Data Wrangling Report

Marc Hilty

07.01.2019

Gather

I was provided with 2 files directly from Udacity: a .csv file (archive.csv) and a .tsv file (predictions.tsv). I downloaded them both programmatically to ensure reproducibility. The third file was gathered using the twitter API and saved to a json file (tweet_json.txt).
Archive.csv is a collection of tweets that was provided by @WeRateDogs and contains information about the tweet like text, rating and the dog categorization they use (doggo, floofer, pupper, puppo).
Predictions.tsv contains information gathered from @WeRateDogs using a neural network to predict the dog breed in the image @WeRateDogs posts together with their tweets.
Tweet_json.txt contains all additional information I could gather using the tweepy, the twitter API. I mostly use retweet_count and favorite_count as an addition to the existing dataset.

Assess

Assessing the data is quiet a task. First I visually assess all three tables with Microsoft Excel. It allows me to get a good first impression. I can already see some quality and tidiness issues. The next step is to dig further into the depths of the data. I programmatically assessed the tables and specifically looked for inconsistencies like duplicated id's using .nunique() and .duplicated(), data type issues and missing entries with .info(). Even though there are more issues in this dataset I settled with the following:

Tidiness:

1. Merging of the tables into one master file
2. Melting doggo, floofer, pupper and puppo into one column dog_category

Quality:

1. Delete retweets and replies in `archive`
2. Replace missing URLs in `expanded_urls`
3. Delete invalid names like 'a', 'the', etc. in `archive` or find a way to replace them properly
4. Delete duplicate IDs in `tweet_addition`
5. Delete duplicate jpg_urls in `predictions`
6. Drop unneccessary columns
7. Some datatypes are messed up e.g. timestamp should be datetime
8. Some pictures are not from dogs, use `predictions` to find and delete them
9. Create a single rating column out of `rating_numerator` and `rating_denominator`

Clean

I learned a lot from this process. At one point I had to start over because I merged the tables first and had a number of consistency issues because I didn't clean the individual tables beforehand. I had to go back to the course material a couple of times and google proofed to be a really good friend when trying to figure out how to optimize a specific task. I provided links to all the source material I used in the end of wrangle_act.ipynp.
Quiet often I had to make a decision if I want to keep data but lose validity to some extent or just remove more to be consistent. Due to my scientific background I chose the second option and removed everything that could possibly compromise the validity of the dataset.