wrangle_report

December 6, 2017

1 Introduction

In this paper we will describe our wrangling effort made in the section of wrangling weRateDog project

Data wrangling consists of:

- Gathering data
- Assessing data
- Cleaning data

1.1 Gathering

Gathering Data for this Project composed from three pieces of data as described below:

- The WeRateDogs Twitter archive. We manually downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv
- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) hosted on Udacity's servers and we downloaded it programmatically using python Requests library on the following (URL of the file: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)
- Each tweet's retweet count and favorite (i.e. "like") count and any additional data we found interesting. Using the tweet IDs in the WeRateDogs Twitter archive, we could 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 stored in a line.

1.1.1 Gather: Summary

Gathering was the first step in the data wrangling process. We could finish the high-level gathering process: - Obtaining data - Getting data from an existing file (twitter-archive-enhanced.csv) Reading from csv file using pandas - Downloading a file from the internet (image-predictions.tsv) Downloading file using requests - Querying an API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy - Importing that data into our programming environment (Jupyter Notebook)

1.2 Assessing

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues was our next step. We could detect and document the following quality issues and tidiness issues.

1.2.1 Quality

Completeness, Validity, Accuracy, Consistency => a.k.a content issues **archive dataset** - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be intergs instead of float - retweeted_status_timestamp, timestamp should be datetime instead of object (string) - The numerator and denominator columns have invalid values - In several columns null objects are non-null (None to NaN) - Name column have invalid names i.e 'None', 'a', 'an' - We only want original ratings (no retweets) that have images - We may want to change this columns type (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and tweet_id) to string because We don't want any operations on them **images dataset** - Missing values from images dataset (2075 rows instead of 2356) - Some tweet_ids have the same jpg_url - Some tweets are have 2 different tweet_id one redirect to the other **json_tweeets dataset** - This tweet_id (666020888022790149) duplicated 8 times

1.2.2 Tidiness

Untidy data => a.k.a structural issues - No need to all the informations in images dataset, (tweet_id and jpg_url what matters) - Various stages of dogs in columns instead of rows archives dataset - We may want to add a gender column from the text columns in archives dataset - All tables should be part of one dataset

1.3 Cleaning

Cleaning our data is the third step in data wrangling. It is where we fixed the quality and tidiness issues that we identified in the assess step.

We used the two types of cleaning, the manual and programmatic even the manual not recommended but the issues were one-off occurrences. Our process was Define, Code and Test and we were always making a copy of tha dataset even we made the copy in file to test the change before applying to the main dataset. We didn't spot all the quality and tidiness assessments at the assessing data section, so we have been iterating and revisiting assessing to add these assessments to our notes.

2 Conclusion

Data wrangling indeed is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. We couldn't be able to make some of the visualization without wrangling (i.e dog gender partition) **So best practices say wrangle. Always.**