Wrangle_report

December 22, 2017

1 Wrangle Report

In this data wrangling project, I use Python Tweepy to collect Tweet messages via Twitter API. The project needs gathering the favorite count and retweet count for each tweet from the tweet list stored in image-prediction.tsv. I write a script to download all tweet status information, extract the favorite count and retweet count from the JSON file. I am able to retrieve most of the tweet status except for those tweets without a status. There is an issue for some tweets, the favorite count appears 0 while retweet count is very large, which is unreasonable. After an investigation, I found these tweets are actually retweets from other tweets, so the original favorite count is printed as zero, but check the retweet status favorite count, which should be the correct favorite count for this tweet.

After gathering the tweets data, I merge all data sets into one and output it to tweet_json.csv. I use Tableau to visually assess data also write Python scripts to look into data programmatically. I've listed the details of my findings in the Jupyter notebook. It's not hard to see several wrong variable types, for example, tweet_id should be character instead of numeric. I also notice some dog names and dog ratings are not correctly extracted. The text column contains hashtag and webpage links, which can be parsed if needed. There is a tidiness issue that I see multiple pictures or a picture with more than one dog, obviously the breed prediction model can identify only one dog per a picture. I don't know how the model's author handles this issue. @dog_rates not only publishes pictures but also occasionally shares dog videos, of which the tweet should be dropped from the final data set.

The next step is to clean the data. It's tricky to find the dog names from text column. It's easy to see most dog names are introduced in a fixed pattern, etc. "This is Charlie". However, there are a few names appearing in the middle of a sentence that needs semantic analysis to determine. Moreover, the dog names are not as simple as one word with first letter capitalized. It's funny I found dogs with both last name and first name, or names with triple words. Anyway, I drop those rows with funny names or more than one name for the dogs.