Wrangle Report

Author: Nguyen Gia Bao Le

October 17, 2022

This is the document for the wrangle data for Udacity Course - Wrangle and Analyze Data Project

Dataset:

- Enhanced Twitter Archive
- Image Prediction File
- Additional Data via the Twitter API

Description:

- Enhanced Twitter Archive: The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. Each record does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo)
- Image Prediction: The images in the tweets forward to a neural network to predict the breed of the dog in image. Each record contains top 3 predictions, includes the class name, the confidence score and the value whether that class is a breed of dog.
- Additional Data via the Twitter API: each record contains much information about the tweet. It contains id of tweet, the created time, favorite count, etc. In this project, I only focus on favorite count, retweet count of the tweet.

Technical:

- Download dataset (if it is available)
- Get dataset from request
- Crawl dataset from API

Requirements:

- Pandas
- NumPy
- Requests
- Tweepy
- json

Detail of wrangling data:

- Image Prediction File:
    - Dataset is a response with CSV structured
    - Using “requests” library to get API and receive the response.
    - Save the response into a CSV file
- Additional Data via the Twitter API:
    - Dataset is in permission of Twitter
    - You have granted the token to get API, by address with Twitter what you are and the purpose you need to get API. If pass, you will be given token to get API
    - Using “tweepy” library to get API. Input is tweet_id and your access token
    - Save multi responses into one text file, each response in a line