WERATEDOGS

BY

Mohamed Khaled Ahmed Nour

Udacity student

A Report on Data Analysis Effort

Abstract

"Valuable insights comes from hard wrangling" Throughout this report we will review the back scene circumstances, steps, actions, encountered challenges and the result during doing the "Data Analysis Process" for "WeRateDogs" Twitter account

Over view:

- WeRateDogs is a Twitter account 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
- Nowadays, WeRateDogs has over 8 million followers and has received international media coverage.
- WeRateDogs profile URL: https://twitter.com/dog_rates?s=20
- Wrangling and analyzing WeRateDogs Twitter account's tweet data (tweet ID, timestamp, text, etc.) for 5000+ of their tweets as they stood on August 1, 2017.
- **Our goal** is to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information

Challenges

1) **Messy data with low quality.** That requires additional hard gathering, then assessing and cleaning is required for "Wow!"- worthy analyses and visualizations.

data Challenge techniques:

1.Take a good time in knowing the datasets and WeRateDogs Twitter account.
2.An organized mindset.
3.Start small-end big way.
4.A lot of practice with "try-error".

2)**Twitter Approval Cycle**. Creating a Twitter Developer Account and getting the approval to use Twitter API. It takes a lot of time with a provability to be declined.

Cycle Challenge techniques:

a. Start approval cycle early.
b. Demonstrate my request purpose clearly with integrity.
c. Answer every piece of question from Twitter regarding my request.

3)**Data Cleaning Complex Code.** Some cleaning actions may require a complex code.

Code Challenge techniques:

a. Search at Udacity FWD Community (Discourse).
b. Use google search.

c. Search at (stackoverflow.com, geeksforgeeks.org, github.com).

Input Datasets

Input File #1: twitter_archive_enhanced.csv – a file on hand

- Manually downloaded from Udacity classroom.
- Contains basic tweet data for all 5000+ of their tweets, but not everything.
- Contains one column named [‘tweet's text’], which Udacity instructor used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo).
- Of the 5000+ tweets, Udacity instructor has filtered for tweets with ratings only (there are 2356).

Input File #2: imagee_prediction.csv – Programmatically downloaded file

- Programmatically downloaded using python request library.
- A neural network output file, which made by Udacity instructor, as he ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*.
- The results: a table full of image predictions alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction.

Input File #3: tweet_json.txt – API file

- Programmatically downloaded using python tweepy library.
- Contains every missing data from twitter_archive_enhanced.csv file, such as: retweet count, favorite count and followers count.

Output Datasets

Output File #1: twitter_archive_master.csv

- The main output file.
- Contains the clean datasets from input_file#1 (twitter_archive_enhanced.csv) and input file#3 (tweet_json.txt – API file).

Output_File #2: imagee_prediction.csv

- The Complementary output file.
- Contains the clean dataset from input_file#2 (imagee_prediction.csv).

Analysis Process

We have 2 main stage in Data Analysis Process :

Stage1: Project Preparation

1. Questions to be answered
2. Import Required

Stage2: Main Phases

1. Data Gathering
2. Data Assessing
3. Data Cleaning and Storing
4. Data Analyzing and Visualizing
5. Making Reports

Stage1: Project Preparation

1. Questions to be answered

Below are what we tried to find out...

- Q1 What are the main devices/apps that WeRateDogs' users use?
- Q2 Is there a relationship between dog rates and retweet count?
- Q3 Is there a relationship between dog rates and favorite count?
- Q4 Is there a relationship between favorite count retweet counts?
- Q5 What time that most of tweets are tweeted at?
- Q6 Is high confidence prediction meet reality more than low ones?

Stage2: Main Phases

1. Data Gathering:
   We will gather data from 3 different source as mention before.
2. Data Assessing

   Assessing data process went through 2 sections:
   1. Assessing Effort.
      After gathering each of the above pieces of data, we assessed for detecting quality and tidiness issues.
      Assessing done in 2 ways:
      1. Visually, using Jupyter Notebook and Spreadsheets.
      2. Programmatically, using Jupyter Notebook.
   2. Assessing Result.

      Assessing efforts ended with 30 detected data issues as below:

      (22) Data quality issues.
      (6) Tidiness issues.
      (2) Feature Engineering actions.

      We will list them in the next section (along with Data Cleaning and Storing Section

3. Data Cleaning and Storing

Data cleaning process went through 3 sections:

1. Cleaning Preparation.

The only and most important task in the preparation for data cleaning is to “Making a a copy from the 3 iput data files”.

2. Cleaning Process

We grouped any data issues that required the same cleaning efforts code in one group.

Below table demonstrates all the details and the cleaning logic for:

(22) Data quality issues.

(6) Tidiness issues.

(2) Feature Engineering action

Conclusion

That is over view about how we wrangling our data , I hope that is useful Thank for your time .