WeRateDogs Twitter Data

Wrangling Report

Kashyap Subramanian S

Introduction

This is a report on data wrangling done for Data Analyst Nanodegree – Data Wrangling @dogrates Twitter page data.

Data Wrangling

Data Wrangling is an important skill for any data enthusiast who wishes to derive meaning from data. Real world data rarely comes clean. Unclean data will lead to false outcomes, wrong decisions and the meaning or information derived will not be credible. Hence it is important to always wrangle data before using it. I have used Python and its libraries to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.

Data wrangling has 3 steps

1. Gather
2. Assess
3. Clean

GATHER

I had to gather data from 3 different sources

1. **comma separated file or .csv** – which contained archived Twitter data for the user '**WeRateDogs**'. The python package 'pandas' was used to load the dataset. It provides built-in functions like `**read_csv**` which simplifies loading process

2. **tab separated file or .tsv** – which contained outcome of image predictions which were results from neural network image classifier. I used `Requests` package which provides built-in functions for downloading datasets from URLs

3. **Twitter data** – which is in `**json**` format. '**Tweepy**' an API for accessing Twitter was used in order to connect to Twitter and download the data for '**WeRateDogs**'

ASSESS

In Assess step, the gathered data was assessed for issues. Assessment can be done in 2 ways.

1. Visual Assessment
2. Programmatic Assessment

Both these assessment methods were used to identify issues. Issues can be classified into 2 as

1. Quality Issues – due to content
2. Tidy Issues – due to structure

I was able to identify 11 quality issues and 6 tidy issues in total when Wrangling was completed.

CLEAN

In clean, I solved the Tidy and Quality issues using in-built functions and operations in **Numpy** & **Pandas**. The cleaning of Tweet text proved to be a difficult task. I had to extract Proper Names [Nouns], Rating value given and Gender. It required knowledge on '**Regex**' for extraction and identification which was difficult since my knowledge on Regex is limited. But thanks to this project, I was able to upskill myself in 'Regex' and now I can safely say, I am 'Regex' Proficient. I have used the python package "**re**" which is used for regex operations.

For extraction of names, gender…etc. I have used "**nltk"** package which is **Natural Language Tool Kit**. NLTK is extensively used in NLP (Natural Language Processing). But for this project, I have used the NER (Named Entity Recognition) of NLTP to identify names, gender and rating.

NLTK Documentation on NER provides more details.

- NLTK

Finally the cleaned dataset was stored in a **.csv** file for further analysis and future usage.