unify

Content Insights & Discovery Accelerator (CIDA)

Showcasing the power of MSFT Knowledge Mining and the power of Cognitive Search

Abstract

Keyword-based search engines are obsolete. Enterprises, researchers, journalists, educators, students—everyone spends time searching for information and answers from files, then attempting to discern relationships and insights across document results.

Content Insights & Discovery Accelerator (CIDA) is an intelligent search solution that showcases the power of Microsoft Knowledge Mining and Cognitive Search. Highlighting the power and breadth of Microsoft's Artificial Intelligence offerings, CIDA ingests documents, videos, audio and images in order to index and mine contents to allow users rapidly explore data by searching for information and gaining insight into patterns.

CIDA illustrates the power and ease of utilizing Microsoft's Knowledge Mining, Cognitive Search, AI Platform and data platforms for employing artificial intelligence at scale while handling multiple disparate data sources and forms to achieve enterprise success. We present case studies in document management and archiving, as well as an application in modern investigative journalism.

Introduction

Real-world data is messy and unstructured. This problem becomes even more untenable with the exponential growth of data and the increasing velocity at which data is being created. Mastering the search for information across a diverse set of document mediums presents a significant problem for enterprises, researchers and journalists, among other professionals. A report sent to an email inbox or a shared repository may be easily lost in the deluge of digital documents. Time spent searching for information means time lost in building insights and finding patterns.

Management and enrichment of data are paramount to employees of today, forced to evaluate larger and larger pools of information. Data or documents by themselves have little to no intrinsic meaning or value.

Forms are used for a purpose, then archived for later analysis or review; however, making use of that information for the collective knowledge of the organization is a challenge. Consider how a windmill channels air flows to power society; similarly, a comprehensive document management and enrichment system guides the flow of documents and automates collection and processing so that knowledge mining may begin. The system may propose relationships and ideas among entities and concepts in a series of documents, but ultimately a human being mines that information to home in on pertinent knowledge. Only by combining the power of artificial intelligence (AI) and human creativity can we maximize the return on stored information.

Microsoft's CIDA aims to support journalists by reducing the barriers to entry for document research and analyses. Powered by Microsoft Knowledge Mining and Cognitive Search, the platform enables AI-driven insights and improves search functionality by drawing connections and relationships within the data. Accessible to individuals and organizations alike, CIDA democratizes the investigation into otherwise impenetrable, massive, collections of documents.

Problem Statement: The Need

Microsoft's CIDA is positioned to solve two general problems in document management: archiving and research. While general AI is still decades away, solving a narrow AI document management problem is a straightforward task on Microsoft's Azure platform. Presented here are two case studies: the first focuses on researching disparate document sets in a modern investigative journalism environment, the second on managing archival data.

Problem Statement: The Idea

The core element inherent in both problems is the need to ingest data of any form, sort the document into its hierarchical class, enrich the document with known relationships, and archive those files. Once the files are ingested, the system needs to manage the disparate document types so that the system can, in near real time, search documents for features identified as important.

The modern investigative journalism case study diverges from archival challenges in that the document types will likely be similar, and the goal is to parse them and gain insight under deadline pressure.

The document management side focuses on extracting more precise details from a known quantity set and structure of documents, although the range may be broader due to historical shifts in medium. Here, the system must incorporate more precise and specific algorithms and solutions to identify elements unique to the document archetypes.

While the modern investigative journalism side attempts to predict and extract new features and patterns in a collection of files, the document management study focuses on adding additional structure to the unstructured collection of files. Two similar, yet distinct tasks.

Problem Statement: The Solution

To solve this problem, Microsoft partnered with Unify Consulting to leverage Azure Cognitive Search functionality and produce a platform with the power and extensibility to address the modern knowledge mining problem. CIDA illustrates the amazing integration capabilities of Microsoft Azure Knowledge Mining and Cognitive Services platforms to complex problem domains.

Key Concepts

Knowledge mining is an evolution of traditional research and thought processes. Its precursor data mining focused on the extraction of pertinent information and patterns from large databases, whereas knowledge mining seeks interesting models from data. By models, we mean complex patterns or archetypes, a higher level of understanding than singular facets or singular discoveries. In knowledge mining, the goal is to extract unknown and useful data groupings. In practice, knowledge mining is machine-supported in its handling of value and relevance while being handicapped by the biases that are present in the source documents. It allows for the rapid decomposition of large databases (both structured and unstructured) to draw together disparate information sets to identify interesting interactions.

The evolution of knowledge mining traces back through the evolution of available database technologies.

- When databases were first introduced en masse in the 1960s, technology was too limited to build complex analytics, so analysts did much of their research by hand or in spreadsheets.
- As the relational data model and database management systems entered into vogue in the 1970s and 1980s, analysts could access data more easily but were still constrained in scope, data volume and processing.
- In the 1990s to 2000s, data mining and data warehousing entered enterprise vernacular; not only was data volume growing but so was the ability to conduct analytics at scale rapidly. At this time, analysts focused on automating and expanding the explorations they had been developing and practicing since the invention of the database. These techniques focused on descriptive summary statistics, association rules, discrimination analysis and other traditional statistical analyses.
- Now, in the 2000s to today, with the availability and reduced cost of cloud computing, analysts are exploring not just larger collections of structured data but also incorporating unstructured data into their processes. As such, there's a greater need for a robust knowledge mining platform that can handle advanced data—such as CIDA—than ever before.

To support the evolution into knowledge mining, Microsoft has combined its diverse and comprehensive collection of cognitive services tools into the Cognitive Search toolkit for AI enrichment. Azure Search is the only cloud search service with built-in AI capabilities that enriches all document types. These enrichments are devised from the integration of the entire cognitive skillset pipeline of both prebuilt cognitive skills and custom cognitive skills.

CIDA was built with extensibility as a core concept; thus, the addition of more precise algorithms to tackle specific problems is a cornerstone feature. To solve one of the open questions proposed by the Microsoft News Labs, the R&D and industry advocate team of Microsoft News, an Azure Custom Vision model was integrated to identify document types from a fixed set of document types derived from a study of large judicial and investigative document collections. When a specific feature extraction from a known document set in the enterprise document management system was needed, another Azure-supported deep learning model was integrated to identify key elements in specific document types.

Sponsorship: Microsoft News Labs

Each of these features is supported by the Microsoft News Labs team and a consortium of newsroom partners.

Knowledge Mining on Azure

Solving Unstructured Data Problems

Solving the unstructured data problem in Azure requires building a system to ingest and handle a diverse set of file mediums. This includes documents, images, audio and videos. However, sending the correct file types to the correct pipeline is just the first step. Once a file enters the system, CIDA identifies documents by type and applies specialized skills to specific file types. For example, a document identified as an email or letter gets sent down a specialized correspondence track to run through additional custom skills to extract the sender, recipient and subject line information. By building advanced document type identification algorithms, CIDA is designed for extensibility of custom extractors to expand the capabilities of the search index.

Cognitive Services

Microsoft Azure's Cognitive Services platform is a collection of world-leading algorithms that can be integrated into any application with the call of a function. They provide the power and insights of a specialized deep-learning data science team without the overhead of building custom solutions. The generalizable efficiency of these algorithms to solve standard problems is continually lowering the barrier of entry in solving AI challenges.

Azure Services Integrated Into CIDA		
App configuration	Cognitive Services	Maps
App Service	Custom Vision	Media Services
Application Insights	Event Grid Domain	Power BI
Azure SQL Database	Azure Functions	Search Service
Azure Storage	Logic Apps	Video Indexer

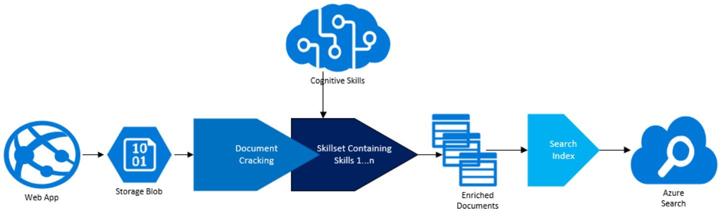
Microsoft has teams of dedicated researchers and data scientists building and improving cognitive service solutions. By leveraging the power of Microsoft Azure Cognitive Services, CIDA can focus data science delivery on solving specialized and unique problems instead of reinventing solutions already solved by Microsoft Azure Cognitive Services. This dramatically reduces the time to insight and delivery, and more time spent generating business insights instead of building custom data science solutions.

Cognitive Search

Microsoft Azure Cognitive Search is the evolution of Azure Cognitive Services platform. While Cognitive Services are a disjointed collection of deep learning algorithms that can be applied to specific problems as needed, the Azure Cognitive Search platform unifies the collection of cognitive skills and presents them as a skillset pipeline. This is then used to extract features and enrich a disparate collection of document types. Azure Cognitive Search's skillset pipeline will extract text from images blobs and other unstructured data sources while simultaneously enriching the data as it flows through the cognitive skills pipeline. As the data passes through each skillset, the data object is enriched with additional information and knowledge.

With text, the skillset pipeline applies natural language processing skills such as entity recognition, language detection, key phrase extraction and sentiment detection, among others, to find entities such as the names of people, places, and things. It may be used to identify locations and highlight synonyms and related phrases to link them. When finding key phrases, it can summarize large documents and segments of the document to identify key topics. All these actions reduce the amount of effort a researcher examining a collection of documents must take to draw parallels and gain insights.

With images, the image processing skills are applied. These skills include skillsets such as optical character recognition (OCR) and the identification of visual features. This leads not only to the extraction of text from images but also the identification of faces, image and object detection, image recognition, orientation and color palette. Any text found can then also be sent through the text processing skillset pipeline to add an additional set of features.



Cognitive Search Pipeline

A preeminent reason for using the Microsoft platform for Cognitive Search is that any updates or improvements generated by specialized talent working to improve Azure Cognitive Services models are then integrated into the application—thereby expanding the power of a data scientist's model-building to anyone who can call the API. This means, rather than having an on-premise team of data scientists rebuilding these solutions, your data scientists can focus on generating direct business value by focusing on the more interesting and specialized problems facing the enterprise.

Video Indexer

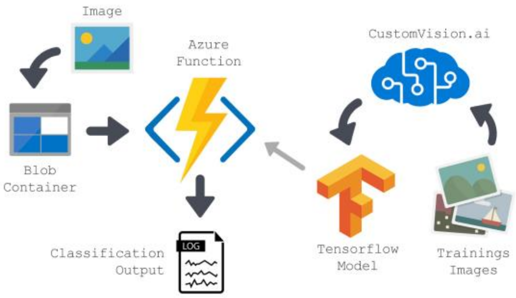
While Cognitive Search addresses static images and text files, Microsoft has added a new cognitive skill in Video Indexer (VI) as part of its Azure Media Services AI solution to address videos. This platform indexes video and extracts insights using Azure machine learning models. It indexes data from the videos based upon the voice, vocals, and videos. Video Indexer can identify emotion, audio effects, language identification and translation, generate transcripts and SRT files, sentiment analysis, unique speaker identification, named entity extraction, keyword extraction, topic modeling, face detection, celebrity identification, best face selection, credits and OCR detection and transcription, scene segmentation, and content moderation. The VI platform is also able to generate video files and clips based upon the ingested data; a skill that allows CIDA to present concise and relevant information to users. These video artifacts are then integrated into the knowledge store database for indexing and searching alongside files of other types for seamless integration within CIDA.

Pipeline Results

The results from the Azure Cognitive Search are split into two storage mediums. The first is the search index which supports the rapid, enriched search functionality of CIDA. The second is the Knowledge Store which contains projections of the data that allow for the retrieval and analysis of the original document with its enriched metadata. Beyond that, the knowledge store is the storage medium that best supports analytics and machine learning on the enriched documents. CIDA utilizes Azure Blob storage for its flexibility and low cost. Storing the Knowledge Store in Azure Blob allows for more advanced analytics using the plethora of tools available there. CIDA's Knowledge Store has been connected to Power BI to support both operational and analytic reporting using Power BI Embedded. The benefit of maintaining the complete knowledge store as opposed to just the final index is the imposed structure on the original documents as well as the ability to recall and serve the entire document. This is important as new skills are added or there is a need to re-index the system.

Azure Data Science Integration

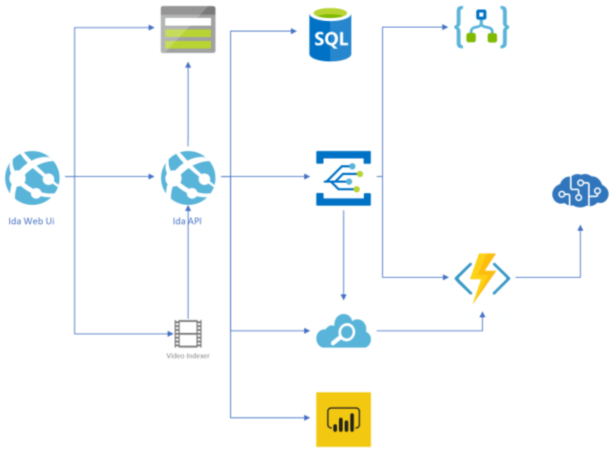
CIDA and the Azure Search functionality are designed for extensibility for increasing the size and granularity of the facets. By utilizing Azure Functions, the system can extend the cognitive skill pipeline by adding additional custom skills to the Azure Search Cognitive Skillset Pipeline. CIDA demonstrates this feature by integrating custom machine-learning models embedded within custom skills to identify document file types, as well as providing custom feature extractors to the pipeline to solve specific archiving needs. These custom skills are integrated directly into the delivery service without the need of creating or maintaining a container instance or machine-learning service to provide a fully scalable delivery system.



CIDA Custom Model Integration

CIDA Architecture

Microsoft's CIDA is designed to leverage Azure Cognitive Search whenever possible; to do that, it follows a typical solution pattern of ingestion of files, enrichment of content via cognitive tools, and exploration via Azure Search. Built as an event-driven platform, all actions pass from API to event grid and actions are taken based upon subscriptions to that grid. These subscriptions trigger actions which process data through the system.



CIDA High Level Architecture

CIDA is required to handle the three Vs of data challenges and opportunities: volume, velocity and variety. To handle the volume of data, the Azure cloud platform is used as it can scale to whatever amount of data submitted to the system. To handle the velocity of files being submitted, an event-driven architecture was built to scale out in parallel whenever possible to avoid bottlenecks in the cognitive skillset pipeline. Finally, the system was built to handle a large array of file types with more document types being integrated regularly.

While building a system to handle data volume, velocity and variety was important, another key component is also how the data is presented back to the users. CIDA secures its data entities at multiple levels: Beyond each instance having access to its own domain and active directory, CIDA controls access to data at a system and dataset level. This allows CIDA administrators to control access for users at various permissions levels.

When using the system, a user will upload, or access sets of documents grouped together as a dataset. These datasets are searchable collections of files that contain sets of documents manually cataloged on upload as an entity of the dataset. The search index is then organized and focused on the dataset level. Users can then create and “save” a set of files in a portfolio. Within the portfolio, users may share, highlight, take notes, make comments and add additional files. The portfolios present the primary suggested use for building a collection of documents that tell a story.

For example, a journalist may upload files containing emails, speeches, and tax records for a group of political candidates as a dataset. That dataset can then be mined for knowledge through the search index: Information or patterns that relate to a specific story or problem can be linked at the portfolio level to share insights across users.

This example extends further in the enterprise archival scenario. Users may upload policy documents relating to specific areas under separate datasets to allow them to be searchable in those contexts; however, if an auditor wants to track a specific policy set and its interaction with other datasets, he or she can pull those threads together through a portfolio view in order to cross the datasets and their data.

Beyond the expansive search capabilities, CIDA is built with native intelligence tools to support research. These include topic mapping, relationship graphs between people, locations, dates and emails, to name a few. This allows users to find related documents or find patterns across multiple documents and entities. These insights are generated to support finding new patterns and relationships between them. For users who want to go beyond the built-in analytics tools, a connector pulls in Power BI to the search index for more complex analytics or analysis.

Applications

Currently, CIDA is being used in investigative journalism scenarios and in building robust and searchable archives; however, this is just the beginning. CIDA infrastructure can be applied to many industry and/or searching application: another last-minute file release from the government before an appointment hearing, a collection of potentially interesting emails, or a subpoenaed collection of randomized documents. Collecting and organizing those files into a searchable index is a first step in shortening the time to insight.

Future applications involve archiving research papers in pharmaceutical research to improve future medical studies or pharmaceutical product tests and archiving a law library to improve the speed and reduce the overhead to finding specific case law and related decisions. Each of these applications are focused on taking collections of knowledge that currently require years of study and memorization and making that information available quicker and with less overhead in order to improve the world we live in.

Conclusions

Microsoft's CIDA aims to provide a structure and blueprint for building intelligent search and knowledge mining capabilities for any organization or industry. The traditional search engine needs a modern tune-up and Microsoft Azure cognitive skillsets are the first step in that process. By generating a complex and specialized skillset pipeline, CIDA supports investigative journalist on tight deadlines, understands massive file repositories, and archives them for use in future stories. Beyond the scope of the current case studies, Azure Cognitive Search and CIDA are applicable in supporting enterprises with data management needs. The ability to automatically enrich and search with natural language is the first step toward a more efficient workforce. By integrating a system that augments the actual content with data insights on relationships and patterns beyond the obvious. Microsoft's partnership with Unify Consulting is propelling journalism into the AI Age.

Contact

Unifyconsulting.com

hello@unifyconsulting.com

8259 122nd Ave NE, Kirkland, WA 98033

206.395.2600