项目整理报告

此次我们分析的数据集是一位推特博主的推文记录，推特昵称为 WeRateDogs。

首先我们拿到的是一份推特档案：twitter_archive_enhanced.csv(分隔符为：“,”)此文件包含推文的一些基本信息(推特 ID、时间、推特文本等)。其次我们还有一份推特图像的预测数据，即根据神经网络，对出现在每个推特中狗的品种（或其他物体、动物等）进行预测的结果，这份数据我们需要用 Python 的 Requests 库解析提供的 URL 链接，可以得到一个.tsv 的文件(分隔符为：“\t”)。另外还有一个名为 tweet_json.txt 的文本文件，这里我们要用到 json 库提取文本内容，至少要包括 tweet_ID、retweet_count 和 favorite_count 字段，这里要了解的是 json 数组相当于 python 列表，而 json 对象则对应 python 字典。然后将这个 .txt 文件逐行读入一个 pandas DataFrame 中。

数据评估这里采用目测评估和编程评估的方法，.csv、.tsv 文件可使用 excel 直接打开，或在 python 中导入 pandas 库使用 df.read()方法，也可以查看数据集的字段和内容。对于一些目测无法得到的信息，我们可使用 df.info()、.value_counts()等方法分析数据。这里共需要至少找出 8 个数据集质量问题和 2 个整洁度问题。

在清洗阶段进行之前，我们一定要记着先备份数据集，接着就可以开始清洗数据了。其中较为棘手的该属于从文本中提取想要的目标信息了，这里使用正则表达式还是较为方便的，但是一定要转化类型，可使用.str 直接转换，再接着.extract 方法就可以直接提取内容了。apply 也是一个相当灵活实用的函数，它可以对 DataFrame 对象进行操作，当参数中存在一个函数(可以是匿名函数)时，它将作用于整列元素。

最后，多个 table 合并我们可使用 merge 函数，这里有两个重要的参数。一个是 on，因为我们三个表都含有同一个字段“tweet_id”,这个需要指定。另一个是 how，它是来指定我们使用哪种方式连结数据。