

Wrangle and Analyze WeRateDogs Data

Wrangle Report

Introduction

Project to display the method and skills used and acquired during Data Wrangling.

The data wrangling would be done for data from twitter account @weratedogs. WeRateDogs is a funny page which rates dog breeds.

Projekt Tasks are as follow:

- **Gathering**
- **Assessing**
- **Cleaning**

Let's take dive in each of the tasks.

Gathering Data

Data was gathered from three different sources for this project.

- Twitter Archive File

twitter-archive-enhanced.csv file was provided by Udacity, I downloaded it manually and using pandas generated a *pandas* dataframe. This dataframe contains information about **tweet ID**, timestamp, text and many other information.

Shape: Dataframe contains 2356 rows and 16 columns

- Image Prediction Data

The data was gathered programmatically from this source using *request* module of python. And loading the data to a tsv (tab separated values) format. This data contains the three top prediction of the dog breed and the confidence level for each of them along with the **tweet ID**.

Shape: Dataframe contains 20175 rows and 11 columns

- Twitter API File

The data was downloaded manually by me which was provided by Udacity and loaded from a json format file. The data contains tweet ID, favorites and retweets.

Shape: Dataframe contains 2354 rows and 3 columns.

Assessing data (Tidiness and Quality Check :P)

As first step in assessing a dataframe always look for unique, NaN and missing values, which was performed for all the three dataframe. And second step is checking for column names, does the name explains the data inside those columns.

- Twitter Archive Data
 - Expanded URLs are checked for existence of more than one URL.
 - Rating denominator is assessed visually and programmatically for values greater than 10. Along with the frequency of numerator values. And visually looking at the text column and finding the misinterpreted values of the numerator and denominator.
 - Check dog types programmatically and find count of dogs categorized into more than one category.
- Image Prediction Data
 - Check if images contain only image type extensions.
 - Prediction values P1, P2 and P3 are checked.

General assessing of Twitter API Data was sufficed.

Cleaning Data (Remedies of untidy and poor quality of data)

The first and most basic step of cleaning a dataframe over the copy of the dataframes.

- Twitter Archive Data
 - 639 duplicated URLs, correct URLs are created using tweet_id field.
 - Here there were two problems in hand.
 - 'Text' column containing the rating values is extracted excluding the 9/11 and robbery days and correct rating.
 - Manually filling the denominator rating value is more than 10.

- Dog types are merged into one column and removing the old columns.

Shape of dataframe after cleaning process: 2176 rows and 10 columns

- Image Prediction Data
 - Removing duplicate image_url
 - New prediction parameter is introduced from P1, P2 and P3

Shape of dataframe after cleaning: 1691 rows and 4 columns

For twitter API data no cleaning process is required so the shape remains the same.

Storing the dataframes into one clean dataframe