RatingDogs

overview

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python

Steps

1-Data Gathering

The WeRateDogs Twitter archive. I am giving this file to you, so imagine it as a file on hand. Download this file manually by clicking the following link: twitter_archive_enhanced.csv

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. 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 file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID

2- assisment

Quality

-Culomns have many useful data because most data is null in tweet archive

-timestamp must be datetime not object in tweet archive

-source apear in html tag in tweet archive

-23 row with wrong rating denominator value (wrong entry data) in tweet archive

-name dog column contains (745 row is null),(55 is a) in tweet archive

-66 row in image prediction is dublicated in image prediction

-in data frame tweet status column source is already in another data frame in tweeter status

-in data frame tweet status useful column is (id,retweet_count,favorite_count) in twitter status tidiness

-3 df must be on a single data frame

-p1,p2 and p3 compain in one column category dogs

-doggo,floofer,pupper and puppo columns must be one column

3- Cleaning Data

Clean steps¶

1-Copy Datat frame in new

2-removeun useful column 'text','in_reply_to_status_id' ,'in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp'

3-convert timestamp to datetime get column month,year

4-Clear html tag From source column

5-remove 23 row wrong rating denominator value

6-compain doggo,floofer,pupper and puppo into stage_dog

7-remove doggo,floofer,pupper and puppo column

8-remove 66row dublicated image url in image prediction dataframe

9-get categroy_dog From image prediction dataframe

10-keep tweet_id and category_dog

11-keap (id,retweet_count,favorite_count) in status dataframe

12-Rename id to tweet_id in status dataframe

13-compain 3 dataframe