10/17/2020

Data Wrangling Report

Abdelrazek Rizk

Data Wrangling Report

About the Dataset(s)

1. Twitter archive for WeRateDogs account.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.
WeRateDogs has over 4 million followers and has received international media coverage.
This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

2. Image Predictions File.

The dataset is 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).

1- Gathering Data for this Project:

a) downloaded the Twitter archive https://support.twitter.com/articles/20170160

b) downloaded the tweet image predictions file using the Requests library

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

c) Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called "tweet_json.txt"

Gathering process:

I. Make Directory for my work space
II. Imported library pandas NumPy requests tweepy json nbconvert pyppeteer pandocfilters
III. Imported functions
IV. Read twitter-archive-enhanced as pandas data frame and quick check to view structure
V. Download tweet image predictions TSV using the Requests library and write it to image_predictions.tsv
VI. Query Twitter API for each tweet in the Twitter archive and save JSON in a text file
read tweet's JSON data line by line and convert to a Data Frame
Create a Data Frame with 'created_at'','tweet_id','place','retweet_count', 'favorite_count', 'display_text_range

2- *Assessing Data for this Project:*

Using panda data frame function df .info (), df.value_counts (), df.columns, df.dtypes
To explore quality issues

3- *Cleaning Data for this Project:*

I. Make copy to my dataset
II. Drop Rows with Missing Values
III. replace unknown name like “a” with name "apple"
IV. Drop empty column with null Values
V. Crete stages_of_dog column to marge 'doggo', 'floofer', 'pupper', 'puppo'
VI. replace URL source by the main source
VII. Marge all data set in one data frame main_df.csv

4- *Reporting for this Project*

Create a word written report called wrangle_report.pdf
describes my wrangling efforts

Create a word-minimum written report called act_report.pdf
that communicates the insights and displays the visualization(s) produced from my wrangled data