Wrangle_report

by Theodora Koutsothanasi

Introduction

In this report we will describe our wrangling process, needed for the wrangling WeRateDogs project.

Data wrangling is the process of

- Gathering,
- Assessing and
- Cleaning data

1. Gathering

We gathered the data from three different sources.

A. **WeRateDogs Twitter archive,** which we downloaded it manually by clicking the following link **twitter_archive_enhanced.csv**

B. **Tweet image predictions,** containing data regarding what dog breed is present in each tweet according to a neural network. The file(image_predictions.tsv) is hosted on Udacity's servers and we downloaded it programmatically using the Requests library. The URL link is:
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

C. **Tweets,** containing for each tweet, retweet count and favorite ("like") count and any additional data we find interesting. We use the tweet IDs in the WeRateDogs Twitter archive, and 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 has been written to its own line. Then we read this .txt file line by line into a pandas DataFrame,quering the info about tweet ID,retweet count, and favorite count.

Completing successfully the first step of the data wrangling process, we obtained the above data from a various number of sources. We downloaded and saved it the twitter archive and then we read it in a jupyter notebook using pandas library. Also, for the Tweet image predictions, we used requests library to download the data and we read it using again pandas library.Finally, by quering an API we managed get the JSON file (Tweet.json) and read it again in our notebook using pandas.

2. Assessing

The next step of our data wrangling process was to assess, the data we have collected, both programmatically and visually, and look for quality and tidiness issues. We found 2 Tidiness issues and 10 quality issues.

Tidiness Issues

1) Merging the tweet data and twitter archive and image predictions into one master dataset in 'tweed_id'.We have single type of observational unit 'tweed_id' spread out over 3 different tables.

2) No need for seperate dog stages columns (e.g 'doggo','floffer','pupper' and 'puppo'). One column should be created,containing the necessary info.

Quality issues

twitter_archive table

1. Columns with missing data.For instance,"in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp" have less than 200 rows of non null values. Delete the columns and the row data for in_reply and retweet data.
2. "expanded_urls" has missing values, meaning that these records had no images.Take into account only the dog ratings that have images.
3. The datatype of "timestamp" is not correct. It is a string (object).
4. The standard number for "rating_denominator" is 10. Some discrepancies appear.
5. The "rating_numerator" also should have 10 as a max value.
6. Incorrect-non logical dog names. E.g (a,an etc)
7. Dog names format
8. Readable source format

image_predictions table

1. Confusing columns'names (p1_conf etc)
2. Format name inside columns (p1,p2,p3)

3) Cleaning

The final step of our data wrangling effort is cleaning.We fixed quality and tidiness issues.I used only the programmatic way to fix the issues, following the process define,code and test. Before starting the cleaning I created copies of our datasets, making sure that that any change will be applied in the original files. Each finding from the assessment process has been resolved and I stored all the necessary data in a master file called twitter_archive_master.csv.

Conclusion

Data wrangling is an important asset, and each data analyst should first follow this process before making the data analysis. Specifically, in this project it would be impossible to find useful insights without the data wrangling process.