

Memoria de Prácticas Externas: Primera Parte

Zhen Feng

1. Descripción concreta y detallada de las tareas desarrolladas

Durante el periodo de prácticas, el estudiante ha estado asignado al área de desarrollo de soluciones basadas en inteligencia artificial aplicadas a la gestión de conocimiento técnico en entornos de ingeniería.

La principal actividad desarrollada ha consistido en el diseño, implementación y optimización de un sistema *Retrieval-Augmented Generation* (RAG) orientado a la consulta de documentación técnica.

En una primera fase, se desarrolló un sistema basado en las siguientes tecnologías:

- LangChain y LangGraph para la orquestación de flujos y agentes
- ChromaDB como base de datos vectorial
- Modelos de lenguaje de OpenAI para la generación de respuestas

Este sistema permite construir una base de conocimiento documental a partir de archivos técnicos y responder automáticamente a consultas de usuarios. Además, se implementó una funcionalidad adicional para la generación automática de informes técnicos (*informes*) a partir de información contextual de proyectos específicos.

Posteriormente, con el objetivo de mejorar el rendimiento del sistema, se exploró una segunda línea de desarrollo basada en búsqueda léxica (*lexical search*) como alternativa o complemento a la recuperación semántica.

Esta aproximación permitió mejorar la precisión de las respuestas en determinados tipos de consultas técnicas y reducir errores en la recuperación de información.

Adicionalmente, se realizaron tareas de evaluación del sistema, análisis de resultados, ajuste de parámetros (como *chunk size*, *top-k* y umbrales de similitud), así como iteración continua del diseño del sistema.

2. Valoración de las tareas en relación con los estudios universitarios

Las tareas desarrolladas durante las prácticas están altamente alineadas con los conocimientos adquiridos en el ámbito de la Ingeniería Informática.

En particular, se han aplicado competencias relacionadas con:

- Inteligencia Artificial y Procesamiento del Lenguaje Natural
- Arquitecturas de software y sistemas distribuidos
- Gestión y análisis de datos
- Desarrollo de aplicaciones backend

Asimismo, el uso de tecnologías emergentes como los modelos de lenguaje y las arquitecturas RAG ha permitido profundizar en áreas de gran relevancia actual dentro del sector tecnológico.

Estas prácticas han supuesto una aplicación directa de conocimientos teóricos en un entorno real, reforzando tanto las competencias técnicas como la capacidad de adaptación a nuevas herramientas.

3. Problemas planteados y procedimientos de resolución

Durante el desarrollo del proyecto se identificaron diversos retos técnicos:

3.1. Baja precisión inicial del sistema RAG

El sistema basado exclusivamente en recuperación semántica presentaba limitaciones en la precisión de las respuestas.

Solución: ajuste de parámetros de recuperación, optimización del proceso de segmentación de documentos (*chunking*) y establecimiento de métricas de evaluación.

3.2. Limitaciones del enfoque semántico

Se detectaron dificultades para recuperar información exacta en documentos técnicos altamente estructurados.

Solución: implementación de un sistema alternativo basado en búsqueda léxica y comparación entre ambos enfoques.

3.3. Evaluación del sistema

Existía la necesidad de medir objetivamente el rendimiento del sistema.

Solución: uso de métricas como *answer accuracy*, empleo de conjuntos de datos con *ground truth* y análisis de resultados.

3.4. Integración de múltiples tecnologías

La coordinación entre diferentes herramientas y frameworks aumentó la complejidad del sistema.

Solución: diseño modular, estructuración en pipelines y desarrollo iterativo.

4. Aportaciones en materia de aprendizaje

Las prácticas han contribuido significativamente al desarrollo académico y profesional del estudiante.

Se ha adquirido una comprensión profunda de:

- Arquitecturas RAG y sistemas basados en LLMs
- Evaluación de sistemas de inteligencia artificial
- Integración de múltiples tecnologías en un sistema complejo

Asimismo, se han desarrollado competencias transversales como el pensamiento crítico, la resolución de problemas y la capacidad de mejora continua.

5. Evaluación de las prácticas y sugerencias de mejora

La valoración global de las prácticas es muy positiva, destacando el uso de tecnologías avanzadas y la participación en un proyecto con aplicación real.

Como sugerencias de mejora del proceso de prácticas, se podrían considerar los siguientes aspectos:

- Definir con mayor claridad los objetivos concretos de cada fase al inicio del periodo de prácticas.
- Proporcionar una documentación inicial más detallada sobre la arquitectura existente y las herramientas utilizadas.
- Planificar con antelación los criterios de evaluación del trabajo desarrollado.

En conclusión, las prácticas han cumplido ampliamente los objetivos formativos, proporcionando una experiencia relevante y alineada con el perfil profesional del estudiante.

Memoria de Prácticas Externas: Segunda Parte

Zhen Feng

1. Descripción concreta y detallada de las tareas desarrolladas

Durante la segunda parte del periodo de prácticas, el estudiante continuó trabajando en el desarrollo de soluciones basadas en inteligencia artificial aplicadas a la gestión y consulta de documentación técnica en entornos de ingeniería.

A diferencia de la primera fase, centrada principalmente en el diseño inicial de un sistema *Retrieval-Augmented Generation* (RAG) basado en recuperación semántica, esta segunda etapa se orientó hacia el diseño, implementación y evaluación de una nueva arquitectura RAG impulsada principalmente por búsqueda léxica (*lexical search*).

El objetivo principal de esta nueva línea de trabajo fue mejorar la precisión del sistema en consultas técnicas donde la coincidencia exacta de términos, códigos, nombres de componentes, referencias documentales o expresiones específicas resulta especialmente importante. Para ello, se diseñó un nuevo sistema de recuperación basado en criterios léxicos, capaz de identificar fragmentos relevantes de documentación técnica de forma más controlada y precisa.

Los resultados obtenidos mostraron una mejora considerable en la precisión de las respuestas, especialmente en comparación con enfoques basados exclusivamente en recuperación semántica. Esta mejora permitió comprobar la utilidad de los métodos léxicos en contextos documentales altamente estructurados, donde pequeñas diferencias terminológicas pueden afectar significativamente a la calidad de la respuesta generada.

Posteriormente, se desarrolló una segunda mejora del sistema mediante la integración de etiquetas (*etiquetas*) asociadas a los documentos o a sus fragmentos. Estas etiquetas permitieron introducir información adicional de clasificación y contexto, facilitando una recuperación más precisa y reduciendo la probabilidad de seleccionar fragmentos irrelevantes. La combinación entre búsqueda léxica y etiquetado documental permitió optimizar todavía más el rendimiento del sistema.

Además del desarrollo del sistema RAG, se llevaron a cabo tareas de integración y

validación técnica. En particular, se realizaron pruebas de integración para comprobar el funcionamiento del sistema completo en un entorno más cercano al despliegue real. Dentro de esta fase, se verificó la posibilidad de desplegar el sistema sobre una base de datos PostgreSQL, lo que permitió avanzar hacia una arquitectura más robusta, persistente y adecuada para su uso en escenarios reales.

Finalmente, se realizaron pruebas relacionadas con tecnologías de reconocimiento óptico de caracteres y comprensión documental. En concreto, se evaluó el uso de Docling para el procesamiento OCR de documentos técnicos, así como la posibilidad de utilizar un modelo VLM desplegado localmente, en este caso NuExtract, para mejorar la extracción de información a partir de documentos visuales o parcialmente estructurados.

Estas pruebas permitieron analizar las capacidades y limitaciones de distintas herramientas de OCR y modelos multimodales en el contexto de documentación técnica, especialmente cuando los documentos contienen tablas, planos, capturas, esquemas o información visual relevante.

2. Valoración de las tareas en relación con los estudios universitarios

Las tareas desarrolladas durante esta segunda parte de las prácticas presentan una relación directa con los estudios universitarios de Ingeniería Informática, especialmente en las áreas de inteligencia artificial, procesamiento del lenguaje natural, bases de datos, recuperación de información y arquitectura de sistemas software.

En primer lugar, el diseño de un sistema RAG basado en búsqueda léxica permitió aplicar conocimientos relacionados con la recuperación de información, el tratamiento de texto y la estructuración de documentos. La comparación entre búsqueda semántica y búsqueda léxica también ayudó a comprender mejor las ventajas y limitaciones de diferentes estrategias de recuperación en sistemas basados en modelos de lenguaje.

En segundo lugar, la incorporación de etiquetas al sistema permitió trabajar con mecanismos de clasificación, filtrado y enriquecimiento de datos. Esta tarea está relacionada con conceptos propios de la gestión de conocimiento, el modelado de información y el diseño de sistemas capaces de aprovechar metadatos para mejorar su rendimiento.

Asimismo, la integración del sistema con PostgreSQL permitió aplicar conocimientos de bases de datos relacionales, persistencia de datos y despliegue de aplicaciones. Esta parte de la práctica resultó especialmente relevante desde el punto de vista de la ingeniería de software, ya que permitió pasar de un prototipo experimental a una solución más estructurada y preparada para un entorno de uso real.

Por otra parte, las pruebas con tecnologías OCR y modelos VLM locales permitieron

profundizar en el procesamiento de documentos no puramente textuales. Esta actividad está relacionada con áreas actuales de la inteligencia artificial, como la comprensión multimodal de documentos, la extracción automática de información y la evaluación de herramientas basadas en modelos avanzados.

En conjunto, las tareas realizadas han supuesto una aplicación práctica de competencias adquiridas durante la formación universitaria, al mismo tiempo que han permitido adquirir experiencia con tecnologías emergentes y metodologías actuales en el desarrollo de sistemas inteligentes.

3. Problemas planteados y procedimientos de resolución

Durante esta segunda parte de las prácticas se identificaron diversos retos técnicos relacionados con la precisión del sistema, la integración de componentes y el procesamiento de documentos complejos.

3.1. Limitaciones de la recuperación semántica en documentación técnica

Uno de los principales problemas observados fue que los sistemas basados únicamente en recuperación semántica no siempre resultaban adecuados para documentos técnicos altamente estructurados. En algunos casos, el sistema recuperaba fragmentos conceptualmente similares, pero no necesariamente exactos, lo que podía afectar a la precisión de las respuestas.

Solución: se diseñó e implementó una nueva arquitectura RAG impulsada por búsqueda léxica. Este enfoque permitió priorizar coincidencias exactas de términos, referencias, códigos y expresiones técnicas, mejorando la precisión en consultas donde la literalidad del contenido era especialmente importante.

3.2. Necesidad de mejorar el contexto de recuperación

Aunque la búsqueda léxica permitió obtener resultados muy positivos, se identificó la necesidad de incorporar información adicional para guiar mejor el proceso de recuperación.

Solución: se integró un sistema de etiquetas asociado a documentos o fragmentos documentales. Estas etiquetas sirvieron como metadatos para clasificar la información y facilitar una recuperación más precisa. La combinación de búsqueda léxica y etiquetas permitió mejorar la selección de fragmentos relevantes y reducir errores en las respuestas generadas.

3.3. Validación del sistema completo

Una vez desarrollado el nuevo sistema, fue necesario comprobar su funcionamiento de forma integrada, evaluando no solo componentes individuales, sino también el flujo completo de procesamiento, recuperación y generación de respuestas.

Solución: se realizaron pruebas de integración para verificar el comportamiento del sistema en conjunto. Estas pruebas permitieron detectar posibles problemas de coordinación entre módulos y comprobar que la arquitectura podía funcionar de manera coherente en un entorno más próximo al despliegue real.

3.4. Despliegue sobre una base de datos PostgreSQL

Otro reto importante fue adaptar el sistema para que pudiera desplegarse sobre una base de datos PostgreSQL. Esto implicaba asegurar la correcta gestión de los datos, la persistencia de la información y la conexión entre los módulos del sistema y la base de datos.

Solución: se realizaron pruebas de despliegue y conexión con PostgreSQL, comprobando la viabilidad de utilizar esta base de datos como infraestructura de almacenamiento para el sistema. Esto permitió avanzar hacia una solución más estable y escalable.

3.5. Procesamiento de documentos técnicos mediante OCR y VLM

Finalmente, se identificaron dificultades asociadas al procesamiento de documentos técnicos con contenido visual, como tablas, capturas, planos o esquemas. Las herramientas OCR convencionales no siempre conseguían extraer correctamente toda la información relevante.

Solución: se realizaron pruebas con Docling para evaluar su capacidad de procesamiento OCR y con NuExtract desplegado localmente como modelo VLM. Estas pruebas permitieron comparar diferentes enfoques de extracción de información y analizar sus ventajas y limitaciones en documentos técnicos reales.

4. Aportaciones en materia de aprendizaje

La segunda parte de las prácticas ha contribuido de forma significativa al aprendizaje técnico y profesional del estudiante.

En primer lugar, se ha adquirido una comprensión más profunda de las arquitecturas RAG y de sus distintos enfoques de recuperación. La comparación práctica entre recuperación semántica, búsqueda léxica y estrategias híbridas permitió entender que no existe una

única solución óptima para todos los contextos, sino que la elección del método depende en gran medida de la naturaleza de los documentos y del tipo de consulta.

En segundo lugar, el desarrollo de un sistema basado en búsqueda léxica permitió reforzar conocimientos sobre recuperación de información, tratamiento de texto, indexación y evaluación de resultados. Además, la integración de etiquetas ayudó a comprender la importancia de los metadatos y de la estructuración previa de la información en sistemas de consulta documental.

También se adquirió experiencia en tareas de integración y despliegue, especialmente mediante el uso de PostgreSQL como base de datos para el sistema. Esta parte de la práctica permitió conectar el desarrollo de inteligencia artificial con aspectos más propios de la ingeniería de software, como la persistencia de datos, la modularidad y la preparación de una arquitectura para entornos reales.

Por otra parte, las pruebas con OCR y modelos VLM permitieron ampliar el conocimiento sobre procesamiento documental avanzado. Se comprendieron mejor las dificultades asociadas a documentos técnicos complejos, así como la importancia de evaluar cuidadosamente las herramientas de extracción antes de incorporarlas a un sistema RAG.

Desde el punto de vista profesional, las prácticas también permitieron desarrollar competencias transversales como la resolución de problemas, la evaluación crítica de resultados, la capacidad de comparación entre alternativas tecnológicas y la adaptación continua durante el desarrollo de un proyecto.

5. Evaluación de las prácticas y sugerencias de mejora

La valoración global de esta segunda parte de las prácticas es muy positiva. El trabajo realizado permitió participar en una fase más avanzada del desarrollo de un sistema inteligente, pasando de la experimentación inicial a una arquitectura más precisa, evaluada e integrada.

Uno de los aspectos más positivos fue la posibilidad de trabajar con tecnologías actuales y de aplicación real, como sistemas RAG, búsqueda léxica, bases de datos PostgreSQL, herramientas OCR y modelos VLM. Esto permitió adquirir una visión más completa del ciclo de desarrollo de una solución basada en inteligencia artificial, desde el diseño del sistema hasta su evaluación e integración.

Además, la práctica permitió afrontar problemas reales relacionados con la precisión de las respuestas, la calidad de la recuperación documental y las limitaciones de las herramientas de procesamiento automático. Esto resultó especialmente formativo, ya que obligó a analizar errores, comparar alternativas y justificar técnicamente las decisiones adoptadas.

Como sugerencias de mejora del proceso de prácticas, se podrían considerar los siguientes aspectos:

- Definir con mayor claridad los objetivos concretos de cada fase al inicio del periodo de prácticas.
- Proporcionar una documentación inicial más detallada sobre la arquitectura existente y las herramientas utilizadas.
- Planificar con antelación los criterios de evaluación del trabajo desarrollado.

En conclusión, esta segunda parte de las prácticas ha resultado una experiencia altamente enriquecedora, tanto desde el punto de vista académico como profesional. Las tareas realizadas han estado estrechamente relacionadas con la Ingeniería Informática y han permitido profundizar en áreas de gran relevancia actual, especialmente en el desarrollo de sistemas RAG, recuperación de información, integración de bases de datos y procesamiento inteligente de documentación técnica.