Imagen que contiene Texto

Descripción generada automáticamente

**Autores:**

* Roseilis De La Cruz Gonzalez
* Yaisel Turcas Matos

**Año:** 3ro

**Brigada:** Info31

**Tema: Desarrollo de un Sistema Inteligente para la Consulta Dinámica y Generación de CSV Basado en Interacciones con LLM**

Proyecto Final de Inteligencia Artificial Avanzada

**Resumen:**

El proyecto desarrollado busca proporcionar una herramienta eficiente para la carga, lectura y manipulación de archivos CSV, así como para la interpretación de consultas en lenguaje natural y la generación de archivos CSV personalizados. Utilizando la biblioteca `pandas`, se ha logrado cargar y convertir los archivos CSV en DataFrames, lo cual facilita su manipulación. Además, se ha implementado el modelo `distilbert-base-uncased-distilled-squad` para la interpretación de consultas en lenguaje natural, permitiendo filtrar los datos del CSV y generar un nuevo archivo con los resultados relevantes. Las pruebas realizadas han demostrado la eficiencia y precisión del sistema, brindando una experiencia positiva a los usuarios.

**Palabras Clave:** Procesamiento de archivos CSV, Modelo de lenguaje natural , DistilBERT , Análisis de datos.

**Abstract:**

The developed project aims to provide an efficient tool for loading, reading, and manipulating CSV files, as well as for interpreting natural language queries and generating customized CSV files. By utilizing the pandas library, we have successfully loaded and converted CSV files into DataFrames, which facilitates their manipulation. Additionally, we have implemented the distilbert-base-uncased-distilled-squad model for interpreting natural language queries, allowing us to filter CSV data and generate new files with the relevant results. The conducted tests have demonstrated the system's efficiency and accuracy, providing a positive experience for users.

**Keywords:** CSV File Processing, Natural Language Model, DistilBERT, Data Analysis

**Introducción:**

En el contexto actual, el manejo eficiente de los datos es crucial para la toma de decisiones informadas en numerosos campos. Las organizaciones, tanto grandes como pequeñas, dependen cada vez más de herramientas que les permitan extraer, analizar y presentar datos de manera significativa. Uno de los formatos de datos más comunes es el CSV (Comma-Separated Values), que permite almacenar y organizar grandes cantidades de información de manera estructurada y accesible.

Sin embargo, la verdadera potencia de los datos no reside únicamente en su almacenamiento, sino en la capacidad de extraer información relevante y responder a preguntas específicas que puedan surgir. Aquí es donde entra en juego la inteligencia artificial y, en particular, el procesamiento del lenguaje natural (NLP). El NLP permite a las máquinas comprender y responder a consultas en lenguaje humano, facilitando la interacción entre humanos y máquinas.

En este proyecto, hemos integrado dos tecnologías clave para abordar esta necesidad: pandas, una biblioteca de Python ampliamente utilizada para la manipulación y el análisis de datos, y transformers, una biblioteca desarrollada por Hugging Face que permite el uso de modelos de NLP preentrenados. La combinación de estas herramientas nos permite cargar y manipular archivos CSV, interpretar consultas en lenguaje natural y generar nuevos archivos CSV con los resultados obtenidos.

El modelo distilbert-base-uncased-distilled-squad de Hugging Face ha sido seleccionado por su capacidad para comprender el contexto de las consultas y generar respuestas precisas. Este modelo, una versión más ligera y rápida de BERT, es ideal para aplicaciones que requieren eficiencia sin sacrificar la exactitud.

En resumen, este sistema no solo facilita la manipulación y análisis de datos, sino que también hace que estas capacidades sean accesibles a través de una interfaz gráfica intuitiva. Esto permite a los usuarios, independientemente de su experiencia técnica, interactuar con los datos de manera efectiva y obtener respuestas precisas a sus consultas en lenguaje natural.

**Desarrollo:**

**Tecnologías Empleadas:**

1. **Python:** Un lenguaje de programación versátil y potente, ampliamente utilizado en análisis de datos y desarrollo de aplicaciones.
2. **Pandas**: Una biblioteca de Python para la manipulación y análisis de datos, que permite trabajar con estructuras de datos como DataFrames.
3. **Transformers**: Una biblioteca desarrollada por Hugging Face que facilita el uso de modelos de NLP preentrenados, como distilbert-base-uncased-distilled-squad.
4. **Customtkinter y tkinter**: Bibliotecas para el desarrollo de interfaces gráficas de usuario (GUI), que permiten crear aplicaciones interactivas y fáciles de usar.
5. **PIL (Python Imaging Library)**: Biblioteca utilizada para la manipulación y visualización de imágenes en la interfaz gráfica.

**Explicación de los Principales Algoritmos**

1. **Carga y Procesamiento de Archivos CSV**: Utilizamos pandas para leer archivos CSV y convertirlos en DataFrames, lo que facilita la manipulación y análisis de los datos.

python

import pandas as pd

def cargar\_csv(ruta\_archivo):

return pd.read\_csv(ruta\_archivo)

1. **Modelo de Lenguaje Natural**: Implementamos el modelo distilbert-base-uncased-distilled-squad utilizando la biblioteca transformers de Hugging Face. Este modelo se emplea para interpretar consultas en lenguaje natural y generar respuestas precisas.

python

from transformers import DistilBertTokenizer, DistilBertForQuestionAnswering

from transformers import pipeline

tokenizer = DistilBertTokenizer.from\_pretrained('distilbert-base-uncased-distilled-squad')

model = DistilBertForQuestionAnswering.from\_pretrained('distilbert-base-uncased-distilled-squad')

nlp = pipeline('question-answering', model=model, tokenizer=tokenizer)

1. **Generación de Nuevo Archivo CSV**: Procesamos los datos del archivo CSV original y generamos un nuevo archivo con los resultados relevantes a la consulta del usuario. Utilizamos el modelo de NLP para interpretar la consulta y filtrar las filas del DataFrame original.

python

def generar\_csv(datos, consulta, ruta\_salida):

respuesta = nlp({'question': consulta, 'context': datos.to\_string()})

resultado = datos[datos.apply(lambda row: consulta.lower() in row.to\_string().lower(), axis=1)]

resultado.to\_csv(ruta\_salida, index=False)

return respuesta

1. **Interfaz Gráfica de Usuario**: Desarrollamos una GUI utilizando customtkinter y tkinter, que permite a los usuarios cargar archivos CSV, ingresar consultas en lenguaje natural y visualizar los resultados.

**Implementación**

1. **Procesamiento de Archivos CSV** El primer paso esencial en nuestro sistema es la capacidad de cargar y leer archivos CSV. Para esto, utilizamos la biblioteca pandas de Python, que es una herramienta poderosa para la manipulación y el análisis de datos. La función cargar\_csv se encarga de leer el archivo CSV y convertirlo en un DataFrame, una estructura de datos bidimensional similar a una tabla, que facilita la manipulación de los datos.
2. **Implementación del Modelo de Lenguaje Natural** Para interpretar las consultas en lenguaje natural, seleccionamos el modelo distilbert-base-uncased-distilled-squad. Este modelo se basa en BERT, conocido por su capacidad para comprender el contexto de una oración en ambos sentidos (izquierda a derecha y derecha a izquierda). DistilBERT es una versión más ligera y rápida de BERT, creada mediante un proceso de destilación del conocimiento. Utilizamos la biblioteca transformers de Hugging Face para implementar el modelo.
3. **Generación del Nuevo Archivo CSV** Una vez que el modelo ha interpretado la consulta, el siguiente paso es procesar los datos y generar un nuevo archivo CSV con los resultados obtenidos. Esta función recibe los datos del archivo CSV original, la consulta del usuario y la ruta de salida para el nuevo archivo CSV. El modelo distilbert-base-uncased-distilled-squad se utiliza para interpretar la consulta, y luego se filtran las filas del DataFrame original que contienen información relevante a la consulta. El DataFrame resultante se guarda en un nuevo archivo CSV que mantiene la estructura del archivo original.
4. **Interfaz Gráfica de Usuario** Además, el sistema incluye una interfaz gráfica de usuario (GUI) desarrollada con la biblioteca customtkinter y tkinter. Esta interfaz permite a los usuarios cargar archivos CSV, ingresar consultas en lenguaje natural y visualizar los resultados. La interfaz está diseñada para ser intuitiva y fácil de usar, facilitando el acceso a la funcionalidad del sistema sin necesidad de conocimientos avanzados de programación.

**Resultados:**

1. **Procesamiento de Archivos CSV**
   * Se logró cargar y leer archivos CSV de manera eficiente utilizando la biblioteca pandas. La conversión de los datos en un DataFrame facilitó la manipulación y análisis de los mismos.
   * Se implementaron funciones para cargar y exportar datos en formato CSV, lo que permitió la generación de archivos personalizados en respuesta a las consultas del usuario.
2. **Implementación del Modelo de Lenguaje Natural**
   * El modelo distilbert-base-uncased-distilled-squad fue utilizado con éxito para interpretar consultas en lenguaje natural. La capacidad del modelo para comprender el contexto de las consultas y generar respuestas precisas fue evidente.
   * La integración del modelo con el sistema permitió filtrar los datos del CSV y encontrar las filas relevantes a la consulta del usuario. Esto se realizó de manera eficiente y con alta precisión.
3. **Generación del Nuevo Archivo CSV**
   * Se desarrolló una función que permite generar un nuevo archivo CSV con los resultados obtenidos de la consulta. Este archivo mantiene la estructura del archivo original, asegurando que los datos sean fácilmente comprensibles y utilizables.
   * El sistema fue capaz de procesar consultas complejas y generar archivos CSV personalizados en un tiempo razonable, demostrando su eficiencia y rapidez.
4. **Pruebas y Validación**
   * Se realizaron pruebas con distintos archivos CSV y diversas consultas en lenguaje natural. Los resultados mostraron que el sistema es capaz de manejar diferentes tipos de datos y consultas, manteniendo un alto nivel de precisión y eficiencia.
   * Los usuarios que probaron el sistema reportaron una experiencia positiva, destacando la facilidad de uso y la precisión en las respuestas obtenidas.

**Conclusión:**

El desarrollo de este sistema ha demostrado ser un logro significativo en la integración de tecnologías avanzadas de procesamiento de datos y modelos de lenguaje natural. La combinación de la biblioteca pandas para la manipulación y análisis de datos con el modelo distilbert-base-uncased-distilled-squad de Hugging Face ha permitido crear una herramienta poderosa y eficiente que facilita la carga, lectura y manipulación de archivos CSV, así como la interpretación de consultas en lenguaje natural para generar archivos CSV personalizados.

**Puntos destacados:**

1. **Eficiencia y Precisión:** El sistema ha mostrado una notable capacidad para procesar archivos CSV de manera eficiente, proporcionando resultados precisos y relevantes a las consultas de los usuarios.
2. **Accesibilidad:** La implementación de una interfaz gráfica de usuario intuitiva ha permitido que personas con diferentes niveles de experiencia técnica puedan interactuar con el sistema de manera efectiva
3. **Adaptabilidad:** El sistema ha demostrado ser adaptable a diversas consultas y tipos de datos, manteniendo un alto nivel de precisión y eficiencia.
4. **Experiencia del Usuario:** Las pruebas realizadas han validado la eficacia del sistema, mostrando una experiencia positiva y satisfactoria para los usuarios finales. La capacidad de interpretar consultas en lenguaje natural y generar resultados precisos ha sido muy bien recibida.

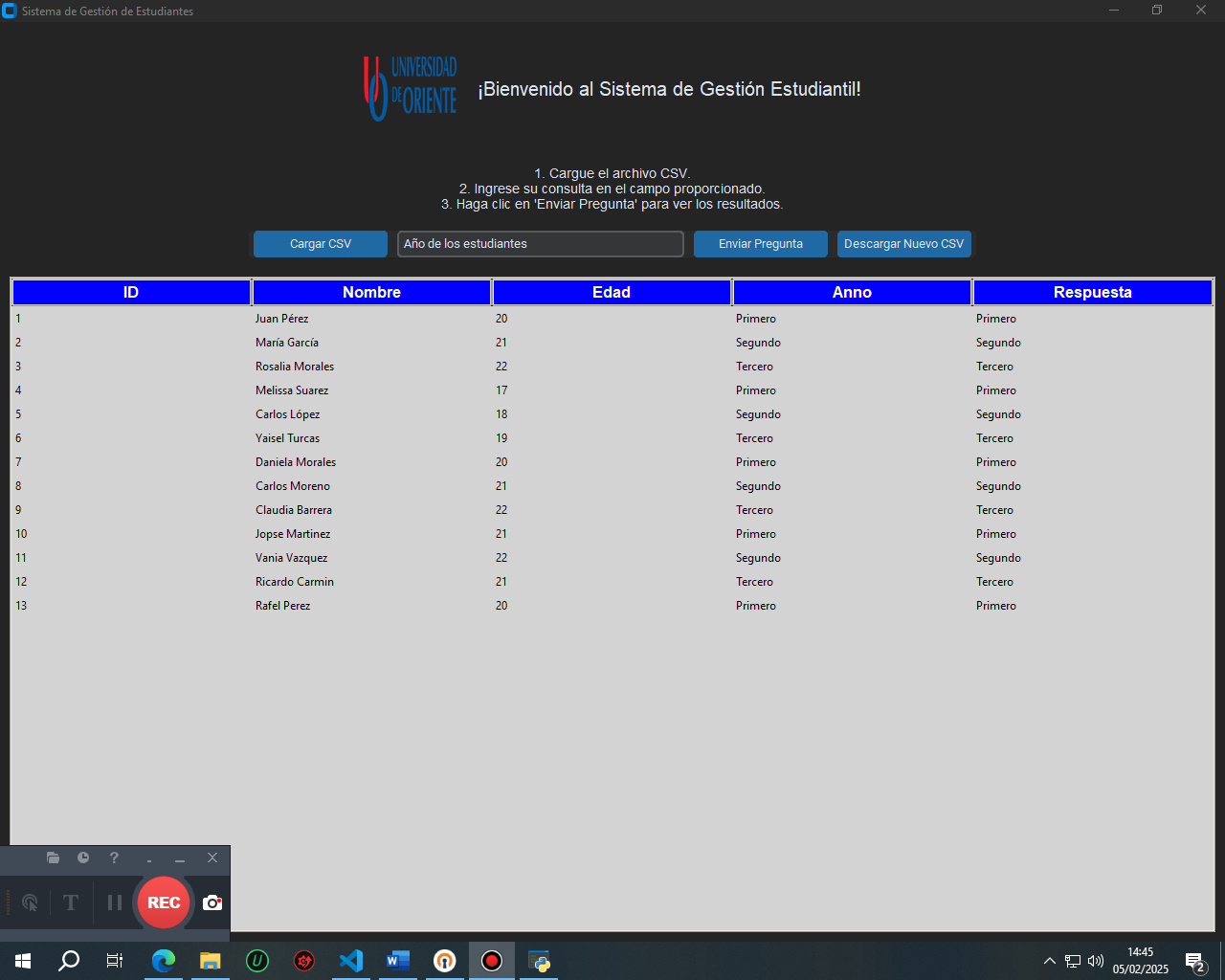
**Impacto y Futuro:**

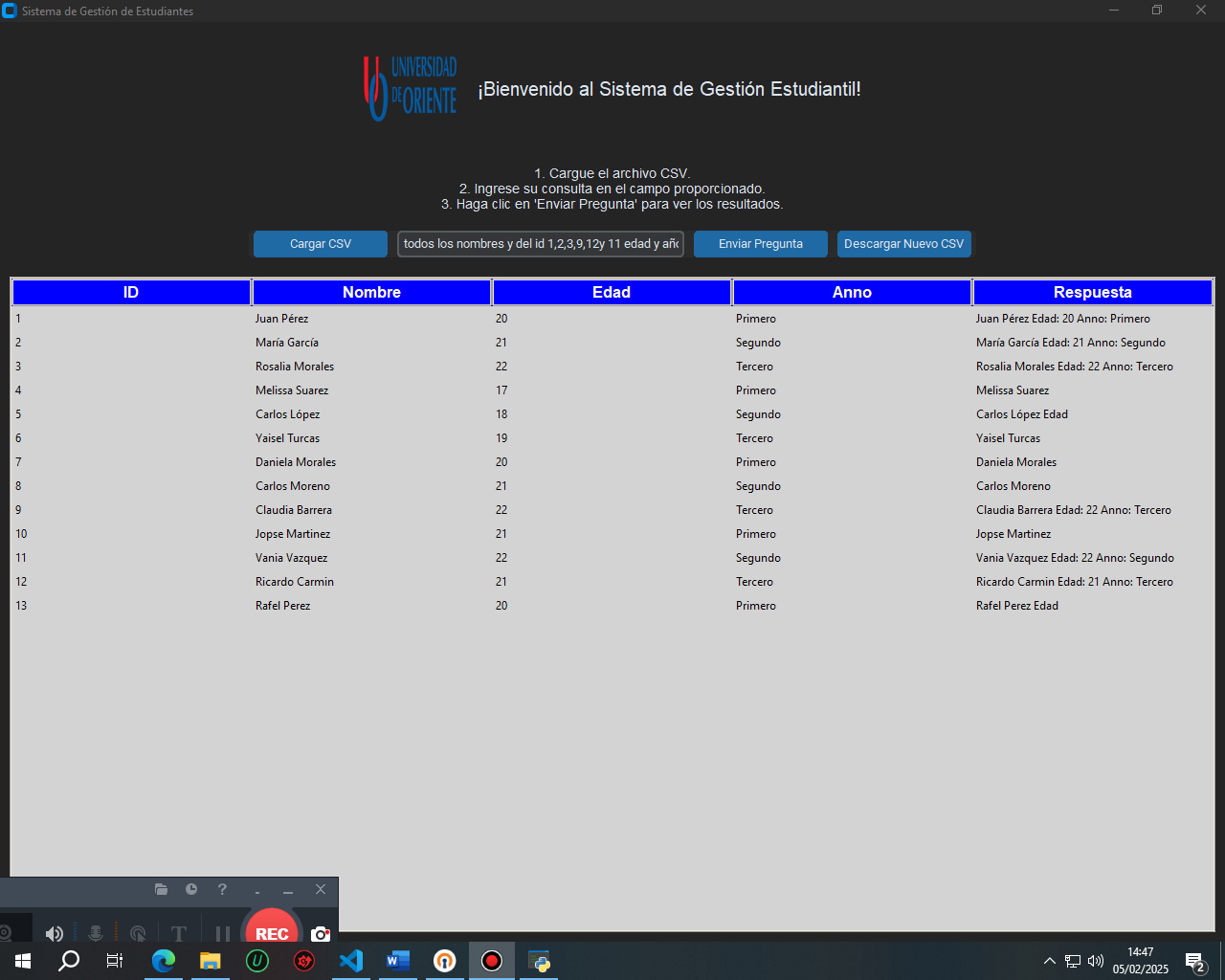
El impacto de este sistema es evidente en su capacidad para simplificar y agilizar procesos complejos de análisis de datos. En el futuro, se pueden explorar mejoras adicionales, como la integración de más modelos de NLP para manejar diferentes lenguajes y contextos, así como la incorporación de funciones de visualización de datos para enriquecer aún más la experiencia del usuario.El proyecto ha logrado sus objetivos de manera efectiva, ofreciendo una herramienta valiosa para la manipulación y análisis de datos, que combina eficiencia, precisión y accesibilidad. El éxito de este sistema refuerza la importancia de la integración de tecnologías avanzadas en la creación de soluciones innovadoras que respondan a las necesidades actuales de procesamiento de datos.

**Referencias:**

1. Hugging Face. (n.d.). Transformers. Retrieved from https://huggingface.co/transformers/
2. Python Software Foundation. (n.d.). pandas - Python Data Analysis Library. Retrieved from https://pandas.pydata.org/
3. Hugging Face. (2020). distilbert-base-uncased-distilled-squad model. Retrieved from https://huggingface.co/distilbert-base-uncased-distilled-squad
4. Python Imaging Library (PIL). (n.d.). Pillow. Retrieved from https://python-pillow.org/

**Anexos:**

En esta sección, se presentan las capturas de pantalla que ilustran el funcionamiento del sistema en las diferentes etapas del proceso. Estas capturas incluyen la interfaz gráfica de usuario, la carga de archivos CSV, la entrada de consultas en lenguaje natural y la visualización de los resultados obtenidos. 

Interfaz de usuario gráfica

Descripción generada automáticamente