

NLP Application

IST-664

Using Amazon's
Textract and
Comprehend to build
an NLP powered
search index



Introduction

Text extraction from a scanned document when it contains formats such as tables, forms, paragraphs, and check boxes can be difficult

Combine that with having to create a ***search index*** that could efficiently sift through millions of documents based on:

key-phrases, language, sentiments or other common elements

Amazon *Textract* and *Comprehend* solve this problem!

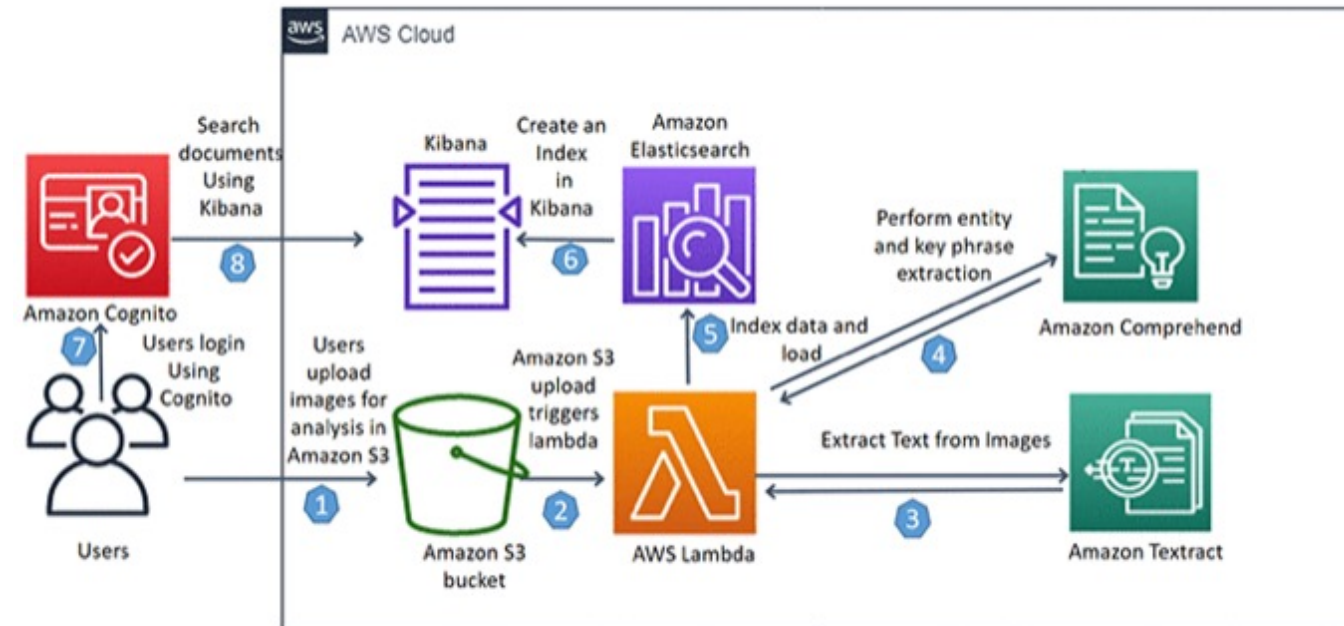
Leveraging Machine-Learning and NLP

Extracting and analyzing text from images or PDFs is a classic machine learning (ML) and natural language processing (NLP) problem.

When extracting the content from a document we want to:

- maintain the overall context
- store the information in a readable and searchable format

Architecture (process pipeline)



The architecture comprises several pieces in Amazon's pipeline:

- *Textract* - machine-learning/deep-learning to extract content from documents & NLP to preserve overall structure
- *Comprehend* – enables an index-based search to extract useful information from documents using NLP
- Elastic search/Kibana
- Cognito
- S3
- Lambda

Comprehend

Uses NLP to scan or parse the content of documents and extract insights

ENTITIES

Provides a list of entities like people, places and locations identified from the document

PII

Capable of extracting personally identifiable information like an individual's address, phone number or bank account

LANGUAGE

Identifies the dominant language in a document

SENTIMENT

Determines the sentiment of a document – Positive, Negative, Neutral or mixed

SYNTAX

Parses each word in the document and determines POS for the word

KEY-PHRASES

Extracts key phrases that appear in a document.

For example, a document about a soccer game might return the names of the teams, the name of the venue, and the final score

Applications and Benefits

Use-Case1: Find documents about a subject (Topic modeling)

Scan a set of documents to determine the topics discussed and to find the documents associated with each topic. Comprehend allows to specify one or more topics and returns corresponding documents

Use-Case2: Find out how customers feel about your products

Comprehend offers a service called *DetectSentiment* where you can send customer feedback and it will tell you whether customers feel positive, negative, neutral or mixed

Use-Case3: Discover what matters to customers

Using Comprehend's topic modeling discover what customers are talking on forums, message boards and then use *Entity detection* to determine people, places and finally apply sentiment analysis to determine how customers feel about the product

Benefits

- Integrate powerful natural language processing into application – Removes complexity of having to build NLP ability in apps and offers service over a simple API
- Deep learning based natural language processing to accurately analyze text
- Scalable natural language processing – Works with millions of documents to discover insights

Demo – *Comprehend* in motion

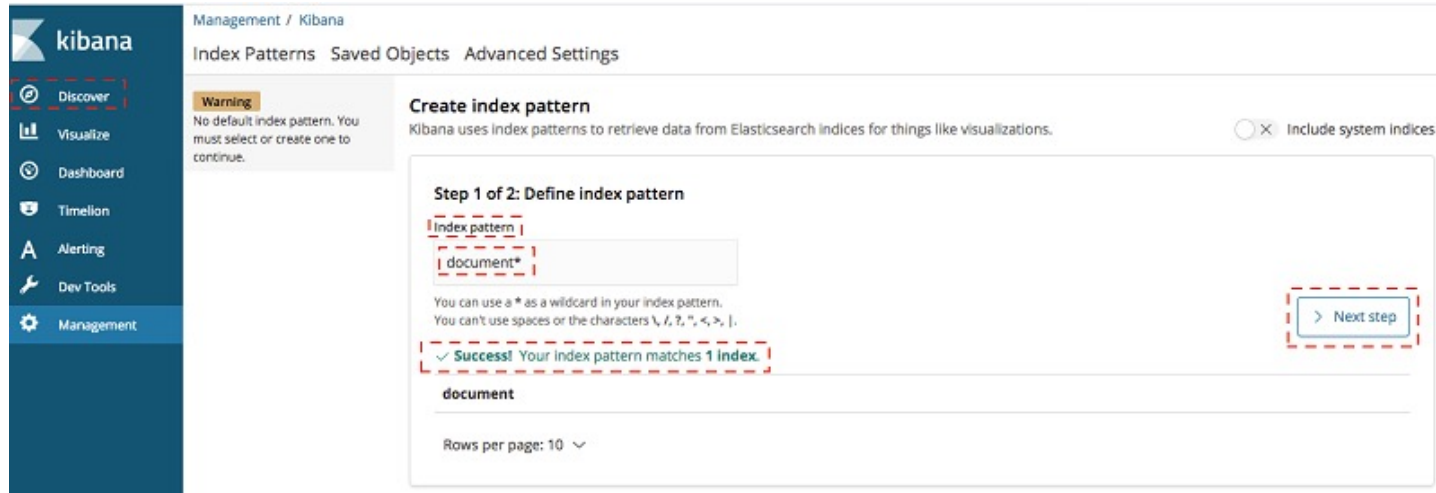


Fig: Search index – Kibana/ES

High-level workflow when you upload a document to Amazon S3 - It triggers a Lambda S3 event that do the following:

1. Extracts text from images using Amazon Textract
2. *Performs key phrase extraction using Amazon Comprehend*
3. Searches text using Amazon ES

```
s3link:s3 location of uploaded documents in S3,  
KeyPhrases: Key phrases from the documents uploaded in S3,  
Entity: it can be DATE, PERSON, ORGANIZATION etc,  
text: raw text from documents,  
table:tables extracted from documents, and forms:form extracted from documents
```

Fetching based on the following document attributes

Demo - Results

The screenshot displays the Kibana web application interface. At the top left, the Kibana logo is visible next to a status bar indicating "6 hrs". A search bar contains the query "Search... (e.g. status:200 AND extension:PHP)". On the right side of the header, there are links for "New", "Save", "Open", "Share", and "Auto-refresh", along with an "Options" menu.

The main navigation sidebar on the left includes icons and labels for "Discover", "Visualize", "Dashboard", "Timeline", "Alerting", "Dev Tools", and "Management". The "Discover" tab is currently selected.

In the center of the screen, a table titled "document*" shows search results. The table has columns for "Entry.LOCATION", "Entry.ORGANIZATION", "Entry.PERSON", "text", "Entry.DATE", and "slink".

- Row 1:** Entry.LOCATION is "-", Entry.ORGANIZATION is "-", Entry.PERSON is "-", text describes extracting data quickly from code/templates, Entry.DATE is "-", and slink points to an Amazon S3 console URL.
- Row 2:** Entry.LOCATION is "Vancouver, BC", Entry.ORGANIZATION is "Amazon.com", Entry.PERSON is "Jeff Bezos", text describes Amazon's history and customer service, Entry.DATE is "July 5th, 1994", and slink points to another Amazon S3 console URL.
- Row 3:** Entry.LOCATION is "Baker N/A", Entry.ORGANIZATION is "Example Corp.", Entry.PERSON is "Jane Doe", text describes employment information, Entry.DATE is "8/15/2013", and slink points to an Amazon S3 console URL.
- Row 4:** Similar to Row 3, with the same data fields.
- Row 5:** Entry.LOCATION is "-", Entry.ORGANIZATION is "-", Entry.PERSON is "-", text describes an expense report, Entry.DATE is "5/13/2013", and slink points to an Amazon S3 console URL.

To the left of the table, a panel titled "Selected Fields" lists fields like "Entry.DATE", "Entry.LOCATION", "Entry.ORGANIZATION", "Entry.PERSON", "slink", "text", and "Available Fields". An "add" button is present at the bottom of this panel. A red dashed box highlights the "add" button and the "text" field in the "Selected Fields" list.

Search results based
on Entities like
location, Org, Date etc.

Search results of
relevant documents
and their attributes

The screenshot shows the Kibana interface with a document search results page. The left sidebar contains navigation links like Home, Visualize, Discover, and Settings. The main area displays a list of search results for a document search. The first result is highlighted, showing fields like _source, _type, and _id. The document content is visible in the right pane, showing a list of documents with fields like name, address, and phone number.

References

Amazon's NLP powered search index project:

<https://aws.amazon.com/blogs/machine-learning/building-an-nlp-powered-search-index-with-amazon-textract-and-amazon-comprehend/>

Introduction to Comprehend:

<https://docs.aws.amazon.com/comprehend/latest/dg/what-is.html>

Introduction to Textract:

<https://aws.amazon.com/textract/>