Embeddings, similitud semántica y LLMs: Cálculo matemático y aplicaciones en procesamiento de lenguaje natural

Adrián Izquierdo Abril

adriania@unex.es

Departamento de Matemáticas. Universidad de Extremadura, Cáceres (España)

Palabras Clave: Modelos de Lenguaje; Word Embeddings; Vectores Semánticos; Procesamiento del Lenguaje Natural

AMS: 60J80

Resumen

La inteligencia artificial, y en particular los modelos de lenguaje de gran tamaño (*Large Language Models*, LLM), están adquiriendo un papel cada vez más relevante en múltiples sectores de la sociedad. Desde el desarrollo de *chatbots* personalizados hasta la automatización de tareas complejas en entornos profesionales, estas tecnologías están transformando la forma en que interactuamos con la información.

Mis líneas de investigación se han centrado en el estudio de los LLMs y en el desarrollo e implementación de una aplicación basada en esta tecnología. Dicha aplicación permite, por un lado, la generación automática de fichas resumen a partir de convocatorias y sus bases reguladoras, y por otro, la incorporación de un chatbot personalizado, capaz de responder de manera precisa a cualquier consulta realizada sobre el contenido de un documento.

El principal desafío en el desarrollo de esta aplicación ha sido garantizar la privacidad de la información proporcionada por el usuario, mediante su ejecución en un entorno local, sin comprometer el rendimiento ni los tiempos de respuesta [1].

Este desafío ha implicado la necesidad de llevar a cabo un estudio detallado de todo el proceso de generación de respuestas, desde la selección y evaluación del modelo de cálculo de embeddings hasta la elección del LLM más adecuado, considerando criterios de precisión, eficiencia y ejecución local.

El primer paso del flujo del programa —y en el que se centrará esta ponencia— consiste en la transformación del documento seleccionado en una representación matemática, mediante el cálculo de sus vectores de embeddings. Esta representación vectorial es fundamental para que el modelo pueda recuperar información relevante y generar respuestas precisas.

Para llevar a cabo este proceso, el documento debe ser fragmentado en secciones más pequeñas, sobre las cuales se aplica un modelo de embedding capaz de reducir la dimensionalidad y generar un vector representativo para cada fragmento [2]. Esta representación permite que fragmentos con contenido semántico similar estén asociados a vectores cercanos en el espacio vectorial, lo cual es esencial para que el modelo pueda recuperar información relevante y generar respuestas coherentes.

Cada vez que se realiza una consulta al modelo, dicha solicitud se transforma en un vector de embedding que permite compararla con los vectores previamente generados para los fragmentos del documento. A partir de esta comparación, basada en la similitud de coseno, se recuperan únicamente los fragmentos más relevantes —normalmente un número reducido— que contienen

la información más relacionada con la consulta [3]. Este proceso no solo mejora la precisión de las respuestas, sino que también permite reducir los costes computacionales, ya que el modelo de lenguaje solo recibe como contexto los fragmentos más significativos, evitando procesar el documento completo.

Todo este proceso pone de manifiesto la relevancia de una representación matemática adecuada del lenguaje como base fundamental para obtener resultados precisos y eficientes en tareas de procesamiento de lenguaje natural.

Agradecimientos: Quiero expresar mi agradecimiento a Carlos Javier Pérez Sánchez y a Miguel Ángel Vega Rodríguez por su constante guía y seguimiento en el proyecto realizado.

Este proyecto ha recibido cofinanciación del Programa Europa Digital (DIGITAL) de la Unión Europea, en virtud del acuerdo de subvención n.º 101083667-Tech4EfficiencyEDIH. Está cofinanciado por el Ministerio de Industria, Comercio y Turismo del Gobierno de España, en el marco de los fondos «Next Generation/PRTR».

Referencias

- [1] Maiya, A. S. (2025). *OnPrem.LLM: A Privacy-Conscious Document Intelligence Toolkit*. Institute for Defense Analyses, Alexandria, VA, USA.
- [2] Zhang, C., Peng, B., Sun, X., et al. (2024). *From Word Vectors to Multimodal Embeddings: Techniques, Applications, and Future Directions for Large Language Models*.
- [3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Google Inc., Mountain View, CA.