Data wrangling

Index

Overview

We have wrangled twitter's data from **WeRateDogs**. The Udacity team starts then I continue.

Data wrangling is to gather data, assess, and then clean it.

Gather

We collected data from **WeRateDogs**. We use three types

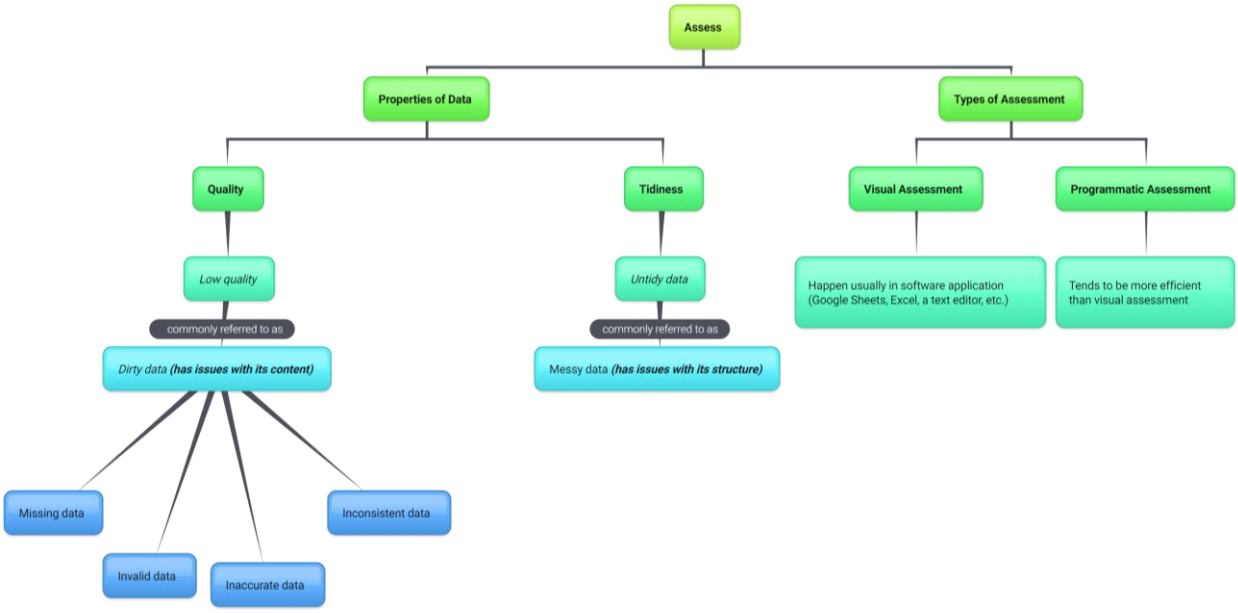
1. CSV
2. TSV
3. JSON

CSV: This is gathered by the Udacity team and we can say that this is the main part, also the Udacity team put it in a file called 'twitter-archive-enhanced.csv'

TSV: This is made by the Udacity instructor. He ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs, and took the first three. He uploaded it to **image-predictions**. I downloaded it using requests or you can just click on it. I have used sep by Tab instead of comma as it is "Tab-separated value".

JSON: Here I used API provided by twitter to collect data with the same tweet_id of 'twitter-archive-enhanced.csv' in order to fix any quality issue if exist.

Asses



I have assessed the data to get the low quality and tidiness issue to fix it. First I divided the data to categorical and numerical data to identify which one should be float or integer and which one should be category.

Then I looked for the columns that we needn't it (retweeted_status_id...), or we need it then remove it as in (in_reply_to_status_id) . We need it to get the duplicated tweets or retweeted. Moreover, I have removed any duplication in url even in rows or in the same row (same record have many link which are the same).

There are many quality problems in the data from 'twitter-archive-enhanced.csv' table, I tried to fix it by the json or -the file from Twitter API-, but I couldn't because I didn't find the record or values that I want in json file.

There are many Tidiness problems that I couldn't fix as missing data in the image-predictions data, because I don't know yet how to use neural network.

I used to assess the way of asking questions to get the most from the data.

Clean

Cleaning is improving quality and tidiness to get the most from this data.

I have tried to solve the problems that have been arisen in the assessing data.

I started with tidiness as it is easier and will help us in the quality issue.

I used Pandas data frame as it helps us tremendously in cleaning and analysis.

I used to organized my steps into three Define, code, test

1. Define is the part we say the problem
2. Code is the part we use code to solve it
3. Test is the part we test whether we solved it or not.

The first and most important part is to copy the data frame and not work with the original one. Because we may damage it, if we use the copied we can reset it through the original one, also we should use copy method as if we use assignment operator we don't make a copy instead we point the same place at the memory. Hence, use the original one and if damaged we will not be able to reset it.

Visualization

This is not a part of data wrangling but I have made it to make sense with the data. I will speak more about that in another PDF (act_report.pdf)

---

storing

Last but not least I have stored data in twitter_archive_master.csv file and all_columns.db (SQL).
