DIGITAL INNOVATION ONE

Amazon Textract - Extraindo Dados Estruturados e Persistindo em um BD

Cassiano Peres
Analista e desenvolvedor de sistemas

DIGITAL INNOVATION ONE

Mais sobre mim

- CTO – Arabyka e Brexbit
- Graduado em TADS – UTFPR MD (2015)
- Liberdade e descentralização
- GitHub: **cassianobrexbit**
- LinkedIn: **peres-cassiano**

DIGITAL INNOVATION ONE

Desafio da Live

Nesta live vamos explorar o Amazon Textract, uma ferramenta de ML que utiliza OCR para extração de dados estruturados e persistindo em uma tabela de banco de dados no Amazon DynamoDB

DIGITAL INNOVATION ONE

Percurso

Etapa 1 O que é OCR?

Etapa 2 Como funciona o Amazon Textract?

Etapa 3 Prática

DIGITAL INNOVATION ONE

Requisitos

- ✔ Conta ativa na AWS
- ✔ Conhecimento básico em Python
- ✔ Vontade
- ✔ Curiosidade

DIGITAL INNOVATION ONE

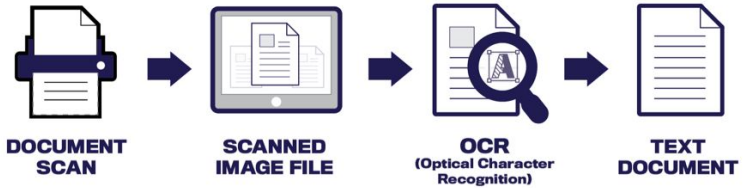
O que é OCR?

OCR é o acrônimo de ***Optical Character Recognition***, ou **Reconhecimento Óptico de Caracteres**, uma tecnologia para reconhecer caracteres a partir de um arquivo de imagem ou mapa de bits escaneados, escritos a mão, datilografados ou impressos. Dessa forma, através do OCR é possível obter um arquivo de texto editável a partir de arquivos não editáveis como imagens ou PDF.

DIGITAL INNOVATION ONE

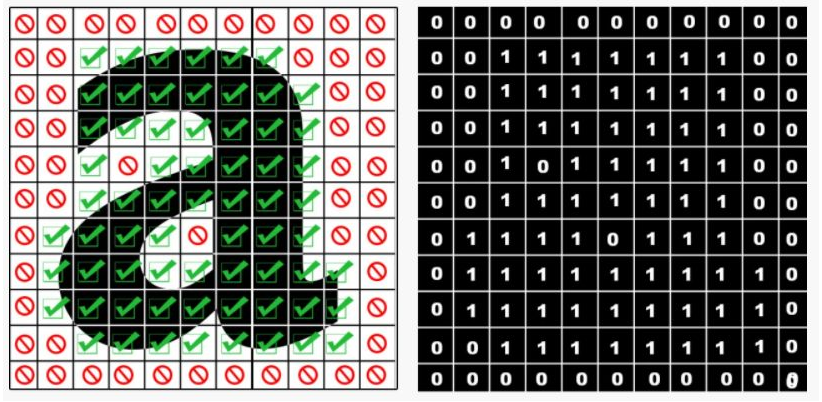
O que é OCR?

O OCR tem sido amplamente utilizado para automatizar processos de cadastro, onboarding e formalização, extraindo informações de documentos de identificação pessoal, contratos e comprovantes de residência e leitura de documentos manuscritos e identificação de placas de carros, por exemplo.



Como funciona o OCR?

O OCR trabalha dividindo a imagem de um caracter de texto em seções e as distinguindo entre regiões vazias e não vazias.



DIGITAL INNOVATION ONE

Fluxo do OCR

- **Aquisição**: ontem um texto não editável, como imagem ou PDF
- **Pré-processamento**: limpa a imagem reduzindo ruídos
- **Segmentação e extração de características**: Varredura do conteúdo da imagem em busca de grupos de pixels que sejam caracteres únicos e atribuição à sua própria classe.
- **Treinamento**: os dados são processados em uma sessão de treinamento de rede neural
- **Validação e retreino**: Após o processamento com correções alimentadas em sessões de treinamento subsequentes.

DIGITAL INNOVATION ONE

O Amazon Textract

O Amazon Textract é um serviço de ML que extrai automaticamente texto, manuscritos e dados de documentos digitalizados, com recursos que vão além do simples OCR, para identificar, entender e extrair dados de formulários e tabelas. Permite adicionar revisões humanas com o Amazon Augmented AI para fornecer supervisão de modelos e revisões de dados confidenciais.

Obs: O nível gratuito (Free Tier) permite a transcrição de 1000 páginas por mês.

DIGITAL INNOVATION ONE

O Amazon Textract

Características e benefícios:

- Extração de dados estruturados e não estruturados
- Já vem adequado às normas internacionais de segurança de dados
- Custo baixo e gerenciável – pay as you go
- Implementaçaõ de revisões feitas por pessoas
- Integração com outros serviços AWS (S3, RDS, DynamoDB, etc)
- Suporta arquivos nos formatos PNG, JPEG e PDF

DIGITAL INNOVATION ONE

O Amazon Textract

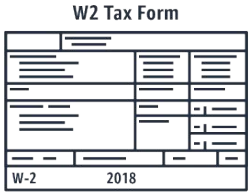
O Amazon Textract fornece operações síncronas e assíncronas que retornam apenas o texto detectado em um documento em múltiplos objetos *Block* (itens reconhecidos em um documento dentro de um grupo de pixels perto uns dos outros).

- As linhas e palavras do texto detectado
- As relações entre as linhas e palavras do texto detectado
- A página em que o texto detectado aparece
- A localização das linhas e palavras do texto na página do documento

DIGITAL INNOVATION ONE

Dados Estruturados no Amazon Textract

Tax, medical, banking, and other form documents



Textract recognizes many forms, such as W2, 1099-MISC, 1040, patient registration, and more

Automatically process documents without data entry or writing extraction rules

Key	Value
Name	John Smith
Address	123 Name St City name, ST 20391
ID	230-740
Company	The Company Name

Automatically extract key-value pairs and retain document context without manual intervention

When extracting text from documents and forms, Textract automatically detects and extracts structured data

Department	Budget	Actual	Difference
Accounting	$12,500.00	$11,293.12	$1,206.88
Finance	$24,000.00	$23,203.29	-$796.71
Human Resources	$13,000.00	$11,832.76	$1,167.24
Marketing	$82,500.00	$84,049.47	-$1.549.47
Sales	$76,000.00	$77,019.38	-$1,019.38

Textract preserves the tabular structure of extracted data, so that text remains grouped within each cell

With the tabular format of the data intact, easily upload extracted data into a database

DIGITAL INNOVATION ONE

Dados Estruturados no Amazon Textract

O Amazon Textract permite detectar **pares de chave-valor** em imagens de documentos automaticamente para reter o contexto inerente do documento sem qualquer intervenção manual.

Um par de chave valor é um conjunto de itens de dados vinculados. Por exemplo: o campo **“Nome”** seria a chave e **“Jane”** seria o valor.

DIGITAL INNOVATION ONE

Dados Estruturados no Amazon Textract

O Amazon Textract preserva a composição dos dados armazenados nas tabelas durante a extração.
Ideal para documentos compostos em grande parte por dados estruturados, como relatórios financeiros ou registros médicos com nomes de **colunas na linha superior da tabela, seguidos por linhas de entradas individuais**.

Este recurso pode ser utilizado para carregar automaticamente os dados extraídos em um banco de dados usando um esquema predefinido.

DIGITAL INNOVATION ONE

Aplicações no mercado

- Compliance
- Serviços financeiros
- Saúde (receituário, prontuário médico)
- Seguros
- Instituições governamentais

DIGITAL
INNOVATION
ONE

Prática

DIGITAL INNOVATION ONE

Mãos à obra

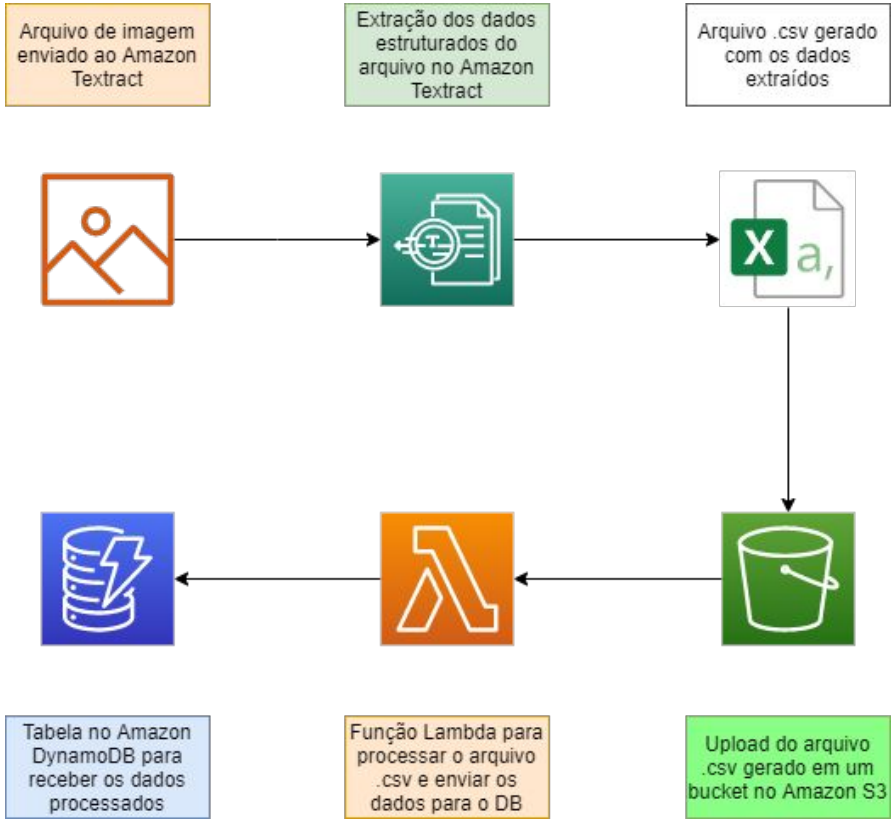
Nesta atividade vamos extrair dados estruturados de uma imagem e gerar um arquivo *.csv* utilizando o console do Amazon Textract. Em seguida vamos desenvolver uma função no AWS Lambda que lê o arquivo gerado e enviar os dados para uma tabela no banco de dados DynamoDB.

Link para o repositório no GitHub:

- https://github.com/cassianobrexbit/dio-live-textract2-05102021

DIGITAL INNOVATION ONE

Arquitetura da prática



DIGITAL
INNOVATION
ONE

Dúvidas?

DIGITAL INNOVATION ONE

Referências

- https://aws.amazon.com/pt/textract/features/
- https://docs.aws.amazon.com/textract/latest/dg/how-it-works-detecting.html
- https://docs.aws.amazon.com/textract/latest/dg/what-is.html
- https://www.itransition.com/blog/ocr-algorithm
- https://aws.amazon.com/getting-started/hands-on/extract-text-with-amazon-textract/
- https://aws.amazon.com/textract/features/