(filename, 'w',encoding='utf-8') as file_object:
                file_object.write(str(result))
                file_object.close()

counts={}#建一个空字典
for i in result:
    if len(i)==1:
        continue
    else:
        counts[i] = counts.get(i,0)+ 1# 在字典中创建或者数量加一
items=list(counts.items())
items.sort(key=lambda x:x[1],reverse=True)
print(items)
```

图 4 删除停用词代码分析

运用上述代码对附件 2 中的读取出来的数据做删除停用词处理。

d） 分词

```
count=0#规定行数为0
workbook1= xlwt.Workbook()#调用xlwt库
sheet=workbook1.add_sheet("城乡建设")#建一个sheet1
for i in range(0,100):
        sheet.write(count,0,items[i][0])
        sheet.write(count,1,items[i][1])
        count=count+1#换行
workbook1.save('城乡建设.xls')
```

图 5 分词代码分析

运用上述代码对附件 2 中的数据进行分词处理。

e） 绘制词云图

```
text=open('taidi.txt','r',encoding='utf-8').read()
f=jieba.lcut(text)
newf=' '.join(f)
w=wordcloud.WordCloud(font_path='msyh.ttc',width=1000,height=700,background_color='white').generate(newf)
w.to_file('城乡建设.jpg')
```

图 6 绘制词云图代码分析

运用上述代码针对附件 2 中的数据做词云图。

2.1.3 问题一模型的建立

针对问题一，本文采用最邻近分类算法对数据进行分类。本文由 K-Means 分类得到聚类中心，利用 KNN 算法找出与各中心相似的元素，根据个数多的判定所属类别，根据向量空间模型，将每一类别文本训练后得到该类别的中心向量记为$C_j(W_1, W_2, \dots\dots, W_n)$，将待分类文本 T 表示成 n 维向量的形式 T$(W_1, W_2, \dots\dots, W_n)$则文本内容被形式化为特征空间中的加权持征向量，即 D=D$(T_1, W_1; T_2, W_2; \dots\dots T_n, W_n)$。对于一个测试文本，计算它与训练样本集中每个文本的相似度，找出 K 个最相似的文本，跟据加权距离和判断测试文本所属的类别，具体算法步骤如下：

（1）对于每一个测试文本，根据特征词形成测试文本向量。

（2）计算该测试文本与训练集中每个文本的文本相似度，计算公式为：

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{\sum_{k=1}^{M} W_{ik}^2} \sqrt{\sum_{k=1}^{M} W_{jk}^2}}$$

式中，d_i为测试文本的特征向量，d_j为 j 类的中心向量；M 为特征向量维数；W_k为向量的第 k 维。K 值的确定一般先采用一个初始值，然后根据实验测试 K 的结果来调整 K 值。

（3）按照文本相似度，在训练文本集中选出与测试文本最相似的 k 个文本。

（4）在测试文本的 k 个近邻中，以此计算每类的权重，计算公式如下：

$$P\ (X, C_j) = \begin{cases} 1, 若 \sum_{d \notin \mathrm{Knn}} \mathrm{Sim(x,d}_i) y(d_i, c_j) - b \geq 0 \\ 0, 其他 \end{cases}$$

式中，d_i为测试文本的特征向量；Sim(x,d_i)为相似度计算公式；b 为阈值，有待于优化选择；而$y(d_i, c_j)$的值为 1 或 0，如果d_i属于C_j，则函数值为 1，否则为 0.

（5）比较类的权重，将文本分到权重最大的那个类别中。

2.2 问题二的方法与过程

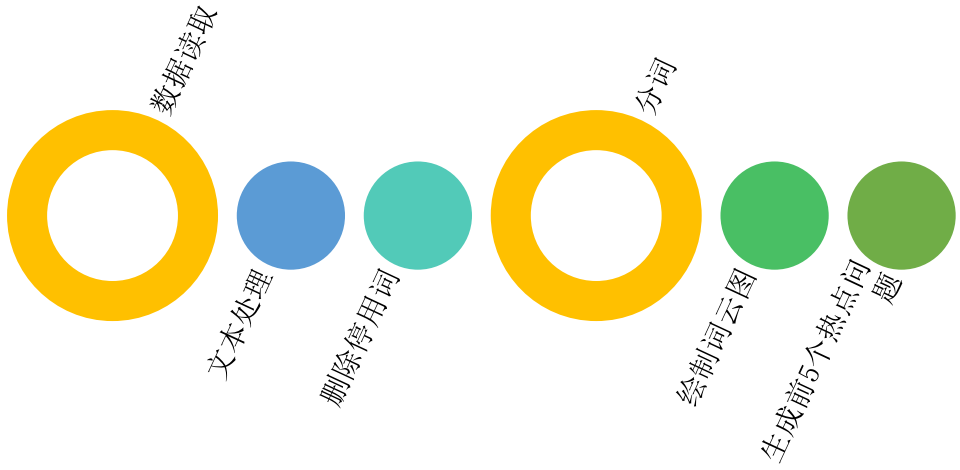

图 7 问题二流程图

2.2.1 问题二的分析

问题二要求将某一时段内反应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，给出评价结果，并保存对应的信息图表。针对问题二，本文首先根据问题一的模型对某一时段内反应特定地点或特定人群问题的留言进行归类，之后运用因子分析的方法定义并筛选出合理的热度评价指标，最后在文章的结尾给出评价结果以及相应的信息图表。

2.2.2 问题二的数据预处理

(1)标记化

标记化是指将文本中的长字符串分割成 tokens 或者小的片段的过程。大段文字可以被分割成句子，句子又可以被分割成单词等等。只有经过了 tokenization，才能对文本进行进一步的处理。Tokenization 同样被称作文本分割或者词法分析。有时，分割用来表示大段文字编程小片段的过程。而 tokenization 指的是将文本变为只用单词表示的过程。

(2)归一化

再进一步处理之前，文本需要进行归一化。归一化指的是一系列相关的任务，能够将所有文本放在同一水平区域上，将所有文本转化成同样的实例，删除标点，将数字转换成相应的文字等等。对文本进行归一化可以执行多种任务，但是对于

我们的框架，归一化有 3 个特殊的步骤：

a) 词干提取：词干提取是删除词缀的过程（包括前缀、后缀、中缀、环缀），从而得到单词的词干。

b) 词形还原：词形还原与词干提取相关，不同的是，词形还原能够捕捉基于词根的规范单词形式。

c) 其他：词形还原和词干提取是文本预处理的主要部分，所以这两项一定要认真对待。他们不是简单地文本操作，需要依赖语法规则和对规则细致的理解。

（3）噪声清除

噪声消除延续了框架的替代任务。虽然框架的前两个主要步骤（标记化和归一化）通常适用于几乎任何的文本或项目，噪声去除是预处理框架中一个更加具体的部分。

2.3 问题三的方法与过程

2.3.1 问题三的分析

问题三要求针对附件 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。针对问题四，本文从答复的相关性、完整性和可解释性的角度分析了相关部门对留言的答复意见，并且给出来一套评价方案。

3.结果分析

3.1 问题一的结果分析

运用 F-Score 对问题一的模型进行评价要用到“准确率”和“召回率”两个指标，这两个指标通常成反比的关系，及二者之间是此消彼长的。一般情况下用参数来控制二者之间的关系，通过修改参数来得到准确率和召回率的 ROC 曲线，这个曲线与 x 轴和 y 轴的面积就形成了 AUG。由于 AUG 的计算比较麻烦，所以运用 F-Score 来代替 AUG，用来评价问题一的模型。

运用 F-Score 对 KNN 最邻近分类算法进行评价的出的结论是：

（1） KNN 适合对文本数据进行一级标签分类。

（2） KNN 对文本数据进行一级标签分类的效果显著。

（3） 该算法比较适用于样本容量比较大的类域的自动分类。

（4） 该算法计算简单易懂，可以处理多分类问题

（5） 该算法对于类域的交叉或重叠较多的待分类样本集来说，该算法更为合适。

3.2 问题二的结果分析

问题二要求将某一时段内反应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，给出评价结果，并保存对应的信息图表。针对问题二，本文首先根据问题一的模型对某一时段内反应特定地点或特定人群问题的留言进行归类，之后运用因子分析的方法定义某问题的点赞数、反对数以及某问题出现的次数作为热度评价指标对热点问题进行评价，并得出图 8。

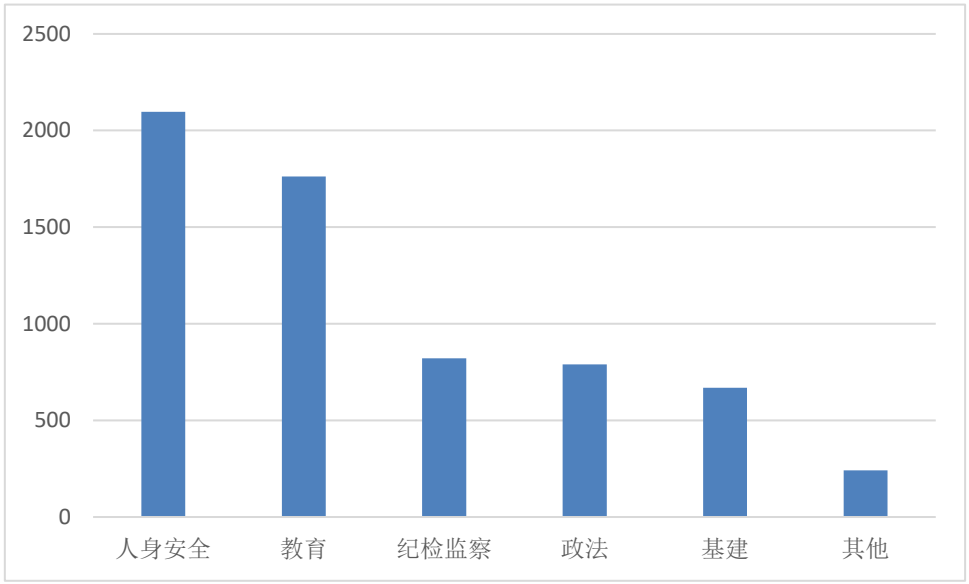

图 8 热点问题排行表

如图 8 所示，排名前 5 的热点问题分别为人身安全问题、教育问题、纪检监察问题、政法问题和基建问题。

根据上述结论，得出了“热点问题表.xls”和“热点问题留言明细表.xls”，如附件所示。

3.3 问题三的结果分析

本文针对相关部门对留言的答复意见，给出以下评价：

（1）从答复的相关性来看，相关部门分别对与本部门职责相对应的问题给出回应，在相关联问题上，将问题进行整合，并依照相关的法律条例，给出较为一致的解决方案。

（2）从答复的完整性来看，群众提出的每一个问题，相关部门都给予详细的回复并给出了恰当的解决方案。答复的内容不仅仅是简单的理论支撑，更是实实在在的解决方案，杜绝“只说不做”的情形出现。在解决问题的同时，对每一个问题提出者的关心、监督和支持表达了真挚的感谢。

（3）从答复的可解释性来看，各个相关部门所给出建议都有一定的法律或者道德条例做支撑，每一个答复的内容都有迹可循，不是无中生有，凭空捏造。在此基础上，某一类问题的答复还包括了相关的咨询电话，以供提出者或者他人

对此问题的解决有一个更好的了解。

针对上述评价，给出以下几条建议：

（1）希望政府部门能够将解决方案落实到实处，而不仅仅只是空谈。

（2）提供相应的咨询电话或服务平台，以供群众对问题的解决有深入的了解，方便群众跟进问题的解决进度。

（3）希望政府部门无论群众提出的事宜大小都能得到妥善解决。

（4）希望政府能够将问题解决的进度及时反馈给群众。

（5）希望政府能够针对群众提出的问题及时给出解决方案并付诸行动。

4.参考文献

[1]饶竹一,张云翔.基于 BiGRU 和注意力机制的多标签文本分类模型[J].现代计算机,2020(01):31-35.

[2]贾旭东,王莉.基于多头注意力胶囊网络的文本分类模型[J].清华大学学报(自然科学版),2020,60(05):415-421.

[3]靳果,朱清智,孟阳,闫奇.基于多层极限学习机的电能质量扰动多标签分类算法[J/OL].电力系统保护与控制:1-10[2020-05-05].

[4]梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.

[5]洪伟俊. 基于 F-score 与支持向量机的砂体静态连通性定量评价[C]. 西安石油大学、陕西省石油学会.2018 油气田勘探与开发国际会议（IFEDC 2018）论文集.西安石油大学、陕西省石油学会:西安华线网络信息服务有限公司,2018:2158-2164.

[6]翟东海,鱼江,高飞,于磊等.最大距离法选取初始簇中心的K_means 文本聚类算法的研究.西南交通大学 2014

[7]朱志远.基于数据挖掘的网络招聘系统是设计与实现.电子科技大学.硕士学位论文.2013

[8]玉千,王成,冯振元,叶金风.K-means 聚类算法研究综述.2012

[9]张晓辉，李莹，王华勇等.应用特征聚合进行中文文本分类的改进 KNN 算法.东北大学 2003

[10]卜凡军.KNN 算法的改进及其在文本分类中的应用.江南大学.硕士学位论文.2009

[11]曹卫峰.中文分词关键技术研究.南京理工大学.硕士学位论文.2009

[12]杨虎.面向海量短文文本去重技术的研究与实现.国防科学技术大学.2007