Módulo 4. Más allá de ChatGPT. Ampliando funcionalidad

Los contenidos de este módulo son provisionales, se cerrarán en los próximos días

Unidad 4.1 Plugins, extensiones y complementos.

Introducción

Ya sabemos que ChatGPT, Gemini y otros chatbots son herramientas de IA diseñados para simular conversaciones en lenguaje natural. Estos chatbots pueden responder preguntas, proporcionar y gestionar información de todo tipo en base a la información con la que han sido entrenados, lo cual no es poco ya que en general se han entrenado con prácticamente todo el conocimiento presente en internet hasta determinada fecha (ChatGPT con información hasta 2021).

Pero estas herramientas también tienen limitaciones, por un lado en la imposibilidad de acceder a información generada posteriormente a la fecha de corte con la que fue entrenado o acceso a datos generados en tiempo real. Por otro lado, en la incapacidad de realizar tareas más allá de su capacidad de gestionar lenguaje natural.

Los esfuerzos en superar estás restricciones han desembocado en la generación de otros sistemas de facilitación e integración de dichos datos y capacidades.

En particular los **plugins, las extensiones de navegadores y los complementos.** Los plugins facilitan a los chatbots el acceso a información on-line , además de proporcionar funcionalidades adicionales. Las extensiones facilitan el acceso a las posibilidades de la IA a la información que encontramos mediante los navegadores web. Por último, los complementos, son similares a las extensiones pero referidas al contexto de aplicaciones concretas, como procesadores de texto o hojas de cálculo, lo que nos permite igualmente incorporar la potencia de la IA en el proceso e interpretación de datos a nuestros propios documentos. Veamos una explicación de su funcionamiento así como ejemplos concretos en cada caso:

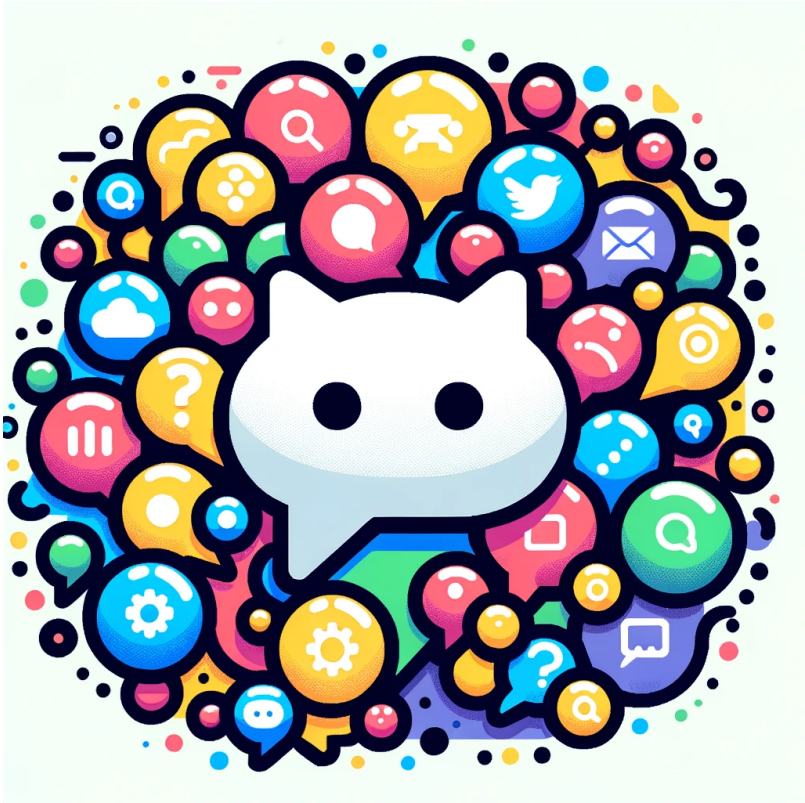


Imagen propia. Generada por Dall-E 3 en ChatGPT

Plugins

Los plugins son programas de software diseñados para ampliar y personalizar la funcionalidad del chatbot. El funcionamiento básico es el de la integración en el chatbot de otras aplicaciones y servicios incluida la posibilidad de navegar por Internet accediendo a información on-line.

Es decir, cuando instalamos un plugin en el chatbot, este **chatbot adquiere una nueva función.**

Para ilustrarlo de manera sencilla veamos un ejemplo. Los chatbots dan resultados convincentes cuando se conversa, se solicita información o se le pide que realice determinadas tareas con un texto (resumir, expandir, traducir, etc..) pero siempre en lenguaje natural. Sin embargo no son capaces muchas veces de realizar operaciones matemáticas sencillas, mucho menos sin son complejas. Esta limitación se ha superado con la posibilidad de instalar plugins específicos de cálculo matemático como **Wolfram.** Este plugin agrega inteligencia adicional al chatbot permitiéndole acceder a cálculos potentes, matemáticas precisas, visualizaciones y datos en tiempo real a través de las tecnologías Wolfram. Haz click en el enlace o en la imagen para acceder a la ***web de Wolfram*** y contemplar sus posibilidades, las cuales, gracias al plugin son accesibles en lenguaje natural desde ChatGPT.

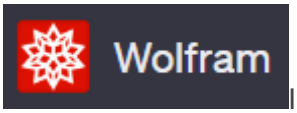


Imagen del plugin de Wolfram en ChatGPT

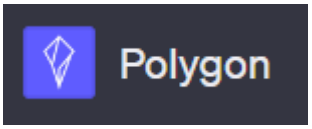
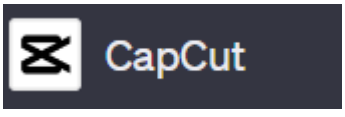
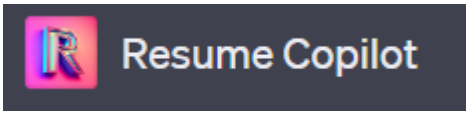
Los chatbots funcionan a través de una serie de principios básicos que se pueden resumir en las siguientes etapas:

- **Competencias.** Los chatbots tienen las habilidades comentadas, comunicarse e interactuar automáticamente con los usuarios para proporcionar información, asistencia o realizar tareas específicas. Si en nuestra interacción con el chatbot, nuestra intención va más allá de sus competencias el chatbot no podrá darnos una respuesta satisfactoria de manera directa.
- **Cognitivo.** Al estar diseñados para interpretar el contexto de la conversación y responder a nuestras preguntas gracias a determinadas herramientas de IA como el Procesamiento del Lenguaje Natural (PNL), el chatbot hará uso de estas habilidades para entender si el texto que introduce el usuario corresponde o no a alguna de sus competencias específicas. Si, el chatbot tiene la capacidad de responder al usuario, lo hará, de no ser así, recurrirá automáticamente al plugin que le permita hacerlo, estableciendo una interfaz de comunicación entre el chatbot y el plugin adecuado.
- **Interfaz de comunicación.** Mediante esta interfaz el chatbot se comunica con un plugin para traducir y trasladar la pregunta del usuario en el lenguaje preciso que necesita el sistema. Este proceso implica la conversión de la entrada del usuario a un formato que el plugin pueda comprender y procesar, para luego recibir la respuesta y traducirla de vuelta a un formato comprensible para el usuario. Esta interfaz por tanto, es de doble sentido, el chatbot traslada la pregunta al plugin y el plugin le devuelve la respuesta al chatbot, que posteriormente trasladará al usuario otra vez en lenguaje natural.

Es decir, el chatbot identifica la intención detrás de la pregunta del usuario y los aspectos relevantes presentes en la consulta, para luego formular la petición adecuada al plugin.

Los plugins han potenciado en gran medida las posibilidades de los chatbots, habilitando la posibilidad de navegar por internet, acceder a servicios de terceros y realizar cálculos matemáticos.

Vemos solo algunos ejemplos de plugins accesibles en ChatGPT

Polygon: Es un plugin que permite a ChatGPT acceder a datos del mercado financiero. Con él, los usuarios pueden obtener información en tiempo real sobre acciones, criptomonedas, noticias y otros detalles financieros y hacer consultas a ChatGPT en relación con esos datos.	
CapCut: Este plugin transforma las solicitudes de texto del usuario en guiones personalizados para videos. Es una herramienta poderosa para aquellos que desean generar contenido visual basado en texto.	
ResumeCopilot: Especializado en la redacción y mejora de currículums. Con él, los usuarios pueden optimizar sus CVs, haciéndolos más atractivos para los empleadores. Su funcionalidad dentro de ChatGPT es muy valiosa para quienes buscan empleo.	

Expedia: Herramienta para planificar viajes de principio a fin. Busqueda de alojamiento, viajes, alquiler de coches o actividades en el destino elegido, todo desde ChatGPT.	Expedia

También hay **plugins educativos,** por ejemplo:

Little professor: Permite crear cuestionarios para el aula ayudando al profesor en su tarea	Little Professor
AskYourPdf: Permite hacer preguntas sobre tus documentos PDF y chatear con ellos. Para investigar y aprender nuevos temas, puedes revisar documentos PDF en un formato conversacional	AskYourPDF
Present it ShowME: Genera diapositivas interesantes en cuestión de segundos introduciendo comandos e indicaciones.	Present It: ShowME
Wolfram: Convirte a ChatGPT en la herramienta perfecta para tareas matemáticas. Usando el plugin Wolfram, puedes analizar algoritmos complejos y mejorar tus habilidades matemáticas.	Wolfram

Estos son solo unos pocos ejemplo de de plugins ya operativos en ChatGPT, pero ya son mas de 500 los disponibles lo que permite a ChatGPT expandir enormemente sus posibilidades.

Si bien, actualmente en ChatGPT solo están disponibles para la versión **ChatGPT plus**, siendo esta de pago.

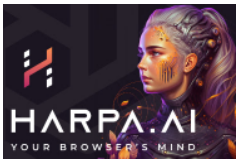
Otra alternativa para ampliar la funcionalidad que nos ofrece la IA son las **extensiones instalables en navegadores y programas.**

Extensiones de IA para navegadores

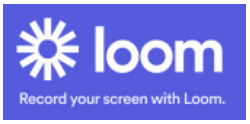
Las extensiones son pequeños programas que personalizan y amplían la experiencia de navegación. Muchas de ellas son aplicaciones web que permiten ser incorporadas al navegador utilizado para poder trabajar con la funcionalidad que proporcionan directamente sobre el contenido de la web en la que estemos navegando. Nos vamos a centrar en aquellas que proporcionan funcionalidades de IA.

Las extensiones de inteligencia artificial para los navegadores de internet ofrecen una variedad de ventajas que pueden mejorar significativamente la experiencia del usuario, la productividad y la accesibilidad.

En el navegador Chrome podemos instalar multitud de extensiones, como por ejemplo:

Voice remaker: Genera voz con IA	
NoteGPT: Transcribe y resume en texto el contenido de videos de youtube.	
Harpa.ai: Muestra las respuestas de ChatGPT en las páginas de los motores de búsqueda. Resume, extrae y monitoriza páginas, precios y datos.	
Read aloud: Lee en voz alta la web o artículo en el que estés navegando con un solo click	
Otter.ai: Transcribe conversaciones en tiempo real utilizando IA. Es ideal para reuniones, entrevistas o cualquier situación donde se requiera una transcripción precisa.	

Al igual que con los plugins, podemos encontrar **extensiones con utilidad para educación**:

Grammarly: Esta extensión utiliza la IA para corregir errores gramaticales y de estilo en textos escritos en línea. Es como tener un editor personal en tu navegador	
Loom: Graba tu pantalla o con la cámara con un solo clik . Comparte el contenido al instante mediante un link	
Google Classroom; herramienta de Google Apps for Education que ayuda a los profesores a crear y organizar tareas rápidamente, proporcionar observaciones de forma eficaz y comunicarse con sus clases con facilidad. A su vez, Classroom ayuda a los alumnos a organizar su trabajo en Google Drive, completarlo y entregarlo, y comunicarse directamente con sus profesores y compañeros.	
Diccionario RAE: como en la página que estés, seleccionar la palabra que no sabemos y pinchar el botón derecho. Ahí te saldrá la extensión y haz clic en ella. Seguidamente, veremos la definición de la RAE (Real Academia Española).	

Las opciones son muchas, piensa en cual es tu necesidad y busca una extensión que la cubra. Para instalarlos, solo tienes que acceder al gestor de extensiones del navegador y buscar la que necesites. En Chrome se accede a través de ***Chrome Web Store*** y buscar por nombre o por categoría.

Complementos de IA

Imagen propia. Generada por Dall-E 3 en ChatGPT

Los complementos permiten **integrar herramientas de IA para utilizarlas directamente una aplicación o programa.**. Por ejemplo en la suite de MS Office (Word, Excel, PowerPoint o Outlook), aunque no son exclusivas de Office para Windows, también están disponibles para otras aplicaciones de escritorio y en diferentes sistemas operativos como macOS.

Los complementos son desarrollados por terceros o por el mismo proveedor del software (Microsoft en el caso de MS Office) que integran capacidades de inteligencia artificial para mejorar la funcionalidad y la experiencia del usuario.

Los complementos se integran directamente en las aplicaciones de oficina apareciendo como botones en la cinta de opciones o incluso como paneles adicionales dentro de la aplicación, posibilitando incrementar la productividad y la calidad del trabajo con funciones como:

- **Análisis de datos y predicciones:** En hojas de cálculo los complementos con IA pueden analizar patrones en los datos y ofrecer predicciones o sugerencias. Pueden ayudar a identificar tendencias, realizar pronósticos y proporcionar información valiosa basada en el análisis de grandes conjuntos de datos.

- **Asistencia en la redacción**. En procesadores de texto, ofrecen correcciones gramaticales, sugerencias de estilo y recomendaciones para mejorar la claridad y la coherencia del texto.

- **Traducción automática**. Herramientas que permiten traducir textos completos o seleccionados a diferentes idiomas directamente dentro del documento.
- **Reconocimiento de imágenes.** Algunos complementos pueden extraer información de imágenes o gráficos y convertirla en texto o datos editables.
- **Automatización de tareas**. En aplicaciones de correo ayudan a automatizar procesos repetitivos, ahorrando tiempo y reduciendo la posibilidad de errores humanos.
- **Accesibilidad**. Ofrecen herramientas para hacer los documentos más accesibles, como lectura en voz alta, descripciones de imágenes y más.
- **Colaboración mejorada**. Facilitan la colaboración en tiempo real, permitiendo a los usuarios trabajar juntos en documentos y hojas de cálculo de manera más efectiva.
- **Personalización y aprendizaje.** Se adaptan al uso y preferencias del usuario, aprendiendo de sus patrones para ofrecer recomendaciones y atajos personalizados

Imagen propia. Generada por Dall-E 3 en ChatGPT.

Los complementos para aplicaciones pueden generalmente encontrarse e instalarse directamente desde el propio programa en la sección de “Complementos” o “Add-ins”. Es importante descargar complementos solo de fuentes confiables para evitar software malicioso.

Al igual que con los plugins y las extensiones, las posibilidades son muy amplias y es interesante hacer un rastreo entre todas las opciones posibles para encontrar aquellos complementos que nos aporten funcionalidades añadidas, incrementando nuestra productividad,

Es importante tener en cuenta, que las tres alternativas están en un momento de expansión en cuanto a oferta y que podemos encontrar múltiples opciones para cada necesidad. Como en cualquier producto o servicio, los habrá mejores y peores, distintas ofertas gratuitas y de pago, lo cual no implica necesariamente unas mayores prestaciones. Será labor de cada uno informarse y probar distintas alternativas hasta dar con las herramientas que realmente nos aporten lo que estamos buscando. Además es importante descargar y utilizar solo herramientas de fuentes confiables para evitar software malicioso.

Unidad 4.2. Ampliando el chat. Hablar con tus datos y generación aumentada

> Con el avance de la tecnología de la realidad virtual, pronto llegaremos a un punto donde no podremos distinguir entre lo que es real y lo que es un juego
>
> **Elon Musk**, CEO de SpaceX y Tesla y cofundador de OpenAI

Introducción

La inteligencia artificial y, en particular, los modelos de procesamiento del lenguaje natural (PLN), han experimentado avances significativos en estos dos últimos años, avances que se traducen en un crecimiento exponencial de aplicaciones en todos los ámbitos y sectores empresariales, gubernamentales y de cualquier entidad u organización.

Modelos de lenguaje, como GPT, BERT, Llama y otros, han demostrado ser extremadamente potentes para comprender y generar texto en lenguaje natural proporcionando mecanismos para facilitar y automatizar la gestión de la información y del conocimiento. Sin embargo, para aprovechar al máximo su potencial, a menudo es necesario personalizarlos y adaptarlos a conjuntos de datos específicos o a dominios particulares. No hay que olvidar que estos modelos se entrenan con datos de fuentes diversas como Wikipedia pero que no están actualizados por lo que en muchas ocasiones deben tener la posibilidad navegar en internet para accder a contenidos más específicos o actuales. No solo eso sino que hay información poco o nada accesible que los modelos desconocen.

En la siguiente tabla podemos apreciar el coste, tanto en tiempo como en dinero del entrenamiento de los modelos de lenguaje más utilizados

Modelo de Lenguaje	Empresa	Año de Creación	Estimación de Tiempo de Entrenamiento	Estimación de Costo de Entrenamiento	Código Abierto
GPT-3	OpenAI	2020	Varios meses	Millones de dólares	No
BERT	Google	2018	Semanas a meses	Cientos de miles a millones de dólares	Sí
T5	Google	2020	Meses	Millones de dólares	Sí
GPT-4	OpenAI	2023	Meses	Decenas de millones de dólares	No
GPT-2	OpenAI	2019	Semanas a meses	Cientos de miles a millones de dólares	Sí
Transformer	Google	2017	Semanas	Decenas a cientos de miles de dólares	Sí
XLNet	Google/CMU	2019	Semanas a meses	Cientos de miles a millones de dólares	Sí
AlphaFold	DeepMind	2020	Meses	Millones de dólares	Sí
MuZero	DeepMind	2020	Meses	Millones de dólares	No
LLaMA	Meta	2023	No disponible	No disponible	Sí

> Es importante tener en cuenta que la disponibilidad de los modelos como código abierto varía significativamente. Algunos modelos, especialmente los más avanzados como GPT-3 y GPT-4 de OpenAI, no son de código abierto, aunque OpenAI ofrece acceso a través de su API. Por otro lado, muchos modelos desarrollados por Google y otros investigadores académicos suelen ser de código abierto para fomentar la investigación y el desarrollo en la comunidad científica. La información sobre el tiempo y el costo de entrenamiento del modelo LLaMA de Meta no está claramente disponible, ya que la compañía no ha divulgado estos detalles.

Dados los costes inasumibles se requieren métodos de actualización de dichos modelos, métodos que no deben pasar por el re-entrenamiento que es absolutamente inasumible por pequeñas empresas o usuarios individuales.

Para ello existen diversas estrategias o como suele decirse 'workarounds' que están implantándose con rapidez en todos los chatbots actuales.

¿Por qué Personalizar?

La personalización de un modelo de lenguaje es crucial cuando trabajamos con datos específicos de un dominio particular o cuando queremos que el modelo realice tareas muy concretas. Los modelos de lenguaje preentrenados son generalistas; han sido entrenados en grandes cantidades de texto de internet, lo que los hace versátiles, pero no necesariamente expertos en áreas específicas. Personalizar estos modelos con nuestros propios datos nos permite ajustarlos para que se alineen mejor con nuestras necesidades particulares, mejorando así su rendimiento y relevancia.

Transfer-Learning: Hablando con tus Datos

Una de las técnicas más comunes para personalizar modelos de lenguaje es el llamado Transfer-Learning o transferencia de conocimiento.

Este proceso implica tomar un modelo preentrenado y continuar su entrenamiento en un conjunto de datos de un dominio específico, evitando del coste de entranar el modelo de nuevo.

El "transfer learning" o aprendizaje por transferencia, es una técnica en el campo de la inteligencia artificial y el aprendizaje automático. Para entenderlo mejor, podemos usar el símil de un chef aprendiendo a cocinar un nuevo tipo de cocina.

Imagina que un chef ya es experto en cocina italiana. Sabe cómo preparar una variedad de platos italianos y entiende los principios básicos de esta cocina. Ahora, si quiere aprender a cocinar comida japonesa, no necesita empezar desde cero. Puede aprovechar muchas de las habilidades y conocimientos que ya posee, como técnicas de corte, manejo de ingredientes frescos y presentación de platos. Este chef solo necesita aprender las diferencias específicas de la cocina japonesa, como trabajar con ingredientes típicos de Japón o técnicas de cocción únicas para esa cocina.

De manera similar, en el aprendizaje por transferencia, un modelo de IA que ha sido entrenado en una tarea (como reconocer objetos en imágenes) puede reutilizar su conocimiento previo para aprender una nueva tarea relacionada más rápidamente y con menos datos. Por ejemplo, si un modelo se ha entrenado para reconocer automóviles en imágenes, y luego se desea entrenar para reconocer motocicletas, no tiene que aprender desde cero. Puede adaptar lo que ya sabe sobre vehículos y características visuales para aprender la nueva tarea con más eficacia.

Esta técnica, que coloquialmente se suele denominar 'habla con tus datos' ha sido una de las principales derviaciones de la IA tectual al permitir a las organizaciones hablar y procesar información propia de manera muchos más intelegiente y específica.

Hasta hace poco este proceso se hacía mediante programación. Hoy en día, herramientas como chatGPT ya permiten la generación de modelos personalizados para campos específicos.

Por ejempo puedo crear un chatBot personalizado y especialista en el campo de la historia medieval y compartirlo con mis alumnos, o centrar mi chatBot en la programación de videojuegos.

En el proceso de creación de estos chats puedo agregar prompts específicos, urls, bases de datos propias e incluso documentos en pdf, vídeos y audios.

Este proceso hace sólo unos meses era muy complejo y requería conocimientos de programación, sin embargo actualmente ya hay herramientas que lo facilitan enormemente, por supuesto también en chatGPT como podemos apreciar en este vídeo dónde el propio SAm Altman (cofundador de chatGPT) desarrolla un sencillo ejemplo de uso de creación de un chatGPT presonalizado:

https://www.youtube.com/embed/q1dcs0biFWU

Vídeo en el que Sam Altman demuestra la nueva funcionalidad de chatGPT para contruir chatBots personalizados

En la siguiente tabla indicamos algunas de las principales herramientas para ello

Herramienta	Tipo de Datos	Descripción	Características Clave
ChatGPT	Texto	Interfaz de chat para interactuar con grandes cantidades de texto, generando respuestas y análisis.	Procesamiento de lenguaje natural, generación de texto.
ChatDoc	Documentos de texto	Herramienta diseñada para analizar y extraer información relevante de documentos de texto.	Extracción de texto, análisis de contenido de documentos.
ChatPDF	Documentos PDF	Especializada en extraer y analizar información de documentos PDF.	Extracción de texto, análisis de contenido de PDF.
PageChat	Páginas web	Permite interactuar con el contenido de páginas web para extraer y analizar información relevante.	Extracción y análisis de contenido web, fácil de usar.
Chatbase	Bases de datos	Herramienta de análisis y consulta de bases de datos mediante una interfaz de chat.	Interfaz de chat para SQL, análisis de datos.
Dante AI	Análisis de texto avanzado	Herramienta para analizar y obtener insights de grandes volúmenes de texto.	Análisis de texto profundo, aprendizaje automático.

Tableau	Datos visuales	Visualización de datos para crear y compartir cuadros de mando y gráficos interactivos.	Visualizaciones interactivas, integración de datos.
Power BI	Datos de negocios	Herramienta de Microsoft para visualizar datos y compartir insights a través de la organización.	Análisis de datos, informes interactivos.
Google Data Studio	Datos web y marketing	Convierte datos en informes y cuadros de mando personalizables e informativos.	Integración con Google Analytics, fácil de usar.
Domo	Datos empresariales	Combina herramientas para la integración, visualización y colaboración en datos.	Visualización de datos, colaboración en tiempo real.

Aumento de Datos o RAG

El aumento de datos es otra estrategia clave para mejorar el rendimiento de los modelos de lenguaje en conjuntos de datos específicos. Consiste en generar variaciones de los datos de entrenamiento para crear un conjunto de datos más amplio y diverso. Esto puede incluir técnicas como la paráfrasis, la traducción a otros idiomas y la vuelta al idioma original, y la manipulación sintáctica.

"Retrieval Augmented Generation" (RAG), que traducido sería "Generación Aumentada por Recuperación", es una técnica en el procesamiento del lenguaje natural que combina la recuperación de información relevante con la generación de texto. Es una metodología avanzada usada en modelos de inteligencia artificial para mejorar la generación de respuestas más informadas y precisas. Aquí te explico con más detalle:

Componentes de RAG

1. **Recuperación de Información**:
   - En la fase de recuperación, el modelo busca en una gran base de datos o repositorio de documentos para encontrar fragmentos de texto que sean relevantes para la pregunta o el prompt dado.
   - Este repositorio puede incluir una amplia gama de documentos, como artículos de Wikipedia, publicaciones de blogs, libros, etc.
2. **Generación de Respuestas**:
   - Utilizando los fragmentos de texto recuperados, el modelo de lenguaje luego genera una respuesta.
   - Esta generación no es una simple repetición de los fragmentos recuperados, sino que el modelo los utiliza como contexto para construir una respuesta coherente y contextualizada.

Funcionamiento de RAG

- **Integración de Recuperación y Generación**:
  - RAG efectivamente integra dos componentes principales de la inteligencia artificial: un sistema de recuperación de documentos (como un motor de búsqueda) y un modelo de generación de texto (como GPT-3).
  - Cuando se formula una pregunta, primero activa su componente de recuperación para encontrar la información relevante. Luego, el modelo de generación utiliza esta información para formular una respuesta informada.
- **Mejora de la Calidad de las Respuestas**:
  - Al basar sus respuestas en información específica y relevante recuperada, RAG puede proporcionar respuestas más precisas, detalladas y contextualizadas.
  - Esto es particularmente útil para preguntas que requieren conocimiento especializado o actualizado.

Aplicaciones de RAG

- **Asistentes Virtuales y Chatbots**: Mejorando la precisión y relevancia de las respuestas en aplicaciones de conversación.
- **Herramientas de Búsqueda y Análisis de Datos**: Proporcionando respuestas más detalladas y contextualizadas a consultas de búsqueda.
- **Educación y Aprendizaje Automático**: Como una herramienta para generar explicaciones educativas o para responder preguntas de estudio.

Ventajas de RAG

- **Respuestas Basadas en Evidencia**: Al usar documentos y datos reales como base para las respuestas, RAG ofrece una forma de generar respuestas que están respaldadas por evidencia concreta.
- **Adaptabilidad**: Puede adaptarse a una amplia gama de temas y preguntas, gracias a su capacidad para buscar y utilizar información de numerosas fuentes.

Desafíos de RAG

- **Dependencia de la Calidad de los Datos**: La efectividad de RAG depende en gran medida de la calidad y actualidad de la base de datos que utiliza para la recuperación de información.
- **Complejidad y Recursos**: Implementar un sistema RAG efectivo puede ser complejo y requerir recursos computacionales significativos.
- La técnica de "Retrieval Augmented Generation" (RAG) se centra principalmente en el procesamiento del lenguaje natural y la generación de texto. Sin embargo, el concepto subyacente de combinar la recuperación de información con la generación o transformación de contenido puede, en teoría, ser aplicado en el campo de la imagen y el video, aunque con diferentes técnicas y tecnologías. En el contexto de imágenes y videos, el proceso sería diferente y se basaría en técnicas de visión por computadora y aprendizaje profundo. Aquí hay un par de aplicaciones hipotéticas en estos campos:

Aplicaciones en Imágenes

1. **Recuperación y Mejora de Imágenes**:
   - Un sistema podría buscar en una base de datos imágenes similares a una dada y utilizar esa información para mejorar o editar la imagen original (por ejemplo, mejorar la resolución, corregir colores, etc.).
   - Por ejemplo, si se tiene una imagen borrosa, el sistema podría buscar imágenes claras y nítidas con características similares y utilizarlas como referencia para mejorar la calidad de la imagen original.
2. **Generación de Contenido Basado en Imágenes Existentes**:
   - Un modelo podría generar nuevas imágenes o modificar las existentes basándose en características y estilos de imágenes recuperadas de una base de datos amplia. Esto sería útil en diseño gráfico, publicidad, y arte digital.

Aplicaciones en Videos

1. **Mejora y Restauración de Videos**:
   - Similar a las imágenes, un sistema podría mejorar la calidad de un video (por ejemplo, resolución, claridad, estabilización) basándose en datos recuperados de videos de alta calidad.
2. **Generación de Secuencias de Video**:
   - Crear nuevas secuencias de video o editar videos existentes basándose en características y estilos de otros videos. Esto podría aplicarse en la producción de películas, publicidad y realidad virtual.

Consideraciones Técnicas

- **Complejidad de Datos**: Los datos de imagen y video son significativamente más complejos que el texto, lo que requiere modelos más sofisticados y más recursos computacionales.
- **Técnicas de Visión por Computadora**: La implementación de una técnica similar a RAG en imágenes y videos requeriría el uso de avanzadas técnicas de visión por computadora y redes neuronales convolucionales.
- **Desafíos en la Recuperación**: La recuperación de información relevante y útil a partir de imágenes y videos es un desafío significativo debido a la variabilidad y riqueza de los datos visuales.

Aunque el concepto de RAG como tal es específico del procesamiento del lenguaje, sus principios fundamentales de combinar la recuperación con la generación o transformación pueden inspirar enfoques similares en otros campos como el de las imágenes y los videos. Sin embargo, estas aplicaciones requerirían un desarrollo tecnológico considerable y enfrentarían desafíos únicos inherentes a estos medios.

RAG representa un paso adelante significativo en la creación de sistemas de IA más sofisticados y útiles, capaces de manejar preguntas complejas y proporcionar respuestas bien informadas y precisas.

https://www.youtube.com/embed/T-D1OfcDW1M

Vídeo introductorio del concepto de RAG

Consideraciones Éticas y de Sesgo

Al personalizar modelos de lenguaje, es importante tener en cuenta las consideraciones éticas y el potencial de sesgo en los datos. Los modelos aprenden de los datos en los que son entrenados, y si esos datos contienen sesgos, el modelo los replicará. Es crucial ser consciente de esto y tomar medidas para mitigar los sesgos tanto como sea posible.

Vectores de datos (embeddings)

Aunque ya hemos comentado este tipo de objetos en el módulo 2 sobre fundamentos, lo retomamos d enuevo ya que además de ser esencales en el entrenamiento de modelos también se usan para tareas típicas de NLP.

Los almacenes de datos que utilizan datos vectorizados están diseñados para mejorar el rendimiento de las consultas y operaciones analíticas en grandes conjuntos de datos. La vectorización es un método de procesamiento de datos en el que se operan vectores enteros de datos, en lugar de procesar un único elemento de datos a la vez. Esto se alinea con las capacidades de las CPU modernas que pueden realizar operaciones en vectores de datos simultáneamente, resultando en un rendimiento significativamente mejorado. A continuación, se describen algunos de los usos y beneficios de los almacenes de datos con datos vectorizados:

¿Qué son los Word Embeddings?

Los word embeddings son, en esencia, una forma de convertir palabras en vectores numéricos. Imagina que cada palabra es una persona y cada persona tiene una lista de características que la describen. En el caso de los word embeddings, estas características son números. Este proceso permite que las computadoras trabajen con palabras y textos, realizando operaciones matemáticas sobre ellos.

¿Cómo Funcionan?

Para entender cómo funcionan, podemos usar un símil: Imagina un mapa de una ciudad donde cada punto en el mapa representa una tienda. Las tiendas que venden productos similares están más cerca unas de otras. De manera similar, en el espacio de word embeddings, palabras con significados similares están "más cerca" unas de otras en términos numéricos. Por ejemplo, "gato" y "perro" estarían más cerca que "gato" y "avión".

Aplicaciones

1. **Búsqueda y Recomendación de Textos**: Ayudan a encontrar textos similares o relacionados.
2. **Análisis de Sentimientos**: Identifican la emoción o el sentimiento detrás de un texto.
3. **Traducción Automática**: Facilitan la traducción de un idioma a otro.
4. **Asistentes Virtuales y Chatbots**: Mejoran la comprensión del lenguaje humano.

Ventajas

- **Mejor Comprensión del Lenguaje**: Permiten a las máquinas entender mejor las sutilezas del lenguaje humano.
- **Versatilidad**: Son útiles en una amplia gama de aplicaciones de NLP.
- **Eficiencia**: Mejoran la eficiencia en el procesamiento de grandes volúmenes de texto.

Desafíos

- **Contexto Limitado**: Pueden no capturar completamente el contexto en el que se usa una palabra.
- **Sesgo en los Datos**: Pueden heredar y amplificar sesgos presentes en los datos con los que fueron entrenados.

Los word embeddings son una herramienta poderosa en el campo del NLP, proporcionando una manera para que las computadoras "entiendan" y trabajen con el lenguaje humano. Al convertir palabras en vectores numéricos, abren un mundo de posibilidades para el procesamiento y análisis de texto, aunque no están exentos de desafíos y limitaciones. Su uso continuará siendo fundamental en el desarrollo de tecnologías relacionadas con el lenguaje.

Conclusión final

En muchos casos, estas técnicas se utilizan juntas en aplicaciones de NLP. Por ejemplo, un modelo de lenguaje podría ser afinado para una tarea específica, y luego las representaciones vectoriales generadas por este modelo podrían ser almacenadas y consultadas utilizando un almacén de vectores de datos como Pinecone. Esto permite tanto la personalización del modelo (a través del fine-tuning) como la búsqueda eficiente y la similitud semántica (a través del almacén de vectores de datos).

El afinamiento (fine-tuning) y el uso de almacenes de vectores de datos son técnicas complementarias más que excluyentes, y cada una tiene su lugar en el procesamiento del lenguaje natural (NLP).

La personalización de modelos de lenguaje para adaptarlos a nuestros propios datos es un paso crucial para aprovechar al máximo el potencial de la inteligencia artificial en el procesamiento del lenguaje natural. Mediante técnicas como el "fine-tuning", la transferencia de conocimientos, el aumento de datos y la inyección de conocimiento, podemos ajustar los modelos para que se alineen mejor con nuestras necesidades específicas, mejorando así su rendimiento y relevancia en tareas concretas. Sin embargo, es importante abordar este proceso con un enfoque reflexivo y crítico, teniendo en cuenta las consideraciones éticas y los potenciales sesgos en los datos. Con un

enfoque cuidadoso y metódico, podemos personalizar los modelos de lenguaje para desbloquear nuevas posibilidades y obtener insights valiosos de nuestros datos.

Unidad 4.3. Nuevos paradigmas de chatbot.

> "Los agentes de inteligencia artificial no son solo programas en una computadora; tienen el potencial de ser compañeros inteligentes y colaboradores en nuestra búsqueda diaria de soluciones a problemas complejos."
>
> **Ben Goertzel,** destacado científico en el campo de la inteligencia artificial (IA). Es conocido principalmente por su trabajo en áreas como la inteligencia artificial general (AGI), que se centra en la creación de máquinas con la capacidad de aprender y aplicar inteligencia de manera amplia y flexible, similar a cómo lo hacen los seres humanos.

Introducción

A lo largo de nuestro viaje explorando el vasto universo de la Inteligencia Artificial, hemos profundizado en la comprensión y aplicación de la IA generativa, prestando especial atención a ChatGPT. Esta herramienta ha demostrado ser un valioso recurso para facilitar el proceso de enseñnza-aprendizaje, ofrecer apoyo personalizado a los estudiantes, ayudar a los docentes a generar material, entre otras cosas. Sin embargo, el campo de la IA está en constante evolución, y hoy nos embarcamos en un nuevo capítulo para descubrir herramientas innovadoras que prometen llevar la IA y a la IA aplicada a la educación a una dimensión distinta.

ChatGPT ha establecido un precedente importante en el campo de la IA conversacional, permitiendo interacciones fluidas y generación de texto coherente. No obstante, las necesidades cambiantes y los avances tecnológicos han dado lugar a la creación de herramientas especializadas que buscan mejorar y expandir las capacidades de ChatGPT. Aquí es donde AgentGPT, y AutoGPT entran en escena, cada uno con características únicas y aplicaciones específicas.

AutoGPT, AgentGPT son herramientas avanzadas basadas en modelos de lenguaje de gran tamaño. Se pueden agrupar dentro del mismo tipo de herramienta, específicamente, **agentes de IA autónomos** que **buscan reducir la cantidad de interacción humana necesaria para completar tareas específicas**, permitiendo que los sistemas de IA trabajen de manera más autónoma hacia un objetivo con mínima o ninguna entrada humana.

Agentes de Inteligencia Artificial (AI)

Los agentes de inteligencia artificial representan un aspecto crucial y cada vez más prominente en el campo de la IA. Son sistemas o programas de software diseñados para realizar tareas específicas de manera autónoma, imitando algunas capacidades humanas como la percepción, el razonamiento, el aprendizaje y la toma de decisiones.

¿Por Qué son Importantes?

1. **Automatización y Eficiencia**: Los agentes de IA pueden manejar tareas repetitivas o complejas, aumentando la eficiencia y permitiendo que los humanos se concentren en actividades más estratégicas o creativas.
2. **Personalización**: Pueden adaptarse a las necesidades y preferencias individuales de los usuarios, ofreciendo servicios y experiencias personalizadas.
3. **Capacidad de Aprendizaje**: Muchos agentes de IA están diseñados para aprender de la experiencia, mejorando su rendimiento y toma de decisiones con el tiempo.
4. **Interacción Natural**: Con el avance de la comprensión del lenguaje natural, estos agentes pueden interactuar con los usuarios de manera más fluida y humana.

Tipos de Agentes de IA

1. **Agentes Reactivos Simples**: Responden directamente a su entorno sin mantener un estado interno. Ejemplo: Un termostato inteligente.
2. **Agentes Basados en Modelos**: Tienen una representación interna del mundo que les rodea y pueden actuar en función de este modelo. Ejemplo: Sistemas de navegación autónoma.
3. **Agentes Basados en Objetivos**: Toman decisiones basándose en metas u objetivos establecidos. Ejemplo: Asistentes virtuales que programan reuniones.
4. **Agentes Basados en el Aprendizaje**: Capaces de aprender de sus interacciones y mejorar con el tiempo. Ejemplo: Sistemas de recomendación personalizada.

Ejemplos y Herramientas

- **Chatbots y Asistentes Virtuales**: Como Siri, Alexa y Google Assistant, que pueden responder preguntas y realizar tareas.
- **Sistemas de Recomendación**: Utilizados en plataformas de streaming como Netflix o Spotify para sugerir contenido.
- **Robots Autónomos**: Utilizados en manufactura, logística y exploración.

Desafíos y Consideraciones Futuras

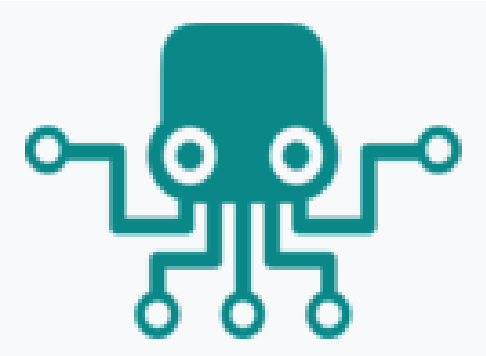
- **Ética y Transparencia**: La toma de decisiones por parte de agentes de IA plantea cuestiones éticas, especialmente en contextos críticos como la salud y la justicia.
- **Interacción Humano-AI**: El diseño de interfaces y sistemas que permitan una colaboración efectiva entre humanos y agentes de IA.

- **Sesgo y Justicia**: Asegurar que los agentes de IA no perpetúen sesgos existentes y operen de manera justa.

Los agentes de inteligencia artificial son una parte integral y en expansión del campo de la IA. A medida que estas tecnologías continúan desarrollándose, su impacto en diversos sectores y aspectos de la vida cotidiana probablemente se ampliará, ofreciendo oportunidades emocionantes y también planteando desafíos significativos.

A continuación comentamos tres de ellos

AutoGPT



AutoGPT es un "agente de IA" que, dada una meta en lenguaje natural, intenta alcanzarla descomponiéndola en subtareas y utilizando internet y otras herramientas en un bucle automático. Es una aplicación de código abierto que interactúa con GPT-4 y GPT-3.5, lo que le permite **automatizar el proceso** de solicitud multi-paso que normalmente se requiere para operar un chatbot como ChatGPT.
AutoGPT también aprende de su propio contenido generado para mejorar sus capacidades lingüísticas, descomponiendo los objetivos en tareas pequeñas para alcanzar el objetivo final.

AutoGPT tiene implicaciones de gran alcance. El diferenciador de otras herramientas de IA convencionales es el circuito de retroalimentación de AutoGPT que le permite planificar, aprender y mejorar.

Con ese enfoque, AutoGPT puede agilizar muchos procesos que requieren dedicación humana. Esto incluye escribir y depurar código, crear contenido, analizar datos y desarrollar planes de negocios. Las personas pueden utilizar el agente autónomo de IA para ayudar con la elaboración de material, la planificación y otras tareas.

Esa es la visión optimista. La otra cara de la historia es que AutoGPT se encuentra hoy en las primeras etapas de su desarrollo. Los resultados y las acciones tomadas por el agente de IA podrían ser potencialmente inexactos o contraproducentes.

AutoGPT también procesa información literalmente, lo que puede resultar problemático en escenarios de toma de decisiones.

En la práctica básicamente, hay que darle una instrucción concreta de un objetivo a AutoGPT y éste planeará y ejecutará los pasos necesarios para finalizar la tarea.

Simplemente se le da un nombre a la tarea, se le asigna una función (por ejemplo, asumir el papel de propietario de una nueva empresa emergente) y asígnele un máximo de cinco objetivos. Por ejemplo, podría utilizar AutoGPT para:

Desarrollar mensajes de chat para atraer a los clientes. Esto puede ayudar a aumentar las ventas, la satisfacción del cliente o las conversiones.

Agiliza y automatiza las tareas del día a día. AutoGPT puede administrar las respuestas de correo electrónico, las respuestas de atención al cliente o el contenido de las redes sociales por usted.

Integre AutoGPT con otras plataformas o herramientas tecnológicas para crear nuevas aplicaciones de procesamiento de lenguaje natural, como la creación de contenido.

Algunos lo comparan con un pasante que puede ayudar a su empresa con tareas simples en las que usted proporciona un objetivo final o una lista de objetivos, y AutoGPT hace el resto.

AgentGPT

AgentGPT es una plataforma de IA autónoma que permite a los usuarios crear y desplegar agentes de IA directamente en el navegador. Genera listas de tareas y luego las ejecuta iterativamente para completar las tareas del usuario.
Es una tecnología basada en NLP que genera texto con una apariencia humana con precisión y fluidez, pudiendo participar en conversaciones, generación de contenido y asistencia para resolver problema.
Se diferencia de AutoGPT en que **no tiene acceso a internet** para buscar información o ejecutar código, pero sigue un proceso iterativo para descomponer y resolver problemas basados en las solicitudes del usuario.

Estas herramientas representan una evolución en el mundo de la IA, buscando trabajar de manera autónoma para alcanzar objetivos definidos por el usuario con mínima intervención humana.

En el siguiente video vemos un ejemplo de uso de AgenGPT

https://www.youtube.com/embed/K6EbB1oSzXI

En este otro vídeo del famoso Dot CSV hay una buena explicación del uso de agentes con chatGPT

https://www.youtube.com/embed/hLYw06LYWIU

Y en este último enlace Mas información y utilización de AgentGPT: https://agentgpt.reworkd.ai/es

Modelos Offline

Los modelos de inteligencia artificial (IA) offline se refieren a sistemas de IA que operan sin necesidad de estar conectados a internet. Estos modelos procesan y analizan datos localmente, en el dispositivo del usuario, en lugar de depender de servidores remotos. Aquí tienes un resumen de sus características y aplicaciones clave:

Características

1. **Procesamiento Local**: Realizan todas las operaciones de procesamiento de datos directamente en el dispositivo del usuario, como un smartphone, una computadora o un dispositivo IoT.
2. **Privacidad Mejorada**: Al no transmitir datos a través de internet, reducen significativamente los riesgos de privacidad y seguridad de los datos.
3. **Funcionamiento Sin Conexión**: Pueden operar en áreas sin acceso a internet o en situaciones donde la conectividad es intermitente o no confiable.
4. **Respuesta Rápida**: Al procesar datos localmente, pueden ofrecer respuestas más rápidas sin la latencia asociada con la transmisión de datos a un servidor remoto y de vuelta.
5. **Menor Consumo de Ancho de Banda**: Al no necesitar enviar datos constantemente a un servidor, reducen el uso del ancho de banda de internet.

Aplicaciones

1. **Dispositivos Móviles**: Aplicaciones de reconocimiento de voz, como asistentes virtuales, que funcionan directamente en teléfonos móviles sin necesidad de una conexión a internet.
2. **Automóviles Autónomos**: Sistemas de conducción autónoma que procesan información de sensores y cámaras en tiempo real para tomar decisiones de conducción.
3. **Robótica**: Robots que operan en entornos remotos o aislados, como robots de exploración en áreas sin cobertura de red.
4. **Salud y Fitness**: Dispositivos de seguimiento de salud y fitness que procesan datos de actividad y salud directamente en el dispositivo.
5. **Seguridad y Vigilancia**: Sistemas de cámaras de seguridad que pueden analizar imágenes y detectar movimientos o actividades sospechosas sin necesidad de enviar datos a un servidor.

Desafíos y Limitaciones

- **Capacidad de Procesamiento**: Los dispositivos deben tener suficiente capacidad de procesamiento para manejar modelos de IA complejos.

- **Actualizaciones de Modelos**: La actualización de modelos offline puede ser más desafiante, ya que requiere la intervención del usuario o mecanismos de actualización automatizados.
- **Complejidad de Implementación**: Desarrollar y optimizar modelos de IA para funcionar eficientemente en un entorno offline puede ser técnicamente desafiante.

Los modelos de IA offline ofrecen ventajas significativas en términos de privacidad, seguridad y accesibilidad. Son especialmente útiles en aplicaciones donde la conectividad es limitada o donde la rapidez y privacidad de los datos son críticas. A medida que la tecnología avanza, es probable que veamos una mayor adopción y evolución de estos modelos en diversos campos.

A continuación indicamos algunas webs y plataformas para experimentar con modelos sin depender de internet

Plataforma	Sitio Web	Descripción	Características Clave
GPT-4All	gpt-4all.com	Permite ejecutar versiones de GPT de manera offline.	- Fácil de usar. - Orientado a usuarios sin experiencia técnica.
OLLAMA	ollama.com	Plataforma especializada en LLM para uso offline.	- Enfoque en privacidad y seguridad de datos. - Personalizable para diferentes aplicaciones.
LangChain	langchain.com	Herramientas y librerías para LLM, con soporte para offline.	- Flexible y modular. - Permite la integración con diferentes LLM.
LLM Studio	llm-studio.com	Plataforma para desarrollar y desplegar LLM de manera offline.	- Interfaz de usuario amigable. - Soporta múltiples modelos de LLM. - Ofrece herramientas para entrenamiento y personalización de modelos.

Estas herramientas y plataformas ofrecen una variedad de opciones para aquellos interesados en trabajar con modelos de lenguaje grandes de manera offline, proporcionando flexibilidad y accesibilidad en diferentes niveles de experiencia y necesidades.

Unidad 4.4. Programación para incautos.

> "La inteligencia general artificial no es simplemente una nueva herramienta que está siendo añadida al conjunto humano de herramientas tecnológicas; es la herramienta que va a rehacer y rediseñar todas las demás herramientas."
>
> *Esta cita de Ben Goertzel refleja su visión sobre el potencial transformador de la inteligencia general artificial (AGI), sugiriendo que su desarrollo no solo aportará una nueva tecnología, sino que también cambiará fundamentalmente la forma en que interactuamos y mejoramos todas las tecnologías existentes.*

Introducción

Aunque no es un curso orientado a programadores hemos considerado interesante añadir esta sección con la única intenciaón de exponer las posibilidades adicionales que ofrecce la IA pra todo aquel que tenga interés o conozca los conceptos básicos de la programación. Probablemente sea el caso de profesores o maestros de ciencias, tecnología etc... Pero también para el resto pues forma parte de la cultura de la IA y como mínimo da una perspecita adicional de la misma.

No hablamos de programar con ayuda de la IA, algo que hacen casi perfectamente los distintos chatbots que hemos visto, sino de usar código para programar y entrenar nuestros propios modelos u otros modelos puestos a disposición del público (llamados modelos Open Source o de código libre).

Las posibilidades son inmensas ya que hay cada vez más modelos y entornos disponibles de manera gratuita así como recursos didácticos y documentación.

La sección se divide en tres partes, una para quién quiera introducirse en programación con herramientas y recursos disponibles, la segunda habla sobre los distintos entornos para programar con IA y modelos de lenguaje y una última en la que mostramos herramientas de creación de aplicaciones sin usar código.

Obviamente las dos primeras son sólamente para aquellos intresados en introducirse o conocer herramientas de programación por lo que las consideramos 'voluntarias' dentro del curso.

Introducción a python

Aunque otros muchos lenguajes, en un gran porcentaje todo lo que se hace y se está haciendo en la actualidad vinculado a la IA y a la llamada ciencia de datos se desarrolla en python. Por ello centraremos esta sección en este lenguaje cada vez más popular.

Para iniciarse en Python, hay una amplia gama de recursos y tutoriales disponibles en línea que pueden ayudar a aprender este lenguaje de programación de manera efectiva. Aquí hay algunas recomendaciones:

Python.org

El sitio web oficial de Python ofrece una sección para principiantes donde puedes encontrar una lista de editores de texto e IDEs recomendados para trabajar con Python, así como libros introductorios y ejemplos de código 27†(Python.org) .

W3Schools

W3Schools proporciona un tutorial interactivo donde puedes aprender Python a través de ejemplos. Este sitio permite editar el código y ver los resultados en tiempo real, lo cual es una forma práctica

de aprender.

DigitalOcean

Ofrece una serie de tutoriales para principiantes en Python. Estos tutoriales exploran el mundo de Python, lo que puede ser una forma útil de obtener una comprensión práctica del lenguaje

FreeCodeCamp

En FreeCodeCamp, hay una lista de 15 cursos gratuitos de Python para principiantes. Entre estos, se incluye un curso completo para principiantes, así como otros recursos como el "Python Handbook" por Flavio Copes.

Microsoft Learn

Microsoft también ofrece un tutorial en español para principiantes en Python, donde podrás descubrir los conceptos básicos de Python, incluyendo el uso de Jupyter Notebook, creación de programas y proyectos, y trabajar con diferentes tipos de datos y estructuras de control en Python.

Estos recursos cubren una variedad de aspectos de Python, incluyendo la sintaxis básica, estructuras de datos, y aplicaciones prácticas del lenguaje. Se proporcionan tanto explicaciones textuales como ejemplos de código interactivos para ayudar a solidificar tu comprensión del material. También es recomendable explorar diferentes plataformas y seleccionar la que mejor se adapte a tu estilo de aprendizaje y necesidades.

Entornos de desarrollo y plataformas

Para la programación y manipulación de modelos de datos, existen varios entornos y herramientas que pueden ser adecuadas dependiendo de tus necesidades y preferencias. A continuación se presentan algunas opciones populares:

Jupyter Notebook

Este es un entorno interactivo que permite la ejecución de código, visualización de datos y documentación todo en uno. Es ampliamente utilizado por científicos de datos y analistas.

RStudio

Es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, que es muy utilizado para la estadística y la visualización de datos.

PyCharm

Este es un IDE para Python, que es un lenguaje de programación muy popular en el ámbito de la ciencia de datos. PyCharm proporciona muchas herramientas y características que facilitan la programación de modelos de datos.

Visual Studio Code (VS Code)

Este es un editor de código fuente gratuito y de código abierto que es extensible y personalizable. Con las extensiones adecuadas, puede ser una herramienta poderosa para trabajar con datos.

SQL Server Management Studio (SSMS)

Si estás trabajando con bases de datos SQL Server, SSMS es un entorno robusto para la gestión y programación de modelos de datos.

DataGrip

Es un IDE de JetBrains para bases de datos y SQL. Permite la conexión a múltiples bases de datos, exploración de datos, y muchas otras funciones útiles para trabajar con modelos de datos.

Google Colab

Es un entorno de notebook en la nube que permite la ejecución de código en servidores de Google, lo cual puede ser útil para la manipulación y análisis de grandes conjuntos de datos.

Databricks

Plataforma basada en la nube que integra un entorno de notebook con capacidades de ejecución distribuida, lo cual es útil para trabajar con grandes conjuntos de datos y modelos complejos.

Tableau

Si bien no es un entorno de programación per se, Tableau es una herramienta poderosa para la visualización y exploración de datos.

MATLAB

Es un entorno para la programación matemática y la manipulación de datos, especialmente útil en el ámbito académico y de investigación.

La elección entre estos entornos dependerá de tus necesidades específicas, la complejidad de tus modelos de datos, y tu familiaridad con los lenguajes de programación y las herramientas mencionadas. También puede ser útil considerar la comunidad y el soporte disponible para cada entorno, así como su integración con otras herramientas y plataformas que puedas estar utilizando.

Creación aplicaciones de IA

A continuación presentamos las herramientas y entornos más importantes relacionadas con la creación y desarrollo de aplicaciones basadas en IA y en el uso de modelos.

HuggingFace

Hugging Face es una empresa conocida por su trabajo en el campo del procesamiento del lenguaje natural (PLN) y el aprendizaje profundo. A continuación, se presentan algunas áreas clave en las que Hugging Face es relevante en el contexto de la programación y gestión de modelos de datos:

Biblioteca Transformers

Hugging Face es famoso por su biblioteca Transformers, que proporciona implementaciones de muchos modelos de lenguaje populares como BERT, GPT-2, T5, y otros. Esta biblioteca facilita el entrenamiento, la evaluación y el uso de estos modelos para diversas tareas de PLN.

Model Hub

Hugging Face también ofrece una plataforma conocida como Model Hub, donde los investigadores y desarrolladores pueden compartir y acceder a modelos preentrenados. Esto facilita la reutilización de modelos y acelera el desarrollo de aplicaciones de PLN.

Datasets Library

Además, Hugging Face proporciona una biblioteca de conjuntos de datos que facilita el acceso a una amplia variedad de conjuntos de datos para entrenamiento y evaluación de modelos.

Tokenizers Library

La biblioteca de tokenizadores de Hugging Face proporciona herramientas para convertir texto en tokens, un paso esencial en el procesamiento del lenguaje natural.

Servicios en la Nube

Hugging Face también ofrece servicios en la nube para entrenar y alojar modelos de lenguaje, proporcionando una plataforma para gestionar el ciclo de vida de los modelos de PLN.

Colaboraciones y Comunidad

Hugging Face colabora con muchas otras organizaciones y comunidades en el campo del aprendizaje profundo y PLN, contribuyendo a la innovación y el avance en estas áreas.

En resumen, Hugging Face proporciona herramientas y plataformas que facilitan la gestión y programación de modelos de datos, especialmente en el ámbito del procesamiento del lenguaje

natural. Su biblioteca Transformers, junto con el Model Hub y otras herramientas, proporcionan un ecosistema robusto para trabajar con modelos de lenguaje y datos relacionados con el texto.

Langchain

LangChain es una plataforma diseñada para facilitar la interacción con modelos de lenguaje grandes (Large Language Models o LLMs) y la integración de estos modelos en aplicaciones y pipelines de datos. A continuación se presentan algunas características clave y capacidades de LangChain:

Integración con LLMs

LangChain (https://www.langchain.com/) proporciona una interfaz estándar que facilita la interacción con una variedad de LLMs de diferentes proveedores como *OpenAI, Cohere, Bloom, Huggingface*, entre otros.

LangChain ofrece una manera estructurada y modular de trabajar con LLMs y aprovechar sus capacidades en una variedad de aplicaciones y escenarios. Esto lo convierte en una herramienta valiosa para los ingenieros de datos y desarrolladores que buscan integrar modelos de lenguaje en sus proyectos.

Documentación

Cómo siempre lo mejor es empezar por la documentación, perfectamente descrita en este enlace

https://docs.langchain.com/docs/

Construcción de Prompts y Gestión de Conversaciones:

Ofrece herramientas para simplificar la creación y gestión de prompts, así como módulos de memoria para gestionar y alterar conversaciones pasadas, lo que es crucial para chatbots 13†(DEV Community) .

Chaining o Encadenamiento de Modelos

Una característica única de LangChain es su capacidad para crear Chains (cadenas) que son enlaces lógicos entre uno o más LLMs, permitiendo crear aplicaciones más complejas al encadenar diferentes componentes 6†(EcoAGI) 8†(Pinecone) .

Agentes Inteligentes e Indexación

Equipa a los agentes con un conjunto de herramientas integral y proporciona métodos para organizar documentos de manera que faciliten la interacción efectiva con los LLMs 13†(DEV Community) .

Aplicaciones Variadas

LangChain se puede utilizar para una amplia gama de aplicaciones como chatbots, sistemas de preguntas y respuestas generativas, resumen de texto y mucho más, proporcionando un marco para incluir IA de LLMs en pipelines de datos y aplicaciones.

Procesamiento de Datos

Descompone grandes cantidades de datos en trozos o *chunks* más manejables, los cuales pueden ser fácilmente incrustados en un vector store. Al recibir un *prompt*, LangChain consulta el Vector Store para obtener información relevante y luego alimenta esta información al LLM para generar o completar la respuesta.

Generación interfaces

streamlit: Creación de aplicaciones basado en *python*, se suele usar en entornos de *LangChain* para generar aplicaciones.

https://streamlit.io/

Cohere

Cohere es una plataforma de inteligencia artificial que se especializa en el procesamiento del lenguaje natural (NLP, por sus siglas en inglés). Esta plataforma ofrece herramientas de IA avanzadas que permiten a los usuarios y desarrolladores comprender, generar y manipular el lenguaje humano de manera eficiente y efectiva. Algunos aspectos clave de Cohere son:

Modelos de Lenguaje Avanzados

Cohere utiliza modelos de lenguaje de última generación, similares a los de OpenAI, para ofrecer capacidades de comprensión y generación de texto.

Aplicaciones Versátiles

Los servicios de Cohere se pueden aplicar en una variedad de casos de uso, como la automatización de respuestas a clientes, la generación de contenido, la traducción automática, la síntesis de información y la moderación de contenido.

Programación de modelos de chat

Lo que significa que se puede utilizar para desarrollar y mejorar aplicaciones de chatbot o sistemas de conversación automatizados. Esta capacidad es particularmente valiosa en áreas como el servicio al cliente, donde los chatbots pueden manejar consultas de manera eficiente y efectiva, proporcionando respuestas en tiempo real y facilitando interacciones fluidas basadas en el lenguaje natural.

APIs Accesibles

La plataforma proporciona APIs (interfaces de programación de aplicaciones) que facilitan la integración de sus capacidades de IA en diversas aplicaciones y sistemas existentes.

Enfoque en la Usabilidad

Cohere está diseñada para ser accesible tanto para desarrolladores experimentados en IA como para aquellos que tienen menos experiencia técnica, con el objetivo de democratizar el acceso a la tecnología de procesamiento de lenguaje natural.

Compromiso con la Ética y la Seguridad

La plataforma también pone énfasis en los aspectos éticos y de seguridad del uso de IA, buscando garantizar que sus herramientas se utilicen de manera responsable.

En resumen, Cohere se posiciona como una plataforma potente y flexible en el campo del procesamiento del lenguaje natural, ofreciendo a desarrolladores y empresas herramientas avanzadas para interactuar y trabajar con el lenguaje humano a través de la IA.

https://cohere.com/

Aplicaciones sin código

Construir aplicaciones con inteligencia artificial (IA) ha sido históricamente una tarea compleja que requiere un conocimiento profundo de algoritmos, estadísticas, y programación. Sin embargo, la aparición de plataformas y herramientas "low-code" o "no-code" ha simplificado significativamente este proceso, permitiendo incluso a los usuarios sin experiencia técnica integrar capacidades de IA en sus aplicaciones. Aquí hay algunas herramientas y plataformas que facilitan la construcción de apps con IA:

Microsoft Power Apps

Con la integración de Microsoft AI Builder, los usuarios pueden añadir inteligencia artificial a sus aplicaciones sin escribir código. Esto incluye capacidades como procesamiento de formularios, predicción, clasificación de objetos en imágenes, y más.

Google AppSheet

AppSheet permite a los usuarios crear aplicaciones móviles y web con funciones de IA como reconocimiento de imágenes, procesamiento del lenguaje natural y modelado predictivo, todo sin necesidad de programar.

OutSystems

Ofrece una plataforma de desarrollo low-code con capacidades de IA, permitiendo a los usuarios integrar servicios de IA y machine learning en sus aplicaciones.

Mendix

Esta plataforma low-code proporciona herramientas para construir aplicaciones inteligentes, integrando servicios de IA y machine learning para mejorar la experiencia del usuario y la eficiencia operativa.

Adalo

Aunque Adalo se centra principalmente en la creación de aplicaciones móviles sin código, los usuarios pueden integrar funcionalidades de IA a través de APIs externas para añadir capacidades avanzadas a sus aplicaciones.

Bubble

Bubble permite a los usuarios construir aplicaciones web sin código y puede integrarse con herramientas de IA mediante el uso de APIs, proporcionando funcionalidades como chatbots, análisis de texto, y más.

Thunkable

Thunkable ofrece la capacidad de crear aplicaciones móviles sin código y permite integrar funciones de IA como reconocimiento de texto, imagen y voz.

Glide

Glide transforma hojas de cálculo de Google en aplicaciones y ofrece integraciones con servicios de IA para añadir funcionalidades como reconocimiento de imágenes y procesamiento de lenguaje natural.

Voiceflow:

Permite a los usuarios diseñar, prototipar y construir aplicaciones de voz interactivas (como Alexa Skills o acciones de Google) sin código, integrando capacidades de procesamiento del lenguaje natural.

Clarifai:

Aunque Clarifai proporciona una API de IA para desarrolladores, también ofrece una interfaz visual que permite a los usuarios crear y entrenar modelos de visión por computadora sin escribir código.

Estas herramientas y plataformas están democratizando el acceso a la inteligencia artificial, permitiendo a más personas aprovechar el poder de la IA para crear aplicaciones avanzadas y personalizadas.

Referencias Módulo 4

Referencias Módulo 4

Web sobre como usar modelos de IA offline

https://www.xataka.com/basics/como-tener-ia-como-chatgpt-local-tu-ordenador-gpt4all

Web AutoGPT

https://www.fool.com/terms/a/autogpt/

https://neilpatel.com/blog/autogpt/#:~:text=Auto,to%20reach%20its%20final%20aim

Vídeo Agentes GPT

https://www.youtube.com/watch?v=K6EbB1oSzXI

Web Agentes GPT

https://agentgpt.reworkd.ai/es

Web LMStudio

https://lmstudio.ai/?s=35

Web para descargar modelos y trabajar fuera de línea

Web Langchain

https://www.langchain.com/

Web Cohere

https://cohere.com/

Web HuggingFace

https://huggingface.co/