Wrangle Report

Download dos dados

Na parte do recolhimento dos dados, foi necessário importar todas as bibliotecas para essa função, sendo elas: **pandas**, **requests**, **tweepy**, **json**, **os** e a **tqdm**.

A **tqdm** não tem uso específico na coleta ou manipulação dos dados, ela se trata apenas de uma interface para criar a barra de progressão durante uma determinada tarefa. Seu uso fica mais claro a frente.

Com **pandas** foi lido o *twitter_archive_enchanced.csv* com informações básicas sobre os tweets.

Utilizando as bibliotecas de **requests** e **os**, é baixado o *TSV* fornecido e salvado como *image_predictions.tsv*, no qual contém previsões de qual raça de cachorro ou objeto inanimado esteja no tweet.

Com a API do Twitter (**tweepy**), foi feita a raspagem dos dados para recolher todas informações necessárias dos tweets fornecidos através do *CSV twitter_archive_enchanced.csv* e armazenagem no arquivo *tweet_json.txt* com a biblioteca **json**. Esse processo demorou aproximadamente 35 minutos, logo, para que fosse possível acompanhar da melhor forma sua progressão, foi utilizado a biblioteca **tqdm**. Os erros foram gerados e armazenados no *CSV errors.csv*.

Carregamento dos dados

A etapa de carregamento dos dados é sucinta, graças à coleta bem-sucedida.

Foi realizado a importação das bibliotecas **pandas**, **numpy**, **json** e **os**. Algumas foram repetidas para que seja possível realizar a execução dessa parte de forma independente.

São criados os DataFrames do **pandas** através dos arquivos disponíveis localmente, com um adendo ao **json**, pois é realizado filtros adicionais no momento da leitura, deixando somente as informações julgadas necessárias.

Avaliação dos dados

O momento da avaliação dos dados é extremamente importante, para que possam ser verificados todos os problemas dos dados. Evitando possíveis retrabalhos. Ainda assim não é incomum esse passo ser realizado diversas vezes durante todo o processo de Data Wrangling.

Nesse processo foram realizadas uma série de visualizações a partir dos DataFrames criados na etapa anterior, visando entender como os dados se relacionam e identificar os pontos de ajustes.

Com isso foram identificados 10 desvios de qualidade e 7 de arrumação que serão corrigidos no próximo passo.

Limpeza

Hora de transformar os pontos identificados em códigos para criar um ou mais DataFrames limpos e ideais para analises.

Por questões de segurança, foram criadas cópias dos DataFrames originais antes de qualquer limpeza. Durante todo o processo só será modificado as cópias.

Nessa etapa são realizadas diversas alterações para chegar no cenário ideal para análise. Algumas dessas são demonstradas abaixo:

- Foram retirados os tweets considerados inúteis, duplicados ou inadequados para a análise. Isso é realizado de forma massiva e pontual, dependendo da quantidade de reincidências verificadas.
- No DataFrame dos tweets, temos informações quantitativas importantes, como as notas, mas para que possamos utilizá-las de maneira confiável, é preciso normalizar seus valores, deixando todos com denominador 10 e removendo notas erradas.
- Foi alterado a "forma" dos DataFrames, os tornando mais sucintos para análises, sendo necessário juntar colunas e deletar outras. Também é feito a atribuição dos tipos de dados ideais para as respectivas colunas.
- Sobre as predições, foi considerado inútil qualquer predição que não tenha identificado um cachorro. Para múltiplas raças identificadas, foi mantida apenas a com mais alta confiabilidade. Com isso foi gerado uma coluna única, contendo apenas as raças de cachorros mais confiáveis.

Finalmente, é gerado um único *CSV* chamado *twitter_archive_master.csv*.

Análise dos dados

Sendo a etapa final, a análise dos dados é normalmente mais prática, uma vez que os dados já estão limpos e você os conhece bem.

No final de tudo, foram gerados alguns insights, tanto em forma de textos como de gráficos. Isso com ajuda de bibliotecas como **matplotlib**, **seaborn** e **pandas**.