2020年第八届“泰迪杯”数学挖掘挑战赛论文

题　　目　“智慧政务”中的文本挖掘应用

关 键 词　留言分类、挖掘、数据清洗

摘　　要：

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，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。本文针对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，对留言进行分类，建立多分类模型，利用自然语言处理和文本挖掘的方法，对数据进行预处理、存储、分析。

首先，针对群众留言分类，需要在python下安装pandas，jieba和word2vec对群众的留言详情进行数据预处理。先提取出所要处理的留言详情数据，对其进行去重去空和分词操作，并去掉停用词，完成初步数据的清洗操作。建立留言内容的一级标签分类模型，使用 F-Score 对分类方法进行评价，得出群众留言分类结果。

关于挖掘热点问题，在对群众留言进行分类的基础上，在python上安装xlwt，xlwings和collections，将群众留言分类的数据清洗结果进行词频统计，将词频按频率大小排序，针对特定的时间、地点和问题进行相似度计算，得出频率前五的结果。从而挖掘出合理的热度评价指标，得出相应的评价结果。

针对相关部门对留言的答复意见，在文本挖掘的基础上，对答复意见的数据进行清洗，用jieba分词，去停用词。再用数学方法选取最具分类信息的特征词，用word2vec筛选出特定文本，构建模型，最后验证模型，得出结果。

Text mining application in "smart government"

Key words: message classification, mining, data cleaning

In recent years, with the development of wechat, microblog, mayor's mailbox, sunshine hotline and other network platforms, which have gradually become an important channel for the government to understand public opinion, gather people's wisdom and gather people's spirit, and 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 become a new trend of social governance innovation and development, which will enhance the government's The level of management and the efficiency of governance play an important role. This paper aims at the records of the public political messages from the Internet and the response opinions of the relevant departments to some of the public messages. Firstly, it classifies the messages, establishes a multi classification model, and uses the methods of natural language processing and text mining to preprocess, store and analyze the data.

First, for the classification of people's comments, we need to install pandas, jieba and word2vec under python to preprocess the details of people's comments.Firstly, extract the message details data to be processed, and carry out the operations of reduplication, emptyness and word segmentation, and remove the stop words to complete the preliminary data cleaning operation.The first-level label classification model of message content was established, and f-score was used to evaluate the classification method, and the classification results of crowd message were obtained.

As for mining hot issues, based on the classification of mass comments, XLWT, xlwings and collections were installed on python. Word frequency statistics were carried out for the data cleaning results of mass comments, word frequency was sorted by frequency, and similarity calculation was carried out for specific time, place and problem to get the top five results of frequency.In this way, a reasonable heat evaluation index is excavated and the corresponding evaluation results are obtained.

In view of the response comments of relevant departments to the message, on the basis of text mining, the data of the response comments to clean, use jieba segmentation, to stop the word.Then, mathematical method is used to select the feature words with the most classified information, and word2vec is used to screen out specific text to build the model, and finally, the model is verified and the results are obtained.

目录

1. 挖掘目标

本次建模针对网络问政平台上关于社情民意的数据，采用自然语言处理和文本挖掘的方法，达到以下三个目标：

（1）对留言内容进行自然语言处理，建立关于留言内容的一级标签分类模型。

（2）对众多留言进行识别，把特定地点或人群合并，把相似的问题归为同一问题，定义热度评价指标及其计算方法。

（3）从相关性、完整性、可解释性等角度进行分析，尝试实现一套相关部门对留言的答复意见的质量评价方案。

2. 分析方法和过程

2.1. 总体流程

本用例主要包括以下几个步骤：

步骤一：读取数据，群众留言的获取是本次数据挖掘分析的第一步。本文中利用 jieba 库对附件 2 进行读取，最后将留言文本批量存进 txt 文件中，得到实验数据。

步骤二:数据预处理，第一步要“去空、去重”；第二步对留言数据进行中文分词，将一句评论分成多个词语进一步分析；第三步进行停用过滤，去掉与情感判定无关词语。

步骤三：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。

步骤四：对问题归并进行相似度计算，并统计人群反映问题词频。

步骤五：对群众留言进行分类，发现某一时间段内反映特定地点和特定人群的热点问题，定义合理的热度指标，得出评价结果，建立模型。

步骤六：针对相关部门对留言的答复意见，尝试实现一套评价方案。

2.2.具体步骤

步骤一：读取数据

在 python 中安装好 pandas，jieba 和 word2vec，分别将附件 1.xlsx，附件 2.xlsx 另存为附件 1.csv，附件 2.csv，使用 pandas 对 csv 文件读取所需的数据，并且按行写入 txt 文本之中。

代码如下：

```
import pandas as pd
import jieba
import jieba.analyse

A = open('C:/Users/Dell/Documents/泰迪杯数据/附件 1.csv')
first_df = pd.read_csv(A)
first_df

B = open('C:/Users/Dell/Documents/泰迪杯数据/附件 2.csv')
second_df = pd.read_csv(B,usecols=[4])
second_df
```

In [22]:
```
import pandas as pd
import jieba
import jieba.analyse
```

In [23]:
```
A = open('C:/Users/Dell/Documents/泰迪杯数据/附件1.csv')
first_df = pd.read_csv(A)
first_df
```

Out[23]:

	一级分类	二级分类	三级分类
0	城乡建设	安全生产	事故处理
1	城乡建设	安全生产	安全生产管理
2	城乡建设	安全生产	安全隐患
3	城乡建设	城市建设和市政管理	园林绿化环卫
4	城乡建设	城市建设和市政管理	城管执法
...	...	...	...
512	劳动和社会保障	退休政策及待遇	内部退养人员待遇
513	劳动和社会保障	退休政策及待遇	其他
514	劳动和社会保障	退休政策及待遇	退休政策
515	劳动和社会保障	退休政策及待遇	退休金发放
516	劳动和社会保障	退休政策及待遇	病退及提前退休人员待遇

517 rows × 3 columns

In [24]:

```
B = open('C:/Users/Dell/Documents/泰迪杯数据/附件2.csv')
second_df = pd.read_csv(B,usecols=[4])
second_df
```

Out[24]:

	留言详情
0	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...
1	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...
2	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...
3	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...
4	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...
...	...
9205	\n \n 我们夫妻都是农村户口，大的是女9岁，小的是儿2岁半，才15斤...
9206	\n \n 本人2015年2月16号在B市中心医院做无痛人流手术，手术后...
9207	\n \n 我们是再婚，很想再要一个小孩，不知我省二胎新政策何时出，如果...
9208	\n \n K8县惊现奇葩证明！ 我是西地省K8县人，想生二孩。被告知要...
9209	\n \n 领导你好，我们属于未婚生子，但是在2013年已经接受处罚，小...

9210 rows × 1 columns

步骤二：数据预处理

文本在经过去重、去空、中文分词后，里面还有一些包含的信息量很低甚至没有信息量的词语，需要将它们过滤掉，否则将会影响分析的正确率，所以需要去停用词。我们将采用词性和停用词表的过滤方法，对数据进行初步清洗。

分词结果示例：

分词前：中国是一个伟大的国家。

分词后：中国/ns 是/v 一个/m 伟大/a 的/uj 国家/n 。/x

去停用词结果示例：

分词后：第一/m 次/qv 在/p 淘宝/nz 购/vg 物/n ，/wd 售/v 后/f 很/d 满意/v ，/wd 下次/t 继续/v 购买/v 。/x

停用词过滤后：第一 淘宝 购物 售 后 很 满意 下次 继续 购买

使用 word2vec 可对数据进行清洗，经过数据预处理后，存入新的 csv 文件。

步骤三：建立分类模型

使用 pandas 对清洗后的数据进行关键词配对，分别使用 merge()函数，

concat()函数，groupby()函数对附件 1.csv 和新的 csv 文件进行配对，建立好分类模型。在建立好分类模型后，使用 F-Score 对分类方法进行评价，得出群众留言分类结果。

步骤四：统计人群反映问题词频

首先提取步骤二的初步清洗后的数据，对其进行词频统计，将词频按频率大小排序，针对特定的时间、地点或问题进行相似度计算，去掉无用的信息，得出的高频词可以反映出留言中的热点问题，及时发现的热点问题，有助于相关部门进行有针对性地处理，提升服务效率。得出频率前五的结果，从而挖掘出合理的热度评价指标。

代码如下：

```
python setup.py install


pip --default-timeout=1000 install -U pip

python -m pip install --user genism


import collections

#coding=utf-8

filename = "C:/Users/Dell/Documents/泰迪杯数据/附件 3.csv"

with open (filename,'rb') as f:

    words_box=[]

    words_box2=[]

    for line in f:
```

```
        line.decode("utf-8")
        words_box.extend(line.strip().split())
    for word in words_box:
        word2 = word.decode("utf-8")
        words_box2.append(word2)
print("词的总数为：%s"%len(words_box2))
    print("词频结果：%s"%collections.Counter(words_box2))
```

步骤五：建立热点问题挖掘模型

针对步骤四得出频率前五的结果，从而建立合适的分类模型，得出附件3.csv 留言文本中排名前五的热点问题，并保存文件为“热点问题表.xls”；得出相应热点问题的留言信息，并保存文件为“热点问题留言明细表.xls”。

步骤六：

针对相关部门对留言的答复意见，在文本挖掘的基础上，对答复意见的数据进行清洗,用 jieba 分词,去停用词。再用数学方法选取最具分类信息的特征词，用 word2vec 筛选出特定文本，构建模型，最后验证模型，得出结果。

3. 总结

总结本次比赛，因为我们是第一次参加这样的比赛，经验不足，刚开始也不知道如何入手,后来根据网上提供的以往的泰迪杯参赛作品的论文以及详细的解释，在观看完泰迪云课堂的题目解读之后，我们有了一点点的思路，然后由简入繁，逐步解决挖掘问题，实现本次数据的目标。

本次数据挖掘分析的过程中，主要分为三个问题：一是对留言内容进行自然语言处理，建立关于留言内容的一级标签分类模型；二是对众多留言进行识别，

把特定地点或人群合并，把相似的问题归为同一问题，定义热度评价指标及其计算方法；三是从相关性、完整性、可解释性等角度进行分析，尝试实现一套相关部门对留言的答复意见的质量评价方案。每一步也都有详细的步骤，各个步骤之间也有密切的联系，实验的每一步也有理有据。但是在数据挖掘的过程中，我们还是遇到了很多瓶颈问题，小的问题我们已尽全力去解决了，大的问题，我们的能力也有限，只能绕圈寻找其他解决办法。

对于此次比赛，我们也更深入地了解了数据挖掘分析给我们的生活带来的便利。我们针对留言进行分类，寻找热点问题，分析热点问题，并从答复的相关性、完整性、可解释性等角度对答复意见的质量给出了一套评价方案，有助于相关部门进行有针对性地处理，提升服务效率。

4. 参考文献

[1]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

[2] 王素格，李书鸣，陈鑫，穆婉青，乔霈，面向高考阅读理解观点类问题的答案抽取方法.山西大学