Intelligent Document Analysis

A solution to document processing

Ashish Khare, Sr. GTM Specialist, AWS

Corey Peters, Sr. Software Developer, Assent Compliance
Agenda

1. Challenges in processing complex documents
2. How AWS can help
3. The solution to complex document processing
4. How Assent Compliance is using ML for Intelligent Document Processing
Our mission at AWS

Put machine learning in the hands of every developer
The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION
- Amazon Rekognition

SPEECH
- Amazon Polly

TEXT
- Amazon Comprehend
- Amazon Translate

SEARCH
- Amazon Kendra

CHATBOTS
- Amazon Lex

PERSONALIZATION
- Amazon Personalize

FORECASTING
- Amazon Forecast

FRAUD
- Amazon Fraud Detector

DEVELOPMENT
- Amazon CodeGuru

CONTACT CENTERS
- Contact Lens
  For Amazon Connect

ML SERVICES

- Amazon SageMaker
- Ground Truth
- Amazon Marketplace for ML
- AWS ML Studio
- SageMaker Studio IDE

- Built-in algorithms
- Notebooks
- Experiments
- Processing
- Model training & tuning
- Debugger
- Autopilot
- Model hosting
- Model Monitor
- Neo
- Augmented AI

ML FRAMEWORKS & INFRASTRUCTURE

- TensorFlow
- mxnet
- PyTorch
- Gluon
- Keras
- Deeplearn
- Graph Library

- Deep Learning
  AMIs & Containers
- GPUs & CPUs
- Elastic
  Inference
- Inferentia
- FPGA

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Organizations face challenges with processing documents

- Extracting text manually is time-consuming, error prone, and expensive
- Manual processes do not scale easily with documents volume
- Current rules-based systems are not intelligent and break with format changes
- Onboarding new documents is effort-intensive with long lead time
AI Services can solve challenges in document processing

Amazon Textract
Quickly automate document workflows and process millions of document pages in hours

Amazon Comprehend
Uncover insights and relationships in unstructured data

Amazon Augmented AI
Build and manage human reviews for machine learning applications
Amazon Textract

Extract text and data from virtually any document
Benefits of Amazon Textract

- Eliminate manual effort (including template development)
- Extract data quickly and accurately, flexible extraction (extract from variable document types)
- Lower document processing costs (average of 10x savings for OCR workloads)
- No ML Experience Required
is washed by waves, and cooled
Amazon Comprehend

A fully managed and continuously trained service that discovers insights and relationships in text
Amazon Comprehend

Discover insights and relationships in text

Documents
Email, chat, social, phone calls and more

Amazon Comprehend
Automatically extract insights from text

Entities + Custom Entities
Sentiment

Key Phrases
Syntax

Language
Topics

Document Classification + Custom Classifiers

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Comprehend Custom Entity Recognizers

1. Prepare examples
   Annotations or Entity List

<table>
<thead>
<tr>
<th>ENTITY VALUE</th>
<th>ENTITY TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>XT2457</td>
<td>PART</td>
</tr>
<tr>
<td>Renew Subscription</td>
<td>ACTION</td>
</tr>
<tr>
<td>SD5578</td>
<td>PART</td>
</tr>
<tr>
<td>Cancel the order</td>
<td>ACTION</td>
</tr>
</tbody>
</table>

2. Train Comprehend
   - Automated annotation
   - Automated algorithm selection
   - Automated tuning and testing
   - SDK or code-free Console UX

3. Extract Entities
   Hello, my name is John Doe and thank you for calling AnyCompany. I understand you are calling about part number XT2457. We have that part on back order and we are expediting it for you.

   Thank you, but I was expediting it last week. At this point I think we should go ahead and cancel the order entirely.

   We are sorry to hear that sir, would you be willing to complete the order if we offered a 10% discount?

   Yes, thank you.
Amazon Augmented AI

Easily implement human review of machine learning predictions
Benefits with Amazon A2I

- Easily implement human review workflows
- Reduce time to market with pre-built workflows and UIs
- Multiple workforce options
- Integrate with your custom ML models
- Increased worker efficiency
Human Review Workforces

Amazon Mechanical Turk
An on-demand 24x7 workforce of over 500,000 independent contractors worldwide, powered by Amazon Mechanical Turk

Private
A team of workers that you have sourced yourself (your own employees or contractors)

Vendors
AWS Marketplace vendors that specialize in human review tasks
Use case: extract and analyze data from documents

Amazon Textract, Amazon Comprehend and Amazon Augmented AI

- **AMAZON TEXTRACT**
  - Process millions of physical documents, quickly and easily

- **DATA EXTRACTED FROM DOCUMENTS**

- **AMAZON COMPREHEND**
  - Labeling and entity extraction
  - Labeling and routing
  - Sentiment Analysis

- **AMAZON AUGMENTED AI**
  - Outputs that don't require high accuracy are complete
  - Human review for outputs that need high accuracy

**Examples of documents:**
- Financial documents
- Legal documents
- Mortgage Applications
Assent Compliance Use Case
Why did Machine Learning start at Assent?

A Typical Month

Assent receives over 20k documents every month from suppliers responding to information requests:

- Reach/RoHS/Prop65 declarations
- Conflict Mineral Reports Templates
- Customer custom surveys
- Technical Specifications
- Test Reports
- Safety Data Sheets / MSDS
- Chemical Substance Safe Use
- Over 30 other regulatory documents

Assent Insights (Assentia Info-Capture) processes every document received.
An original declaration document submitted by a Supplier in response to a request to certify that parts are REACH compliant.

- Declarations are submitted in many file formats:
  - PDF
  - MS Word and other editors
  - PNG/GIF/JPEG images
  - Facsimiles / Scans
- Various methods:
  - File upload to Assent SUPO website
  - Email attachment
  - Email text
The Results

AI To The Rescue

Assent Insights (Assentia Info-Capture) processes every document received.

1. Logo/Letterhead of Customer
2. Targeted Contact Name
3. Document Date
4. Salutation
5. KeyWord and Phrase Identification
   a. Products
   b. HST
   c. Article
   d. Regulation
   e. REACH/ROHS/etc
   f. Not Subject
   g. SVHC
   h. ECHA
   i. etc...
6. Date of List (DSL)
7. Signature
8. Signature Name and Position
Architecture

1000ft View

Assentia: Architecture Diagram

Assentia Data Sources
- Stack Bucket
- Redis
- Dynamo
- SageMaker
- Textract
- Comprehend

Transform
- Jpeg
- Textract
- Text
- Classify

Assentia Amplify
- API
- Assentia GraphQL
- Amazon Cognito

Assent Data Sources
- Assentia SQL
- GCC
- Zendesk
- SalesForce
- Neptune
- Production Files

Request Queue
- AppSync

Events Queue
- Reach
- RoHS
- Prop65
- RM
- SDS
- Others
- Augmented AI
- HumanLoop
- Website

Analyzers

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Human In The Loop Interface

Assent Participation
Active Learning

Confidence In A Zone

From “Human-in-the-Loop Machine Learning” book by Robert Munro
Active Learning

Random Documents

From “Human-in-the-Loop Machine Learning” book by Robert Munro
Thank you!