Analyzing Text on AWS with Amazon Comprehend

INTRODUCING AMAZON COMPREHEND & NATURAL LANGUAGE PROCESSING



Mark Nunnikhoven AWS COMMUNITY HERO @marknca markn.ca

Overview

Amazon Comprehend basics

Key natural language processing concepts

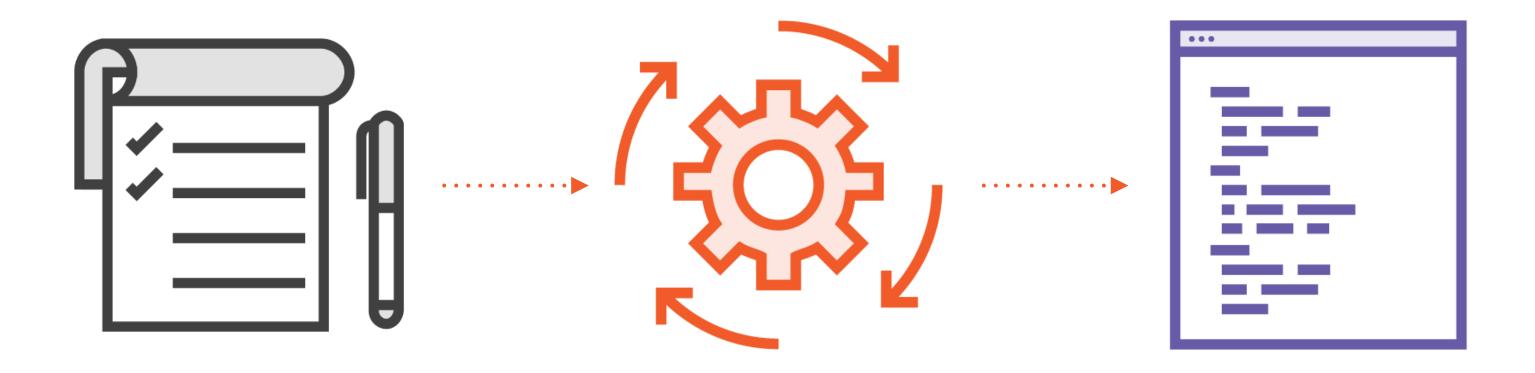
Confidence scores

What is Amazon Comprehend?

Amazon Comprehend

A service that "uses natural language processing (NLP) to extract insights about the content of documents." Given text in a supported language in UTF-8, Amazon Comprehend will provide a number of insights about that text.

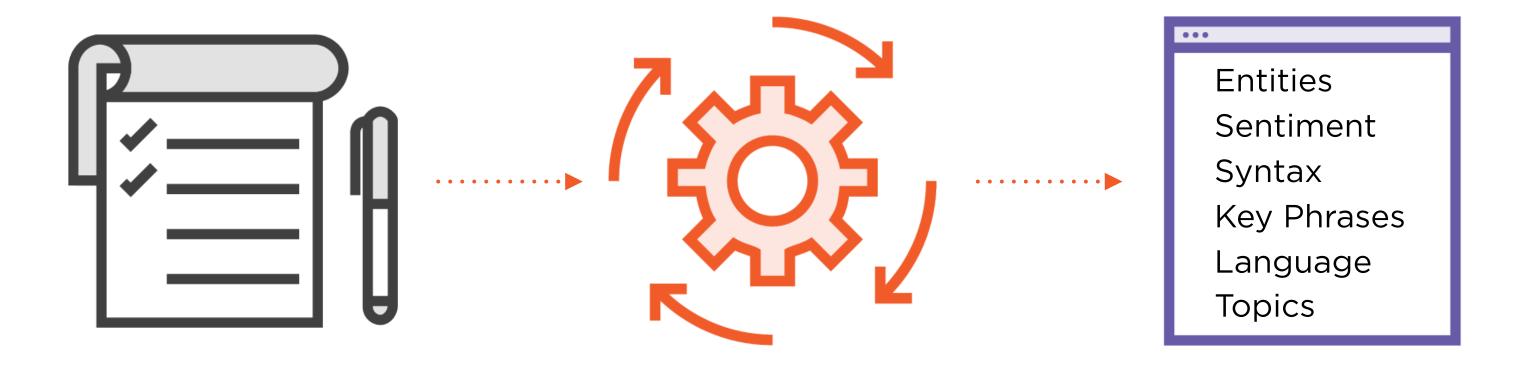
Amazon Comprehend Workflow



Text snippets or documents

Amazon Comprehend Document and language structure and analysis

Amazon Comprehend Workflow



Text snippets or documents

Amazon Comprehend Document and language structure and analysis

English

Spanish

French

German

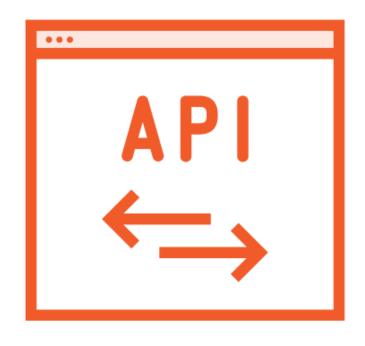
Italian

Portuguese



Document Sources







Web UI

Best for testing and getting acquainted with the service.

API

Handy for small text samples and some workflows.

S3

Most common source.
Best for batches and longer documents.

Suggested Use Cases



Customer analytics



Improving search results



Knowledge management



Classifying support tickets

Amazon Comprehend helps you understand documents

Demo

Introducing the Amazon Comprehend Management Console

Natural Language Processing Basics

Natural Language Processing

Natural language processing (NLP) is a branch of artificial intelligence that focuses on deconstructing human language to help computers better process and analyze language.

The Need for NLP

Turn on the light

English - Canadian

Ouvrir la lumière

French - Canadian

Open the light

English - Nonsense

Understanding the meaning of language is critical to using it successfully



Grammar checkers

Language translation

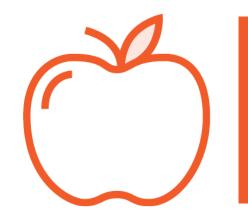
Sentiment analysis

Autocomplete

Related keywords/content

Conversational interfaces

Long History



First work started in the 1950's with various research projects



Initial ideas were to recreate the rules of language in code



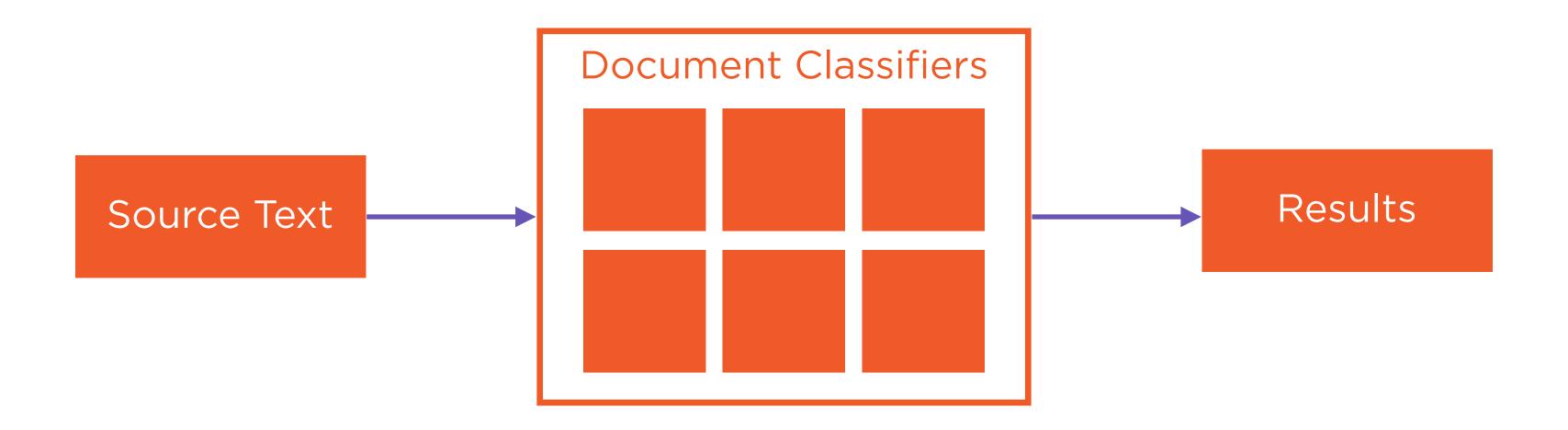
As computer science and artificial intelligence techniques improved, so did natural language processing

As machine learning rapidly evolve, natural language processing became usable

Document Classifier

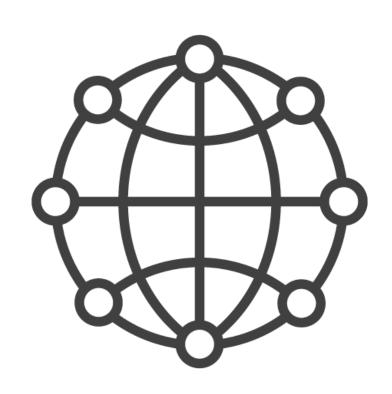
Predicting the classification of various components of a document. For example, predicting—within the provided context—that Amazon refers to a company and not a location.

NLP—In Action



Confidence Scores

Entity Analysis Example



"AWS" is an organization

Isn't it?

99.97115325927734

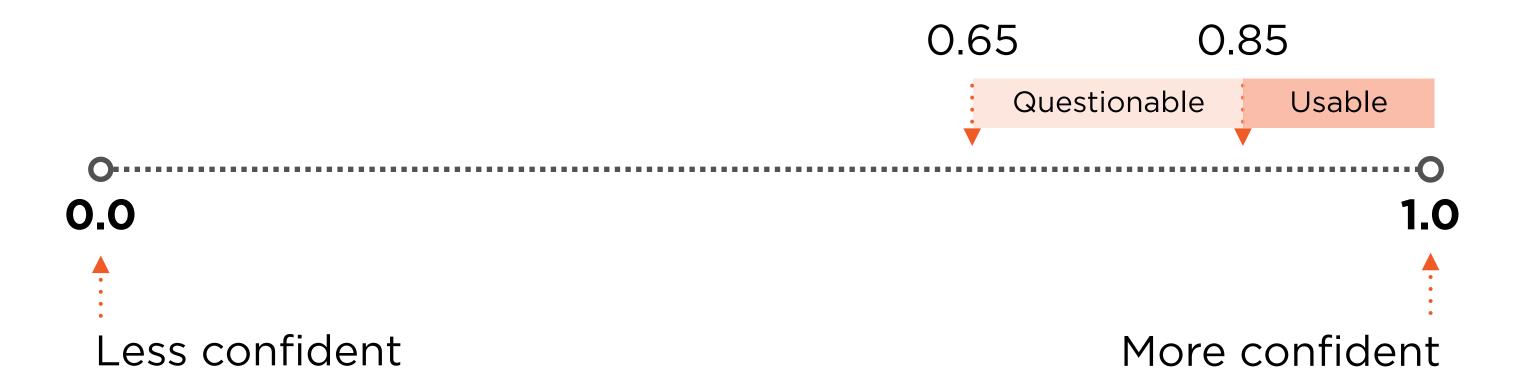
▲::
::

14 significant digits



NLP results are the best *guess* of the given model or classifier used

Confidence Score



Confidence vs. Accuracy

Confidence

Probability that a data point belongs to a specific classification



The classifier determines that the term ["AWS" is an entity of type "organization"] with a confidence score of 99.16326904296875

Accuracy

Whether or not a data point actually belongs to it's predicted classification



The classification that ["AWS" is an entity of type "organization"] has an accuracy of 100.0

Each project will use a confidence score threshold that aligns with that project's risk tolerance

Review



Amazon Comprehend uses natural language processing (NLP) to attempt to discover the meaning and structure of text



NLP uses various classifiers to detect different aspects of the text. Amazon Comprehend offers six different classifiers



Nothing is certain. All classifications in the service come with a confidence score that indicates the likelihood that a determination is correct