

C 题：“智慧政务”中的文本挖掘应用

摘 要：随着互联网的发展，消息的传播变得越来越快，越来越便捷。近年来，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。因此我们利用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进行相关数据的分析与挖掘。其主要内容包括：

对于问题一，利用现有的群众留言信息，将留言信息进行分类处理。首先是将数据进行数据清洗，剔除无用信息，再将信息内容进行标签分类。建立 k 近邻(k-Nearest Neighbor, kNN)模型对文本进行分类，最后利用 F-Score 对分类方法进行评价其有效性。

对于问题二，利用命名实体识别中的地点识别和相似度计算中的问题归类分贝对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，并给出排名前 5 的热点问题文件“热点问题表.xls”；给出相应热点问题对应的留言信息文件“热点问题留言明细表.xls”。

对于问题三，针对相关部门对留言的答复意见，首先要了解相关部门如何回答这个问题，回答的质量怎么样。因此先利用 jieba 库进行分词，然后提取关键词，找出相关部门对该类问题回答所使用的词如何，最后结合实际给出一套评价方案。

关键词：F-Score 数据清洗 命名实体识别 相似度计算 jieba 库

kNN 算法

C: Application of Text Mining in "Smart Government"

Absrtact: with the development of the Internet, the spread of news becomes faster and more convenient. In recent years, WeChat, Weibo, Mayor's mailbox, Sunshine Hotline and other online political platforms have gradually become an important channel for the government to understand public opinion, gather people's wisdom and gather people's spirit. The amount of text data related to all kinds of social sentiment and public opinion has been rising. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government system based on natural language processing technology has been a new trend of social governance innovation and development, which has a great role in promoting the management level and governance efficiency of the government. So we're taking advantage.

For question one, using the existing mass message information, the message information will be classified processing. The first is to clean the data, eliminate useless information, and then the information content label classification. Finally, the F-Score is used to evaluate the effectiveness of the classification method.

For the second question, the location identification in named entity identification and the problem classification in similarity calculation are used to classify the messages that reflect the specific location or the specific population problem in a certain period of time, define the reasonable heat evaluation index, and give the top 5 hot issue file "hot issue table ". xls "; give the corresponding hot issues corresponding message information file "hot issues message list. xls".

For question 3, in response to the response comments of the relevant departments to the message, first of all, to understand how the relevant departments to answer this question, how the quality of the answer. Firstly, we use the jieba library to use the word segmentation, then extract the key words,

find out how the relevant departments to answer this kind of questions, and finally give a set of evaluation scheme combined with the actual situation.

Key words: F-Score data-cleaning named entity recognition similarity calculation jieba library kNN algorithm

目 录

1. 挖掘目标.....	5
2. 分析方法与过程.....	5
2.1. 具体步骤.....	5
2.2. 对答复意见的评价方案.....	7
3. 结论.....	8
4. 参考文献.....	9

1. 挖掘目标

从互联网上收集而来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将群众留言进行级标签分类、热点问题挖掘并用 Excel 表呈现，最后结合所给数据就相关部门对留言的答复意见，从其答复的相关性、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。

2. 分析方法与过程

2.1. 具体步骤

步骤一：

1. 运用 Python 自然语言处理先将所得数据进行数据清洗，再将信息进行分类处理得出运行结果

2. 根据附件二的数据，建立了基于统计的分类模型：k 近邻 (k-Nearest Neighbor, kNN) 模型，是比较好的文本分类算法之一。(kNN 分类模型的主要思想：通过给定一个未标注文档 d，分类系统在训练集中查找与它距离最接近的 k 篇相邻(相似或相同)标注文档，然后根据这 k 篇邻近文档的分类标注来确定文档 d 的类别。) 分类实现过程：

1) 将训练集样本转化为向量空间模型表示形式并计算每一特征的权重；

2) 采用类似步骤 1 的方式转化未标注文档 d 并计算相应词组元素的权重；

3) 计算文档 d 与训练集样本中每一样本的距离(或相似度)；

4) 找出与文档 d 距离最小(或相似度最大)的 k 篇训练集文本；

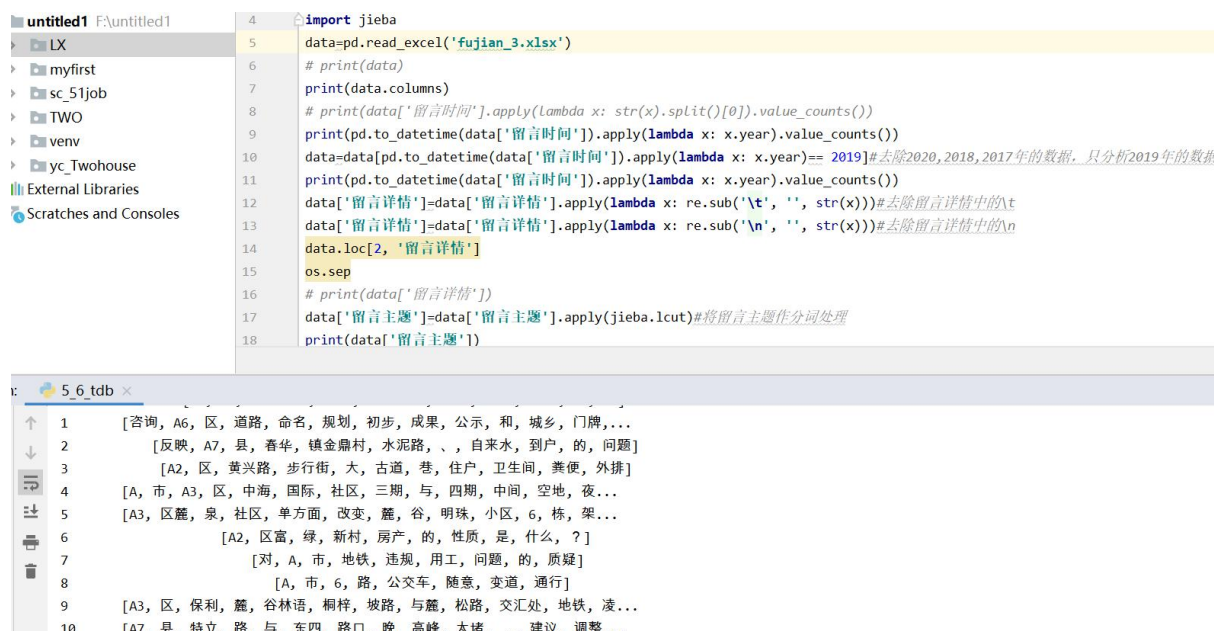
5) 统计这个 k 篇训练集文本的类别属性，一般将文档 d 的类归为 k 中最多的样本类别。(kNN 分类模型是一种“懒学习”算法，实质上它没有具体的训练学习过程。分类过程只是将未标注文本与每一篇训练集样本进行相似度计算，kNN 算法的时间和空间复杂度较高。因而随着训练集样本的增加，分类的存储资源消耗大，时间代价高。一般不适合处理训练样本较大的分类应用。)

3. 一位所得出的结果与模型进行对比评估

步骤二：使用 python 语言，进行数据预处理，将 2019 年的数据进行提取，再用数据清洗的方法将其他年份的数据、没用的数据和符号\n\t 全部剔除，分词，以便于后续分类，然后通过 jieba 分词库对数据做分词处理。然后用实体命名识别进行提取问题的三要素（特定时间，特定地区和发生的时间），再用停用词进行相似度计算将所有数据中的相似回答进行合并规整得到相似文本，计算相同词出现次数，再进行文本数据按照由

多到少进行排序，最后保存 Excel 文本得到热点问题表和热点问题留言明细表。

利用 jieba 分词库将留言详情进行分词操作，完成分词操作后将提取关键字，进行关键字的相似度计算，然后再进行热点排名，得到热点数据表



```

4 import jieba
5 data=pd.read_excel('fujian_3.xlsx')
6 # print(data)
7 print(data.columns)
8 # print(data['留言时间'].apply(lambda x: str(x).split()[0]).value_counts())
9 print(pd.to_datetime(data['留言时间']).apply(lambda x: x.year).value_counts())
10 data=data[pd.to_datetime(data['留言时间']).apply(lambda x: x.year)== 2019]#去除2020,2018,2017年的数据，只分析2019年的数据
11 print(pd.to_datetime(data['留言时间']).apply(lambda x: x.year).value_counts())
12 data['留言详情']=data['留言详情'].apply(lambda x: re.sub('\t', '', str(x)))#去除留言详情中的\t
13 data['留言详情']=data['留言详情'].apply(lambda x: re.sub('\n', '', str(x)))#去除留言详情中的\n
14 data.loc[2, '留言详情']
15 os.sep
16 # print(data['留言详情'])
17 data['留言主题']=data['留言详情'].apply(jieba.lcut)#将留言主题作分词处理
18 print(data['留言主题'])

```

```

1 [咨询, A6, 区, 道路, 命名, 规划, 初步, 成果, 公示, 和, 城乡, 门牌,...
2 [反映, A7, 县, 春华, 镇金鼎村, 水泥路, , , 自来水, 到户, 的, 问题]
3 [A2, 区, 黄兴路, 步行街, 大, 古道, 巷, 住户, 卫生间, 粪便, 外排]
4 [A, 市, A3, 区, 中海, 国际, 社区, 三期, 与, 四期, 中间, 空地, 夜...
5 [A3, 区, 麓, 泉, 社区, 单方面, 改变, 麓, 谷, 明珠, 小区, 6, 栋, 架...
6 [A2, 区, 富, 绿, 新村, 房产, 的, 性质, 是, 什么, ? ]
7 [对, A, 市, 地铁, 违规, 用工, 问题, 的, 质疑]
8 [A, 市, 6, 路, 公交车, 随意, 变道, 通行]
9 [A3, 区, 保利, 麓, 谷林语, 桐梓, 坡路, 与麓, 松路, 交汇处, 地铁, 凌...
10 [A7, 县, 特立, 路, 与, 东四, 路口, 晚, 高峰, 太堵, , , 建议, 调整...

```

步骤三：

将附件 4 的数据进行留言主题分类——>将回复内容按留言主题的分类进行分类，再制定一个评价指标：相关部门如何回答这个问题。从而需要对该问题进行一个量化，相关部门怎样去答复该问题，质量怎么样。因此对答复意见的归类整理，相似度计算，分类回答意见。具体如下：

①对留言分类（去除停用词（文本缩减）——> $Tf-idf = \text{词频}(TF) * \text{逆文档频率}(IDF)$ （关键词提取）——> $IDF(\text{逆文档频率}) = \log(\frac{\text{词料库的文档总数}}{\text{包含该词的文档数} + 1})$ ）

```

                                m_titles_S \
0                                [A2, 区景蓉华苑, 物业管理, 有, 问题]
1                                [A3, 区潇楚, 南路, 洋湖, 段, 怎么, 还, 没, 修好, ? ]
2                                [请, 加快, 提高, A, 市, 民营, 幼儿园, 老师, 的, 待遇]
3                                [在, A, 市买, 公寓, 能, 享受, 人才, 新政, 购房, 补贴, 吗, ? ]
4                                [关于, A, 市, 公交站点, 名称, 变更, 的, 建议]
...
2810 [为, 促进, 张界, 家中, 湖乡, 旅游, 有序, 发展, , , 呼吁, 在, 中, 湖...
2811 [汽车, 北站, 进站, 口, 附近, 居民, 强烈, 反对, 建设, I, 市, 平康, ...
2812 [强烈, 反对, I, 市, 9, 路, 公交车, 改, 线路]
2813 [对, G7, 县文盛, 小学, 特色, 班, 的, 一点, 质疑]
2814 [燃油, 税费, 改革, 政策, 的, 咨询]

                                responses_S
0    [现将, 网友, 在, 平台, 《, 问政, 西地省, 》, 栏目, 向, 胡华衡, 书记,...
1    [网友, “, A00023583, ”, : , 您好, !, 针对, 您, 反映, A3,...
2    [市民, 同志, : , 你好, !, 您, 反映, 的, “, 请, 加快, 提高, 民营,...
3    [网友, “, A000110735, ”, : , 您好, !, 您, 在, 平台, 《, ...
4    [网友, “, A0009233, ”, , , 您好, , , 您, 的, 留言, 已, 收悉...
...
2810 [你好, , , 你, 所, 反映, 的, 问题, 已, 转交, 相关, 部门, 调查, 处置...
2811 [您, 的, 留言, 已, 收悉, 。, 关于, 您, 反映, 的, 问题, , , 已转, ...
2812 [ “, UU008194, ”, 您, 的, 留言, 已, 收悉, 。, 关于, 您, 反映...
2813 [ “, UU0082115, ”, 您好, !, 获悉, 关于, “, 对, G7, 县文盛...
2814 [西地省, 平台, 《, 问政, 西地省, 》, 栏目组, : , , , 网民, ...

```

利用 pandas 中的 DataFrame 将留言主题和留言回复进行提取出来。

```

A2 区景蓉华苑 物业管理
A3 区潇楚 洋湖
幼儿园 民营 待遇
市买 公寓 新政
公交站点 变更 名称
业主大会 业委会 业主 停车 胡华衡
施工 坪塘 排水 A3 换填
民办 幼儿园 待遇 教师 学前教育
购房 房屋交易 补贴 管理中心 首次
马坡岭 来信 小学 公交站点 市民

```

利用 jieba 分词中的 analyse 提取关键字即要提取关键字的个数。

②然后将相应的答复意见归并在同一类留言中（可以利用正则表达式进行提取）

③相似度计算

分词（）——>语料库（词集）（清洗）——>词频——>词频向量

④针对该分类的同类留言进行评价一致化，从而形成一套评价方案。

2.2. 对答复意见的评价方案

通过互联网收集群众留言意见已经成为相关部门或企业发展的主要途径之一，随着网络的不断发展和普遍，几乎所有地方都被网络覆盖，使得

人们对生活、生产上的不满意能够很好的向相关机构进行随时随地地反映具体情况，相关机构设置了专门的信息收集、向上级反映、回复群众、提供解决方案的部门为人民群众提供服务咨询。使得相关部门可以了解民声、贴近民生，从而得到更好地发展；也使得民众得到切实的益处。

故对其答复意见的评价方案可以从以下几个方面进行参考：

首先，确定群众留言内容是关于哪一方面的留言，了解他们反映的具体内容，所留言内容是否迫切需要得到答复。

其次，看相关机构或部门是否及时的对群众反映的内容进行具体回复，回复的方案是否切合实际，能够得到快速实施解决相关问题，回复之后是否及时采取相应措施。

最后，根据具体回复内容看实施之后是否得到预期的成效，群众对回复及措施的满意度。

3. 结论

总结本次比赛，通过互联网问政服务的出现来促使政府更加方便地了解民意、汇聚民智、凝聚民气，各类社情民意相关的文本数据量不断攀升，使民众反映的问题不断细化，处理不断智能化，给社会带来极大益处，而在以往数据量较大，且主要依靠人工来进行留言划分和热点整理，使得相关部门的工作进展缓慢。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。因此本次文本数据挖掘任务一、二、三中均使用到自然语言处理技术和数据挖掘等技术来完成挖掘目标，实现任务，解决问题。

4. 参考文献

- [1]. 赵琳瑛. 基于隐马尔科夫模型的中文命名实体识别研究. 西安电子科技大学. 2007
- [2]. 范庆春. 基于中文分词技术的文本相似度检测研究. 池州学院学报. 2019
- [3]. 黄丹丹. 基于深度学习的中文分词和关键词抽取模型研究. 北京邮电大学
- [4]. 解宇涵. 基于深度学习的中文分词模型应用研发. 重庆大学. 2017
- [5]. 康鲲鹏. 基于大数据的数据清洗研究. 商丘师范学院信息技术学院. 2018
- [6]. 朱少杰, 基于深度学习的文本情感分类研究. 哈尔滨工业大学:硕士学位论文. 2014

附录

代码 1:

```
import pandas as pd
import re
import os
import jieba
data=pd.read_excel('fujian_3.xlsx')
# print(data)
print(data.columns)
# print(data['留言时间'].apply(lambda x:
str(x).split()[0]).value_counts())
print(pd.to_datetime(data['留言时间']).apply(lambda x:
x.year).value_counts())
data=data[pd.to_datetime(data['留言时间']).apply(lambda x:
x.year)== 2019]#去除 2020,2018,2017 年的数据,只分析 2019 年的数据
print(pd.to_datetime(data['留言时间']).apply(lambda x:
x.year).value_counts())
data['留言详情']=data['留言详情'].apply(lambda x: re.sub('\t', '',
str(x)))#去除留言详情中的\t
data['留言详情']=data['留言详情'].apply(lambda x: re.sub('\n', '',
str(x)))#去除留言详情中的\n
data.loc[2, '留言详情']
os.sep
# print(data['留言详情'])
data['留言主题']=data['留言主题'].apply(jieba.lcut)#将留言主题作
分词处理
print(data['留言主题'])
# print(data)
```

代码 2:

```
import pandas as pd
```

```

df_datas=pd.read_excel(r'C:\Users\ASUA\Desktop\课\比赛\泰迪\C 题
全部数据\C 题全部数据\附件 4.xlsx',names=['m_numbers', 'm_ids',
'm_titles', 'm_times', 'm_details', 'responses', 'response_times'])
df_datas=df_datas.dropna()
df_datas.head()

print(df_datas)

print(df_datas.shape)
#
m_title=df_datas.m_titles.values.tolist()
print(m_title[0:2815])
response=df_datas.responses.values.tolist()
print(response[0:2815])

#
import jieba
m_titles_S=[]
responses_S=[]
#遍历
for line in m_title:
    current_segment=jieba.lcut(line)
    if len(current_segment)> 1 and
current_segment!='\n\t\t\t\t\t\n\t\t\t\t\t':#换行符
        m_titles_S.append(current_segment )
print(m_titles_S)

for line in response:
    current_segment=jieba.lcut(line)
    if len(current_segment )> 1 and
current_segment!='\n\t\t\t\t\t\n\t\t\t\t\t':#换行符
        responses_S.append(current_segment)
print(responses_S)
#
df_m_titles=pd.DataFrame({'m_titles_S':m_titles_S})
print(df_m_titles.head(2815))
df_responses=pd.DataFrame({'responses_S':responses_S})
print(df_responses.head(2815))

```

```
#
#TFD 提取(m_titles 中)关键字
import jieba.analyse
for i in range (0,2815):
    indexs=i
    # print(df_datas['m_titles'][indexs])
    m_titles_S_str="".join(m_titles_S[indexs])
    print("
".join(jieba.analyse.extract_tags(m_titles_S_str,topK=3,withWeight=False)))

for i in range (0,2815):
    indexs=i
    # print(df_datas['responses'][indexs])
    responses_S_str = "".join(responses_S[indexs])
    print("
".join(jieba.analyse.extract_tags(responses_S_str,topK=10,withWeight=False)))
```