wrangle_report

June 24, 2020

1 Data Wrangle Report

1.1 Introduction

This report briefly describes the various tasks performed for the Data Wrangling process of the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

1.2 Gather

- The data was extracted from three sources.
  - The given csv file from The WeRateDogs Twitter archive.
  - The tweet image predictions file downloaded from Udacity server.
  - json data querying Twitter API.
- WeRateDogs twitter archive.
  - This data is downloaded to a DataFrame using the pandas read_csv command.
- Image predictions.
  - The tweets image prediction file is downloaded from Udacity's servers programmatically using the Requests library.
- json Data.
  - First, sign up for a Twitter account. Then set up a developer account.
  - Create an app, Then go to the Keys and Tokens tab on this page to find or generate the Consumer API keys, and the Access Token and Access Token Secret that will be required to get the json data using Twitter API.
  - Write Code to obtain json data and store it in a file.

1.3 Assess

Quality

- Manually screen the data copied to the Data Frames.
  - the columns having Nan values. For example there are missing expaned_urls in the twitter archive DataFrame.

  - the columns having irrelevant data. For example the name column has wrong data (it has words such as a, the, actually). Also, the source column has lengthy data it has the full href tag and it can be reduced to something appropriate.

- Using info() look at the data types of each field.
- Using describe() look for values which looks odd.
  - Check the mean, max and other values. For example the max of rating numerator is high, so this needs to be cleaned.
- Using duplicated() check for duplicates.
- Programmatically screen data.
  - Check if retweet and reply data is included.
  - Check if all the tweets are twitter archive data are present in the image prediction data.
  - Check for rating values that are too high or too low which will hamper the outcome.
  - In the image predictions file look for rows where all predictions have failed.

Tidiness

- Drop columns which don't have proper values and are not useful in analysis. For Example, Since we are not interested in retweets and replies, the columns related to them can be dropped.
- The Compactness of data, For example, The various stages of dogs can be made into a single column.
- Merge data if required. For example, json data can be combined with twitter archive data

1.4 Clean

- The goal of this step was to get a clean dataset which will help in analysis and visualizations.
- Each data assessment from the above step was divided into three parts. Define, Code and Test.
- First a copy of the existing data frames were obtained. This helps to protect your information in case you make a mistake.
- Define, here we write our approach to how this data assessment will be handled.
- Code, here we write code to fix the shortcomings.
- Test, here we test if the cleaning process was successful.
- Cleaning is an important part of data analysis. More data helps in useful predictions but each data should also be clean and improvise our analysis. A detailed scanning should be performed to pick the right data from the obtained data.

1.5 Store Data

- Store the clean data obtained from the Data Wrangling Process.
- Here Data is stored in a csv file.