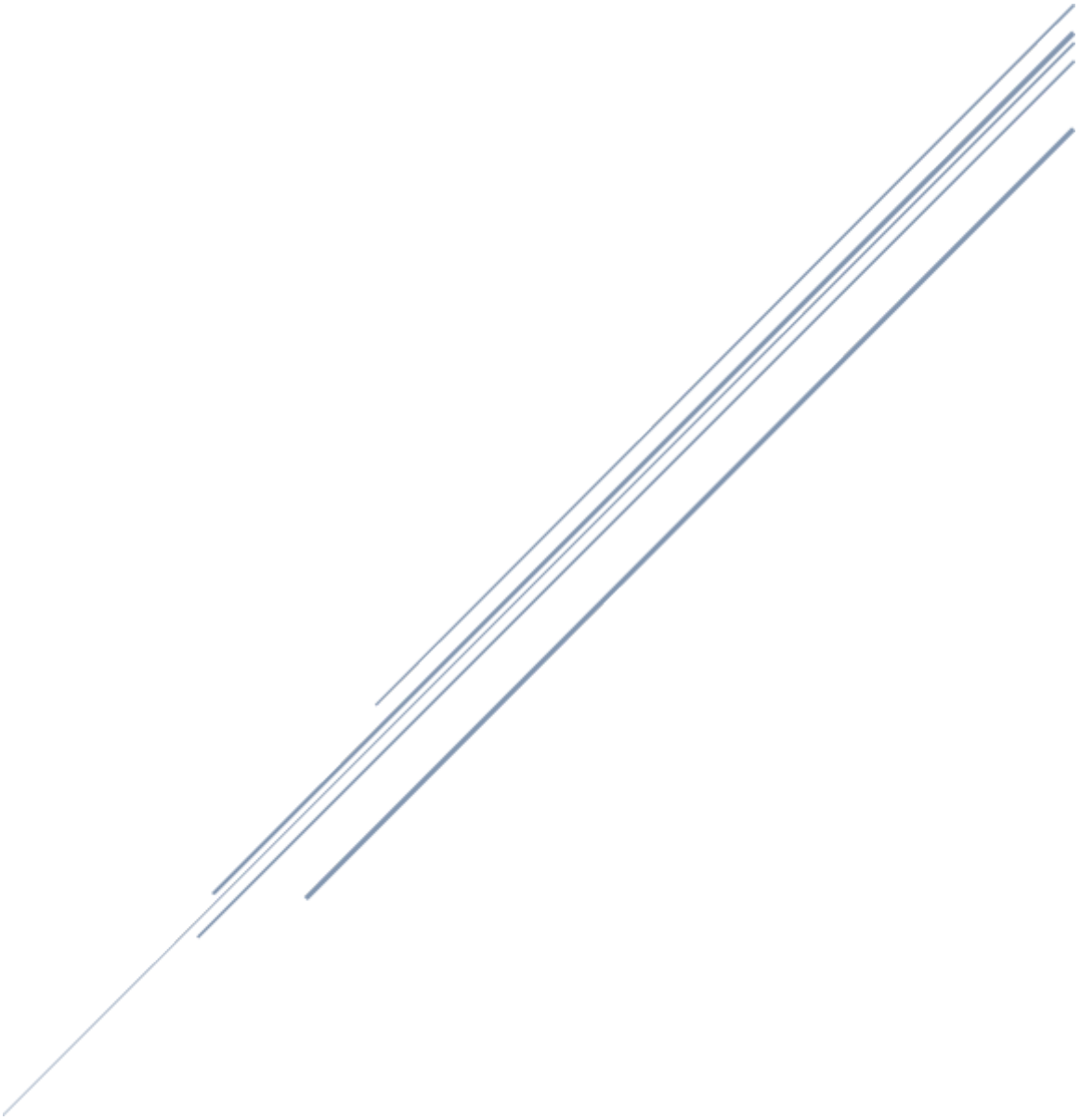
ANTEPROYECTO

UCASAL



Alumno: Alejo Torres

Título del trabajo

El título del trabajo a desarrollar será: “Síntesis de texto usando Modelos de Lenguaje”.

Profesor guía propuesto

La Mg. Lic. María Lorena Talamé de la Universidad Católica de Salta ha sido seleccionada como directora propuesta para liderar el seguimiento y acompañamiento del proyecto, Esto debido a la relación con el docente y su amplio conocimiento en materia de investigación e inteligencia artificial.

Breve descripción del trabajo

El presente trabajo tiene por objetivo la construcción de una aplicación que pueda resumir y sintetizar archivos PDF, como lo pueden ser Papers, revistas científicas, libros de literatura etc. Aprovechando la tecnología conocida como *Large Language Models (LLM)* originada gracias a la arquitectura Transformer, junto a la aplicación de algoritmos de *clustering* o clasificación sobre espacios de alta dimensionalidad como lo son los *Word Embeddings.*

El resultado final de este trabajo es desarrollar una aplicación web full-stack utilizando las últimas herramientas y marcos de trabajos orientados al desarrollo de aplicaciones impulsadas por IA. Que cumpla con el objetivo de exponer una interfaz al usuario quien podrá cargar un archivo en formato PDF y gracias a las técnicas y algoritmos mencionados anteriormente, recibirá un resumen conciso y certero de la información más relevante.

Objetivo General

El objetivo principal será producir una aplicación web que permita resumir la información de un archivo en formato PDF a través del poder de inferencia de un LLM. Formando así una potente herramienta que permita agilizar y mejorar los procesos de estudio e investigación sobre cualquier tema puntual.

Objetivos Específicos

Los objetivos propuestos en el marco del presente trabajo serán:

1. **Investigación sobre la arquitectura Transformer y algoritmos de Clustering:** El trabajo requiere investigar principalmente sobre los papers originales, complementandose con artículos científicos, masterclasses, seminarios y charlas informativas, con el objetivo de recopilar toda la información necesaria para obtener un dominio completo sobre el tema.

2. **Explorar el ecosistema de desarrollo de aplicaciones impulsadas por IA:** Contando con un entendimiento bien formado sobre los LLM me dispondré a explorar el ecosistema de herramientas y librerías como LangChain, CTransformer, BitsAndBytes, StreamLit, ChainLit, entre otras. Para comprender cuál es el paradigma a seguir a la hora de pensar y desarrollar servicios, aplicaciones y abstracciones sobre modelos del lenguaje de estas características. Que luego me permitirán abordar el desarrollo de la aplicación full-stack

3. **Diseñar la arquitectura del sistema:** utilizando la herramienta de *Jupyter Notebooks* se apunta crear la arquitectura y las conexiones necesarias para integrar cada una de las herramientas, servicios y APIs necesarias para cumplir la tarea en cuestión

4. **Validar los resultados del software:** El objetivo sería evaluar la precisión y la utilidad del software desarrollado.Sometiendo al Software a pruebas de síntesis sobre textos conocidos,

exponiendo sus resultados a expertos del tema en particular para que brinden su apreciación y puedan puntuarlos con una escala de valor

5. **Diseñar una interfaz adecuada:** Una vez conseguidos estos objetivos, una de las últimas etapas será desarrollar una interfaz de usuario intuitiva y amigable que permita a los usuarios de la aplicación interactuar con los servicios de inferencia del modelo.

Alcances del trabajo

El alcance del trabajo para el desarrollo de un software de síntesis de texto incluirá los siguientes aspectos:

1. **Exploración de las técnicas de Machine Learning involucradas :** Se requerirá revisar los papers originales, así como artículos científicos, seminarios y charlas informativas para obtener un conocimiento completo y profundo sobre estas técnicas.

2. **Exploración del ecosistema de desarrollo de IA:** Se explorará el ecosistema de herramientas, frameworks y abstracciones creadas por la comunidad para aprovechar los modelos de lenguaje y Machine learning en aplicaciones impulsadas por inteligencia artificial. Se estudiarán librerías como LangChain, CTransformer, BitsAndBytes, StreamLit y ChainLit, con el objetivo de seleccionar herramientas y marcos de trabajo necesarios para construir los servicios e interfaces necesarios para el Software.

3. **Diseño y Desarrollo del Software:** Se incluirá el diseño y el desarrollo del software de síntesis de texto utilizando Clustering e inferencia producto de LLMs

4. **Diseño de la interfaz de usuario:** El alcance del trabajo incluiría el diseño de una interfaz de usuario intuitiva y suficiente para un uso primario, que permita a los usuarios interactuar con los servicios de inferencia y que presente la información y configuraciones de manera clara y comprensible, en un formato útil.

Institución, empresa u organización

Los avances, documentos, Software y pruebas de concepto obtenidos serán puestos a disposición del Departamento de Ingeniería en Informática de la Universidad Católica de Salta. Con el objetivo de servir como una guía de estudio del tema.

Metodología y/o procedimiento a utilizar

La estrategia metodológica que se va a implementar para el desarrollo de este proyecto se basará en el enfoque de "desarrollo en cascada". Este enfoque implica la realización de actividades secuenciales y claramente agrupadas en fases o ciclos. La decisión de utilizar este enfoque se debe al hecho de que el equipo de trabajo será reducido, contando con solo una persona dedicada al desarrollo del proyecto, por lo que no es necesario realizar énfasis en las comunicaciones internas dentro del equipo

El objetivo principal de este proyecto es desarrollar un software que permita sintetizar el contenido de un archivo en PDF generando así un resumen conciso y útil del mismo, utilizando algoritmos de clustering e inteligencia artificial. A lo largo de las fases de esta metodología, se contempla una investigación profunda de los conceptos y tecnologías relacionadas con la arquitectura que sustenta a los LLMs y las técnicas de Machine Learning asociadas. Posteriormente, se llevará a cabo un análisis detallado de las necesidades y problemáticas las cuales este software pretende solucionar.

Una vez completado el análisis, se procederá al diseño del software de síntesis de texto, definiendo los requisitos y características necesarias. A continuación, se llevará a cabo el desarrollo del software utilizando las tecnologías pertinentes. Finalmente, se realizarán pruebas exhaustivas con respecto a las inferencias realizadas por el mismo con el objetivo de validar la capacidad de la aplicación

Con este enfoque metodológico, se espera lograr un desarrollo estructurado y efectivo de un Software de Síntesis de texto que pueda constituirse como una herramienta útil que complemente el proceso de estudio y enseñanza dentro de la Universidad Católica de Salta

Plan de trabajo

A continuación se proporciona un plan de trabajo tentativo para el proyecto

1. Investigación y familiarización con las técnicas y tecnologías(4 semanas)
    - Revisión de librerías, proyectos y estudios relacionados con el Procesamiento del lenguaje natural, técnicas de clustering y LLMs.
    - Familiarización con las tecnologías y herramientas existentes en el campo.
2. Definición de funcionalidades y alcance de la aplicación (1 semana)
    - Definir y acotar el alcance de los servicios de inferencia
3. Pruebas y depuración (2 semanas)
    - Realización de pruebas de concepto de las distintas herramientas en diferentes escenarios y con corpus de texto de prueba.
    - Identificación y corrección de errores y problemas encontrados durante las pruebas.
4. Diseño del sistema y de la interfaz de usuario (2 semanas)
    - Diseño de la arquitectura del sistema
    - Creación de las directivas y establecer valor a los hiperparametros
5. Desarrollo del software (8 semanas)
    - Implementación del software de Síntesis de Texto utilizando inferencia de LLMS
    - Integración de algoritmos de Machine Learning.
6. Validación y optimización (4 semanas)
    - Evaluación del rendimiento y la precisión del software utilizando datos reales y expertos en los temas específicos.
    - Optimización del software para mejorar su eficiencia y rendimiento.

Duración Total del Proyecto: Aproximadamente 21 semanas

Este plan de trabajo permitirá abordar de manera efectiva la investigación, el desarrollo de aplicaciones y la presentación final, abarcando el objetivo del proyecto en su totalidad. Cabe aclarar que las estimaciones de tiempo son tentativas y pueden sufrir modificaciones.

Bibliografía

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Pytorch(2023) Guía de documentación de la librería: https://pytorch.org/docs/stable/index.html

LangChai (2023) Guía de documentación de la librería: https://docs.langchain.com/docs/

StreamLit (2023)Guía de documentación de la librería: https://docs.streamlit.io/

NextJs (2022)Guía de documentación de la librería: https://nextjs.org/docs

Neural Networks: Zero to Hero (2022), Andrej Karpath, Blog Personal: https://karpathy.ai/zero-to-hero.html

transformers(2022)Repositorio de codigo de la libreria: https://github.com/huggingface/transformers

Material de estudio utilizado en la Universidad Politécnica de Madrid en las asignaturas de Aprendizaje automático II y Sistemas inteligentes, en la carrera de Data Science e inteligencia Artificial dictadas por el Dr. Francisco García Serradilla

Camacho, J. D. V. (2020). Desarrollo en cascada (waterfall) VS desarrollo agile-SCRUM.