wrangle_act

July 31, 2018

1 Data Wrangling Project

In this project I will wrangle and analyze the tweet archive of Twitter user @dog_rates. @dog_rates that rates people's dogs with a some tweet about the dog. These ratings almost always have a denominator of 10. @dog_rates asks people to send photos of their dogs which are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as ''13/10''.

1.1 Data Set

- The First Data set is a csv file which we need to access it by downloading it

- The Second Data set is a 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 use code by Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv

- The Third Data set is need to be access using twitter API in and save in a file in json format after that it need to be converted to dataframe

The Data was downloaded using python in Jupyter Notebook using pandas, requests and tweepy libraries. The tricky part is tweepy libraries because we need the approval of twitter for app and then we need to get key and then access it

1.2 Access

Access is second part of data wrangling and most important one for Model. The process of the Access is an iterative process which help for data to be more clean the issue is divided into 2 part:

- Data quality issue
- Lack of tidiness issue

The following issue was there in the data

Quality

Df Dataframe

- Text column should not have links
- Timestamp column is in string
- missing name of the dogs
- dog's name are wrong
- The rating_denominator should not be more than 10
- text column should not start with RT @dogs:
- ''&amp'' characters present in text

Df2 DataFrame

- The predicted breed of dogs should be in capital letter

Tidiness

- json_tweet column name id should be changed to tweet_id to align it to other
- column name TimeStamp should be named tweet_timestamp
- duplicate data on json_tweet dataframe
- merger of all the dataframe

1.3 Clean

This is the third and last part of the data wrangling and this is the part where the issues identified are fixed. The dataframe are first copied to make sure if we did some wrong step the original dataframe does not changes.The challenging part is to know how to solve the issue but by searching it on the web I could able to know and That's the best part because we get to know the new things. After fixing the issue I copied the dataframe to a csv file