Wrangle_Report

April 23, 2019

1 Data Gathering

I have used the provided `twitter-archive-enhanced.csv` and `image-predictions.tsv` and read the using panads `pd.read_csv('twitter-archive-enhanced.csv')` and `pd.read_csv('image-predictions.tsv', sep='\t')`.

Then I used my twitter developer account to create an App to query WeRateDogs, I have used tweepy to connect to my twitter APIs, I have queries all the tweets using statuses_lookup function, that allowed me to query for 100 per API call and 30 calls in total (3000 tweets), then I converted from JSON to oandas Dataframe using `pandas.DataFrame()`, exported from the dataframe id, retweet_count and favorite_count, and finally stored it in `queried_tweets_using_tweepy.csv` using `pd.to_csv('queried_tweets_using_tweepy.csv')`

2 Data Assessment

I have assessed the data both visually using excel and (pd.head()) and progmatically using several functions of python, pandas framework and Numpy; and the functions I used:

```
df.head()
df.info()
df.describe()
df.duplicated()
df.Series.values_count()
df.Series.str
df.Series.replace()
df.Series.unique()
df.Series.sample()
df.isna()
df.isis()
df.merge()
df.join()
df.sample()
df.to_csv()
```

I started with tidiness check in mind and found the following issues by visually inspecting the dataset:

1. dog stages should be one column instead of 5, as this breaks the rule of a tidy dataset.

2. rating should be in one column instead of 2, as this breaks the rule of a tidy dataset.
3. Joing queried tweets to the orginal dataset, as it belongs to the orginal archive.

Then I started to look for quality issues in the dataframs here is what I found:

1. Drop retweets columns and rows (as advised in the project motivation we should look into Original tweets not retweets)
2. Source column needs clean up as the 2 unique values are iphone and vine.
3. There are some missing names (invalid names like a, an, the, None, such, quite etc).
4. Timestamp datatype.
5. Fix tweet_id 883482846933004288 numerator.
6. Tweet_id 666287406224695296 and 835246439529840640 has wrong numerator and denominator.
7. Tweet_id 810984652412424192 doesn't have a rating.
8. Remove duplicates from ip dataset
9. Change datatype of rating numerator and denominator to string (was done as part of tidiness fix).

At first I wrote down these issue, but when I reached the cleaning step, I re assessed and found that there are not issues:

1. Denominator should be only 10, there are pleny of other values, its either text parsing error or something else.
2. Numerator should be equal or more than 10 according the account style.
3. Index(2260) of twetter archive DF expanded URL is duplicated.
4. Index(2272) of twetter archive DF two dogs instead of one.(requires spliting)
5. Tweet_id 695064344191721472 2 ratings.(probalbly the 2nd rating)
6. Index(1598) of twetter archive DF denomater of 20?. (Correct info)
7. Remove rows with no expanded_urls and no in_reply_to_status_id and in_reply_to_user_id.(no need to clean, as its a reply)

3 Data Cleaning

I started with the tidiness fix for my dataframes:

1. Melting dog stage names into one column, metling is an option here but I chose to extract the data again from the source 'text' since its available and its easier.

2. Rating column also needs to be in one column instead of and it should be string, I have used pandas cat() to concatinate the 2 columns with “" between”, then I droped the orginal 2 Denominator and Numerator.

3. joining the id, retweet_count and favorite_count from “queried_tweets_using_tweepy.csv” with my existing dataframe of tweets using pd.merge, then droped the extra id column in the merged DF.

Quality cleaning for the dataframes:

1. Drop retweets columns and rows (as advised in the project motivation we should look into Original tweets not retweets), I have used df.text.str.extarct() to identfy the retweets then removed them from the dataframe, then df.drop() to drop the associated columns.

2. Source column needs clean up as the 2 unique values are iphone and vine. I have cleanup the source column to be readable to have the source without the URL.

3. There are some missing names (invalid names like a, an, the, None, such, quite etc). I have used a better regexp to get the names form the tweet text, though even my regexp didn't get all of the names correctly like O'Malley, I have fixed this name later manually.

4. Timestamp datatype. I used pd.to_datetime().

5. Fix tweet_id 883482846933004288 numerator. the original value was 5/10 though in the text it was 13.5/10, so I guess the initial extraction missed this. since its only one occurance I fixed it manually.

6. Tweet_id 666287406224695296 and 835246439529840640 has wrong numerator and denominator, I have corrected this manually also since it will be hard to catch progrmatically.

7. Tweet_id 810984652412424192 doesn't have a rating, 24/7 is not a rating, hence I set the rating for this tweet to NaN manually.

8. Remove duplicates from ip dataset, img_url had 66 duplicats, I have removed them using duplicated() function.