收集清洗说明

首先从项目信息得知，数据至少要收集 twitter 转发数和喜欢数，然后还有 twitter id，狗名，狗地位及狗评分和图片信息，原始的推特档案里提取的狗名，狗地位及狗评分数据多少有点问题，故重新从文本里提取，对于狗名，阅读档案 csv 文件，大致有这么几种介绍方式：

'This is '

'Meet '

'name is '

'Say hello to '

'named '

故采用正则表达式'(?:This is|Meet|name is|Say hello to|named) ([A-Z][a-z]{2,12})'来进行匹配，以提取狗名，对于评分，大部分分母评分是 10 分，分子评分大于 10，但少量分子评分低于 10，而这部分属于无效评分，另外还有一些分母不为 10 的，在提取时即做限制，只提取分母为 10，分子大于 10，但不至于大到异常的评分，通过观察，最终确定分子评分范围限定在 11~16，对于狗的地位，大部分文本中没有，只能提取到有的。

对于转发数和喜欢数，可以从 tweet_json 文件中提取。

提取上述数据后，再将其与图片信息合并来进行分析，由于在提取时就尽可能的提取干净的数据，所以合并后的数据框要做的清洗不是很多。