数据清洗报告

这个项目，分析 WeRateDogs 推特数据，从收集数据开始，到导入数据，再到评估、可视化数据，这一系列的步骤，从不同的角度对 WeRatedog 对狗狗打分这一举动有了更清楚直观的了解。

导入数据

先导入库

导入 WeRateDogs 的推特档案数据

再导入推特图像的预测数据，对狗狗属于哪种品种进行预测。

– 用 requests 库下载推特图像的预测数据

– 加载下载好的推特图像预测数据，转换为 df2

```
df2=pd.read_csv('image-predictions.tsv', sep='\t')
df2.head()
```

然后再加载推特额外数据,提取转发数(retweet_count)和喜欢数(favorite_count)**

再来检测看载入数据是否成功

```
df_json.head()
```

备份原始数据集，在备份数据集里对数据作出修改

导入、备份数据成功后，现在开始清理数据。

数据分析清理主要清理有整洁度和质量问题的数据。

先看 3 个数据是否各自有整洁度问题。

找出缺失值

df1_copy.isnull().any()

df2_copy.isnull().any()

df_json.isnull().any()

结果有明显，df1 有数据缺失，其他两个数据都不存在数据缺失，所以只填补 df1 的缺失数据。

再次检测看，缺失数据是否填补成功了

这次，缺失数据全部修复好了。

然后看是否有重复值

结果无重复值。

每个数据集转化为列表。

查看数据的分布情况

每个数据抽样 20 个看看

清理完整洁度问题后，现在准备清理质量问题。

先设置一下列宽。

首先，查看推特档案数据中转发信息是否正确

然后，查看 text 中转发信息的个数与 retweeted_status_id 转发编号的数量有没有差异。

结果说明，转发信息个数与 retweeted_status_id 转发编号的数量有差异。

因为普遍狗狗评分分子都大于 10，所以去看狗狗评分中分子小于 10 的数据有多少个

再看分子小于 10 的 text。

同时，来看另一种很特殊的情况，分母不等于 10。

查看相对应的 text

查看 rating_numerator 和 rating_denominator 列中数据的有效性。

查看 name 列的有效性

再查看 tweet_id 和 jpg_url 列是否存在重复值。

jpg_url 列存在重复值，但是因为对数据没什么影响，所以不做处理。

先处理其他数据问题，最后再来看 jpg_url 列重复值

因为三个数据集有相同的列 tweet_id, 所以现在通过 tweet_id 合并三个数据集; 分成两个、两个合并，先合并 df1_copy 和 df2_copy,命名 df_merge1,再将 df_merge1 和 df_json 合并。

现在筛出转发数据，将除此之外的数据重新赋值给 df1

然后测试看看不包含转发信息的数据概览结果

因为这两列缺失数据太多，没什么有效值，所以直接删除这两列

检测一下删除之后的结果

删除成功了

同时，相关转发数据已失效，所以同时删除其他相关的转发列

再删掉 expanded_urls, source 这两列用不上的数据，只留下有用数据。

提取 doggo，floofer，pupper，puppo 列不为空的数据到 df_stage

查看 dog_type 数据信息

然后将 doggo, floofer, pupper, puppo 四列转移到 stage 列里

去掉 stage_name 列

将 dog_tyoe 和 df_tweet 合并在一起，去掉 df_tweet 中的 doggo, floofer, pupper, poppo 四列，只留下 stage 列

现在根据名字前会出现的单词、字段在 text 中用正则表达式提取狗狗名字

测试结果

查看名字列的有效值

通过正则表达式在 text 中提取评分，分为 text_numerators, text_denominators，分子分母两列

查看结果

来看评分分母不等于 10 的数据

对应的把这些行的评分分子改动

查看它们的 text 信息

然后根据 text 信息中的分数，来对应修改评分分母，统一所有分母为 10，像 60/50 就变为 12/10

将这些行的分母都修改为 10

因为经过修改后的 rating_numerator 和 rating_denominator 变成了 object 类型，所以重新修改为数字型。

再次查看和提炼分子评分小于 10 的数据

再次用正则表达式提取评分分子小于 10 范围内的评分分数，命名 df_tweet2

设置行宽和列款

再来看看有多个评分的行

因为刚才修改过很多行的评分分子分母，导致修改后的数据类型变成了 object,所以现在把 df_tweet 中的评分分子分母数据类型修改为数字型。

下一步来用 findall 提取 text 中狗狗地位。

用正则表达式选取任意有四个名字中的一个出现的

再修改 timestamp 函数，修改为 datetime64 类型。

– 再来看 jpg_url 列的重复值是否被删除

保存数据到本地文件。

总结

从整个质量和整洁度数据清理过程来看，需要逐个清理存在的数据问题。对于数据集影响很大的数据问题，每一个都需要认真清洗干净，而对于数据集没什么影响的问题，可以不用重点关注。对于这个数据集来说，质量问题比整洁度问题更多，主要的质量问题有：很多转发数据缺失，转发信息失去有效性，狗狗地位列大量数据缺失、有效性较少，狗狗名字、评分分子分母列也有严重缺失，timestamp 类型错误，jpg_url 列重复值较多等。整洁度问题包括存在无用列 source 等，doggo，floofer，pupper，puppo 列单独四列可合并为一列，推特档案、推特图像预测数据、推特额外附加数据三个数据集科可合并在一起。

对于画图和分析来说，前面的数据收集、整理、清洗是非常关键的步骤，可能还存在一些数据问题可以继续清理，因为对后续分析的问题已经清理完，剩下的较小的问题对后面的分析已经没什么影响了，可以忽略。