TWITTER数据爬取，清洗与分析

数据评估结果和清洗思路

数据质量所存在的问题

推特档案数据：

- 01.我们只需要含有图片的原始评级，存在转发的条目需要删除
- 02.多列数据缺失严重
- 03.source table: 数据需要提取，整理成Ipone客户端,Web客户端,VINE, tweet deck四个选项
- 04.评分的提取不太准确
- 05.name table: 有的name提取不正确
- 06.timestamp table : 数据类型转换为时间类型
- 07.有时同一条数据中， 含有多个狗的地位数据， 正常情况下只能有doggo，floofer，pupper，puppo 中的一个

额外附加数据：

- 08.我们需要将预测结果不是狗的数据删除，且p1的预测概率远远大于p2和p3, 含有p2,p3的数据列也需要删除

数据清洗思路

1. 我们只需要含有图片的原始评级，存在转发的条目需要删除

如果retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp这三列中有非空值，说明是转发推特, 我们需要将该行删除。

我们将依次提取出与转发相关的三列为空的行，完成筛选后，将这三列删除。

最后，我们需要删掉不含有图片的行。

2.多个列数据缺失严重

有数据较多缺失的列与核心数据分析关联不大，故我将‘in_reply_to_status_id’，‘in_reply_to_user_id’列删掉¶。

3. source栏数据需要提取成Ipone客户端,Web客户端,VINE, tweet deck四个选项¶

我们利用pandas.series.str与正则表达式，重新建立规则，将标签Twitter for iPhone等从HTML格式内提取出来¶。

4. 评分提取不太准确

两栏的特殊值统计全是整数，未能提取text中浮点数评分， 我们利用正则表达式重新建立新的提取规则。之后查看特殊值数据所在的行，与原文字对比，手动修改一些特殊值。

5. name table: 有的name提取不正确

显示name的特殊值列，将名字开头是小写的数据和None值删掉。

6. timestamp table : 数据类型不正确¶

数据类型转换为时间类型， pd.to_datetime(df_copy[‘timestamp’])。

7. 有时同一条数据中， 含有多个狗的地位数据， 正常情况下只能有doggo，floofer，pupper，puppo 中的一个

写了一个函数，遍历dataFrame,找出同时有多个地位数据的行， 与文字栏对比，进行修改。¶

8.我们需要将预测结果不是狗的数据删除，且p1的预测概率远远大于p2和p3, 含有p2,p3的数据列也需要删除

9.三份数据需要合并到同一表格内

将三个dataFrame融合为一个并重建索引,对图片预测数据集进行merge时选择inner方式，以完成对无图片的推特需要删除，因为没有图片就不会有图片预测数据。

10. 推特数据中，doggo，floofer，pupper，puppo栏应当删掉，建立新的地位栏，并填入相应数据¶

遍历每一行，提取出地位数据并添加到新的status列中，并删去原有的doggo等四列。

此时，我们已经完成了基本的数据清洗，将所有的数据保存至twitter_archive_master.csv文件内