Wrangle Report

收集

使用 pandas 读入项目需要的 3 个文件（格式为 txt 和 tsv），其中图像预测文件使用 requests 从 internet 上获取。

评估

目测发现各表中都存在一些空值，但大部分属于合理范围（如 reply, retweet）相关，一些变量呈现为不友好的格式，但大部分不需要用到。经过进一步编程评估，记录下发现的各表数据格式相关问题以备后续清理。

一个突出的问题是, stage 是一个变量, 但却分置在 4 个列中, 并且 archive 表中提取的 stage 信息其实并不完整。

三个表的数据可以合并到一个表中。

清理

刚开始清理时其实对哪些数据将要参与分析并不完全清楚，所以尽量少 drop 掉数据，对于拿不准是否要清理的数据可以暂时先不清理。动手之前先对 DataFrame 进行备份十分重要。

先按常规处理各表中的缺失值、格式等问题，然后处理更为复杂的部分，比如重新从 tweet 文字中提取 stage 数值，并归到一列中。
图像预测数据主要用到 p1，但因为想对比一下可信度较低的预测情况，所以仍然保留了 p2 和 p3 的数据。这些数据存放方式不够整洁，比如 p1_dog, p2_dog, p3_dog 实质属于同一变量，但因为不打算对它们之间作太多分析，所以没有作 flat 化处理。

分析

对时间、评分、名字、Tweet 来源等作简单表格或图表呈现。

本想对 stage 做些有趣的分析，但数据量太少，很难做进一步解读。

因为对机器判定“不是狗”的情况比较在意，所以筛选了一些图片作目测分析（编写了函数 ImageMatrix 来显示图片阵）。
有一些导致判别失误的原因很容易看出来，有一些则不是那么明显。