wrangle_report

简介

本次项目选取推特昵称为 WeRateDogs 的档案为数据集进行数据清洗分析和可视化。因为这份档案只包含基础信息，还需要另外收集相关数据，一起进行评估和清洗，得出结论。

一、 收集数据

- twitter_archive_enhanced.csv 推特档案
- tweet_json.txt 包含转发数(retweet_count)和喜欢数（favortite_count）
- image_predicitions.tsv 图像预测文件,对推特中中的品种（或其他物体）进行预测的结果

收集方式

- twitter_archive_enhanced.csv 可以直接下载
- tweet_json.txt 从 API 下载
- image_predicitions.tsv 从 url 下载

二、评估

通过目测和编程发现数据需要清理：

1.质量

twitter_archive_enhanced 表格

- retweeted_status_id 含有转发数据
- tweet_id 是整数，不是字符串
- rating_numerator 分子中有低于 10 分的数据
- rating_denominator 分母有不等于 10 分的数据
- name 中有空值，且含 a、an 等错误
- doggo floofer pupper puppo 有全空的
- 与 image_predicitions 合并后 twitter 表中 jpg_url 包含无图片数据

tweet_json 表格

- tweet_id 是整数，不是字符串

image_predicitions 表格

- tweet_id 是整数，不是字符串

2.整洁度

- twitter_archive_enhanced 表中狗狗的地位（stage）:doggo,pupper,puppo,floof(er)成为一列。
- 把 twitter 表、tweet 表、image 表合成一张表，三者的联系是 tweet_id
- 删除 twitter 的无用列，只保留 tweet_id,text,rating_numerator,rating_denominator, name，stage，tweet_count,favorite_count

三、清洁

1.备份：

通过 copy()对三个数据集合进行备份

2.处理缺失数据

删除 twitter_archive_enhanced 表中 doggo、floofer、pupper、puppo 这些列

3.整洁度

- 对 twitter_archive_enhanced 表新增一个变量 stage，暂时以 np.nan 代替，后续会从 text 中提取
- 利用函数 mege 把 twitter_archive_enhanced 表和 tweet_json 合成一张表
- 删除 twitter_archive_enhanced 表无用列，只保留 tweet_id,text,retweeted_status_id,rating_numerator,rating_denominator, name，stage,retweet_count,favorite_count

4.质量

- 使用 as.type()把 tweet_id 的数据类型修改为字符串。因为 tweet 表格已经合并到 twitter 表格里，因此只需对 df_twitter_clean 和 df_image_clean 进行操作即可。
- 利用 drop 删除 df_twitter_clean 里 retweeted_status_id 的非空值，删除转发数据
- 使用正则表达式和 pandas 的 `str.extract` 方法，从 text 中提取 rating_numerator,rating_denominator。提取出的分子 rating_numerator 的,去除特别高的，把范围限定在 11~16。分母 rating_denominator 为 10。
- 使用正则表达式和 pandas 的 `str.extract` 方法，从 text 中提取 name
- 使用 str.findall 从 text 中提取 stage

- 把 df_twitter_clean 和 df_image_clean 合并在一起，方式是 inner，去除不包含图片的。

5.保存数据

把清洗好的数据存在 twitter_archive_master.csv 里。