一、数据集基本情况简介

- **twitter_archive_enhanced.csv**：推特用户 @dog_rates 的档案，推特昵称为 WeRateDogs。WeRateDogs 是一个推特主，他以诙谐幽默的方式对人们的宠物狗评分。这些评分通常以 10 作为分母。但是分子则一般大于 10：11/10、12/10、13/10 等等。twitter_archive_enhanced.csv 里面包含 2356 条包含评分的推特数据。

字段名称	描述	字段名称	描述
tweet_id	推特 ID	in_reply_to_status_id	回复 ID
in_reply_to_user_id	被回复的 ID	timestamp	发送推文的时间戳
source	发送推文来源（使用设备）	text	推文内容
retweeted_status_id	转发推文的 ID	retweeted_status_user_id	转发用户 ID
retweeted_status_timestamp	转发时间戳	expanded_urls	推文链接
rating_numerator	评分分子	rating_denominator	评分分母
name	狗名	doggo	大狗
floofer	毛好看的狗	pupper	小狗
puppo	青春期的狗		

- **image-predictions.tsv**：对狗狗种类进行分类的神经网络，运行这份推特档案中的所有图像。获取的结果为一份图像预测结果表格，其中包含了预测结果的前三名，推特 ID，图像 url 以及最可信的预测结果对应的图像编号（由于推特最多包含 4 个图片，所以编号为 1 到 4）

字段名称	描述	字段名称	描述
tweet_id	推特 ID	jpg_url	预测的图像资源链接
img_num	最可信的预测结果对应的图像编号	p1	算法对推特中图片的一号预测
p1_conf	算法的一号预测的可信度	p1_dog	一号预测该图片是否属于“狗”
p2	算法对推特中图片预测的第二种可能性	p2_conf	算法的二号预测的可信度
p2_dog	二号预测该图片是否属于“狗”	p3	算法对推特中图片预测的第三种可能性
p3_conf	算法的三号预测的可信度	p3_dog	三号预测该图片是否属于“狗”

- **tweet_json.txt**：每条推特的额外数据，需要从中提取推文的转发量和点赞量，提取出来的数据集名为 tweet_additional，里面包括 id、retweet_count(转发数)、favorite_count(点赞量)

二、数据评估

1. 质量

***twitter_archives* 表格**

- 数据类型问题
  - timestamp 列的数据类型应该为 datetime
  - doggo、pupper、puppo、floofer 列的数据类型应为分类数据
  - tweet_id 列数据应为字符串，image_predictions 数据集、tweet_additional 数据集也有同样的问题
- 数据抓取错误
  - name 列存在大量空值，以及'the'、'a'等不正确的名字
  - 对狗评级数据的抓取存在错误，有些推文中有多个‘/’形式的数据
  - 有些推文里描写了狗的两种 status，有 14 行数据记载了狗的两种 status 分类
- 无关、冗余数据
  - 有些推文主题并未涉及狗，抓取的评级分数为日期简写或者其他与评级无关的数据。
  - expanded_urls 列有 59 个空值，有 137 个重复值
  - 数据集中包含转发数据 181 个，数据重复数据
  - 删除 source 列中多余的字符
- 其他
  - 有些分母评级数据显著高于 10(如'150'、'170')，这是对多只狗一起打分的结果
  - 给狗的评分进行分类
  - 存在 1976 个狗的 status 数据缺失(无法处理)

***image_predictions* 表格**

- jpg_url 列有 66 个重复项
- 三次预测结果可以综合为一个

2. 整洁度

- twitter_archives 数据集中 doggo、floofer、pupper、puppo 四列都是描述狗的 status 的，应该并为一列
- tweet_additional 数据集的 columns'id'应该改为'tweet_id'
- 三个数据集合并为一个数据集

三、数据清理

1. 质量问题

处理无关、冗余数据

- 删除 twitter_archives 数据集中所有转发的数据
- 删除 twitter_archives 数据集中 expanded_urls 的空值和重复值的数据
- 删除与狗无关的推文数据，其 tweet_id 为 749981277374128128、810984652412424192
- 删除 image_predictions 数据集中 jpg_url 列的重复值和空值
- 用 extract()提取 twitter_archives 数据集 source 列中具有'><'形式的字符

处理数据抓取错误

- 优化正则表达式，从 twitter_archives 数据集 text 列中重新提取狗的名字，名字第一个字应为大写，前面一般跟有'This is'、'Meet'、'Say hello to'等语句，对于一条推文中出现两只或三只狗的名字，第二、三个名字前面一般为'and'或者',。
- 从 twitter_archives 数据集 text 列重新抓取狗的状态，筛选出推文中存在多个状态的数据进行查看。通过观察，有些推文对两只状态不同的狗甚至三只打了同样的评分，还有一些推文中只有一条狗，但却描述了狗的两种状态，因此进行人工判断，找出只有一条狗却描述了多种状态的数据，将其数据存入字典 misread_status（字典的键为 twitter_archives 数据集，value 为狗的正确的状态），然后在 twitter_archives 数据集中新建一列 status，将狗的状态全部抓取进 status 列，有多个状态的用&连接，利用 misread_status 字典将 status 列中错误的数据修正。同时删除 doggo、floofer、pupper、puppo 列，解决了整洁度问题里面 twitter_archives 数据集中 doggo、floofer、pupper、puppo 应该并为一列的问题。
- twitter_archives 数据集 text 文本中有些分子为小数形式，而原数据值提取了小数点以后的数据，因此修改正则表达式，重新提取数据；对于推文文本存在多个‘/’形式的数据情况，将其筛选打印出来进行人工修正

处理数据类型问题

- 使用 pd.to_datetime 将 twitter_archives 数据集中 timestamp 列数据类型转换为 datetime
- 使用 astype 将 twitter_archives 数据集中 status 数据类型转换为分类数据
- 将三个数据集中 tweet_id(tweet_additional 数据集中为 id)列转换为字符串

其他质量问题

- 对于分母评级数据显著高于 10 的数据，通过观察可以看到分母评级不等于 10 的数据均为 10 的倍数，因此将分子分母评级数同时除以一个数，使分母评级为 10
- 为方便后续分析，在 twitter_archives 数据集中新建一列 rate_group 给狗的评分进行分类，大于等于 10 的为 high，小于等于 5 的为 low，中间评分为 medium
- image_predictions 数据集中，从中选中置信度最高的一次预测作为狗的品种的预测

2. 整洁度问题

- 使用 rename 将 tweet_additional 数据集中的'id'改为'tweet_id'
- 用 pd.merge()将三个数据集合并为一个

四、存储清理后数据

将清理后的数据保存为 twitter_archive_master.csv