简介

一、数据集说明

数据集是推特用户 @dog_rates 的档案, 推特昵称为 WeRateDogs。WeRateDogs 是一个推特主，他以诙谐幽默的方式对人们的宠物狗评分。这里共有三个数据集：

1. **twitter_archive_enhanced.csv。**这是最主要的一个数据集，包括基本的推特数据（推特 ID、时间戳、推特文本等），包含了截止到 2017 年 4 月 1 日的 5000 多条推特。其中评分、地位和名字等数据是从text原文中提取。其包含以下几列：

1）tweet_id : 推特的ID
2）in_reply_to_status_id:： 回复推文的ID
3) in_reply_to_user_id: 被回复推文的原始ID
4) timestamp: 发twitter的时间
5) source: 发twitter的设备
6) text: 发twitter的文本
7) retweeted_status_id: 转发twitter的ID
8) retweeted_status_user_id: 被转发的原始用户ID
9) retweeted_status_timestamp: 转发时间
10) expanded_urls: 推文链接
11) rating_numerator: 评分分子
12) rating_denominator: 评分分母
13) name: 宠物名
14) doggo: 狗的成长阶段（分类变量）
15) floofer: 狗的成长阶段（分类变量）
16) pupper:狗的成长阶段（分类变量）
17) puppy:狗的成长阶段（分类变量）

2. image-predictions.tsv：用神经网络对狗狗种类进行分类的结果数据。这个结果中包含了预测结果的前三名，twitter ID，图像url，以及最可信息的预测结果对应的图像编号。

1) tweet_id: 推特链接的最后一部分，位于 "status/" 后面 → https://twitter.com/dog_rates/status/889531135344209921
2) jpg_url: 预测的图像资源链接
3) img_num:可信的预测结果对应的图像编号 → 1 推特中的第一张图片
4) p1: 算法对推特中图片的一号预测 → 金毛犬
5) p1_conf: 算法的一号预测的可信度 → 95%
6) p1_dog: 一号预测该图片是否属于“狗”（有可能是其他物种，比如熊、马等） → True 真
7) p2: 算法对推特中图片预测的第二种可能性 → 拉布拉多犬
8) p1_conf: 算法的二号预测的可信度 → 1%
9) p2_dog: 二号预测该图片是否属于“狗” → True 真

3. tweet_json.txt: 一些额外的twitter信息，如点赞数、转发数等

二、主要数据问题

1.JSON数据载入处理过程中，发生了一些问题，已解决。主要资料可以参考以下部分：

1） [JSONDecodeError: Extra data: line 2 column 1](https://blog.csdn.net/u011318077/article/details/88550775)
2） [JSON读取大量数据错误：JSONDecodeError: Extra data: line 2 column 1或者ValueError: Extra data: 类似错误处理](https://blog.csdn.net/u011318077/article/details/88550775)
3） [如何区别python中的json模块loads和load方法](https://jingyan.baidu.com/article/86f4a73ebade9337d65269d1.html)
4） [JSON](https://docs.python-guide.org/scenarios/json/)
经过编程查看，发现数据集主要有以下一些问题：

2. 数据质量问题

1）twitter_archive_enhanced.csv 数据集
- tweet_id、retweeted_status_user_id 应该是字符型，而不是数值型
- 数据包含转发的twitter，应予以删除
- 狗的推文链接（expanded_urls）有缺失值
- 评分分母有问题：有些评分分母不为10
- 评分的分子有问题：有些分子大于14，有些分子为0
- 狗的种类数据标签严重缺失
- 狗的名字有缺失值，并且有一些不是名字的值（如a,an,the等）
- timestamp,retweeted_status_timestamp这两列应该是datetime格式
- source列有额外的字符串

2）predictions.tsv 数据集
- tweet_id应该是字符型，而不是整数型(int64)
- jpg_url列有66行重复数据

3）tweet_json.txt
- id列的名字应为tweet_id,以使其与前面两个表格一致；另外，id这列的格式应为字符

3. 数据整洁度问题

- df1中狗的地位（成长阶段）是分类数据，应该将4类合成一类
- 可以以tweeter ID为关键列，将三个表格的数据进行整合

三.数据清理

1.质量问题

twitter_archive_enhanced.csv数据集
1）用str函数，将tweet_id、retweeted_status_id、retweeted_status_user_id、转换为字符型
2）用数据筛选的方式，删除retweeted_status_id这列中的181行非空缺数据
3） 用is.null函数进行数据筛选，删除expanded_urls列的缺失值

4） 找出rating_denominator 这列值不等于10的行，查看原因。如果以下打分都属于特殊情况，则予以删除。
5）删除rating_numerator列评分大于14的行和评分为0的行
6）用 findall 函数结合正则表达式，从原始text中查找狗的种类标签(doggo,floofer,pupper,puppo),然后将其填写到stage这一列中。同时，将缺失数据用np.nan进行替代。
7）将缺失值，不属于名字的值(a,an,the等）改为np.nan
8) 将timestamp,retweeted_status_timestamp改为datetime格式

predictions.tsv 数据集
1）用astype函数将tweet_id 这列转换为字符型
2）用drop_duplicated函数，将jpg_url这列的重复数据删除

tweet_json.txt

将id列的名字替换为tweet_id,以使得与前面两个表格一致，并将这列转为字符型格式

2. 整洁度问题

用merge函数，将三个数据合并，存在文件中。命名为 “clean.csv”