Twitter-cute-dog 数据整理报告

数据收集

1. 收集推特UP主**WeRateDogs**的推特数据，保存为 dataframe（twitter-archive-enhanced）
2. 收集狗狗种类图像预测结果，保存为 dataframe（image_predictions）
3. 收集推特 UP 主推特的评论转发数据，保存为 dataframe（tweet_json）

数据评估

目测评估数据

在 notebook 中滚动查阅收集部分载入的三个 dataframe。检查数据质量与清洁度问题。

编程评估数据

使用 info，value_counts，describe 查看数据字段类型，空值以及重复的列名等问题。

数据清理

备份三个数据集

- twitter_archive_enhanced_clean
- image_predictions_clean
- tweet_json_clean

清洁度整理

1. `tweet_json`的转发数和喜欢数是`twitter_archive_enhanced`表格的一部分
2. `image_predictions`预测出的狗狗种类是`twitter_archive_enhanced`表格的一部分
3. `twitter_archive_enhanced`表格代表狗狗地位的变量（doggo,floofer,pupper,puppo）应该合并到一列
4. `twitter_archive_enhanced`表格 text 应该拆分为两个变量：tweet 内容与 tweet 链接

质量整理

1. 一条推特中狗狗会含有多个身份的情况(tweet_id:770093767776997377,733109485275860992)(因为 tweet 中同时包含了对两只狗狗的评价)
2. 狗狗身份判断错误(tweet_id:855851453814013952)(应该是 puppo)
3. 狗狗身份判断错误(tweet_id:817777686764523521)(应该是 pupper)
4. 狗狗身份判断错误(tweet_id:785639753186217984)(应该是 doggo)
5. 狗狗身份判断错误(tweet_id:801115127852503040)(应该是 pupper)
6. 狗狗身份判断错误(tweet_id:854010172552949760)(应该是 doggo)
7. 狗狗身份判断错误(tweet_id:751583847268179968)(应该是 doggo)
8. 分数提取不合理
9. tweet_id:820690176645140481 12/10(分值有问题,手动修复)
10. tweet_id:731156023742988288 12/10(分值有问题,手动修复)
11. tweet_id:710658690886586372 10/10(分值有问题,手动修复)
12. tweet_id:704054845121142784 12/10(分值有问题,手动修复)
13. tweet_id:686035780142297088 2/10(分值有问题,手动修复)
14. tweet_id:677716515794329600 12/10(分值有问题,手动修复)
15. tweet_id:832088576586297345 删除(非狗狗评分的相关的 tweet)
16. tweet_id:810984652412424192 删除(非狗狗评分的相关的 tweet)
17. tweet_id:682808988178739200 删除(非狗狗评分的相关的 tweet)
18. 部分名字提取错误（出现了大量的 a 和 None，可以用*named* *name is*关键字匹配）
19. 狗狗名称用了缩写(Tyr->Tyrone)（tweetid:796865951799083009）
20. 小写字母开头的狗狗名字都不对（such,an,very,a,just,not,one）
21. 单个字母的狗狗名字也有问题（O,a）
22. 错误的数据类型（ *in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_id*, *retweeted_status_user_id*浮点改为整型）
23. 错误的数据类型（*timestamp*, *retweeted_status_timestamp* 字符串改为时间类型）

存储清理后的主数据集

将清理后的主数据集（twitter_archive_enhanced_clean）保存为 csv 文件（twitter_archive_enhanced_clean.csv）