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
10、观察检查文档中含有html的项，必须要的项需要清洗

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、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rating_numerator、rating_denominator存在异常值，需清除
（7）twitter_extra的dataframe中retweet_count为0的异常值
（8）image_predictions的dataframe中p1、p2、p3中明显错误的异常值
（9）twitter的dataframe中source含有html元素，需要去掉

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