Propuesta Técnica: Sistema Inteligente de Consulta de Subvenciones

Sistema RAG (Retrieval-Augmented Generation)

Para: Pedro (InfoSubvenciones) **De:** Cristian Rojas **Fecha:** 24 de Noviembre de 2025 **Asunto:** Implementación de sistema de IA para búsqueda y consulta inteligente de convocatorias de subvenciones

---

Resumen

Esta propuesta detalla la implementación de un sistema de Inteligencia Artificial que transformará aproximadamente 143,000 PDFs de convocatorias de subvenciones en una base de datos consultable e inteligente. El sistema utilizará modelos de Google (Gemini 2.5 Flash-Lite y Gemini Embedding 001) para comprender el contenido de las convocatorias y responder preguntas precisas con citación de fuentes.

Enfoque de desarrollo: Comenzaremos con un prototipo funcional (1,000 PDFs) para validar la precisión del sistema antes de procesar el volumen completo de información.

---

1. Arquitectura del Sistema

El sistema se compone de cuatro módulos principales:

1.1. Sistema de Ingesta Inteligente

Objetivo: Descargar, procesar y almacenar las convocatorias de subvenciones.

Proceso:

1. Conexión a la API de InfoSubvenciones
2. Descarga de PDFs de convocatorias
3. Extracción de metadatos estructurados (región, beneficiarios, plazos, cuantías)
4. Generación de resúmenes de alta calidad usando **Gemini 2.5 Flash-Lite**
5. Deduplicación (evitar procesar documentos repetidos)

Tecnologías:

- Python para scripts de automatización
- API oficial de InfoSubvenciones
- Gemini 2.5 Flash-Lite (Google AI)
- Checksum MD5 para deduplicación

1.2. Sistema de Vectorización

Objetivo: Convertir el texto en representaciones matemáticas que permitan búsquedas por significado.

¿Qué son los embeddings? Los embeddings convierten palabras y frases en coordenadas numéricas en un espacio matemático multidimensional. Esto permite que el sistema entienda el *significado* del texto, no solo las palabras exactas:

- **Búsqueda tradicional:** Solo encuentra documentos con las palabras exactas ("ayuda para contratar personal")
- **Búsqueda semántica:** Encuentra documentos con significados similares, aunque usen palabras diferentes ("subvención para staff", "apoyo para empleados", "financiación para contratación")

Esta capacidad permite que el sistema comprenda las matices del lenguaje y encuentre información relevante incluso cuando la pregunta del usuario no coincide palabra por palabra con el documento.

Estrategia de optimización: En lugar de vectorizar los 90GB completos de PDFs (extremadamente costoso), solo vectorizamos los resúmenes de 200 palabras generados por el LLM. Esto reduce los costes de procesamiento en más del 90% manteniendo la precisión.

Tecnología:

- Gemini Embedding 001 (Google AI)

1.3. Base de Datos Híbrida (Supabase)

Objetivo: Almacenar información estructurada y vectorial en un solo lugar.

La base de datos combina dos tipos de búsqueda:

Tipo de Dato	Columnas	Función
Datos estructurados (SQL)	región, fecha_límite, tipo_beneficiario, cuantía, número_convocatoria	Filtros precisos (ej: "solo Madrid", "solo PYMES")

Datos vectoriales	embedding del resumen	Búsqueda por significado (ej: "ayudas para teatro")
Documento fuente	PDF completo o enlace	Fuente de verdad para la respuesta final

Tecnología:

- Supabase (PostgreSQL + pgvector)

1.4. Sistema de Consulta Inteligente ("El Bibliotecario")

Objetivo: Responder preguntas del usuario con precisión y citar las fuentes.

Proceso en 4 pasos:

1. **Filtro SQL:** Aplicar filtros duros según los criterios del usuario (región, tipo de beneficiario, fechas)
2. **Búsqueda semántica:** Entre los documentos filtrados, buscar los 3-5 más relevantes usando vectores
3. **Recuperación del documento original:** Obtener el PDF completo o las secciones relevantes
4. **Generación de respuesta:** Un modelo más potente (**Gemini 2.5 Flash**) lee el documento original y genera una respuesta precisa con citación de artículos específicos

Ejemplo de consulta:

```
Usuario: "¿Qué ayudas hay para autónomos de artes escénicas en Madrid que
cierren en 2026?"

Sistema:
1. Filtra: región = Madrid, beneficiario = autónomos, deadline > 2026
2. Busca semánticamente: "artes escénicas"
3. Recupera los PDFs de las 3 convocatorias más relevantes
4. Responde: "La convocatoria 'Cultura Madrid 2026' (ID: 12345) cubre
proyectos
   de artes escénicas. Según el Artículo 5.2, autónomos pueden solicitar
hasta
   15,000€. Plazo: 31/03/2026. [enlace al PDF]"
```

Tecnología:

- Gemini 2.5 Flash (Google AI) para respuestas finales

1.5. Sistema de Actualización Continua

Objetivo: Mantener la base de datos actualizada automáticamente.

Funciones:

- Consulta diaria/semanal a la API de InfoSubvenciones para detectar nuevas convocatorias
- Procesamiento automático de nuevas subvenciones (ingesta + vectorización)
- Detección y eliminación de convocatorias expiradas (basado en fechas límite)

Tecnología:

- Script Python con programación temporal (cron jobs)
- API de InfoSubvenciones

1.6. Interfaz de Usuario

Objetivo: Permitir a los usuarios consultar el sistema.

Opciones disponibles (a definir por Pedro):

- API REST para integración con sistemas existentes
- Interfaz web simple tipo chat
- Integración con aplicaciones existentes

Tecnología: A definir según requisitos

---

2. Modelos de Lenguaje y Alternativas

Modelos Propuestos (Google AI)

Modelo	Uso	Coste por 1M tokens
Gemini 2.5 Flash-Lite	Generación de resúmenes (ingesta masiva)	Input: $0.10 / Output: $0.40
Gemini Embedding 001	Vectorización de texto	$0.02
Gemini 2.5 Flash	Respuestas finales al usuario	Input: $0.30 / Output: $2.50

Alternativas Disponibles

El sistema puede configurarse con otros modelos según preferencias de coste/calidad:

Alternativas de LLM:

- OpenAI GPT-4o / GPT-4o-mini
- Claude (Anthropic)
- Modelos open-source (Llama, Mixtral) para mayor control

Alternativas de embeddings:

- OpenAI text-embedding-3-small/large
- Cohere embeddings
- Modelos open-source

Nota: Los costes varían significativamente según el modelo elegido. Podemos ajustar la arquitectura para optimizar la relación coste/calidad según las necesidades.

---

3. Estimación de Costes Operativos Mensuales

Infraestructura (Supabase)

- Plan Pro: ~$25 USD/mes
- Incluye: 8GB base de datos, 100GB almacenamiento, pgvector

Costes de IA (estimación para 1,000 consultas/mes)

- Ingesta inicial: Una sola vez (ver propuesta de desarrollo)
- Consultas de usuarios: ~$5-10 USD/mes
- Actualizaciones periódicas: ~$2-5 USD/mes

Total operativo estimado: $30-50 USD/mes

Nota: Estos costes escalan gradualmente con el uso. El diseño del sistema minimiza costes recurrentes.

---

4. Propuesta de Desarrollo en Dos Fases

FASE 1: Prototipo Funcional

Objetivo: Demostrar que el sistema comprende el material y responde con precisión.

Alcance:

- Procesamiento de 1,000 PDFs de muestra
- Sistema completo de ingesta (descarga, análisis, vectorización)
- Base de datos híbrida en Supabase
- Sistema de consulta funcional ("el bibliotecario")
- Demostración de precisión con casos de prueba reales

Entregables:

- Sistema funcional con 1,000 convocatorias
- Documentación del proceso
- Demo interactiva para pruebas
- Informe de precisión y calidad de respuestas

Inversión Fase 1: €1,500 + IVA

Términos de pago:

- 30% al inicio (€450 + IVA)
- 70% a la entrega (€1,050 + IVA)

Duración estimada: 2-3 semanas

FASE 2: Sistema Completo de Producción

Objetivo: Escalar a las 143,000 convocatorias y automatizar el mantenimiento.

Alcance:

- Procesamiento de las ~143,000 convocatorias completas
- Sistema de actualización automática (nuevas convocatorias)
- Sistema de detección y limpieza de convocatorias expiradas
- Interfaz de usuario (tipo a definir)
- Optimización de rendimiento y costes
- Documentación técnica completa
- Pruebas exhaustivas de producción

Entregables:

- Sistema completo en producción
- Scripts de actualización automatizada

- Interfaz de usuario operativa
- Documentación técnica y de usuario
- Manual de mantenimiento

Inversión Fase 2: €2,000 + IVA

Términos de pago:

- 50% al inicio de Fase 2 (€1,000 + IVA)
- 50% a la entrega final (€1,000 + IVA)

Duración estimada: 3-4 semanas

INVERSIÓN TOTAL: €3,500 + IVA

Flujo de pagos:

1. €450 + IVA - Inicio Fase 1 (30%)
2. €1,050 + IVA - Entrega prototipo
3. €1,000 + IVA - Inicio Fase 2 (tras aprobación del prototipo)
4. €1,000 + IVA - Entrega sistema completo

---

5. Ventajas del Enfoque Propuesto

Validación Temprana

El prototipo con 1,000 PDFs permite validar la calidad del sistema antes de invertir en el procesamiento completo.

Eficiencia de Costes

- Vectorización solo de resúmenes (no texto completo) reduce costes >90%
- Uso de modelos optimizados para cada tarea
- Infraestructura escalable que crece con el uso

Precisión Garantizada

- El sistema siempre cita fuentes específicas
- Respuestas basadas en el documento original (no en resúmenes)

- Combinación de filtros SQL + búsqueda semántica para máxima precisión

Flexibilidad

- Posibilidad de cambiar modelos de IA según necesidades
- Base de datos que permite consultas SQL tradicionales y búsquedas semánticas
- Arquitectura modular y mantenible

Automatización

- Sistema de actualización automática
- Limpieza de convocatorias expiradas
- Bajo mantenimiento operativo
