NLP Application

IST-664

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

Sharat Sripada
MS – Applied Data-science program

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!

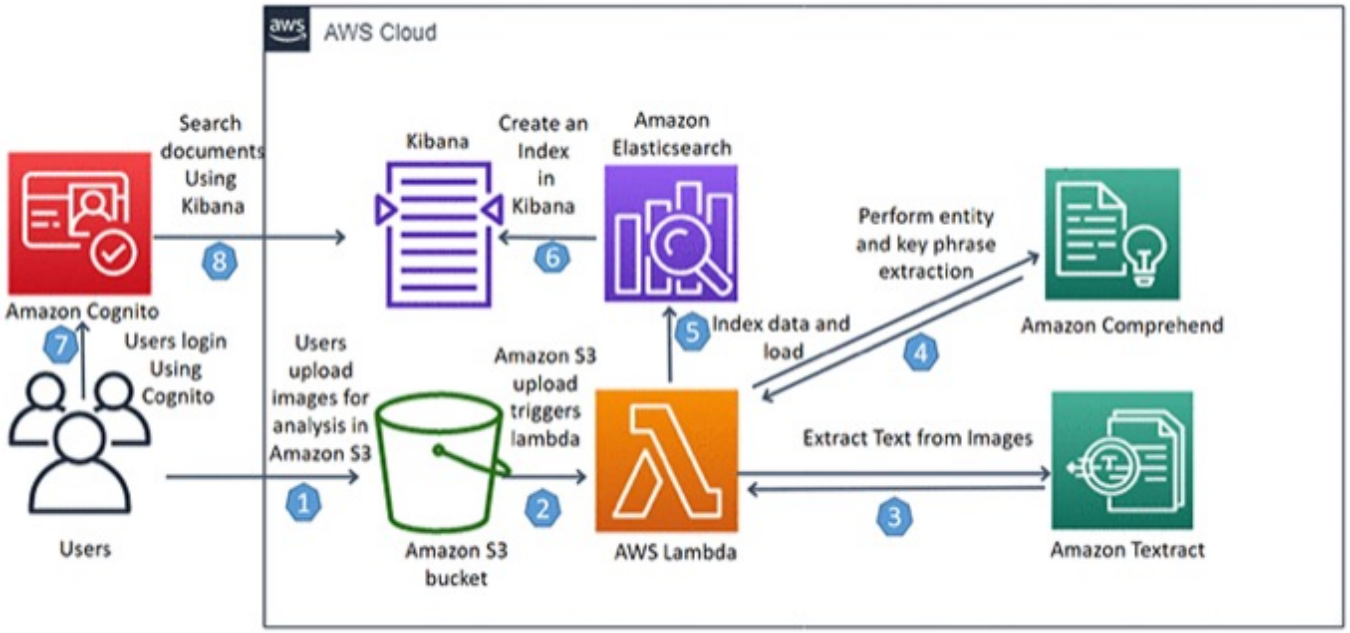
Leveraging Machine-Learning and NLP

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

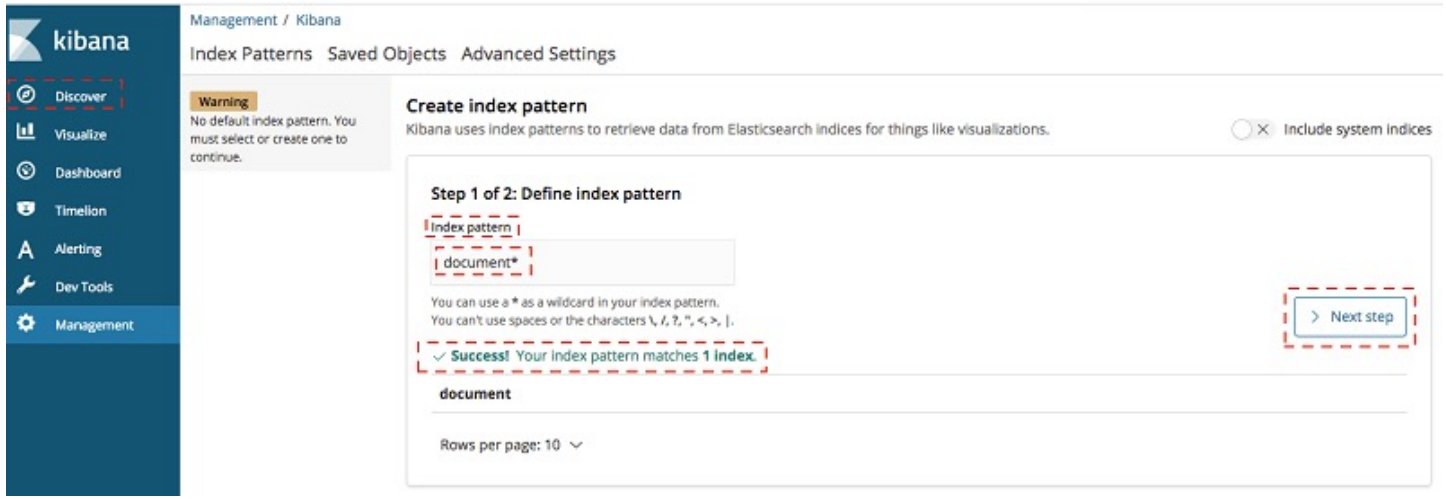


Fig: Search index – Kibana/ES

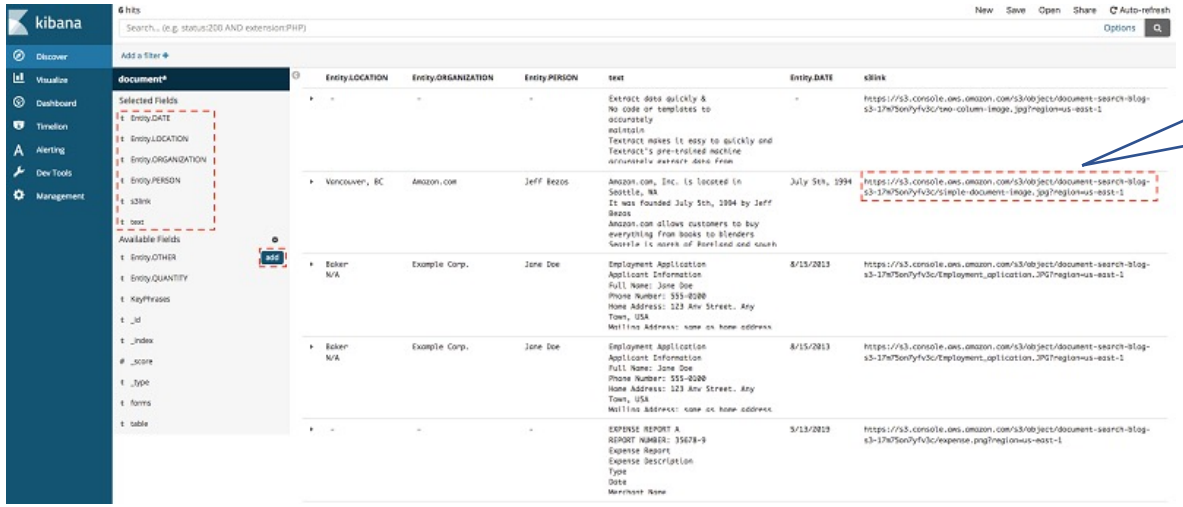
High-level workflow when you upload a document to Amazon S3 - It triggers a Lambda S3 event that do the following:

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
```

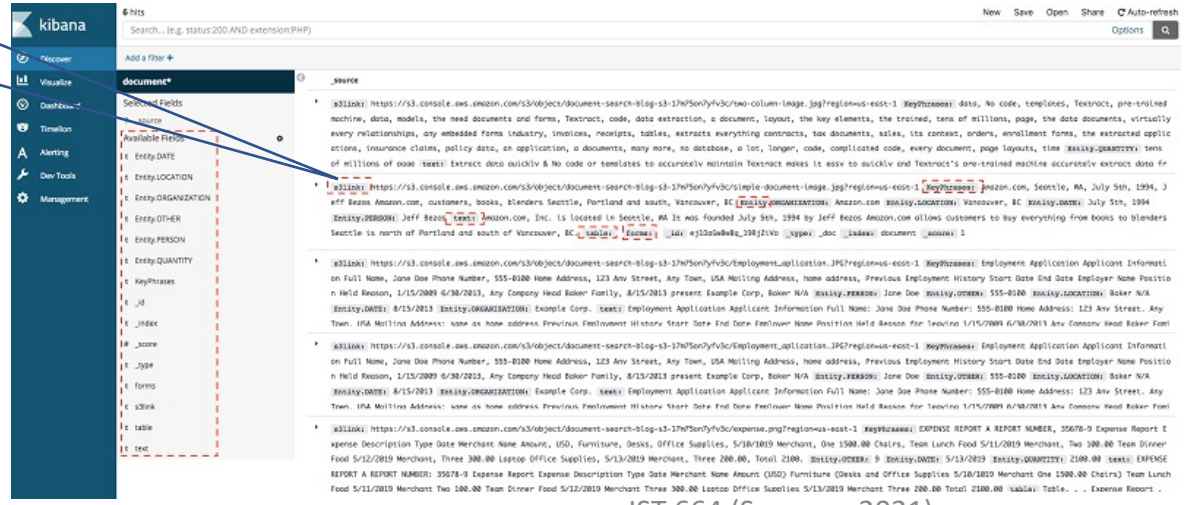
Fetching based on the following document attributes

Demo - Results



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

Search results of relevant documents and their attributes



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/