Data Wrangling Report

September 16, 2019

0.1 Gathering the data

The first part of the Data Wrangling process is gathering of data. Following were the steps followed for gathering the data.

0.1.1 Step 1:

The first source of data was the csv file provided - "twitter_archived_enhanced.csv". This was manually downloaded from the link provided and then manually uploaded to the project workspace.

0.1.2 Step 2:

The second source of data was the "image_predications.tsv" file. This file was hosted on Udacity's servers and was downloaded programatically using Requests library and the URL.

0.1.3 Step 3:

The third source of data was Twitter. Tweepy was used to gather this data. Prior to using Tweepy, Twitter developer account was created and an application setup in the twitter developer's site to gain access to the keys used for the authentication for using the twitter API to gather the data. The csv file 'twitter-archive-enhanced.csv' was read and the data was imported into a pandas DataFrame 'twitter_archived_enhanced_df'. Then the next step followed was to iterate through the tweet_ids in the DataFrame 'twitter_archived_enhanced_df'and fetch the JSON documents for the each tweet_id and save it to a list 'tweet_json_list'. The JSON document corressponding to each 'tweet_id' was stored in a separate line in the text file 'tweet_json.txt'.

0.1.4 Step 4:

The file 'tweet_json.txt' was read and fields 'retweet_count' and 'favorite_count' was extracted parsing through the JSON documents for each tweet_id. This was stored in the pandas DataFrame 'api_data_df'

0.2 Assessing the data

The second part of the Data Wrangling process is assessing the data. Following were the steps followed for assessing the data.

0.2.1 Step 1:

Check for Quality Issues

0.2.2 Step 2:

Check for Tidiness Issues
Following were the datasets assessed:

1. twitter_archived_enhanced.csv (data present in pandas DataFrame 'twitter_archived_enhanced_df')
2. image_predictions.tsv (import data into 'image_predictions_df')
3. Data gathered using Twitter API (data present in pandas DataFrame 'api_data_df')

0.2.3 Assessment Summary for each of the data sources is provided below.

DataFrame - twitter_archived_enhanced_df Quality Issues:

- Out of 2356 entries, 181 entries are re-tweets
- There are 78 tweets which are replies to original tweets, this is signified by the columns 'in_reply_to_status_id' and 'in_reply_to_user_id'
- The values in the column 'name' is not consistent, the dog names which are not capitalized seem to be errorneous data
- 'tweet_id' 835246439529840640 has rating_denominator 0, as per the field 'text'; the valid rating is 13/10
- 59 rows with missing images, i.e. missing values for 'expanded_urls' field
- Values in the field 'source' are not very meaningful
- Following tweets have dogs classified as more than one stage, the 'text' field has been checked to determine the correct dog stage:

  tweet_id 855851453814013952 - should be puppo

  tweet_id 854010172552949760 - should be floofer

  tweet_id 817777686764523521 - should be pupper

  tweet_id 808106460588765185 - multiple dogs - change it to None

  tweet_id 802265048156610565 - multiple dogs - change it to None

  tweet_id 801115127852503040 - should be pupper

  tweet_id 785639753186217984 - should be doggo

  tweet_id 781308096455073793 - multiple dogs - change it to None

  tweet_id 775898661951791106 - multiple dogs - change it to None

  tweet_id 770093767776997377 - multiple dogs - change it to None

  tweet_id 759793422261743616 - multiple dogs - change it to None tweet_id 751583847268179968 - should be None

tweet_id 741067306818797568 - multiple dogs - change it to None

tweet_id 733109485275860992 - multiple dogs - change it to None

Tidiness Issues:

- the fields ′doggo′, ′floofer′, ′pupper′ & ′puppo′ should actually be values under the field ′stages′
- Following columns are not required in the final dataset which will contain only the original tweets (replies & retweets and coressponding fields are not required)

  in_reply_to_status_id

  in_reply_to_user_id

  retweeted_status_id

  retweeted_status_user_id

  expanded_urls
- the different DataFrames ′twitter_archived_enhanced_df′, ′image_predictions_df′ & ′api_data_df′ should be joined into a single DataFrame

DataFrame - api_data_df Quality Issues:

- There are only 2331 entries, data for some tweets could not be retreived using Tweepy
- The tweet_ids for the data which could not be retreived have been stored in the DataFrame ′missing_tweets_df′

DataFrame - image_predictions_df Quality Issues:

- There are only 2075 entries, looks like some are missing
- The field representing the dog breeds are not capitalized

Tidiness Issues:

- The Tidiness rule has been violated. Retain only the prediction with highest confidence and those that have identified dogs.

0.3 Cleaning the data

Cleaning the data involved the following:

- Delete the retweets
- Delete the replies
- Cange errorneous values in the field ′name′ to null
- Change datatype for field ′retweeted_status_timestamp′ from ′object′ to ′date time′

- Correction required - ’tweet_id’ 835246439529840640 has rating_denominator 0, as per the field ’text’; the valid rating is 13/10
- Remove rows with missing images, i.e. missing values for ’expanded_urls’ field
- Values in the field ’source’ are not very meaningful. Use regular expression to retain only the required text.
- Following tweets have dogs classified as more than one stage, the ’text’ field has been checked to determine the correct dog stage. Correct these as per the below:

  tweet_id 855851453814013952 - should be puppo

  teet_id 854010172552949760 - should be floofer

  tweet_id 817777686764523521 - should be pupper

  tweet_id 808106460588765185 - multiple dogs - change it to None

  tweet_id 802265048156610565 - multiple dogs - change it to None

  tweet_id 801115127852503040 - should be pupper

  tweet_id 785639753186217984 - should be doggo

  tweet_id 781308096455073793 - multiple dogs - change it to None

  tweet_id 775898661951791106 - multiple dogs - change it to None

  tweet_id 770093767776997377 - multiple dogs - change it to None

  tweet_id 759793422261743616 - multiple dogs - change it to None

  tweet_id 751583847268179968 - should be None

  tweet_id 741067306818797568 - multiple dogs - change it to None

  tweet_id 733109485275860992 - multiple dogs - change it to None
- The fields ’doggo’, ’floofer’, ’pupper’ & ’puppo’ should actually be values under the field ’stage’. Use melt functionality to remove redundant fields.
- Following columns are not required in the final dataset which will contain only the original tweets (replies & retweets and corressponding fields are not required). Drop these fields:

  in_reply_to_status_id

  in_reply_to_user_id

  retweeted_status_id

  retweeted_status_user_id

  expanded_urls
- Retain only the Breed with highest confidence in the image_prediction_df
- Drop Redundant columns in the image prediction dataframe
- Join the different DataFrames ’twitter_archived_enhanced_df’, ’image_predictions_df’ & ’api_data_df’ into a single DataFrame

- Rename the following colums in the final DataFrame ′twitter_archive_master_df′:
  {"tweet_id":"Tweet_Id"
  ,"timestamp":"Timestamp"
  ,"source":"Source"
  ,"text":"Text"
  ,"rating_numerator":"Rating_Numerator"
  ,"rating_denominator":"Rating_Denominator"
  ,"name":"Dog_Name"
  ,"stage":"Dog_Stage"
  ,"retweet_count":"Retweets"
  ,"favorite_count":"Likes"
  ,"jpg_url":"Image_URL"
  ,"img_num":"Image_Number"
  ,"dog_breed":"Dog_Breed"
  ,"confidence":"Breed_Prediction_Confidence"}
- Capitalize the values in the field "Dog_Breed"
- Write master data to csv

0.4 Gathering insights from the Wrangled Data

1. Top 10 Dog Breeds Tweeted (excluding the Unclassified Dog Breeds)

```
Golden_retriever      156
Labrador_retriever    106
Pembroke              94
Chihuahua             90
Pug                   62
Toy_poodle            50
Chow                  48
Samoyed               42
Pomeranian            41
Malamute              33
```

2. Which Dog Stages are most tweeted? (excluding the ones which are not classified)

```
pupper     222
doggo      73
puppo      24
floofer    10
```

3. Dogs identified with highest confidence by the Neural Network Program (excluding dogs without names)

```
Dog_Name    Dog_Breed               Breed_Prediction_Confidence
----------  ----------------------- ----------------------------
Bob         Pug                     0.997445
Sarge       Saint_bernard           0.998830
Ulysses     Schipperke              0.997953
Louis       Pomeranian              0.997210
Olaf        Chow                    0.999837
Panda       Pomeranian              0.997750
Claude      French_bulldog          0.998544
Kloey       Pomeranian              0.998275
Ben         Blenheim_spaniel        0.998335
Ozzy        Pug                     0.999365
Derek       Chow                    0.999823
Kyle        Pug                     0.996952
Oscar       Samoyed                 0.998021
Stanley     Great_pyrenees          0.997692
Bell        Pug                     0.997310
Buddy       Chow                    0.999953
Pete        Old_english_sheepdog    0.999715
Cooper      Dalmatian               0.999828
Hubertson   Pug                     0.999044
Roscoe      French_bulldog          0.999201
```

0.5 Visualizations from the Wrangled Data

Visualizations from Wrangled Data

1. Top 10 Dog Breeds Tweeted (excluding the Unclassified Dog Breeds)
2. Pie Chart Showing the distribution of Dog Stages
3. Popular Dog Stages (excluding the ones which are not classified)
4. Average Retweets and Likes for the top 10 Dog Breeds
5. Retweets for a particular Dog_Stage
6. Visualize Number of Tweets from different Sources
7. Visualize the average number of Retweets/Likes for tweets from different Devices(Sources)
8. Visualize the correlation between the number for Retweets vs Likes