数据整理过程报告

一、收集数据

整个数据整理过程中，收集了三个不同来源的数据集。分别是：项目提供的twitter-archive-enhanced.csv数据集、从提供的URL中编程下载prediction数据集和读取的JSON文件数据集。下面是这三个数据的收集过程：

> **archive数据：**利用pandas的read_csv功能直接读取；

> **prediction数据：**载入requests库，向指定的URL发送请求，之后直接创建一个可写入的文件把下载的数据放进去，再用pandas读取；

> **tweet数据：**逐行读取下载的tweet_json.txt文件，提取我们需要的数据内容写入到新的可写入文件中，之后在用pandas读取。

二、评估数据

收集好了三个数据集后，开始逐个进行数据的评估。首先先利用Excel打开archive数据，通过目测进行评估，把发现的问题写入到wrangle_act中；之后利用pandas读取数据，编写代码查找、确认所有数据中的有误部分。整个评估过程从数据的质量和清洁度两方面进行，具体过程如下：

> **质量：**数据质量方面主要是一些数据类型错误、数据录入出错（比如archive数据中的一些分子和分母值与推特文本中有偏差）。通过代码*.head()*、*.info()*、*.value_counts()*等方法可以评估出数据的质量问题；

> **清洁度：**通过目测和代码，可以看出数据的一些清洁度问题（比如archive数据中的狗狗的“地位”数据分为4列，不符合清洁度要求，应该合并）。

三、清洗数据

对评估数据中发现的质量和清洁度问题进行清洗。首先，copy一份收集的原始数据，在对copy的数据进行清理；对照评估数据中出现的质量和清洁度问题，进行对数据的错误类型修改、错误数据修改和数据合并整理等。

> **错误的数据类型：**用astype函数修改为正确的类型；

> **错误的数据内容：**分辨出错误的数据的样式特征，匹配出错误数据，修改为指定的值（比如name列，元素应该为首字母大写的格式，在评估中发现一些不是这种格式，匹配查看后确认这些不是名字，之后修改为None）；

> **合并列：**一些列不符合清洁度原则，通过melt函数将其变为一列数据（比如，狗狗的“地位”数据占用了4列，其列名为“地位”的分类，利用melt创建一个分类列，删掉之前的4列，在把对应的结果匹配到原数据中）；

> **合并数据：**为了便于最后的研究，需要把三组数据的部分列内容合并一起，利用merge函数，通过tweet_id这列唯一数据进行合并。