WeRateDog 项目数据清洗

1. 数据收集

1.1. WeRateDogs 推特档案数据直接导入 twitter-archive-enhanced.csv 文件

1.2. 通过 URL 编程下载推特图像的预测数据，即根据神经网络，对出现在每个推特中狗的品种（或其他物体、动物等）进行预测的结果。

1.3. 通过 Tweepy 库下载额外的推特数据，至少要包含转发数（retweet_count）和喜欢数（favorite_count）

2. 数据评估

通过目测评估和编程评估的方式对数据进行质量及整洁度的评估

2.1. 质量：

2.1.1. twitter_archive 表格

- tweet_id 是整数格式
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 列有太多缺失值
- 只需要含有图片的原始评级 (不包括转发)
- timestamp 是字符串格式不是时间格式，并且都有+0000
- source 列太多多余的信息，如 http 连接地址
- expanded_urls 列中有缺失值或者一行中内容重复
- rating_numerator 和 rating_denominator 的最大值最小值异常
- name 有很多 None 或者 a/an
- doggo, floofer, pupper, puppo 中有很多 None
- 不需要 2017 年 8 月 1 日后的数据

2.1.2. *image_predictions* 表格

- tweet_id 和 img_num 是整数格式
- jpg_url 列有重复项

2.1.3. tweet_extra_data 表格

- tweet_id 是整数格式

2.2. 整洁度：

- 狗狗阶段分为了四列 (doggo, floofer, pupper, puppo)
- 三张表格可以整合

3. 数据清理

3.1. 数值缺失问题

- 只需要含有图片的原始评级 (不包括转发)：删除包含 in_reply_to_status_id, in_reply_to_user_id 的行
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 列有太多缺失值：删除列

3.2. 整洁度

- 狗狗阶段分为了四列 (doggo, floofer, pupper, puppo)：重新从文中提取 stage 的信息，放入 stage 列中，删除原有的四列
- 三张表格可以整合：放在最后合并

3.3. 质量

3.3.1. Twitter_archive 表格

- tweet_id 是整数格式：修改 tweet_id 的格式为字符串
- timestamp 是字符串格式不是时间格式，并且都有+0000：用 pd.to_datetime 转换成时间格式，并只选取+0000 之前的字符
- source 列太多多余的信息，如 http 连接地址：从 source 中截取正确的信息，比如 Twitter for iphone, Twitter Web Clien 等
- expanded_urls 列中有缺失值：保留 expanded_urls 为 notnull 的行
- rating_numerator 和 rating_denominator 的最大值最小值异常：从文中重新提取分数，检查异常值并修改
- name 有很多 None 或者 a/an：重新从文中匹配狗狗名字
- doggo, floofer, pupper, puppo 中有很多 None：四列已删除
- 不需要 2017 年 8 月 1 日后的数据：在合并三张表后删除 2017/8/1 之后的数据

3.3.2. image_predictions 表格

- tweet_id 是整数格式：修改 tweet_id 的格式为字符串
- jpg_url 列有重复项：用 drop_duplicates 删除重复项

3.3.3. tweet_extra_data 表格

- tweet_id 是整数格式：修改 tweet_id 的格式为字符串

4. 数据合并及储存

- 合并三张表格并删除 2017/8/1 之后的数据
- 储存数据到 twitter_archive_master.csv 中