###### ANEXO F: Manual de Usuario - Aplicación Streamlit

## F.1 Introducción

Este anexo proporciona las instrucciones detalladas para utilizar la aplicación web desarrollada con Streamlit que permite interactuar con el sistema RAG de recuperación semántica de documentación técnica de Microsoft Azure. La aplicación ofrece cuatro funcionalidades principales organizadas en páginas independientes.

## F.2 Requisitos del Sistema

### F.2.1 Requisitos de Hardware

* **Memoria RAM**: Mínimo 8 GB (16 GB recomendado)
* **Almacenamiento**: 20 GB libres para modelos y base de datos
* **Procesador**: CPU multinúcleo (GPU opcional para aceleración)

### F.2.2 Requisitos de Software

* **Python**: 3.8 o superior
* **Sistema Operativo**: Windows 10/11, macOS, o Linux
* **Navegador Web**: Chrome, Firefox, Safari o Edge (versiones recientes)

### F.2.3 Dependencias Principales

* Streamlit 1.46.1
* ChromaDB 0.5.23
* Sentence-Transformers 2.3.0
* OpenAI API (para modelo Ada)
* Matplotlib/Plotly para visualizaciones

## F.3 Instalación y Configuración

### F.3.1 Clonar el Repositorio

git clone https://github.com/sirharold/SupportModel.git

cd SupportModel

### F.3.2 Instalar Dependencias

pip install -r requirements.txt

### F.3.3 Configurar Variables de Entorno

Crear archivo .env en la raíz del proyecto:

OPENAI\_API\_KEY=your\_openai\_api\_key

GOOGLE\_API\_KEY=your\_google\_api\_key # Si se usa Gemini

### F.3.4 Iniciar la Aplicación

streamlit run src/apps/main\_qa\_app.py

La aplicación se abrirá automáticamente en el navegador predeterminado en http://localhost:8501

## F.4 Navegación Principal

### F.4.1 Sidebar de Navegación

La aplicación presenta un menú lateral (sidebar) con dos secciones principales:

#### \*\*🧭 Navegación\*\*

Contiene las cuatro páginas disponibles:

1. 🔍 Búsqueda Individual

2. 📈 Análisis de Datos

3. ⚙️ Configuración Métricas Acumulativas

4. 📊 Resultados Métricas Acumulativas

#### \*\*⚙️ Configuración\*\*

Permite seleccionar:

* **Modelo de Embedding**: Ada, MPNet, MiniLM, E5-Large
* **Modelo Generativo**: TinyLlama, GPT-4, Gemini, Mistral

## F.5 Página: 🔍 Búsqueda Individual

### F.5.1 Propósito

Permite realizar consultas individuales sobre documentación de Azure y obtener respuestas generadas por el sistema RAG.

### F.5.2 Elementos de la Interfaz

#### \*\*Campo de Consulta\*\*

* **Ubicación**: Parte superior de la página principal
* **Placeholder**: "Ejemplo: How to configure Azure Virtual Network?"
* **Función**: Ingrese su pregunta sobre Azure en lenguaje natural

#### \*\*Parámetros de Búsqueda\*\*

* **Número de documentos (k)**: Slider para seleccionar cuántos documentos recuperar (1-20)
* **Enable CrossEncoder Reranking**: Checkbox para activar/desactivar reranking
* **Usar RAG (Retrieval Augmented Generation)**: Checkbox para generar respuestas completas

#### \*\*Botón de Búsqueda\*\*

* **Texto**: "🔍 Buscar"
* **Acción**: Inicia el proceso de recuperación y generación

### F.5.3 Proceso de Uso

1. **Ingresar Consulta**: Escriba una pregunta específica sobre Azure

* + Ejemplo: "How to configure SSL certificates in Azure Application Gateway?"

2. **Configurar Parámetros**:

* + Ajuste el número de documentos según necesidad (default: 10)
  + Active reranking para mejorar relevancia (recomendado)
  + Active RAG para obtener respuestas generadas

3. **Ejecutar Búsqueda**: Haga clic en "🔍 Buscar"

4. **Revisar Resultados**:

* + **Respuesta Generada**: Aparece primero si RAG está activo
  + **Documentos Recuperados**: Lista de documentos relevantes con:

Título del documento

Score de relevancia

Contenido del fragmento

Enlace a la documentación oficial

### F.5.4 Interpretación de Resultados

#### \*\*Métricas Mostradas\*\*

* **Antes del Reranking**: Scores originales del modelo de embedding
* **Después del Reranking**: Scores ajustados por CrossEncoder
* **Diferencia**: Cambio en el ordenamiento y relevancia

#### \*\*Visualización de Scores\*\*

* Gráfico de barras comparativo antes/después del reranking
* Colores indican mejora (verde) o degradación (rojo) en posición

### F.5.5 Casos de Uso Típicos

1. **Consulta de Procedimientos**: "How to create a virtual machine in Azure?"

2. **Resolución de Problemas**: "Troubleshooting Azure Storage connection issues"

3. **Mejores Prácticas**: "Best practices for Azure SQL Database security"

4. **Configuración**: "Configure Azure Application Gateway with SSL"

## F.6 Página: 📈 Análisis de Datos

### F.6.1 Propósito

Proporciona estadísticas y visualizaciones sobre el corpus de documentos y las colecciones en ChromaDB.

### F.6.2 Secciones Principales

#### \*\*📊 Estadísticas del Corpus\*\*

* **Total de Documentos**: Número total de documentos únicos
* **Total de Chunks**: Fragmentos procesados
* **Promedio de Chunks por Documento**: Indicador de granularidad
* **Tamaño Promedio de Chunk**: En caracteres

#### \*\*📈 Distribución de Documentos\*\*

* **Histograma**: Distribución de chunks por documento
* **Interpretación**: Documentos con muchos chunks son más extensos/complejos

#### \*\*🗄️ Colecciones en ChromaDB\*\*

* **Tabla de Colecciones**:
  + Nombre de la colección
  + Número de documentos
  + Modelo de embedding usado
  + Dimensionalidad de vectores

### F.6.3 Uso de la Información

1. **Verificar Integridad**: Confirmar que todas las colecciones están pobladas

2. **Comparar Modelos**: Ver diferencias en número de documentos por modelo

3. **Identificar Problemas**: Detectar colecciones vacías o incompletas

4. **Planificar Mejoras**: Identificar documentos que necesitan mejor segmentación

### F.6.4 Métricas Clave

* **Cobertura**: Porcentaje de documentos procesados exitosamente
* **Balance**: Distribución uniforme entre colecciones
* **Calidad**: Tamaño apropiado de chunks para recuperación efectiva

## F.7 Página: ⚙️ Configuración Métricas Acumulativas

### F.7.1 Propósito

Permite configurar y generar archivos de configuración para evaluaciones exhaustivas del sistema en Google Colab.

### F.7.2 Elementos de Configuración

#### \*\*📝 Configuración General\*\*

* **Nombre del Experimento**: Identificador único para la evaluación
* **Descripción**: Detalles sobre el propósito de la evaluación
* **Número de Preguntas**: Cantidad de preguntas a evaluar (10-1000)

#### \*\*🎯 Selección de Modelos\*\*

* **Modelos de Embedding**: Checkboxes para seleccionar modelos a evaluar
  + ✅ Ada (OpenAI)
  + ✅ MPNet
  + ✅ MiniLM
  + ✅ E5-Large

#### \*\*📊 Métricas a Calcular\*\*

* **Métricas Tradicionales**: Precision@k, Recall@k, F1@k, NDCG@k
* **Métricas RAG**: Faithfulness, Answer Relevancy, Context Precision
* **Métricas Semánticas**: BERTScore (Precision, Recall, F1)

#### \*\*🔧 Parámetros de Evaluación\*\*

* **Top K**: Número de documentos a recuperar (default: 10)
* **Enable Reranking**: Activar CrossEncoder para todos los modelos
* **Batch Size**: Tamaño de lote para procesamiento (default: 50)

### F.7.3 Proceso de Configuración

1. **Definir Experimento**:

```

Nombre: evaluacion\_completa\_agosto\_2025

Descripción: Evaluación exhaustiva con 1000 preguntas por modelo

```

2. **Seleccionar Modelos**:

* + Marcar todos los modelos para comparación completa
  + O seleccionar subconjunto para evaluación específica

3. **Configurar Métricas**:

* + Mantener todas las métricas activas para análisis completo
  + Desactivar métricas costosas si hay limitaciones de tiempo

4. **Generar Configuración**:

* + Click en "💾 Generar Configuración"
  + Se descarga archivo evaluation\_config\_[timestamp].json

### F.7.4 Archivo de Configuración Generado

{

"experiment\_name": "evaluacion\_completa\_agosto\_2025",

"description": "Evaluación exhaustiva con 1000 preguntas por modelo",

"timestamp": "2025-08-03T10:30:00",

"models": ["ada", "mpnet", "minilm", "e5-large"],

"num\_questions": 1000,

"metrics": {

"traditional": ["precision", "recall", "f1", "ndcg", "mrr"],

"rag": ["faithfulness", "answer\_relevancy", "context\_precision"],

"semantic": ["bertscore\_precision", "bertscore\_recall", "bertscore\_f1"]

},

"parameters": {

"top\_k": 10,

"enable\_reranking": true,

"batch\_size": 50,

"seed": 42

}

}

### F.7.5 Uso del Archivo en Google Colab

1. **Subir a Colab**: Cargar el archivo JSON generado

2. **Ejecutar Notebook**: Cumulative\_Ticket\_Evaluation.ipynb

3. **Monitorear Progreso**: La evaluación mostrará progreso en tiempo real

4. **Descargar Resultados**: Al finalizar, descargar cumulative\_results\_\*.json

## F.8 Página: 📊 Resultados Métricas Acumulativas

### F.8.1 Propósito

Visualiza y analiza los resultados de evaluaciones completas ejecutadas en Google Colab.

### F.8.2 Carga de Resultados

#### \*\*📁 Cargar Archivo de Resultados\*\*

1. Click en "Browse files" o arrastrar archivo

2. Seleccionar archivo cumulative\_results\_\*.json

3. El sistema valida y carga automáticamente

#### \*\*Validación del Archivo\*\*

* Verifica estructura JSON correcta
* Confirma presencia de métricas requeridas
* Detecta modelos evaluados
* Calcula estadísticas de completitud

### F.8.3 Visualizaciones Disponibles

#### \*\*📊 Comparación de Modelos\*\*

* **Gráfico de Barras**: Precision@5, Recall@5, F1@5 por modelo
* **Antes/Después Reranking**: Impacto del CrossEncoder
* **Tabla Resumen**: Todas las métricas en formato tabular

#### \*\*📈 Análisis de Rendimiento\*\*

* **NDCG@k**: Calidad del ranking (1-15)
* **MRR**: Mean Reciprocal Rank
* **MAP**: Mean Average Precision

#### \*\*🎯 Métricas RAG\*\*

* **Faithfulness**: Fidelidad de las respuestas generadas
* **Answer Relevancy**: Relevancia de las respuestas
* **Context Precision**: Precisión del contexto recuperado

#### \*\*📉 Análisis Temporal\*\*

* **Tiempo por Consulta**: Distribución de latencias
* **Throughput**: Consultas procesadas por minuto
* **Bottlenecks**: Identificación de cuellos de botella

### F.8.4 Interpretación de Resultados

#### \*\*Identificar Mejor Modelo\*\*

1. Revisar Precision@5 (métrica principal)

2. Considerar balance Precision/Recall

3. Evaluar impacto del reranking

4. Verificar consistencia en métricas RAG

#### \*\*Análisis de Reranking\*\*

* **Mejora Positiva**: Verde, el reranking ayuda
* **Impacto Negativo**: Rojo, el reranking degrada
* **Neutral**: Gris, cambio mínimo

#### \*\*Significancia Estadística\*\*

* Valores p < 0.05 indican diferencias significativas
* Intervalos de confianza no solapados confirman diferencias
* Tests de Wilcoxon para comparaciones pareadas

### F.8.5 Exportación de Resultados

#### \*\*📄 Formatos Disponibles\*\*

1. **CSV**: Tabla de métricas para análisis externo

2. **PNG**: Gráficos de alta resolución

3. **JSON**: Datos crudos para procesamiento adicional

4. **PDF**: Reporte completo con visualizaciones

#### \*\*🎨 Personalización\*\*

* Seleccionar métricas específicas
* Filtrar modelos de interés
* Ajustar escalas y colores
* Agregar anotaciones

### F.8.6 Casos de Uso Avanzados

1. **Comparación Multi-Experimento**:

* + Cargar múltiples archivos de resultados
  + Comparar evolución temporal
  + Identificar mejoras/degradaciones

2. **Análisis por Subgrupos**:

* + Filtrar por tipo de pregunta
  + Analizar por servicio de Azure
  + Segmentar por complejidad

3. **Optimización de Parámetros**:

* + Variar top\_k y analizar impacto
  + Ajustar umbrales de reranking
  + Encontrar configuración óptima

## F.9 Troubleshooting Común

### F.9.1 Errores de Inicio

**Error**: "ModuleNotFoundError"

* **Solución**: Verificar instalación de dependencias con pip install -r requirements.txt

**Error**: "ChromaDB connection failed"

* **Solución**: Verificar que ChromaDB esté instalado y la ruta sea correcta

### F.9.2 Problemas de Rendimiento

**Síntoma**: Búsquedas muy lentas

* **Causa**: Modelo de embedding no optimizado
* **Solución**: Usar MiniLM para pruebas rápidas, Ada para producción

**Síntoma**: Memoria insuficiente

* **Causa**: Modelos generativos grandes
* **Solución**: Usar TinyLlama en lugar de Mistral

### F.9.3 Resultados Inesperados

**Síntoma**: Métricas en cero

* **Causa**: Configuración incorrecta del modelo
* **Solución**: Verificar prefijos para E5-Large, normalización de embeddings

**Síntoma**: Documentos irrelevantes

* **Causa**: Query muy general o ambigua
* **Solución**: Ser más específico, incluir nombres de servicios Azure

## F.10 Mejores Prácticas

### F.10.1 Para Búsqueda Individual

1. Usar preguntas específicas con contexto

2. Activar reranking para mejor precisión

3. Ajustar k según necesidad (más documentos = más contexto)

### F.10.2 Para Evaluaciones

1. Usar mínimo 100 preguntas para significancia estadística

2. Evaluar todos los modelos para comparación completa

3. Guardar configuraciones para reproducibilidad

### F.10.3 Para Análisis

1. Comparar métricas antes/después de cambios

2. Documentar configuraciones utilizadas

3. Exportar resultados para respaldo

## F.11 Conclusión

La aplicación Streamlit proporciona una interfaz intuitiva y poderosa para interactuar con el sistema RAG de documentación Azure. Las cuatro páginas cubren el ciclo completo desde búsquedas individuales hasta evaluaciones exhaustivas, permitiendo tanto uso operacional como investigación sistemática del rendimiento del sistema.