



## **Maestría en Inteligencia Artificial Aplicada**

### **Proyecto Integrador**

### **Ingeniería de características**

<b>Alumno</b>	<b>Matrícula</b>
Johanna Rodríguez Jaramillo	A01794010
Katherine Vanegas Salamanca	A01794113
Marcos Chávez Chávez	A01688507

#### **Profesores Titulares**

**Dra. Grettel Barceló Alonso / Luis Eduardo Falcón Morales**

#### **Profesor Asistente**

**Horacio Martínez Alfaro**

12 de mayo de 2024

# Ingeniería de características

## Contexto

En el análisis de aproximadamente 150 documentos sobre normas para la fabricación de cables, es fundamental identificar patrones y estructuras que nos permitan extraer información crítica para la producción y el cumplimiento de normativas. Utilizaremos la metodología CRISP-ML para organizar el proceso de preparación de datos en nuestro proyecto de inteligencia artificial, asegurando que cada paso sea justificado y contribuya efectivamente a los objetivos del proyecto.

Las actividades de preprocesamiento serán las descritas a continuación:

## Segmentación de Documentos

La capacidad de segmentar documentos en partes como títulos, listas y párrafos es crucial para estructurar la información de manera que el motor de búsqueda pueda indexar y recuperar contenido específico rápidamente. Esto facilita una búsqueda más eficiente y relevante, permitiendo a los usuarios encontrar exactamente lo que necesitan sin tener que revisar documentos completos.

## Extracción de Términos

Extraer términos técnicos, pasos de fabricación y reglas específicas es esencial para enriquecer la base de datos de búsqueda semántica. Al alimentar al sistema con estos términos clave, se potencia la capacidad del motor de búsqueda para asociar consultas de los usuarios con los documentos más pertinentes y precisos, asegurando que la información relevante sea fácilmente accesible.

## Análisis de Tablas

Dado que muchos documentos de normas contienen especificaciones y requisitos en formatos tabulados, poder extraer y procesar esta información numérica y específica es fundamental. Esto asegura que el buscador pueda interpretar y presentar datos cuantitativos, algo valioso en contextos donde los detalles numéricos critican la toma de decisiones o la verificación de cumplimiento.

## Clasificación de Textos

Clasificar los textos según categorías como información de seguridad, necesidades de calidad y pasos de fabricación mejora la precisión de los resultados de búsqueda. Esto permite que el sistema ofrezca respuestas más afinadas a las consultas específicas de los usuarios, mejorando la relevancia y la utilidad de la información recuperada.

## Herramientas aplicables en MS OpenAI

Para el preprocesamiento de documentos del proyecto utilizando las soluciones de OpenAI en Azure, podremos emplear varias herramientas eficaces disponibles a través de Azure AI Document Intelligence y Azure OpenAI.

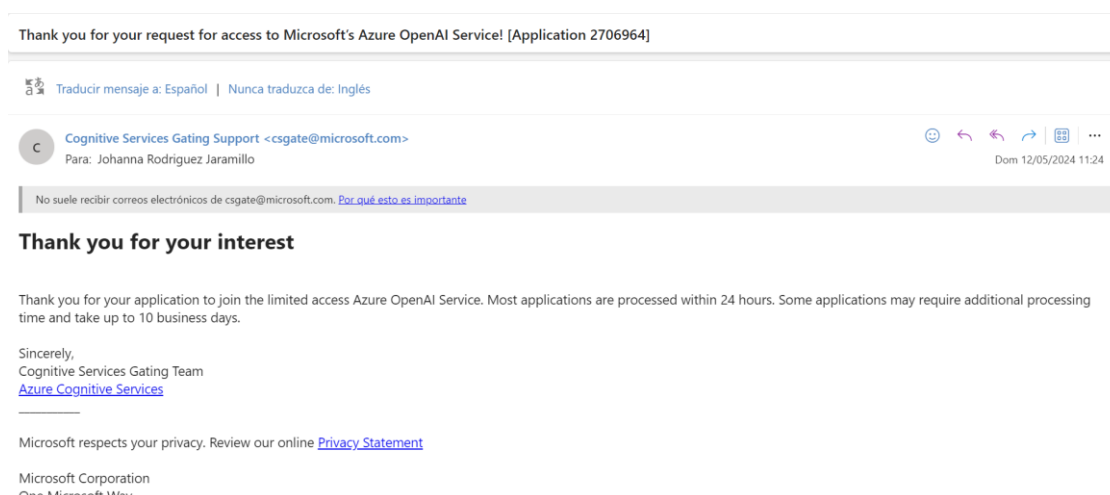
**Azure AI Document Intelligence:** Esta herramienta de inteligencia de documentos de Azure, diseñada específicamente para analizar y entender la disposición y estructura de documentos complejos. Puede manejar distintos elementos como texto, tablas y marcas de selección, lo cual es invaluable para procesar documentos que contienen datos en formas variadas y complejas

**Conversión a Markdown:** Utilizando el modelo de layout pre-establecido de Azure AI, se podrá cambiar la estructura de los documentos a formato Markdown. Esto simplifica la estructura del documento, facilitando su posterior análisis y extracción de datos utilizando modelos NLP. Esta capacidad es útil especialmente para mantener la jerarquía del contenido durante la extracción

**Azure OpenAI:** Después de estructurar los datos en Markdown, podríamos utilizar Azure OpenAI para realizar solicitudes de completado y extraer datos estructurados específicos en formato JSON, que es fácilmente manipulable y almacenable para usos posteriores.

Azure proporciona una infraestructura que permite escalar las operaciones de extracción de datos conforme al volumen de documentos y la demanda de procesamiento. Esto es esencial para manejar grandes volúmenes de documentos sin degradar el rendimiento ni la precisión de la extracción de datos (MS Learn).

En el marco del uso de estos servicios realizamos el registro para acceso a la herramienta OpenAI dado que es requerido por Microsoft, adjuntamos las evidencias:



RV: Thank you for your request for access to Microsoft's Azure OpenAI Service! [Application 2701368]

**De:** Cognitive Services Gating Support <csgate@microsoft.com>

**Enviado:** martes, 7 de mayo de 2024 9:16 p. m.

**Para:** Marcos Chávez Chávez <A01688507@tec.mx>

**Asunto:** Thank you for your request for access to Microsoft's Azure OpenAI Service! [Application 2701368]


No suele recibir correos electrónicos de csgate@microsoft.com. [Por qué esto es importante](#)

## Thank you for your interest

Thank you for your application to join the limited access Azure OpenAI Service. Most applications are processed within 24 hours. Some applications may require additional processing time and take up to 10 business days.

Sincerely,  
Cognitive Services Gating Team  
[Azure Cognitive Services](#)

Thank you for your request for access to Microsoft's Azure OpenAI Service! [Application 2706961]

 Cognitive Services Gating Support <csgate@microsoft.com>  
Para: Katherine Vanegas Salamanca

 Dom 12/05/2024 11:17

No suele recibir correos electrónicos de csgate@microsoft.com. [Por qué esto es importante](#)

## Thank you for your interest

Thank you for your application to join the limited access Azure OpenAI Service. Most applications are processed within 24 hours. Some applications may require additional processing time and take up to 10 business days.

Sincerely,  
Cognitive Services Gating Team  
[Azure Cognitive Services](#)

Microsoft respects your privacy. Review our online [Privacy Statement](#)

Microsoft Corporation  
One Microsoft Way  
Redmond, WA, USA 98052

Inicialmente, un integrante del equipo realizó la solicitud, la cual fue aprobada por Microsoft, posterior a esto, los demás integrantes solicitaron dicho acceso, ya todas las solicitudes fueron aceptadas.

**Enviado:** miércoles, 8 de mayo de 2024 4:02 a. m.  
**Para:** Marcos Chávez Chávez <A01688507@tec.mx>  
**Asunto:** Welcome to the Azure OpenAI Service, Marcos! [ApplicationID 2701368]

No suele recibir correos electrónicos de csgate@microsoft.com. [Por qué esto es importante](#)

## Onboarding to Azure OpenAI Service

Congratulations! After reviewing your application, we are pleased to inform you that you have been onboarded to the Azure OpenAI Service.

As provided in the terms, you may only use this service for the use case provided in your application.

### Getting Started

- [Review the quick start documentation for text and code](#)
- [Review the quick start documentation for OpenAI's Whisper model on Azure](#)
- [Review the responsible AI documentation and requirements](#)
- [Find additional product documentation](#)
- Questions? Need help? Visit Support and Troubleshooting within the Azure Portal

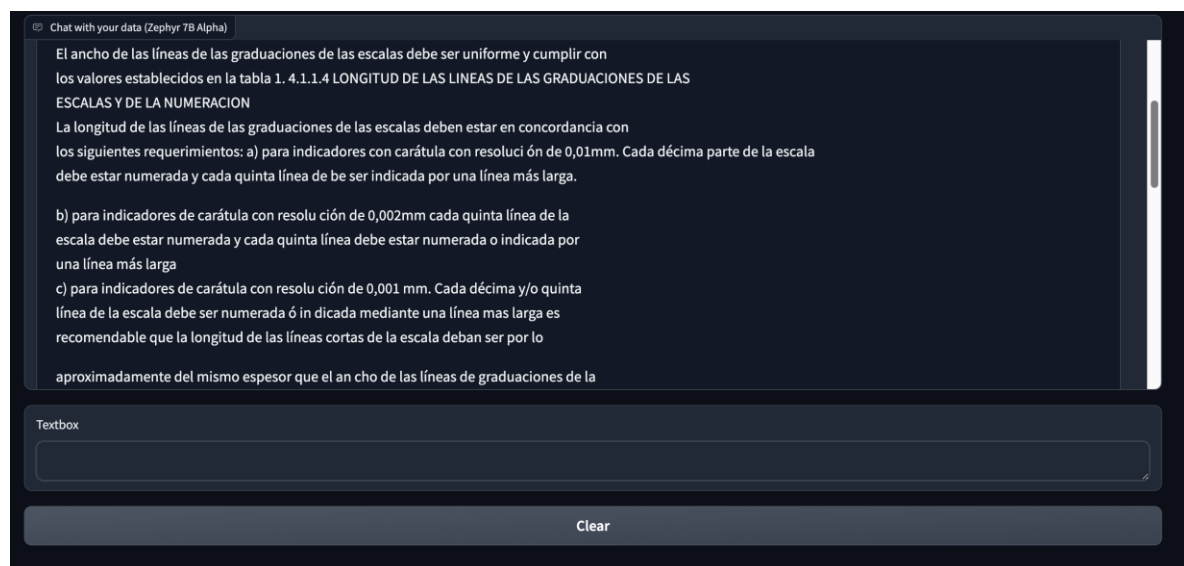
## Analisis otros modelos

Quisimos acercarnos a otras opciones como huggingface, más exactamente el modelo:

`anakin87/zephyr-7b-alpha-sharded`

<https://www.e2enetworks.com/blog/deploying-and-using-zephyr-7b-alpha>

El código del modelo se encuentra en nuestro repositorio del GitHub, la idea es explorar diferentes opciones para validar cual se ajusta más a nuestro problema.



## Conclusiones

La integración de estas técnicas de NLP y análisis de documentos no solo mejora la capacidad de búsqueda y recuperación de información relevante de manera eficiente y precisa, sino que también ayuda a mantener la solidez del sistema frente a las normas vigentes y futuras modificaciones, minimizando el riesgo de incumplimiento. Esto, a su vez, optimiza la eficiencia operativa y reduce costos, alineándose con los objetivos centrales de tu proyecto.

Para nuestro caso, la ingeniería de características no es aplicable al problema que planteamos, el paso a seguir luego del preprocesamiento y limpieza de los documentos es la tokenización y vectorización de los documentos para posteriormente ser alimentados en el modelo [5].

Adicionalmente, seguiremos explorando diferentes modelos de generación de texto que puedan cubrir las necesidades que plantea nuestro problema.

## Bibliografía

- [1] Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., y Plöd, M. (2023). CRISP-ML(Q). The ML Lifecycle Process. MLOps. INNOQ <https://ml-ops.org/content/crisp-ml>
- [2] Azure OpenAI Service documentation - Quickstarts, Tutorials, API Reference - Azure AI services. Microsoft Learn. <https://learn.microsoft.com/en-us/azure/ai-services/openai/>
- [3] Owczarek, D, (diciembre 2023), Generative Question Answering over Documents with LLMs, <https://nexocode.com/blog/posts/generative-question-answering-llms/>
- [4] Deshbhratar, S, (diciembre 2023), Building a Private Data-Driven Question-Answering System with Large Language Models, <https://medium.com/@nbasatish/building-a-private-data-driven-question-answering-system-with-large-language-models-a0b4d4c2385c>
- [5] De Tender, P, (agosto 2023), Build an Azure AI chatbot using your own data from blob storage, <https://pdtit.medium.com/build-an-azure-ai-chatbot-using-your-own-data-from-blob-storage-e372be207ed>