Análisis de Big Data: Informes AGN

Descripción del Proyecto

Este proyecto consiste en una serie de cuadernos Jupyter diseñados para llevar a cabo un análisis completo de big data, desde la obtención de datos hasta el análisis, visualización, informe y la creación de un prototipo de chatbot. El objetivo es proporcionar un flujo de trabajo integral que pueda ser utilizado para extraer insights valiosos de grandes conjuntos de datos.

Objetivo del Proyecto

El objetivo principal de este proyecto es analizar el funcionamiento y los informes de la Auditoría General de la Nación (AGN) de Argentina. Esto implica la obtención de información disponible en su página web, así como el análisis de las resoluciones y los informes de auditoría generados por esta entidad.

Tabla de Contenidos

Instalación y Configuración

Requisitos

Antes de comenzar, asegúrese de tener instalado Jupyter Notebook y todas las bibliotecas necesarias para ejecutar los cuadernos. Puede instalar Jupyter Notebook siguiendo las instrucciones en la página oficial de Jupyter.

Instrucciones de Instalación

Clone el repositorio a su máquina local y ábralo en Jupyter Notebook:

```
git clone https://github.com/tu_usuario/tu_repositorio.git
cd tu_repositorio
jupyter notebook
```

Uso

A continuación, se proporciona una breve descripción de los cuadernos Jupyter incluidos en este proyecto:

- **00_Análisis_API.ipynb:** Realiza un análisis de diferentes APIs para determinar la más adecuada para la obtención de datos.

- **01_DB_Extracción_de_datos_API.ipynb**: Extrae datos utilizando la API seleccionada y los almacena en una base de datos SQLite.
- **02_Pipeline_datos.ipynb:** Construye un pipeline de datos para procesar y limpiar los datos extraídos.
- **03_Data_Analytics.ipynb: (Sin terminar)** Realiza análisis de los datos obtenidos, incluyendo análisis exploratorio y visualizaciones.
- **04_Chatbot.ipynb: (Sin terminar)** Desarrolla un prototipo de chatbot que puede responder preguntas basadas en la base de datos.

Licencia

Este proyecto es un portafolio personal y está destinado a ser utilizado como una demostración de habilidades. No está licenciado para su uso en otros proyectos.

Contacto

Para cualquier pregunta o comentario, no dude en ponerse en contacto a través de:

- Correo Electrónico: miguelhdg@egmail.com
- LinkedIn: www.linkedin.com/in/miguelhdg

Estado del Proyecto

- **00_Análisis_API.ipynb:** Completo.
- **01_DB_Extracción_de_datos_API.ipynb:** Completo.
- **02_Pipeline_datos.ipynb:** Completo.
- **03_Data_Analytics.ipynb:** En progreso.
- **04_Chatbot.ipynb:** En progreso.

Uso de Datos de la Página AGN

Este proyecto utiliza datos obtenidos de la página de la Administración General de la Nación (AGN). Se han respetado las directrices proporcionadas en el archivo /robots.txt del sitio, permitiendo el acceso y análisis de su contenido.

Es importante destacar que el uso de estos datos es únicamente con fines educativos y de demostración para mi portafolio personal, y no tiene fines comerciales. Se han seguido todas las políticas de uso y privacidad del sitio y se proporcionan los debidos créditos a AGN como la fuente de los datos.

Este proyecto no está licenciado para su uso en otros proyectos, y cualquier persona interesada en utilizar los datos de AGN debe dirigirse directamente al sitio y cumplir con sus políticas y licencias correspondientes.

Detalles de los Cuadernos

A continuación, se describen los cuadernos Jupyter incluidos en el proyecto:

00_Análisis_API.ipynb

Objetivo:

Este cuaderno está dedicado al análisis detallado de la API de la Auditoría General de la Nación (AGN), enfocándose en las auditorías disponibles en su sitio web. El objetivo principal es entender la estructura y funcionalidad de la API para facilitar la extracción de informes de auditoría.

Librerías Utilizadas:

- **requests:** Para realizar solicitudes HTTP a la API.
- **json:** Para codificar y decodificar datos en formato JSON.

Contenido y Estructura de la API:

Se exploran diversas URLs asociadas a la API de AGN, identificando que se utiliza el formato JSON:API. Se destaca la importancia de tener en cuenta aspectos como la limitación en solicitudes, paginación, profundidad de los recursos, campos específicos, errores de validación y actualizaciones de la API.

URLs Analizadas:

Se realiza una selección crítica de 10 URLs esenciales para el análisis y posterior creación de la base de datos, proporcionando información relevante sobre informes, infografías, archivos y categorías descriptivas utilizadas como filtros en el sitio web de AGN.

Componentes Clave de la API:

Se analizan los componentes "attributes" y "relationships" dentro de la estructura JSON de las URLs, proporcionando detalles sobre informes, archivos y categorías descriptivas.

Consideraciones Especiales:

- Se menciona la necesidad de gestionar múltiples solicitudes para obtener todos los datos debido a la paginación.
- Se destaca la importancia de tener en cuenta posibles limitaciones y errores al interactuar con la API.

01_DB_Extracción_de_datos_API.ipynb

Objetivo:

Este cuaderno se centra en la extracción de datos de la API de la Auditoría General de la Nación (AGN) y su posterior almacenamiento en una base de datos SQLite. Forma parte de un proyecto más amplio que abarca desde la obtención de datos hasta la creación de un prototipo de chatbot.

Librerías Utilizadas:

- **requests:** Para realizar solicitudes HTTP a la API.
- **pandas:** Para la manipulación y análisis de datos.
- **sqlite3:** Para trabajar con la base de datos SQLite.
- Otras librerías específicas para procesamiento de archivos y texto.

Descripción del Contenido y Estructura:

Este cuaderno aborda los siguientes aspectos clave:

- Función para Suprimir Advertencias: Se presenta una función que permite suprimir advertencias, útil al hacer solicitudes a la API.
- Función para Hacer Solicitudes a la API: Define una función para realizar solicitudes HTTP a la API de AGN, con manejo de intentos y tiempo de espera.
- Función para Estilizar y Mostrar DataFrames: Permite mostrar DataFrames de manera estilizada en entornos Jupyter.
- Función para Guardar DataFrames en SQLite: Presenta una función para guardar DataFrames en la base de datos SQLite, reemplazando datos existentes en las tablas específicas.
- Base de Datos SQLite: Describe la elección de SQLite como la base de datos para este proyecto y destaca sus ventajas.
- Tablas en la Base de Datos: Menciona las tablas creadas en la base de datos, como "detalle_Informes," "codigo_archivos," "codigo_tid," y "codigo_infografias," junto con sus respectivas descripciones.
- Obtención de Tablas en la Base de Datos: Proporciona funciones para obtener información sobre las tablas y columnas en la base de datos.
- Visualización de Datos en la Base de Datos: Presenta una función para mostrar datos de las tablas en la base de datos, ya sea de manera aleatoria o basada en el "codigo_nid."

Este cuaderno sienta las bases para la extracción de datos desde la API de AGN y su almacenamiento en una base de datos SQLite, facilitando el análisis y la creación del chatbot basado en estos datos.

02_Pipeline_datos.ipynb

Objetivo:

El cuaderno Jupyter 02_Pipeline_datos.ipynb se centra en el análisis y modelado de datos relacionados con las resoluciones e informes de la Auditoría General de la Nación (AGN). El objetivo principal de este cuaderno es automatizar y estandarizar el proceso de adquisición, limpieza, transformación y modelado de estos datos con el propósito de obtener insights valiosos y, posiblemente, desarrollar un modelo predictivo o clasificatorio.

Librerías Utilizadas:

- **sqlite3:** Para trabajar con bases de datos SQLite y acceder a los datos almacenados en la base de datos "informes_agn.db".
- **IPython.display:** Se emplea esta herramienta para mostrar y formatear la salida en entornos Jupyter.
- **pandas:** La librería pandas se utiliza para la manipulación y análisis de datos, especialmente para estructuras de datos tabulares.

Descripción del Dataset:

El dataset utilizado principalmente proviene de la tabla "detalle_Informes" de la base de datos "informes_agn.db". Además, se complementó con datos de otras tablas de la base de datos, lo que resultó en un conjunto de datos más completo y adecuado para análisis detallados.

Base de Datos:

- Nombre: **"informes_agn.db"**

Tablas Utilizadas:

1. **detalle_Informes:** Esta tabla es la fuente principal de la mayoría de los datos utilizados en el análisis.
2. **codigo_archivos:** Contiene información sobre los archivos relacionados.
3. **codigo_tid:** Ofrece códigos y descripciones que se usaron para enriquecer la información.
4. **codigo_infografias:** Proporciona información sobre las infografías relacionadas.
5. **resolucion_informe_texto: Contiene el texto de las resoluciones e informes.**

Dataset - Columnas Principales:

El dataset contiene diversas columnas que incluyen información relevante sobre los informes de la AGN, como el título, estado, gerencia responsable, tipo de auditoría, organismo auditado, actuación, año de aprobación, ventanas de trabajo y más. Además, se realizaron transformaciones y sustituciones de códigos por descripciones para enriquecer las palabras clave del informe.

Proceso de Análisis y Modelado:

El cuaderno Jupyter detalla paso a paso el proceso de análisis y modelado de datos, incluyendo la carga de datos desde la base de datos SQLite, la transformación de columnas, la detección de situaciones en el texto y la creación de un nuevo dataset denominado "df_DA_informes".

Guardar Resultados:

El resultado final del análisis se guarda en un archivo CSV llamado "df_DA_informes.csv", que contiene el dataset procesado y listo para su posterior análisis o modelado.

Este cuaderno es una parte fundamental del proyecto global de Análisis de Big Data y Prototipo de Chatbot, y contribuye significativamente al procesamiento y preparación de los datos para su uso en etapas posteriores del proyecto.

Nota Importante:

Este archivo "léeme" está actualmente en proceso de mejora en lo que respecta a su redacción y formato. Sin embargo, es importante destacar que dentro de cada cuaderno (notebook), los códigos están debidamente explicados y los procesos están justificados en detalle.

Además, cabe señalar que el análisis de la información se realiza en castellano para su utilización en Argentina, por lo que las descripciones y los datos de los cuadernos se mantienen en ese idioma.

Agradecemos su comprensión mientras trabajamos en la mejora continua de este archivo "léeme" para brindar una experiencia más clara y efectiva a nuestros usuarios. Si tiene alguna pregunta o necesita información adicional, no dude en ponerse en contacto.