wrangling project

by Moamen Mohasseb

Dec , 2020

The "We Rate Dogs" Twitter archive contains over 5000 tweets, which have been filtered to create the enhanced archive that forms the basis of this analysis. The goal of this project is to wrangle the data - gather, assess, and clean

gathering data

a) from csv file(twitter-archive-enhanced) twitter archive whitch downloaded manually from udacity server.
b) gathering TSV file from url download programmatically from udacity server.
c) gathering data from json file by using Twitter API, then read json file.

Assessing

Assessing

Quality issues:

- Many rows containing null values and other contain word None which mean null value too.
- For column name many non-name values there like (‘such', 'a', 'quite', 'not', 'one', 'incredibly', 'very')
- Some value should correct manually by visual analysis (HI. MY. NAME. IS. BOOMER) to boomer name.
- Validity issue we should only keep original tweet not retweet or reply to tweet.
- Data type of timestamp is string so we need to change it date time.
- Data type of tweet id in three data frame to string instead of int.
- Data completeness the three data frames has different number of rows in twitter_archive_clean = 2097

number of rows in image_predictions_clean = 2075 number of rows in json_df_clean = 2354

- In image prediction we assume that each 3 false prediction mean it’s not dog image, although in about 15 image it is dog .
- So we will not drop any data in this table and keep it for future analysis. But in the last twitter_archive_master.csv we will keep only most confidence and true value data and set all negative prediction to null .
- After merge dog stages we notice validity issue as for some dogs there where multiple stage by revising data we decide to keep both with comma as separator.
- Get highest confidence with true value for dog types

Tidiness issues:

- columns doggo,floofer,pupper,puppo is variables not column names , so we should merge them to dog_stage column.
- many columns has ambiguous names which suppose to be rows not column

Cleaning

First we make copy of default data frames .

we use python packages to implement issues in assessment section pandas, numby , matplotlip and others we first Define problem then code and finally test .

by using lessons and many resources like https://www.geeksforgeeks.org/

https://stackoverflow.com/ , which can be found in coding file.

At last I Merge all Data to create twitter_archive_master.csv last file

To reduce data redundancy we only get statistical the best confidence and true prediction

and at the same time we will keep file image prediction for future analysis as I notice that there some false values is for dogs by visual assessment.