本数据集收集的是一个推特博主叫WeRateDog的每条推特。（超级有意思的博主）这个博主就是对狗狗打分 传送门->https://twitter.com/dog_rates

 Image Name

** 我在做的事情就是 **

- 1.数据进行收集
- 2.评估数据质量问题和清洁度（结构）问题，
- 3.清洗数据集 ☺

收集

- 收集文件 1 保存为 twitter_archive_enhanced
- 收集文件 2 保存为 tweet_json
- 收集文件 3 保存为 image_predictions

评估

目测评估

twitter_archive_enhanced数据中有大量的空值
twitter_archive_enhanced数据中有些列是不需要用的比如 in_reply_to_status_id

tweet_json数据中包含了大量的空值

image_predictions数据中对狗狗品种认定，但是给出了三种结果，取其中最高的即可。

质量问题

`twitter_archive_enhanced` 表格

- 转发的(即retweets)的数据和回复的数据需要删除
- source 列应当只包含iphone web内容
- timestamp 列的时间数据不是datatime类型
- name列数据异常
- doggo floofer pupper puppo 列数据缺失
- rating_denominator列中有异常值（等于0）
- 没有图片的数据需要删除
- tweet_id应当为object类型

`tweet_json` 表格

`image_predictions` 表格

整洁度问题

- 合并`twitter_archive_enhanced`、`tweet_json`和`image_predictions`。以`twitter_archive_enhanced`为主要数据集合，保留`favorite_count`和`retweet_count`以及`jpg_url`、`img_num`、`p1`、`p1_conf`和`p1_dog`
- doggo floofer pupper puppo 可以合并成一个列

清理

问题描述一

定义

in_reply_to_user_id 有78条，这些数据是回复。将其删除。
retweeted_status_user_id 是转发数据。将其删除
先查找出retweeted_status_user_id不为空的行，然后批量将其删除。in_reply_to_user_id也是如此

问题描述二

定义

source列应当只包含iphone web等内容
解析出每一个html标记语言中的内容 `<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>` 提取出Twitter for iPhone

问题描述三

定义

timestamp 列的时间数据不是datatime类型
用pd.astype()方法直接转化数据类型

问题描述四

定义

`name`列数据异常。根据`text`列修补

- 使用正则表达式匹配名字，并且名字第一个字应该是大写.
- 如果`text`中没有写出名字那么用空字符串代替

问题描述五

定义

`doggo floofer pupper puppo`等 列数据缺失

- 查询text中是否存在对应的单词（doggo floofer pupper puppo）
- 存在则放入对应的列

问题描述六

定义

rating_denominator列中有异常值（等于0）
使用正则表达式从新提取分数

问题七

定义

没有图片的数据需要删除

- 直接用drop删除img列为空的行

问题八

定义

tweet_id应当为object类型

- 使用astype直接转化数据类型

问题描述九

定义

合并twitter_archive_enhanced、tweet_json和image_predictions。以twitter_archive_enhanced为主要数据集合，保留favorite_count和retweet_count以及jpg_url、img_num、p1、p1_conf和p1_dog

合并jpg_url、img_num、p1、p1_conf和p1_dog的任务在**问题描述七**中完成了

问题描述 十个

定义

`doggo floofer pupper puppo` 可以合并成一个列

- 直接用字符串拼接方法将列合并成一个字符串，然后存放在`nickname`列中。