

# 数据清洗报告

## 一、数据收集

数据收集包含三个表格，1) twitter\_WERATEGODS, 2)

1. “twitter\_WERATEGODS”: WeRateDogs 的推特档案，包含了此博主的微博数据，从官网直接下载
2. “image\_predictions”: 神经网络数据，通过机器学习判断图片中狗的类型，为也是从官网提供的链接下载，但采用的是 Python 的 Requests 方法
3. “tweet\_data”: 每条推特的数据，包括 tweet\_id，被赞次数和被转次数，数据集从官网直接下载，使用 pandas 中的 read\_json 读取

twitter\_WERATEGODS 基本信息:

```
twitter_WERATEGODS.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id          2356 non-null int64
in_reply_to_status_id  78 non-null float64
in_reply_to_user_id  78 non-null float64
timestamp         2356 non-null object
source            2356 non-null object
text              2356 non-null object
retweeted_status_id  181 non-null float64
retweeted_status_user_id 181 non-null float64
retweeted_status_timestamp 181 non-null object
expanded_urls     2297 non-null object
rating_numerator   2356 non-null int64
rating_denominator 2356 non-null int64
name              2356 non-null object
doggo             2356 non-null object
floofer           2356 non-null object
pupper           2356 non-null object
puppo            2356 non-null object
```

image\_predictions 基本信息

```
image_predictions.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id          2075 non-null int64
jpg_url           2075 non-null object
img_num           2075 non-null int64
p1                2075 non-null object
p1_conf           2075 non-null float64
p1_dog            2075 non-null bool
p2                2075 non-null object
p2_conf           2075 non-null float64
p2_dog            2075 non-null bool
p3                2075 non-null object
p3_conf           2075 non-null float64
p3_dog            2075 non-null bool
```

## tweet\_data 基本信息

```
tweet_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 31 columns):
contributors      0 non-null float64
coordinates       0 non-null float64
created_at        2352 non-null datetime64[ns]
display_text_range 2352 non-null object
entities          2352 non-null object
extended_entities  2073 non-null object
favorite_count    2352 non-null int64
favorited         2352 non-null bool
full_text         2352 non-null object
geo              0 non-null float64
id               2352 non-null int64
id_str           2352 non-null int64
in_reply_to_screen_name 78 non-null object
in_reply_to_status_id 78 non-null float64
in_reply_to_status_id_str 78 non-null float64
in_reply_to_user_id 78 non-null float64
in_reply_to_user_id_str 78 non-null float64
is_quote_status   2352 non-null bool
lang             2352 non-null object
place            1 non-null object
possibly_sensitive 2211 non-null float64
possibly_sensitive_appealable 2211 non-null float64
quoted_status     28 non-null object
quoted_status_id  29 non-null float64
quoted_status_id_str 29 non-null float64
retweet_count     2352 non-null int64
retweeted         2352 non-null bool
retweeted_status  177 non-null object
source           2352 non-null object
truncated        2352 non-null bool
user             2352 non-null object
```

## 二、数据评估

通过对三个表格的观察，可以发现如下问题：

### 1. 数据质量：

表格“twitter\_WERATEGODS”中有如下质量问题：

- 1) Tweet id 的数据类型应该是 string
- 2) In\_reply 和 retweet\_id 数据为 float，应该改成 int 并且转化为 string
- 3) Time-stamp 在里面是 string 类型，应该转化为 datetime 类型
- 4) 表格中 content source 一栏中的<ahref =xxx, >来自于提取中的问题，为多余信息，并且 content source 包含了网址和实用平台两个信息
- 5) dog's name 中有提取错误，比如 ‘a’, ‘an’, ‘such’ 明显不是狗的名字
- 6) Rating Denominator 中有 2 和 0 这两个错误，来源于提取了错误的信息，这两行的 rating 数据都要修改

- 7) Rating numerator 没有正确处理小数，比如 9.75 被错误的取成了 75
- 8) Text 中“/r/n”这样的冗杂信息会影响分析

## 2. 数据整洁度：

- 1) “twitter\_WERATEDOGS”: 最后 4 列的信息应该被合并成一列“dog\_stage”
- 2) “image\_predictions”: 每个变量应该为 1 列，所有(p1,p2,p3), (p1\_conf, p2\_conf, p3\_conf), (p1\_dog, p2\_dog, p3\_dog) 应该合并成 3 列
- 3) “tweet\_data”: 中包含太多不需要的信息，提取其中的 tweet\_id , favourite\_count 和 retweet\_count 来制作一张新的表格

## 三、数据清洗

经过清理后的三张表格的简要形式如下：

```
1 twitter_WERATEDOGS_clean.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 16 columns):
tweet_id                2356 non-null object
in_reply_to_status_id    78 non-null object
in_reply_to_user_id      78 non-null object
timestamp                2356 non-null datetime64[ns]
source                  2356 non-null object
text                    2356 non-null object
retweeted_status_id      181 non-null object
retweeted_status_user_id 181 non-null object
retweeted_status_timestamp 181 non-null datetime64[ns]
expanded_urls            2297 non-null object
rating_numerator         2356 non-null float64
rating_denominator       2356 non-null int64
name                    1563 non-null object
dog_stage                380 non-null object
webSource                2356 non-null object
appSource                2356 non-null object
dtypes: datetime64[ns](2), float64(1), int64(1), object(12)
memory usage: 294.6+ KB
```

```
1 image_predictions_clean.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 6225 entries, 0 to 6224
Data columns (total 7 columns):
tweet_id                6225 non-null object
jpg_url                 6225 non-null object
img_num                 6225 non-null int64
prediction_type          6225 non-null object
result                  6225 non-null object
confidenceLevel          6225 non-null float64
isDog?                  6225 non-null bool
```

```
1 tweet_data_clean_new.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2073 entries, 0 to 2072
Data columns (total 3 columns):
tweet_id      2073 non-null object
favorite_count 2073 non-null int64
retweet_count 2073 non-null int64
dtypes: int64(2), object(1)
memory usage: 48.7+ KB
```

## 四、构建新表 twitter\_archive\_master

### 1. 重新检查

#### 1) 质量问题

- a) 整理过程中发现 Tweet\_data\_clean\_new 有重复的 tweet\_id
- b) 表中有 79 个 favorite\_count 为 0。0 与其他数值相差太远

#### 2) 整洁度

- a) 将三表以 tweet\_id 为 key 合并成一张 twitter\_archive\_master 表格

### 2. 清理

- 1) 发现质量问题的 a) 和 b) 的解决办法是一样的，只要将 favorite\_count 的行去除就行，因为所有重复 tweet\_id 的行都是只重复了 1 次，并且其中一行的 favorite\_count 数值为 0
- 2) 用 pd.merge 的方法，以 twitter\_WERATEDOG\_cl0065an 表格为基础，建立新表格

```
1 twitter_archive_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 22 columns):
tweet_id                2356 non-null object
in_reply_to_status_id   78 non-null object
in_reply_to_user_id     78 non-null object
timestamp               2356 non-null datetime64[ns]
source                 2356 non-null object
text                   2356 non-null object
retweeted_status_id     181 non-null object
retweeted_status_user_id 181 non-null object
retweeted_status_timestamp 181 non-null datetime64[ns]
expanded_urls          2297 non-null object
rating_numerator        2356 non-null float64
rating_denominator      2356 non-null int64
name                   1563 non-null object
dog_stage              380 non-null object
webSource              2356 non-null object
appSource              2356 non-null object
favorite_count         1994 non-null float64
retweet_count          1994 non-null float64
prediction_type        2075 non-null object
result                 2075 non-null object
confidenceLevel        2075 non-null float64
isDog?                 2075 non-null object
```