数据清洗报告

本项目的数据清洗的对象包括三个数据集：推特文档数据集，推特附加文档数据集，推特图像预测数据集。这三个数据集都有推特 id 一列，可作为三个表合并的主键。由于本项目主题是关于狗狗评分的推特数据，在数据清洗时着重观察评分，狗狗的地位，推特发表时间等列。

首先导入数据集后，对每个数据集的内容进行抽样的观察，了解每一列数据表示什么内容，可以采用目测评估和编程评估的方式。通过 dataframe.info()的方式，观察数据是否有缺失并进行相应的填充；观察每一列数据的类型是否有误，如 timestamp 列的数据为字符串，需要更改为时间戳类型。另外注意到很多列都是从 text 列提取出来，可以观察下提取的内容是否准确，并对有误的内容进行修正。接下来着观察数据的一致性问题，比如狗狗的评分的分母值很多记录都不是 10，可以统一折算到 10，分子进行相应折算，便于比较。

关于数据整洁度的问题，遵循 tidy data 规则，单个变量自成一列，可以发现狗狗的 4 个地位用分别用 4 列来表示，实际上应该合并成一列。此外注意数据是否有重复，比如说推特附加档案的 id 和 id_str 列，都表示推特 id，因此可以删掉其中一列。

重要一步的是按照项目的基本要求，需要提取出含有图片且非转发的原始记录。

最后每个数据表清洗完毕后，以推特 id 为主键合并成一个主数据集，删除所有内容重复的列，并保存为 csv 文件。