wrangle_report

收集

1. 多次向twetter申请API权限，但申请一直被驳回。为了能快速完成project，直接现在udacity提供成数据。
2. 通过github链接下载数据，将数据存到twitter_archive_enhanced.csv和image_predictions.tsv文件中。
3. 读取twitter_archive_enhanced.csv，image_predictions.tsv和tweet_json.txt文件，其中tweet_json.txt只读取tweet_id,retweet_count,favorite_count等3列。
4. 讲数据存入3个不同的dataframe中，df_weratedogs存we_rate_dog数据，df_predicted_image存入图像识别数据，df_tweet_json存retweet_count和favorite_count数据。

评估

对数据苹果采取程序和视觉两种方式，这两种方式交替使用。通常先做视觉评估，然后再进行程序评估。

首先对df_weratedogs进行视觉评估。通过观察数据，我们发现如下问题：

1. rating_numerator和rating_denominator与text中的信息不符合
2. timestamp结尾显示方式存在问题
3. source显示有问题，其中包含多余的html代码
4. 缺失狗种类严重缺失数据

然后进行程序评估，发现问题如下：

1. expanded_urls缺少数据,其中还存在数据重复现象。
2. in_reply_to_status_id和in_reply_to_user_id大量缺少数据。
3. tweet_id相关的属性类型为int64,而不是string。
4. floofer,pupper,puppo属于同一种数据，但他们被分别例如3个不同的列中。

以上是df_weratedogs中的主要问题。其中8好选项为数据结构问题，其余均为质量问题。

然后对df_predicted_image进行评估，通过观察数据，我们发现这个dataframe中的问题数据多，也不主要。目前只发现两个问题：

1. tweed_id类型为int64，而不是string。
2. jpg_url数据重复

最后对df_tweet_json中的数据进行评估。根据项目要求，我只需要获取retweet_count和favorite_count的数据，讲这两列数据与df_weratedogs合并。 所以我只需要截取部分dataframe，部分数据包括retweet_count和favorite_count以及对应的tweet_id。

评估数据后发现，retweet_count和favorite_count两列数据不存在问题，可以直接合并至主df_weratedogs。

清洗

先备份数据，在备份数据上，根据上述列出的问题，对数据进行清理。先解决结构问题，在处理质量问题。

结构问题：

通过merge函数，将df_tweet_json中的retweet_count和favorite_count列与df_weratedogs和df_predicted_image合并。

质量问题：

处理数据质量问题可以细分5各维度：

1. 完整性
2. 唯一性
3. 准确性
4. 有效性
5. 一致性

首先处理完整性，尽可能将重要的缺失数据补充完整。通常先试图通过其他数据找回丢失数据，如果参考其他数据，可选择用常用数据补充。 例如数值类型可以用平均数补充。如果某一属性数据严重缺失，可考虑放弃放弃此类数据，直接删除。 本项目中：

1. 主要缺失数据为retweet_count和favorite_count。这个问题可以也可以被列入数据结构问题。幸运的是这两类数据可以通过访问API或者tweet_json.txt文件获得。通过pandas.merge()或者pandas.concat()完整数据。
2. expanded_urls中缺失少量数据，可以通过用NaN值填充。
3. in_reply_to_status_id和in_reply_to_user_id大量确实数据。这些数据目前对分析没有任何帮忙，可以直接删除这两列。

然后处理数据唯一性问题。换句话说，删除重复数据。

1. 删除expanded_urls中存在重复数据
2. 删除数据中包含转发tweet，项目中明确说明，转发数据无效。通过判断retweeted_status_id是否为空决定是否为转发数据。其次通过判断text中是否存在“RT @”进行二次验证。
3. 删除jpg_url中存在重复数据.

处理准确性问题。数据类型错误为最常见的准确性问题。

1. 所有的tweet_id猎德数据类型为int，tweet_id常规意义上为数字类型，但这里不需要对id做运算处理，所以应使用string类型，将所有的tweet_id改为string类型。
2. 将timestamp由原来的object类型改为datetime64。
3. 将source中包含多余的html代码删除。处理方式是通过正则表达式的方式提取soruce信息。

处理有效性问题。在实际生活中，数据类型和解雇无误，但是信息内在表达错误。

1. 在name列中，我们发现一些狗的name显示为a,an,the等英文冠词。通常没人会用冠词来命名宠物，所以这部信息是无效的。通过观察，我们发现text中会出现name的信息，可以通过正则表达式的方式提取信息。提取是应该注意英文语法和习惯使用规则。通常介绍有方式为”This is“，”Meet“后面接name. 需要注意的是” This is a/an“ 后面接的是狗的品种而不是name。虽然在text中提取名字，但是对这列中的数据矫正效果甚微！name还是存在大量a,an和None。最后决定用NaN替换掉这个无意义的数据。
2. 将rating_denominator这一列的数据全都改为数字10。

最后是一致性问题。评估和评价数据需要统一的标准。

1. 上文中，我们将rating_denominator列的数据都改为10就是为了时间一致性。
2. 通过rating_numerator和rating_denominator判断狗的评分不直观，用户需要一个已读性的评分。创建一个新的类，命名为rating， rating_numerator/rating_denominator。

清理部分的最后，我们需要将整体数据，导出至twitter_archive_master.csv文件中。

以上是清理部分的主要是思路与处理方式。

In [ ]: