wrangle_report

February 14, 2019

1 Wrangle Report

1.1 Introduction

The aim of this report is to walk through this project's wrangling process. We will look at this in separate steps, i.e. gathering data, defining the scope, assessing and cleaning data.

1.2 Gathering Data

Three files were needed for this project and the details on their nature and means to obtain them are described below.

Enhanced Twitter Archive I have downloaded this file manually by clicking on the link and uploaded it into the Jupyter directory so that I could read it with the function *pd.read_csv*.

Tweet Image Predictions The file containing the tweet image predictions is in a flat structure or tabular form. I was provided with the url, which I accessed thanks to the **requests** library. I have then created and opened a new file in the directory and written the url content in it.

Tweet Info For this piece of data I had to first set up a Twitter Developer account (note I already had a normal Twitter account). I have then created and described a new application within the Twitter Developer portal and obtained Keys and Tokens.

In Python, I imported **tweepy** and accessed Twitter API through the keys and tokens previously mentioned. I opened a text file in order to append the output and looped through each of the tweet ids in the archive csv file. For each tweet I get the retweet and favorite count and store it in a dictionary together with the corresponding tweet id.

With the function *json.dumps* I store this dictionary in a json file, which I then read line by line in order to load the data in a Pandas DataFrame.

1.3 Defining the Scope

Here, I have filtered out tweet ids that are retweets or have no images, in order to work directly with the pool of data that we will need to analyse. This should make running the code and the cleaning process faster.

1.4 Assessing Data

Through visual (scrolling) and programmatic assessment (using functions such as **info, describe, value_counts**), I have identified 6 Quality issues and 2 Tidiness issues. I have decided to work on these first and reassess once I had given the data a "first clean". In fact, I identified other 2 Quality issues at a later stage.

1.5 Cleaning Data

While the *Assessing Data* step looks somewhat unstructured, the cleaning process consists in focusing on one issue at the time and defining the problem and action that will be taken to resolve the problem, code it and test it.

This was done throughout the different dataframes, and the steps taken were around:

Quality

- Data types
- Data format (for instance, lower or upper case)
- Wrong data (for instance, showing random words instead of dog names)
- Messy/unclear data (irrelevant information can be deleted)

Tidiness

- Dataframes can be merged
- One variable per column

As mentioned earlier, I then performed again the assessment steps to finally perform two last quality-improving actions. After that, I have stored the data and started my analysis.