WeRateDogs 数据整理报告

本次项目，整理的数据集是推特用户 @dog_rates 的档案，WeRateDogs 是推特主，他以诙谐幽默的方式对人们的宠物狗评分。本次整理工作内容如下：

1、数据收集

在本阶段，共收集了 3 份数据，其中一份是 WeRateDogs 推特档案信息（twitter-archive-enhanced.csv），这一部分数据可直接从 GitHub 上下载得到；另一份数据集是推特图片预测文件（image-predictions.tsv），利用 Python 的 Requests 库和告知的 URL 来获取；还有一份是推特档案的附加信息 add_infos.csv，该数据集是是从已知的 JSON 文件中通过编程的方式获取的。

2、数据评估

在本阶段，通过视觉评估和编程评估两种方式对已经获取到的数据集进行了评估。提出了若干关于数据整洁度和数据质量的若干问题：

（1）质量问题

1) archives 表

- 回复转发相关列（如 in_reply_to_status_id 等）有缺失值，扩展 URL 列有缺失值

- 宠物狗等级列（doggo 等）信息不准确

- 宠物狗名字列（name）不准确，除了正确的名字之外，还有 None, an,a 这些不正确的值

- source 列代表了推特信息的来源，是 HTML 文本的样式，只需要其中的内容即可

- 回复转发相关列（如 in_reply_to_status_id 等）数据类型不对，是浮点型，应将其改为 int 型，使其和 tweet_id 列的数据类型保持一致

- 评分相关列的值也不准确，有些列的值是小数（如 9.75），其数据类型应该也改为浮点型

- timestamp 和 retweeted_status_timestamp 列应该是 datetime 类型

- 经过合并后的等级列（stage）数据类型应该是 category

2) predictions 表

- p1,p2,p3 预测结果应该统一，首字母统一大写,分隔符统一用下划线‘_’

（2）整洁度问题

1) archives 表

-该表中包含了推特信息和推特中宠物狗的信息，应将其分为两个表，推特信息表和宠物狗信息表

-应该将 doggo，floofer，pupper，puppo 这几列合并为 stage 代表地位

2) addinfos 表

- 附加信息表中包含了推特信息的转发数和喜爱数，应同推特信息表合并

- 应该将 id 列改名为 tweet_id，使其和其他两个表统一

3、数据清理

该阶段主要针对评估阶段提出的问题进行整理，在清理之前先对原始数据进行备份。

（1）预处理阶段

根据项目要求，删掉转发的信息，以及 2017 年 8 月 1 日之后的信息。

（2）缺失值的处理

- 回复转发列及扩充 URL 列无法填充
- 名字，评分，等级列重新构建正则表达式从推特文本中提取

（3）整洁度处理

- 将 archives 表拆分成两个表推特信息表 tweet_archives_clean 和宠物狗信息表 dog_archives_clean
- 将 doggo，floofer，pupper，puppo 这几列合并为 stage 代表"地位"
- 将 addinfos_clean 中的转发数和喜欢数加到推特信息表中
- 将 addinfos_clean 的 id 列改成 tweet_id，和其他表统一

（4）质量问题处理

- source 列中将 HTML 文本中的内容提取出来
- 回复转发列的数据类型用 astype()转成 int
- 时间相关列的数据类型用 pd.to_datetime 转成 datetime
- 宠物狗的等级列转成 category 类型
- predictions 表中的 p1,p2,p3 预测结果统一，首字母统一大写，分隔符统一用下划线'_'

4、数据存储

将清理好的数据用 to_csv 存储至 twitter_archive_master.csv。

5、可视化分析及结论

（1）当前数据集中，转发最多的一条推特信息是关于拉布拉多寻回犬的，转发数高达 79116 次，喜欢数最多的一条推特信息是关于拉克兰猎犬的，喜欢数高达 132318 条。

（2）通过统计分析，发现当前数据集中，与宠物狗相关的推特信息的喜爱数要明显高于不含宠物狗的推特信息，但这两类的转发数相差不大

（3）当前数据集中，最受欢迎的宠物狗等级是 doggopuppo，平均转发数和平均喜爱数都要高于其他等级的宠物狗，而等级为 pupper 的宠物狗的转发数和喜爱数都要比其他等级的要低。