Wrangle Report 清理报告

Gather 收集数据

库的导入

导入pandas库，供后续使用pandas进行数据集清理

导入requests库，供后续从互联网爬取文件

导入json库，供之后解析json文件

导入matplotlib库，供后续可视化分析

导入seaborn库，并设置风格为darkgrid，加强可视化的效果

资料来源：手头文件

twitter-archive-enhanced.csv是已经给的，所以用pandas中的read_csv打开

得到 dogs_rate 数据集

资料来源：从互联网下载文件

image_predictions.tsv 文件是使用 Python 的 **Requests** 库和提供的 URL 来进行编程下载，一份推特图像的预测数据，用read_csv将文件在pandas中打开

得到 image 数据集

资料来源：json文件

使用 Python **Tweepy 库**查询 API 中每个推特的 JSON 数据，把所有 JSON 数据存储到一个名为 tweet_json.txt 的文件中，将txt文件转成csv文件在pandas中打开

得到 tweet_df 数据集

Access 评估数据

检查数据集是否存在两个问题：数据质量问题（即，内容性问题）和 缺乏整洁度（即，结构性问题）

将之前打开的三个数据集事用常规编程方法评估

常规编程方法包括：

head

tail

sample

info

评估 dogs_rate 数据集：

先用head发现其中in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id, retweeted_status_timestamp 等列没有任何数据，source列数据比较杂乱

再用tail发现，index 从2352到2354列的 name 中使用了“a”代替名字，使用 sample 发现其他 index 也存在相同情况，使用 value_counts 查看name列的具体情况，发现代替名字的有“a”，“the”，“an”，“by”还有存在没有名字的情况

rating_denominator 列中有值为0，2，7的情况，通过索引查看详情，从text的文本中发现是合理的数据

使用 info 查看所有列是否有空值和所有列的数据类型

发现有转发的推特列也在数据集中

发现 tweet_id 数据类型是int，最好是str

发现 expanded_urls 列有空值，使用 isnull().sum() 查看空值数量

对非常重要的评分系统 rating_numerator 和 rating_denominator使用 value_counts

评估 image 数据集：

使用head、tail、sample、info查看空值和数据类型

发现 tweet_id 数据类型是int，str更为合适

评估 tweet_df 数据集：

使用head、tail、sample、info查看空值和数据类型

发现 tweet_id 数据类型是int，str更为合适

Clean 清理数据

Quality 质量问题

保存原本的数据集，使用copy将即将整理的数据集后缀加clean

清理 dogs_rate 数据集：

先将转发的推特列从数据集中清理出去，将dogs_rate_clean.retweeted_status_id为空值的筛选出来，只保留博主自己的原始推文

将与数据分析无关的内容，并且也为空值的几列in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 清除

将名字中，使用代替名字的“a”，“the”，“an”，“by”统一改成NaN，看之后能不能通过其他资料补全

用 dropna 将 expanded_urls 没有图片的清理

使用 replace 将 source 列中替换为最简短的来源

将rating_numerator 和 rating_denominator 列中的 datatype 改为 float

用 to_datetime 将 timestamp 改成 datetime 以备后续分析的方便

清理 image 数据集：

用 astype 将 tweet_id 列的数据类型变更为 str

清理 tweet_df 数据集：

用 astype 将 tweet_id 列的数据类型变更为 str

Tidiness 整洁度

使用 extract 将 doggo，floofer，pupper，puppo 的文本提取出来填充到新列 stage

再将之前在 dogs_rate_clean 中的 doggo，floofer，pupper，puppo 四列从数据集中丢掉

使用 merge 将 image_clean 和 tweet_df_clean 合并到 dogs_rate_clean 数据集中

使用 head 和 info 查看数据集详情和数据类型