- **收集**

1. 通过 url 下载数据 image-predictions.tsv
2. 生成访问 tweepy API 的密钥，从该网站下载缺少的数据，即下载所有匹配到 image-predictions.tsv 文件中 tweet_id 的 favorite_count 和 retweet_count。将下载到的新的文件写入 tweet_json.txt。

- **评估**

目测评估

twitter-archive-enhanced.csv 中的 rating_numerator 列有非数字形式的值

expanded_urls 列有重复两次甚至多次

doggo，floofer，pupper，puppo 应该作为数据一列中的四个变量

利用 info 函数评估

twitter-archive-enhanced.csv 中有几列出现了严重的缺失，如 in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp

expanded_urls 列有空值，name 列有空值

利用 value_counts 函数评估

name 列有异常值，即与姓名不符，如 a ,an,the

rating_numerator 列有不符合实际的评分

rating_denominator 列并不全是是 10 。

- **清洗**

1. 将转发的数据整行删除，即 retweeted_status_id, retweeted_status_user_id 或 retweeted_status_timestamp 中存在非空值的整行删除。
2. in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp 本来存在缺失，对研究目的没有影响，可以删除。
3. 从 text 中匹配狗的 doggo，floofer，pupper，puppo ，新生成一列名为 stage，并将 doggo，floofer，pupper，puppo 这四列删除。
4. name 列存在类似于 a,an,the 的不是名称的异常值，将其改为正确的缺失值形式。

5. name 列的空值未被正确表示，应将其正确表示。
6. expanded_urls 每一个逗号之后又会重复一次，应截取第一个逗号前的字符串。
7. rating_denominator 规定是 10，存在异常值，将所有值改为 10。
8. rating_numerator 中存在明显脱离现实的数据，将其用 75%的值来填充。