第八届“泰迪杯”数据挖掘挑战赛

——C 题：“智慧政务”中的文本挖掘应用

论 文

2020 年 5 月 7 日

目录

摘　要

许多网络平台的群众留言是政府了解民意的重要渠道，但是对于大量的留言问题，不仅要知道它是属于什么类型的问题，还要知道哪个问题是群众反映最多最关心的，然后政府给予相关解决方法以及回复，所以需要建立基于相关的计算机语言处理技术系统来对这些数据进行自动的分析划分，以提高政府管理水平及施政效率。本次文本挖掘主要是通过群众留言进行热点问题的整理，并针对热点问题进行回复。由于数据数量巨大，人工数据分析工作量过大易出错效率低，所以需要用到数据分析工具进行分析，从大量的数据中抽取有用的数据并建模，为了提高数据分析的操作效率，我们主要根据理解利用 Python 语言来解决问题。Python 用于处理大量数据的爬取、清洗、整理、分析等工具，另外我们还使用到了 R 语言进行数据分析，R 语言较 Python 更易操作，但其分析数据的功能没有 Python 强大。

关键词：留言划分 热点整理 自然语言处理技术

1 概述

1.1 问题背景

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。面对大量的数据，人工数据分析面临巨大问题，数据量多导致的效率低下使得问题不能最有效的被解决。针对这一现象，我们对于大量数据进行分析的解决方法是使用数据分析工具进行处理。数据分析工具，我们主要采取了两种，Python 和 R 语言。

1.2 研究目标

建立基于自然语言处理技术的系统，利用相关模型对留言内容进行数据清洗评价，提高政府的管理水平和施政效率。

1.3 解决思路

1、群众留言分类

根据附件 1 和附件 2 的数据首先建立关于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。

2、热点问题分类

识别并归类相似留言（归并特定地点或人群的数据），然后根据分好类的留言中点赞数和反对数的数据定义一个热度评价指标的定义和计算方法，对指标排名。

3、答复意见评价

根据答复内容评价其内容的相关性、完整性、可解释性。

2 具体解决步骤

1、先读取附件数据，再进行数据的抽取。

主要运用 pandas 包对于数据进行读取。

```
def data_process(file='附件2.xlsx'):
    data = pd.read_excel(file)  # 导入数据
    data.columns = ['number', 'user', 'theme', 'time', 'message', 'lev1lab']  # 更换列名

    data_dup = data['message'].drop_duplicates()  # 去除重复行
    data_qumin = data_dup.apply(lambda x: re.sub('x', '', x))  # 去除敏感字符

    data_cut = data_qumin.apply(lambda x: jieba.lcut(x))    #分词

    stopWords = pd.read_csv('stopword.txt', encoding='utf-8', sep='hahaha', header=None)  #导入停用词表
    stopWords = [' '] + list(stopWords.iloc[:, 0])
    data_after_stop = data_cut.apply(lambda x: [i for i in x if i not in stopWords])   #去停用词
```

```
9185    [\n, \n, 外省, 中专, 毕业证, 可以, 参加, K, 市, 助理, 医师, 考试...
9186    [\n, \n, 新手, 准妈妈, 每次, 医院, 产检, 过程, 体验, 不好, 产检, ...
9187    [\n, \n, 7, 父亲, 心梗, 住, 中南大学, 楚雅, 医院, 时间, 多名, 医...
9188    [\n, \n, 尊敬, 卫健委, 领导, 您们好, 我于, 中专, 文凭, 报考, 乡村,...
9189    [\n, \n, 尊敬, 领导, 你好, 妇女, 主任, 催人, 检查, 唐氏, 筛查, 报...
9190    [\n, \n, 第一个, 建议, 建议, 医检, 检验, 增加, 透明度, 血常规, 检查...
9191    [\n, \n, 您好, 咨询, 问题, 请, 百忙中, 回复, 盼, 国家, 卫生, 健康...
9192    [\n, \n, 请问, K, 市, K1, 区, 二级, 二级, 公立医院, 请, 提供,...
9193    [\n, \n, K, 市中心, 医院, 龚卫平, 被选为, 副, 主任, 龚卫平, 这样,...
```

```
    data_after_stop.index  #获取序号
    numbers = data.loc[data_after_stop.index, 'number']
```

2、数据清洗，对数据进行预处理。

主要结合 beautifulsoup 和正则表达式。首先进行处理编码，然后将文档分割成句子，将句子分割成词，用 pyenchant 拼写错误纠正，正则表达式去掉标点符号，去掉长度过小 len<3 的单词。

```
29          4079
30          4082
31          4361
          ...
9180      348079
9181      348614
9182      349308
```

```
9180    \n \n 打扰 想 请教 老婆 K2 区 蔡 市镇 K1 区 邮亭 圩镇 可以 老婆 户口...
9181    \n \n 4 14.15 到区 妇幼保健 院 看病 久 始终 不见 开门 请问 这样 工作...
9182    \n \n 网上 得知 狂犬病 死亡率 几近 100% 请问 真的 预防 狂犬病 措施 \n...
9183    \n \n M2 县 印塘 乡 一名 村民 朱灿 发现 儿子 朱诺 杭 3 岁 误服 几粒 ...
9184    \n \n 2014 老公 K 市 办理 准生证 生 第一个 孩子 户口 迁到 市 居住地 ...
9185    \n \n 外省 中专 毕业证 可以 参加 K 市 助理 医师 考试 毕业证 全国 中等职业...
9186    \n \n 新手 准妈妈 每次 医院 产检 过程 体验 不好 产检 项目 繁多 流程 希望 ...
```

```
adata = data_after_stop.apply(lambda x: ' '.join(x))    #空格隔开
```

3、分词。

利用相关库（jieba 库等）对留言的内容进行关键提取，主要使用了 jieba 分词包是效率高，最简单的分词包。也可以使用 HankLP 分词和 pyltp 分词等方式进行分词和词性。去停用词，去除没有意义的词语和字符。用数字方法选取最具分类信息的特征进行文本特征选择。区分文本相似度。

4、进行建模准备、数据准备。

5、根据已知数据进行建模。

使用文本分类模型，传统的机器学习方法，深度学习。进行文本聚类，自动摘要，情感分析，命名实体类别等。

情感分析，是对带有感情色彩的主观性文本进行分析、处理、归纳和推理的过程。按照处理文本的类别不同，可分为基于新闻评论的情感分析和基于产品评论的情感分析。其中，前者多用于舆情监控和信息预测，后者可帮助用户了解某一产品在大众心目中的口碑。目前常见的情感极性分析方法主要是两种：基于情感词典的方法和基于机器学习的方法。[1]

6、进行模型验证，优化。

根据结果进行评价与优化，部署。优化分词，去停用词，优化特征工程，算法调优。

3 总结

通过数据分析工具的编写,我们大致进行了数据的读取、清洗,分词等操作，对于问题的解决有了一定的理解。

[1] 百度百科 Python 情感分析