WeRateDogs 数据整理项目

一、项目背景

项目整理的数据集是推特用户 @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。

二、整理过程

本次数据整理主要分三步：收集数据、评估数据、清理数据。具体分步骤描述如下：

（一）收集数据。

数据来源：（1）项目提供的 WeRateDogs 推特数据；（2）从给定网址下载的图像预测数据；（3）项目提供的 txt 文件。

处理方法：（1）用 pandas 的 read_csv()函数导入，命名为‘twitters’，共 2356 条；（2）用 request 下载 tsv 格式文件，再用 pandas 读入，命名为‘images’，共 2075 条；用 json 解析给定的 txt 格式文件，再逐行构建表，命名为‘twitter_additions’，共 2352 条。

（二）评估数据。

评估方式：先用目视评估，寻找线索；再用编程评估，深入探索；反复上述过程。

评估结果：依次对上述 3 个表进行评估，得到 18 个问题，其中 3 个数据缺失问题无法处理。对余下 15 个问题（12 个质量问题、3 个结构问题），按照数据质量维度进行归类并适当归并。

（三）清理数据。

处理顺序：按**“缺失数据、结构问题、质量问题”**的顺序依次处理问题。

处理方法：按“**define-code-test** 框架”依次处理每个问题。

主要问题：（1）数据合并为 1 个表；（2）部分列合并；（3）正文 text 列拆分；（4）处理无效内容；（5）处理不是狗的条目、转发的条目、没有图片的条目；（6）处理表达不一致、表达不准确的内容等。

三、整理结果

将 3 个表以 twitter_id 为关键词合并为 1 个主表，并对上述问题进行了清理，清理后的数据命名为‘df_master’，并输出为‘twitter_archive_master.csv’文件。