Gmail Rafael Rodrigues Marquesi <rafael.4marcos.nt@gmail.com>

#7DaysOfCode - Ciência de Dados 1/7: Data Cleaning and Preparation

Paulo Vasconcellos <paulo.vasconcellos@7daysofcode.io> Fri, Apr 22, 2022 at 7:05 AM
Reply-To: reply-caelum57945.activehosted.1373.8447.1043955@d32a.emsend3.com
To: Rafael Rodrigues Marquesi <rafael.4marcos.nt@gmail.com>

Olá, Rafael Rodrigues Marquesi.

Chegou o grande momento!

Você vai passar os próximos 7 dias praticando Dados comigo.

Antes de começar, me responda uma pergunta: **com o que você acha que Data Scientists gastam mais tempo durante seu trabalho?**

Eu queria poder mentir pra você e dizer que passamos a maior parte do nosso tempo criando modelos altamente complexos e brincando com o que há de mais avançado tecnologicamente.

Eu queria, ainda, poder dizer que as bases de dados que você irá trabalhar já estão limpas, bem processadas e prontas para que você possa criar visualizações e análises poderosas.

Mas, na verdade, não é bem isso. O que passamos a maior parte do tempo fazendo é a preparação do dado, **mais especificamente nas partes de limpeza e transformação**.

Em uma pesquisa com cerca de 80 cientistas de dados, as

etapas de preparação de dados foram responsáveis por consumir mais de 80% do tempo deles, reforçando a **necessidade de desenvolver a habilidade de tratar dados** e deixá-los prontos para a parte mais divertida: criar modelos e análises.

Por isso, quero te dar um ótimo desafio para desenvolver e aprimorar essa skill tão importante para toda pessoa cientista de dados.

Neste primeiro dia, **eu te proponho baixar um dataset do portal do CEAPS** (Cota para Exercício da Atividade Parlamentar dos Senadores) **e aplicar processos de tratamento e limpeza de dados nele** (processo conhecido como Data Wrangling).

Basicamente, **o CEAPS contém todos os gastos que senadores brasileiros declararam, divididos por ano**.

Esse tipo de dado é tão importante que já criou várias iniciativas interessantes, como a Operação Serenata de Amor, que aplica Inteligência Artificial para analisar gastos de deputados brasileiros, e que já foi capaz de identificar vários usos indevidos do dinheiro público.

Imagina as coisas que você pode criar!

Os dados do CEAPS contêm uma série de problemas que podem dificultar a criação de análises mais aprofundadas.

Uma das primeiras coisas que você pode fazer é identificar tais inconsistências, como campos que possuem valores nulos ou duplicados, converter campos de data que estão sendo carregados como texto, corrigir valores monetários, nomes incorretos, formatar campos de CNPJ, etc.

DICA

Os dados do CEAPS estão divididos por ano.

Que tal juntar dados de vários anos em um grande dataset e aplicar técnicas de limpeza e processamento dos dados?

Você poderia pegar dados dos últimos quatro anos e aplicar o que usou nesse exercício.

Além disso, também proponho que você **documente o seu processo de tratamento dos dados** (pode ser no próprio Jupyter Notebook). Assim, qualquer pessoa que consumir o seu trabalho saberá qual foi sua intuição e as técnicas utilizadas, além de facilitar a reprodutibilidade.

Tire o maior proveito possível dos comentários no código!

Não existe uma receita de bolo para quais técnicas utilizar na limpeza de dados, isso varia de projeto para projeto. Contudo, existem algumas coisas que você pode fazer inicialmente.

Lidar com dados nulos (deletar ou imputar um valor novo, por exemplo); remover colunas que não trazem nenhuma informação; processar datas que estão em formato incorreto; alterar o tipo da coluna (uma coluna que é numérica está como texto no Pandas); remover duplicados; dentre outras.

Esse post da Tableau traz algumas dicas sobre técnicas que você pode utilizar.

EXTRA

O Portal da Transparência é um ótimo site para **encontrar novos datasets para explorar e criar projetos**. Esse é o site que contém informações detalhadas sobre a execução orçamentária e financeira da União.

Devido a cada sistema ter sua forma de exportação dos dados, é possível identificar uma série de oportunidades de melhora nos dados exportados. Que tal expandir o seu trabalho explorando outros datasets?

Ou, ainda: e se você hospedar esse seu dataset limpo em uma plataforma que outras pessoas possam utilizar?

Existem várias plataformas legais para isso, desde o Kaggle, onde você pode submeter datasets para a comunidade, até iniciativas brasileiras de dados abertos, como o Brasil IO.

Bom trabalho!

Paulo Vasconcellos
Cientista de Dados **e Alura Star**

---

alura