Table of Contents

RESUMEN  
ABSTRACT  
CONTENIDOS  
ÍNDICE DE TABLAS  
ÍNDICE DE FIGURAS

CAPÍTULO I: INTRODUCCIÓN Y FUNDAMENTOS DEL PROYECTO  
1. Formulación del Problema  
2. Alcances  
3. Delimitaciones  
4. Limitaciones  
5. Objetivos  
5.1. Objetivo General  
5.2. Objetivos Específicos  
6. Estructura del Documento

CAPÍTULO II: ESTADO DEL ARTE  
1. Introducción  
2. NLP Aplicado a Soporte Técnico  
2.1. Evolución del Procesamiento de Lenguaje Natural  
2.2. Aplicaciones en Soporte al Cliente  
3. Bases de Conocimiento como Entrada para Recuperación de Información  
3.1. Microsoft Learn como Fuente de Conocimiento  
3.2. Estructuración de Documentación Técnica  
4. Comparación de Enfoques Vectoriales y Clásicos  
4.1. Modelos de Embeddings  
4.2. Búsqueda Léxica (BM25)  
4.3. Enfoques Híbridos  
5. Casos Empresariales Relevantes  
6. Medidas de Evaluación en Recuperación de Información  
6.1. Métricas Tradicionales (Precision, Recall)  
6.2. Métricas RAG Específicas

CAPÍTULO III: MARCO TEÓRICO  
1. Fundamentos de Recuperación de Información  
2. Modelos de Embeddings  
2.1. OpenAI Ada  
2.2. Sentence-BERT (MPNet, MiniLM)  
2.3. E5-Large  
3. Arquitecturas RAG (Retrieval-Augmented Generation)  
4. CrossEncoders y Reranking  
5. Bases de Datos Vectoriales

CAPÍTULO IV: ANÁLISIS EXPLORATORIO DE DATOS (EDA) 1. Características del Corpus de Documentos 1.1. Composición General del Corpus 1.2. Análisis de Longitud de Documentos 1.3. Distribución Temática del Corpus 1.4. Análisis de Calidad del Corpus 2. Características del Dataset de Preguntas 2.1. Composición del Dataset de Preguntas 2.2. Análisis de Longitud de Preguntas 2.3. Distribución de Tipos de Preguntas 2.4. Análisis de Ground Truth 3. Análisis de Correspondencia entre Datos 3.1. Mapping Preguntas-Documentos 3.2. Distribución Temática de Ground Truth 3.3. Calidad de la Correspondencia 4. Hallazgos Principales del EDA 4.1. Características Estructurales 4.2. Implicaciones para el Sistema RAG 4.3. Benchmarking del Corpus 5. Recomendaciones para Mejoras 5.1. Mejoras al Corpus de Documentación 5.2. Mejoras al Dataset de Preguntas 5.3. Mejoras Metodológicas 6. Conclusiones del EDA

CAPÍTULO V: METODOLOGÍA  
1. Diseño de la Investigación  
2. Recolección y Preparación de Datos  
2.1. Extracción de Microsoft Learn  
2.2. Procesamiento de Documentos  
2.3. Generación de Embeddings  
3. Arquitectura del Sistema  
3.1. Pipeline de Recuperación  
3.2. Componentes de Reranking  
4. Configuración Experimental  
4.1. Modelos Evaluados  
4.2. Parámetros de Búsqueda  
5. Proceso de Evaluación

CAPÍTULO VI: IMPLEMENTACIÓN  
1. Tecnologías Utilizadas  
2. Arquitectura del Sistema  
2.1. Componente de Indexación  
2.2. Componente de Búsqueda  
2.3. Componente de Evaluación  
3. Extracción Automatizada de Datos desde Microsoft Learn  
3.1. Herramientas y Técnicas de Web Scraping  
3.2. Proceso de Extracción de Documentación  
3.3. Proceso de Extracción de Preguntas y Respuestas  
3.4. Consideraciones Éticas y Legales del Uso de Documentación Técnica  
4. Implementación de ChromaDB  
5. Pipeline de Procesamiento  
6. Interfaz de Usuario (Streamlit)  
7. Optimizaciones y Mejoras

CAPÍTULO VII: RESULTADOS Y ANÁLISIS  
1. Resultados por Modelo de Embedding  
1.1. Ada (OpenAI)  
1.2. MPNet  
1.3. MiniLM  
1.4. E5-Large  
2. Análisis Comparativo  
2.1. Métricas de Precisión  
2.2. Métricas de Relevancia Semántica  
2.3. Tiempos de Respuesta  
3. Impacto del CrossEncoder  
4. Análisis de Casos de Uso  
5. Discusión de Resultados

CAPÍTULO VIII: CONCLUSIONES Y TRABAJO FUTURO  
1. Conclusiones Principales  
2. Contribuciones del Trabajo  
3. Limitaciones Encontradas  
4. Trabajo Futuro  
4.1. Mejoras en Modelos  
4.2. Expansión de Fuentes de Datos  
4.3. Optimización de Pipeline  
5. Recomendaciones para Implementación en Producción

REFERENCIAS BIBLIOGRÁFICAS

ANEXOS  
A. MODELOS DE EMBEDDING RELEVANTES (HASTA PRINCIPIOS DE 2025) B. Código Fuente Principal  
C. Configuración de Ambiente  
C. Ejemplos de Consultas y Respuestas  
D. Resultados Detallados por Métricas