Twitter 数据整理

一、收集

1. 直接下载：twitter-archive-enhanced.csv，读取该数据集，命名为 df1。
2. 用 Resquests 库进行编程下载，下载 URL 地址为：https://raw.githubusercontent.com/udacity/new-dand-advanced-china/master/%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97/WeRateDogs%E9%A1%B9%E7%9B%AE/image-predictions.tsv，读取该数据集，命名为 df2。
3. 从 tweet_json.txt 文件中读取每个推特 JSON 数据，利用 json 库函数读取 json 文件中的 id_str、retweet_count、favorite_count 信息，将读取到的所有信息依次叠加到数据集 df3。

二、评估

1.质量

df1 数据集:

1.1 rating_denominator 列的值并非全是 10，还有其他值。

1.2 name 列有多个名字为'a'。

1.3 狗狗 stage 的 4 列 doggo\floofer\pupper\puppo，且有大量数据缺失，以及存在同时处于 2 个 stage。

1.4 转发的 tweet 即 retweeted_status_id 有 181 条，需要删除。

1.5 source 列的信息冗余，只需要提取'> <'之间的有用信息。

1.6 tweet_id 列数据类型是 int，应该转化为 str。

1.7 把 rating_numerator、rating_denominator 两列数据相除，得到 rating 列，除去 rating 列存在的异常值。

df2 数据集:

1.8 jpg_url 列有 66 个重复。

1.9 df2 的 tweet_id 列 int 数据应转为 str。

df3 数据集:

1.10 df3 的 tweet_id 列 int 数据应转为 str。

2.整洁度

2.1 df1 狗狗的阶段 stage:doggo\floofer\pupper\puppo 4 列可以整合成一列。

2.2 rating_numerator、rating_denominator 两列数据可以去掉，只需留下 rating 列即可。

2.3 3 个数据集有共同的 tweet_id,可以合并成一个数据集。

三、清洗

分别复制 3 个数据集，为 df1_clean、df2_clean、df3_clean。

针对以上评估中标注的序号，依次对号填写清洗过程如下:

1.质量

1.1 使用 extract()函数及正则表达式重新提取 rating_numerator 和 rating_denominator

的数值。

1.2 使用 extract()函数及正则表达式重新从 text 列中提取宠物的 name。

1.3 使用 findall()从 text 中重新提取狗狗的地位分类，形成新的一列 df['stage']。

1.4 删除转发的推特的行的办法是，只需要留下 retweeted_status_id 为空值的行。

1.5 使用 extract()及正则表达式提取'><'之间的关键信息。

1.6 使用 astype('str')将 df1_clean 中 tweet_id 的整数型数据转化成字符串数据。

1.7 添加 rating 列，其数值为 rating_numerator/rating_denominator，查看 rating 列数值分布情况，并去除异常值

1.8 使用 drop_duplicates()，除去 jpg_url 列的重复数据。

1.9 使用 astype('str')将 df2_clean 中 tweet_id 的整数型数据转化成字符串数据。

1.10 使用 astype('str')将 df3_clean 中 tweet_id 的整数型数据转化成字符串数据。

2.整洁度

2.1 因为已经添加了 stage 列，所以只要删除 doggo\floofer\pupper\puppo 4 列即可。

2.2 已经添加了 rating 列，所以删除 rating_numerator、rating_denominator 两列即可。

2.3 先合并 df1_celan 与 df2_clean 方式是 inner，合并后的数据集与 df3_clean 以 left 方式合并，生成数据集 df_clean。

整理完成后的数据集 df_clean 保存到 CSV 文件，命名为 twitter_archive_master.csv，以备分析。