数据整理过程

数据收集

总共收集了 3 个数据集。

1. WeRateDogs 的推特档案:

用课程 3-项目细节里提供的链接下载 twitter_archive_enhanced.csv，然后将 twitter_archive_enhanced.csv 导入名为 twitter_archive_enhanced 的自建 Pandas DataFrame。

2. 推特图像的预测数据:

将在课程 3-项目细节里提供的 URL 赋值于 url 变量，然后用 requests 库将 url 赋值于名为 response 的 requests.models.Response 对象。用 with 语法打开（创建）一个 image-predictions.tsv 文件，然后将 response 里面的数据写进 image-predictions.tsv 里，然后将 image-predictions.tsv 导入名为 image_predictions 的自建 Pandas DataFrame

3. 每条推特的额外附加数据:

由于不能科学上网，所以在课节 4-Twitter API 中下载了名为'tweet_json.txt'的文件。用 with 语法打开 tweet_json.txt，然后用逐行读取里面的 json 数据并存为 data 列表。

创建 3 个空列表：id_list，retweet_count_list,favorite_count_list，然后遍历 data 列表中的数据并把每个数据转为 json 格式，然后读取 key 值分别为'id','retweet_count'和'favorite_count'的值，把'id'相对的值放在 id_list，'retweet_count'相对的值放在 retweet_count，'favorite_count'相对的值放在 favorite_count 。

然后创建一个 Dataframe，把 id_list，retweet_count_list 和 favorite_count_list 读入为 3 列数据，相对的列名是'tweet_id','retweet_count','favorite_count'。

数据评估

把评估分成了两部份去完成，先在 jupyter notebook 中目测评估 3 个在数据收集步骤中生成的数据集，然后把那 3 个数据集导出为 excel 文件，再在 Excel 中打开目测数据。把评估出来的问题分为质量和整洁度两类记录下来。

第二步是编程评估，用代码评估 3 个数据集并把评估出来的问题分为质量和整洁度两类记录下来。

数据清洗

把评估出来的问题以质量和整洁度两类依次清洗。