Technology Review

Amazon Comprehend: How it derives and understands valuable insights from text within documents

Author: ratuls2@illinois.edu

Introduction

In today's fast-paced world, where 2.5 quintillion bytes of data is generated each day, there were 44 zettabytes of data in the world in 2020, which is expected to grow to 175 zettabytes by 2025. Google processes more than 20 petabytes of data every day. This includes around 3.5 billion search queries. With the growing popularity of IoT (Internet of Things) and Social Media, this data creation rate will become even greater.

From 80% to 90% of data generated and collected by organizations today is unstructured, and its volumes are growing rapidly — many times faster than the rate of growth for structured databases. Unstructured data stores contain a wealth of information that can be used to guide business decisions.

With this large growth of data volume, there is a growing need for faster information gathering. Which leads to better decision making, improved experience, solving critical problems and sometimes saving lives.

This topic covers and important tool Amazon Comprehend, provided as an Internet based service by Amazon Web Services (AWS), that helps us to gather insightful information from unstructured data (documents) at scale.

What is Amazon Comprehend

Natural Language Processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language. The goal is a computer capable of understanding the contents of documents, including the contextual nuances of the language within them. It's a great technology and process to understand language and its structure by reading texts at massive scales underpinned by machine learning, which can analyze documents far quicker than human capacity.

Amazon Comprehend is a fully managed and continuously trained NLP service backed by machine learning, which is used to analyze and detect meaningful insights from any text in UTF-8 format (an encoding system for Unicode), or in a semi-structured document, such as a Word doc or a PDF file.

Comprehend falls under the machine learning category of Amazon Web Services (AWS), and it uses a continuously pre-trained model to identify and extract valuable insights from within the text of documents using NLP. It can produce insights and meaningful data, which allows us to use

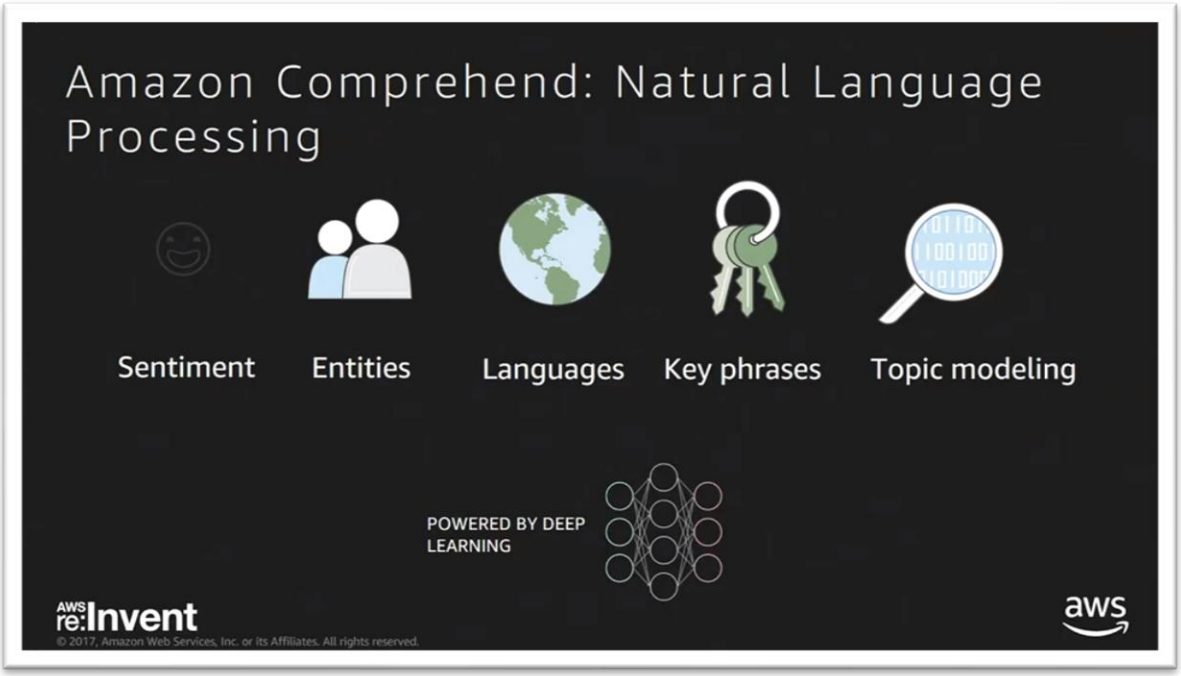
the knowledge to make changes and adjustments to our businesses and capitalize on this valuable data gained. As an example, customer user experience can be enhanced by detecting customer sentiment, which would also allow to determine what actions could lead to the most positive customer experience and outcomes.

How Comprehend Works

Comprehend is fully managed by AWS and is continuously trained. This is the main value proposition by AWS for this product. This reduces the burden of Data Annotation, finding best NLP model and training the model from data scientist and data engineering teams.

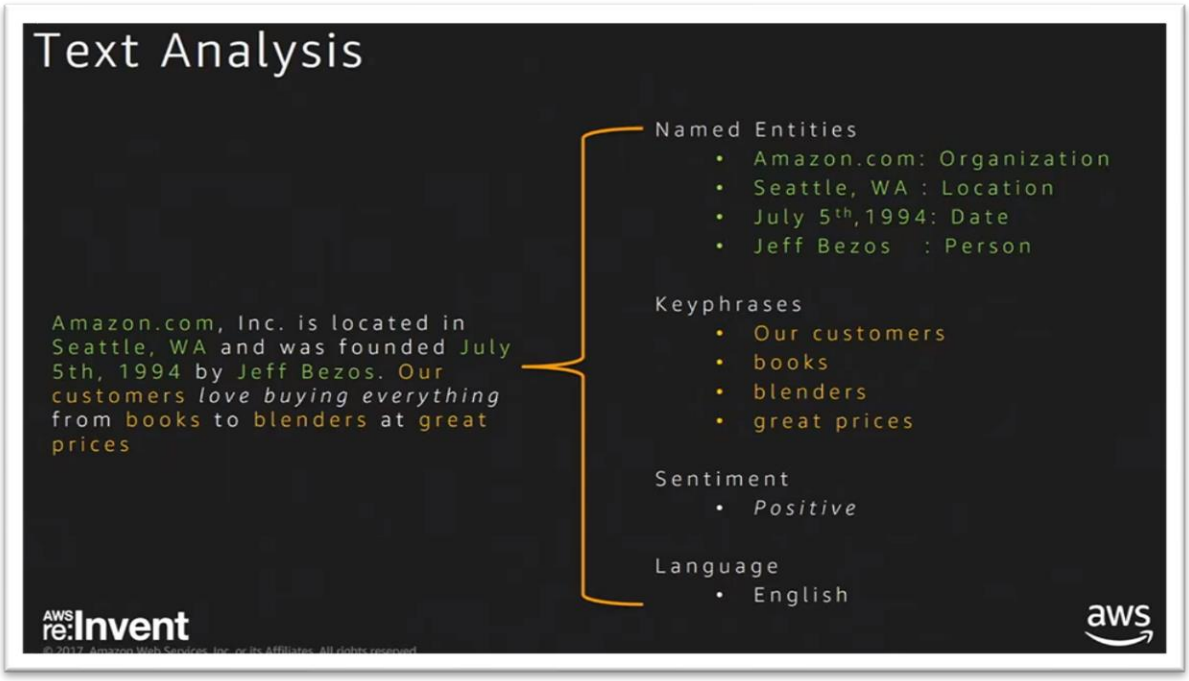
The service has 5 core capabilities.

- Sentiments: understanding positive/negative sentiment from reviews
- Entities: identify entities like organization, place, date, products
- Languages: identify what language the text is presented to (up to 100)
- Key Phrases: identify common noun phrases and their attributes
- Topic Modeling: organize documents into topics



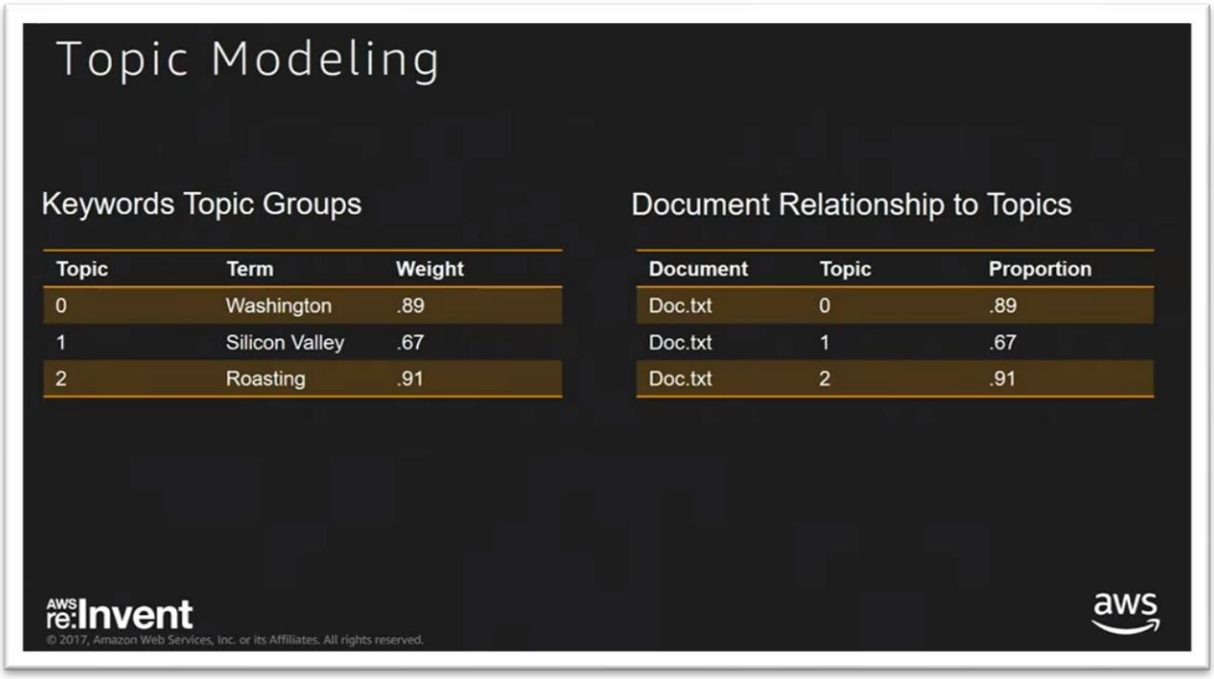
All the above 5 capabilities are built on Deep Learning and packaged nicely into individual services in itself which then integrates with each other to show great insights about the textual data.

A simple example of processing a document with textual information:



Here, the above text is processed, to provide information about the entities attributed as Organization, Location, Data and a Person. Key Phrases identified as books, blenders etc. Overall sentiment of the text was positive when English language was used to present it.

The topic modeling feature, when run, provides 2 views from the corpus of documents provide to the model.



First view contains the list of keyword topic groups (buckets of words) that have been identified from all the documents. Each topic group contains the terms associated to it. This also accompanies with a weight factor for that term associated to that group.

The second view shows which documents are mapped to which topic groups. This view is associated with a proportion factor that tells how strong the association of the topic to the documents inside the corpus is.

Features

This service can scan documents at scale and understand the content in them. It can extract key components of data such as key phrases, entities, and sentiment and more, using a range of different APIs.

Below are some highly sought uses of Comprehend:

- Using NLP service performing over machine learning, to uncover information from unstructured data
- Identify what language a text has (short or long text)
- Textual analysis
- Sentiment analysis - analyze texts, decipher negativity or positivity of language used, understand the expressed feelings
- Key Phrase extraction
- Named Entity identification
- Topic modeling at scale
- Extract medical information from documents

Detailed below are some of the features of Amazon Comprehend:

Key Phrases: A key phrase is a combination of words that contain a noun phrase and is associated with some identifiers. E.g., *her new red dress*. In this noun phrase, dress is noun, new and red are adjectives (attributes of the noun). For every key phrase that is detected by Comprehend, it will issue a score, and this score determines how confident Comprehend is that the string of text being referenced is a noun phrase. One can build a custom application to determine whether the key phrase can be considered.

Sentiment: The sentiment relates to the emotional context of the text. Comprehend will tries to determine the underlying sentiment of the text language. E.g., if a customer review about a product review being scanned was positive, negative, neutral, or even mixed, it will generate a percentage score rating for each of these four emotions to determine the overall sentiment of the comment. This is extremely useful to determine if buyers were generally pleased with the product or not.

Entities: An entity within Comprehend can be described as a reference to a person, a place, an event, a specific date and time, in addition to commercial items and quantities.

E.g., Jack and Jill, went to Aspen, on the Christmas eve, of year 2010. Here, Jack and Jill are people; Aspen is recognized as a location; Christmas eve of 2010 would be seen as a date.

Each of these identified classifications is associated with score that spells Comprehend's confidence of its selection of the text as an entity and its type. Other than name, date-time, event, location, It can also identify commercial item (a brand / product), organization, quantity (amount), title (official name tagged with a work / creation), Personally identifiable information (PII).

Amazon Comprehend uses a large list of PII entities to help identify this data. As PII contains sensitive information, Comprehend can either identify the PII information and classify the PII identity type it has found and present that data, or it can redact the PII data that it has found from within the document. E.g., "Dear Mr. X, your account has a charge of $120 for a purchase at Macys, at Charleston, South Carolina". A copy of your statement has also been emailed to your home address at 123 Main St, Barrington, Illinois.". Here, Mr. X (a name) and 123 Main St, Barrington, Illinois (an address) would be considered PII. With Comprehend, one can redact this information, and can still use the information for processing.

It also can understand a wide range of different languages. And based on the text being analyzed, it can determine which is the most dominant language that the text was written in. A percentage rating is used to determine the confidence level of Comprehend in its understanding of the text language.

Syntax: Comprehend passes each word to determine the syntactic function of the word. Which in turn build up a detailed understanding of the words in the document and their relationship to each other. It does this by classifying each word as a noun, adjective, verb, pronoun etc.

Topic Modeling: This helps determining the different common topics or themes that exist among a large text. E.g., Comprehend would return the topics like time travel, teleportation, telekinesis, aliens, and space travel by processing many Science Fiction stories. Using a specific learning model, Amazon Comprehend can detect and analyze every word, its meaning, and its context. It can detect that if the same word is consistently used in the same context throughout the text, it will be used to determine a topic. Topic modeling is used to help you organize and classify large set of documents into different categories.

<u>Real-Life Use Cases</u>

Below are some real-life use cases where Comprehend has proved itself to be useful

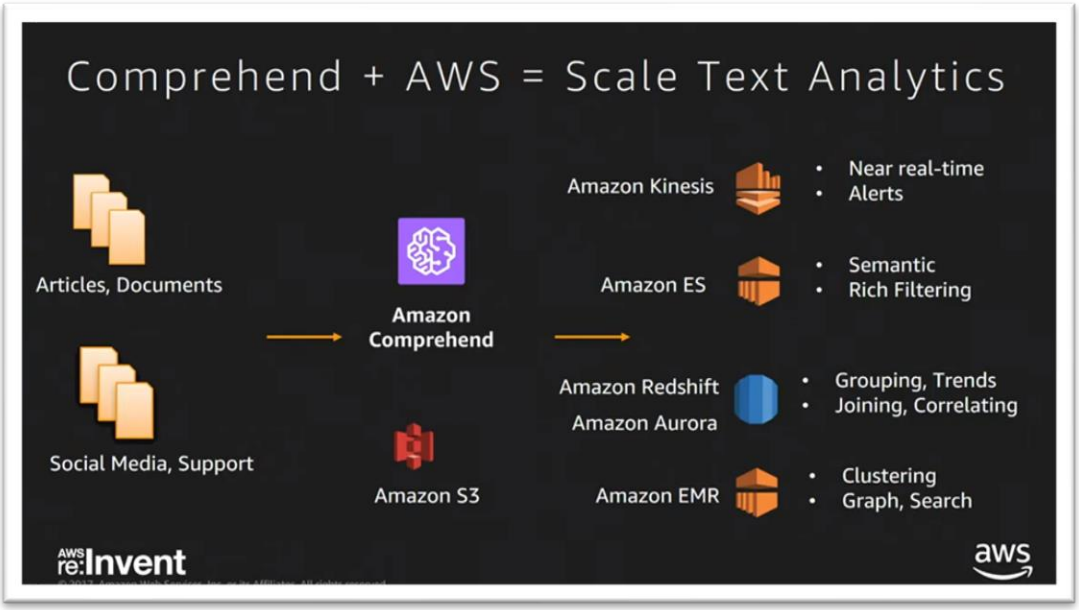
- Analyze customer emails, support tickets, product reviews, social media, and advertising copy gathers insights into customer sentiment that can be put to work for business. This

helps to customize future customer communications. This also helps to identify main features to highlight, and problematic areas to improvement.

- Identify the first update from a customer in a ticketing system to set the future language on the ticket to be used and translate future updates based on that
- Analyze the behavior of the persons in the call and extract sentiments from the utterances.
- Analyze product reviews and determine whether it is positive or negative
- Analyze user feedback submissions to compare their ratings with the sentiment detected from the text. It drives accurately classify and group the feedback. This also allows the product owner to easily and accurately prioritize improvements needed.
- Extract medical information from large set of medical records.
- Recognize Medical Entities (medical condition, treatment, etc.) from a medical text record
- NLP is used to identify potential health risk problems for a patient by examining their historical clinical records resulting in faster diagnosis, quicker treatment decisions, recovery, and rehabilitation

One such example is Elementum. It is a global supply chain company supporting fortune 100 companies, provides a real-time-end-to-end platform that unifies procurement, logistics, manufacturing, and inventory operations.

Their AWS Comprehend solution starts reading news articles across the globe over all supply chain streams; understands the impact of those on global supply chain process; and supply that insight into other systems. The later systems then process that knowledge to actions to proactively recommending of alternate routing when needed that improves time to market, delivers right products on time and monitors actual process performance.



Advantages

Comprehend is recommended for its usefulness on performing various functions with ease:

- Easy to use UI; Fast to build; Rapid deploy; Easy integration with other AWS solutions
- Comprehend Custom allows creating and training new / custom models with new set of training data
- Simple and flexible way to incorporate and integrate a highly sophisticated text analysis tool into existing applications
- Compatible with a range of other AWS services, allowing to seamlessly collate, process, and analyze the textual results at scale; full-stack development possible
- Managed by state-of-the-art access control and encryption methodologies
- PII module manages handling of sensitive and personal information (E.g., Medical Records)
- Analysis at scale: capability to handle millions of documents; fast turnover

Dislikes

- The pre-built models are generic, so if someone is looking to work with data that is very specific, the results would not be encouraging. In this regard, a custom-built model and deploy using AWS Sagemaker would be beneficial.
- The models are basic and don't have all the robust features. There are only few use cases against which this service can be used.
- If someone is processing millions or billions of documents, it might be wise to build a custom NLP pipeline, since price might be steep.

Conclusion

With its simple user interface (UI) and pre-built features, there is a much less learning curve for this AWS service. Within almost 15 mins one can build a model and start using it and generate insights about their data. The data handling is simple, and the outcomes are very precise, reliable. Comprehend also seamlessly integrates with all the Amazon tools like S3, Glue etc. APIs are generous and useful for full stack development. It easily integrates with other tools like Nodejs. Integration of Python API makes it more useful and favorable to programming fraternity.

Though it has its own limitations, but as this a fairly new product, we expect to see more improvements to this offering over time.

References

- Product Overview: https://aws.amazon.com/comprehend/
- Product Reviews: https://www.g2.com/products/amazon-comprehend/reviews
- Product Resources: https://aws.amazon.com/comprehend/resources/
- Product documentation: https://docs.aws.amazon.com/comprehend/
- API Reference: https://docs.aws.amazon.com/comprehend/latest/APIReference/index.html
- Developer Guide: https://docs.aws.amazon.com/comprehend/latest/dg/index.html
- Real Life Customer Use Cases: link here