清洗过程

获取数据后，有twitter、twitter_extra、image_predictions3个dataframe。首先是需要解决什么问题，哪些数据是需要的。然后，3个dataframe以什么为基准合并。检查发现有相同的tweet_id，以此为基准合并dataframe。然后开始系统清洗。去找质量问题和整洁度问题。

根据思路整理出需要验证的问题：

1、在twitter、twitter_extra、image_predictions中tweet_id是不是唯一的。
验证3个dataframe中个tweet_id是否唯一，3个dataframe是否有交集。
2、检查twitter、twitter_extra、image_predictions是否都含有同tweet_id。
3、检查twitter中丢失数据量。通过检查非空的数量。（in_reply_to_status_id、in_reply_to_user_id、retweeted_status_id、retweeted_status_user_id、retweeted_status_timestamp）
4、检查是否所有tweet_id都能对应有url。
5、检查狗名是否正确
6、检查每列中格式是否正确
7、检查twitter dataframe中rating_numerator、rating_denominator是否正确
8、清除twitter_extra中retweet_count为0的项
9、image_predictions中p1、p2、p3中明显错误的项目

验证问题后发现问题如下：

数据质量问题

（1）twitter的dataframe中不需要处理的数据in_reply_to_status_id、in_reply_to_user_id、r（1）twitter的dataframe中不需要处理的数据in_reply_to_status_id、in_reply_to_user_id、retweeted_status_id、retweeted_statu。

（2）twitter的dataframe中狗的name字段含有None,a,an,a,such,quite,the，需要去掉或填充None类型。

（3）twitter的dataframe中狗的name字段None为字符串格式，需要修改为python中None。

（4）twitter的dataframe中timestamp、retweeted_status_timestamp应该为时间格式。

（5）twitter的dataframe中doggo、floofer、pupper、puppo中None不是str应该为None格式

（6）twitter的dataframe中，有的推送的内容中含有2条狗，rating_numerator、rating_denominator有2个评分，这种情况就取2个多平均值。

（7）twitter的dataframe中，有的推送的内容中含有多个“数字/数字”格式内容，提取内容了错误的rating_numerator、rating_denominator。

（8）twitter的dataframe中，有的rating_denominator>10的部分数据中，有的有用数据，需要修改。例如为99/90，可以理解为9条狗总分，修改为11/10

（9）twitter_extra的dataframe中retweet_count为0的异常值

（10）image_predictions的dataframe中p1、p2、p3中明显错误的异常值

数据整洁问题

（1）所有的数据在多个dataframe中，需要合并到一个dataframe进行分析。以twitter、twitter_extra、image_prediction中对应id为基准合并。

（2）twitter的dataframe中expanded_urls含有多个url，需要清理掉

（3）twitter的dataframe中oggo、floofer、pupper、puppo需要合并成1列