INTRODUÇÃO À CIÊNCIA DE DADOS

Arquivos

CARGA DE DADOS

- Projetos de DS precisam de dados!
- Grande parte da tarefa de um DS é carregar, limpar, organizar e transformar dados.
- Muitas vezes esses dados vêm em formas variadas de arquivos e precisam ser carregados para dentro do projeto.
- Também é fato que dados analisados e tratados, em algumas situações precisam ser devolvidos para o contexto externo do projeto, nos mais variados formatos.
- Nesta aula vamos focar no processo de carga de dados, tanto de arquivos texto como de arquivos em formatos estruturados.

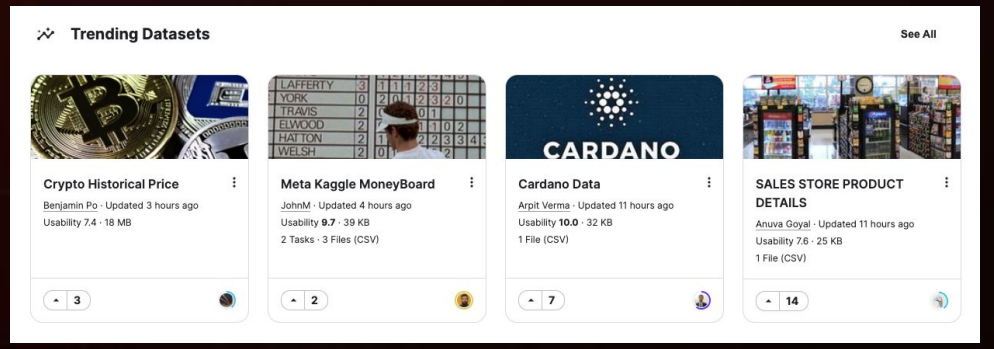
TIPOS DE ARQUIVOS

- Formatos comuns como CSV, XLS (Excel), XML, Json e os próprios arquivos em formato de texto são bem comuns para programadores de um modo geral.
- O contexto de DS e BigData apresenta alguns outros formatos que não são tão comuns para quem está iniciando no mundo do DS.
- Entre esses formatos, podemos encontrar: arquivos binários MessagePack, pickle, dados do sistema SAS, arquivos HDF, dados binários Feather, arquivos STATA, entre outros.
- Muitas vezes não há como escolher o tipo de dado que o projeto precisará consumir.

PANDAS

- A biblioteca Pandas é uma grande aliada e oferece muitas funcionalidades para lidar com arquivos.
- Há funções prontas já no Pandas para realizar o consumo de dados, veja algumas opções que podem ser utilizadas e facilitam muito o processo de carga de dados:
  - read_csv: lê dados que utilizam vírgula como delimitador, o arquivo pode vir de um arquivo ou de um endereço URL.
  - read_excel: lê dados tabulares de um arquivo Excel XLS ou XLSX
  - read_html: lê as tabelas que estão em um arquivo HTML especificado.
  - read_json: lê dados de uma representação em string JSON (Java Script Object Notation)
- Há funções para vários outros tipos (veja na documentação do Pandas)

FONTES DE DADOS



- Há uma infinidade de fontes de dados públicas e disponíveis para serem utilizadas. Há dados verdadeiros e dados fabricados.
- Aproveite e use para estudar e criar seus primeiros projetos:
  - Kaggle
    - https://www.kaggle.com/datasets
  - Dados Governo Brasileiro
    - https://dados.gov.br/dataset
  - Instituto Johns Hopkins
    - https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data
  - UCI Machine Learning
    - https://archive.ics.uci.edu/ml/datasets.php
  - Datasets no GitHub (muito bom!)
    - https://github.com/awesomedata/awesome-public-datasets

FONTES DE DADOS

- Muitas vezes a fonte de dados pode ser uma API.
- Na nossa demonstração, vamos usar uma api de teste, acessando um endereço, ela gera um dado aleatório de uma pessoa!
  - https://randomuser.me/api
  - Resultado em Json!
  - Observaremos isso na demonstração

GRAVAR DADOS / DEMONSTRAÇÃO

- Assim como é feita a carga de dados, oriundos do mundo externo, também é necessário gravar dados muitas vezes.
- Há funções para realizar essas gravações, incluindo as opções de gravação de dados que a própria biblioteca Pandas faz (assim como faz também para ler dados).
- Vamos fazer uma demonstração tanto de gravação, quanto de leitura de dados.

FINALIZANDO

- Coletar e armazenar dados em arquivos é uma tarefa do cotidiano de quem trabalha com DS.
- Dados em formatos TXT, CSV e oriundos de muitos outros formatos farão sempre parte da vida de um cientista de dados.
- É muito importante conhecer esses tipos de formatos e compreender como é possível lidar com eles nos mais variados momentos de um projeto de DS.

INTRODUÇÃO À CIÊNCIA DE DADOS

Arquivos