



Big Data Real-Time Analytics Com Python e Spark 3.0

Big Data Real-Time Analytics Com Python e Spark Versão 3.0

Estudo de Caso 1 Limpeza e Pré-Processamento de Dados com NumPy



Formação Cientista de Dados 3.0

Big Data Real-Time Analytics com Python e Spark

Estudo de Caso 1

Limpeza e Pré-Processamento de Dados com NumPy

Imagine que em um determinado projeto de Ciência de Dados você receba um dataset extremamente complicado, contendo dados com muitas strings, caracteres especiais, problemas de encoding, datas mal formatadas, números e textos na mesma coluna, url's contendo Ids importantes para análise, valores ausentes, coluna que contém informação que deveria estar distribuída em três ou mais colunas. E como se não bastasse tudo isso, parte dos dados necessários para análise está em outro dataset, que deve ser combinado com o primeiro.

Seu trabalho seria limpar e pré-processar esse dataset, preparando-o para a sequência do processo de análise.

É exatamente este cenário que estamos reproduzindo no Estudo de Caso 1. A partir de dados complexos e com diversos problemas, iremos fazer um extenso trabalho de limpeza e pré-processamento. E tudo isso usando apenas o NumPy, poderoso pacote da Linguagem Python para computação e processamento de dados.

Este Estudo de Caso traz uma quantidade incrível de conhecimento sobre manipulação de dados em Python. Aproveite!