Data Wrangling Report

By David Samy Mahrous

December 2020

As an assignment for the Udacity Data Analyst Nanodegree; This Report illustrates the main steps involved in the data-wrangling of Twitter account “WeRateDogs”.

Data Gathering

In this step, collecting data takes place. For this project, there were three main sources for the data to deal with:

1. Twitter_archive_enhanced.csv, this file was delivered by email and downloaded manually to our working directory and then imported into our working environment using Pandas funciotn ‘pd.read_csv’
2. Image_predictions.tsv is the second file that has been hosted on a webage and downloaded using the Request library get function and pd.read_csv pandas function. This file encompassed image predictions for the dogs breeds obtained through a neural network on most of the tweets in the archive file.
3. The final dataset was gathered from twitter API via the Tweepy library by querying the API to obtain extra information to the tweets ids in the first file, e.g. retweets count and favorite count.

Data Assessment

In this step, we investigate our imported datasets visually and programmatically for quality and tidiness issues.

1. The visual assessment done on spreadsheet application like excel sheet and then the programmatic assessment is done in Jupyter notebook.
2. Missing data were addressed first then messy structures. Then the rest of the quality issues that is validity, accuracy and consistency issues.
3. Some of the data cleaning efforts were guided by the scope of the project that mandated the exclusion of retweets and replies and tweets featuring no images.

Issues found:

Issue	solution
Quality issues:	
archive_df table:	
1- there are data for retweet and replys	Drop the rows had had data in this cells
2- unnessary columns of rewteets (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)	Drop these columns
3- timestamp is string not datetimestamp	Convert type to datetime stamp
4- some photos are not for dogs	Find there tweet_id from image_predictions table and drop these rows
5- in consistence of Null values as'None' in the name column, has 745 'None' cells	Missing values convert into NaNs
6- ther is 23 rating_denominator != 10	Fix every one of them
7- rating_numerator have errors	Check the right rating at text column
8- errornes name like 'a' or 'an'	If exist need to extract
9- some wrong data with wrong urls	Remove the row with missing or wrong data of url
10- text column have some more data not nessesry	Clean text column from urls and any other unnecessary data
image_predictions_df table:	
11- there is 66 duplicated jpg_url and there data are duplicated	Remove duplicated rows
13- columns head not describable	Rename the column head with describable name

Tidiness issues:	
1- doggo, floofer, pupper, puppo is a variables	Make a new column called dog_stage with the stage of the dog and remove these four column
2- api_df should be part of the archive_df	Append the api_df with archive_df at archive_df

Output is two tables: archive_df has 1773 rows, 11 columns and image_predictions_df has 2009 rows, 12 columns