En la era digital actual, donde la información se genera y se acumula a un ritmo sin precedentes, el análisis efectivo de datos se ha convertido en un pilar fundamental para la toma de decisiones informadas en múltiples sectores. La proliferación de tecnologías digitales y la interconexión global han facilitado el acceso a vastos volúmenes de datos, que abarcan desde estadísticas económicas hasta indicadores de salud pública. En este contexto, los datos no solo son esenciales para comprender realidades complejas, sino que también actúan como un recurso estratégico que permite a gobiernos, empresas y organizaciones no gubernamentales formular políticas y estrategias que respondan de manera eficaz a las necesidades cambiantes de la población.

La vasta cantidad de información generada se presenta en diversos formatos, cada uno con sus propias características y desafíos. **Los datos estructurados**, organizados en tablas y bases de datos, permiten un acceso y análisis más sencillo, facilitando la identificación de patrones y tendencias; ejemplos de estos datos incluyen hojas de cálculo y bases de datos relacionales, que son ampliamente utilizados en entornos empresariales y académicos. Por otro lado, **los datos no estructurados** representan una porción significativa de la información disponible, abarcando correos electrónicos, documentos de texto, imágenes y publicaciones en redes sociales; la naturaleza variable y compleja de estos datos plantea retos importantes para su análisis, ya que requieren técnicas avanzadas para extraer información útil. Además, el fenómeno del **Big Data** ha transformado el panorama del análisis de datos, con conjuntos de datos extremadamente grandes y complejos que superan las capacidades de las herramientas tradicionales, el Big Data se caracteriza por las "3 V": volumen, velocidad y variedad. Esta realidad exige nuevas estrategias y tecnologías que permitan procesar y analizar información en tiempo real.

En este contexto multifacético, la información también fluye a través de **plataformas digitales** que fomentan la interacción y el intercambio constante de datos. Redes sociales, blogs y sitios web son ejemplos donde los usuarios generan y acceden a contenido continuamente, contribuyendo a un ecosistema completamente dinámico. En este vasto mar de información, las estadísticas oficiales emergen como una fuente confiable que proporciona una base sólida sobre la cual se pueden construir análisis significativos. Instituciones como la Oficina Nacional de Estadística e Información (ONEI) desempeñan un papel crucial al compilar y presentar anuarios que reflejan las dinámicas sociales y económicas del país. Estos anuarios estadísticos no solo ofrecen cifras y estadísticas organizadas, sino que también incluyen texto no estructurado que proporciona contexto y análisis, enriqueciendo la comprensión de los datos.

Sin embargo, el análisis tradicional enfrenta obstáculos significativos: la complejidad y el volumen creciente de información pueden dificultar la extracción de patrones relevantes. En este sentido, los modelos de lenguaje grande (LLM) surgen como herramientas innovadoras con el potencial de transformar nuestra interacción con los datos tabulares y textuales. Estos modelos están diseñados para procesar lenguaje natural y pueden interpretar datos estructurados con mayor precisión y eficiencia. La capacidad de los LLM para analizar grandes volúmenes de información podría revolucionar cómo abordamos los anuarios estadísticos y otros conjuntos de datos complejos. Al integrar tanto los datos numéricos como el texto explicativo, los LLM permiten una exploración más profunda y holística de la información, facilitando la identificación de tendencias, correlaciones y insights que pueden ser cruciales para la formulación de políticas informadas.

Motivacion

A medida que avanzamos hacia un futuro donde los datos son cada vez más abundantes e influyentes en nuestras vidas cotidianas, es esencial adoptar enfoques innovadores que maximicen su valor. Esta investigación busca explorar cómo los LLM pueden ser utilizados para mejorar el análisis estadístico en contextos específicos, particularmente en el ámbito de los anuarios estadísticos. En última instancia, este esfuerzo no solo busca potenciar e enriquecer el campo del análisis estadístico, sino que también puede impactar significativamente en la formulación de políticas públicas basadas en evidencia sólida. La implementación efectiva de estas tecnologías podría contribuir a un enfoque más eficaz y basado en datos en la toma de decisiones, promoviendo así un desarrollo sostenible y adaptativo en un mundo cada vez más interconectado.

Antecedentes

Problemantica

la problemática radica que a pesar de la creciente disponibilidad de datos, el análisis tradicional enfrenta limitaciones significativas que obstaculizan la extracción de información valiosa, especialmente en el ámbito de los anuarios estadísticos. Estos anuarios contienen tanto datos estructurados, como tablas y cifras, como texto no estructurado, que incluye descripciones y análisis contextuales. Esta dualidad presenta un desafío considerable, ya que los métodos convencionales suelen especializarse en uno u otro tipo de datos, lo que limita la capacidad para realizar un análisis integral y holístico. Lo que subraya la necesidad urgente de explorar soluciones innovadoras como los modelos de lenguaje grande (LLM). Estos modelos tienen el potencial no solo de mejorar el análisis estadístico, sino también de facilitar una toma de decisiones más informada y basada en evidencia sólida, contribuyendo así a un desarrollo

Objetivo

Objetivo General

Explorar y desarrollar metodologías que integren modelos de lenguaje grande (LLM) para mejorar el análisis de datos en los anuarios estadísticos de la Oficina Nacional de Estadística e Información (ONEI), facilitando así la interpretación tanto de datos estructurados como de texto no estructurado, con el fin de proporcionar insights más profundos y relevantes basadas en evidencia sólida.

Objetivos Específicos

1. **Desarrollar un marco metodológico** que integre modelos de lenguaje grande (LLM) para el análisis de datos en los anuarios estadísticos de la Oficina Nacional de Estadística e Información (ONEI), permitiendo una interpretación más efectiva de los datos estructurados y no estructurados.
2. **Implementar un prototipo de análisis** basado en LLM que mejore las respuestas brindadas sobre el contexto específico de los anuarios estadísticos, optimizando la interacción con la información contenida en ellos.
   * **2.1. Implementar un sistema de recuperación de información** que permita acceder a los datos presentes en los anuarios estadísticos de manera eficiente.
   * **2.2. Realizar la segmentación adecuada** de la información, diferenciando entre datos estructurados (tablas) y no estructurados (texto).
   * **2.3. Crear una base de datos vectorial** para almacenar los embeddings correspondientes a la información procesada, facilitando su recuperación.
   * **2.4. Recuperar la información más relevante** almacenada en función de las consultas específicas realizadas por los usuarios.
   * **2.5. Generar respuestas precisas mediante LLM**, integrando el contexto ampliado con la información recuperada y las consultas planteadas.
3. **Evaluar la efectividad del modelo**, midiendo su precisión, eficiencia y relevancia en la respuesta a preguntas específicas basadas en el contexto proporcionado por los anuarios estadísticos.

Propuesta de Solución

La solución propuesta para optimizar el análisis de datos en los anuarios estadísticos de la Oficina Nacional de Estadística e Información (ONEI) se basa en la integración de modelos de lenguaje grande (LLM). Este enfoque busca mejorar la recuperación y análisis de información, permitiendo una interpretación más precisa tanto de los datos estructurados como de los no estructurados.

El primer paso consiste en implementar un sistema eficiente para extraer y segmentar la información contenida en los anuarios. Esto incluye procesar datos tabulares (estructurados) y texto descriptivo (no estructurado), asegurando que ambos tipos de información sean tratados adecuadamente. Posteriormente, se generarán embeddings que representen semánticamente el contenido procesado, lo cual permitirá almacenar esta información en una base de datos vectorial diseñada para facilitar su recuperación rápida y relevante.

Para responder consultas específicas, se desarrollará un mecanismo basado en LLM que combine el contexto ampliado con la información recuperada y las consultas del usuario. Este sistema generará respuestas precisas y contextualizadas, optimizando la interacción con los datos estadísticos. Finalmente, se evaluará el modelo utilizando métricas claras que midan su precisión, eficiencia y relevancia en la generación de respuestas, asegurando así su efectividad en el contexto específico de los anuarios estadísticos.

Estructura

Introducción al Estado del Arte

En este capítulo se presenta el **estado del arte**, cuyo propósito es analizar y sintetizar los avances más relevantes en el ámbito de estudio, proporcionando un marco teórico y metodológico que respalde la investigación. Este apartado permitirá identificar las principales contribuciones, enfoques y limitaciones en este campo.

TAT-LLM: A Specialized Language Model for Discrete Reasoning over Tabular and Textual Data

El artículo investiga la capacidad de los modelos de lenguaje grandes (LLM) como GPT-3.5 y GPT-4 para comprender datos de tablas estructuradas. Si bien los LLM han demostrado ser prometedores para el procesamiento del lenguaje natural, sus capacidades para manejar datos semiestructurados como las tablas siguen sin explorarse. El estudio presenta un punto de referencia llamado SUC (Structural Understanding Capabilities), que consiste en varias tareas destinadas a evaluar el desempeño de los LLM en la comprensión de tablas. Los autores destacan que el desempeño está influenciado por factores como el formato de entrada, las técnicas de solicitud y los métodos de serialización de datos. También proponen una solicitud autoaumentada para mejorar la comprensión estructural, demostrando la eficacia de sus métodos a través de evaluaciones empíricas en varios conjuntos de datos. Los hallazgos sugieren que, si bien los LLM tienen una comprensión fundamental de las estructuras de las tablas, aún tienen dificultades con las tareas complejas relacionadas con las tablas.

Towards Quality Benchmarking in Question Answering over Tabular Data in Spanish

Este artículo de investigación aborda la creciente necesidad de metodologías de evaluación efectivas en las capacidades de respuesta a preguntas (QA) de los grandes modelos lingüísticos (LLM) cuando interactúan con datos tabulares, específicamente en español. Presenta Spa-DataBench, un nuevo punto de referencia que consta de diez conjuntos de datos relacionados con varios temas dentro de la sociedad española, e incluye un marco estructurado para evaluar la eficacia con la que los LLM pueden responder a las preguntas derivadas de estos conjuntos de datos. Los hallazgos indican que, si bien los LLM pueden realizar razonamientos sobre datos tabulares, su desempeño en español es notablemente menos efectivo que en inglés, lo que sugiere la necesidad de mejoras continuas en el manejo de las indicaciones en español por parte de estos modelos. Las implicaciones de este estudio subrayan el potencial para futuros avances en el modelado multilingüe, en particular para dominios en los que los puntos de referencia establecidos son limitados.

Investigar capacidades

Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study

El artículo tiene como objetivo investigar las capacidades de los modelos de lenguaje grandes (LLM), específicamente GPT-3.5 y GPT-4, para comprender datos de tablas estructuradas. A pesar de los avances en las tareas de procesamiento del lenguaje natural (NLP), ha habido estudios limitados que evalúen la competencia de los LLM con formatos de entrada de datos estructurados, en particular tablas. Los autores proponen un punto de referencia llamado Capacidades de comprensión estructural (SUC) para evaluar la comprensión de tablas estructuradas a través de siete tareas únicas, lo que revela que el rendimiento de los LLM fluctúa, influenciado por múltiples factores, incluido el diseño de entrada y la estrategia de prompt. Sugieren emplear prompts junto con formatos de entrada bien concebidos para mejorar la comprensión estructural, lo que conduce a mejoras notables en el rendimiento en varias tareas tabulares.

nooooooooooooo

TableLLM: Enabling Tabular Data Manipulation by LLMs in Real Office Usage Scenarios

El artículo presenta un modelo de lenguaje de gran tamaño diseñado específicamente para la manipulación eficaz de datos tabulares, en particular en situaciones relacionadas con la oficina. A diferencia de los modelos tradicionales que tienen dificultades con la naturaleza diversa de los datos tabulados, TableLLM funciona de manera eficiente con tablas integradas en documentos y hojas de cálculo. Los autores detallan un método de supervisión a distancia único para crear datos de entrenamiento especializados, lo que permite que el modelo desarrolle capacidades de razonamiento sólidas adaptadas a diversas tareas, como la consulta, la actualización y la visualización de tablas de datos. A través de evaluaciones exhaustivas, TableLLM demuestra un rendimiento superior a los LLM existentes en el manejo de tareas tabulares complejas, incluida una interacción significativa a través de una aplicación web fácil de usar. Esta investigación enfatiza la importancia de adaptar los modelos de lenguaje a necesidades prácticas específicas, mejorando la productividad en entornos profesionales.

Usar llm para datos tabulares

Unleashing the Potential of Large Language Models for Predictive Tabular Tasks in Data Science

El artículo investiga la aplicación de modelos de lenguaje grandes (LLM) como Llama-2 para manejar de manera eficaz las tareas predictivas asociadas con datos tabulares. Al reconocer las limitaciones de los LLM convencionales para interpretar conjuntos de datos estructurados, el estudio propone un régimen de entrenamiento específico que utiliza un corpus especialmente seleccionado de Kaggle que comprende aproximadamente 13 mil millones de ejemplos. Los autores tienen como objetivo cerrar la brecha de conocimiento al entrenar a los LLM para que realicen clasificación, regresión e imputación de valores faltantes con una precisión excepcional, mostrando mejoras sustanciales con respecto a los puntos de referencia existentes. Los hallazgos subrayan que el entrenamiento previo personalizado en tareas tabulares mejora significativamente el rendimiento en diversos escenarios de modelado predictivo, lo que destaca el potencial de los LLM en aplicaciones de ciencia de datos.

Conversion de tablas

Exploring the Impact of Table-to-Text Methods on Augmenting LLM-based Question Answering with Domain Hybrid Data

El estudio investiga el impacto de diferentes métodos de conversión de tablas a texto en la mejora de los sistemas de preguntas y respuestas basados ​​en modelos de lenguaje extensos (LLM), en particular con datos específicos de dominio que a menudo consisten en una combinación de texto y tablas semiestructuradas. Dados los desafíos que plantean estos formatos híbridos, la investigación examina cuatro métodos representativos para transformar tablas en formatos de texto: Markdown, serialización de plantillas, métodos basados ​​en modelos de lenguaje preentrenados tradicionales (TPLM) y basados ​​en LLM. Los experimentos se llevaron a cabo utilizando un conjunto de datos del mundo real, ICT-DATA, con especial atención a las variaciones de rendimiento en dos marcos de control de calidad: ajuste fino específico del dominio (DSFT) y generación aumentada por recuperación (RAG). Los hallazgos indican que, si bien los métodos basados ​​en LLM generalmente arrojaron resultados superiores, la selección del método correcto influyó drásticamente en el rendimiento, con implicaciones tanto para la investigación académica como para las prácticas de la industria, es decir, que depende del campo en específico la elección del método a usar.

nooooooooooooooooooo

LLMClean: Context-Aware Tabular Data Cleaning via LLM-Generated OFDs

El artículo presenta un nuevo método diseñado para la limpieza automatizada de datos en función del contexto mediante modelos de lenguaje grandes (LLM). El objetivo es mejorar la calidad de los datos, algo esencial para las aplicaciones de aprendizaje automático eficaces, en particular con el rápido crecimiento de los datos generados por la Internet de las cosas (IoT). Los métodos tradicionales de limpieza de datos suelen basarse en modelos de contexto creados manualmente que incorporan conocimientos específicos del dominio para extraer dependencias funcionales. LLMClean automatiza este proceso de generación de modelos, implementando LLM para analizar conjuntos de datos, extraer estructuras relevantes y aplicar técnicas de limpieza sensibles al contexto. El estudio demuestra la eficacia de LLMClean para identificar y corregir errores de datos en conjuntos de datos de IoT, atención médica e Industria 4.0, logrando un rendimiento comparable al de los modelos generados por humanos, pero con una eficiencia y una adaptabilidad superiores.

preambulo

Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey

El artículo ofrece un análisis exhaustivo del uso de los LLMs en el modelado de datos tabulares en diversas aplicaciones, incluidas la predicción, la generación de datos y la respuesta a preguntas. A pesar de los recientes avances en los LLMs que mejoran su capacidad para el modelado de datos estructurados, existe una brecha significativa en la literatura que consolida técnicas, métricas y metodologías en este dominio. Los autores pretenden abordar esta brecha mediante el análisis de las tendencias de investigación actuales, la identificación de las fortalezas y limitaciones de los enfoques existentes y la extracción de información para futuras direcciones de investigación. En última instancia, la encuesta tiene como objetivo proporcionar a los profesionales herramientas y conocimientos relevantes para aprovechar eficazmente los LLMs para aplicaciones de datos tabulares, facilitando una mejor comprensión de los desafíos involucrados.

noooooooooooooooooooo

Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning

El artículo presenta un nuevo marco conocido como OCTree (Optimizing Column feature generator with Decision Tree reasoning) que integra modelos de lenguaje grandes (LLMs) para mejorar el proceso de generación automática de características para datos tabulares. Históricamente, los modelos basados ​​en árboles han superado a los enfoques de aprendizaje profundo en tareas tabulares, en gran medida debido a la eficacia de los métodos tradicionales de ingeniería de características. Sin embargo, estos métodos enfrentan desafíos como la dependencia de espacios de búsqueda predefinidos y la incorporación inadecuada de información de experimentos anteriores. OCTree aborda estas limitaciones al permitir que los LLM propongan y refinen iterativamente reglas de generación de características basadas en el razonamiento de árboles de decisión, mejorando así la eficiencia y la eficacia de la extracción de características. Los resultados empíricos demuestran que OCTree supera a los métodos de ingeniería de características automatizados existentes, lo que conduce a un mejor rendimiento de predicción en varios puntos de referencia.

embeding

LLM-Augmented Retrieval: Enhancing Retrieval Models Through Language Models and Doc-Level Embedding

El artículo presenta un nuevo marco para la recuperación de información que emplea grandes modelos de lenguaje (LLM) para mejorar el rendimiento de los sistemas de recuperación existentes. Al integrar los LLM, los autores introducen un método de incrustación a nivel de documento que no depende de ningún modelo y que mejora las técnicas tradicionales de recuperación densa. El estudio enfatiza la eficacia de enriquecer las incrustaciones de documentos con consultas sintéticas y títulos de documentos, mejorando así las tasas de recuperación de los modelos de recuperación ampliamente utilizados, incluidos Contriever y DRAGON. A través de experimentos extensos en conjuntos de datos de referencia como LoTTE y BEIR, el método de aumento propuesto demuestra resultados de vanguardia, lo que indica posibles implicaciones para sistemas de recuperación más robustos y efectivos.

Recuperación d info

ENHANCING KNOWLEDGE RETRIEVAL WITH IN-CONTEXT LEARNING AND SEMANTIC SEARCH THROUGH GENERATIVE AI

Este artículo presenta un enfoque innovador para mejorar la recuperación de conocimiento mediante la integración de la IA generativa, el aprendizaje en contexto y la búsqueda semántica. Los autores identifican desafíos significativos con los sistemas de recuperación tradicionales que se basan en modelos de lenguaje extensos (LLM) de propósito general, en particular a la hora de proporcionar respuestas precisas a consultas específicas de un dominio. Para abordar estos problemas, proponen un modelo novedoso, Generative Text Retrieval (GTR), que recupera información de manera eficiente tanto de datos estructurados como no estructurados sin requerir ajustes costosos. Los resultados experimentales muestran que GTR logra una precisión y eficiencia excepcionales, superando a los modelos existentes y presentando potencial para democratizar el acceso a herramientas de IA avanzadas para la recuperación de información y la consulta de bases de datos.

metricas

Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions∗

El artículo explora la evaluación de los modelos de lenguaje grandes (LLM) dentro del procesamiento del lenguaje natural (NLP). El estudio enfatiza la importancia de las métricas para evaluar el desempeño de los LLM, en particular a medida que surgen diversos modelos en el ámbito académico y la industria. Las conclusiones clave resaltan la necesidad de una comprensión integral de las métricas de evaluación y sus interpretaciones matemáticas y estadísticas. Con un enfoque específico en los LLM biomédicos, los autores presentan comparaciones de varias métricas y ofrecen orientación práctica para los investigadores en la selección de métodos de evaluación apropiados en medio de desafíos significativos, incluidas las limitaciones estadísticas y los estándares de oro imperfectos.