Wrangle-Report

May 13, 2020

1.Wrangling Report

prepared by Lakshya Mutneja

1.1 Gathering and Loading Data

In gathering step we have to gather 3 different tables using 3 different ways below:
The below files are obtained by the help guidelines provided by udacity step by step-

1. twitter-archive-enhanced.csv: this file is downloaded manually from the udacity site.
2. image-predictions.tsv: this file is obtained by downloading it programmatically using the python's requests library
3. tweet_json.txt: this file is obtained by calling the twitter API for each tweet, using the tweet id as parameter, which we obtained from the twitter-archive-enhanced.csv data.

After that, we make sure that all the data are available in the same directory as the Wrangle-act.ipynb

Once I downloaded all the files then my task was to convert them into 3 data frames respectively as shown below

1) archive_df - this is a dataset "twitter-archive-enhanced.csv" which gives info. Regarding basic tweet data and was converted into a dataframe.
2) tweets_info_df - This dataset contains information like tweet_id, no of retweets and no of favorites etc.
3) Image_predictions_df - All the info. In regard to predictions is stored in this dataset.

1.2 Assessing

In the assessing step, I try to gather some quality and tidiness issue from the data collected in the gathering step. Initially , I try to assess each file separately before moving to another file.
In the end, I try to find issue related to redundancy by assessing all the data in the 3 files together.

Below, each column of each table in this twitter dataset is described as follows:

1)Enhanced Twitter Archive

WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column contain : each tweet's text, which used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced."

archive_df columns with description:

- tweet_id: unique identifier for every tweet
- in_reply_to_status_id: status id to the tweet id
- in_reply_to_user_id: the status id for the reply given to the tweet id (w.r.t user id)
- timestamp: Date and time the tweet was created,
- source: the web link
- text: tweets text
- retweeted_status_id: the status id to the tweet id
- retweeted_status_user_id: the status id the tweet id (w.r.t user id) i.e., for the retweeted id
- retweeted_status_timestamp: Date and time the tweet was created
- expanded_urls: Expanded version of url1; URL entered by user and displayed in Twitter.
- rating_numerator: tuser given ranking
- rating_denominator: The reference ranking given by the user
- name: the breed or dog's name
- doggo, floofer, pupper, puppo -- The stage of the dog

Quality - archive_df

1. There are some Missing values in columns from in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id etc.
2. rating_numerator and rating_denominator have some inconsistent values in the numerator and denominator (some of them showing as high as 1776, 170 respectively).
3. tweet id 835246439529840640 has a rating of denominator = 0
4. Crazy names found for dogs - **'infuriating', 'just', 'life', 'light', 'mad', 'my', 'not', 'officially', 'old', 'one', 'quite', 'space', 'such', 'the', 'this', 'unacceptable', 'very'**
5. timestamp and retweeted_status_timestamp must be of datetime instead of the object.

2. Tweets_info_df

Tweets_info_df columns and their description:

- **tweet_id**: Identifier for each tweet
- **retweets**: Number of retweets done by user.
- **favorites**: Favourite count done by user.
- **followers**: Number of followers
- **friends**:Number of friends.

Quality - tweets_info_df table

- 25 tweet ids information is Missing

3. Quality - image_predictions_df :

A table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

image_predictions_df columns:

- **tweet_id**
- **jpg_url**
- **img_num**
- **p1**
- **p1_conf**
- **p1_dog**
- **p2**
- **p2_conf**
- **p2_dog**
- **p3**
- **p3_conf**
- **p3_dog**

Quality - image_predictions_df table:

- only 2075 tweetIds have images

III. Cleaning

To clean all the 3 dataframes following steps were been followed:

Step1:

Convert the datatype of "tweet_id" into string

Step2:

Creating a universe dataset gathering all dataframes on basis of tweet_id

Step3:

Converting the dog stage into one column instead of multiple columns.

By following steps we got some duplicated rows approx. 334 because the count was increased and occured due to the multiple tagging done with dog_status.

We have to clean rows with only one dog_status column value and following steps have to be followed:

- Removing the ambiguity b/w dog_stages
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id -- Convert all the above into objects and strings.
- retweeted_status_timestamp - Conversion into datetime format.
- Observing that information of text is truncated to 50 characters.Increase the text format representation.
- Names found for dogs- 'infuriating', 'just', 'life', 'light', 'mad', 'my', 'not', 'officially', 'old', 'one', 'quite', 'space', 'such', 'the', 'this', 'unacceptable', 'very'. Clean to the ideal name by looking at the text given above.
- rating_numerator and rating_denominator have few inconsistent values in the numerator and denominator. In one tweet_id , the rating for the denominator is shown as 0.

  We are seeing the wide range of values. We would not disturb the ratings provided here.

- retweeted_status_timestamp - has the null values.

IV. Store

Stored the final dataframe into csv file with name **twitter_archive_master.csv** with final data.