

Amazon Textract

Extraia automaticamente a caligrafia, o texto ou os dados de qualquer documento usando machine learning

O Amazon Textract é um serviço que detecta e extrai automaticamente texto e dados de documentos digitalizados. Ele vai além do simples optical character recognition (OCR – Reconhecimento óptico de caracteres) para também identificar o conteúdo de campos em formulários e informações armazenadas em tabelas.

O Amazon Textract ajuda você a adicionar detecção e análise de texto de documentos aos seus aplicativos. Usando o Amazon Textract, você pode fazer o seguinte:

- Detecte texto digitado e manuscrito em uma variedade de documentos, incluindo relatórios, registros médicos e formulários fiscais.
- Extraia texto, formulários e tabelas de documentos com dados estruturados, usando o API de análise de documentos do Amazon Textract.
- Especifique e extraia informações de documentos usando o recurso Consultas em a API do Amazon Textract Analyze Document.
- Processe faturas e recibos com a API AnalyzeExpense.
- Processar documentos de identificação, como carteiras de motorista e passaportes emitidos pelos EUA. government, usando a API AnalyzeID.
- Carregar e processar pacotes de empréstimos hipotecários, através do encaminhamento automático do páginas do documento para as operações de análise apropriadas do Amazon Textract usando o Analisar Fluxo de trabalho de empréstimo. É possível recuperar os resultados da análise para cada página do documento ou pode recuperar resultados resumidos para um conjunto de páginas do documento.
- Use Consultas Personalizadas para personalizar o recurso de Consultas pré-treinadas usando seus dados para atender às suas necessidades de processamento downstream.

AWS Textract

The screenshot displays the AWS Textract console interface. The main area shows the analysis results of a document titled "paystub". The results are organized into two columns: "Text simples" and "Formas". The "Formas" column contains a table of extracted form fields.

Formas	Text simples
CO.	FILE
ABC	126543
DEPT.	CLOCK
123456	12345
NUMBER	Period ending:
00000000	7/18/2008
Pay date:	Social Security Number:
7/25/2008	987-65-4321
JOHN STILES	Taxable Marital Status:
101 MAIN STREET ANYTOWN, USA 12345	Married
Federal:	State:
3. \$25 Additional Tax	2
Local:	total to date
2	27.00
Loan Amt Paid	Gross Pay
840.00	\$ 452.43

Below the main results, there is a section for "Escolha uma amostra de documento" (Choose a sample document) and "Upload de documento" (Upload document). The "Upload de documento" section includes a button "Escolha o documento" (Choose the document) and a link "Fazer upload de documentos em massa" (Upload documents in bulk).

Processos para iniciar a análise de documentos através do portal AWS

- Definir o escopo de análise a ser realizada. Qual tipo de documento, texto e dados a serem extraídos.
- Fazer o upload do arquivo ou lotes de arquivos em massa (batches) para plataforma, nesse caso uma opção seria o Amazon S3 (storage), conforme figura abaixo.

Análises

A aplicação do método de OCR em documentos utilizando a ferramenta Amazon Textract para conversão de textos de imagens ou documentos digitalizados em textos legíveis, permite dinamizar e automatizar tarefas.

Os recursos de oferecidos de índices de pesquisa inteligentes nos permite criar bibliotecas de texto através dos arquivos em PDF, ao mesmo tempo que é possível realizar extração de texto através de linhas completas, agrupamentos de células de tabelas, formulários.

Os dados extraídos podem ser tanto estruturados como não-estruturados pois a ferramenta tem a capacidade de normalizar os mesmos, baseando-se em diferentes fontes.

A ferramenta também possui recursos de classificação e extração de documentos utilizando-se de uma API do Analyze Lending permitindo a classificação automatizada dos documentos.

Através da API do Amazon Textract é possível implementar o recurso em aplicações desenvolvidas que necessitem desse recurso. Com isso o serviço se torna mais escalável permitindo um ganho maior de produtividade, alinhando o baixo custo.

Exemplos de Aplicações

- **Automação de Processos Bancários** – Extração de dados de cheques, faturas e contratos para reduzir entrada manual de dados.
- **Gerenciamento de Registros Médicos** – Conversão de formulários e prontuários médicos digitalizados em dados es
- **Processamento de Faturas e Notas Fiscais** – Extração automática de valores, datas e detalhes de fornecedores para sistemas financeiros.

- **Análise de Documentos Jurídicos** – Digitalização de contratos e petições com estruturação de informações relevantes.
- **Cadastro de Clientes (KYC - Know Your Customer)** – Extração de informações de RG, CNH e passaportes para validação de identidade.
- **Pesquisa e Indexação de Documentos** – Conversão de documentos em texto pesquisável para sistemas de gerenciamento de conteúdo.
- **Educação e Arquivos Acadêmicos** – Extração de textos de provas, currículos e certificados para processamento digital.
- **Seguros e Processamento de Sinistros** – Análise automática de apólices, formulários de sinistro e relatórios de inspeção.

AWS CLI – AWS Command Line Interface

É a ferramenta que fornece uma interface de interação com todo o ecossistema do Amazon Web Services. Cada serviço diferente da AWS, possui seus comandos a serem executados dentro do AWS CLI. Esses comandos podem ser consultados diretamente no guia do usuário, incluindo descrições, sintaxes e exemplos de uso na documentação oficial da AWS.

IAM – Identity and Access Management

Para que se possa executar remotamente os comandos via API da AWS é necessário criar um usuário específico para acesso à determinados serviços da Amazon Web Services. O processo se divide em 3 etapas e é feito diretamente no site da AWS logado com a conta do Administrador de serviços:

1. Criar o usuário e especificando detalhes do mesmo
2. Definir as permissões que serão concedidas para esse usuário
3. Revisão e criação

As **Chaves de Acesso** (Chave de Acesso / Chave de Acesso Secreta) deveram ser guardadas em arquivo (.CSV) e armazenadas em local seguro de único acesso.

Os exemplos aplicados nesse curso em Python através do AWS CLI, estão nas pastas descritas dentro deste material.