Introducción

Antecedentes y Contexto

El sector de la tecnología financiera (fintech) ha experimentado un crecimiento sin precedentes en la última década, revolucionando la forma en que las personas y las organizaciones gestionan sus finanzas. Los métodos tradicionales de planificación financiera, que a menudo implicaban la contabilidad manual, las consultas en persona con asesores financieros o las herramientas de presupuestación estáticas, están siendo reemplazados rápidamente por soluciones digitales innovadoras. Estos avances están impulsados por la creciente adopción de la inteligencia artificial (IA), el procesamiento del lenguaje natural (PLN) y las tecnologías de big data.

Hoy en día, los asistentes financieros personales, los chatbots y los robo-asesores dominan el panorama. Empresas como Mint, YNAB (You Need a Budget) y Wealthfront proporcionan recomendaciones personalizadas y conocimientos automatizados, permitiendo a los usuarios rastrear gastos, establecer metas financieras y tomar decisiones de inversión informadas. De manera similar, los sistemas de IA conversacional como Cleo y Erica (desarrollado por Bank of America) permiten a los usuarios interactuar con sus datos financieros de manera fluida a través de interfaces basadas en chat. Estos sistemas se basan en modelos de aprendizaje automático sofisticados y en un extenso entrenamiento de datos para ofrecer asesoramiento personalizado.

Esta transición a enfoques digitales está impulsada por la demanda de conveniencia, accesibilidad y asistencia en tiempo real. A diferencia de la planificación financiera tradicional, que requería un tiempo y una experiencia significativos, los sistemas modernos democratizan la gestión financiera al hacer que las herramientas sean intuitivas y fácilmente disponibles. Los usuarios ahora pueden acceder a conocimientos desde cualquier lugar, recibir respuestas instantáneas y beneficiarse de estrategias impulsadas por IA que anteriormente estaban limitadas a consultas de expertos. Esto es particularmente beneficioso porque empodera a todos los usuarios, independientemente de su alfabetización financiera, en tiempos pasados estos eran servicios adaptados y disponibles únicamente para una minoría adinerada de la población.

A pesar de estos avances, la mayoría de las soluciones existentes dependen de un extenso entrenamiento de modelos de IA, lo que requiere grandes conjuntos de datos y recursos computacionales. Esta dependencia introduce desafíos en la escalabilidad, la adaptabilidad a las necesidades específicas de los usuarios y el mantenimiento de bases de conocimiento actualizadas. La próxima frontera en la innovación fintech radica en aprovechar sistemas más dinámicos, eficientes y adaptables que puedan generar respuestas precisas y contextualmente relevantes sin el reentrenamiento tradicional de modelos.

Problema y Objetivos de Investigación

La adopción de sistemas de Generación Aumentada por Recuperación (RAG) en la planificación financiera ha mejorado significativamente la eficiencia y precisión de los sistemas de IA conversacional. La mayoría de las implementaciones existentes dependen de sistemas RAG basados en vectores, donde las incrustaciones de texto se indexan y buscan por similitud semántica. Aunque efectivo, este enfoque tiene limitaciones inherentes para capturar las relaciones estructuradas intrincadas que son críticas para ciertos campos como la planificación financiera, donde las interdependencias entre conceptos como gastos, inversiones, metas y riesgos son altamente relacionales.

Los sistemas RAG basados en grafos, que se centran en búsquedas de relaciones semánticas a través de estructuras de grafos, ofrecen una alternativa al codificar relaciones de manera explícita y dinámica. Esta capacidad puede potencialmente mejorar la comprensión contextual de los datos

financieros, permitiendo una representación más matizada e interconectada de las metas del usuario, patrones de gasto y estrategias de inversión. Sin embargo, los sistemas RAG basados en grafos aún no se han explorado en el dominio de la planificación financiera.

Además, las soluciones de IA actuales en este dominio a menudo requieren un tiempo, recursos y experiencia sustanciales para entrenar y ajustar modelos de lenguaje grandes (LLMs). Esta dependencia crea barreras para individuos y organizaciones que no tienen acceso a vastos conjuntos de datos y poder computacional, limitando la democratización de estas herramientas avanzadas. Al investigar la viabilidad de un enfoque de “zero-shot"—confiando únicamente en la afinación de prompts sin reentrenamiento de modelos—esta investigación busca reducir estas barreras y hacer que los chatbots de planificación financiera sean más accesibles y asequibles.

Esta tesis aborda la brecha introduciendo un sistema RAG basado en grafos para la planificación financiera y explorando cómo un enfoque de afinación de prompts puede desempeñarse en este contexto. Se espera que los hallazgos contribuyan al desarrollo de herramientas de IA prácticas y rentables que estén verdaderamente disponibles para una audiencia más amplia, mientras se exploran las ventajas únicas que las estructuras de grafos pueden aportar a este dominio.

El objetivo principal de esta tesis es diseñar y evaluar un chatbot de planificación financiera basado en un sistema RAG de grafos, utilizando solo la afinación de prompts para generar salidas significativas. Los objetivos secundarios incluyen:

1. Evaluar el rendimiento del sistema RAG de grafos en comparación con los sistemas RAG basados en vectores en la provisión de recomendaciones financieras relevantes y precisas.
2. Explorar la usabilidad y practicidad de un enfoque de zero-shot en aplicaciones del mundo real.
3. Identificar las ventajas y limitaciones de la búsqueda semántica basada en grafos para capturar relaciones financieras.

Al abordar estos objetivos, esta investigación busca contribuir al campo más amplio de la IA en fintech, ofreciendo ideas sobre la aplicabilidad de los sistemas RAG basados en grafos y el potencial de los métodos zero-shot para hacer que las soluciones de IA sean más accesibles e impactantes.

Alcance y Delimitaciones

Esta investigación se centra en diseñar y evaluar un chatbot de planificación financiera utilizando un sistema de Generación Aumentada por Recuperación (RAG) basado en grafos, confiando exclusivamente en un enfoque de ajuste de prompts “zero-shot”. Aunque las técnicas exploradas en este estudio pueden eventualmente ser útiles para desarrollar chatbots genéricos específicos de dominio, el enfoque principal está limitado al dominio de la planificación financiera. El alcance no se extiende a la creación de un marco generalizado para chatbots basados en grafos, ni esta investigación abordará la implementación de sistemas RAG utilizando arquitecturas alternativas como los enfoques basados en vectores, ya que estos están ampliamente cubiertos en la literatura existente.

En cambio, este trabajo adopta un enfoque exploratorio, experimentando con varias técnicas y configuraciones necesarias para el desarrollo del sistema. Esto incluye:

- Investigar diferentes métodos de búsqueda en grafos para mejorar la recuperación de relaciones semánticas.
- Explorar configuraciones de prompts y estrategias de ajuste para mejorar la calidad de las respuestas del chatbot.
- Experimentar con técnicas de generación de respuestas adaptadas al contexto de la planificación financiera.

La investigación adopta una metodología de "libre para todos", permitiendo flexibilidad en la prueba e integración de diferentes enfoques, herramientas y algoritmos según sea necesario para lograr los objetivos del proyecto. Sin embargo, el estudio se mantendrá estrictamente dentro de los límites de los sistemas RAG basados en grafos y no profundizará en tecnologías o marcos no relacionados.

Al mantener este enfoque, la investigación pretende ofrecer conocimientos específicos sobre la aplicación de sistemas RAG basados en grafos para la planificación financiera, sin intentar generalizar los hallazgos más allá del dominio o arquitectura especificados.

Preguntas de Investigación e Hipótesis

Esta investigación busca explorar el potencial de un sistema de Generación Aumentada por Recuperación (RAG) basado en grafos para la planificación financiera abordando las siguientes preguntas clave:

1. Efectividad del Sistema RAG Basado en Grafos
   - ¿Con qué efectividad puede un sistema RAG basado en grafos capturar las relaciones semánticas necesarias para ofrecer asesoramiento financiero personalizado y contextualmente preciso?
2. Viabilidad del Ajuste de Prompts Zero-Shot
   - ¿Puede un enfoque de ajuste de prompts zero-shot generar recomendaciones financieras significativas y precisas sin requerir un extenso entrenamiento del modelo?
3. Comparación con Sistemas RAG Basados en Vectores
   - ¿Cómo se compara el rendimiento de un sistema RAG basado en grafos con los sistemas basados en vectores en términos de relevancia, comprensión contextual y adaptabilidad dentro del dominio de la planificación financiera?
4. Practicidad y Escalabilidad del Sistema
   - ¿Cuáles son las ventajas prácticas, limitaciones y posibles desafíos de usar un sistema RAG basado en grafos para aplicaciones de planificación financiera en el mundo real?

Estas preguntas guían la exploración y evaluación del chatbot propuesto, ayudando a identificar tanto su potencial técnico como sus implicaciones prácticas.

Metodología

Este estudio se llevó a cabo en dos etapas principales: investigación y diseño y desarrollo del sistema. Cada etapa jugó un papel crucial en la realización de un chatbot de planificación financiera basado en un sistema de Generación Aumentada por Recuperación (RAG) basado en grafos, asegurando tanto una base teórica como una implementación práctica.

Investigación

La fase de investigación comprendió dos objetivos principales:

1. Encuesta de Sistemas RAG Basados en Grafos

Para comprender los principios y metodologías subyacentes para implementar un sistema RAG basado en grafos, se realizó una revisión exhaustiva de la literatura académica relevante. Se analizaron críticamente artículos que discutían técnicas para la recuperación, indexación y generación de respuestas dentro de arquitecturas basadas en grafos. Esto fue esencial para obtener una comprensión integral de los enfoques existentes y su aplicabilidad para construir el sistema propuesto.

2. Revisión de Sistemas de Chatbots Fintech

Una revisión secundaria se centró en explorar trabajos existentes en el dominio de los chatbots fintech, particularmente aquellos que emplean sistemas RAG. Esta investigación proporcionó valiosos conocimientos sobre cómo se abordaron los desafíos específicos del dominio, como la

relevancia contextual, la comprensión de la intención del usuario y la representación de datos financieros. Los hallazgos de esta revisión informaron las decisiones de diseño adaptadas al dominio de la planificación financiera.

Estas dos corrientes de investigación aseguraron que el diseño del sistema estuviera basado en técnicas establecidas mientras se abordaban los requisitos únicos de la aplicación objetivo.

Diseño y Desarrollo del Sistema

El sistema se implementó en varias fases, cada una abordando componentes distintos necesarios para la funcionalidad del chatbot.

1. Extracción de Conocimiento

Se utilizaron herramientas y bibliotecas como LangChain, Beautiful Soup y APIs de Gemini y OpenAI para extraer conocimientos de planificación financiera de varias fuentes web. La información se procesó en una estructura de grafo con nodos y aristas, donde ambos contenían anotaciones textuales. Por ejemplo:

- Nodo: “ARRENDAMIENTO DE COCHE”
- Arista: “ES UNA FUENTE DE”
- Nodo Objetivo: “DEUDA”

Además, el grafo capturó múltiples relaciones por nodo, como se ilustra en:

- “ARRENDAMIENTO DE COCHE” -> “REQUIERE” -> “PAGOS MENSUALES”.

La estructura del grafo se almacenó en Neo4j, mientras que las representaciones incrustadas del contenido textual se almacenaron en una base de datos de vectores utilizando MongoDB. Esta representación dual permitió flexibilidad en las estrategias de recuperación.

2. Recuperación de Conocimiento

El enfoque principal fue recuperar conocimiento de las incrustaciones del grafo en lugar de la estructura del grafo en sí. Se emplearon técnicas como el clustering k-means y la similitud coseno para identificar nodos y aristas relevantes. Estos métodos permitieron una recuperación eficiente de puntos de datos semánticamente similares basados en las consultas de los usuarios.

3. Extracción de Subgrafos

Los nodos y aristas recuperados formaron la base para construir un subgrafo, representando el subconjunto de información más relevante para una consulta dada. Se exploraron varias técnicas para la extracción de subgrafos, que se detallarán en secciones posteriores.

4. Generación de Respuestas

El subgrafo extraído se procesó para generar respuestas en lenguaje natural. Esto se logró utilizando la API de Gemini, aprovechando varios modelos para transformar los datos estructurados del grafo en salidas coherentes y contextualmente apropiadas.

5. Evaluación y Pruebas

Para evaluar el rendimiento del chatbot, se creó un conjunto de consultas elaboradas a mano que cubrían diversos escenarios financieros. Las respuestas del sistema se compararon con las de otros sistemas y modelos de propósito general. La evaluación se basó en las siguientes métricas:

- Precisión: La corrección de la información proporcionada en relación con la consulta.
- Corrección Factual: Si la respuesta se alineaba con el conocimiento financiero verificado.
- Eficiencia: La velocidad y el costo computacional de generar una respuesta.
- Robustez: La capacidad del sistema para manejar consultas ambiguas o incompletas.

Este marco metodológico aseguró una exploración e implementación sistemática del sistema RAG basado en grafos propuesto, proporcionando una base para contribuciones tanto técnicas como prácticas al campo de la planificación financiera.

Estructura de la Tesis

Esta tesis está organizada en varios capítulos, cada uno abordando un aspecto específico del proceso de investigación y contribuyendo al objetivo general de desarrollar un chatbot de planificación financiera basado en un sistema de Generación Aumentada por Recuperación (RAG) basado en grafos. La estructura es la siguiente:

- Capítulo 1: Introducción

La introducción proporciona una visión general de la investigación, incluyendo el contexto, la declaración del problema, los objetivos de la investigación y la motivación del estudio. También se describen el alcance y las delimitaciones del trabajo, y se presentan las preguntas de investigación y las hipótesis que guían el estudio.

- Capítulo 2: Revisión de la Literatura

Este capítulo revisa la literatura relevante en las áreas de chatbots fintech, sistemas RAG y representaciones de conocimiento basadas en grafos. Incluye una discusión de los enfoques existentes para el desarrollo de chatbots en el dominio de la planificación financiera y una encuesta de diferentes técnicas para sistemas RAG basados en grafos, proporcionando la base para la investigación.

- Capítulo 3: Metodología

El capítulo de metodología describe el diseño de la investigación y los métodos empleados para construir el chatbot de planificación financiera propuesto. Incluye una explicación detallada de los pasos de recopilación y preprocesamiento de datos, el diseño y desarrollo del sistema, las técnicas de recuperación utilizadas y el proceso de generación de lenguaje natural. La estrategia de evaluación y las métricas de rendimiento también se describen en este capítulo.

- Capítulo 4: Arquitectura e Implementación del Sistema

Este capítulo proporciona una descripción detallada de la arquitectura del sistema, incluyendo los componentes involucrados en la extracción de conocimiento, la construcción de grafos, la recuperación y la generación de lenguaje natural. También se discute el proceso de implementación, con especial atención a las herramientas, marcos y bases de datos utilizados para desarrollar el sistema.

- Capítulo 5: Evaluación y Resultados

En este capítulo se presentan los resultados de la evaluación del sistema. Se evalúa el rendimiento del chatbot basado en varias métricas como precisión, corrección factual, eficiencia y robustez. Los hallazgos de los experimentos de evaluación se discuten en detalle y se analizan las implicaciones de los resultados.

- Capítulo 6: Conclusión y Trabajo Futuro

El capítulo de conclusión resume los hallazgos clave de la investigación, reflexiona sobre la contribución del trabajo y discute sus limitaciones. También sugiere posibles vías para futuras investigaciones, incluyendo mejoras al sistema y aplicaciones adicionales de enfoques RAG basados en grafos en otros dominios.