DATA WRANGLING REPORT

By Kevin Foong

Gather

I used 3 files in this project. They were:

- twitter-archive-enhanced.csv – This file was provided to us and it was loaded in using read_csv.
- image-predictions.tsv – This file was hosted on the Udacity server. It was downloaded using the requests package.
- tweet_json.txt – This file was downloaded one line at a time via the Twitter API. The steps performed were:
    - Applied for a Twitter developer account.
    - Download JSON object of tweets.
    - Save each tweet as one row into a file.
    - Read each tweet from the file into a dataframe. Only select tweet_id, retweet_count and favorite_count

Assess

I opened each file in Excel and took a quick look to get an overview feel of the data and to perhaps identify any "quick-wins" problems. I then went through each file one at a time programmatically in more detail in Jupyter notebook.

For each file I explored each field independently, exploring it from various perspectives, where one discovery often led to another. A typical process would be:

- Take a look at the overall table structure, variable names, data type and if there are any missing values. (df.info)
- Assess the data visually by taking a look at a small portion of the data. (df.head or df.sample)
- If it is appropriate, examine the spread of data. (df.field.value_counts)
- Through this process if we find certain issues stand out then we will dive into these issues by looking at the data from various perspectives to try and identify further problems. Sometimes we will find further problems and sometimes we will hit a dead end.
- If problems are found it will be noted down. At this stage we are only describing the problem not how it is to be solved.

Clean

I go through the list from the assess stage and for each problem, or groups of problems, we define how it is to be addressed. This is where we provide the specifics such as naming the field and what operation we will be performing on it.

Once it is defined we then proceed to implementation. I have included Clean & Test in one section as I find that cleaning and testing are often intermingled. For example, in a typical process, we might see what the dataframe looks like before, perform the operation, then see what it looks like after to prove the operation has succeeded.

This process is repeated through all the issues identified in the assess stage. At the end we end up with one clean dataframe. This dataframe is then saved as a new CSV file ready for analysis.