## We Rate Dogs 数据清洗项目

对WeRateDogs的推特档案数据 twitter-archive-enhanced.csv，通过推特API获取的附加数据 tweet\_json.txt和图像预测文件 image-predictions.tsv进行清洗，最终获得一个可以后续进行数据分析的数据集twitter\_archive\_master.csv。

整个清洗过程由收集数据、评估数据、清洗数据三部分组成。

一、收集数据

首先导入本清洗项目所需要的库，包括pandas/numpy/requests/simplejson/os等。按照需要导入所需要的三个数据集，分别命名为df\_enhanced、df\_json和df\_image\_pre。

二、评估数据

#### df\_enhanced数据集：

利用观察法和info()、describe()、dupliceted()、value\_counts()、sample()等函数评估数据集发现有以下问题：

1、expanded\_urls有数据缺失59个；

2、expanded\_urls有数据重复137个；

3、部分expanded\_urls有多个值；

4、部分expanded\_urls中不含id信息，无法提取id；

5、部分从expanded\_urls中的id与tweet\_id不符，说明是转发；

6、部分expanded\_urls中无数据无图片信息，说明不含图片；

7、name列有一些是a、the、an等；

8、有一些分母不为10，观察对应的text中的信息，有一些是提取错误，有一些是评价了多条狗狗；

9、doggo、floofer、pupper、puppo应该作为一列出现。

#### df\_json数据集：

1. 运用info()、sample()函数发现含有很多信息，保留id、favorite\_count和retweet\_count数据即可。

#### df\_image\_pre数据集：

1. 运用duplicated()函数发现jpg\_url有数据重复66个；
2. 在数据整合阶段，预测结果不是狗狗的的数据不使用。

三、清理数据

针对评估出现的问题，设计清理方案。将三个数据集copy，分别命名为df\_enhanced\_clean、df\_json\_clean、df\_image\_pre\_clean，清理均在备份数据集上完成。

**（一）define**

**质量问题**

#### df\_enhanced\_clean数据集

1、应该删除expanded\_urls缺失的数据，完整性问题；

2、应该删除expanded\_urls重复的数据，完整性问题；

3、应该删除 2017 年 8 月 1 日之后的数据，完整性问题；

4、应该仅保留含有图片的原始评级 (不包括转发），完整性问题；

5、应该将expanded\_urls有多个值的数据更改为1个，有效性问题；

6、应该删除无法从expanded\_urls中提取id的数据，有效性问题；

7、应该从text中提取评分值，有两条数据评分分别为420/10和1776/10，将其作为异常值处理，删除，有效性问题；

8、修改提取的评分与rating\_numerator和rating\_denominator不符的值，准确性问题；

9、评分分母不为10的改为10，多条狗狗一起的评分应该取平均值，一致性问题；

10、修改name列，保留名字，其余的内容为a, the, an的内容删除，准确性问题；

#### df\_json\_clean数据集

11、应该将favorite\_count, retweet\_count为0的数据作为异常值处理，有效性问题；

**整洁度问题**

1、df\_enhanced数据集中从text中提取狗狗的种类信息，将doggo, floofer, puppo, pupper四列转换为一列status;

2、将三个数据片段合成一个数据集。

**（二）code（见ipynb文件）**

**（三）test**

运用info()、describe()、dupliceted()、sample()等函数评估df\_master数据集，清洗完成。可以进行后续分析。