

项目：清洗与分析数据

本项目使用的库：

- pandas
- numpy
- requests
- json
- matplotlib

数据收集

- 一、收集来自本地的文件

使用`pd.read_csv`语句，读取文件 *twitter_archive_enhanced.csv*，并将读取的数据存入数据集 *twitter* 中。

- 二、收集来自互联网的文件

使用`requests.get(url)`语句，收集特定网址上的信息，而后将收集的信息存入文件 *image-predictions.tsv*，最后使用`pd.read_csv('image-predictions.tsv', sep='\t')`语句将读取的数据存入数据集 *prediction* 中。

- 三、收集来自API的文件

由于该计算机无法访问Twitter，因此本项目直接使用课程提供的Twitter返回数据 *tweet_json.txt*。然后将这个 *tweet_json.txt* 文件逐行读入一个数据集 *extra* 中，该数据集包含 **tweet ID**、**retweet_count** 和 **favorite_count** 字段。

数据评估

- 首先评估数据集 *twitter*。先通过目测，查看一下数据集的前5行和后5行，对数据集有一个整体的印象。再通过编程方法查看是否有重复统计的*tweet_id*、*rating_numerator*、*rating_denominator* 和 *name* 值的分布情况。最后再查看一下该数据集的整体情况，看看是否存在空值等问题。
- 而后评估数据集 *prediction* 和 *extra*。主要是通过目测评估两个数据集的抽样样本，然后编程评估两个数据集的整体情况。

通过数据评估，发现上述三个数据集存在以下问题：

质量问题

twitter

- *tweet_id*列值得类型为int，*extra*中*tweet_id*列的值为str（一致性问题）
- *in_reply_to_status_id*、*in_reply_to_user_id*等列含有缺失值（完整性问题）
- *timestamp*列的值类型为str（一致性问题）
- *retweeted_status_timestamp*列的值类型为str（一致性问题）
- *name*列中包含None、a、an、the等值（有效性问题）
- *rating_denominator*列中存在非10的值，而且出现次数非常少（有效性问题）
- *rating_numerator*列中有一些值仅出现1至2次，应该是错误值（准确性问题和有效性问题）

prediction

- *tweet_id*列值得类型为int，*extra*中*tweet_id*列的值为str（一致性问题）
- *p1*、*p2*、*p3*的值大小写不一致（一致性问题）

整洁度问题

twitter

- *doggo*、*floofer*、*pupper*、*puppo*是同一种分类变量，但是占据了四个列

extra

- *retweet_count*、*favorite_count*是*twitter_clean*表格的一部分

数据清理

为了防止修改原始数据，给三个数据集建立副本。

首先，清理缺失值的问题。先去掉大部分值为空值的列

in_reply_to_status_id, *in_reply_to_user_id*, *retweeted_status_id*, *retweeted_status_user_id*, *retweeted_status_timestamp*, 接着去掉 *expanded_urls* 列中含有空值的行

接着，处理一致性的问题。将数据集 *twitter* 和 *prediction* 的 *tweet_id* 列值的类型改为 *str*。将 *timestamp* 列的值类型改为 *datetime*，并分成 *year*, *month*, *day*, *hour* 四列，而后删除 *timestamp* 列。将列 *p1*、*p2*、*p3* 的值全部改为小写字母。

再处理有效性问题。将 *name* 列中的 *None*、*a*、*an*、*the* 等值替换为 *Unnamed*。*rating_denominator* 列的值均改为10。调整 *rating_numerator* 列中出现次数少于3的值，如果是三位数除以100，两位数除以10，结果不保留小数位。

最后处理整洁度问题。将 *doggo*、*floofer*、*pupper*、*puppo* 作为分类变量放入一列，列名称为 *level*，没有定义狗“地位”的，赋值 *Nolevel*。将数据集 *extra* 合并到数据集 *twitter* 中。