Data Wrangle Report

by Beiran Chen

Jan.11th, 2019

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators always greater than 10. 11/10, 12/10, 13/10, etc. WeRateDogs has over 4 million followers and has received international media coverage.

The data I got in this project comes unclean. Using Python and its libraries, I will gather data from 3 kind of sources which are in a variety of formats. Then I will assess its quality and tidiness, then clean it using Python (and its libraries) .

Gather

The data of this project has three different data source:

1. archive: the tweitter_archive_enhanced2.csv was provided by Udacity and downloaded manually. Then I can read csv file by pandas library.

2. image_prediction: downloaded programmatically using the Requests library from Udacity servers.

3. tweet_json: I applied a Twitter API and get data by Tweepy library and stored it in the local as Json file. Then read json file by pandans library.

Assess

After gather three tables, I assessed the data visually and programmatically.

1. Visually, I printing the three entire datasets in jupyter notebook, and I also check them in Excel

2. Programmatically, I using info(), value_counts(),sample(), duplicated(), etc.

Clean

In this part, each clean is presented in three steps : define, code, and test.

First, I made copy of each original data frame, and began to do clean on the copy data.

1.archive table

*Keep only original ratings;

*Timestamp column should be in datetime format;

*Split 'datetime' column into 'date' column and 'time' column;

*Transform rating_numerator and rating_denominator columns type into float;

*Change some error in numerator;

*Correct (archive_clean.rating_numerator) fraction values;

*Correct denominators.

2.image_prediction table

*Delete duplicated jpg_url;

*Merge prediction column together.

3.tweet_json table

*Rename the id column to "tweet_id" to match the other 2 tables.

Tidiness

1.archive table

*Remove some columns won't be used in anaylze.

*Melt 'doggo','floofer','pupper','puppo'column into a new column 'dog_stage'.

2.image_prediction table

*Delete columns won't be used in analyze.s.

3.tweet_json table

*Remove the columns that won't be used

Joint all 3 tables together

The most challenging cleaning step for me is to correct some rating numerators. Those wrong rating number always came from the tweet text, such as extract from date time or even the fraction part of the real rating number. So I found out the strange numerators, and check their full text. And change them.

Conclusion

Data wrangling is a core skill in data science. I used Python and its libraries to gather, assess, and clean the original data which is needed in the future analyze.

1. In this project, we use three ways to gathering data: reading from local file, scraping data from web , and using API to collect data.

2. Assessing data visually and programmatically can help us to get known about data.

3. Cleaning data made the data quality better and more tidy.