LICDIA
Laboratorio de Investigación en Ciencia de Datos & Inteligencia Artificial

JORNADAS 2026 DE CIENCIA, TECNOLOGÍA Y EXTENSIÓN
DEPARTAMENTO DE CIENCIAS BÁSICAS

Línea de I+D - Diálogo y RAG

Consulta conversacional de normativa universitaria sobre SUDOCU: un sistema RAG para las universidades nacionales

Juan Manuel Fernández[1] - Francisco Tamasi[1] - Sebastián Marchetti[1] - Mario Oloriz[1] - Marcelo Errecalde[2]

[1] LICDIA, Departamento de Ciencias Básicas, Universidad Nacional de Luján - [2] LIDIC, Universidad Nacional de San Luis

{jmfernandez, ftamasi, smarchetti, moloriz}@unlu.edu.ar - merreca@unsl.edu.ar

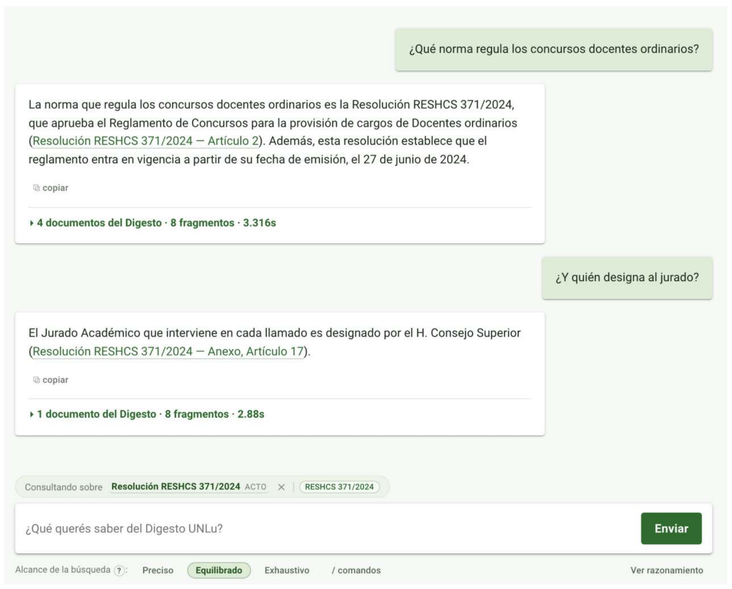
Un sistema RAG responde en lenguaje natural sobre los 21.344 actos que la UNLu publica en SUDOCU, cita cada afirmación con enlace al documento oficial y lleva la cobertura en repreguntas elípticas de 0,12 a 0,93.

1 Problema

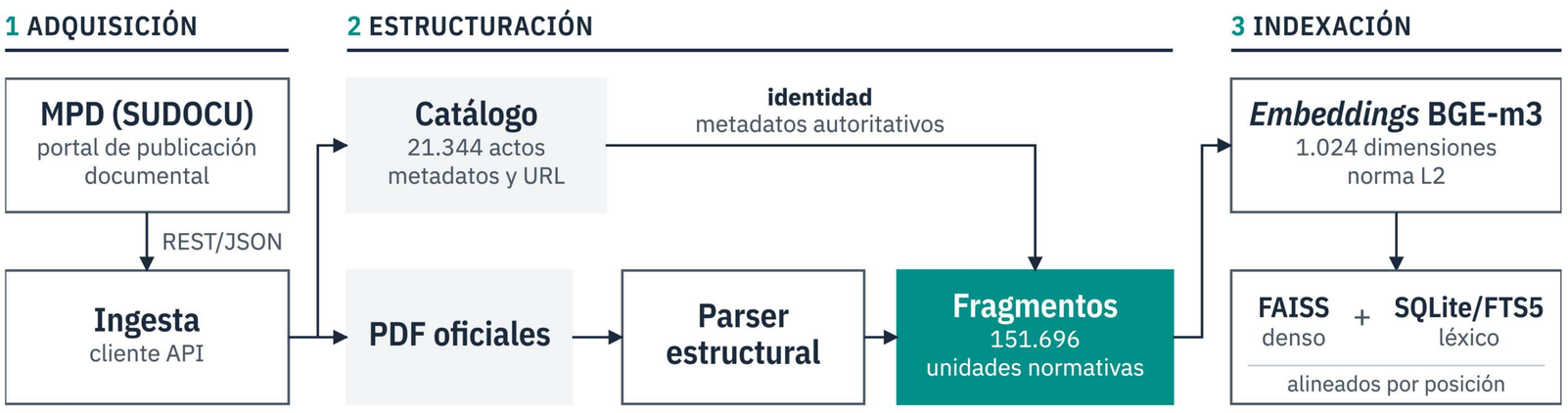
Los portales de normativa filtran por tipo, número u órgano del acto, datos que conoce la gestión pero no quien busca qué norma regula su situación. SUDOCU, usado por 37 instituciones, ofrece una base común.

2 Sistema

Cada afirmación cita el acto y enlaza al PDF oficial. Sobre el campo de escritura se muestra el estado de diálogo, que quien consulta puede corregir o descartar.



3 Arquitectura



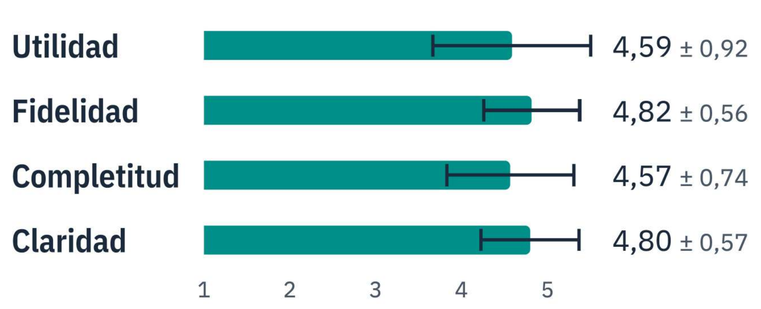
Construcción del corpus. La operación completa se ejecuta desde el panel de administración del sistema.

4 Recuperación

R@5	Identificador n = 30	Contenido n = 131	Conversacional n = 58
Solo *embeddings*	0,10	0,78	0,10
Solo BM25	0,83	0,79	0,09
Híbrido	1,00	0,82	0,12
Híbrido + estado	**1,00**	**0,82**	**0,93**

El híbrido resuelve los identificadores y el estado lleva las repreguntas de 0,12 a 0,93.

5 Validación humana



Cinco evaluadores puntuaron 15 respuestas de 1 a 5. El 97,0 % de los actos citados está entre lo recuperado.

6 Conclusiones

IDENTIFICADORES
Anclaje normativo
R@5 de 1,00 frente a 0,10 con solo *embeddings*.

DIÁLOGO
Estado con procedencia
Repreguntas elípticas de 0,12 a 0,93.

PORTABILIDAD
Solo configuración
Segunda instancia sobre el portal de la UNSL, con el mismo código.

SOBERANÍA
Infraestructura propia
Corpus, índice y búsqueda en la UNLu; modelo de generación intercambiable.

Sigue: generar las respuestas con modelos abiertos adaptados a la normativa universitaria y servidos en infraestructura propia, desplegar el sistema en otras instalaciones de SUDOCU e incorporar el digesto histórico de la UNLu (116.895 documentos).

Referencias

1. Fernández et al. Un RAG portable sobre SUDOCU. CACIC 2026.
2. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS 2020.
3. Chen et al. M3-Embedding: multi-linguality, multi-functionality, multi-granularity. Findings of ACL 2024.
4. Robertson y Zaragoza. The probabilistic relevance framework: BM25 and beyond. FnTIR 2009.
5. Cormack et al. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. SIGIR 2009.

Agradecimientos

Los autores agradecen al CIDETIC de la UNLu por la infraestructura de despliegue; al SNCAD, a Clementina XXI (IPAC, proyecto PCI-146) y a la SICyT por el apoyo en los experimentos, y a quienes participaron como evaluadores.

A su vez, un agradecimiento especial a los evaluadores externos: Humberto Soria, Bianca Chiaramello y Facundo Bobadilla.

Código

Sistema

Líneas de I+D - Investigación aplicada en ciencia de datos e IA, con fuerte impronta en los problemas regionales

Modelos de lenguaje abiertos · IA en educación · **Diálogo y RAG** · Gobernanza y transformación digital · Blockchain y credenciales · Imágenes y datos geoespaciales

licdia.unlu.edu.ar