Coleta de dados

Para este projeto em questão foram necessários usar 3 fontes de dados, cada um sendo adquiridos de uma forma distinta.

- twitter-archive-enhanced.csv, fornecido pelo projeto, baixado na Udacity;
- image-predictions.tsv, baixado via programação de um servidor na nuvem;
- tweet_json.txt, baixado via api oficial da Twitch, tweepy;

Avaliando dados

Os dados das fontes de dados foram documentados e classificados de acordo com dois critérios, problemas de qualidade e falta de organização.

Problemas de qualidade

twitter-archive-enhanced.csv

1. coluna tweet_id formato int64, alterado para string;
2. coluna rating_numerator com valores maiores que 15, foram conferidos manualmente (acessando o post em questão) e corrigidos por programação;
3. coluna rating_denominator com valores diferentes de 10, foram conferidos manualmente (acessando o post em questão) e corrigidos por programação;
4. coluna name com nomes de cãe representado erroneamente por “None”, “a” e “the”, esses dados foram corrigidos para dados vazio (none), não é contado no comando value_counts();
5. colunas in_reply_to_status_id e in_reply_to_user_id, quase todos valores nulos, portanto foram retiradas;
6. colunas retweeted_status_id e retweeted_status_user_id apresentava quase todos os valores nulos, porém os valores não nulos indicavam quais eras retweets e como o critério era eliminar os retweets foram excluídos as linhas de dados que continham esses valores não nulos, ao final as duas colunas foram retiradas por não agregar mais informações;

image-predictions.tsv

1. total de linhas diferente dos arquivos twitter-archive-enhanced.csv, indicando problemas com falta dados, corrigido através da função merge onde uniu as 3 fontes de dados e o critério foi existir a mesma tweet_id nas 3 fonte de dados, assim o total de linhas após a união das fonte de dados foi o menor valor entre as 3 fonte de dados;
2. coluna tweet_id formato int64, alterado para string;
3. colunas p1, p2 e p3 apresentavam palavras que inicia com minúscula e outras com maiúscula, adotou de transformar todas as palavras em minúscula, seguindo as normas da língua portuguesa, a correção foi feita na coluna breed que foi originada da reorganização das colunas p1, p2 e p3;

tweet_json.txt

1. coluna id formato int64, alterado para string;

2. valor 0 em favorite_count, indicava falta de dados e essas linhas foram removidas automaticamente durante o processo de juntar (merge) as 3 fonte de dados;
3. valor 0 em retweet_count, indicava falta de dados e essas linhas foram removidas automaticamente durante o processo de juntar (merge) as 3 fonte de dados;

Problemas de arrumação

twitter-archive-enhanced.csv

1. As seguintes colunas doggo, floofer, pupper, puppo foram reorganizadas em uma só coluna chamada de stage;
2. Coluna timestamp apresentava na mesma célula data e hora, como hora não era necessária ela foi eliminada e a coluna foi renomeada para date;

image-predictions.tsv

1. A coluna breed apresenta a raça mais provável, foi criada a partir do critério de o maior valor dentre as colunas p1_conf, p2_conf e p3_conf determinava a raça mais provável indicada pela coluna p1, p2 ou p3;