A Seminar Report

on

"PROCESSING OF SCANNED DOCUMENTS THROUGH AWS TEXTRACT"

submitted by

Kiran S. Kolte
(47)

under the guidance of
Prof. S.S. Muddalkar

Department of Information Technology
Shri Sant Gajanan Maharaj College of Engineering,
Shegaon-444203 (MS)
(Affiliated to Sant Gadge Baba Amravati University, Amravati,
Recognized by AICTE New Delhi, Accredited by NBA and ISO 9001:2000 Certified)

(Academic Session 2021-22)

Shri Sant Gajanan Maharaj College of Engineering, Shegaon-444203

CERTIFICATE

*This is to certify that the seminar work entitled “**PROCESSING OF SCANNED DOCUMENTS THROUGH AWS TEXTRACT**” is bonafide work carried out by **Kiran S. Kolte**, in partial fulfilment of the requirements for the award of the degree of **Bachelor of Engineering in Information Technology, Sant Gadge Baba Amravati University, Amravati (MS)** during the year 2021-2022. The Seminar report has been approved as it satisfies the academic requirements in respect of seminar work prescribed for the Bachelor of Engineering degree.*

Prof. S.S. Muddalkar	**Prof. F I Khandwani**	**Prof. A S Manekar**
Guide	Seminar Coordinator	HOD

ACKNOWLEDGEMENT

It is my proud privilege and duty to acknowledge the kind of help and guidance received from several people in preparation for this report. It would not have been possible to prepare this seminar in this form without their valuable help, cooperation and guidance.

First and foremost, I wish to record my sincere gratitude to the **Management of this college** and to our beloved **Principal, Dr S B Somani,** for their constant support and encouragement in the preparation of this seminar and for making available internet, library and laboratory facilities needed to prepare this seminar.

Further my sincere thanks to **Prof. A S Manekar, Head of the Department, Information Technology,** for his valuable suggestions and guidance throughout this seminar.

I express my sincere gratitude to my guide, **Prof. S.S. Muddalkar,** for guiding me in investigations for this seminar and in carrying out relevant work. Our numerous discussions were extremely helpful. I received his/her esteem guidance, encouragement and inspiration.

I sincerely thanks **Prof. Faizan Khandwani,** Seminar Coordinator for supporting this seminar work. His contribution and technical support in preparing this seminar are greatly acknowledged.

Last but not the least, I wish to thank my **parents** for financing my studies in this college as well as for constantly encouraging me to learn engineering. Their sacrifice in providing this opportunity to learn engineering is gratefully acknowledged.

Place: Shegaon

Date: Kiran S. Kolte

ABSTRACT

Scanning of documents and extracting information from them is crucial part of every business. Here in this report, I have discussed the necessity of the AWS Textract and to automate the document analysing process. The report focuses on the using of the AWS Textract and the implementation of the same.

AWS Textract is a compelling proposition when used as part of a broader application or in conjunction with other AWS services for digital transformation purposes. Seen in isolation, Textract is just a modern twist on traditional OCR technology. But when Textract is seen as a tool to read a document, understand its structure, and extract information from it in the form in which it was meant to be read, that is surely just the starting point for transformation and automation.

Keywords: - AWS: Amazon Web Services, OCR: Optical Character Recognition, API: Application Programming Interface

CONTENTS

PROCESSING OF SCANNED DOCUMENTS THROUGH AWS TEXTRACT

1. INTRODUCTION

1.1 Overview

Amazon Textract is a machine learning (ML) service that uses OCR to automatically extract text, handwriting, and data from scanned documents such as PDFs. Amazon Textract makes it easy to add document text detection and analysis to your applications.

1.2 Why AWS Textract?

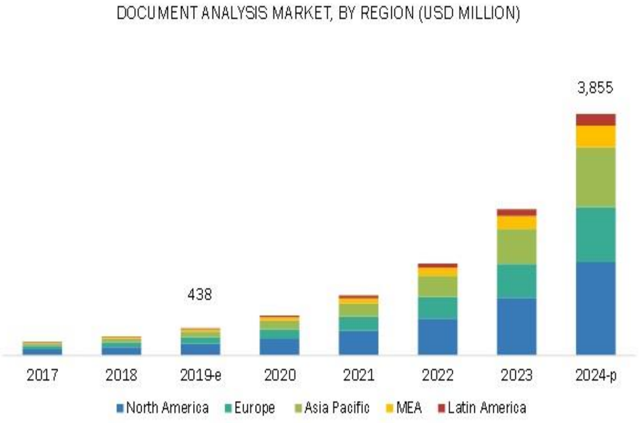
Amazon Textract is based on the same proven, highly scalable, deep-learning technology that was developed by Amazon's computer vision scientists to analyze billions of images and videos daily. Amazon Textract includes simple, easy-to-use APIs that can analyze image files and PDF files. Amazon Textract is always learning from new data, and Amazon is continually adding new features to the service.

1.3 Effectiveness of AWS Textract

In many companies and organizations, plenty of valuable business data is stored in documents. This data is at the heart of digital transformation. Unfortunately, according to statistics, 80% of all this data is embedded in unstructured formats like business invoices, emails, receipts, PDF documents, and many more. Therefore, to extract and make the most out of information from these documents, companies slowly started relying on Artificial intelligence (AI) based services. Out of those that provide AI-based services, Amazon had been one of the most prominent players for a long time. It had its wings spread across different solutions like document processing, speech recognition, text analytics, and much more.

2. LITERATURE SURVEY

Highlight in 1950's, applied throughout the spectrum of industries resulting into revolutionizing the document management process. Optical Character Recognition or OCR has enabled scanned documents to become more than just image files, turning into fully searchable documents with text content recognized by computers. It extracts the relevant information and automatically enters it into electronic database instead of the conventional way of manually retyping the text. It is a vast field with a number of varied applications such as invoice imaging, legal industry, banking, health care industry etc. It is also widely used in many other fields like Captcha, Institutional repositories and digital libraries, Optical Music Recognition without any human correction or human effort, Automatic number plate recognition and Handwritten Recognition.



e-estimated, p-projected
Source: MarketsandMarkets Analysis

There are 16.3M US mortgage applications from 20161 and the nearly quarter of a billion W2 tax forms expected to be processed in the US in 2018. Same scenario in our country too and in various department and sector.

2.1 AWS Textract is redefining how companies process documents in a Digital World

by Andrea Morton-Youmans (AWS Machine Learning Blog)

Think about the last time you opened a bank account, applied for insurance, or refinanced your home. It was probably done on paper. The number of documents in a mortgage packet alone is over 100 pages long. What do you do with all that paper? For many companies across a variety of industries, including financial services, healthcare, and manufacturing, processing these documents is painstaking. It's manual, slow, expensive, and error-prone, and data is often spread across disparate sources. As a result, creating and managing a document processing pipeline remains a challenge for many companies.

According to Ritu Jyoti from IDC, "Supporting document processing requires an AI-native platform that helps improve accuracy, performance, agility and flexibility while supporting a broad set of document types. Artificial

Intelligence (AI), can help streamline document automation providing better business outcomes, improved ROI, and reduce manual efforts."[

AWS has launched a solution to help organizations extract insights and automate processing documents of different formats (PDF, Word, raw text) and layouts (bullets, lists) using Amazon Comprehend. This new launch combines the power of natural language processing (NLP) and Optical Character Recognition (OCR) to help reduce the amount of pre-processing or post-processing required to process documents. You can now use custom named entity recognition (NER) on more document types without needing to convert your files to raw text.

AWS has been innovating in the intelligent document processing (IDP) space for years to convert data in documents into usable information for document-centric processes. AWS launched AI services like Amazon Textract, Amazon Comprehend, and others to help with the automation of extracting insights from documents. Since the launch of those services, improvements in accuracy and speed have been tenfold. These services offer new APIs like specialized support for invoices and receipts, handwriting and language support, plus improvements in latency.

2.2 Automate Document Processing in Logistics using AI

by Manikanth Pasumarti, Santosh Mohanty, and Narcisse Zekpa (AWS Architecture Blog)

Multi-modal transportation is one of the biggest developments in the logistics industry. There has been a successful collaboration across different transportation partners in supply chain freight forwarding for many decades. But there's still a considerable overhead of paperwork processing for each leg of the trip. Tens of billions of documents are processed in ocean freight forwarding alone. Using manual labor to process these documents (purchase orders, invoices, bills of lading, delivery receipts, and more) is both expensive and error-prone.

We need to automate the document processing in the logistics industry.

2.3 Taxes, governments, and great experiences using the cloud

Prasad Alavilli (AWS Public Sector Blog)

Government agencies focused on tax, revenue, employment security, and labour that are responsible for collecting various types of taxes from individuals and businesses such as income, payroll and unemployment insurance are often challenged by heavy workload. These challenges were exacerbated by COVID-19, newly enacted benefit programs (e.g., Pandemic Unemployment Assistance or PUA) and increased fraud and cybercrime. This forced governments to work within the constraints of reduced revenue, tighter budgets, and workforce shortages. Information technology and cloud services can be enablers for modernization of these business processes and tax systems, and improved constituent experience. For public administrators, these tools can also shorten the window between the enactment of a new law and its timely and successful implementation.

2.4 Using AI to rethink document automation and extract insights

by Neil Mackin, Aaron Sengstacken, and Nieves Garcia (AWS Public Sector Blog)

Documents have come a long way from being inscribed with ink on papyrus or scratched in runes by ancient civilizations. They are now a fundamental tool of modern life, with documents used to capture and record essential information in many ways including application forms, certificates and licenses, purchase orders and invoices, and legal contracts**.**

While digital transformation made strides in automating many processes, automating document management and entry has not been done until recently. The maturing of artificial intelligence (AI) has brought ready-made services that organizations can use, not only to automate data entry work but also to apply intelligence into the business process. Using modern AI capabilities on Amazon Web Services (AWS), organizations can transform approaches to document management. This allows public sector organizations to save time (enabling faster throughput especially during higher volume paperwork times), so they can help get constituents their services faster, and focus on the most valuable work of the high-touch or high-need cases. Document automation helps reduce human entry error and provide backup services in case of natural disaster.

Extracting insights from huge volumes of historic documents: An Australian federal agency and Cloudten

Cloudten, an AWS Advanced Consulting Partner, worked with an Australian federal agency to automate the ingestion and analysis of scanned documents, and to gain deeper business insights by helping extract the intent and sentiment of processed data. The agency held nearly a million pages of historic scanned PDFs and JPGs of archived data that needed to be ingested and processed. CloudTen delivered a solution designed around Amazon Textract, as well as a range of other cloud-native serverless services such as AWS Lambda to extract, process, and present analyzed data so that it delivered insights with the benefits of scaling in line with the workload.

In addition to using Amazon Textract to convert scanned images into digitized text, the solution also expanded the data ingestion pipeline to incorporate machine learning (ML) capabilities that facilitated natural language processing (NLP). This includes using Amazon Comprehend and Amazon SageMaker to perform advanced pattern analysis and interpretation on the extracted dataset to deliver actionable insights. The agency’s officers were quickly able to extract detailed information and intent from archived business documents that previously required lengthy manual analysis.

Improving citizen services with AI driven form processing: Firemind

Firemind, an AWS Partner based in the United Kingdom, helped local government agencies, such as Maidstone Borough Council and Tunbridge Wells, improve staff efficiency using AI to automate mundane aspects of their

workflow, like having to rotate images that arrive in the wrong orientation and then copy key information from scanned forms. Everyday, their citizens upload documents, such as photographs of completed applications related to local parking and housing. In turn, the council's staff review the uploaded forms and extract key information to drive the relevant business processes to resolution and deliver the citizen services.
Firemind automated this workflow by using Amazon Textract to analyze the uploaded documents and give immediate feedback to the citizen if they've mistakenly uploaded the wrong photograph, so they don't need to wait days until a human worker reaches the item in the queue and flags their error. Once the correct image is processed, Amazon Textract pulls the key information from the form, including the name, address, and vehicle details automatically, and identifies the rotation of the document based on the text orientation. If its the incorrect orientation, it flips the document a number of ways until the service determines that the document is the correct orientation.

The Firemind solution reduced the number of manual tasks that council staff needed to undertake by one-third, enabling them to become more time efficient and focus on more value added work as well as providing the key services to the citizens more quickly.

Supporting citizens through COVID-19: Arizona State University Cloud Innovation Center (CIC)

The Arizona State University Cloud Innovation Center (CIC) built an open source asset to refine the document processing technology of Amazon Textract for utility bill and drivers license data extraction. This solution was recently used by Wildfire, a state association for Community Action Agencies, and Prefix Health Technologies (Prefix), an AWS Partner Network (APN) Technology Partner, to help provide relief to citizens during the COVID-19 pandemic.

The Arizona benefits portal allows COVID-19 impacted households to pre-screen and apply for assistance with rent, mortgage, gas, electric, and water. Applicants can attach document images to the benefit applications using their mobile phone camera. Amazon Textract captures the data from the images and populates or verifies the data entered, which eliminates the need for manual verification and speeds up the processing time. In many cases, eligibility is determined at the point of entry and funds are credited to the customer's account with little or no delay. For additional details on the solution developed read, "A streamlined, mobile-first approach to service delivery for counties and states" where the solution they developed for Arizona resulted in 49% of the applications to be automatically approved, therefore reducing the time required to verify and distribute funds.

Why Amazon AI and ML services for document automation

Using AWS AI and ML services to automate document processing helps organizations:

- Reduce effort tailoring to each form type: Amazon Textract ingests and reads documents and forms without requiring any extensive pre-work to understand the form’s layout. Instead, the AI-based approach understands the content based on the physical layout, even extracting the data held in tables or forms and mapping that into machine readable structures to indicate what has been written in each part of a form by mapping those values to their respective data fields.
- Scale up and down as needed: Business operations are often challenged by managing peaks in demand, for example, during application deadlines or during events like the COVID-19 pandemic. Scalability and modern serverless cloud architectures are key, which help quickly ramp up to process large volume of documents and then to immediately scale down, minimizing the on-going costs.
- Combine human and AI expertise to confirm or correct data entry more easily and quickly: Tightly integrated augmented AI flags to a human reviewer the aspects of forms which the AI couldn’t read confidently. The combination of AI and a human working together delivers a highly robust approach to efficiently automating a document workflow.
- Recognize sustainability benefits: Organizations can reduce the carbon and energy expended in physically moving tons of physical paper documents between sites, and then storing the same in physical archives. Shifting to electronic document processing, with digital mailrooms ingesting and scanning the media, liberates the workforce away from being physically co-located with the documents. A wider shift for the workforce that AI brings, is the ability to rely on the AI for the mundane tasks and allow the human workforce to focus on more value adding tasks that require uniquely human skills.
- Extract more value from data to improve processes: Using ML techniques also raises the bar on how much value can be extracted from documents. Amazon Rekognition is used to identify and extract images or diagrams embedded within documents, saving time and manual effort by identifying and cropping out images. The text within documents is processed through services Amazon Translate, making it possible to support 55 languages and variants from Afrikaans to Vietnamese, without requiring in-house translators. Amazon Comprehend uses natural language processing techniques to help understand a document. This is often used to triage inbound correspondence by understanding the nature of the request, and directing the task to the best work queue. These can be fed directly into robotic process automation driven workflows to partially or fully undertake work that would require human teams.
- Gather data insights to improve services: Extracted data can be pushed into a graph database, such as Amazon Neptune, for subsequent network analysis. This approach helps detect application fraud where networks of associates, addresses, and businesses are identified from the graph that might be otherwise very hard to recognize.

3. METHODOLOGY

A crucial part of many business transactions today is the exchange and processing of scanned documents. One such example is invoice processing, which is an end-to-end process with many different tasks to handle invoices that are received. After an invoice is received, the information it contains is entered into an enterprise resource planning (ERP) system, either manually or by using optical character recognition (OCR) software. Other downstream processing tasks complete the processing of the invoice based on the information extracted, such as transaction date, who ordered it, what were the items, how much did each item cost, and the amount of taxes paid. Two of the challenges customers face with invoice processing are:

• How to digitally store invoices in a secure, centralized location that not only provides easy access but also is cost-effective.

There can be several one can apply as per their respective needs. We will be focusing on a particular aspect and the best one wherein a user can be able to specify the needs for getting the desired results.

3.1 Processing Documents with Synchronous Operations

Amazon Textract can synchronously detect and analyze documents that are provided as images in JPEG or PNG format. Synchronous operations return results in near real time. For more information about documents, see Text Detection and Document Analysis Response Objects.
This section covers how you can use Amazon Textract to detect process text using synchronous operations. These operations process one page of a document at a time. To process multipage documents, it is reccommended to use asynchronous operations. For more information about document processing using asynchronous operations, see Processing Documents with Asynchronous Operations.
You can use Amazon Textract synchronous operations for the following purposes:

- Text detection – You can detect lines and words on a single-page document image by using the DetectDocumentText operation. For more information, see Detecting Text.

- Text analysis – You can identify relationships between detected text on a single-page document by using the AnalyzeDocument operation. For more information, see Analyzing Documents.

- Invoice and Receipt Analysis – You can identify financially-related relationships between detected text on a single-page document using the AnalyzeExpense operation. For more information, see Analyzing Invoices and Receipts.

3.1.1 API Reference for Synchronous Operations

i) analyze_document(kwargs):**

Analyzes an input document for relationships between detected items.

The types of information returned are as follows:

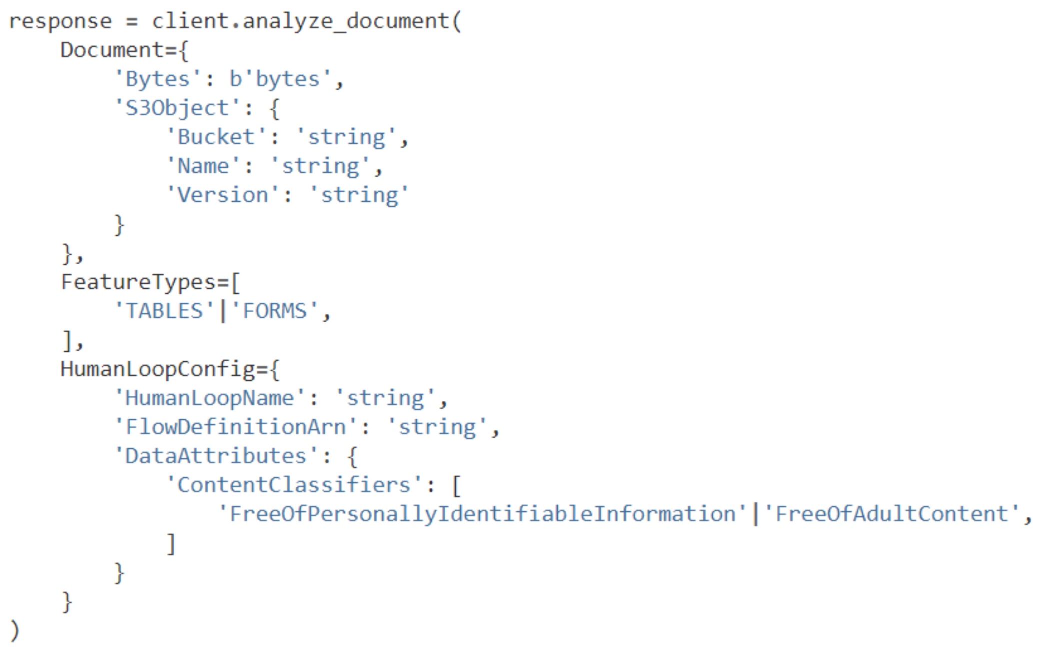
- Form data (key-value pairs). The related information is returned in two Block objects, each of type KEY_VALUE_SET : a KEY Block object and a VALUE Block object. For example, *Name: Ana Silva Carolina* contains a key and value. *Name:* is the key. *Ana Silva Carolina* is the value.
- Table and table cell data. A TABLE Block object contains information about a detected table. A CELL Block object is returned for each cell in a table.
- Lines and words of text. A LINE Block object contains one or more WORD Block objects. All lines and words that are detected in the document are returned (including text that doesn't have a relationship with the value of FeatureTypes).

Selection elements such as check boxes and option buttons (radio buttons) can be detected in form data and in tables. A SELECTION_ELEMENT Block object contains information about a selection element, including the selection status.

You can choose which type of analysis to perform by specifying the FeatureTypes list.

The output is returned in a list of Block objects.

Request Syntax



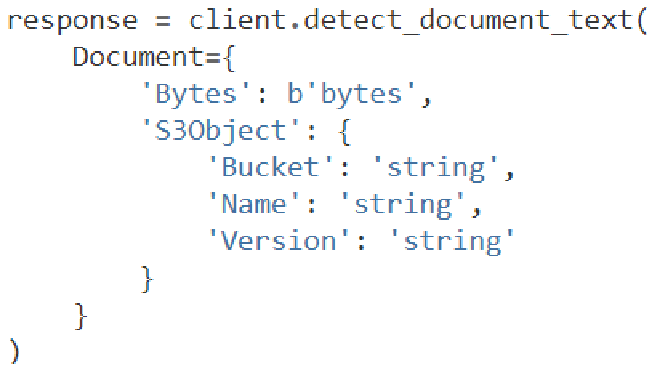
```
response = client.analyze_document(
    Document={
        'Bytes': b'bytes',
        'S3Object': {
            'Bucket': 'string',
            'Name': 'string',
            'Version': 'string'
        }
    },
    FeatureTypes=[
        'TABLES'|'FORMS',
    ],
    HumanLoopConfig={
        'HumanLoopName': 'string',
        'FlowDefinitionArn': 'string',
        'DataAttributes': {
            'ContentClassifiers': [
                'FreeOfPersonallyIdentifiableInformation'|'FreeOfAdultContent',
            ]
        }
    }
)
```

ii) detect_document_text(***kwargs*):

Detects text in the input document. Amazon Textract can detect lines of text and the words that make up a line of text. The input document must be an image in JPEG or PNG format. DetectDocumentText returns the detected text in an array of Block objects.

Each document page has as an associated Block of type PAGE. Each PAGE Block object is the parent of LINE Block objects that represent the lines of detected text on a page. A LINE Block object is a parent for each word that makes up the line. Words are represented by Block objects of type WORD.

Request Syntax



```
response = client.detect_document_text(
    Document={
        'Bytes': b'bytes',
        'S3Object': {
            'Bucket': 'string',
            'Name': 'string',
            'Version': 'string'
        }
    }
)
```

3.2 Processing Documents with Asynchronous Operations

Amazon Textract can detect and analyze text in multipage documents that are in PDF format. Multipage document processing is an asynchronous operation. Asynchronous processing of documents is useful for processing large, multipage documents. For example, a PDF file with over 1,000 pages takes a while to process. Processing the PDF file asynchronously allows your application to complete other tasks while it waits for the process to complete.
This section covers how you can use Amazon Textract to asynchronously detect and analyze text on a multipage or single-page document. Multipage documents must be in PDF format. Single-page documents processed with asynchronous operations can be in JPEG, PNG, or PDF format.
You can use Amazon Textract asynchronous operations for the following purposes:

- Text detection – You can detect lines and words on a multipage document. The asynchronous operations are StartDocumentTextDetection and GetDocumentTextDetection . For more information, see Detecting Text.

- Text analysis – You can identify relationships between detected text on a multipage document. The asynchronous operations are StartDocumentAnalysis and GetDocumentAnalysis . For more information, see Analyzing Documents.

3.2.1 API Reference for Asynchronous Operations

i) analyze_expense (kwargs):**

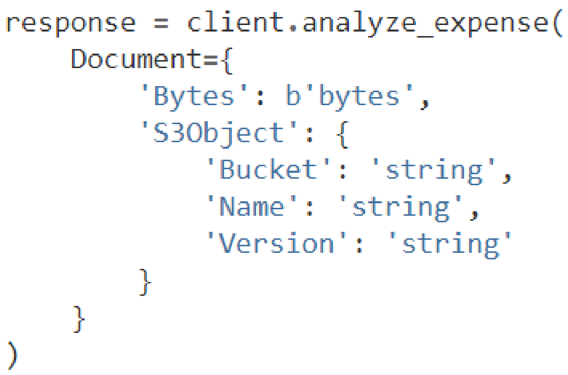
Analyzes an input document for financially related relationships between text.

Information is returned as ExpenseDocuments and seperated as follows.

- LineItemGroups - A data set containing LineItems which store information about the lines of text, such as an item purchased and its price on a receipt.

- SummaryFields - Contains all other information a receipt, such as header information or the vendors name.

Request Syntax



```
response = client.analyze_expense(
    Document={
        'Bytes': b'bytes',
        'S3Object': {
            'Bucket': 'string',
            'Name': 'string',
            'Version': 'string'
        }
    }
)
```

ii) get_document_analysis (***kwargs*):

Gets the results for an Amazon Textract asynchronous operation that analyzes text in a document.

You start asynchronous text analysis by calling StartDocumentAnalysis , which returns a job identifier (JobId). When the text analysis operation finishes, Amazon Textract publishes a completion status to the Amazon Simple Notification Service (Amazon SNS) topic that's registered in the initial call to StartDocumentAnalysis . To get the results of the text-detection operation, first check that the status value published to the Amazon SNS topic is SUCCEEDED . If so, call GetDocumentAnalysis , and pass the job identifier (JobId) from the initial call to StartDocumentAnalysis.

- Form data (key-value pairs). The related information is returned in two Block objects, each of type KEY_VALUE_SET : a KEY Block object and a VALUE Block object. For example, *Name: Ana Silva Carolina* contains a key and value. *Name:* is the key. *Ana Silva Carolina* is the value.

- Table and table cell data. A TABLE Block object contains information about a detected table. A CELL Block object is returned for each cell in a table.

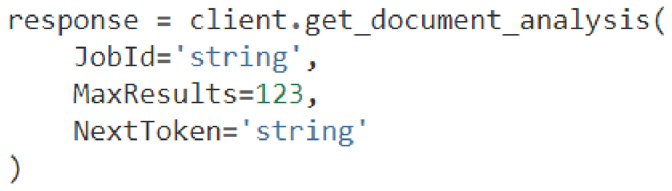
- Lines and words of text. A LINE Block object contains one or more WORD Block objects. All lines and words that are detected in the document are returned (including text that doesn't have a relationship with the value of the StartDocumentAnalysis FeatureTypes input parameter).

Selection elements such as check boxes and option buttons (radio buttons) can be detected in form data and in tables. A SELECTION_ELEMENT Block object contains information about a selection element, including the selection status.

Use the MaxResults parameter to limit the number of blocks that are returned. If there are more results than specified in MaxResults , the value of NextToken in the operation response contains a pagination token for getting the next set of results.

To get the next page of results, call GetDocumentAnalysis , and populate the NextToken request parameter with the token value that's returned from the previous call to GetDocumentAnalysis .

Request Syntax



```
response = client.get_document_analysis(
    JobId='string',
    MaxResults=123,
    NextToken='string'
)
```

iii) get_document_text_detection (***kwargs*):

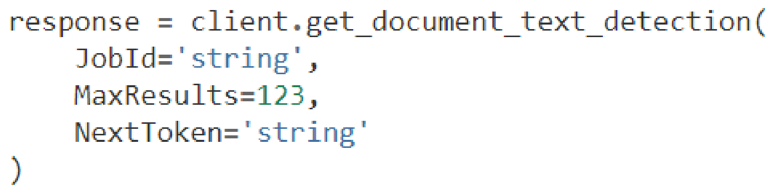
Gets the results for an Amazon Textract asynchronous operation that detects text in a document. Amazon Textract can detect lines of text and the words that make up a line of text.

You start asynchronous text detection by calling StartDocumentTextDetection , which returns a job identifier (JobId). When the text detection operation finishes, Amazon Textract publishes a completion status to the Amazon Simple Notification Service (Amazon SNS) topic that's registered in the initial call to StartDocumentTextDetection . To get the results of the text-detection operation, first check that the status value published to the Amazon SNS topic is SUCCEEDED . If so, call GetDocumentTextDetection , and pass the job identifier (JobId) from the initial call to StartDocumentTextDetection.

Each document page has as an associated Block of type PAGE. Each PAGE Block object is the parent of LINE Block objects that represent the lines of detected text on a page. A LINE Block object is a parent for each word that makes up the line. Words are represented by Block objects of type WORD.

Use the MaxResults parameter to limit the number of blocks that are returned. If there are more results than specified in MaxResults , the value of NextToken in the operation response contains a pagination token for getting the next set of results. To get the next page of results, call GetDocumentTextDetection , and populate the NextToken request parameter with the token value that's returned from the previous call to GetDocumentTextDetection .

Request Syntax



```
response = client.get_document_text_detection(
    JobId='string',
    MaxResults=123,
    NextToken='string'
)
```

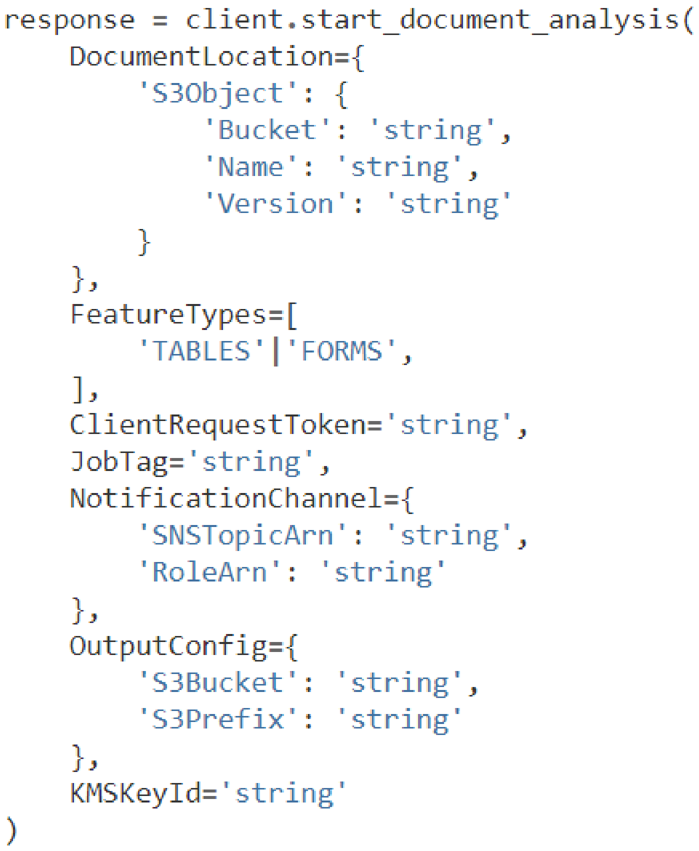
iv) start_document_analysis (***kwargs*):

Starts the asynchronous analysis of an input document for relationships between detected items such as key-value pairs, tables, and selection elements.

- StartDocumentAnalysis can analyze text in documents that are in JPEG, PNG, and PDF format. The documents are stored in an Amazon S3 bucket. Use DocumentLocation to specify the bucket name and file name of the document.

- StartDocumentAnalysis returns a job identifier (JobId) that you use to get the results of the operation. When text analysis is finished, Amazon Textract publishes a completion status to the Amazon Simple Notification Service (Amazon SNS) topic that you specify in NotificationChannel . To get the results of the text analysis operation, first check that the status value published to the Amazon SNS topic is SUCCEEDED . If so, call GetDocumentAnalysis , and pass the job identifier (JobId) from the initial call to StartDocumentAnalysis .

Request Syntax



```
response = client.start_document_analysis(
    DocumentLocation={
        'S3Object': {
            'Bucket': 'string',
            'Name': 'string',
            'Version': 'string'
        }
    },
    FeatureTypes=[
        'TABLES'|'FORMS',
    ],
    ClientRequestToken='string',
    JobTag='string',
    NotificationChannel={
        'SNSTopicArn': 'string',
        'RoleArn': 'string'
    },
    OutputConfig={
        'S3Bucket': 'string',
        'S3Prefix': 'string'
    },
    KMSKeyId='string'
)
```

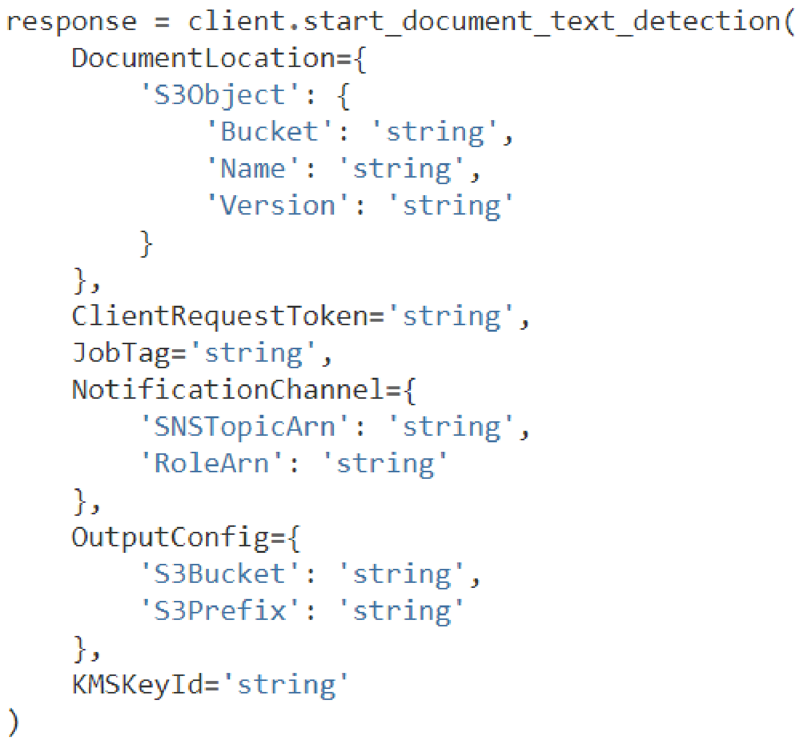
v) start_document_text_detection (*kwargs*):**

Starts the asynchronous detection of text in a document. Amazon Textract can detect lines of text and the words that make up a line of text.

- StartDocumentTextDetection can analyze text in documents that are in JPEG, PNG, and PDF format. The documents are stored in an Amazon S3 bucket. Use DocumentLocation to specify the bucket name and file name of the document.

- StartTextDetection returns a job identifier (JobId) that you use to get the results of the operation. When text detection is finished, Amazon Textract publishes a completion status to the Amazon Simple Notification Service (Amazon SNS) topic that you specify in NotificationChannel . To get the results of the text detection operation, first check that the status value published to the Amazon SNS topic is SUCCEEDED . If so, call GetDocumentTextDetection , and pass the job identifier (JobId) from the initial call to StartDocumentTextDetection.

Request Syntax



```
response = client.start_document_text_detection(
    DocumentLocation={
        'S3Object': {
            'Bucket': 'string',
            'Name': 'string',
            'Version': 'string'
        }
    },
    ClientRequestToken='string',
    JobTag='string',
    NotificationChannel={
        'SNSTopicArn': 'string',
        'RoleArn': 'string'
    },
    OutputConfig={
        'S3Bucket': 'string',
        'S3Prefix': 'string'
    },
    KMSKeyId='string'
)
```

4. FUNCTIONAL DIAGRAM

We will be implementing the AWS Textract with AWS Lambda where the processing of the pdf file will take place and hence further analysis can be done.

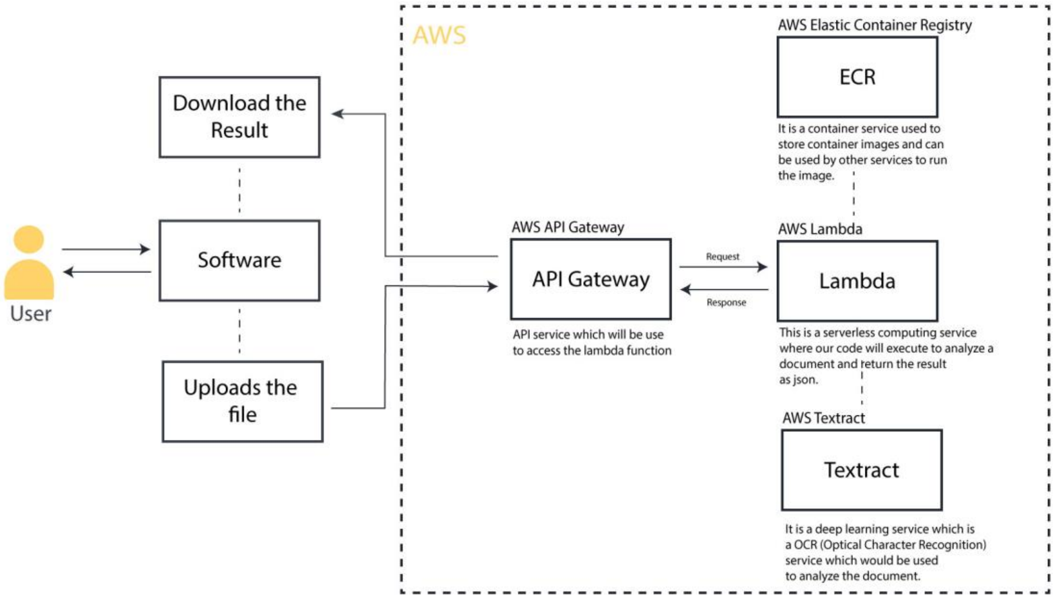
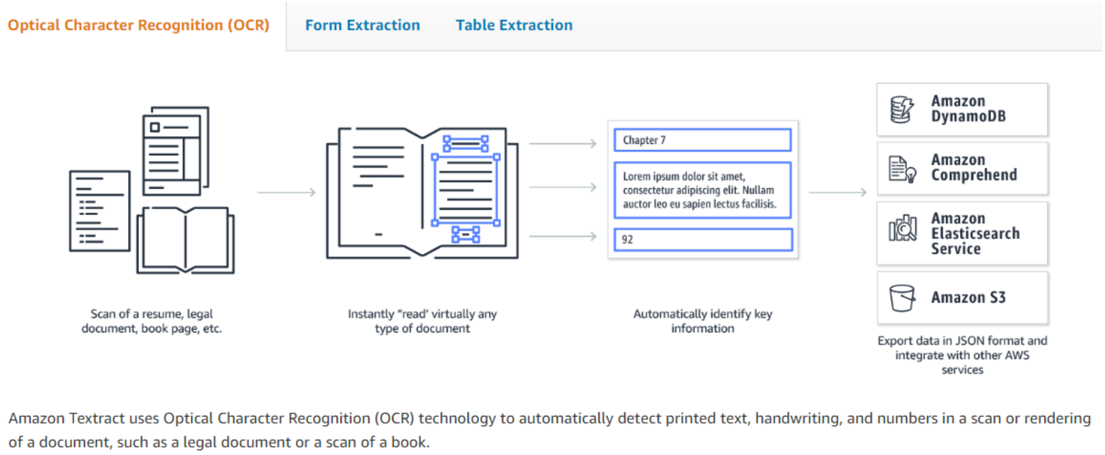
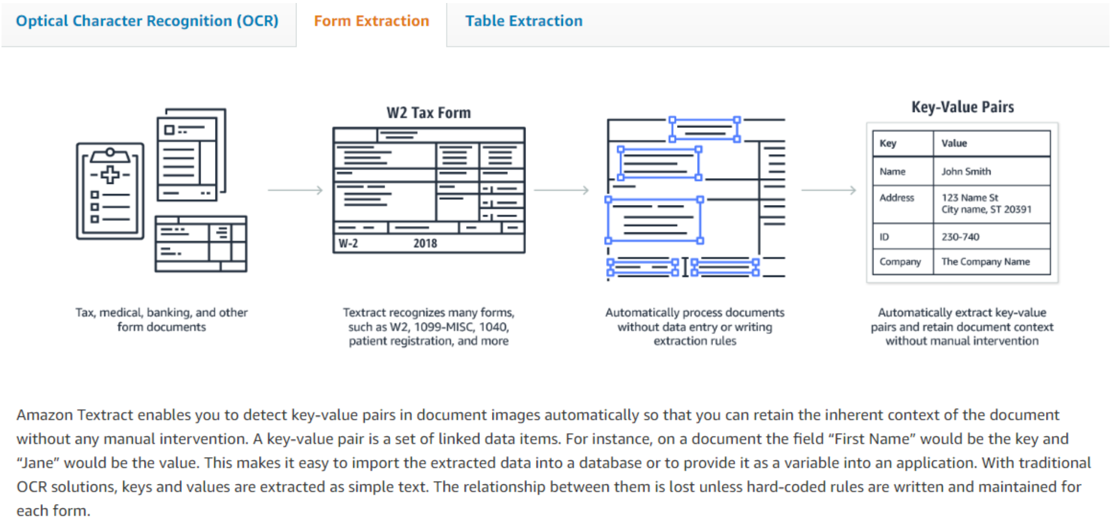

AWS
AWS Elastic Container Registry
ECR
It is a container service used to store container images and can be used by other services to run the image.
Download the Result
Software
User
Uploads the file
AWS API Gateway
API Gateway
API service which will be use to access the lambda function
Request
Response
AWS Lambda
Lambda
This is a serverless computing service where our code will execute to analyze a document and return the result as json.
AWS Textract
Textract
It is a deep learning service which is a OCR (Optical Character Recognition) service which would be used to analyze the document.

Figure 4.1 Software Architecture


Optical Character Recognition (OCR)
Form Extraction
Table Extraction
Chapter 7
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam auctor leo eu sapien lectus facilisis.
92
Amazon DynamoDB
Amazon Comprehend
Amazon Elasticsearch Service
Amazon S3
Scan of a resume, legal document, book page, etc.
Instantly "read' virtually any type of document
Automatically identify key information
Export data in JSON format and integrate with other AWS services

Amazon Textract uses Optical Character Recognition (OCR) technology to automatically detect printed text, handwriting, and numbers in a scan or rendering of a document, such as a legal document or a scan of a book.

Figure 4.2 OCR Scanning of Document



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

Tax, medical, banking, and other form documents

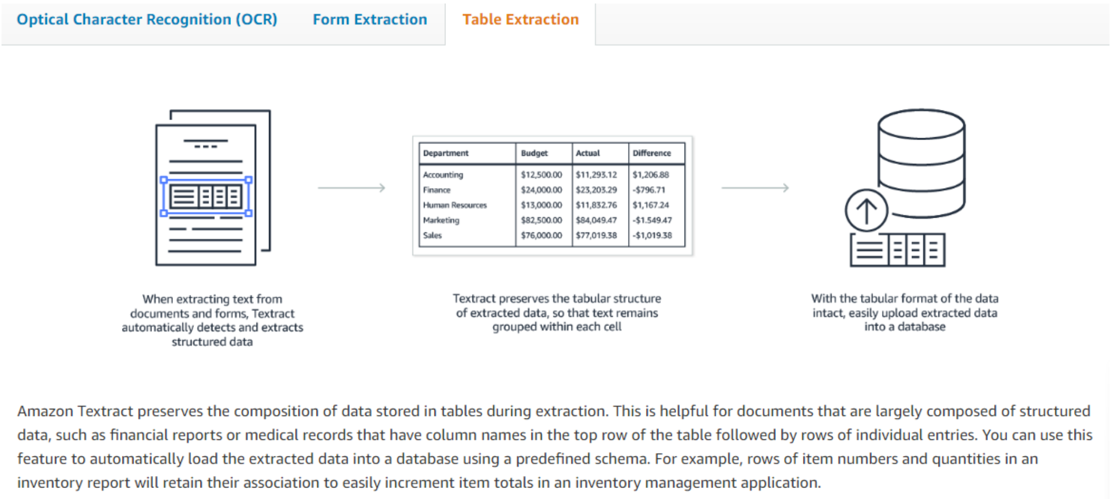
Textract recognizes many forms, such as W2, 1099-MISC, 1040, patient registration, and more

Automatically process documents without data entry or writing extraction rules

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

Amazon Textract enables you to detect key-value pairs in document images automatically so that you can retain the inherent context of the document without any manual intervention. A key-value pair is a set of linked data items. For instance, on a document the field "First Name" would be the key and "Jane" would be the value. This makes it easy to import the extracted data into a database or to provide it as a variable into an application. With traditional OCR solutions, keys and values are extracted as simple text. The relationship between them is lost unless hard-coded rules are written and maintained for each form.

Figure 4.3 Form Extraction Scanning of Document



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

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

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

Amazon Textract preserves the composition of data stored in tables during extraction. This is helpful for documents that are largely composed of structured data, such as financial reports or medical records that have column names in the top row of the table followed by rows of individual entries. You can use this feature to automatically load the extracted data into a database using a predefined schema. For example, rows of item numbers and quantities in an inventory report will retain their association to easily increment item totals in an inventory management application.

Figure 4.4 Table Extraction Scanning of Document

As seen from the figure 4.1 of our project, we can easily understand the nature and implementation of the AWS Textract into the software.

Firstly, a user will be provided with the GUI based C# application where he will be interacting with the software and choose the options and available features from the software.

The user can upload PDF or Image of the invoice/bill/receipt with the help of the software. Once he selects the file, he can then be able to choose from options provided to how he wants the results. For better results he can also specify what type of file it is.

Once the user selects to analyze the file, then the process shown inside the dotted boundary box starts.

As one can see the Request and Response through API Gateway*, it means that the PDF is request sent to API with certain parameters and response is the results we want from the PDF.

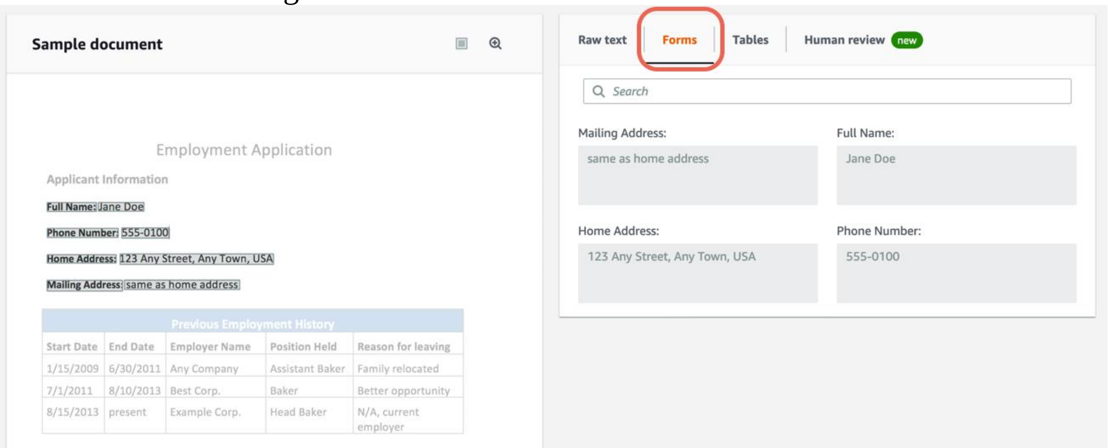
*API Gateway: Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. APIs act as the "front door" for applications to access data, business logic, or functionality from your backend services. Using API Gateway, you can create RESTful APIs and WebSocket APIs that enable real-time two-way communication applications. API Gateway supports containerized and serverless workloads, as well as web applications.

After the pdf gets passed through the request it will then be used by AWS Lambda, the lambda runs on Docker based container which runs a Python based application where it will process the PDF file and the desired results will be analyzed.

Once after successful completion of the application the user then will be able to download the results in Text/JSON/CSV file.

5. COMPARATIVE STUDY AND ANALYSIS

AWS Textract is a deep learning-based service that converts different types of documents into an editable format. Consider we have hard copies of invoices from different companies and store all the vital information from them on excel/spreadsheets. Usually, we rely on data entry operators to manually enter them, which is hectic, time-consuming, and error-prone. But using Textract, all we need to do is upload our invoices to it and in turn, it returns all the text, forms, key-value pairs, and tables in the documents in a more structured way. Below is a screenshot of how AWS does intelligent information extraction:



Employment Application

Applicant Information

Full Name: Jane Doe

Phone Number: 555-0100

Home Address: 123 Any Street, Any Town, USA

Mailing Address: same as home address

Previous Employment History

Start Date	End Date	Employer Name	Position Held	Reason for leaving
1/15/2009	6/30/2011	Any Company	Assistant Baker	Family relocated
7/1/2011	8/10/2013	Best Corp.	Baker	Better opportunity
8/15/2013	present	Example Corp.	Head Baker	N/A, current employer

Mailing Address: same as home address

Full Name: Jane Doe

Home Address: 123 Any Street, Any Town, USA

Phone Number: 555-0100

Figure 5.1 Information Extraction on AWS Textract

Not just typed text, AWS Textract also identifies handwritten texts in the documents. This makes information extraction more useful, as in some cases extracting handwritten text is more complicated to extract than typed ones. Now let’s see some of the common use cases for using Textract:

Robust and Normalised Data Capture: Amazon Textract enables text and tabular data extraction from a wide variety of documents, such as financial documents, research reports, and medical notes. However, these are not custom-made APIs, but they learn from a vast amount of data every day, and with this continuous learning, extracting unstructured and structured data from your document will be much easier.

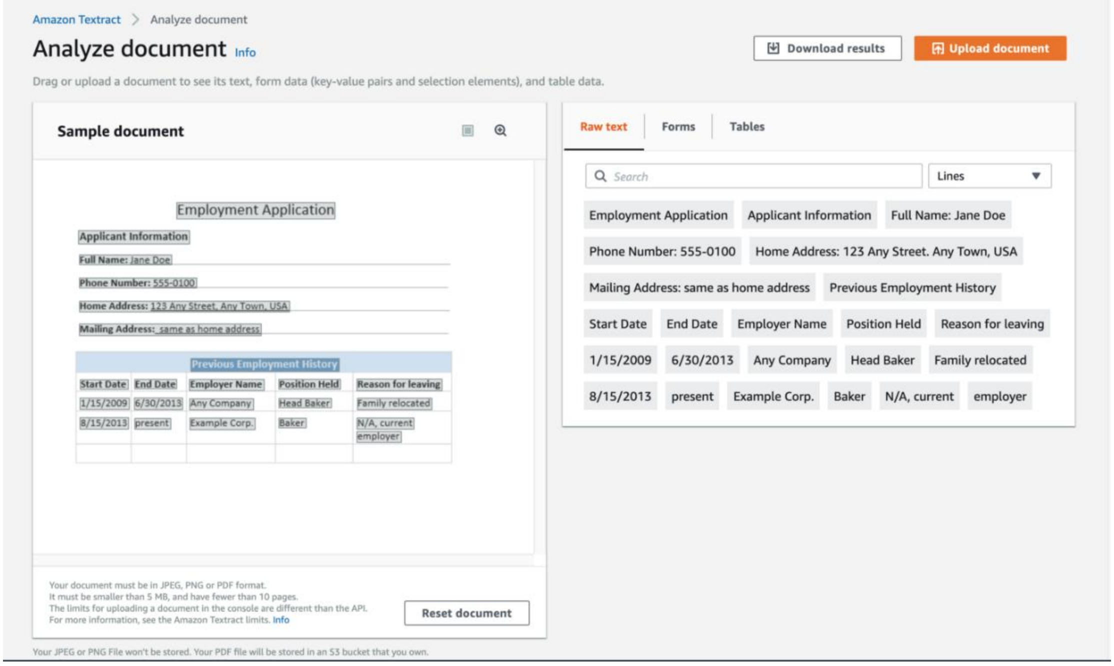
Key-Value Pair Extraction: Key-Value pair extraction has become a common problem for document processing but with Amazon Textract this can be easily solved. We can build pipelines for key-value pair extraction using Textract that automates document processing right from scanning documents to pushing data to excel sheets etc.

Creating an intelligent search index: Amazon Textract enables you to create libraries of text detected in image and PDF files.

Using intelligent text extraction for Natural Language Processing (NLP) – Amazon Textract enables you to extract text into words and lines. It also groups text by table cells if Amazon Textract document table analysis is enabled. Amazon Textract provides you with control over how text is grouped as input for NLP.

Amazon describes its machine learning (ML) technology stack as a three-layered cake. At the bottom is a layer for advanced practitioners to work with deep learning frameworks such as Tensorflow or Pytorch. The next layer is for developers to work with Amazon SageMaker to build out managed ML capabilities. The top layer provides pre-built artificial intelligence (AI) services that can be used immediately via an API integration point. Textract is one of these AI services offered by AWS. In simple terms, Textract provides data extraction from scanned documents. It does this by delivering ML-based optical character recognition (OCR) tools to read the document and accurately extract data and text.

OCR tools have been on the market for decades but came into the mainstream in the 1990s. However, though the concept of reading a letter or number in a scanned document is simple, it is extremely difficult to do with any degree of accuracy. The resolution quality of the scan and differing font sizes and types – not to mention handwriting quality – make the task of data extraction infinitely complex. No OCR does the job perfectly, including Textract; there will always be an error rate and exceptions to deal with. Yet even though Textract has been on the market less than a year, tests show its accuracy rate is as good as, and possibly better and more sophisticated than, many older systems on the market. And it should be noted that this accuracy rate will only improve as the system is used more extensively and the underlying machine learning has the opportunity to learn and improve.



Analyze document Info

Drag or upload a document to see its text, form data (key-value pairs and selection elements), and table data.

Employment Application

Applicant Information

Full Name: Jane Doe

Phone Number: 555-0100

Home Address: 123 Any Street, Any Town, USA

Mailing Address: same as home address

Previous Employment History				
Start Date	End Date	Employer Name	Position Held	Reason for leaving
1/15/2009	6/30/2013	Any Company	Head Baker	Family relocated
8/15/2013	present	Example Corp.	Baker	N/A, current employer

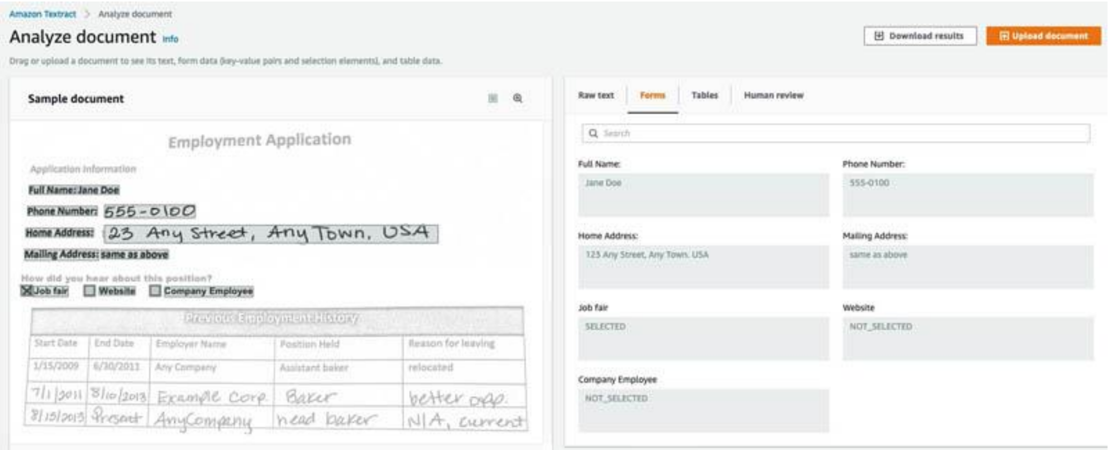
Your document must be in JPEG, PNG or PDF format.
It must be smaller than 5 MB, and have fewer than 10 pages.
The limits for uploading a document in the console are different than the API.
For more information, see the Amazon Textract limits. Info

Your JPEG or PNG File won't be stored. Your PDF file will be stored in an S3 bucket that you own.

```
Employment Application    Applicant Information    Full Name: Jane Doe
Phone Number: 555-0100    Home Address: 123 Any Street. Any Town, USA
Mailing Address: same as home address    Previous Employment History
Start Date    End Date    Employer Name    Position Held    Reason for leaving
1/15/2009    6/30/2013    Any Company    Head Baker    Family relocated
8/15/2013    present    Example Corp.    Baker    N/A, current    employer
```

Figure 5.1 Text Extraction from Image

5.1 Form Data based Analysis Samples



Amazon Textract > Analyze document

Analyze document Info

Drag or upload a document to see its text, form data (key-value pairs and selection elements), and table data.

Sample document

Employment Application

Application Information

Full Name: Jane Doe

Phone Number: 555-0100

Home Address: 23 Any Street, Any Town, USA

Mailing Address: same as above

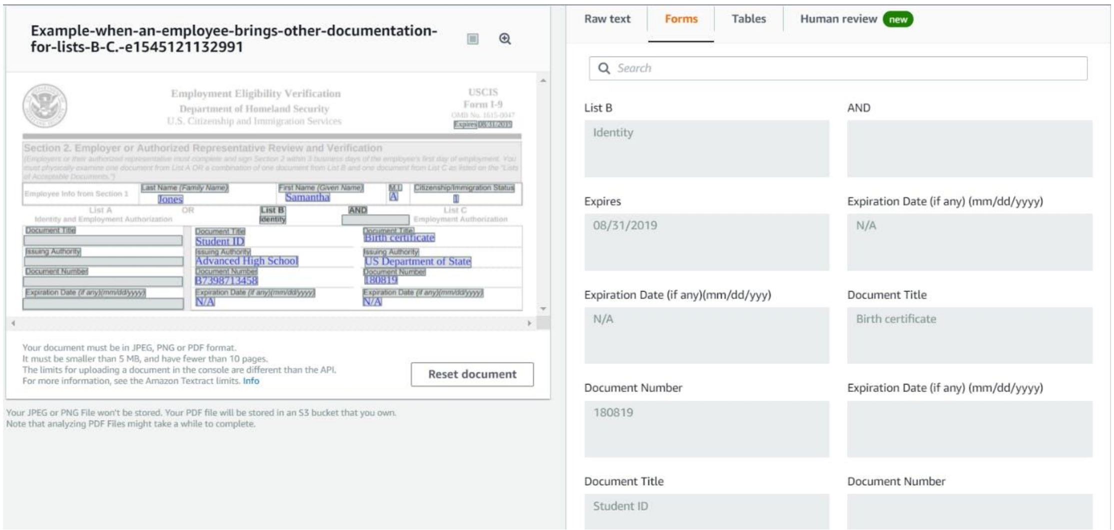
How did you hear about this position?

☒ Job fair ☐ Website ☐ Company Employee

Previous Employment History				
Start Date	End Date	Employer Name	Position Held	Reason for leaving
1/15/2009	6/30/2011	Any Company	Assistant baker	relocated
7/1/2011	8/10/2013	Example Corp.	Baker	better opp.
8/15/2013	Present	AnyCompany	head baker	N/A, current

Full Name:	Phone Number:
Jane Doe	555-0100
Home Address:	Mailing Address:
123 Any Street, Any Town, USA	same as above
Job fair	Website
SELECTED	NOT_SELECTED
Company Employee	
NOT_SELECTED	

Figure 5.1.1 Sample



Example-when-an-employee-brings-other-documentation-for-lists-B-C.-e1545121132991

Employment Eligibility Verification
Department of Homeland Security
U.S. Citizenship and Immigration Services

USCIS
Form I-9
OMB No. 1615-0047
Expires 08/31/2019

Section 2. Employer or Authorized Representative Review and Verification

(Employers or their authorized representative must complete and sign Section 2 within 3 business days of the employee's first day of employment. You must physically examine one document from List A OR a combination of one document from List B and one document from List C as listed on the "Lists of Acceptable Documents.")

Employee Info from Section 1	Last Name (Family Name)	First Name (Given Name)	M.I.	Citizenship/Immigration Status
	Jones	Samantha	A	1

List A Identity and Employment Authorization	OR	List B Identity	AND	List C Employment Authorization
Document Title		Document Title Student ID		Document Title Birth certificate
Issuing Authority		Issuing Authority Advanced High School		Issuing Authority US Department of State
Document Number		Document Number B7398713458		Document Number 180819
Expiration Date (if any)(mm/dd/yyyy)		Expiration Date (if any)(mm/dd/yyyy) N/A		Expiration Date (if any)(mm/dd/yyyy) N/A

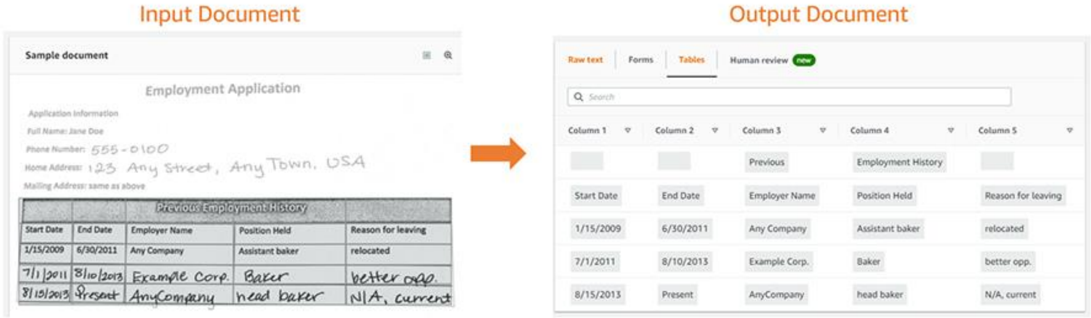
Your document must be in JPEG, PNG or PDF format.
It must be smaller than 5 MB, and have fewer than 10 pages.
The limits for uploading a document in the console are different than the API.
For more information, see the Amazon Textract limits. Info

Your JPEG or PNG File won't be stored. Your PDF file will be stored in an S3 bucket that you own.
Note that analyzing PDF Files might take a while to complete.

List B	AND
Identity	
Expires	Expiration Date (if any) (mm/dd/yyyy)
08/31/2019	N/A
Expiration Date (if any)(mm/dd/yyy)	Document Title
N/A	Birth certificate
Document Number	Expiration Date (if any) (mm/dd/yyyy)
180819	
Document Title	Document Number
Student ID	

Figure 5.1.2 Sample

5.2 Table Data based Analysis Samples



Input Document

Sample document

Employment Application

Application Information

Full Name: Jane Doe

Phone Number: 555-0100

Home Address: 123 Any Street, Any Town, USA

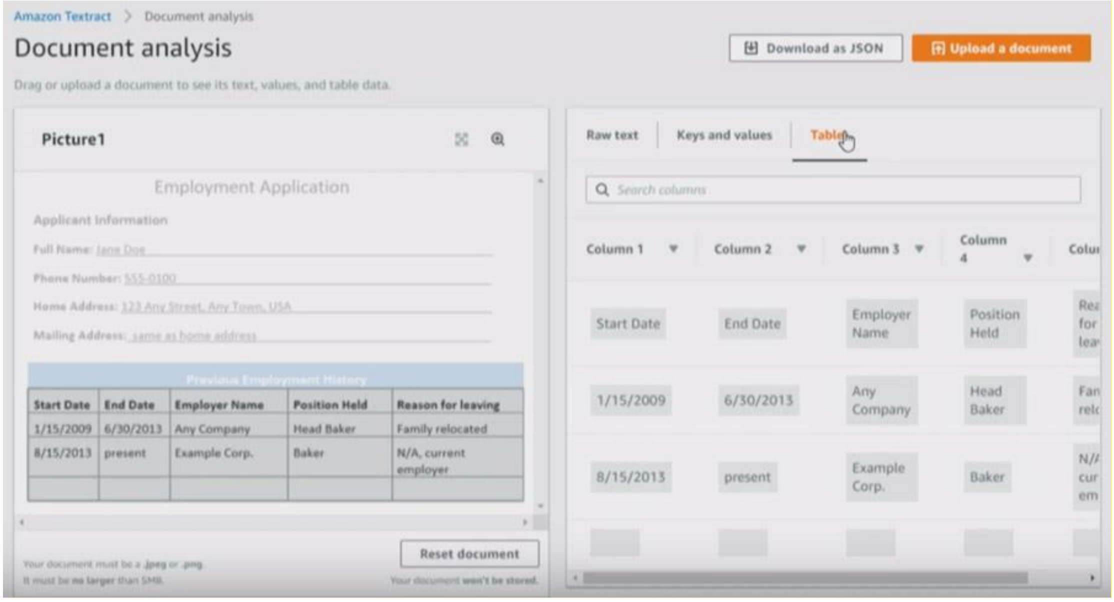
Mailing Address: same as above

Previous Employment History				
Start Date	End Date	Employer Name	Position Held	Reason for leaving
1/15/2009	6/30/2011	Any Company	Assistant baker	relocated
7/1/2011	8/10/2013	Example Corp.	Baker	better opp.
8/15/2013	Present	AnyCompany	head baker	N/A, current

Output Document

Column 1	Column 2	Column 3	Column 4	Column 5
		Previous	Employment History	
Start Date	End Date	Employer Name	Position Held	Reason for leaving
1/15/2009	6/30/2011	Any Company	Assistant baker	relocated
7/1/2011	8/10/2013	Example Corp.	Baker	better opp.
8/15/2013	Present	AnyCompany	head baker	N/A, current

Figure 5.2.1 Sample



Document analysis

Drag or upload a document to see its text, values, and table data.

Picture1

Employment Application

Applicant Information

Full Name: Jane Doe

Phone Number: 555-0100

Home Address: 123 Any Street, Any Town, USA

Mailing Address: same as home address

Previous Employment History

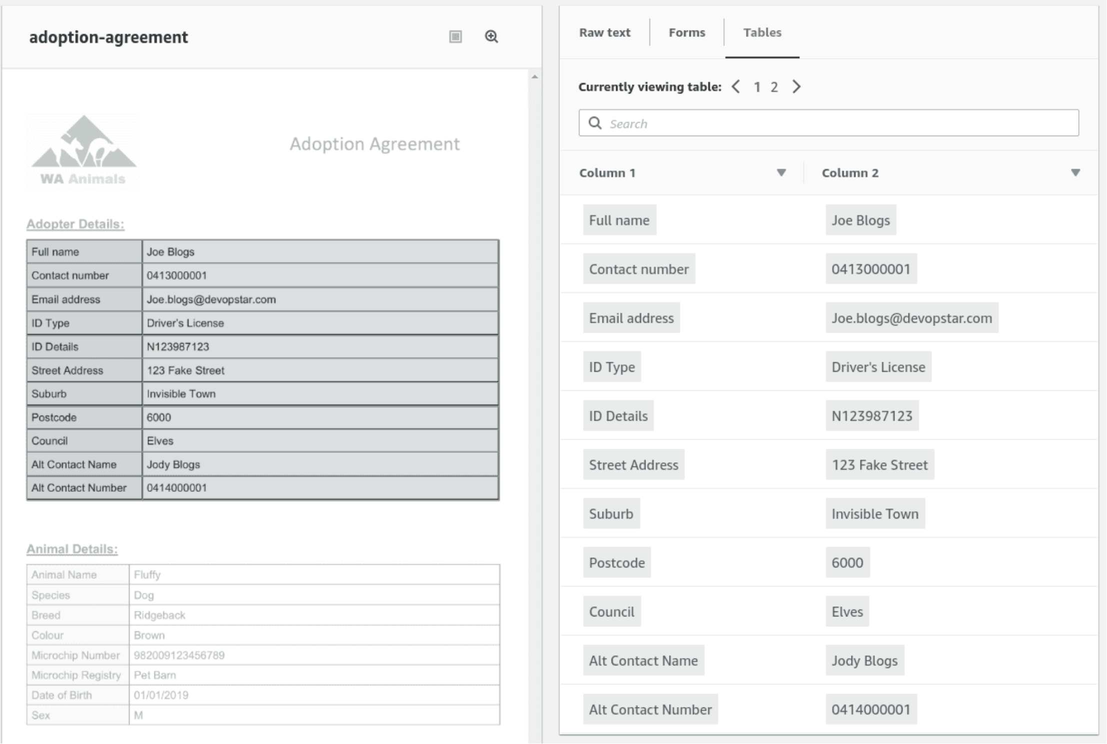
Start Date	End Date	Employer Name	Position Held	Reason for leaving
1/15/2009	6/30/2013	Any Company	Head Baker	Family relocated
8/15/2013	present	Example Corp.	Baker	N/A, current employer

Your document must be a .jpeg or .png.
It must be no larger than 5MB.

Your document won't be stored.

Column 1	Column 2	Column 3	Column 4	Colu
Start Date	End Date	Employer Name	Position Held	Rea for lea
1/15/2009	6/30/2013	Any Company	Head Baker	Fan relc
8/15/2013	present	Example Corp.	Baker	N/A cur em

Figure 5.2.2 Sample



adoption-agreement

Adoption Agreement

Adopter Details:

Full name	Joe Blogs
Contact number	0413000001
Email address	Joe.blogs@devopstar.com
ID Type	Driver's License
ID Details	N123987123
Street Address	123 Fake Street
Suburb	Invisible Town
Postcode	6000
Council	Elves
Alt Contact Name	Jody Blogs
Alt Contact Number	0414000001

Animal Details:

Animal Name	Fluffy
Species	Dog
Breed	Ridgeback
Colour	Brown
Microchip Number	982009123456789
Microchip Registry	Pet Barn
Date of Birth	01/01/2019
Sex	M

Currently viewing table: < 1 2 >

Column 1	Column 2
Full name	Joe Blogs
Contact number	0413000001
Email address	Joe.blogs@devopstar.com
ID Type	Driver's License
ID Details	N123987123
Street Address	123 Fake Street
Suburb	Invisible Town
Postcode	6000
Council	Elves
Alt Contact Name	Jody Blogs
Alt Contact Number	0414000001

Figure 5.2.3 Sample

6. IMPLICATIONS

AWS Textract has been applied to a number of applications. Some of them have been explained below.

A. **Invoice Imaging**

Invoice imaging is widely used in many businesses applications to keep track of financial records and prevent a backlog of payments from piling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes. As the technology continues to develop, more and more applications are found for OCR technology, including increased use of handwriting recognition. Furthermore, other technologies related to OCR, such as barcode recognition, are used daily in retail and other industries.

B. **Legal Industry**

Legal industry is also one of the beneficiaries of the OCR technology. OCR is used to digitize documents, and directly entered to computer database. Legal professionals can further search documents required from huge databases by simply typing a few keywords.

C. **Banking**

Another important application of OCR is in banking, where it is used to process cheques without human involvement. A cheque can be inserted into a machine where the system scans the amount to be issued and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well reducing the waiting time in banks.

D. **Healthcare**

Healthcare has also seen an increase in the use of OCR technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. Form processing tools, powered by OCR, are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded.

E. **Handwriting Recognition**

Handwriting recognition is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition) or intelligent word recognition. Alternatively, the movements of the pen tip may be sensed "on line", for example by a pen-based computer screen surface.

7. LIMITATIONS

7.1 Pros:

1. **Easy Setup with AWS Services:** Setting up Textract with another AWS service is an easy task compared to other providers. For example, storing extracted document information with Amazon DynamoDB or S3 can be done by configuring an add-on.
2. **Secure:** Amazon Textract conforms to the AWS shared responsibility model, which includes regulations and guidelines for data protection. AWS is responsible for protecting the global infrastructure that runs all the AWS services; therefore, we need not worry about our data being leaked or used by any others.

7.2 Cons:

1. **Inability to Extract Custom Fields:** There could be multiple data fields in a given invoice, say Invoice ID, Due Date, Transaction Date etc. These fields are something that are common in most invoices. But if we want to extract a custom field from an invoice, say, GST number or bank information, Textract does a poor job.
2. **Integrations with upstream and downstream providers:** Textract doesn't allow you to integrate with different providers easily, say, for example, we'll have to build an RPA pipeline with a third-party service; it would be difficult to find appropriate plugins that suit Textract.
3. **Ability to define table headers:** For table extraction tasks, textract doesn't allow you to define table headers. Therefore, it would be not easy to search or find a particular column or a table in a given document.
4. **No Fraud Checks:** Modern OCRs are now able to find if a given document is original or fake by validating dates and finding pixelated regions. AWS Textract doesn’t come with this; its only job is to pick all the text from an uploaded document.
5. **No Vertical Text Extraction:** In some of the documents, invoice numbers or addresses can be found in a vertical alignment. At present, AWS only supports horizontal text extraction with a slight in-plane rotation.
6. **Language Limit:** Amazon Textract supports English, Spanish, German, French, Italian, and Portuguese text detection. Amazon Textract will not return the language detected in its output.
7. **Everything’s Cloud:** Any document processed with Textract goes into the cloud, only supporting a few regions. More information here. However, some companies might not be interested in taking their documents to the cloud for reasons like confidentiality or legal requirements. Still, unfortunately, AWS Textract does not support any on-premise deployment for document processing.

7.3 Provide an Optimal Input Document

The following is a list of a few ways that you can optimize your input documents for better results.

- Ensure that your document text is in a language that Amazon Textract supports. Amazon Textract supports: English, Spanish, German, Italian, French, and Portuguese.
- If using an image of a document use a high quality picture, ideally at least 150 DPI.
- If your document is already in one of the file formats that Amazon Textract supports (PDF, JPEG, and PNG), don't convert or downsample the document before uploading it to Amazon Textract.

For the best results when extracting text from tables in documents, ensure that:

- Tables in your document are visually separated from surrounding elements on the page. For example, the table isn't overlaid onto an image or complex pattern.
- Text within the table is upright. For example, the text isn't rotated relative to other text on the page.

When extracting text from tables, you might see inconsistent results when:

- Merged table cells that span multiple columns.
- Tables with cells, rows, or columns that are different from other parts of the same table.

7.4 Use Confidence Scores

You should take into account the confidence scores returned by Amazon Textract API operations and the sensitivity of their use case. A confidence score is a number between 0 and 100 that indicates the probability that a given prediction is correct. It helps you make informed decisions about how you use the results.

In applications that are sensitive to detection errors (false positives), enforce a minimum confidence score threshold. The application should discard results below that threshold or flag situations as requiring a higher level of human scrutiny.

The optimal threshold depends on the application. For archival purposes, such as documenting handwritten notes, it might be as low as 50%. Business processes involving financial decisions might require thresholds of 90% or higher.

8. CONCLUSION

The ability to automate document processing remotely has proven essential as companies face new challenges in this pandemic. Demand for services like loan processing and grocery delivery has spiked in areas that no one could have predicted and the ability to quickly respond to those demands remains vital.

New challenges will inevitably arise. When time was of the essence, these organizations looked to ML technology and automation to serve their customers' needs and find new ways to operate. This use of new technology will not only help them respond to the pandemic today, but also set them up to thrive in the future.

Textract is a compelling proposition when used as part of a broader application or in conjunction with other AWS services for digital transformation purposes. And this is where it gets interesting. Seen in isolation, Textract is just a modern twist on traditional OCR technology. But when Textract is seen as a tool to read a document, understand its structure, and extract information from it in the form in which it was meant to be read, that is surely just the starting point for transformation and automation.

The AWS Textract is hence a revolutionary service which will be used by almost every company to innovate and automate their document processing challenges. It can be integrated with several other technologies and bring up new solutions for real world problems.

REFERENCES

1. https://aws.amazon.com/blogs/machine-learning/aws-is-redefining-how-companies-process-documents-in-a-digital-world/
2. https://aws.amazon.com/blogs/machine-learning/announcing-expanded-support-for-extracting-data-from-invoices-and-receipts-using-amazon-textract/
3. https://aws.amazon.com/blogs/machine-learning/tc-energy-builds-an-intelligent-document-processing-workflow-to-process-over-20-million-images-with-amazon-ai/
4. https://aws.amazon.com/blogs/machine-learning/intelligent-governance-of-document-processing-pipelines-for-regulated-industries/
5. https://aws.amazon.com/blogs/machine-learning/process-text-and-images-in-pdf-documents-with-amazon-textract/
6. https://aws.amazon.com/blogs/machine-learning/store-output-in-custom-amazon-s3-bucket-and-encrypt-using-aws-kms-for-multi-page-document-processing-with-amazon-textract/
7. https://aws.amazon.com/blogs/machine-learning/deploying-and-using-the-document-understanding-solution/
8. https://aws.amazon.com/blogs/machine-learning/improved-ocr-and-structured-data-extraction-with-amazon-textract/