Wrangling Report of

Udacity Wrangling & Analyzing Twitter Data

Soner Nefsiogullari

Introduction

The aim of this project is to practice what I learned in the lessons from Udacity Data Analyst Nanodegree Program. I will use tweet archive of Twitter user @dog_rates, also known as WeRateDogs. The account rates people's dogs and always give points with denominator of 10. However, since they think every dog deserves good rating, the account usually gives more than 10 points as numerator.

Actually they have a popular meme about this rating system called "*they're good dogs Brent".* When a person namely Brant, complain about their rating logic and they gave this humorous answer.

The analysis consists of ratings, retweets and favorite counts overtime.

Gathering Data

The data consists of three different sources for this project. One is coming from WeRateDogs, which is csv file of twitter archive. I read this file using `pandas.read_csv()` function. Second is a tsv file that contains convolutional neural network predictions which I programmatically downloaded using the Python `Requests` library. Last part of data is coming from twitter which I queried with twitter API using Python's `Tweepy` library. I read this file using `pandas.read_json`.

First I learned how to use Twitter API by searching *stackoverflow.com.* The links that shared in the course page also helped me to get my tokens and keys.

Assessing Data

After gathering process, I want to assess data visually and programmatically. First, I printed the dataframes and check how they looks like. After that, I use info, describe, unique, value_counts

functions to understand the general information about dataframes. By reading these functions output, I was able to interpret the type of data stored in columns and which columns have missing values. Hence, I convert the columns in a proper format such as; I changed the tweet_id column as string, timestamp column to datetime, rating column to float. I changed the column names if it is necessary in Image prediction dataframe.

Cleaning Data

Back-up first! Before I change the data every time I get used to take copy of original files to prevent data loss in case of any mistake. First I merged three dataframe into one dataframe with using inner join on tweet_id. Then, changed type of dog columns into a one column. After that, I dropped unnecessary columns for this project. Instead of storing numerator and denominator, I use rating as score. I also find miswritten numerator and denominator and fixed them by checking actual tweets. I rename some names such as “a”, “an”, “the”, into None. I filtered the image predictions to find the actual dogs.

After renaming and cleaning needless columns, I used some visualization tools to see summary of data and correlation between attributes. Seaborn and matplotlib libraries was very helpful with scatter and barplot functions. In addition, I found top rated and top favorited dogs in my cleaned dataframe.

References

http://knowyourmeme.com/memes/theyre-good-dogs-brent