twitter 数据集整理

文件说明:

- wrangle_act_cn.ipynb 中使用了 pandas，numpy，requests，tweepy，json，matplotlib 库
- twitter_archive_master.csv 是 整理完成的文件

想重现清理过程请运行 wrangle_act_cn.ipynb,

1.收集

数据集有三个

- 本地文件：twitter_archive_enhanced.csv
  包含 twitter 帖子数据，优达学城提供，具体数据请运行 wrangle_act_cn.ipynb 查看
- 下载文件：image_predictions.tsv
  包含 twitter 中狗的图片种类预测，具体数据请运行 wrangle_act_cn.ipynb 查看
- json 文件：tweet_json.txt
  从 twitter 接口获取或者使用优达学城提供，包含每条 twitter 的详细信息，本次整理只提取了 id 点赞和转发数，具体数据请运行 wrangle_act_cn.ipynb 查看

2.评估

在数据质量和数据整洁度两个方向进行了评估，找出数据集中的部分问题，并在评估部分做了记录。根据项目要求在数据中去掉了转发的记录，只保留了回复的评分和原始的评分记录。三个文件都是关于评分的，于是合并成了一个表。twetter_archive_enhanced.csv 的原始文件中的 text 列包含了评分，由于评估过程中发现原始的评分数据某个记录有错误，所以重新在原始的 text 中提取了评分数据(分子和分母)。

3.清理

清理过程没有严格按照清理缺失值->清理整洁问题->清理其他质量问题的步骤，因为在清理整洁问题时发现需要先解决质量问题。由于在评估时已经大概想好了要问的问题，有些看似没有用的列没有着重清理。