WeRateDogs Twitter 数据清理报告

数据清理过程主要分为三步：首先收集数据，其次对数据进行评估，评估部分主要分为目测评估和编程评估两种手段，然后总结出评估出的所有问题，第三步针对第二步评估出的问题进行一一对应处理，即数据清理。

1. 收集

- twitter_archive_enhanced.csv 已提供。
- image-predictions.tsv，从以下 URL 下载。

  URL ：https://raw.githubusercontent.com/udacity/new-dand-advanced-china/master/%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97/WeRateDogs%E9%A1%B9%E7%9B%AE/image-predictions.tsv
- tweet_json.txt 已提供。

收集到上述三个数据集之后，通过 Python 的 Pandas 库生成三个 DataFrame 格式的表格:

- twitter_dog 表格
- image_prediction 表格
- tweet_json 表格

2. 评估

- 通过目测评估和编程评估汇总出以下数据问题，并从质量和整洁度两个角度进行分类。

2.1 质量

- twitter_dog 表格

- 根据项目要求，需要删除转发数据，只保留原始 tweets
- 根据项目要求，只保留有图片的 tweets

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 这几列有很多的空值
- 有些行数据狗狗四种地位的值均为 None，表示缺少地位值
- timestamp 不是 datetime 格式
- 从记录的 text 可以看到分子有 13.5 这样的数，所以 rating_numerator 类型应该为 float
- 评分分母通常是 10，不是 10 的那些数据可能是文本提取错误
- source 列的值为 html 的 a 标签格式
- source 总共有四种分类，应该转为 category 数据类型

- image_prediction 表格

- 狗狗图片的预测结果 p1, p2, p3 三列有的首字母大写，有些没有，不符合一致性

- tweet_json 表格

- contributors, coordinates, geo 三列没有非空值，place 只有一个非空值
- id 列名对应于 twitter_dog 的列名 tweet_id，列名应该统一
- tweet_json 表格的 in_reply_to_status_id, in_reply_to_user_id, source 三列在 twitter_dog 表格中已存在

2.2 整洁度

- twitter_dog 表格中狗狗的地位（stage）分成了多列，可以合并为一列 stage
- image_prediction 表格的图片预测结果和 tweet_json 表格的转发数和喜爱数是 twitter_dog 表格的一部分
- twitter_dog 表格需要增加新的一列计算评分比值

3. 清理

- 清理过程主要是针对上一步评估出的所有问题制定出对应的处理方法进行数据清理，但需要首先进行备份。

3.1. 备份三个数据集

3.2. 清理质量问题

- twitter_dog 表格

- retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 这三列属于转发信息，也就是说这三列为空值的属于原始 tweets，所以删除这三列为空值的行数据即可，原始数据为 2356 条，删除完应该剩余 2175 条数据
- 由于 image_prediction 表格没有空值，将 image_prediction 表格与 twitter_dog 表格以 inner jion 的方式合并即可只保留有图片的 tweets
- 由于现在 retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 这三列没有非空值，可以直接删除这三列，in_reply_to_status_id, in_reply_to_user_id 两列暂时保留，不作处理
- 观察发现，很多文本里确实就没有狗狗的地位值，所以缺少地位值是正常情况，暂不做处理
- 将 timestamp 转为 datetime 格式
- 从 text 中重新提取评分数据
- 观察评分分母值不为 10 的数据，如是提取错误，进行手动替换
- 提取出>和之间的内容
- 将 source 转为 category 数据类型

- image_prediction 表格

- 由于 image_prediction 表格已经与 twitter_dog 表格合并，直接在 twitter_dog 表格中将 p1, p2, p3 三列改为全小写形式

- tweet_json 表格

- 删除 contributors, coordinates, geo 三列， place 暂时保留
- 修改列名 id 为 tweet_id
- 为避免重复，删除这三列：in_reply_to_status_id, in_reply_to_user_id, source

3.3. 清理整洁度问题

- 由于有的一个狗狗有两种地位，所以重新从文本中提取值作为新的一列 stage，并删除原来的四列：doggo, floofer, pupper, puppo
- 计算 rating_numerator 和 rating_denominator 比值作为新的一列 rating，并删除 rating_numerator 和 rating_denominator 两列
- 由于上面的处理，image_prediction 表格已经合并到 twitter_dog 表格，由于 tweet_json 含有较多列，我们暂且先根据项目需求提取 favorite_count 和 retweet_count 两列合并至 twitter_dog 表格

3.4. 存储清理后的主数据集为 twitter_archive_master.csv