Wrangle Report

by Paul Stancliffe

May 2018

This short report describes the wrangling efforts involved in completing the "WeRateDogs" project as part of Udacity's Data Analysis Nanodegree.

The Data Wrangling process consists of:

1. Gathering the data
2. Assessing the data
3. Cleaning the data

1. Gathering

Gathering Data for this Project involved obtaining three different datasets from three different sources. Each one testing a different way of obtaining a dataset.

The first was to download a file manually and be able to open a csv file. In this case the file was called twitter_archive_enhanced.csv and was the file consisting of the largest amount of data.

The second was to be able to download a file programmatically using Python Requests library. The file contained image predictions on the breed of the dog coming from a neural network on some of the tweets already downloaded in the archive file. The file was in tsv format and tested your ability to open this type of file successfully.

The final dataset tested your ability to query Twitter's API and use a Python library called Tweepy to obtain further data on the tweets in the archive file using the tweet id. The Tweepy library returned the data in json format, from which it was possible to iterate through and append data to a file as a list of dictionaries and then a pandas data frame. A copy was saved in csv format of the data frame created.

2. Assessing

The three saved data frames were then assessed visually inside a jupyter notebook with pandas and because the datasets were not too large, a copy of each was exported into one Excel workbook. This allowed quick scanning through the rows and use of filters to identify areas for more detailed investigation. Following this a programmatic assessment was made inside jupyter with pandas using the following functions, df.info(), df.head(), df.sample(10) (several different samples were taken), df.value_counts().

The datasets were accessed under two criteria, quality and tidiness. When an issue was detected it was documented under one of these two criteria.

Quality refers to issues related to the content of the data, sometimes called dirty data. The standard criteria of completeness, validity, accuracy, and consistency of the data were used to identify quality issues. These issues were varied and are listed in the assessment section of the "wrangle_act.ipynb" jupyter notebook.

Tidiness refers to issues related to the structure of the data, sometimes called messy data. The basis for assessment is that each variable forms a column, each observation forms a row and each type of observational unit forms a table. After assessing the three datasets, it was decided to marge them into a single data frame, reducing superfluous columns that wouldn't be needed in any future analysis.

3. Cleaning

The final step in the wrangling process is cleaning the data for quality and tidiness issues. The cleaning followed the standard process of define, code and test for each of the issues and they were tackled in a logical

order, which is reflected in the numbering order in the “wrangle_act.ipynb” notebook and closely followed standard practice of cleaning missing data first, then cleaning for tidiness and finally quality.

Most of the cleaning was performed using programmatic tools, such as def functions or pandas built-in functions (merge, melt etc), but some manual cleaning was also performed to correct ratings and dog type row errors.

Conclusion

Data wrangling provides a clean data frame for future analysis and visualization, in our case we concluded with the “twitter_archive_master.csv’. This file can also be shared with others without having to wrangle the data.