整理报告

一、数据背景

本次清洗的（分析和可视化）的数据集是推特用户 @dog_rates 的档案，也叫做 WeRateDogs。推特用户 WeRateDogs 以诙谐幽默的方式对人们的宠物狗评级。这些评级通常以 10 作为分母。但是分子呢？分子一般大于 10。 11/10、12/10、13/10 等，为什么呢？因为 "Brent 它们是好狗。" WeRateDogs 拥有四百多万关注者，曾受到国际媒体的报道。[1]

二、数据采集和载入

本次数据清洗和分析的平台采用的是 python。首先，对数据集进行采集。由于一些原因，不能登陆推特，因此直接下载了 tweet_json.txt 文件。剩下的 image_prediction.tsv 和 twitter_archive_enhanced.csv 文件则通过 request 包进行下载。之后调用 pandas 包中的函数将数据集载入 dataframe 中。

三、数据初探

tweet_json 的每一行即是每条推特数据。twitter_archive_enhanced 是从 tweet_json 提取到的一些信息，其中较为重要的包括：推特 id、评价文本（text）、姓名（name）、评分（rating_numerator 和 rating_denominator）和评级（doggo、floofer、pupper 和 puppo）。其中，姓名、评分和评级都是从评价文本中提取到的，但是有的条目明显提取错误，因此要重新提取。表格中评级的结构也较为冗余，因此也要调整。另外 tweet_json 中的转发数量和点赞数量却没有包括进去，处理的时候也要进行收集。
image_prediction 主要包括 tweet_id 和应用神经网络对图片中狗狗品种的预测，主要包括了最有可能的三个预测，需要注意的是，有时模型并没有将图片中的狗辨认为狗，例如辨认成了“蚊帐”，因此提取信息的时候要尤为注意。

四、数据清洗

在进行了初步分析后，发现的问题以及解决方式为：
首先，因为我们分析的是原始数据，即不包含转发的数据，因此在 archive 的 retweeted 记录不为空（转发条目）的都要删除。
archive 表格：
- tweet_id 类型应为 str。
利用 astype 函数修改类型为 str。

---

[1] 简介来自优达学城的项目介绍：
https://classroom.udacity.com/nanodegrees/nd002-cn-advanced-career/parts/baa15c02-fe97-4999-9bc0-6cd7578d037e/modules/6029470c-61d2-465a-a252-ce5d11b21932/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/5e3db54a-1a5f-41a6-8e20-fd99f201861d

- in_reply_to_status_id、in_reply_to_user_id、retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp 类型不对。
这些特征感觉对于分析暂时用处不大，而且有的特征缺省值过多，因此考虑直接删除。

- in_reply_to_status_id、in_reply_to_user_id、expanded_urls 及转发的信息缺省。（无法更改，转发信息也无需更改）

- timestamp 的命名与 json 不统一，且类型错误。
删除，只以 tweet_id 为唯一的识别特征。

- name 有的条目提取错误，且有缺省值，但类型不是 np.nan。
运用 str.extract 函数和正则表达式重新提取姓名。

- doggo、floofer、pupper、puppo 有缺省值，但类型不是 np.nan。
- doggo、floofer、pupper、puppo 应综合为一列。
- doggo、floofer、pupper、puppo 有的条目并列存在。
删除这四列，并用 str.findall 函数和正则表达式重新提取“stage”信息，再用 apply 函数针对 stage 这一列做出修改，将未提取出来的修改为 NaN，有多重地位的将这些地位信息用逗号连接。

- text 的文本模板有的没有统一。（无法更改）

- rating_numerator 存在异常值。
- rating_denominator 通常应为 10，但是有 23 条记录不是。
这两列一起修改。先通过正则表达式和 extract 函数提取两项评分。将正确提取的 rating 转换成 int 类型。再次检查异常值，查看异常值是否有异常规律，并且检查 text，然后分别进行分析和处理。对于 numerator，最后有实在无法确定的（text 中没有体现，或者评分者的评分太异常，则用众值来进行修正）。

prediction 表格：
- 缺少记录。(无法补充)
- tweet_id 类型应为 str。
利用 astype 函数修改类型为 str。

- 应补充‘kind’列，且类型为‘category’。然后依据 tweet_id 加入 archive 表格中。
利用 apply 函数提取新的 kind 列，然后利用 pd.merge 函数以 tweet_id 为识别特征，将 kind 列加入到 archive_clean 表格中，因为我们只分析有图片信息的狗狗，因此在 merge 时 how 为 inner。

json 表格：
- 缺少记录。(无法补充)
- created_at 类型错误。

利用 astype 函数修改类型为 timestamp。

- id_str 应更名为 tweet_id，和 archive、prediction 表格保持一致。
利用 rename 函数将 id_str 更名为 tweet_id。

- full_text 更名为 text，与 archive 保持一致。
利用 rename 函数将 full_text 更名为 text。

- 将 favorite_count 和 retweet_count 加入 archive 表格中，类型为 int 即可。
先利用 astype 函数将这两列的类型改为 int，再提取 json_clean 中的 tweet_id、favorite_count 和 retweet_count 列，形成新的 dataframe，然后利用 pd.merge 函数根据 tweet_id 为识别特征，将 favorite_count 和 retweet_count 特征加入 archive_clean 中。