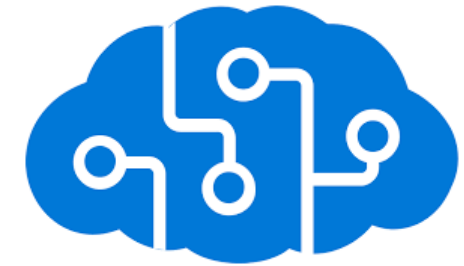


Overview of Azure AI Services and Their Integration with RAG Applications

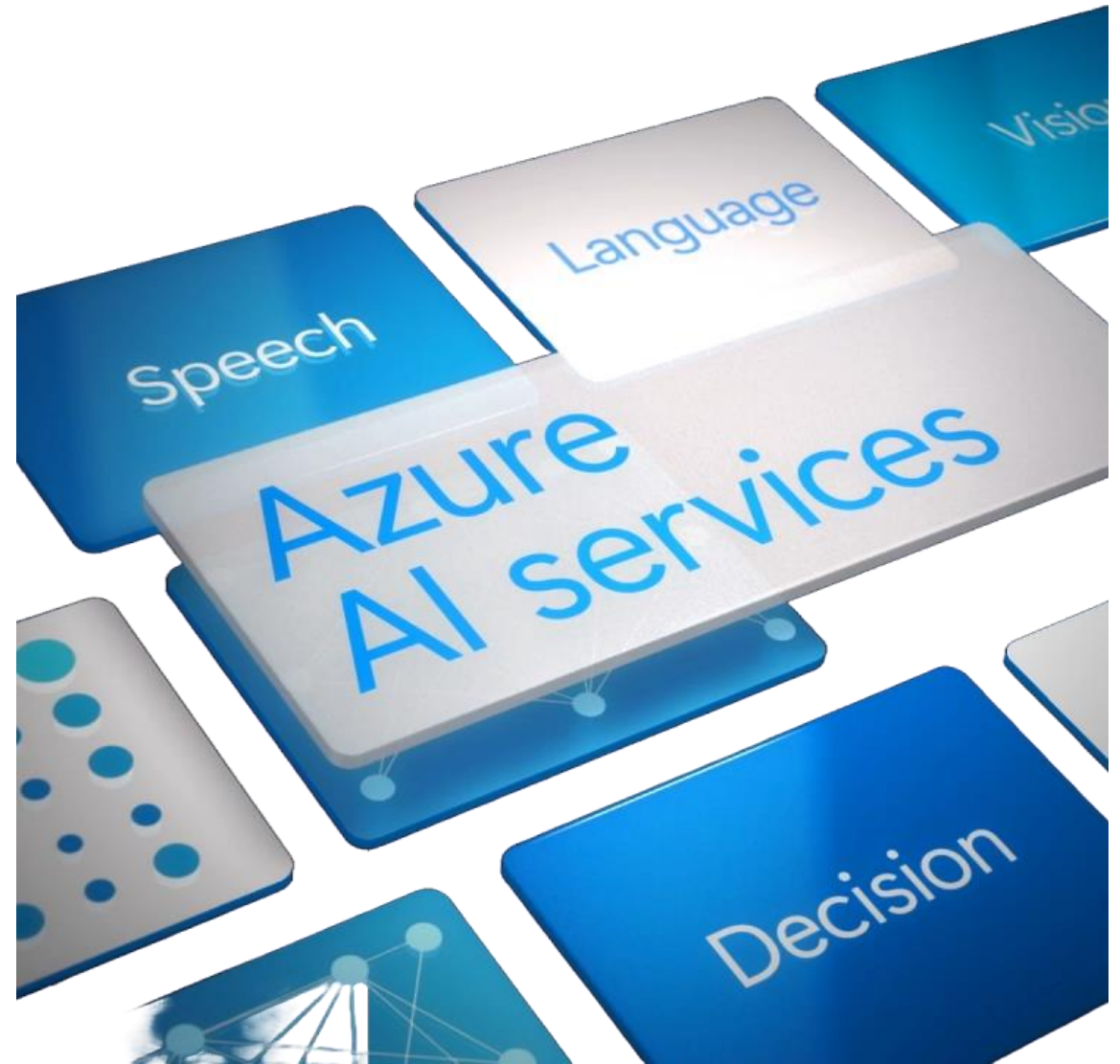
By RDP Team



Data & AI
Azure

Agenda

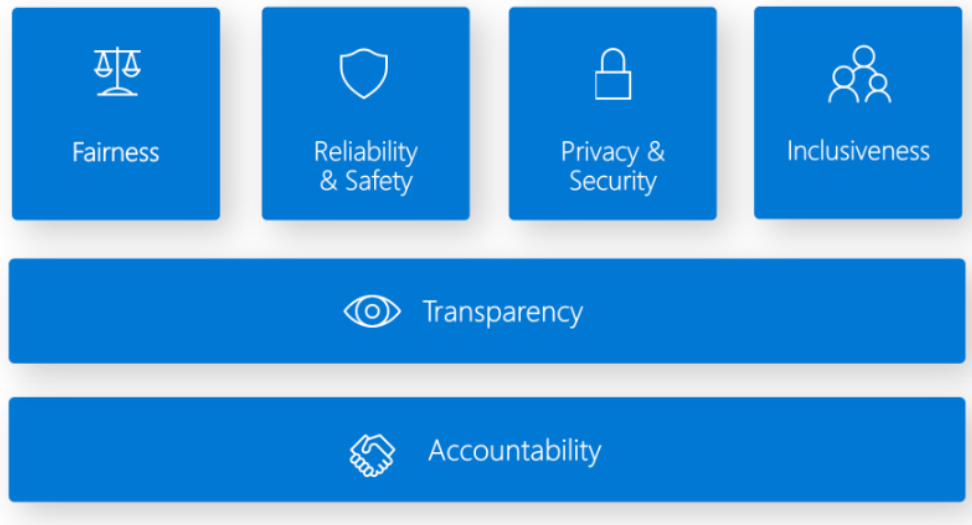
- Introduction to Azure AI Services
- Using Azure Portal with various Azure AI Services
- Code based demo for working with Azure AI Services
- LLM/RAG introduction and benefits of integrating with Azure AI services



Introduction to Azure AI Services

- Prebuilt AI models allow for fast development without the need to train your own models.
- Easily integrate advanced capabilities like natural language processing, computer vision, and speech recognition.
- Quickly deploy scalable and reliable AI-powered applications without custom model training.
- Ensures high standards of security, compliance, and performance with minimal setup.



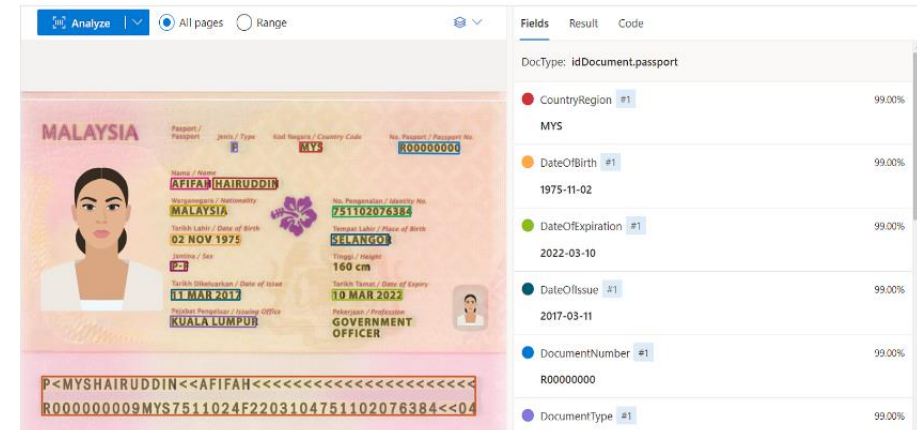


Responsible AI

- Fairness:** AI systems should treat all individuals equitably, avoiding bias or discrimination.
- Reliability and Safety:** AI systems must function consistently and safely across various scenarios.
- Transparency:** AI decisions should be clear and explainable to users and stakeholders.
- Privacy and Security:** AI systems must protect personal data and comply with privacy regulations.
- Inclusiveness:** AI systems should consider diverse users and ensure accessibility for all.
- Accountability:** Developers and organizations must be responsible for AI outcomes and maintain oversight throughout the system's lifecycle.

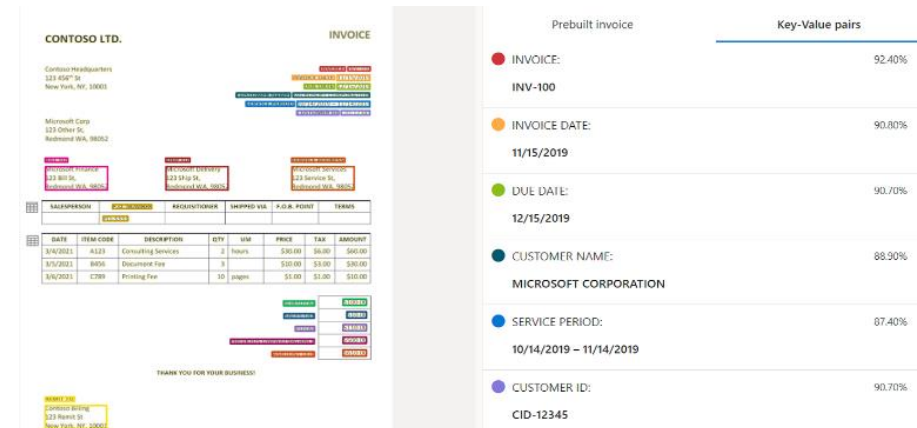
Computer Vision: OCR for Document Digitization

- OCR extracts text from printed and handwritten documents, converting them into digital formats.
- Supports multiple languages, making it ideal for global document digitization.
- Useful for digitizing invoices, receipts, and business cards, improving document management.
- Uses deep learning models to ensure high accuracy and scalability for large projects.



The screenshot shows a Malaysian passport with various fields highlighted by colored boxes. To the right, a table lists the extracted fields and their confidence scores.

Fields	Result	Code
Country/Region #1	MYS	99.00%
DateOfBirth #1	1975-11-02	99.00%
DateOfExpiration #1	2022-03-10	99.00%
DateOfIssue #1	2017-03-11	99.00%
DocumentNumber #1	R00000000	99.00%
DocumentType #1		99.00%



The screenshot shows an invoice from Contoso Ltd. with various fields highlighted by colored boxes. To the right, a table lists the extracted key-value pairs and their confidence scores.

Key-Value pairs	
INVOICE:	92.40%
INV-100	
INVOICE DATE:	90.80%
11/15/2019	
DUE DATE:	90.70%
12/15/2019	
CUSTOMER NAME:	88.90%
MICROSOFT CORPORATION	
SERVICE PERIOD:	87.40%
10/14/2019 – 11/14/2019	
CUSTOMER ID:	90.70%
CID-12345	

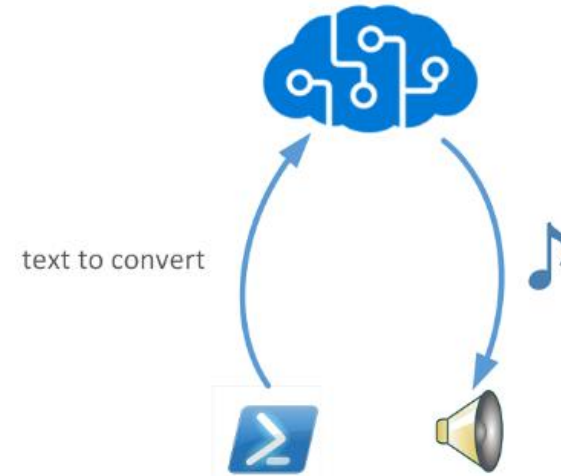
Speech Services: Speech-to-Text and Text-to-Speech

- Speech-to-Text:**

- Converts spoken language to text using advanced recognition technology. Supports multiple languages and accents, ideal for transcription and voice commands.

- Text-to-Speech:**

- Turns text into natural-sounding speech with prebuilt and custom neural voices. Uses deep neural networks for high-quality synthesis and supports SSML for customization.



AI Translator

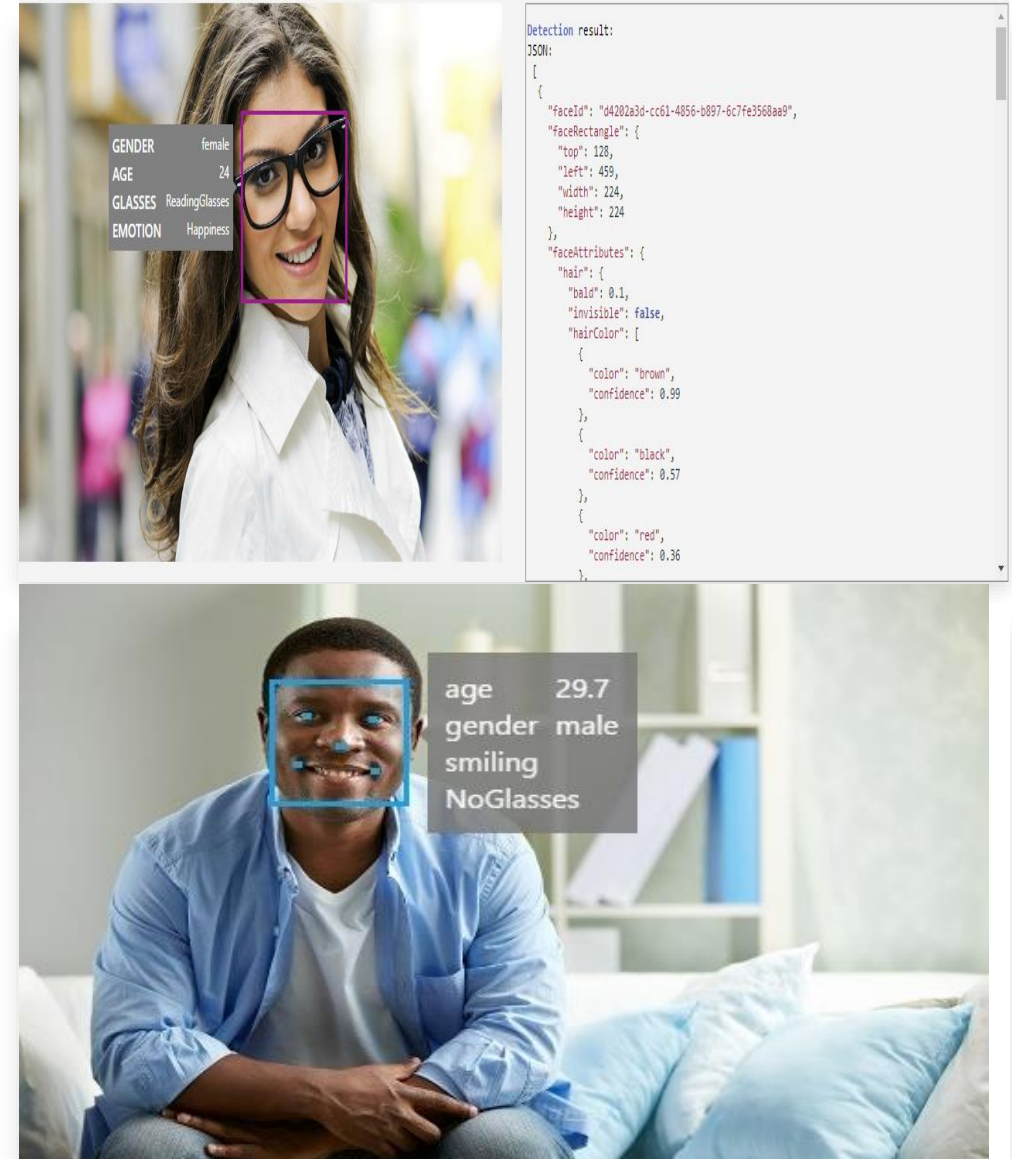
- **Azure Translator** provides real-time translation across multiple languages, making it ideal for global communication and multilingual applications.
- Easily integrates into apps, websites, and services to offer seamless language translation with high accuracy.



From	To
French	English
Où puis-je trouver un bon restaurant/café/la plage/le centre-ville?	Where can I find a good restaurant/café/beach/downtown?
<pre>{ "text": "Où puis-je trouver un bon restaurant/café/la plage/le centre-ville?" }</pre>	<pre>{ "translations": [{ "text": "Where can I find a good restaurant/café/beach/downtown?", "to": "en" }] }</pre>

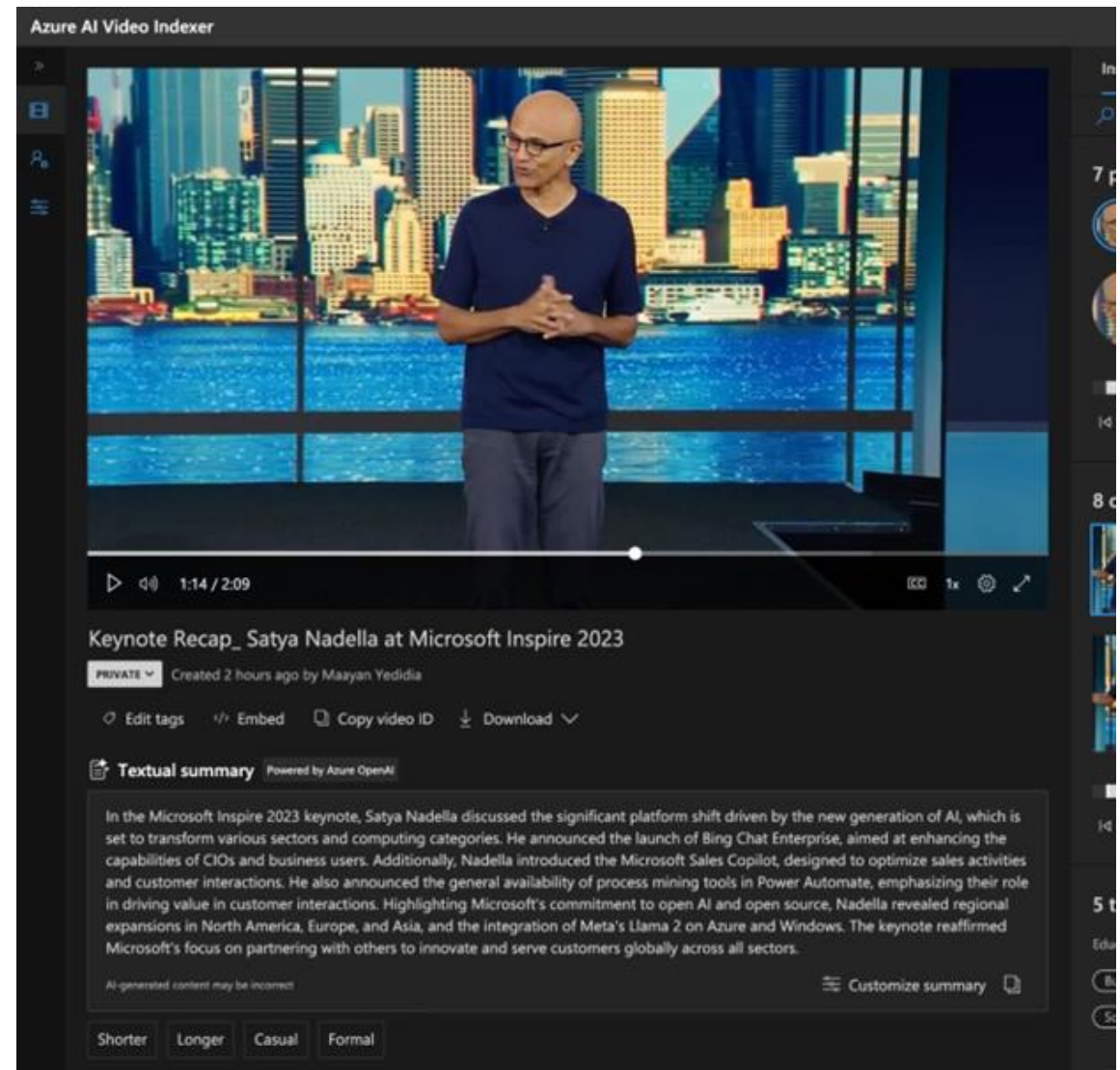
Face API

- **Azure Face API** detects and analyzes human faces in images, providing features like face recognition, emotion detection, and facial attribute extraction.
- It enables applications to perform tasks such as identity verification, emotion analysis, and facial authentication with high accuracy.



Video Indexer

- **Transcribes speech** and translates it into multiple languages with subtitles.
- **Detects objects, faces, and emotions** within video content.
- **Extracts metadata** like keywords, scene segmentation, and sentiment analysis.
- **Performs OCR** to recognize and extract text appearing in videos.



Video Indexer

- By enabling AI features with an Azure OpenAI resource, you unlock additional capabilities such as:
- **Automated Video Summarization:** Generate concise summaries of video content.
- **Enhanced Accuracy:** Benefit from more coherent and human-like summarization and metadata extraction.
- **Customization:** Tailor the tone and length of video summaries to your needs.
- **Broader Use Cases:** Suitable for education, media, and corporate applications.



Unlock generative AI: Connect your Azure Video Indexer resource to Azure OpenAI for more capabilities. [Learn about the benefits of connecting to Azure OpenAI](#)

Azure OpenAI Resource ⓘ

Text Analytics: Key Phrase Extraction and Entity Recognition

- Key Phrase Extraction:**

- Automatically identifies and extracts the main concepts and phrases from unstructured text, helping to summarize and categorize content efficiently.

- Named Entity Recognition (NER):**

- Detects and categorizes entities such as people, places, and dates within text, aiding in the organization and understanding of large volumes of information.

Key phrases

Pike place market, favorite Seattle attraction

Original text

Pike place market is my favorite Seattle attraction.

Key phrase

Key phrase

Date

Entity value: 6/29/2018
Confidence: 100.00%

BorrowerName

Entity value: Parker McLean
Confidence: 100.00%

BorrowerAddress

Entity value: 1234 Main Rd
Confidence: 96.00%

BorrowerCity

Entity value: Frederick
Confidence: 100.00%

BorrowerState

Entity value: Nebraska
Confidence: 100.00%

LenderName

Entity value: Dylan Williams
Confidence: 100.00%

LenderAddress

Entity value: 7890 May Street
Confidence: 100.00%

LenderCity

Entity value: Winchester
Confidence: 100.00%

LenderState

Entity value: Kentucky
Confidence: 100.00%

Original text

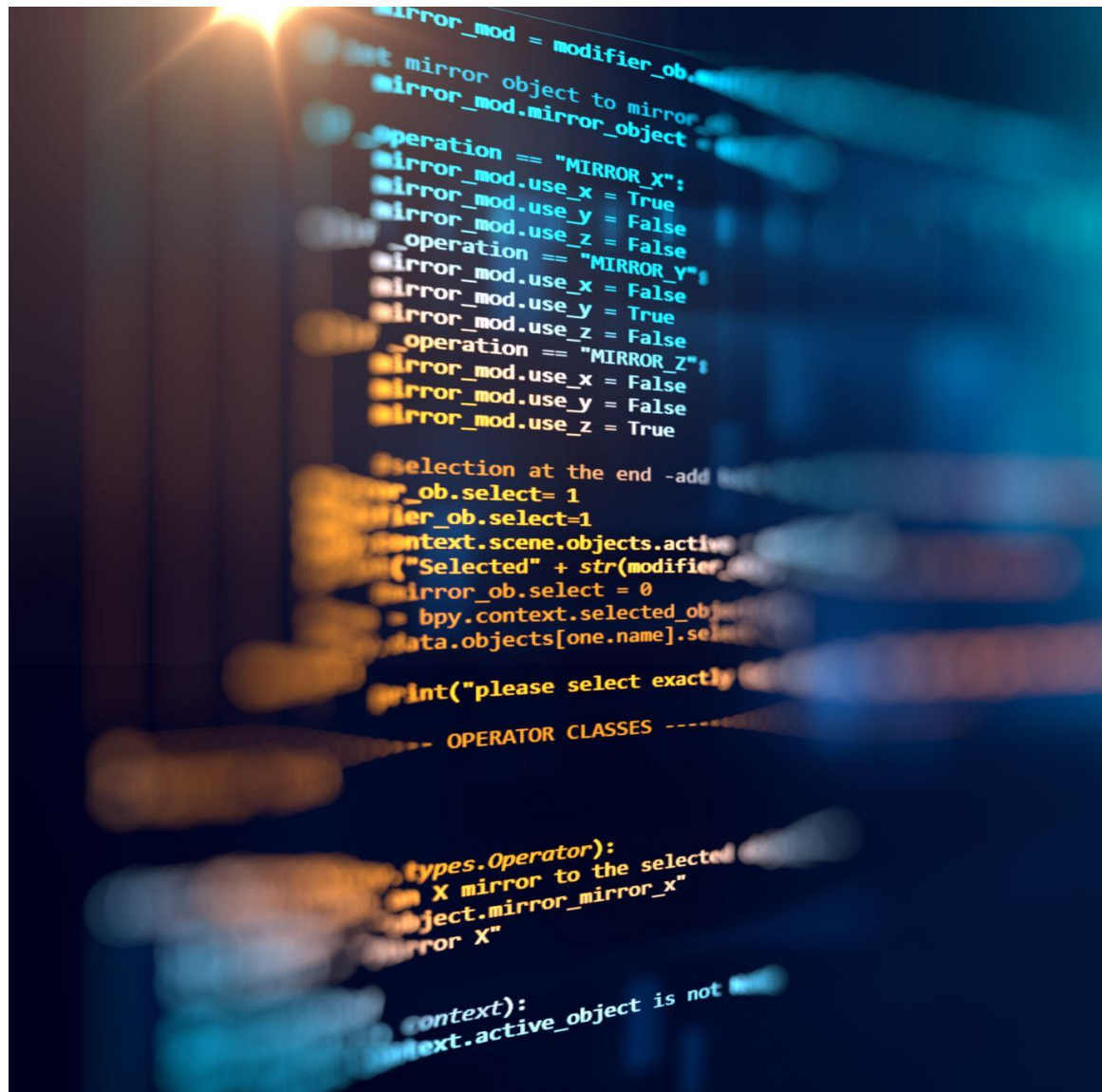
Date 6/29/2018
Date

This is a loan agreement between the two individuals mentioned below in the parties section of the agreement.

I. Parties of agreement

- Parker McLean with a mailing address of 1234 Main Rd, City of Frederick, State of Nebraska, (the "Borrower")
BorrowerName BorrowerAddress BorrowerCity BorrowerState

- Dylan Williams with a mailing address of 7890 May Street, City of Winchester, State of Kentucky (the "Lender")
LenderName LenderAddress LenderCity LenderState

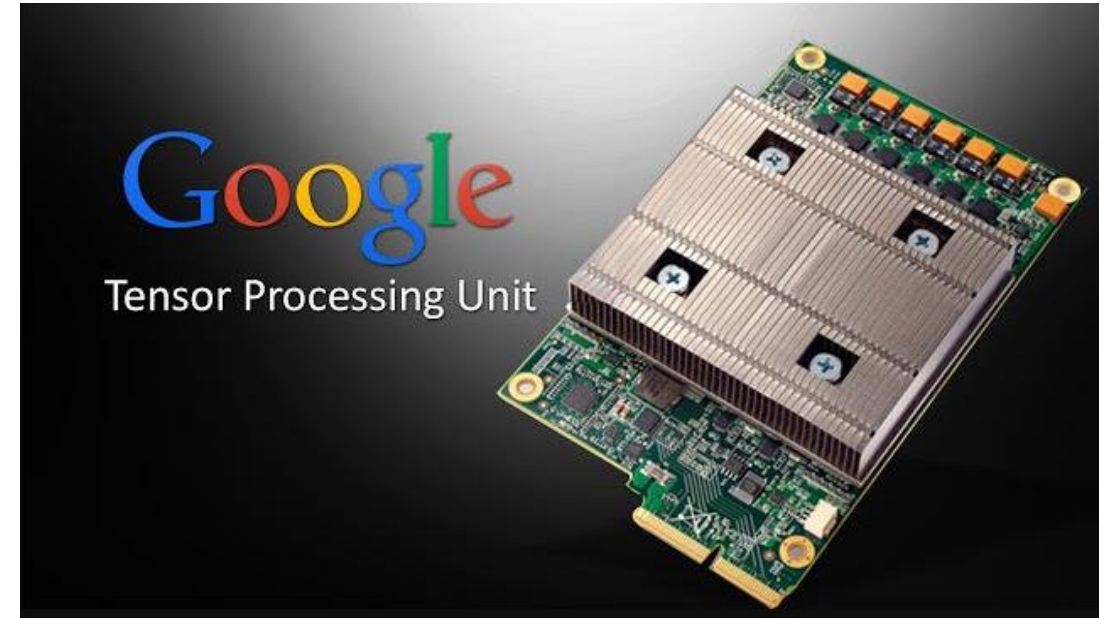


```
mirror_mod = modifier_ob.  
Set mirror object to mirror  
mirror_mod.mirror_object  
operation == "MIRROR_X":  
mirror_mod.use_x = True  
mirror_mod.use_y = False  
mirror_mod.use_z = False  
operation == "MIRROR_Y":  
mirror_mod.use_x = False  
mirror_mod.use_y = True  
mirror_mod.use_z = False  
operation == "MIRROR_Z":  
mirror_mod.use_x = False  
mirror_mod.use_y = False  
mirror_mod.use_z = True  
  
selection at the end -add  
mirror_ob.select= 1  
modifier_ob.select=1  
context.scene.objects.active  
("Selected" + str(modifier_ob.  
mirror_ob.select = 0  
= bpy.context.selected_object  
data.objects[one.name].select  
print("please select exactly  
  
-- OPERATOR CLASSES ----  
  
types.Operator):  
X mirror to the selected  
object.mirror_mirror_x"  
mirror X"  
  
context):  
context.active_object is not
```

Text Analytics API – Python Code Based Demo



PaLM 2	BERT
DALL·E	Google
GPT- 4	OpenAI
LLaMA	Meta
Claude	ANTHROPIC
Dolly	databricks
RedPajama	TOGETHER
MPT- 7B	mosaic ^{ML}



Foundation Models

- It is highly impractical, if not impossible to train your own LLM due to AI expertise needed and the amount of computing power needed.
- It is more practical to use pre-trained models, called Foundation Models from the big tech companies

✓ Response

— Markdown

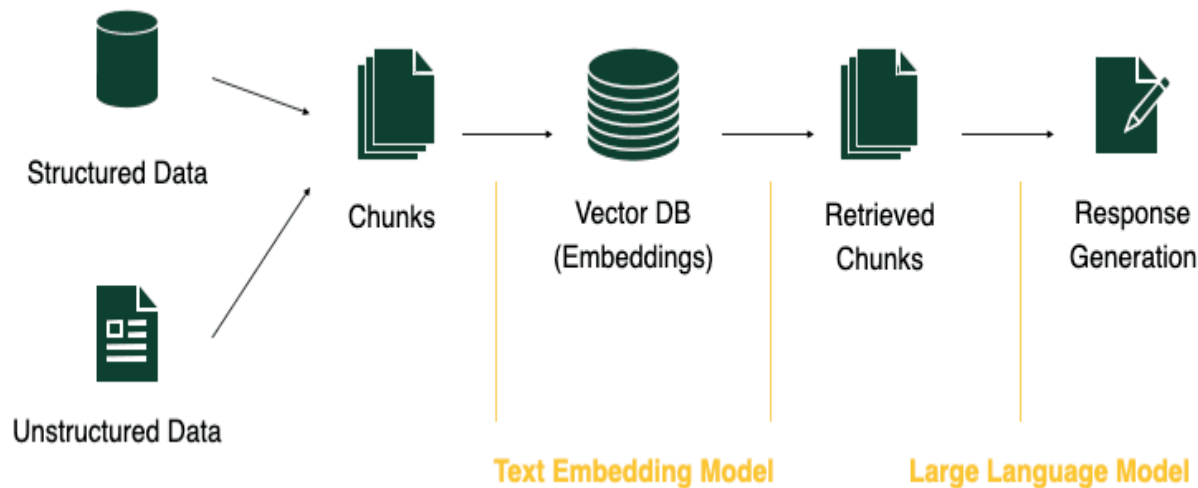
According to the New York Times, Joe Biden won the 2024 Presidential election. Biden received 306 electoral votes, while his opponent, Donald Trump, received 232 electoral votes. Biden also won the popular vote, receiving 74.2 million votes to Trump's 70.5 million votes.

Who was the sole survivor of Titanic?

The sole survivor of the Titanic disaster was Charles Joughin. He was a British baker and the chief baker aboard the Titanic. Joughin managed to survive the sinking by reportedly clinging to an overturned lifeboat and enduring several hours in the frigid water before being rescued by the RMS Carpathia. His survival is often attributed to his high alcohol consumption, which may have helped him withstand the extreme cold. Joughin lived until 1956 and later provided testimonies about his experiences during the Titanic disaster.

LLM Problems - Hallucinations

- LLMs often generate confident answers on topics they weren't trained on, leading to incorrect information (hallucinations).
- RAG helps reduce hallucinations by providing additional, real-time context to ground the LLM's responses.

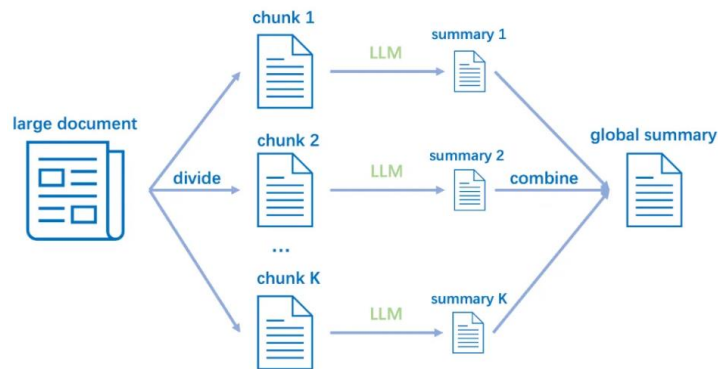


RAG (retrieval-augmented generation)

- **Convert Text to Numbers:** The input text is turned into a list of numbers (a vector) that represents its meaning.
- **Find Similar Text:** This vector is compared to other vectors in a database using cosine similarity to find similar texts.
- **Retrieve Similar Text:** The most similar text from the database is retrieved.
- **Use in LLM:** The retrieved text is added to the large language model's context, improving the accuracy of its response.

Chunking

- Chunking breaks long texts into manageable sections
- A common strategy is to split text by a number of words or sentences
- Chunking strategies are actively being researched



Input Text:

•"Data is becoming a critical asset for businesses. Companies need efficient ways to manage and analyze it. Cloud platforms provide scalable solutions for handling large datasets. Azure offers various services for data storage, processing, and analysis. Machine learning models can extract valuable insights from this data. These insights help drive better decision-making across industries."

•**Chunk 1:** "Data is becoming a critical asset for businesses. Companies need efficient ways to manage and analyze it."

•**Chunk 2:** "Cloud platforms provide scalable solutions for handling large datasets. Azure offers various services for data storage, processing, and analysis."

•**Chunk 3:** "Machine learning models can extract valuable insights from this data. These insights help drive better decision-making across industries."

Vector Representation

- You can turn text into numbers (called vectors) using different methods. In LLMs (Large Language Models), an embedding model converts the text into a vector.

- Example:

- Text: "I love cake"

- Vector: [0.01, 0.02, 0.062, ...]

- Text: "Cats are cute"

- Vector: [0.05, 0.08, 0.09, ...]

- Text: "Let's go hiking"

- [0.03, 0.07, 0.11, ...]

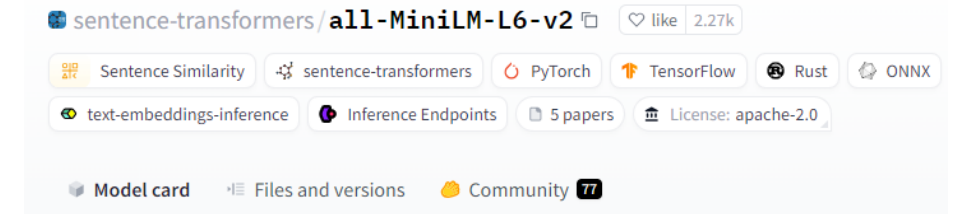
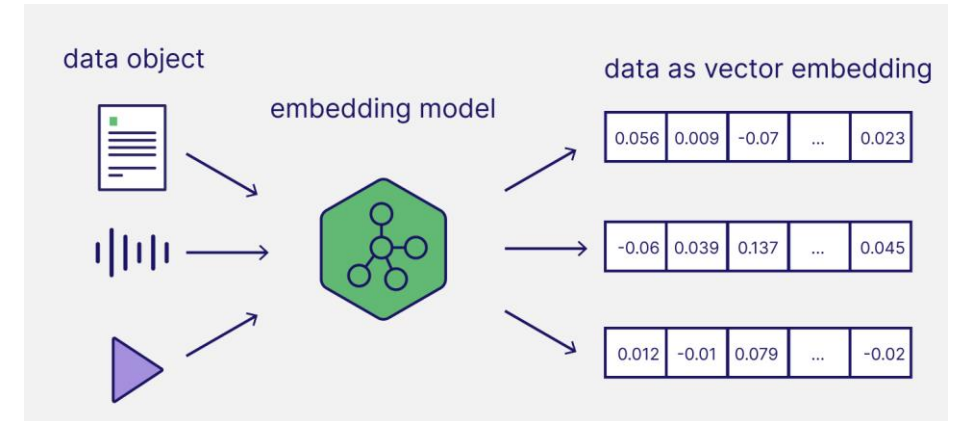
Method	Pros	Cons
TF-IDF	Simple, interpretable	Doesn't capture context, large vectors
Bag of Words	Simple, fast	Ignores grammar and word order
Word2Vec	Captures semantic meaning	Requires large corpus to train well
GloVe	Captures both local and global context	Requires memory for co-occurrence matrix
FastText	Works well with rare words	Slightly more complex than Word2Vec
Doc2Vec	Useful for document-level tasks	More complex to train and fine-tune
LSA	Reduces dimensionality, finds topics	Needs tuning of SVD, may lose some interpretability
BERT	Context-aware embeddings	Computationally intensive

Embedding Model

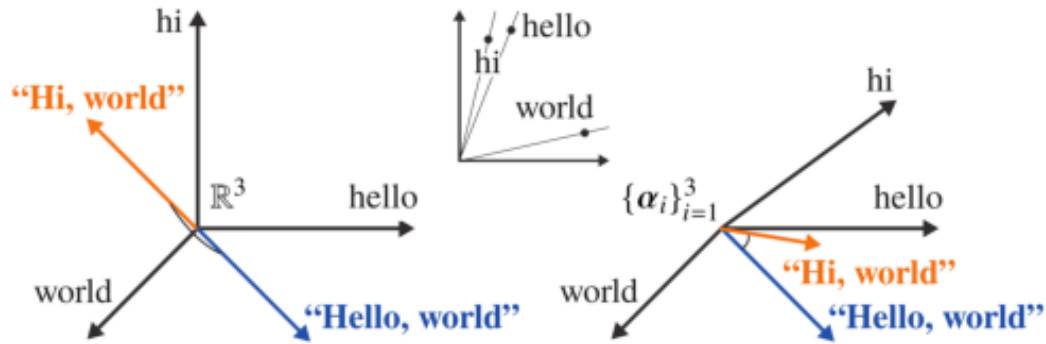
- Preferred Approach for LLMs:** Embedding models are a common method to convert text into vectors for LLMs.

- Open Source Availability:** You can find open-source embedding models on platforms like Hugging Face.

- Versatile Use Cases:** Embedding models are crucial for tasks like semantic search, clustering, and similarity analysis by transforming text into dense vector spaces.



Cosine Similarity



Cosine Similarity shows how similar two sentences are in meaning by turning them into numbers (vectors).

•**Example:**

- "I like cats" and "I like dogs" would have a high similarity because they mean almost the same thing.
- "Johnny Bravo" would have low similarity with both because it's unrelated.

The score goes from **0** (completely different) to **1** (exactly the same).

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Vector A = [2,3,4] Vector B = [5,6,7]

Step 1: Dot Product of A and B

$$\vec{A} \cdot \vec{B} = (2 \times 5) + (3 \times 6) + (4 \times 7) = 10 + 18 + 28 = 56$$

Step 2: Magnitude of A

$$\|\vec{A}\| = \sqrt{(2^2 + 3^2 + 4^2)} = \sqrt{4 + 9 + 16} = \sqrt{29} \approx 5.385$$

Step 3: Magnitude of B

$$\|\vec{B}\| = \sqrt{(5^2 + 6^2 + 7^2)} = \sqrt{25 + 36 + 49} = \sqrt{110} \approx 10.488$$

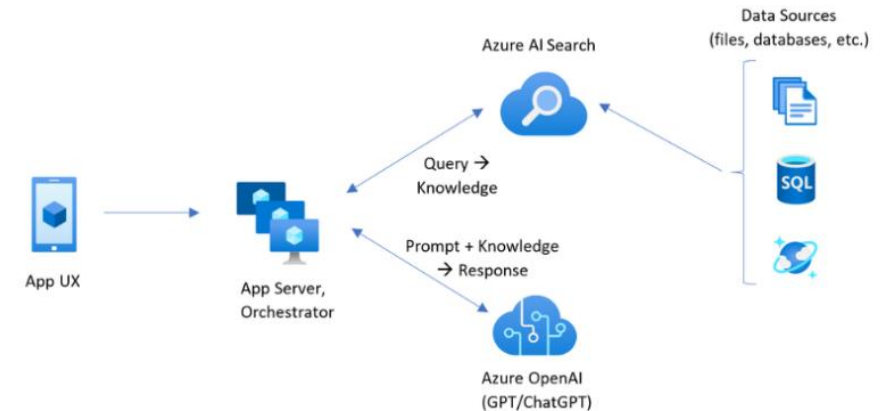
Step 4: Cosine Similarity Calculation

$$\text{Cosine Similarity} = \frac{56}{5.385 \times 10.488} = \frac{56}{56.478} \approx 0.9915$$

Integration with RAG Applications

• Benefits:

- **State of the Art Models:** Microsoft's partnership with OpenAI ensures access to cutting-edge models.
- **All-in-One Platform:** Azure provides LLMs via Azure OpenAI, vector databases with Azure AI Search, and storage with Azure Blob.
- **Seamless Integration:** Use other Azure AI services like Document Intelligence to extract and feed data into your vector databases.
- **Scalable & Secure:** Easily scale RAG systems with integrated, enterprise-ready AI services.
- **Production Ready:** Gain other benefits from the Azure ecosystem such as access control, networking, security, and data encryption.



Azure AI Search

Azure AI Search is a vector store for semantic search.

Supports full-text search and vector embeddings.

Enables k-nearest neighbor (KNN) similarity search.

Integrates with AI-powered features like OCR and entity recognition.



Azure AI Search

- Built-in Embedding:** Azure AI Search offers a built-in embedding model for vectorizing your data.
- Time-saving:** Eliminates the need for managing your own embedding models, reducing overhead.
- Seamless Integration:** Easily connect to Azure services like Blob Storage, Data Lake, and OneLake for data import and processing.

Vectorize your text

Connect to an Azure OpenAI, AI Studio or an Azure AI service and select an embedding model or multi-service account for vector generation. [Learn more](#)

Kind	<div>Azure OpenAI</div>
Subscription *	<div>Stephen-VS</div>
Azure OpenAI service * ⓘ	<div>RDPOpenAI123</div> <div>Create a new Azure OpenAI service</div>
Model deployment * ⓘ	<div>embeddingtest123</div>
Authentication type ⓘ	<div><input checked="" type="radio"/> API key <input type="radio"/> System assigned identity <input type="radio"/> User assigned identity</div>
<div><input type="checkbox"/> I acknowledge that connecting to an Azure OpenAI service will incur additional costs to my account. View pricing</div>	

Search explorer

rdpasearch

Index

vector-1726067000869

Query options

View

specialists best practices to prepare for the behavioral shift towards increased reliance on mobile Web Development Intern Wisconsin Public Televisio...

Search

Results

```
1 {
2   "@odata.context": "https://rdpasearch.search.windows.net/indexes('vector-1726067000869')/$metadata#docs/*",
3   "@search.answers": [],
4   "value": [
5     {
6       "@search.score": 0.03306011110544205,
7       "@search.rerankerScore": 3.193474292755127,
8       "@search.captions": [
9         {
10          "text": "specialists best practices to prepare for the behavioral shift towards increased reliance",
11          "highlights": "<em>specialists best practices to prepare for the behavioral shift towards increase",
12        }
13      ],
14      "chunk_id": "0ee117f73a61_aHR0cHM6Ly9yZHBzdG9yYXd1MTIzNDU2LmJsbnIuY29yZS53aW5kb3dzLm5ldC9yZXN1bWVzLzY1",
15      "parent_id": "aHR0cHM6Ly9yZHBzdG9yYXd1MTIzNDU2LmJsbnIuY29yZS53aW5kb3dzLm5ldC9yZXN1bWVzLzY1Zjk4NTZmYTNh",
16      "chunk": "specialists best practices to prepare for the behavioral shift towards increased reliance on",
17      "title": "65f9856fa3aa7b7a8214fa6b_simple-1.png",
18      "text_vector": [
19        0.0028651152,
20        -0.018100582,
21        0.03197816,
22        -0.032144856,
23        -0.006497734,
24        0.019142443,
25        -0.017586598,
26        -0.016308583,
27        -0.0064873155,
28        -0.02186517,
29        0.02725506,
30        0.0030474407,
31        0.0047057346,
```

Query options

API version

2024-05-01-preview

Vector search

On

Hide vector values in search results

Off

Semantic ranker

On

Semantic configuration

vector-1726067000869-semantic-configuration

Create new

Spell correction (Preview)

Off

Query language

American English (en-...

Close

Azure AI Search

- You can integrate AI Search with AI Services using AI Search using Skillsets
- This lets you leverage the AI services directly before loading your data into AI Search
- Example: There are documents that contained unstructured data and you need to load the data into Azure AI Search for your LLM/AI application. Instead of writing the code for the data extraction, leverage Azure AI Search Skillsets to automatically extract the relevant data and load it into the data store.

Import data ...

^ Add enrichments

Run cognitive skills over a source data field to create additional searchable fields. [Learn about additional skills and extensibility here.](#)

Skillset name * ⓘ

azureblob-skillset

☒ Enable OCR and merge all text into **merged_content** field ⓘ

Source data field *

merged_content

Enrichment granularity level ⓘ

Source field (default)

☐ Enable incremental enrichment ⓘ

Checked items below require a field name.

<input type="checkbox"/> Text Cognitive Skills	Parameter	Field name
<input checked="" type="checkbox"/> Extract people names		people
<input checked="" type="checkbox"/> Extract organization names		organizations
<input checked="" type="checkbox"/> Extract location names		locations
<input checked="" type="checkbox"/> Extract key phrases		keyphrases
<input checked="" type="checkbox"/> Detect language		language
<input type="checkbox"/> Translate text	Target Language <input type="text" value="English"/>	translated_text
<input checked="" type="checkbox"/> Extract personally identifiable information		pii_entities
<input checked="" type="checkbox"/> Image Cognitive Skills		Field name
<input checked="" type="checkbox"/> Generate tags from images		imageTags
<input checked="" type="checkbox"/> Generate captions from images		<input type="text" value="imageCaption"/>



Results

```
1  {
2    "@odata.context": "https://rdpaisearch.search.windows.net/indexes('azureblob-index1')/$metadata#docs(*)",
3    "value": [
4      {
5        "@search.score": 3.231768,
6        "content": "\n",
7        "metadata_storage_path": "aHR0cHM6Ly9yZHBzdG9yYWdlMTIzNDU2LmJsb2IuY29yZS53aW5kb3dzLm5ldC9yZXN1bWVzL210
8        "locations": [
9          "Houston",
10         "TX",
11         "College Station"
12       ],
13       "organizations": [
14         "Loomis Armored US, LLC",
15         "Texas A&M University",
16         "MCSE",
17         "CCNA"
18       ],
19       "people": [
20         "DENNIS",
21         "SCHERRER"
22       ],
23       "masked_text": " ***** WORK EXPERIENCE ***** ***** ***** 2020 -
24       "language": "en",
25       "keyphrases": [
```





Thank You