wrangle_report

December 11, 2019

The main challenge I experienced when wrangling this dataset was a relatively open brief. Having no specific focus made it particularly challenging to make decisions which data to keep / discard and which format is best suited for particular types of data. Resisting the urge to keep cleaning and rearranging data was an exercise in self-control.

Data wrangling process naturally lends itself to iterative approach, however this becomes a particular challenge when the outcome needs to be structured in a linear narrative report form.

Practically speaking, one of the first challenges I faced was when downloading twitter archive and saving it into a properly structured json file. Making sure all brackets and parentheses are positioned correctly took some debugging time.

Another peculiar challenge arose when looking into int and str representations of tweet id's. When reading the json with read_json function with default set of arguments - there was considerable amount of mismatches between the str and int columns. Eventually it appeared that the function's type inference option was behaving in a very strange manner, corrupting some of the strings when converting them into ints. With the dtype=False flag turned off things came back to normal. There is an open issue #20608 on pandas github about this: https://github.com/pandas-dev/pandas/issues/20608

Significant amount of data in the merged dataset was duplicated, particularly using 'int' and 'str' types for data integrity purposes. Once basic checks were done that data is intact - all the duplicate data was removed.

Some other columns were nearly empty and as such were difficult to make use of in our analysis. Therefore data from them was saved to separate variables and columns deleted from the dataset.

One interesting detailed discovered in the process is how Twitter JSON uses British and American spellings of the word 'favourite', namely: 'favourites_count' inside user object to indicate the count of favourites user has given over the account's lifetime and 'favorite_count' in the tweet object to indicate how many favourites this particular tweet has accumulated.

Exporting the report into a PDF has also presented a number of challenges in resolving dependencies and reading issue logs on github. Namely issue #1105 in jupyter/nbconvert tool repository.