WRANGLE REPORT

ANEESH CHOPRA

GATHERING

We are supposed to acquire our data from 3 sources

1. The WeRateDogs Twitter archive. This file is given to us locally under the name: ‘twitter_archive_enhanced.csv’

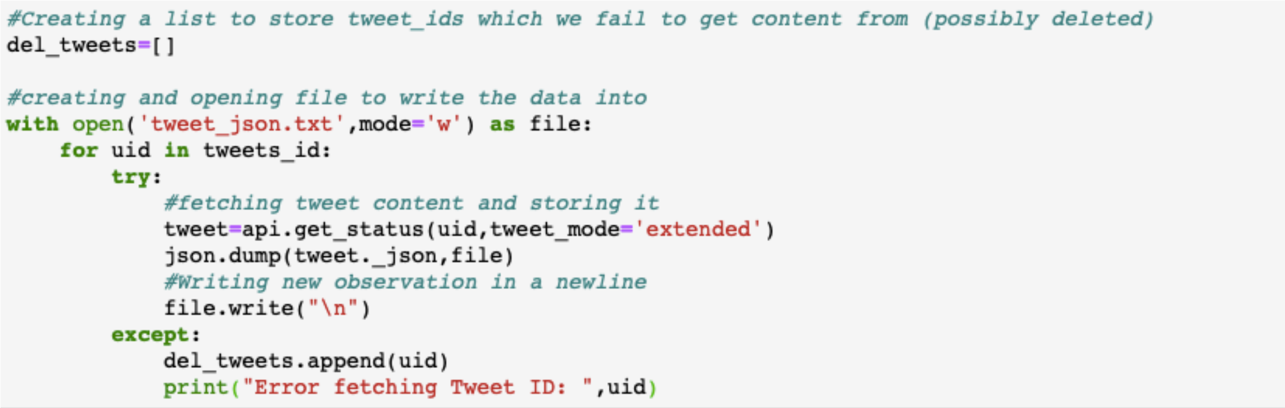
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file is hosted on Udacity's servers and should be downloaded programmatically

3. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's **Tweepy** library and store each tweet's entire set of JSON data in a file called tweet_json.txt

First file was simply read in using the pandas library (df_enhanced)

The Image prediction dataset was read in to csv as well by passing on the URL provided to us. (df_img_pred)

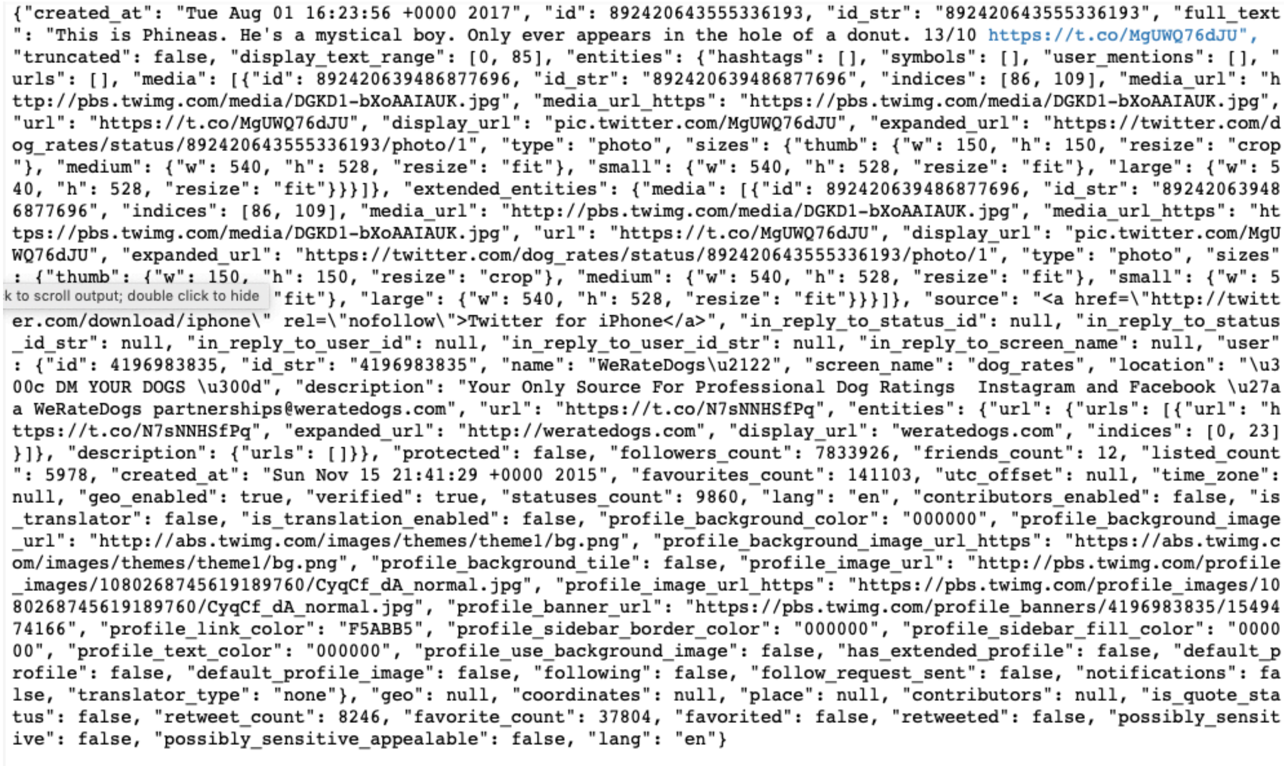
For the third dataset, we had to authenticate ourselves using our API and Consumer key.
Then, we had to retrieve the list of tweet_ids to find from the df_enhanced dataset and parse through them using the api.getstatus() and make a list of tweets which have been deleted



```
#Creating a list to store tweet_ids which we fail to get content from (possibly deleted)
del_tweets=[]

#creating and opening file to write the data into
with open('tweet_json.txt',mode='w') as file:
    for uid in tweets_id:
        try:
            #fetching tweet content and storing it
            tweet=api.get_status(uid,tweet_mode='extended')
            json.dump(tweet._json,file)
            #Writing new observation in a newline
            file.write("\n")
        except:
            del_tweets.append(uid)
            print("Error fetching Tweet ID: ",uid)
```

After getting the JSON data from the api, we had to dump it into a file: tweet_json.txt'. After getting all the data , we printed a sample to understand its sample for understanding how to parse through it



```
{"created_at": "Tue Aug 01 16:23:56 +0000 2017", "id": 892420643555336193, "id_str": "892420643555336193", "full_text
": "This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU",
"truncated": false, "display_text_range": [0, 85], "entities": {"hashtags": [], "symbols": [], "user_mentions": [], "
urls": [], "media": [{"id": 892420639486877696, "id_str": "892420639486877696", "indices": [86, 109], "media_url": "h
ttp://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg", "media_url_https": "https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg",
"url": "https://t.co/MgUWQ76dJU", "display_url": "pic.twitter.com/MgUWQ76dJU", "expanded_url": "https://twitter.com/d
og_rates/status/892420643555336193/photo/1", "type": "photo", "sizes": {"thumb": {"w": 150, "h": 150, "resize": "crop
"}, "medium": {"w": 540, "h": 528, "resize": "fit"}, "small": {"w": 540, "h": 528, "resize": "fit"}, "large": {"w": 5
40, "h": 528, "resize": "fit"}}}]}, "extended_entities": {"media": [{"id": 892420639486877696, "id_str": "89242063948
6877696", "indices": [86, 109], "media_url": "http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg", "media_url_https": "ht
tps://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg", "url": "https://t.co/MgUWQ76dJU", "display_url": "pic.twitter.com/MgU
WQ76dJU", "expanded_url": "https://twitter.com/dog_rates/status/892420643555336193/photo/1", "type": "photo", "sizes"
: {"thumb": {"w": 150, "h": 150, "resize": "crop"}, "medium": {"w": 540, "h": 528, "resize": "fit"}, "small": {"w": 5
k to scroll output; double click to hide "fit"}, "large": {"w": 540, "h": 528, "resize": "fit"}}}]}, "source": "<a href=\"http://twitt
er.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>", "in_reply_to_status_id": null, "in_reply_to_status
_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null, "in_reply_to_screen_name": null, "user"
: {"id": 4196983835, "id_str": "4196983835", "name": "WeRateDogs\u2122", "screen_name": "dog_rates", "location": "\u3
00c DM YOUR DOGS \u300d", "description": "Your Only Source For Professional Dog Ratings  Instagram and Facebook \u27a
a WeRateDogs partnerships@weratedogs.com", "url": "https://t.co/N7sNNHSfPq", "entities": {"url": {"urls": [{"url": "h
ttps://t.co/N7sNNHSfPq", "expanded_url": "http://weratedogs.com", "display_url": "weratedogs.com", "indices": [0, 23]
}]}, "description": {"urls": []}}, "protected": false, "followers_count": 7833926, "friends_count": 12, "listed_count
": 5978, "created_at": "Sun Nov 15 21:41:29 +0000 2015", "favourites_count": 141103, "utc_offset": null, "time_zone":
null, "geo_enabled": true, "verified": true, "statuses_count": 9860, "lang": "en", "contributors_enabled": false, "is
_translator": false, "is_translation_enabled": false, "profile_background_color": "000000", "profile_background_image
_url": "http://abs.twimg.com/images/themes/theme1/bg.png", "profile_background_image_url_https": "https://abs.twimg.c
om/images/themes/theme1/bg.png", "profile_background_tile": false, "profile_image_url": "http://pbs.twimg.com/profile
_images/1080268745619189760/CyqCf_dA_normal.jpg", "profile_image_url_https": "https://pbs.twimg.com/profile_images/10
80268745619189760/CyqCf_dA_normal.jpg", "profile_banner_url": "https://pbs.twimg.com/profile_banners/4196983835/15494
74166", "profile_link_color": "F5ABB5", "profile_sidebar_border_color": "000000", "profile_sidebar_fill_color": "0000
00", "profile_text_color": "000000", "profile_use_background_image": false, "has_extended_profile": false, "default_p
rofile": false, "default_profile_image": false, "following": false, "follow_request_sent": false, "notifications": fa
lse, "translator_type": "none"}, "geo": null, "coordinates": null, "place": null, "contributors": null, "is_quote_sta
tus": false, "retweet_count": 8246, "favorite_count": 37804, "favorited": false, "retweeted": false, "possibly_sensit
ive": false, "possibly_sensitive_appealable": false, "lang": "en"}
```

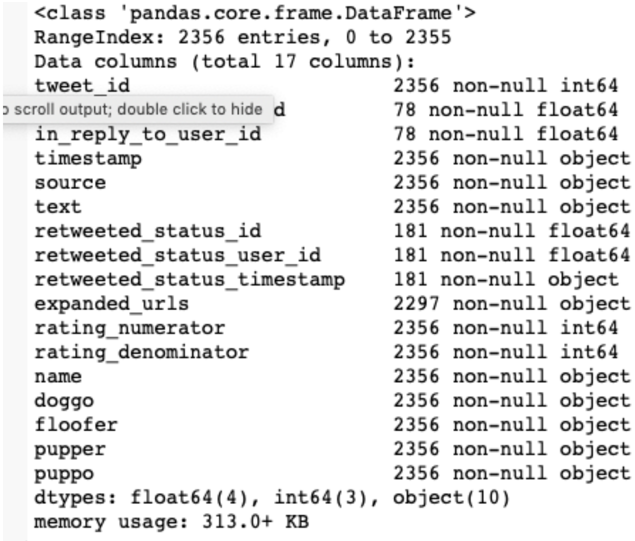
After printing out an example, we realised that every form of data is to be presented as on observation/record

We then open this file and read it into a dataframe record by record. (df_tweet)

Assessing

I. df_enhanced

We first get a gist of our data by printing top 5 records of it and by gaining its datatype and missing information using info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null int64
o scroll output; double click to hide d       78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null int64
rating_denominator            2356 non-null int64
name                          2356 non-null object
doggo                         2356 non-null object
floofer                       2356 non-null object
pupper                        2356 non-null object
puppo                         2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

- Data type wrongly interpreted of many columns(in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, timestamp and retweeted_status_timestamp)
- Retweets are present in the dataset as well
- Some of the expanded URLs are missing
- Missing Values in dog stages columns (doggo,floofer,pupper,puppo) are registered as None

We also noticed a problem in the dog names upon manual visualisation of a sample. We further verify this using value_counts() method.

- Names have been incorrectly entered, such as 'a','such','an' etc.
- upon further analysis we find out Instances where stages of dog hasn't been mentioned
- Some dog breeds are Capitalized, others are in all small case letters
- Types of sources not human readable/clearly understood

II. df_img_pred

1)Some dog breeds are Capitalized, others are in all small case letters

Also,
2)Some objects appear to be not dogs at all, by the image prediction algorithm, which suggests they might not be a dog afterall.

Tidiness Issues

- 3 datasets instead of 1 master dataset

df_enhanced

- Indiviual dogs stages columns need to be converted into a single categorical column
- Unnecessary columns need to be removed

Cleaning Process

Quality

1. Twitter Enhanced Dataset

1.1

columns in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id and retweeted_status_user_id will be converted into string.
Whereas timestamp and retweeted_status_timestamp will be converted into datetime

1.2

Filter out the records which have non-null values in the retweet_status_id

1.3

For places where expanded URL is null, we will add the https://twitter.com/dog_rates/status/ and tweet_id column, and convert into a string to store in.

1.4

Change the missing values in dogs_stage columns "None" in all 4 columns into np.nan

1.5

converting all the names beginning with lowercase letters into np.nan

1.6¶

Extract the text between the anchor tags to extract the actual source used

1.7

We will divide the numerator column by the denominator column and then multiply it by 10 to get the rating out of ten, and then drop the denominator column which will no longer be needed

1. Image Prediction Dataset

- We will Filter out the instances where the algortihm predicts that the image is not of a dog in all 3 instances
- Capitalising every string under columns p1,p2,p3 using .capitalize() function

Tidiness

- While keeping other columns constant, the columns ('doggo' ,'floofer', 'pupper', 'puppo') will be melted into one column. We have to keep in mind that one dog can be in more than one stage since the terms are vaguely defined

- Merging all tables using tweet_id as the primary key and by inner join, so we have a tight dataset of tweets for which we have all data available
- Removing Redundant Columns