Table of Contents

# 1. INTRODUCCIÓN Y FUNDAMENTOS DEL PROYECTO

## 1.1 Formulación del Problema

Las organizaciones tecnológicas enfrentan un desafío creciente en la gestión eficiente de tickets de soporte técnico, particularmente en el contexto de plataformas complejas como Microsoft Azure. Los equipos de soporte deben atender miles de consultas diarias que requieren conocimiento especializado distribuido en vastas bases documentales. Esta situación genera tres problemas fundamentales:

Primero, existe una **brecha entre el conocimiento disponible y su accesibilidad efectiva**. Para este proyecto se extrajeron 62,417 documentos únicos de Microsoft Learn relacionados con Azure (segmentados en 187,031 chunks para procesamiento según el análisis de las colecciones ChromaDB del proyecto), evidenciando la vasta cantidad de información técnica disponible. Sin embargo, los agentes de soporte frecuentemente dependen de su memoria y experiencia personal, resultando en respuestas inconsistentes y tiempos de resolución prolongados.

Segundo, la **duplicación de esfuerzos es sistemática**. Un análisis de cobertura realizado sobre más de 3,000 preguntas únicas del dataset que contenían enlaces a documentación oficial identificó que 2,067 de estas consultas (aproximadamente 67%) ya tenían sus respuestas disponibles como documentos indexados en la base de conocimiento de Microsoft Learn (external\_helpers/verify\_questions\_links\_match.py, 2025). A pesar de esta alta disponibilidad de información documentada, cada ticket continúa siendo abordado como un caso único, lo que sugiere un desperdicio significativo de recursos.

Tercero, los **sistemas tradicionales de búsqueda léxica son insuficientes** para el dominio técnico. Las consultas de usuarios emplean lenguaje natural y terminología variada que no siempre coincide con los términos exactos utilizados en la documentación oficial, limitando severamente la efectividad de los motores de búsqueda convencionales.

Este proyecto propone desarrollar un sistema de recuperación semántica de información que integre técnicas avanzadas de procesamiento de lenguaje natural (NLP) para vincular automáticamente tickets de soporte con documentación técnica relevante, mejorando la eficiencia y consistencia en la resolución de consultas técnicas.

## 1.2 Alcances

### 1.2.1 Alcance Temático

Este proyecto se enmarca específicamente en la intersección de tres dominios tecnológicos: el procesamiento de lenguaje natural aplicado a dominios técnicos especializados, los sistemas de recuperación de información basados en semántica vectorial, y la gestión automatizada de conocimiento para soporte técnico. La investigación abarca el diseño, implementación y evaluación de un sistema RAG (Retrieval-Augmented Generation) completo, incluyendo:

* Extracción y procesamiento de más de 62,000 documentos técnicos únicos de Microsoft Learn (segmentados en 187,031 chunks)
* Implementación de múltiples modelos de embeddings (Ada, MPNet, MiniLM, E5-Large)
* Desarrollo de arquitecturas de búsqueda híbrida combinando recuperación vectorial y reranking
* Evaluación sistemática mediante métricas específicas de recuperación de información

### 1.2.2 Alcance Temporal

El desarrollo del proyecto se ejecutó durante el período académico 2024-2025, con una fase intensiva de implementación de 16 semanas. Los datos de documentación técnica y preguntas fueron recolectados durante marzo de 2025, creando una imagen estática del conocimiento disponible en Microsoft Learn y Microsoft Q&A hasta esa fecha, lo que permite una evaluación consistente y reproducible del sistema sin variaciones temporales en el corpus de datos.

## 1.3 Delimitaciones

### 1.3.1 Delimitación Geográfica

Aunque el sistema está diseñado para operar globalmente, la implementación se enfoca exclusivamente en documentación y consultas en idioma inglés. Los datos provienen de fuentes públicas internacionales (Microsoft Learn y Microsoft Q&A), sin restricciones geográficas específicas, pero el procesamiento lingüístico se optimiza para terminología técnica en inglés.

### 1.3.2 Delimitación de Dominio

El proyecto se delimita estrictamente al ecosistema de Microsoft Azure, excluyendo otros productos de Microsoft o plataformas cloud competidoras. Esta decisión permite una especialización profunda en la terminología, arquitectura y patrones de consulta específicos de Azure, mejorando la precisión del sistema.

### 1.3.3 Delimitación Funcional

El proyecto tiene un enfoque exclusivamente académico, centrado en la evaluación de técnicas de recuperación de información más que en el desarrollo de un producto comercial. Aunque se implementa una arquitectura RAG completa, el énfasis está en la medición y análisis de métricas de recuperación (Precision@k, Recall@k, MRR, nDCG) más que en la utilización práctica de las respuestas generadas. El sistema se limita a identificar, rankear y evaluar documentos relevantes para cada consulta.

## 1.4 Limitaciones

### 1.4.1 Limitaciones de Datos

Por consideraciones de privacidad y confidencialidad empresarial, el proyecto utiliza exclusivamente datos públicos. No se tuvo acceso a tickets reales de soporte empresarial, lo que impide validar el sistema con casos de uso industriales confidenciales. Las 13,436 preguntas del dataset provienen de Microsoft Q&A (foros públicos), con solo 2,067 preguntas conteniendo enlaces validados a documentación oficial, lo que puede no representar completamente la complejidad de tickets corporativos internos.

### 1.4.2 Limitaciones Técnicas

El procesamiento se limita a contenido textual, excluyendo elementos multimedia como imágenes, diagramas arquitectónicos y videos instructivos que forman parte significativa de la documentación técnica moderna. Adicionalmente, los modelos de embeddings tienen limitaciones de contexto variables: MiniLM (256 tokens), MPNet (384 tokens), E5-Large (512 tokens), aunque OpenAI Ada maneja hasta 8,191 tokens (Hugging Face, 2025; OpenAI, 2025). Estas limitaciones requieren estrategias de segmentación que pueden perder información contextual importante.

### 1.4.3 Limitaciones de Evaluación

La validación del sistema se basa en enlaces explícitos entre preguntas y documentos identificados en respuestas aceptadas por la comunidad. Este criterio, aunque objetivo, es más estricto que escenarios reales donde múltiples documentos pueden ser igualmente válidos para resolver una consulta.

## 1.5 Objetivos

### 1.5.1 Objetivo General

Desarrollar y evaluar un sistema de recuperación semántica de información basado en técnicas avanzadas de procesamiento de lenguaje natural, utilizando la documentación de Microsoft Azure como caso de estudio para simular un entorno de soporte técnico especializado, con el fin de medir y analizar la efectividad de diferentes modelos de embeddings y arquitecturas de recuperación en la identificación precisa de documentos relevantes para consultas técnicas específicas de dominio.

### 1.5.2 Objetivos Específicos

1. **Construir un corpus comprehensivo de conocimiento técnico** mediante la extracción, procesamiento y estructuración de la documentación completa de Microsoft Learn y las consultas históricas de Microsoft Q&A, estableciendo un dataset de referencia con 62,417 documentos únicos (segmentados en 187,031 chunks) y 13,436 preguntas con sus respuestas.
2. **Implementar y optimizar múltiples arquitecturas de embeddings** evaluando comparativamente modelos de código abierto (Sentence-BERT variantes MiniLM y MPNet, E5-Large) y propietarios (OpenAI Ada) para determinar la representación vectorial óptima del contenido técnico especializado de Azure.
3. **Diseñar e implementar un sistema de almacenamiento y recuperación vectorial** utilizando ChromaDB como base de datos especializada, configurando índices optimizados para búsquedas de similitud semántica a escala, con soporte para más de 800,000 vectores de alta dimensionalidad distribuidos en 8 colecciones especializadas y 1 colección auxiliar para preguntas con enlaces validados.
4. **Desarrollar mecanismos avanzados de reranking** implementando CrossEncoders especializados y técnicas de normalización (Min-Max) para mejorar la precisión en el ordenamiento final de documentos recuperados, optimizando específicamente para consultas técnicas complejas.
5. **Evaluar sistemáticamente el rendimiento del sistema** mediante un framework comprehensivo de métricas que incluye medidas tradicionales de recuperación de información (Precision@k, Recall@k, MRR, nDCG) en etapas pre y post reranking, métricas específicas para sistemas RAG (Answer Relevancy, Context Precision, Faithfulness), y validación semántica utilizando tanto RAGAS como BERTScore, lo que permite analizar el impacto de cada componente del pipeline de recuperación.
6. **Establecer una metodología reproducible y extensible** documentando exhaustivamente el proceso de implementación, creando pipelines automatizados de evaluación con métricas reales y verificables, y desarrollando herramientas auxiliares (incluyendo una interfaz Streamlit) que faciliten la ejecución de pruebas y la visualización de resultados para futuras investigaciones en el dominio de recuperación semántica de información técnica.

## 1.6 Referencias del Capítulo

Gómez, H. (2025). *verify\_questions\_links\_match.py: Script de verificación de cobertura entre preguntas y documentación* [Código fuente]. https://github.com/[repositorio]/SupportModel/blob/main/external\_helpers/verify\_questions\_links\_match.py

Microsoft. (2025). *Microsoft Learn Documentation*. https://learn.microsoft.com/

Microsoft. (2025). *Microsoft Q&A*. https://learn.microsoft.com/en-us/answers/

Hugging Face. (2025). *Sentence Transformers: Model documentation and specifications*. https://huggingface.co/sentence-transformers/

OpenAI. (2025). *Embeddings API documentation*. https://platform.openai.com/docs/guides/embeddings

## 1.7 Nota sobre las fuentes de datos

Los datos cuantitativos presentados en este capítulo provienen de análisis realizados sobre el dataset del proyecto: - Las estadísticas de cobertura (67% o 2,067 de 3,000+ preguntas con enlaces) se calcularon mediante el análisis de las colecciones ChromaDB y scripts de verificación - Las cantidades de documentos (62,417 únicos, 187,031 chunks) y preguntas (13,436 totales) provienen del análisis directo de las colecciones ChromaDB - Las observaciones sobre la dependencia del conocimiento tácito y la insuficiencia de búsquedas léxicas son inferencias basadas en la literatura de recuperación de información y la naturaleza del problema abordado