Data Wrangling Project: Internal Report

Brief Description on Wrangling Effort

One of the major efforts I have made for the data wrangling in this project is on how to correct some data in the datasets downloaded through internet or read from the csv file. Some data in the datasets such as the rating of a tweet and name of the dog in the tweet are extracted from the original text of the tweet. However, the program used for it seems to be too simple and sometimes fails to extract the actual data (as I have found in the data wrangling process, some ratings and names of dogs are obviously wrong). I thus tried to look back the original texts for the tweets, extract the correct data, and replace the incorrect data in the columns for rating ("*rating_numerator*" and "*rating_denominator*") and name of the dogs ("*name*") by the correct ones. It was an interesting experience that through the data wrangling process, I could somehow understand how Twitter extracts the various informations from the text of tweets.

Another effort I made for the wrangling is the choice of the order for cleaning the data. For example, when we need to remove some rows from the dataset at some point, it is efficient to do so at the early stage. Otherwise, even I cleaned up one row for some issue, that row could be removed during the cleaning for other issues. (For example, the retweet must be removed in the data wrangling process for this project. Thus, it is not meaningless to clean up the issues in the retweets. Removing them first is the quickest way to proceed the data wrangling process.) I expect that the nice ordering for the cleaning will also help the other people to grasp the detail of the cleaning process easily

As mentioned in the project instruction, since there are a lot of issues to be corrected in the datasets, it takes a while to clean up completely. Therefore I selected various types of issues. For the assessment I tried to use a lot of programming methods This helped me to grasp what the data sets are like in the course of the assessment.

(LaTeXwas used for making this report)