清洗报告

1.收集

通过编程下载收集数据分别保存到 3 个数据表格(padas DataFrame 类型)中：tweet 表格、predictions 表格、tweet_json 表格

2.评估

1.1 目测评估

通过对三份文件数据进行目测观察，发现以下数据问题：

（1）狗的名字不准确，有 None、a、an、the 等非名字的字符被提取；

（2）有的是转发的数据；

（3）狗的地位有 4 列，应该属于一个变量；

（4）第三个数据文件中 retweet count 和 favorite count 是第一个数据的一部分。

1.2 编程评估

通过编程进一步发现的问题如下：

（1）有 181 条转发数据和 78 条回复，项目要求是不含转发数据；

（2）评分数据的分母不都是 10；

（3）tweet_id 数据类型是整型；

（4）图片预测结果对应的图像编号有 4，而 4 的数据不存在。

综上，存在问题如下：

质量问题

`tweet` 表格

1. 狗的名字 name 数据不正确，有 None、a、an 等非名字的字符 【处理:正则表达式重新提取】

2. 有转发的数据 retweeted_status_user_id 列中有非空 181 条 【处理：删除】

3. 狗的评分数据部分不准确 【处理：重新提取】

4. 评分数据分母不一致，不方便比较 【处理：统一计算为以 10 为分母】

5. tweet_id 数据类型是整型 【处理：修改为字符串型】

`predictions` 表格

1. 1. tweet 数据 2356 个而 predictions 中 2079 个，说明有些狗没有图片
【处理：合并两个数据集就能得到都是有图片的数据】

2. 预测图片不是狗 【处理：合并图片预测数据，提取预测结果是狗的数据】

`tweet_json` 表格

1. tweet_json 表格中变量名 id 与其他表格中的变量名 tweet_id 不一致 【处理：修改为 tweet_id】

清洁度问题

1. `tweet` 表格中的狗的地位 doggo,pupper,puppo,floofer 是同一属性的变量，而不是 4 个变量却占用了 4 列 【处理：删除这 4 列并重新提取狗的地位存储 status】

2. `tweet_json` 表格中的 retweet_count 和 favorite_count 两列是 tweet 表格中的一部分 【处理：合并到 `tweet` 表格中】

3.清洗

通过编程对以上数据问题进行清洗,清洗后得到干净的主数据集 df_clean(padas DataFrame 类型)存入 df_clean.csv 文件中