4-3-2025

Lervis: Un RAG para publicaciones académicas.

Propuesta TFG – Palabras clave: RAG; LLM; OCR

Roi Pereira Fiuza
GRADO EN CIENCIA DE DATOS

1. Contenido

2. Ilustraciones

3. Contexto y Justificación del Trabajo

En la actualidad, el ritmo al que avanza la tecnología es exponencial, lo que genera un entorno de aprendizaje complejo y muy demandante en la mayoría de los campos de estudio. Esto ha provocado que el volumen de información que se debe leer y estudiar resulte abrumador, tanto por su cantidad como por su complejidad.

Es por ello, que la principal problemática a resolver de este TFG[1] es simplificar la forma en la que los usuarios acceden y navegan este tipo de información, específicamente centrado en el ámbito de las publicaciones académicas. Buscando un formato de acceso y de comunicación con los datos más natural y sencillo para el usuario, ya que el lenguaje natural es el factor común de comunicación para todos los seres humanos, por lo tanto, es un formato de comunicación universal y sencillo de entender.

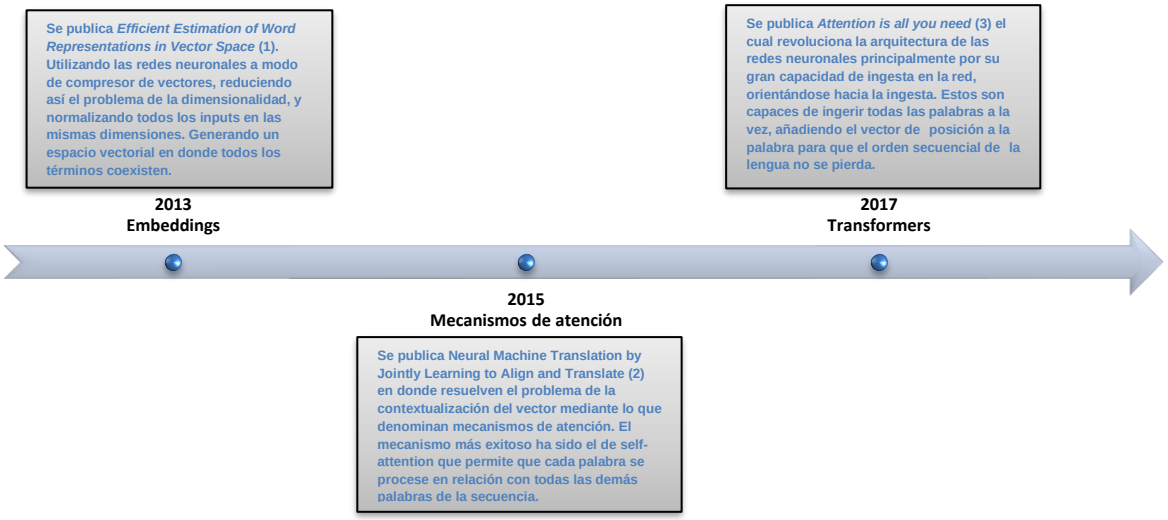
A su vez, la elección de la creación de un RAG ha sido también dada la gran versatilidad y utilidad que puede tener este tipo de proyectos en el ámbito empresarial. La comunicación mediante lenguaje natural con los datos está revolucionando nuestra sociedad y por ello la importancia del desarrollo de este proyecto, el cual posee una amplitud de usos enorme. Además, los proyectos que integran NLP están ganando cada vez más relevancia, ya que se está reduciendo la barrera de acceso a los datos para usuarios sin conocimientos avanzados en TIC o programación.

La principal motivación para realizar esta propuesta radica en su enfoque pragmático y en que el concepto es fácilmente exportable a cualquier área de negocio. En este caso, el objetivo es consultar publicaciones académicas, pero el mismo enfoque se puede extrapolar a una empresa para que pueda consultar sus datos mediante un este sistema.

Esta propuesta se sitúa entre la emergente tendencia de las inteligencias artificiales y la significativa mejora de los modelos de procesado natural del lenguaje que ha ocurrido recientemente, se enumeraran cronológicamente los tres principales avances que han permitido el desarrollo de este proyecto.

Dentro del cronograma, cada texto explicativo posee un hipervínculo a un diagrama explicativo facilitando así la explicación de cada evento.

[1] **TFG**: Trabajo final de grado



4. Descripción general

A modo de descripción general del proyecto y su arquitectura, se presenta el diagrama de flujo y una breve descripción de este:

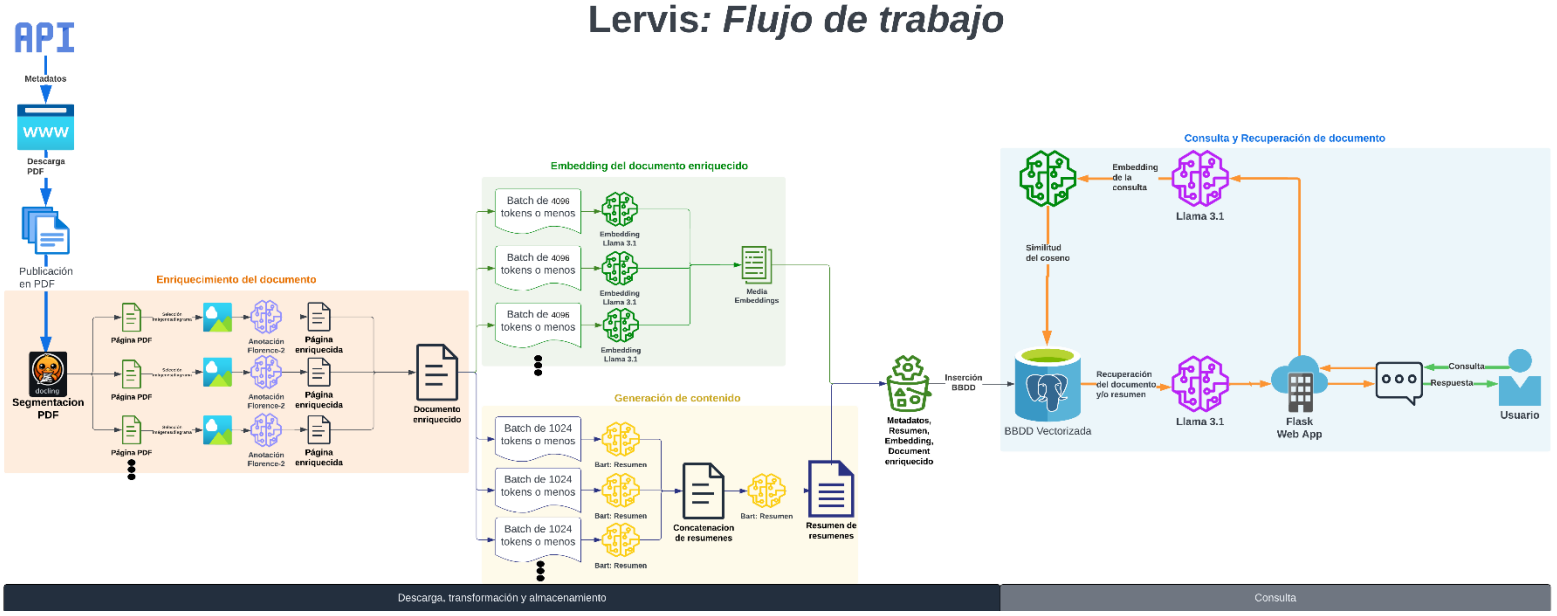


Ilustración 1: Diagrama de flujo

Como se puede observar, se puede segmentar en dos grandes bloques (parte inferior del diagrama). Dado que aún no se ha evaluado los rendimientos con el volumen de datos y procesamiento real, puede que los procesos de *Embedding* y generación de contenido se vean alterados en su posicionamiento, en lo que refiere al proceso de ejecución cambiando de un procesamiento en paralelo a un hilo único.

1. Descarga, transformación y almacenamiento:
Se accede a la API [2] de ArXiv (4) para extraer publicaciones y descargar los PDF. Luego, se segmenta cada documento con Docling (IBM) (5), convirtiéndolo en objetos JSON[3] etiquetados.

Las imágenes extraídas se anotan con Florence-2 (Microsoft) (6) para convertirlas en texto, logrando un formato unificado. Generando así una versión del PDF enriquecida mediante la anotación de imágenes.

Se genera un embedding del documento enriquecido con Llama 3.1 (7), dividiendo el contenido en bloques de máximo 4096 tokens y calculando un promedio de embeddings.

Luego, se genera un resumen utilizando bart-large-cnn (8), aplicando una estrategia de "resumen de resúmenes" para superar la limitación de 1024 tokens y optimizar la métrica ROUGE (9).

Finalmente, los datos se almacenan en una base de datos, incluyendo metadatos, el embedding del documento y el resumen generado.

1. Consulta
Se desarrolla una Flask Web App con un chat con un chatbot basado en Llama 3.1 para facilitar las consultas en un lenguaje natural. Se utiliza Ollama (10) para optimizar el modelo en un entorno con recursos computacionales limitados.

El chatbot mediante el análisis del contexto del input del usuario, ejecutará funciones específicas, como la generación de embeddings del input y la búsqueda en la BBDD[4] mediante similitud del coseno. Además, del contexto predefinido y actualizado con la versión más reciente de la BBDD y otros parámetros, facilitan la resolución de consultas sin la necesidad de contactar a la BBDD.

5. Objetivos

Objetivos generales

1. **Base de datos actualizada**: Se generarán actualizaciones de la BBDD para que englobe todas las publicaciones más recientes, estas actualizaciones serán automáticas.

[2] **API**: Interfaz de aplicación informática en inglés *Application Programming Interface*.

[3] **JSON**: formato ligero de intercambio de datos, fácil de leer y escribir para los humanos y simple de procesar para las máquinas. Se basa en la sintaxis de objetos de JavaScript, pero es independiente de cualquier lenguaje de programación. De las siglas en inglés *JavaScript Object Notation.*
[4] **BBDD**: Base de datos

2. **Consulta mediante lenguaje natural**: Facilitar la capacidad de búsqueda del usuario. Al ser mediante una conversación con el *chatbot* [5], facilita y simplifica el formato de acceso a los datos.
3. **Recomendación de publicaciones y sus resúmenes**: Sugiere artículos relacionados con la consulta ejecutada por el usuario. Además, ofrece resúmenes textuales que ayudan a la comprensión de las publicaciones rápida y eficientemente. Algo fundamental en la exploración de nuevas publicaciones no conocidas.

Objetivos específicos:

1. **Base de datos actualizada:**

    a. Actualización programada:
    b. Optimizar el almacenamiento y recuperación de datos
    c. Monitorización de la BBDD
    d. Actualización del contexto

2. **Consulta mediante lenguaje natural**

    a. Consultas en varios idiomas
    b. Tono cordial y profesional.

3. **Recomendación de publicaciones y sus resúmenes**

    a. Estructura limpia y definida
    b. Capacidad conversacional

6. Enfoque y método a seguir

Estrategia	Ventaja	Desventaja
Modelos propios	▪ Máxima personalización ▪ Modelos especializados en los datos del entreno	▪ Gran coste temporal y computacional ▪ Alto riesgo de errores ▪ Difícil acceso a datos para el entrenamiento
Modelos pre entrenados	▪ Cero costes de entreno ▪ Modelos de alta calidad ▪ Ajuste fino para mejorar la calidad ▪ Facilidad de integración ▪ Destilación de modelos en caso de ser necesario	▪ Sin control sobre los datos de entrenamiento ▪ Dependencias de terceros ▪ Potenciales problemas de compatibilidad

[5] **Chatbot**: Sistema automatizado que interactúa con usuarios a través de texto o voz, utilizando técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático (ML).

La elección de la estrategia para este proyecto ha sido el uso de modelos pre entrenados. Esta decisión se fundamenta principalmente en la gestión de los recursos disponibles, ya que el equipo está compuesto por una sola persona y el proyecto tiene un carácter académico con restricciones de tiempo y sin financiación externa.

Dado este contexto, los modelos pre entrenados permiten entregar un producto funcional y escalable, lo cual no sería posible con el enfoque de desarrollo de modelos propios, debido a los elevados costes temporales y computacionales.

Además, al optar por modelos pre entrenados, se puede aprovechar el ajuste fino para mejorar la calidad y aplicabilidad del sistema sin tener que invertir recursos significativos en la creación de un modelo desde cero.

Esta estrategia también se alinea con la necesidad de entregar un producto que sea fácilmente exportable al mundo empresarial, aprovechando las herramientas y tecnologías ya establecidas en la industria

7. Planificación

Tareas:

1. Planificación y diseño:
    a. Definición de objetivos
    b. Definición de requisitos
    c. Herramientas clave
    d. Diseño de la arquitectura y la web
    e. Configuración entornos de desarrollos y repositorios

2. Implementación de componentes principales:
    a. Creación de la BBDD
    b. Desarrollo Evaluación y métricas

3. Desarrollo de función y componentes de la ETL
    a. Funciones para extracción de datos API.
    b. Funciones segmentación y enriquecimiento del documento
    c. Funciones generación de resúmenes
    d. Funciones para generación de Embeddings
    e. Funciones de inserción de datos en la BBDD

4. Desarrollo del chatbot y web app:
    a. Creación y configuración del chatbot
    b. Función de contexto dinámico

   c. Creación web app

5. Carga de datos y evaluación funcional de consultas:

   a. Carga de datos
   b. Evaluación general del sistema
   c. Evaluación precisión y rendimiento de los modelos

6. Documentación y Entrega Final:

   a. Documentación técnica
   b. Documentación funcional
   c. Revisión de la memoria
   d. Desarrollo Presentación
   e. Desarrollo Video Presentación

Hitos:

- BBDD Creada
- Desarrollo del trabajo Fase 1: PEC 1
- Todas las funciones completadas
- Web app creada
- Carga de datos completada
- RAG funcional
- Desarrollo del trabajo Fase 2: PEC 2

Análisis de riesgos:

Se han detectado los siguientes riesgos:

- **Falta de capacidad computacional**: Individualmente se han evaluado los distintos componentes, pero no se ha evaluado el proceso en funcionamiento y es claramente un riesgo y algo que centrar la atención.
- **Dependencia de modelos externos**: Se están empleando modelos que dependen de terceras partes, principalmente mediante API. En función del tipo de fuente que fallase, el grado de impacto en el proyecto podría ser importante, especialmente si Florence-2 o Docling fallasen, ya que son componentes vitales que generan un valor añadido único en el proyecto.
- **Métricas automatizadas**: El uso de ROUGE y de BERTScore, pueden tener efectos no contemplados en por ejemplo el tono o la intencionalidad en la generación de contenido., por lo que como mitigación de riesgos se analizará periódicamente respuestas o resúmenes generados para una validación humana.
- **Riesgos de integración con Web app**: Dada la pequeña experiencia en interfaces de usuario, se detecta un riesgo menor de no obtener el resultado deseado o de que la compatibilidad con el proyecto no sea optima.
- **Riesgo de versiones**: Dado la gran entropía de nuestro proyecto en términos plataformas empleadas, existe un riesgo importante en que terceras partes realicen actualizaciones y estas generen errores o problemas en las ejecuciones.

Calendario (Diagrama Gantt):

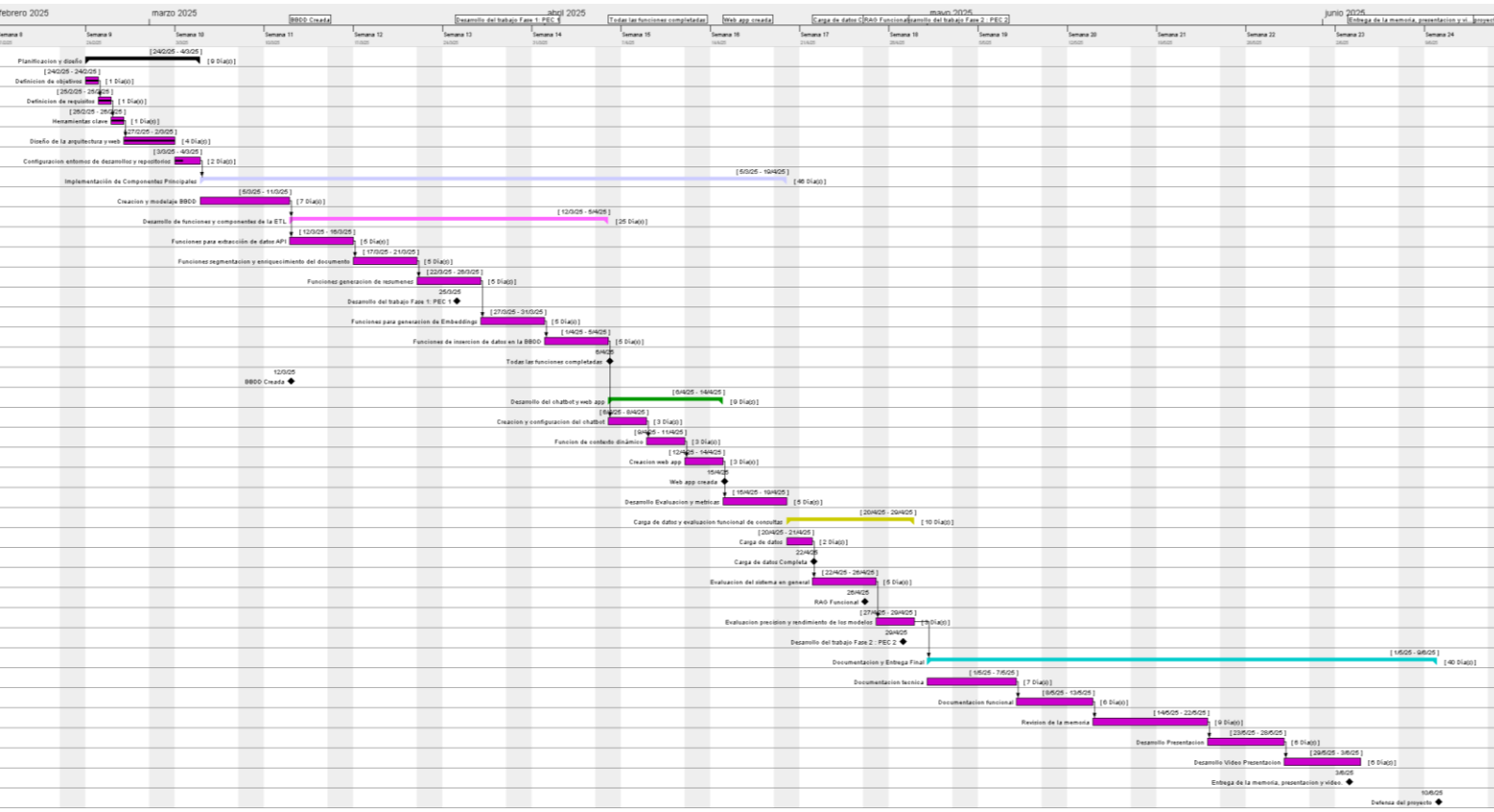


Ilustración 2: Diagrama Gantt

En el caso de que se prefiera visualizar en el navegador para una mejor resolución, acceda mediante este enlace.

8. Resultados esperados

Los entregables del proyecto serán:

- **Archivo de Python** en donde se encapsulará con un entorno virtual, todo el proyecto.
- **Documentación técnica**.
- **Documentación funcional**.
- **Memoria del proyecto**.
- **PowerPoint del proyecto**.

9. Bibliografía

1. **Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean.** Efficient Estimation of Word Representations in Vector Space. [En línea] 7 de 9 de 2013. [Citado el: 01 de 03 de 2025.] https://arxiv.org/abs/1301.3781.

2. **Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio.** Neural Machine Translation by Jointly Learning to Align and Translate. *https://arxiv.org/.* [En línea] 1 de 9 de 2014. [Citado el: 01 de 03 de 2025.] https://arxiv.org/abs/1409.0473v7.

3. **Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin.** Attention Is All You Need. [En línea] [Citado el: 01 de 03 de 2025.] https://arxiv.org/abs/1706.03762.

4. **University, Cornell.** Arxiv. [En línea] [Citado el: 30 de 12 de 2024.] https://arxiv.org/.

5. **al, Christoph Auer et.** Docling Technical Report. *ArXiv.org.* [En línea] [Citado el: 04 de 12 de 2024.] https://arxiv.org/abs/2408.09869.

6. **Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, Lu Yuan.** Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. *ArXiv.org.* [En línea] [Citado el: 03 de 12 de 2024.] https://arxiv.org/abs/2311.06242.

7. **Ollama.** llama3.1. *Ollama.* [En línea] [Citado el: 02 de 01 de 2024.] https://ollama.com/library/llama3.1.

8. **Facebook.** facebook/bart-large-cnn. *huggingface.* [En línea] [Citado el: 01 de 01 de 2025.] https://huggingface.co/facebook/bart-large-cnn.

9. **Antholgy, ACL.** ROUGE: A Package for Automatic Evaluation of Summaries. *ACL Antholgy.* [En línea] [Citado el: 30 de 11 de 2024.] https://aclanthology.org/W04-1013/.

10. **Ollama.** Ollama. [En línea] 02 de 01 de 2025. https://ollama.com/.

11. **ArXiv.org.** API Access. *ArXiv.* [En línea] [Citado el: 01 de 01 de 2025.] https://info.arxiv.org/help/api/index.html.

12. **Microsoft.** Hugging Face. *all-MiniLM-L6-v2.* [En línea] [Citado el: 01 de 11 de 2024.] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

13. **IBM.** IBM watsonx. *IBM Documentation.* [En línea] [Citado el: 04 de 12 de 2024.] https://www.ibm.com/docs/en/watsonx/saas?topic=developing-generative-ai-solutions.

14. **Wikipedia.** Escala Likert. *Wikipedia.* [En línea] [Citado el: 31 de 10 de 2024.] https://es.wikipedia.org/wiki/Escala_Likert.

15. **Huggingface.** BERTScore. [En línea] [Citado el: 03 de 01 de 2025.] https://huggingface.co/spaces/evaluate-metric/bertscore.

16. **Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin.** Attention Is All You Need. [En línea] [Citado el: 12 de 10 de 2024.] https://arxiv.org/abs/1706.03762.

17. **Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio.** Neural Machine Translation by Jointly Learning to Align and Translate. *https://arxiv.org/.* [En línea] 1 de 9 de 2014. [Citado el: 12 de 10 de 2024.] https://arxiv.org/abs/1409.0473v7.

18. **Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean.** Efficient Estimation of Word Representations in Vector Space. [En línea] 7 de 9 de 2013. [Citado el: 12 de 10 de 2024.] https://arxiv.org/abs/1301.3781.

19. **Scrum.org. [En línea] [Citado el: 09 de 01 de 2025.] https://www.scrum.org/resources/what-scrum-module.**

20. Wikipedia. Burndown Chart. [En línea] [Citado el: 09 de 01 de 2025.] https://en.wikipedia.org/wiki/Burndown_chart#:~:text=A%20burndown%20chart%20or%20burn,with%20time%20along%20the%20horizontal..

21. Springer. Springer. [En línea] [Citado el: 10 de 01 de 2025.] https://www.springer.com/gp?.

22. (IEEE), Institute of Electrical and Electronics Engineers. ieee.org. [En línea] [Citado el: 07 de 11 de 2024.] https://www.ieee.org/.