

# twitter 数据集整理

## 文件说明:

- wrangle\_act\_cn.ipynb 中使用了 pandas, numpy, requests, tweepy, json, matplotlib 库
- twitter\_archive\_master.csv 是 整理整理完成的文件

想重现清理过程请运行 wrangle\_act\_cn.ipynb,

## 1.收集

数据集有三个

- 本地文件: twitter\_archive\_enhanced.csv  
包含 twitter 帖子数据, 优达学城提供, 具体数据请运行 wrangle\_act\_cn.ipynb 查看
- 下载文件: image\_predictions.tsv  
包含 twitter 中狗的图片种类预测, 具体数据请运行 wrangle\_act\_cn.ipynb 查看
- json 文件: tweet\_json.txt  
从 twitter 接口获取或者使用优达学城提供, 包含每条 twitter 的详细信息, 本次整理只提取了 id 点赞和转发数, 具体数据请运行 wrangle\_act\_cn.ipynb 查看

## 2.评估

在数据质量和数据整洁度两个方向进行了评估, 找出数据集中的部分问题, 并在评估部分做了记录, 其中有些问题在解决其他问题时自动解决了, 例如评分分母为 0 的记录, 在删除转发的记录时被删掉了。

## 3.清理

清理过程没有严格按照清理缺失值->清理整洁问题->清理其他质量问题的步骤, 因为在清理整洁问题时发现需要先解决质量问题。由于在评估时已经大概想好了要问的问题, 有些看似没有用的列没有着重清理。