

Taller IA Generativa

Mayerly Johana Tellez Gamez Michael Caicedo

I. PRIMER PUNTO

Se usó el documento "Capitulo7Punto1.ipynb" como base en el ejercicio ya que ya tiene incorporada la api de ollama. Antes, el código fallaba porque intentaba leer archivos desde una carpeta local (*./Textos/*) que no existía en el entorno de ejecución y, además, solo estaba preparado para procesar archivos de texto plano (.txt). Esto generaba el error FileNotFoundError y posteriormente un ValueError al intentar indexar una lista vacía. Ahora, el sistema fue corregido para permitir la carga manual de archivos directamente desde el entorno de Google Colab, guardándolos automáticamente en una carpeta temporal, y se amplió la función de lectura para soportar tanto archivos .txt como .pdf utilizando la librería PyPDF2. Gracias a estos cambios, el flujo completo —desde la carga de documentos hasta la generación de embeddings— funciona correctamente. De modo que al final este documento se debe ejecutar celda por celda y sirve a partir de la funcionalidad de subir documentos directamente desde el computador.

Descripción del código

El código inicia con la instalación y configuración del entorno de ejecución. Se instala el servidor Ollama, encargado de ejecutar los modelos de lenguaje localmente, y se descargan los modelos llama3.1 (para la generación de texto) y nomic-embed-text (para crear embeddings o representaciones vectoriales del contenido). Además, se instalan las bibliotecas de LangChain y Chroma, que permiten la creación del flujo RAG y la gestión de la base vectorial.

Posteriormente, el código define funciones para cargar y procesar documentos desde archivos locales. A través de la librería PyPDF2, se extrae texto de archivos PDF, y los archivos de texto plano se leen directamente. Estos textos se convierten en vectores mediante el modelo de embeddings de Ollama y se guardan en una base de datos local gestionada por Chroma, en una colección denominada IA-minirobots. Este paso garantiza la persistencia de los datos y la posibilidad de realizar búsquedas semánticas eficientes. Luego, se define un *prompt* del sistema, donde se instruye al modelo a comportarse como un asistente académico experto.

Finalmente, el sistema integra los componentes de recuperación y generación en una cadena RAG mediante las funciones `create_stuff_documents_chain` y `create_retrieval_chain`. Una vez configurado, el asistente es capaz de responder consultas académicas tales como: "¿Qué es la inteligencia?".

II. SEGUNDO PUNTO

Hacer un estudio de diferentes chatbots de IA, cuyo costo sea 0, para actividades como: resumen de texto, generación

de texto, generación de imágenes, generación de audio y generación de video.

- Resúmenes de texto: Se le pidió a cada chatbot generar un resumen de los conceptos más importantes del documento "5.Redes Neuronales1 2025-2.pdf". Prompt: Hola, dame un resumen de los conceptos más importantes del documento. Luego dime cuantas palabras usaste, cuantos y cuales conceptos abordaste, cuantas ecuaciones colocaste, y más importante cuante te demoraste generando la respuesta de ese resumen. Y al final luego de generar la respuesta, dame la respuesta en código latex, solo la respuesta sin la estructura del documento.

Respuestas:

- Deepseek 3.2:

1. Introducción a las RNA: Modelos inspirados en el cerebro para reconocimiento de patrones y aprendizaje a partir de datos.

2. Neurona Artificial:

$$Z = \sum_{j=1}^n w_j x_j + b, \quad A = g(Z)$$

Funciones de activación: Sigmoid, ReLU.

3. Aprendizaje Supervisado: Entrenamiento con pares (X, Y_r) .

4. Backpropagation:

$$\frac{\partial E}{\partial W^{(l)}} = -\delta^{(l)} A^{(l-1)^T}, \quad \delta^{(l)} = \text{diag}(\sigma'(Z^{(l)})) \cdot e$$

5. Estructura de RNA:

$$Z^{(i)} = W^{(i)T} A^{(i-1)} + B^{(i)}, \quad A^{(i)} = g(Z^{(i)})$$

6. Deep Learning y CNN: Múltiples capas, convoluciones, pooling.

7. Función de Pérdida:

$$J(W) = \frac{1}{2m} \sum_{i=1}^m (h_w(X_{(i)}) - Y_{(i)})^2$$

Detalles de la respuesta:

Tiempo de respuesta: 4min

Palabras usadas: 398

Conceptos abordados: 10

Ecuaciones incluidas: 4

Análisis: Abordó 10 conceptos del documento, entre conceptos matemáticos como conceptuales. No generó bien las estadísticas, se demoró unos segundos en generar la respuesta y puso 4 minutos,

tampoco logró generar su propia respuesta en el formato latex puesto que como se puede apreciar quedó incompleta, no muestra ni conceptos textuales ni todos los conceptos. No hizo el listado neto de conceptos que pedí.

- ChatGPT-5: El documento “*Introducción a las Redes Neuronales Artificiales*” explica el funcionamiento, estructura y aprendizaje de las *redes neuronales artificiales (RNA)*, inspiradas en el cerebro humano.
 1. *Modelo biológico*: La neurona natural recibe impulsos eléctricos a través de las dendritas, los procesa en el soma y transmite la señal por el axón mediante sinapsis.
 2. *Modelo artificial*: Representa matemáticamente la neurona biológica. Cada entrada x_i se pondera con un peso w_i , se suma con un sesgo b , y se obtiene una salida $z = b + \sum w_i x_i$.
 3. *Funciones de activación*: Introducen no linealidad; las más comunes son la *sigmoide* $\sigma'(z) = \sigma(1 - \sigma)$ y la *ReLU*, utilizada en redes profundas.
 4. *Clasificación lógica*: Se muestran ejemplos con compuertas *AND*, *OR*, *NOT*, *NAND* y *NOR*, demostrando cómo los pesos definen el comportamiento de la red.
 5. *Aprendizaje supervisado*: La red ajusta sus pesos comparando salidas calculadas con salidas esperadas mediante una *función de pérdida* $E = \frac{1}{2}(Y_r - Y)^2$.
 6. *Retropropagación (Backpropagation)*: Método para minimizar el error propagándolo hacia atrás en la red y actualizando los pesos con la tasa de aprendizaje η .
 7. *Propagación hacia adelante*: En cada capa i , las salidas se calculan como

$$Z^{(i)} = B^{(i)} + (W^{(i)})^T A^{(i-1)}, \quad A^{(i)} = g(Z^{(i)}).$$

8. *Aprendizaje profundo (Deep Learning)*: Se extiende el concepto a redes con muchas capas ocultas, aplicando los mismos principios de retropropagación.
9. *Redes convolucionales (CNN)*: Aplican filtros a imágenes mediante *convolución* y *pooling* (como *max pooling*) para extraer características relevantes y reducir tamaño.
10. *Función de costo general*:

$$J(W) = \frac{1}{2m} \sum_{i=1}^m (h_W(X) - Y)^2,$$

mide el error promedio en un conjunto de entrenamiento.

Detalles de la respuesta:

Palabras usadas: 273.

Conceptos abordados: 10.

Ecuaciones incluidas: 3.

Tiempo de generación: 7.6 segundos.

Análisis: Generó de mejor forma el recorrido de los conceptos, no usó tantas palabras como el anterior, tampoco recurrió a muchas ecuaciones para explicar el documento. Le faltó generar el listado neto de los conceptos y marcó bien las estadísticas.

- Gemini 2.5: Las *Redes Neuronales Artificiales (RNA)* son modelos matemáticos simplificados inspi-

rados en el cerebro animal, siendo la columna vertebral de la Inteligencia Artificial moderna, incluyendo la arquitectura *Transformer*.

La unidad básica de una RNA es la *Neurona Artificial o Perceptrón*, la cual simula a la neurona natural. Recibe un vector de entradas X (rasgos o *features*), las pondera con un vector de *Pesos W* (que representan las *Sinapsis*) y produce una señal de salida Z . La neurona incorpora un *Sesgo (b)* que representa el umbral de disparo. La ecuación fundamental para la entrada neta Z es $Z = \sum_{j=1}^n w_j x_j + b$.

Para introducir complejidad (no linealidad) a la red, se utiliza la *Función de Activación g(Z)*. Las funciones derivables más comunes son la *Sigmoide* y *ReLU (Rectified Linear Activation)*, siendo la elección de la función la que permite a las redes no ser solo combinaciones lineales.

Las funciones principales de las RNA son la *Clasificación* y la *Predicción*. El *Aprendizaje Supervisado* es el proceso de *entrenamiento* donde se proporcionan pares de entrada-salida (características X y rótulos Y). El objetivo es minimizar el *Error* (diferencia entre la salida de la red y el rótulo deseado) ajustando los pesos. Este proceso busca la *Generalización*, evitando el *underfitting* o *overfitting*. Las redes más simples son las de *Propagación Directa (Feedforward)*. El algoritmo clave para el ajuste de pesos en el entrenamiento es la *Retropropagación (Backpropagation)*, que utiliza la *Regla de la Cadena* para calcular la derivada de la *Función de Pérdida (E)* respecto a los pesos en cada capa, desde la capa de salida hacia atrás, con el fin de minimizarlos.

Detalles de la respuesta:

Tiempo de respuesta: 6 segundos

Palabras usadas: 163

Conceptos abordados: 17

Ecuaciones incluidas: 1

Análisis: Abordó varios conceptos sobre todo textuales, logró generar de forma adecuada las estadísticas y su enfoque fue mucho mejor ya que al ser un resumen las ecuaciones no son el mejor enfoque para dar información de primera mano y no recurrió tanto a ello y en su lugar generó una cadena de ideas con los conceptos que estaban en el documento. Además, se incluye a continuación el listado de conceptos y generó bien las estadísticas.

Lista de conceptos:

- * Redes Neuronales Artificiales (RNA)
- * Transformer
- * Neurona Artificial/Perceptrón
- * Vector de Entradas (X)/Features
- * Pesos (W)/Sinapsis
- * Señal de Salida (Z)
- * Sesgo (b)/Bias/Umbral
- * Función de Activación ($g(Z)$)
- * Función Sigmoide
- * Función ReLU

- * Clasificación
- * Predicción
- * Aprendizaje Supervisado/Entrenamiento
- * Error
- * Generalización (*Underfitting/Overfitting*)
- * Propagación Directa (*Feedforward*)
- * Retropropagación (*Backpropagation*)
- * Función de Pérdida (E)
- * Regla de la Cadena

- Generación de texto e imagen

Prompt: Genera un cuento de 100 palabras acerca de una niña con ojos verdes que se encuentra con un oso tomando café con leche en un bar. Genera la imagen que representa el cuento.

- Deepseek 3.2:

El oso del café con leche

Clara, de ojos verdes como esmeraldas olvidadas, encontró al oso en la última mesa del bar. Sobre la mesa, una taza minúscula temblaba en su zarpaz.

“¿Me invitas a uno?”, preguntó la niña.

El oso, de traje ráido, asintió. Le sirvieron una espuma blanca con manchas de café.

“Extraño la miel del bosque”, confesó él con voz de trueno apagado.

Clara le ofreció su pastel de miel. Por un instante, entre sorbos, el oso no fue una rareza, sino un compañero de viaje que también añoraba su hogar. Aquí tienes la imagen que representa el cuento:
![Una niña con ojos verdes y un oso enorme, pero amigable, sentados en una mesa pequeña de un bar acogedor. El oso viste un traje ráido y sostiene con cuidado una taza de café con leche. La niña le ofrece un pedazo de pastel de miel. La escena tiene un ambiente mágico y nostálgico, iluminada por la luz cálida del local.]

Análisis: Tuvo en cuenta todos los parámetros brindados pero no puede generar la imagen así que la describió

- ChatGPT-5:



Fig. 1: Imagen generada por ChatGPT

Análisis: Se demoró un montón generando la imagen y no generó texto, además el contexto de la imagen no está en un bar

- Gemini 2.5: Conocí a una niña con ojos verdes que se encontró con un oso tomando café con leche en

un bar. El oso estaba leyendo el periódico y ella no podía creer lo que veía. La niña se acercó al oso y le preguntó si podía sentarse con él. El oso levantó la vista y sonrió. Le ofreció a la niña un sorbo de su café con leche y los dos se hicieron amigos. La niña y el oso hablaron durante horas y el oso le contó historias de la vida en el bosque. La niña estaba fascinada y sabía que nunca olvidaría a su nuevo amigo.

Aquí tienes una imagen que representa la historia:



Fig. 2: Imagen generada por Gemini

Análisis: Generó la imagen y el cuento muy bien, aunque no parece estar en un bar, la imagen está muy bien detallada.

- Generación de audio y video

Prompt: Genera el audio y video de un gato haciendo algo tierno

- Deepseek 3.2:

No, no puedo generar videos, animaciones ni ningún contenido multimedia en formato de video.

No, no puedo generar archivos de audio ni sonido.

Soy un modelo de texto puro y no tengo capacidades de síntesis de voz o generación de audio.

- ChatGPT 2.5:

Actualmente no puedo generar ni adjuntar archivos de video directamente aquí (solo imágenes y audio con instrucciones).

Actualmente no puedo generar ni adjuntar archivos de audio directamente desde este entorno.

- Gemini 2.5:

Lo siento, como un modelo de lenguaje, no tengo la capacidad de generar archivos de audio o de hablar. Solo puedo responder con texto escrito.

No, no puedo generar videos. Soy un modelo de lenguaje y mi capacidad se limita a producir y procesar texto.

En general las tres herramientas dieron la información que requería, en especial ChatGPT y Gemini cumplían con la mayoría aunque Deepseek se quedó un poco atrás para procesar la información. Ninguna logra generar por sí misma audio o video así que no tiene sentido la comparación en ese aspecto.

III. TERCER PUNTO

Mejorar el sistema RAG para crear un chat que permita acceder a la información actualizada de la Universidad Nacional, u otra empresa o entidad

En este punto el objetivo principal de este código es construir un sistema que permita responder preguntas relacionadas con la Universidad Nacional de Colombia (UNAL) utilizando información actualizada proveniente en este caso de la página "Agencia de Noticias oficial" <https://agenciadenoticias.unal.edu.co/>.

Descripción del código

En primer lugar, el script instala y configura las dependencias necesarias, incluyendo los modelos llama3.1 (para generación de texto) y nomic-embed-text (para la creación de embeddings o representaciones vectoriales del texto). Posteriormente, se importan las bibliotecas LangChain, requests, y BeautifulSoup, las cuales permiten la conexión entre el modelo de lenguaje y las fuentes de información externas.

A continuación, se configura el modelo de lenguaje llama3.1 para la generación de respuestas y el modelo de embeddings nomic-embed-text para transformar los textos de las noticias en vectores numéricos. Estos vectores se almacenan en una base de datos local llamada Chroma, cuya carpeta de persistencia se define como chroma_unal.

Luego, se define la función `get_unal_news()`, encargada de extraer automáticamente las noticias más recientes desde el sitio web de la Agencia de Noticias UNAL (<https://agenciadenoticias.unal.edu.co/>). Para ello, se realiza una solicitud HTTP a la página, se analiza el contenido HTML con la librería BeautifulSoup y se seleccionan los artículos que contienen títulos, resúmenes y enlaces. La función devuelve una lista de textos concatenados con esta información.

Posteriormente, se define un *prompt* del sistema que instruye al modelo para responder como un asistente especializado en la Universidad Nacional de Colombia, utilizando exclusivamente información proveniente de la Agencia de Noticias UNAL. Este mensaje se combina con la entrada del usuario mediante una cadena de procesamiento (`create_retrieval_chain`), la cual implementa la arquitectura RAG al unir el componente de recuperación de información con el modelo generativo.

Finalmente, se define la función `chat_unal(query)`, que recibe una consulta en lenguaje natural, invoca la cadena RAG y devuelve la respuesta generada por el modelo.

IV. CUARTO PUNTO

Hacer un estudio sobre las posibilidades de la IA Generativa en el mundo físico y sus posibilidades en nuestro país.

En Colombia, la implementación de la Política Nacional de Inteligencia Artificial – CONPES 4144 (2025) del Departamento Nacional de Planeación (DNP) marca una oportunidad estratégica al promover la investigación y adopción responsable de la IA con una inversión pública significativa (DNP, 2025). Además, estudios nacionales sobre drones agrícolas (Universidad UDCA, 2023) y gemelos

digitales educativos (RedALyC, 2023) demuestran el interés por integrar IA generativa en entornos reales, mientras que investigaciones internacionales como Evolution 6.0: Evolving Robotic Capabilities Through Generative Design (Khan et al., 2025, arXiv) evidencian cómo esta tecnología permite a robots crear herramientas y adaptarse autónomamente. Así, nuestro país tiene el potencial de convertirse en un referente regional en innovación aplicada, siempre que fortalezca su infraestructura digital, fomente la formación en ingeniería de IA y promueva alianzas entre Estado, academia y sector privado para trasladar la creatividad artificial al desarrollo físico y productivo nacional. [4] [5]

REFERENCIAS

- [1] OpenAI. (2025). *ChatGPT (versión GPT-5)* [Modelo de lenguaje de gran escala]. Disponible en: <https://chat.openai.com>
- [2] Google DeepMind. (2025). *Gemini (versión 1.5)* [Modelo de lenguaje multimodal]. Disponible en: <https://gemini.google.com>
- [3] DeepSeek AI. (2025). *DeepSeek (versión R1)* [Modelo de lenguaje de código abierto]. Disponible en: <https://www.deepseek.com>
- [4] Departamento Nacional de Planeación (DNP). (2025). *Política Nacional de Inteligencia Artificial (Documento CONPES 4144)*. Bogotá, Colombia. Disponible en: <https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/4144.pdf> :contentReference[oaicite:3]index=3
- [5] Universidad de Ciencias Aplicadas y Ambientales (UDCA). (2023). *[Título del estudio sobre drones agrícolas]*. Bogotá, Colombia. [Datos verificados pendientes].