WeRateDogs Twitter Feed

Extracting dog data and predictions from tweets

Background
This project took on the challenge of processing unstructured tweet text and images from a popular dog followers Twitter feed: WeRateDogs.

When parsing and analysing the feed data, we are looking for the following information:

- Dog name
- Dog stage (doggo, floofer, pupper, puppo)
- A rating for the dog out of 10, typically the numerator is bigger than 10 (e.g.; 13/10)!
- Up to 4 dog images, used to predict the dog's breed
- The number of retweets and likes given to each tweet
- The top 3 dog breed predictions, including the confidence factor

Data Enrichment
The core tweet data set we are provided with is enriched with tweet data that it is missing (retweet and like counts). To do this, we use the Tweepy library to integrate with Twitter's API and retrieve the original tweet data, then extract the missing fields.

Image Data

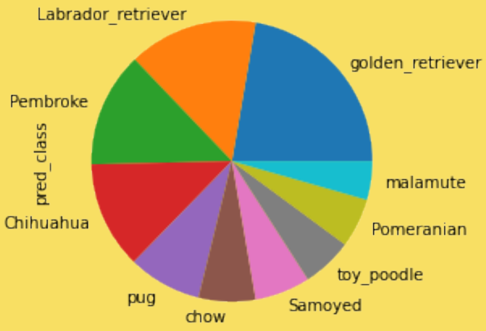
The image data set holds the top 3 dog breed prediction from an image classifier used to predict dog breed. It is worth noting that this classifier must be trained on a set of images that includes things other than dogs, and on occasions may generate prediction label that don't describe dogs. There is a boolean flag in each prediction to highlight this scenario. And the data set links to the image for the most confident prediction.

Visualisations

Below you can see a couple of the visualisations produced, showing the top 10 most popular dog breeds, first based on tweet count, and second based on the favourite counts against the relevant tweets.

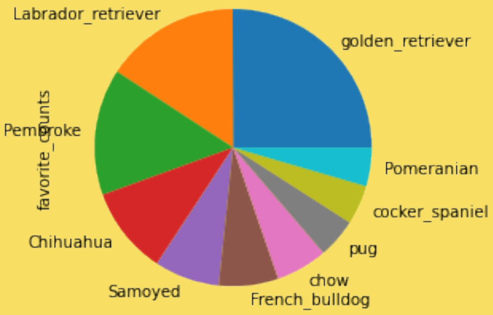
Distribution by number of tweets

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f87295fa3d0>
```



Distribution by number of favorites

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f87295fad10>
```



You can see that whilst they are similar at the top end of the ranking, dominated by retrievers, the mix of less dominant breeds changes significantly between the two.

Challenges

Parsing unstructured data is always challenging, and some of the scores, dog names or stages and dog images may be misinterpreted or misclassified. On occasions multiple dog stages are parsed, which may or may not make sense in the context of the tweet text. And some images are classified not as dogs!

Closing remarks

The objective of this project was to wrangle the tweet feed data, and deliver a clean data version suitable for further processing or analysis. This includes applying transformations to meet the Tidy Data principles. We believe the submitted notebook achieves these objectives, and have shown a couple of sample visualisations to demonstrate how the data may be used.