Aula 6: Limpeza e Processamento de Dados Coletados

Eduardo Ribeiro Silva de Oliveira

07 de Outubro de 2024

Estratégias de Limpeza de Dados

- Técnicas para remover duplicatas e tratar caracteres especiais.
- A limpeza de dados é essencial para garantir a qualidade das análises.
- Dados sujos podem levar a conclusões incorretas, tornando a limpeza um passo crucial.

Remoção de Entradas Duplicadas

- Identificar e remover entradas duplicadas para evitar redundância.
- Entradas duplicadas podem distorcer resultados e análises.
- Utilizar ferramentas como Pandas para identificar e remover duplicatas de forma eficiente.

Tratamento de Caracteres Especiais

- Tratar caracteres especiais e problemas de codificação.
- Corrigir problemas que dificultam o processamento dos dados.
- Caracteres especiais podem surgir devido a diferentes fontes de dados e codificações, como UTF-8.

Lidar com Valores Ausentes

- Lidar com valores ausentes (missing values).
- Decidir entre descartar valores ausentes ou imputá-los.
- Técnicas de imputação incluem usar a média, mediana ou valores mais frequentes para preencher as lacunas.

Normalização e Padronização dos Dados

- Como garantir que os dados estejam prontos para análise.
- Padronizar formatos e nomenclaturas para evitar inconsistências.
- A padronização facilita a comparação entre diferentes conjuntos de dados e garante que todos os dados sejam interpretados da mesma forma.

Padronização de Formatos

- Converter todos os dados para um formato padronizado.
- Exemplos: datas no mesmo formato e valores monetários na mesma moeda.
- Padronizar formatos evita erros durante a análise e facilita a agregação de dados de múltiplas fontes.

Normalização para Análise Estatística

- Normalizar os dados para facilitar comparações e análises estatísticas.
- Importante para garantir que diferentes variáveis estejam na mesma escala.
- A normalização é particularmente útil para algoritmos de aprendizado de máquina, que podem ser sensíveis a escalas diferentes.

Uso de DataFrames para Organização

- Introdução ao uso de bibliotecas como Pandas.
- DataFrames ajudam na manipulação e organização dos dados.
- DataFrames são estruturas de dados altamente eficientes e permitem operações como filtragem, agregação e transformação de dados de forma prática.

Armazenamento em Estruturas como CSV

- Como armazenar os dados em arquivos CSV.
- CSV é um formato amplamente utilizado para facilitar o acesso e análise.
- CSV é legível tanto por humanos quanto por máquinas e é compatível com muitas ferramentas de análise de dados.

Organização dos Dados para um Fluxo de Trabalho Eficiente

- Organização dos dados em bases de dados para um fluxo de trabalho mais eficiente.
- A boa organização dos dados facilita análises e reduz erros.
- Armazenar os dados em bases de dados relacionais ou NoSQL permite escalabilidade e acesso eficiente a grandes volumes de informações.