wrangle_report

2020 年 4 月 27 日

1 项目背景

项目名称：推特狗狗数据清洗与分析

项目内容：本项目需要收集、清洗、整理和分析数据集里的推特用户对狗狗的喜爱程度。

2 数据文件

- tweet_json.txt 数据
- twitter-archive-enhanced.csv 原始数据
- image-predictions.tsv 为图像预测

3 项目目标

- 对项目数据进行评估收集上述三个数据集后，使用编程评估的方式对数据集的质量和清洁度进行评估，在 wrongle_act_cn(1).ipynb 记录评估的过程和结果。
- 对项目数据进行清洗对评估时列出的问题进行清洗，在 wrangle_act_cn(1).ipynb 里记录清洗过程。理论上，清洗后应该会整理出一个干净整洁并符合项目分析要求的新数据集。
- 对项目数据进行存储、分析和可视化将清理后的数据集存储为 twitter_archive_master.csv。
- 在 wrongle_act_cn(1).ipynb 中对清洗后的数据进行分析和可视化。
- 项目汇报建立名为 wrangle_report.pdf 的书面报告，主要描述数据整理过程。

4 收集

1. 收集手头文件'twitter-archive-enhanced.csv', 其中包含了一些主要的推特信息，是本次清洗的主要数据，其中的评分、地位和名字等数据是从'text' 原文中提取的，但是提取的并不好。评分并不是正确的，狗的名字和地位也有不正确的。

2. 编辑下载互联网文件：'image-predictions.tsv'，其中包含了推特图像预测信息，根据推特中的图片预测出狗狗种类。

3. 查询 API 收集额外推特信息'tweet_json.txt' 从中提取所需数据，至少需要提取转发数（'retweet_count') 和喜欢数 ('favorite_count')。

5 评估

质量

'twitter 表格

- 'expanded_urls' 处在空值;-'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id', 存在大量空值。
- 'timestamp' 数据应该为时间
- 'expanded_urls' 有 137 个重复
- 狗狗分类 'doggo','floofer','pupper','puppo' 需要合并并删除重复数据

'image_predict 表格

- 'tweet_id' 是整数值，需要清理一些重复列
- 'jpg_url' 有 66 个重复值
- 'p1','p2','p3' 数据大小写不一致

tweet 表格

- 目前没有发现数据异常

整洁度

- twitter 表格里的狗分类模式有问题
- 三个表格可以根据'tweet_id' 合并，以便更方便整理数据的完整性。

6 清理

问题描述一

twitter，image_predict，tweet 三个表格合并和删除不含图片的推特信息 使用 merge() 将三个表格合并为一个数组，合并后三个表格的数据都可以在新建的 df 表格里找到，方便后续清理。

问题描述二

各项 'retweet' 信息无效 str.find() 在 text 列里找出'RT @'信息，将多余的无效数据删除。删除 doggo,floofer,pupper,puppo 列