数据收集

通过编程下载image-prediction.tsv

通过tweepy API 收集推文点赞和转发的数据

把json格式的文本转换成csv，并保存为名叫tweet_favourite_retweet.csv的文档

数据评估清理

1.先把数据通过copy（）函数保存副本

质量

twitter

- source 含有格式字符 （使用正则表达式对字符串提取文本内容，把格式字符去掉）
- 一部分狗狗status 提取错误，有多种status的狗狗（筛选出有多种status的狗狗，通过查看对应text，修改正确的status，对于一条推文有两种狗狗的一小部分数据，保留不变）
- timestamp 时间格式错误 （用pandas的to_datetime函数修改str to date time）
- name 名字提取错误，有英文助词‘a’，‘an’，etc （用str.extract把This is |Meet |name is |Say hello to |named 后面狗的名字提取出来）
- in_reply_to_status_id、retweeted_status_user_id、retweeted_status_id、retweeted_status_user_id 数据格式错误 （这几列在后面已经删除，不需要修改）
- in_reply_to_status_id,in_reply_to_user_id,，retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,只需要原始数据，不需要转发的数据（把in_reply_to_status_id,in_reply_to_user_id,retweeted_status_user_id、retweeted_status_id、retweeted_status_user_id为non-null的所有条目删除（non-nullin_reply_to_status_id,in_reply_to_user_id,的retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp的所有行都被删除之后，这五列都为空，然后移除掉这五列））
- 部分rating_numerator/rating_denominator 提取错误,rating_numerator 没提取浮点数（通过正则表达式把rating_numerator的小数部分提取出来,，同时re.findall 提取第二次出现的xx/xx，通过打印text判断多次出现的值，选择正确的评分）
- tweet_id列是int数据类型（通过astype（'object'）函数，把int改成object类型） #### image_pre
- p1/p2/p3 名字首字母大小写没有统一 （运用str.lower(),把名字改成全小写的）
- p1_conf/p2_conf/p3_conf 数据不是百分比表示 （运用apply函数对 p1_conf/p2_conf/p3_conf的数据都乘以100）
- p1，p1_conf,p2,p2_conf,p3,p3_conf描述性不够强（通过rename函数修改列名，使其更具描述性,confidence加上%符号指明数值的表示方法）

清洁度

- rating_numerator与rating_denominator分开两列（把rating_numerator/rating_denominator合成一列'rating',作为实际WeRateDog评出的得分值）
- twitter/image_pre/tweet_favourite_retweet分开不同的表格,缺favourite_count，retweet_count （通过pd.merge函数，how='inner'(可以删掉没有图片的推文条目)，将tweet_favourite_retweet的favourite_count,retweet_count，image_pre_clean,按照tweet_id合并到twitter_id_clean，编程下载favourite_count，retweet_count数据）
- 狗的status分开独立的列（ 把doggo，floofer，pupper，puppo合并成一列status，重复状态的用“，”隔开表示两种不同的状态，然后删除doggo，floofer，pupper，puppo）