# 数据整理报告

数据整理的过程包括如下三个步骤：

- 数据收集阶段
- 数据评估阶段
- 数据清理阶段

**在数据收集阶段，我们主要通过以下三个方式收集数据**：

- 通过本地数据收集数据
- 通过 Requests 库，收集网页数据
- 通过 API 借口，截取 Json 格式的有用数据

**在数据评估阶段，我们总共识别了 12 个数据问题，其中包括了 10 个数据质量问题以及 2 个数据整洁度问题。具体问题如下：**

数据质量问题：

1 分子数据存在不符合事实的情况，其中有分子小于 10,也有分子大于 1000 的情况。需要关注。
2.分母数据存在不符合事实的情况,分母存在不等于 10 的情况。
3.'in_reply_to_status_id' 数据格式为浮点数，应该为整数
4.'in_reply_to_user_id'数据格式为浮点数，应该为整数
5.'timestamp'的格式并非时间格式，需要修改
6.'retweeted_status_id' 数据格式为浮点数，应该为整数
7.'retweeted_status_user_id'数据格式为浮点数，应该为整数
8.'retweeted_status_timestamp'的格式并未时间格式，需要修改
9.'Name'列中存在若干狗的名字为 a 的情况，需要关注。同时，存在没有名字的情况。
10.狗的地位若干列中存在较多空值，需要关注。

数据整洁度问题：

1.twi_add_info 可以同 twi_arc_en 合并
2.根据变量应为一列的原则，在 twi_arc_en 中可以将狗的 stage 合并为一列。

**在数据清理阶段，我们主要使用了编程的方式，针对上述问题逐一进行清理测试。最后数据清理为 2 张表格，其中表格名为 twi_arc_en 共有 2356 行、17 列数据，表格名为 image_predict 共有 2075 行，12 列数据。**