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。