清洗与分析数据

Kipmin | 数据分析（进阶） | 2018-02-11

收集数据

下载已有的数据文件 **twitter_archive_enhanced.csv**

从网络中收集其他数据 **image-predictions.tsv**

从 twitter-archive-enhanced.csv 文件中获取的 tweet_id 来获取每条推特详细信息保存到 **tweet_json.txt**

评估数据

将数据文件用 pandas 读取到 jupyter 中，生成 3 张表，分别是：

twitter_archive_enhanced.csv→twitter_archive_enhanced

image-predictions.tsv→image_prediction

tweet_json.txt→tweet_json

通过 pd.options 将表修改成能展示全部内容

查看数据类型发现：twitter_archive_enhanced 中的 timestamp，tweet_json 中的 id，retweet_count，favorite_count 类型不符

解决方案：timestamp 列可以转换为 date 类型数据，id，retweet_count，favorite_count 中的所有列都可以转化为 int 类型

目测观察发现：不存在的值都被写成了 None，source 列有很多无用 html 代码，doggo,floofer,pupper,puppo 是分类变量，不应该有 4 列，in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 总数太小

解决方案：将 None 都替换成 NaN，将 source 列中的无用 html 代码除去，floof(er)有两种，分别是 floofer 和 floof，在 twitter_archive_enhanced 的 text 中重新获取地位信息，放到 stage 一列中，删除 doggo,floofer,pupper,puppo 这四列，删除 in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 这些列。

数据中不能包含转发和没有图片的推文：删除转发的推文 181 条，没有图片的推文 59 条

查看是否有重复数据和错误数据：发现 image_prediction 中有 66 条信息照片地址重复，评分的分母不都是 10，狗狗的名字不可能是 a,an,the

解决方案：删除重复照片地址，从 text 中重新获取评分，把狗狗名字为 a,an,the 的改为空值。