Udacity Project - Data Wrangling

Data Wrangling Report

Rachita R. Puri

Introduction

The aim of this project is to use Python and its libraries to gather data from a variety of sources in different formats and learn to assess its quality and tidiness and then clean it. This is because real world data rarely comes clean and correctly formatted and therefore we should be familiar with how to complete the data wrangling process.

This project requires us to analyse data from the Twitter user WeRateDogs. WeRateDogs is a Twitter accounts that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc.

Gathering Data Process

Three data sources will be used as part of the project.

1. The WeRateDogs Twitter Archive (Enhanced Twitter Archive)

This file consists of the tweets archive of the WeRateDogs twitter user. The archive concsists of basic tweet data (for example tweet ID, timestamp, text, etc.) for all 5000+ tweets as they stood on August 1st 2017.

2. Image Predictions File

A neural network was created that helped classify the breeds of dogs. As part of the project we were provided with a file where every image in the WeRateDogs Twitter archive has been run through this neual network. The image predictions file is a table which predicts the top 3 breeds the dog in the image could possibly be; along with confidence intervals as to how confident it is that the dog in the image is that specific breed. It also contains the tweet ID for each image as well as the image url.

3. Data Downloaded from the Twitter API (Retweets/Favourite Counts)

The last data source consisted of us creating a developer account on Twitter to be able to download the data from the Twitter API. The Twitter API enabled us to download the 3000 most recent tweets. I was able to download from the API the retweet count and favourites count which was useful in creating the insights in the analysis report.

Assessing Data

As part of the project we were required to analyse all three datasets visually and programmatically to understand the issues with the data. The requirement was that we find a minimum of 8 quality issues in the data and 2 tidiness issues which we wil be required to fix within the cleaning process.

Quality Issues

1. Enhanced Twitter Archive

Issue Number	Issue Details
Issue 2	Name column contains incorrect names such as(i.e. a, an Bo, My). Some names are also appearing as None.
Issue 3	The source column is still including the html tags in the rows of data.
Issue 11	There are rows of data where denominator is not 10. This needs to be changed.

2. Image Prediction File

Issue Number	Issue Details
Issue 6	Remove the underscores in the dog breed columns P1, P2 and P3.
Issue 7	Remove the rows where P1_dog, P2_dog or P3_dog appear as False.
Issue 8	Delete the 66 jpg_urls that are duplicated in the table.
Issue 9	Combine all dog breed names and confidence interval columns into one column for each.

3. Twitter API Data

Issue Number	Issue Details
Issue 10	Only keep the original tweets

Tidiness Issues

Issue Number	Issue Details
Issue 1	All the tables should be part of one dataset, therefore will need to be combined into one table.
Issue 4	Doggo, floofer, pupper, puppo should be combined to one column.
Issue 5	The tweet_id columns should be a string and not an integer.

Cleaning Data

Using different libraries in Python I was able to fix the quality issues with the three data sources. I intially combined all three data sources together and then performed the task of correcting the data quality issues and then when creating the final dataset I have dropped any unneccessary columns which would not be required to do the analysis.