数据分析(进阶)

WeRateDogs分析文档

收集

利pandas的read_csv，将项目提供的数据读入到jupyer notebook中，注意读取tsv文件时，spe参数为'\t'。

评估

主要从数据质量和数据清洁度两个方面来评估数据。

清洁度

项目中3个dataframe都是tweets的相关信息，所以应该在一张表中。

关于狗的地位stage在表中显示为四列，太冗余了，合并为一列。

质量

表中部分列的数据类型不正确。

从文本中提取评分和地位时，数据有错误还有些不完整，都需要重新清理。评分分母应为10

source列的数据有html文本难以阅读，text列的部分'＆amp';转义符要替换为'&'。

部分狗的小写名字从text中提取错误。

表的None是字符串，应更改为np.nan类型

没有图片tweets，不作为分析数据。

部分tweets是转发的tweet， 不是作者原创。

清理

重新从text文本中提取狗的地位dog_stage,删除'doggo','pupper','puppo','floofer'四列。

利用merge和tweet_id将三张表合并在一起。

删除没有图片的tweets。

删除转发的tweets。

利用正则表达式从text中提取dog_stage、狗的name和修改错误的评分，并将source中的html文件去掉。

利用replace将错误的数据替换为正确的。

最后，用astype变更错误的数据类型。储存为twitter_archive_master.csv。