Checklist de Limpeza de Dados

Bem-vindo à Era da Inteligência Artificial (IA), em que a maneira como você faz negócios depende de tecnologias que consomem muitos dados, como machine learning e deep learning. Para tirar proveito dessas novas ferramentas de IA, você precisa verificar se a "casa" dos dados da sua organização precisa de uma faxina.

Confira o checklist para você dar os primeiros passos na limpeza do seu armazenamento de dados. Ela se divide em duas fases principais: treinamento e inferência.

Siga estas etapas para se tornar mestre em IA. Para mais informações sobre como levar a IA da prova de conceito à produção e escala total, confira este relatório da IDC, *Acelere e operacionalize implantações de IA usando infraestrutura otimizada para IA.*

Treinamento

No treinamento de preparação para a IA, você desenvolverá algoritmos para entender um conjunto de dados. Sua principal preocupação será coletar dados existentes e usar a IA para aprender um novo recurso.

- [ ] Descubra o problema de negócio específico que você quer resolver usando a IA (comece com projetos menores para aprender)
- [ ] Localize os dados que podem resolver esse problema usando fontes relevantes (provavelmente não estarão todos localizados no mesmo lugar)
- [ ] Prepare seus dados com tags de metadados para reduzir significativamente o tempo necessário para encontrar dados pertinentes
- [ ] Verifique se seus dados estão sincronizados e vinculados adequadamente em todos os conjuntos de dados que você usará (inclusive a sincronização de horários)
- [ ] Sinalize dados particulares e confidenciais do cliente para garantir que você os mantenha absolutamente seguros e cumpra toda a governança e regulamentação apropriadas (o processo de marcação de metadados pode ajudar nisso)
- [ ] Escolha o ambiente de desenvolvimento certo para o tipo de dados que você está usando e a forma como eles serão formatados (ou seja, imagens, vídeo, texto e áudio em formato livre, cada um geralmente tem um tipo de ambiente)
- [ ] Puxe conjuntos de dados do seu repositório e leve-os ao seu ambiente de desenvolvimento
- [ ] Divida seus dados em dois grupos para ajudar a melhorar o processo de desenvolvimento do modelo (mantenha um conjunto em uma pasta chamada "treino" e outro em uma chamada "teste")
- [ ] Mantenha a rastreabilidade dos dados acompanhando de onde eles vieram (considere o uso de ferramentas que podem ajudar a automatizar o processo)
- [ ] Execute tarefas básicas de higiene de dados para prepará-los para a construção de um modelo (por exemplo, inclua o preenchimento de entradas de dados ausentes e a remoção de entradas nulas)
- [ ] Use uma amostra de subconjunto de dados para a qual você já conhece a resposta da atividade de previsão (isso é chamado de "conjunto de treinamento") e identifique todas as etapas de pré-processamento necessárias para preparar os dados para fazer uma previsão
- [ ] Use seu conhecimento desse conjunto de treinamento para calcular pontuações de precisão que podem dar a você a confiança necessária para aplicar o mesmo modelo a novos dados para os quais ele nunca foi explicitamente treinado

Inferência

Depois de desenvolver um modelo que funciona para resolver seu problema comercial, você passa do treinamento para a inferência. Nessa fase, você adota esse modelo bem-sucedido e o aplica a novos dados, o que exige algumas tarefas de limpeza em andamento também.

- [ ] Localize seu modelo de IA próximo aos seus dados para diminuir a latência, reduzir os requisitos de largura de banda e melhorar a performance geral do modelo
- [ ] Desenvolva um processo eficiente de pipeline de dados e aplique a rotulagem de metadados aos dados à medida que eles entram, para que novos dados possam ser coletados e usados para aprimorar o modelo dali em diante
- [ ] Marque os dados de uma maneira que seja vinculada e sincronizada, por exemplo, se os dados tiverem uma sequência de horários, você poderá sincronizar entre conjuntos de dados ou vincular escolhendo um campo, como o nome de um cliente, em todos os dados que entrarem
- [ ] Desenvolva um plano de armazenamento de ciclo de vida de dados de longo prazo para saber como você gerenciará o volume e a velocidade deles à medida que entram e você os arquiva
- [ ] Considere contratar um diretor de dados para manter o gerenciamento de dados da sua organização para IA, deep learning e outros futuros projetos baseados em dados

IBM