Amazon Textract

Amazon Textract is a service that automatically extracts text and data from scanned documents. Amazon Textract goes beyond simple optical character recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

Many companies today extract data from documents and forms through manual data entry that's slow and expensive or through simple optical character recognition (OCR) software that is difficult to customize. Rules and workflows for each document and form often need to be hard-coded and updated with each change to the form or when dealing with multiple forms. If the form deviates from the rules, the output is often scrambled and unusable.

Amazon Textract overcomes these challenges by using machine learning to instantly "read" virtually any type of document to accurately extract text and data without the need for any manual effort or custom code. With Textract you can quickly automate document workflows, enabling you to process millions of document pages in hours. Once the information is captured, you can take action on it within your business applications to initiate next steps for a loan application or medical claims processing. Additionally, you can create smart search indexes, build automated approval workflows, and better maintain compliance with document archival rules by flagging data that may require redaction.

Use cases

Create smart search indexes

Extract structured data from documents and create a smart index using Amazon Elasticsearch Service to allow you to search through millions of financial statements quickly. For example, a mortgage company could use Amazon Textract to process millions of scanned loan applications in a matter of hours and have the extracted data indexed in Amazon Elasticsearch. This would allow them to create search experiences like "search for loan applications where applicant name is John Doe," or "search contracts where the interest rate is 2 percent."

Build automated document processing workflows

Amazon Textract can provide the inputs required to automatically process forms without human intervention. For example, a bank could write code to read PDFs of loan applications. The information contained in the document could be used to initiate all of the necessary background and credit checks to approve the loan so that customers can get instant results of their application rather than having to wait several days for manual review and validation.

Maintain compliance in document archives

Because Amazon Textract identifies data types and form labels automatically, it's easy to maintain compliance with information controls. For example, an insurer could use Amazon Textract to feed

a workflow that automatically redacts personally identifiable information (PII) for their review before archiving claim forms by automatically recognizing the important key-value pairs that require protection.