Fundamentos de Data Science II

Projeto 2 – Data Wrangling

Evandro Silva

Wrangle Report

Sumário

Introdução

Nesse relatório vou descrever os passos e os métodos para fazer o projeto Wrangle and Analyze Data com foco no twitter WeRateDogs, utilizando Python e Bibliotecas especificas para:

- Coletar os dados
- Avaliar os dados
- Limpar e Arrumar os dados
- Armazenar os (aquivo ".CSV")

Este projeto tem como base 3 arquivos, dois fornecidos através de link disponibilizados e um que deve ser montado através de API que faz leitura de contas no Twitter (mais detalhes serão vistos nos passos carga dos de cada arquivo):

- twitter_archive_enhanced.csv
  Este arquivo contém dados básicos de tweets (WeRateDogs) para os tweets em 1º de agosto de 2017.

- image_predictions.tsv
  As previsões de imagens em tweets, é uma previsão de qual raça de cachorro está presente em cada tweet.

- tweet_json.txt
  A contagem de re-tweets e favoritos de cada tweet.

Coletar os dados

A coleta de dados está dividida em subetapas:

1. Carregar o arquivo **twitter_archive_enhanced.csv**, após download manual. O arquivo está disponível no endereço **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv**.

2. O arquivo **image_predictions.tsv** está hospedado nos servidores da Udacity e deve ser baixado programaticamente e está localizado na seguinte URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**

3. Fazer uso de uma API de consulta no Twitter (usando as ID de tweets do arquivo **twitter_archive_enhanced.csv**) para cada ID tweet usar a biblioteca Tweepy do Python, coletar os dados e armazenar cada retorno de dados em uma linha no arquivo **tweet_json.txt**.

Avaliando os dados

Durante a coleta de dados já é possível fazer algumas validações visuais de problemas nos dados. Porém, neste ponto é que vamos aprofundar nossa avalição para identificar/constatar erros nos dados. Que podem ser de qualidade, integridade, validade e arrumação. De posse destas informações, podemos partir para a limpar e arrumar os dados.

- A coluna "name" possui alguns registos com valores: 'a', 'an' e 'the'.
- As colunas doggo,floofer,pupper,puppo(dog_stage) tem 1976 linhas como valor None em topas as colunas.
- As colunas doggo,floofer,pupper,puppo(dog_stage) tem 14 linhas onde mais de um dog_stage foi atribuido.
- As colunas in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, deveriam ser int64.
- Muitos valores nulos para as colunas in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id e retweeted_status_timestamp.
- A coluna status_timestamp, deveria ser datetime.
- Existem menos registro que a tabela df_twitter_arc, eliminar registros sem imagem e retweets.
- 324 imagens não foram identificadas como cachorro
- A mesma imagem sendo utilizada para tweet_id's diferentes (66 ocorrências)

Limpar e Arrumar os dados

Esta é a última etapa do Data Wrangler.
Os dados dos arquivos **twitter_archive_enhanced.csv**, **image-predictions.tsv** e **tweet_json.txt**, foram agrupados em um Dataframe.
Uma cópia do dataframe(unifcado) foi feita para aplicarmos os processos de Limpeza e Arrumação.

Caso tenhamos algum problema com a execução de alguma etapa de limpeza/arrumação, podemos voltar aos dados originais para refazer os passos.

Armazenar os Dados (arquivo “.CSV”)

Agora, vamos gerar o arquivo twitter_archive_master.csv a fim de deixarmos armazenado as informações para possibilitar as análises numéricas e gráficas das informações após o Wrangling.

Conclusão

Para quem trabalha com análise de dados e/ou ciência de dados, o processo de Wrangling é um dos passos mais importantes.

Com este projeto, isso ficou muito evidente e será uma das etapas que vou dedicar cada vez mais atenção para que minhas atividades no trabalho sempre tenham a melhor qualidade possível para o conjunto de dados a analisar.

Sem falar, que análise gráfica dos dados após o processo, fica mais intuitiva e assertiva.