Projeto WeRateDogs

Documento Interno

Wrangle Report

Este documento descreve de forma breve os esforços realizados para a atividade de *wrangling* no projeto WeRateDogs. O arquivo inicial disponibilizado, identificado como (1) twitter_archive_enhanced.csv, continha muitos ajustes a serem realizados. Nesta etapa foi realizada parcialmente a limpeza, visto que alguns problemas identificados necessitam de mais tempo para codificação e teste. É necessário um futuro esforço para ajuste e correção dos nomes dos cães e do rating, além de uma nova análise exploratória com o intuito de verificar outras possíveis pendências.

Teve-se acesso pleno as outras duas fontes de dados: o arquivo (2) image_predictions.tsv, originário do servidor da Udacity via download por url previamente identificada; (3) dados do Tweeter obtidos via API pelo aplicativo Tweepy. No processo de importação via API, o sistema relatou 13 erros.

Os três conjuntos de dados puderam ser acessados e visualizados em sua plenitude. O arquivo twitter_archive possuía inicialmente 2.356 elementos, o image_predictions possuía 2.075, e os dados oriundos via API, armazenados no Data Frame tweet_df, totalizaram 2.343. Observa-se, assim, uma discrepância entre as fontes de dados.

A fusão da massa de dados inicial com o arquivo adicional image_predictions.tsv e com os dados extraídos do Tweeter via API foi realizada com sucesso. Passou-se, portanto, a ter um conjunto de dados muito mais rico do que o incialmente fornecido.

A análise exploratória encontrou problemas de qualidade e de estrutura. Dentre os problemas de qualidade, tem-se: colunas com valores NaN, coluna ‘name’ com valores estranhos, colunas com formatos incorretos, necessidade de remoção de retweets, necessidade de remoção de tweets sem imagem, e nomes de cães incorretos.

O processo de limpeza foi iniciado com a cópia dos arquivos originais e a fusão dos 3 arquivos em único dataframe. Buscou-se executar a limpeza das atividades relacionadas previamente. As ações de limpeza foram identificadas cada uma com 3 etapas: definição (define), codificação (code) e teste (test).

Na análise dos aspectos estruturais:

- Verificou-se que os estágios (classificação) dos cães estavam em 4 colunas ao invés de apenas uma.
- Foi criada uma única coluna no dataframe que armazenou os estágios dos cães em formato categórico.
- Um novo procedimento foi executado para identificar os numeradores e denominadores do rating dos cães e a construção de um novo coeficiente para o rating.

- Colunas desnecessárias foram apagadas do dataframe.
- Algumas colunas precisaram ter seus formatos estabelecidos de forma adequada.
- Tweets que não possuíam imagem foram removidos.
- Colunas identificadas com retweets foram removidas.

Por fim, análises sobre o rating e a contagem de favoritos foram realizadas, incluindo visualizações gráficas.