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。