INSTITUTO POLITÉCNICO NACIONAL

UNIDAD PROFESIONAL INTERDISCIPLINARIA DE INGENIERÍA CIENCIAS SOCIALES Y ADMINISTRATIVAS

Materia: Aplicación de Redes

Secuencia: 4NV81

Integrantes:

- Castro Hernandez Alan Daniel
- Dominguez Navarrete Eduardo
- Pacheco García Víctor Alejandro
- Herrera Cano Edwin

Fecha de entrega: 20/03/2024

ASISTENTE DE DOCUMENTACIÓN.

Alcancé

Se espera que este proyecto proporcione una solución innovadora y eficiente para facilitar el acceso y comprensión de información contenida en documentos PDF mediante el uso de técnicas avanzadas de inteligencia artificial generativa (AI generative).

Alcance del Desarrollo:

- Investigación y selección de técnicas apropiadas de IA generativa para el desarrollo del modelo de LLM.
- Recopilación y preparación de datos para el entrenamiento del modelo, incluyendo corpus de documentos PDF y preguntas relevantes.
- Diseño e implementación del modelo de LLM, optimizando su capacidad para comprender y generar respuestas precisas.
- Desarrollo de la interfaz de usuario del bot interactivo, incluyendo funciones de carga de documentos y formulación de preguntas.
- Pruebas exhaustivas del sistema para validar la precisión y coherencia de las respuestas generadas.
- Implementación de mejoras y ajustes basados en retroalimentación de usuarios y pruebas adicionales.

Entregables:

- Modelo de LLM entrenado y optimizado para comprender y generar respuestas sobre documentos PDF.
- Bot interactivo funcional integrado con el modelo de LLM.
- Documentación detallada que describe la arquitectura del sistema, instrucciones de instalación y uso, y recomendaciones para futuras mejoras.

Implicaciones Técnicas

SOFTWARE

1. Plataforma Watson AI de IBM:

- Para este proyecto, se empleará la plataforma Watson AI de IBM, la cual proporciona una amplia gama de servicios y herramientas avanzadas de inteligencia artificial y procesamiento del lenguaje natural.

2. Utilización del Modelo Fundacional “llama-2-13b-chat" de Meta:

- Se basará en el modelo fundacional “llama-2-13b-chat" este LLM se encuentra enfocado en la función de chat, contando con 13 billones de tokens que nos ayudarán a una mayor comprensión del lenguaje español e inglés, este modelo fundacional es creado por Meta (anteriormente Facebook), se hará adaptado y optimizado para la tarea específica de consulta de documentos PDF con la función de chat.

3. Integración de Streamlit para la Interfaz de Usuario:

- Para desarrollar la interfaz de usuario del bot interactivo, se utilizará Streamlit, una herramienta de código abierto que permite crear aplicaciones web interactivas con Python de manera sencilla y eficiente.

4. Lenguaje de programación Python y librerías:

- El proyecto se desarrollará utilizando el lenguaje de programación Python, aprovechando su flexibilidad y amplio soporte en la comunidad de desarrollo.

Se hará uso de diversas bibliotecas de Python, como pdf2, pandas, y scikit-learn, entre otras, para realizar tareas específicas como la extracción de texto de documentos PDF, manipulación de datos, y aplicación de técnicas de aprendizaje automático, respectivamente.

HARDWARE

1. Computadoras portátiles o de escritorio: Cada miembro del equipo (4) necesitará una computadora con suficiente potencia para ejecutar las herramientas de desarrollo y realizar tareas relacionadas con el proyecto. Estas computadoras deberían tener al menos un procesador moderno y suficiente memoria RAM para manejar las cargas de trabajo de desarrollo, como el entrenamiento del modelo de LLM y la ejecución del bot interactivo.

2. Especificaciones de almacenamiento y procesadores Se recomienda que las computadoras tengan al menos un procesador multi-núcleo moderno (por ejemplo, Intel Core i5 o superior, o procesador AMD equivalente), al menos 8 GB de RAM (preferiblemente 16 GB o más para un rendimiento óptimo), y suficiente espacio de almacenamiento para instalar software y almacenar datos relacionados con el proyecto.

3. Conexión a Internet estable: Dado que el proyecto implica el uso de servicios en la nube y la colaboración en línea, es importante que cada miembro del equipo tenga acceso a una conexión a Internet estable y rápida para comunicarse, compartir archivos y acceder a recursos en línea de preferencia vía ethernet, por lo cual sería requerido un cable y un modem.

4.Monitores adicionales (opcional): Para mejorar la productividad y la comodidad visual durante largas sesiones de trabajo, cada miembro del equipo podría considerar usar uno o más monitores adicionales junto con su computadora principal.

5. Dispositivos periféricos: Teclado, mouse u otros dispositivos periféricos que cada miembro del equipo prefiera para su comodidad personal durante la codificación y otras tareas.

INFRAESTRUCTURA

Entorno de Desarrollo:

Se requerirá un entorno de desarrollo integrado (IDE) compatible con Python, como Jupyter Notebook a través de Conda, entorno de trabajo para proyectos de AI y Python.

Plataforma Watson x AI de IBM:

Acceso a la plataforma Watson x AI de IBM para utilizar sus servicios de inteligencia artificial y procesamiento del lenguaje natural. Esto implica la configuración de credenciales (API, ID proyecto) y permisos de acceso.

Bibliotecas de Python:

Instalación de las bibliotecas de Python necesarias, incluyendo pdf2, pandas, scikit-learn, así como otras bibliotecas adicionales para tareas específicas como procesamiento de texto, manejo de datos y aprendizaje automático.

Modelo "LLAMA" de Meta:

Descarga e implementación del modelo fundacional "LLAMA" creado por Meta, adaptándolo y optimizándolo para la consulta de documentos PDF en el contexto del proyecto.

Streamlit:

Configuración de Streamlit para el desarrollo de la interfaz de usuario del bot interactivo, lo que implica la instalación de la biblioteca y la configuración de la aplicación web en el entorno de desarrollo.

Recursos de Hardware:

Utilización de la laptop con las características nombradas anteriormente (Intel Core i5 o superior, o procesador AMD equivalente) como la principal infraestructura de hardware para el desarrollo y ejecución del proyecto. Es importante asegurar que las laptop cuenten con suficientes recursos de memoria y capacidad de procesamiento para manejar las tareas de entrenamiento del modelo y ejecución del bot.

Almacenamiento de Datos:

Para el almacenamiento de los datos se realizará a través de la caché al momento de usar el modelo de IA Generativa, para ir acorde a la privacidad de datos de los documentos que se vayan a manejar.

Control de Versiones:

Configuración de un sistema de control de versiones como Git, junto con un repositorio en una plataforma como GitHub, para colaboración entre miembros del equipo y seguimiento de cambios en el código fuente.

Documentación y Gestión de Proyectos:

Utilización de herramientas de documentación y gestión de proyectos como Trello para organizar tareas, mantener registros y colaborar de manera efectiva en el desarrollo del proyecto.

Estándares Aplicados

Normas de Programación:

Se seguirán las convenciones de estilo de código de Python, como PEP 8, para garantizar la legibilidad y consistencia del código.
Se utilizarán nombres descriptivos y significativos para variables, funciones y clases, siguiendo las mejores prácticas de nomenclatura.

Documentación del Código:

Todo el código desarrollado estará debidamente documentado utilizando comentarios claros y concisos.
Se proporcionará documentación adicional en forma de docstrings para describir el propósito y la funcionalidad de las funciones y métodos.

Control de Versiones:

Se empleará un sistema de control de versiones como Git para gestionar el código fuente del proyecto.
Se seguirán prácticas de control de versiones como ramificación (branching) y fusiones (merges) para gestionar cambios en el código de manera organizada y colaborativa.

Gestión de Proyectos:

Se utilizará una metodología de gestión de proyectos, como Scrum, para planificar y seguir el progreso del proyecto.
Se asignarán roles y responsabilidades claras dentro del equipo de desarrollo para garantizar una distribución efectiva del trabajo.

Pruebas y Validación:

Se realizarán pruebas unitarias y de integración de manera continua durante el desarrollo del proyecto para garantizar la funcionalidad y fiabilidad del sistema.
Se llevarán a cabo pruebas de aceptación con usuarios finales para validar la usabilidad y efectividad del bot interactivo.

Seguridad de Datos:

Se implementarán medidas de seguridad para proteger la confidencialidad e integridad de los datos, especialmente aquellos relacionados con documentos PDF cargados por los usuarios.

Se seguirán prácticas recomendadas de seguridad de datos, como encriptación y autenticación, para proteger el acceso no autorizado a la información del sistema. Optimización del Rendimiento:

Se realizarán esfuerzos para optimizar el rendimiento del sistema, tanto en términos de velocidad de respuesta como de consumo de recursos, para garantizar una experiencia fluida para los usuarios.
Se llevarán a cabo pruebas de carga y rendimiento para identificar y abordar posibles cuellos de botella y mejorar la escalabilidad del sistema.
Mantenimiento y Soporte:

Se establecerá un plan de mantenimiento y soporte para garantizar la continuidad del funcionamiento del sistema después de su implementación.
Se proporcionará documentación detallada y capacitación para facilitar futuras actualizaciones y modificaciones del sistema.

Conclusiones

En conclusión, este proyecto representa un esfuerzo integral para desarrollar un modelo de LLM de AI generativa y un bot interactivo de consulta de documentos PDF, aprovechando técnicas avanzadas de inteligencia artificial y procesamiento del lenguaje natural.
Desde la investigación y selección de técnicas apropiadas hasta la implementación de la interfaz de usuario y las pruebas exhaustivas del sistema, se ha tomado en cuenta diversos estándares, así como también se han considerado implicaciones técnicas y de seguridad que podrían ser necesarios en distintas etapas del desarrollo. Con un enfoque en la innovación, la eficiencia y la seguridad de los datos, este proyecto busca ofrecer una solución eficaz para facilitar el acceso y comprensión de información contenida en documentos PDF, con potencial para impactar positivamente en diversas áreas, y con un amplio rango de usuarios, un ejemplo de esto son los estudiantes quienes deben de leer una gran cantidad de información durante sus estudios.

Bibliografía

Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play, David Foster, Editorial O'Reilly Media, Año de publicación: 2019

Practical Natural Language Processing, Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta && Harshit Surana, Editorial O'Reilly Media, Año de publicación: 2020

What are foundation models?, IBM Research, https://research.ibm.com/blog/what-are-foundation-models , Mike Murphy, 09 Mayo 2022

What is generative AI?, IBM Research , https://research.ibm.com/blog/what-is-generative-AI , Kim Martineau , 20 Abril 2023