MNA Maestría en Inteligencia Artificial Aplicada

Materia: TC5035.10 Proyecto Integrador
(Gpo 10)

Prof. Titulares:
Dra. Grettel Barceló Alonso
Dr. Luis Eduardo Falcón Morales

Avance 2. Ingeniería de Características.

EQUIPO 2

Alumnos - Matrícula

Matthias Sibrian Illescas - A01794249

José Ramiro Adán Charles - A00174646

Genaro Ramos Higuera - A00351269

INTRODUCCIÓN

Por la naturaleza del proyecto, consideramos relevante presentar una propuesta inicial de la arquitectura a implementar, con el objetivo de ilustrar de una mejor manera la ingeniería de características.

A dicha propuesta hemos llegado una vez completados los siguientes pasos:

- Sesión con equipo funcional para delimitar expectativas.
- Arquitectura teórica aplicable en base a investigación.
- Pasos para la extracción de características.
- Azure como plataforma de servicios de dicha arquitectura.
- Conclusión.

RESUMEN DE LA SESIÓN CON EQUIPO FUNCIONAL DE LA SOLUCIÓN

El día jueves 9 de mayo 3:00pm tuvimos una reunión con la Funcional de administración de Políticas y Normativa del Tec de Monterrey Lic. Liliana Flores Rosales y con el Ing. Manuel Terán, responsable del área de inteligencia artificial del Tec de Monterrey, en la que nos comentó la Lic. Liliana Flores que le interesa principalmente la búsqueda inteligente de documentos que se encuentran alojados en un servidor sharepoint, el cual se tendrá acceso posiblemente en las siguiente semana para la extracción masiva de los mismos.

Actualmente en el portal https://correosoficiales.tec.mx/ realiza búsquedas muy básicas, que sólo les entregan documentos de manera masiva sin una relación clara sobre el tema en cuestión. Por lo que desea una aplicación que le permita conocer qué documentos (búsqueda inteligente) son los actuales en cuanto a una normativa o tema en cuestión, así mismo solicita como otra tarea a realizar de la aplicación el que se puedan hacer propuestas para las normas, que es una fase más avanzada.

Adicionalmente, cabe mencionar que, en el contexto de nuestra solución, la Ingeniería de Características se ve como un pre-procesamiento adecuado de los documentos y los vectores numéricos que representan su contenido con el fin de generar respuestas más acertadas y precisas. En ese contexto es que se definen los pasos relacionados a la Ingeniería de Características a continuación.

ARQUITECTURA TEÓRICA

La Ingeniería de Características en el contexto de la solución que se creará debe tomar en cuenta el preprocesamiento de la información necesaria para la creación de un contexto útil para que el modelo LLM pueda “interpretar” preguntas de cierta manera abiertas y dar una respuesta acertada en cada caso. Es también importante notar que se pretende que el sistema sea de DOMINIO CERRADO, es decir, limitado a la temática de normativas académicas, con el fin de lograr una eficiencia y precisión altos.

En la siguiente imagen 1 se muestra una arquitectura general:

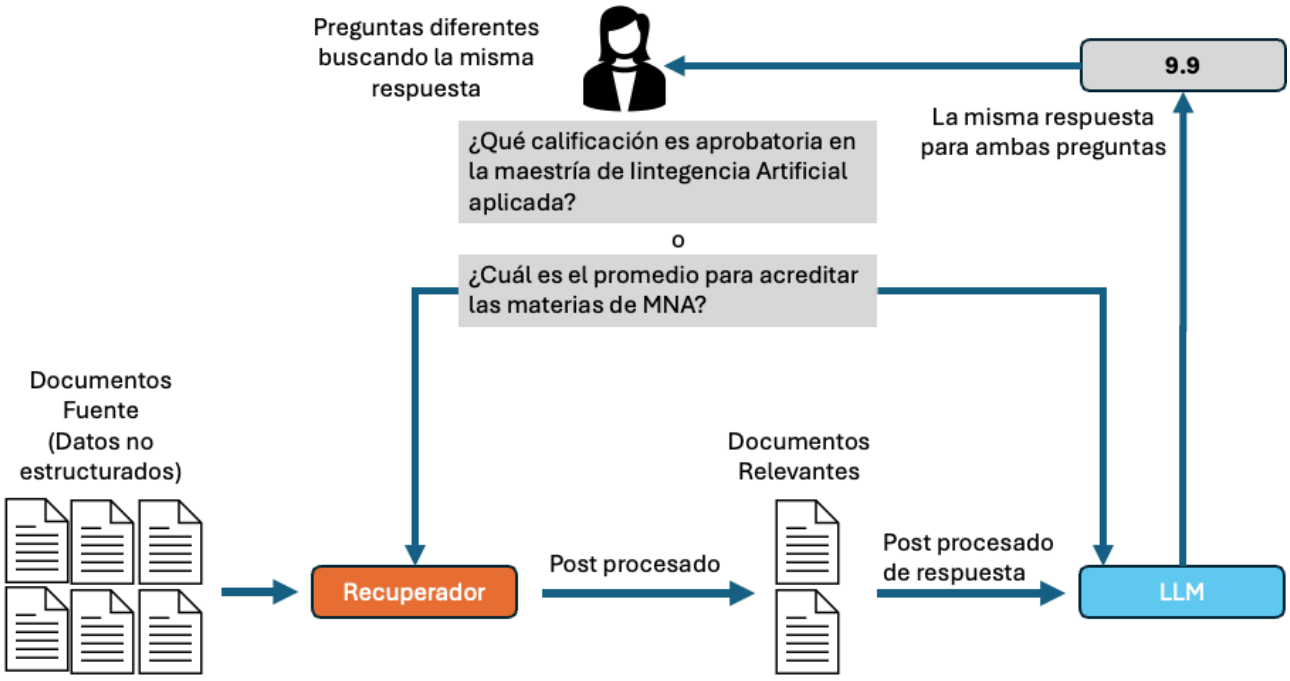

Imagen 1: Arquitectura general para la solución

Usuario: Este elemento representa al usuario funcional que hace una pregunta. Acá, se registra su pregunta en lenguaje natural.

Documentos fuente: Estos documentos son los documentos en formato PDF que provee el Tecnológico de Monterrey relacionados a la comunicación oficial que se desea utilizar para contestar las preguntas del usuario funcional, en el proceso de consulta de políticas antiguas contenidas en una amplia base de documentos institucionales.

Recuperador: Este elemento es responsable de recuperar los documentos relevantes en base a la pregunta realizada por el usuario funcional. Se dividen en dos tipos, esparzo o denso. Si es esparzo, el recuperador usará la frecuencia de términos para representar cada documento y cada pregunta como un vector esparzo.

La relevancia de un documento para una consulta luego se calcula como un producto interno entre los dos vectores. En cambio, si es denso, se usan codificadores como modelos transformadores para representar la consulta y el documento como embeddings contextualizados, los cuales contienen un significado semántico. Esto permite que se mejore la precisión de la búsqueda al entender el contexto completo de la consulta.

Posprocesamiento de documentos: Esta etapa debe ser capaz de generar un repositorio de representaciones vectoriales útiles del cuerpo de documentos relevantes para responder las preguntas del usuario funcional. Debe definirse una manera de representar cada uno de ellos en un registro distinto con una clasificación utilizable.

Large language model (LLM): Este elemento es responsable de extraer una respuesta a la pregunta del usuario funcional, utilizando los documentos relevantes que le proveyó el Recuperador. Usualmente, está compuesto de otro modelo desplegado como tal.

Posprocesamiento de la respuesta: La respuesta debe proveerse en lenguaje que el usuario funcional esté usando. Además, deberá cumplir con reglas de gramática y de ortografía relevantes, así como ajustarse al contexto que sugiere el usuario en su consulta.

PASOS PARA EXTRACCIÓN DE CARACTERÍSTICAS

Para la Extracción de Características, que es la etapa análoga a la Ingeniería de Variables para la solución desarrollada, se cuenta con dos grandes etapas: manipulación del texto y generación de embeddings.

1. **Manipulación del texto**
    a. **Extracción del texto desde los documentos:** En este paso, se extrae el contenido textual relevante de los documentos de origen, para su posterior procesamiento. En el caso de nuestra solución, se extrae de PDFs alojados en el portal oficial.
    b. **Limpieza:** La limpieza implica eliminar ruido y elementos no deseados del texto, como símbolos especiales, números, o palabras irrelevantes, con el objetivo de preparar el texto para pasos posteriores.
    c. **Segmentación de párrafos:** En este paso, se divide el texto en párrafos individuales para identificar unidades de contenido significativas, lo que facilita el análisis y la extracción de características específicas de cada sección.
    d. **Generación de tokens:** Aquí, el texto se descompone en unidades más pequeñas llamadas tokens, que pueden ser palabras individuales o partes de palabras, que omiten palabras conectoras del lenguaje que no representan el mensaje central de la consulta.

2. **Generación de embeddings**
    a. **Representación de vectores numéricos:** En este paso, cada token o palabra del texto se representa como un vector numérico, que captura su significado semántico y contexto en relación con otros tokens. Acá es donde se enfoca la ingeniería de características, pues la representación adecuada de la información es muy importante para representar correctamente los documentos que sirven de contexto y las consultas como tal.
    b. **Incorporación de metadatos:** Además de la representación de tokens, se incorporan metadatos relevantes del documento (como la fecha, autor, tipo de documento, etc.) en la generación de embeddings. Esto mejora el contexto, pues provee información como fechas, clasificaciones del tipo de documento o el autor.

SERVICIOS AZURE

El Tec de Monterrey nos solicita Microsoft Azure Cloud Services para la solución, debido a que tienen prácticamente la gran mayoría de sus servicios de nube en esta plataforma y para garantizar la seguridad de la información.

Con base en investigación bibliográfica y sitios de Microsoft, consideramos que los modelos RAG (Retrieval Augmented Generation) basados en Azure pueden responder preguntas específicas a un contexto determinado (dominio cerrado) sobre dichos documentos, pero requieren un modelo de ingeniería de características adecuado para optimizar su rendimiento, el cual presentamos a continuación a manera de propuesta junto con las herramientas de Microsoft para llevar a cabo la Extracción de Características de forma masiva de la fuente de información de Sharepoint.

1. Extracción de texto PDF:

- **Azure Cognitive Services - Text Analytics API:** Convierte documentos PDF a texto sin formato, elimina metadatos y corrige errores.

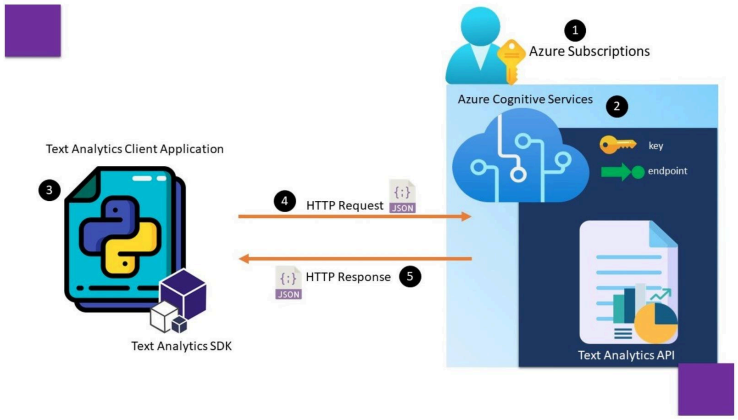

Imagen 2: Conversión de documentos usando Azure Cognitive Services (Gupta, 2021)

- **Azure Storage Blob:** Almacena los documentos PDF y el texto extraído de forma segura y escalable.

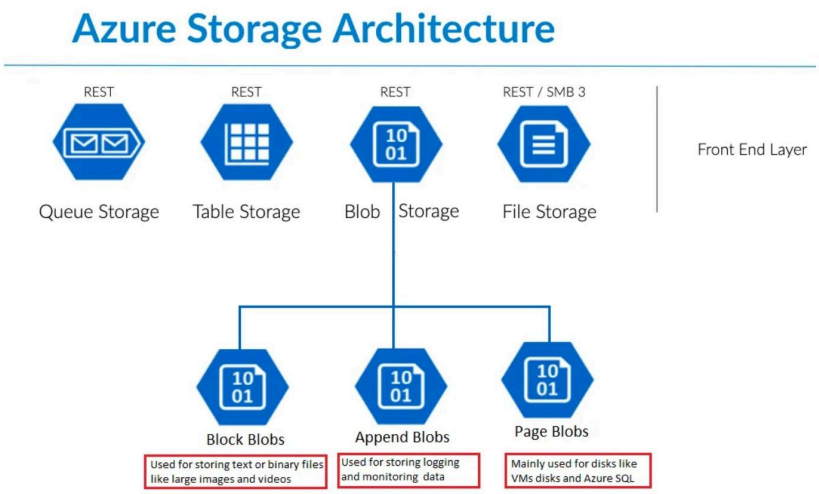

Imagen 3: Arquitectura de Azure Storage (Verma, 2022)

2. Extracción de características léxicas:

- **Azure Cognitive Services - Text Analytics API:** Identifica palabras clave, términos relevantes y calcula la frecuencia de palabras.
- **Azure Machine Learning:** Crea vectores de características TF-IDF (Term Frequency-Inverse Document Frequency) para representar cada documento.

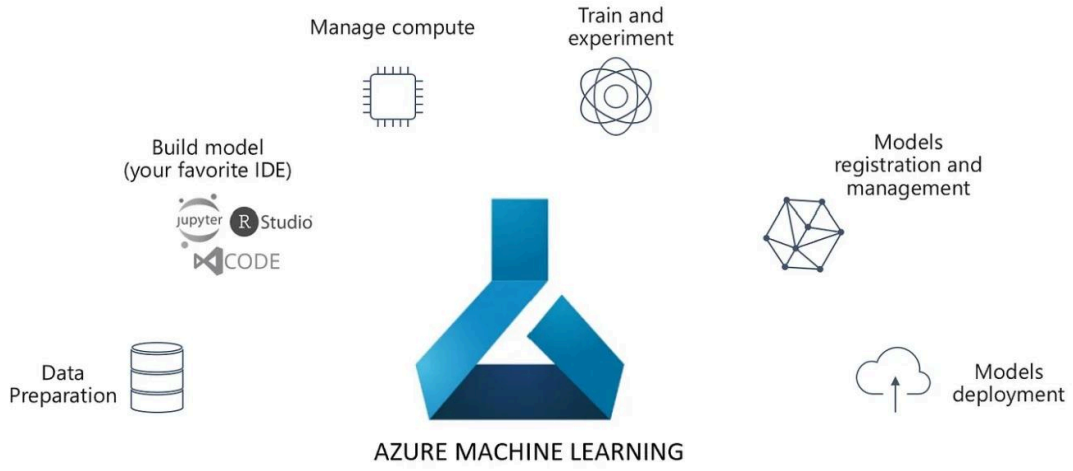

Imagen 4: Características de Azure Machine Learning (Alto, 2022)

3. Extracción de características contextuales:

- **Azure Cognitive Services - Language Understanding API:** Identifica la entidad objetivo y extrae el contexto de la pregunta.

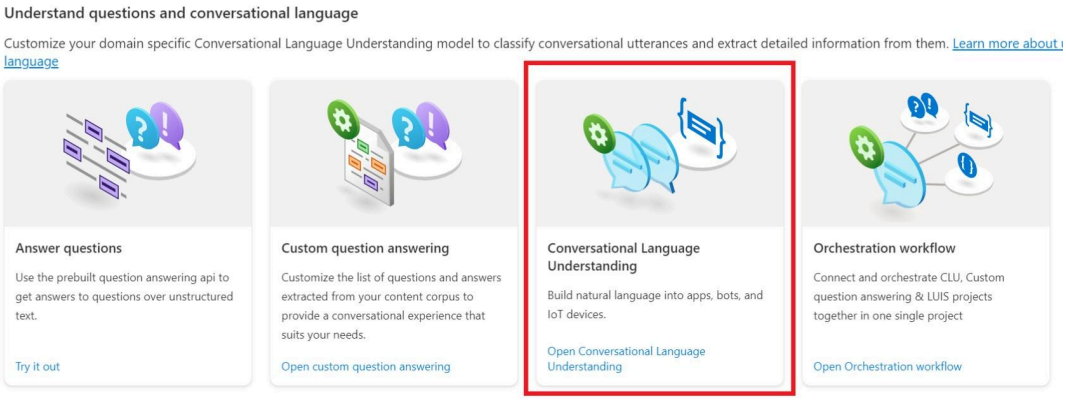

Imagen 5: Herramientas para la comprensión del lenguaje conversacional de Azure (Jboback, 2023)

- **Azure Knowledge Graph:** Extrae entidades mencionadas en los documentos PDF y crea vectores de representación para ellas.

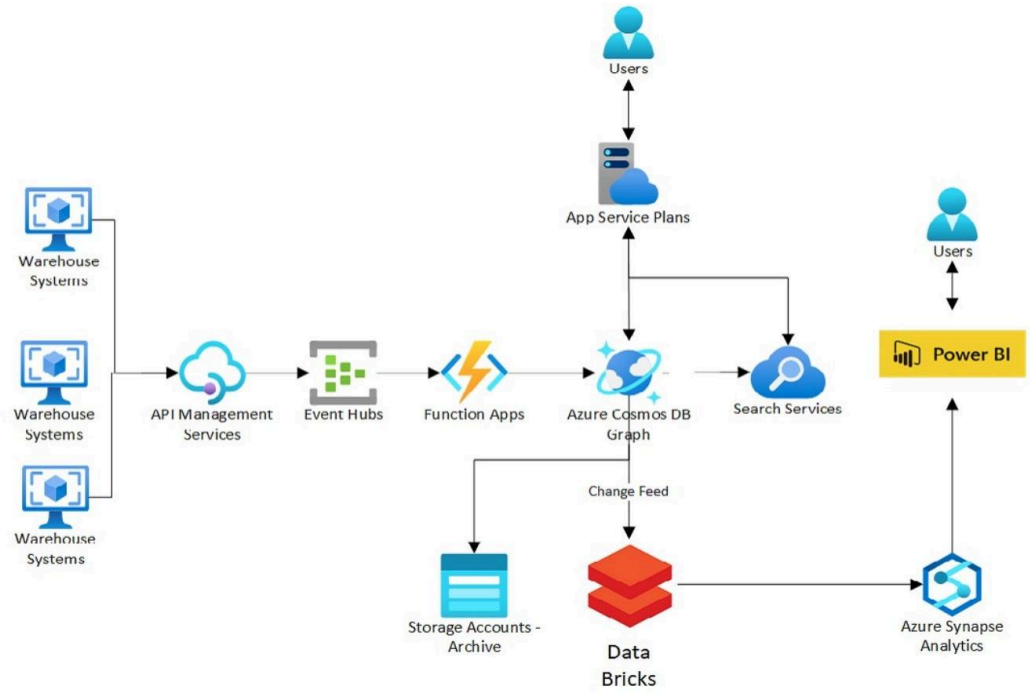

Imagen 6: Extracción de relaciones entidades con Azure Knowledge Graph (*Microsoft Azure Marketplace*, n.d.)

4. Extracción de características de relevancia:

- **Azure Cognitive Services - Text Analytics API:** Calcula la similitud semántica entre la pregunta y los documentos PDF.
- **Azure Machine Learning:** Identifica las características más relevantes para la tarea de IR utilizando técnicas de aprendizaje automático.

5. Modelo RAG:

- **Azure Cognitive Services - Text Generation API:** Implementa un modelo RAG para generar respuestas a las preguntas.
- **Azure Machine Learning:** Entrena el modelo RAG utilizando las características extraídas y evalúa su rendimiento.

6. Evaluación y mejora:

- **Azure Machine Learning:** Evalúa el modelo RAG utilizando métricas como MAP (Mean Average Precision) o NDCG (Normalized Discounted Cumulative Gain).
- **Azure Machine Learning Pipeline:** Ajusta los parámetros del modelo y las técnicas de ingeniería de características para optimizar el rendimiento.

Herramientas adicionales:

- **Azure Databricks:** Procesamiento y análisis de datos a gran escala.
- **Azure Data Factory:** Orquestación de flujos de trabajo de datos.
- **Azure Cognitive Services - QnA Maker:** Crea bases de datos de preguntas y respuestas.

CONCLUSIÓN

Microsoft Azure ofrece una amplia gama de herramientas y servicios para implementar un modelo de ingeniería de características robusto para un modelo RAG de recuperación de información en documentos PDF. La combinación de estas herramientas con técnicas de extracción de características léxicas, contextuales y de relevancia, junto con un modelo RAG bien entrenado, permite a los usuarios encontrar información relevante de manera eficiente y precisa en grandes volúmenes de documentos PDF, lo cual se realizará masivamente para la información del portal del Tecnológico de Monterrey.

REFERENCIAS

Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. "O'Reilly Media, Inc.".

Chris Seferlis, Christopher Nellis, Andy Roberts (2023). Practical Guide to Azure Cognitive Services. "O'Reilly Media, Inc.".

Gupta, J. (2021, June 2). *Azure Cognitive Services:Text mining and Sentiment Analysis*. Machine Learning | AI | Data Science. https://connectjaya.com/azure-cognitive-servicestext-mining-and-sentiment-analysis/

Verma, P. (2022, January 6). Microsoft Azure Storage Services. - Pawan Verma - medium. Medium. https://radheradhepawan.medium.com/microsoft-azure-storage-services-25606cca95ae

Alto, V. (2022, November 28). Introduction to Azure Machine Learning - Microsoft Azure - Medium. Medium. https://medium.com/microsoftazure/introduction-to-azure-machine-learning-13143ccd19b2

Jboback. (2023, December 19). Quickstart - create a conversational language understanding project - Azure AI services. Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/language-service/conversational-language-understanding/quickstart?pivots=language-studio

Microsoft Azure Marketplace. (n.d.). https://azuremarketplace.microsoft.com/en-us/marketplace/apps/infosysltd.infosys-traceability-knowledge-graph?tab=overview