Podcast

Disciplina: Linguagens de programação para ciência de dados
Título do tema: Manipulando dados com python
Autoria: Yuri Vasconcelos de Almeida Sá
Leitura crítica: Henrique Salustiano Silva

Abertura:

Olá! No podcast de hoje vamos falar sobre o primeiro passo na criação e desenvolvimento de um modelo ciência de dados em Python: a aquisição dos dados.

É a partir da importação dos dados brutos que se dá início ao processo de criação de um produto de dados, seja para avaliar métricas estatísticas e definir assim os caminhos para atingir os objetivos do modelo, seja para adequar os dados de entrada para produção.

Esta fase é a responsável por fazer o tratamento inicial dos dados, observando valores fora da curva e tratando impurezas que possam atrapalhar o fluxo de trabalho.

Durante esse processo é importante que dados nulos ou com tipos incompatíveis devem ser avaliados e só então convertidos ou removidos, conforme regra. Dentre os exemplos clássicos temos de encontrar *strings* onde deveria haver números, números fora da escala, remoção de duplicados, má formação do registro etc.

Esta etapa não deve se limitar somente a qualidade objetiva dos registros, mas deve também atender as particularidades do modelo. Essa afirmação se torna mais clara quando pensamos em modelos de séries temporais por exemplo, onde os dados devem ser contínuos e em ordem onde a falta ou incongruência dos dados trazem prejuízos à análise. Ainda sobre séries temporais, o agrupamento por período de análise é executado logo após a primeira validação de qualidade, podendo inclusive conter múltiplos agrupamentos em períodos diferentes, semanais, mensais e semestrais, por exemplo.

Os critérios gerais para avaliação da qualidade dos dados são: validade, completude, consistência e uniformidade. Outros critérios podem ser avaliados e até mesmo criados, conforme necessidade.

Este processo forma a base para o desenvolvimento e utilização do produto de dados e está presente durante todo o ciclo de vida do projeto.

Também é importante a redução de dados para eficiência do projeto, ou seja, manter somente os dados que serão utilizados sem que haja desperdício de recursos, principalmente memória, que é o item de hardware mais caro da operação. Mesmo poucos bytes por registro podem se tornar um grande custo computacional durante o treinamento e execução de um modelo quando ele for escalado.

Muitas vezes pensamos que a aquisição e importação de dados ocorre com arquivos tabulares ou banco de dados, mas a própria natureza dos sistemas

prevê inúmeras fontes de dados, um só modelo pode adquirir dados de arquivos, bancos de dados e interfaces REST, por exemplo. E todos esses dados devem ser organizados, selecionados, limpos, relacionados e consolidados.

A própria estrutura dos dados para análise é fluida, podendo conter atributos e outros tipos de metadados que enriquecem e orientam o projeto. Ainda sobre metadados, podemos criar nossos próprios, contando caracteres de alguma *string*, por exemplo, e fazer disso uma dimensão do modelo.

Não é raro que existam desenvolvedores especializados em aquisição de dados, é uma área rica e essencial!

Os próximos passos depois da aquisição inicial dependem do tipo de técnica e modelo a serem utilizados, mas geralmente envolvem transformações, normalizações e reduções, que incluem uma grande dose de estudo do objetivo do produto.

Fica claro então que uma aquisição de dados bem feita melhora e facilita todo o processo.

Fechamento:

Este foi nosso podcast de hoje! Até a próxima!