Twitter 上狗狗的信息整理——工作报告

关于项目

这次项目的任务是：清洗 WeRateDogs 推特数据，并就清洗干净的数据做一些有价值的分析，并将分析结果做一个可视化的长线。我们得到的信息比较全面，包括推特档案，但这些数据是来自后台数据库，可能与前端页面中的显示有一些区别。如果我们要得到更出色的分析，我们需要结合一些推特网站中的信息来呈现更好的效果。因此我们需要整理这些数据，并清洗、评估、分析，来实现我们的目标。

信息收集

我们通过不同的方式获取了网络上关于狗狗信息的三张表：

第一张表：是推特的档案数据，他们包括一些基本信息，并且从网页文本中简单提取了一些姓名、评分和狗狗阶段等信息。

第二张表：是通过推特的 API 获取的附加数据，这些数据主要包含了转发和点赞的数据。

第三张表：是图像的预测数据，这是通过神经网络技术得出的对图像可能的文字描述，然后列举出了概率的前三名。

信息评估

收集到以上三张表，我们需要对表中的数据和信息进行简单的评估，知道这些表中的问题所在，并将这些问题记录下来，为下一步的清洗做准备。

我们发现的质量问题包括：缺少数值、填写错误、格式不对、信息重复冗余、评价标准不统一等。

我们发现的整洁度问题有：信息分散、填列方式不利于分析等。

数据清洗

我们同 python 的 numpy、pandas 库中的工具对三张表的上述问题进行清理，得到一张干净的没有质量和整洁度问题的表，方便我们进行下一步的分析。

分析与可视化

我们通过制作散点图可以查看喜欢数和转发数的关系。通过选择评分最高或者喜欢数最高的 ID 查看狗狗的姓名，品种等信息。通过计数的方式查看预测最多的狗的品种和狗狗成长阶段的数量信息。