Big Data Analysis: AGN Reports

Project Description

This project consists of a series of Jupyter notebooks designed to perform comprehensive big data analysis, from data acquisition to analysis, visualization, reporting, and the creation of a chatbot prototype. The aim is to provide a comprehensive workflow that can be used to extract valuable insights from large datasets.

Project Objective

The main objective of this project is to analyze the operations and reports of the Auditoría General de la Nación (AGN) of Argentina. This involves obtaining information available on its website, as well as analyzing the resolutions and audit reports generated by this entity.

Table of Contents

Installation and Configuration

Requirements

Before getting started, make sure you have Jupyter Notebook installed along with all the necessary libraries to run the notebooks. You can install Jupyter Notebook by following the instructions on the official Jupyter page.

Installation Instructions

Clone the repository to your local machine and open it in Jupyter Notebook:

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

Usage

Here is a brief description of the Jupyter notebooks included in this project:

- **00_Análisis_API.ipynb**: Conducts an analysis of different APIs to determine the most suitable for data retrieval.

- **01_DB_Extracción_de_datos_API.ipynb**: Extracts data using the selected API and stores it in a SQLite database.
- **02_Pipeline_datos.ipynb**: Builds a data pipeline to process and clean the extracted data.
- **03_Data_Analytics.ipynb** (Incomplete): Performs data analysis, including exploratory analysis and visualizations.
- **04_Chatbot.ipynb** (Incomplete): Develops a chatbot prototype that can answer questions based on the database.

License

This project is a personal portfolio and is intended for use as a skills demonstration. It is not licensed for use in other projects.

Contact

For any questions or comments, please feel free to contact me at:

- Email: miguelhdg@email.com
- LinkedIn: www.linkedin.com/in/miguelhdg

Project Status

- **00_Análisis_API.ipynb**: Complete.
- **01_DB_Extracción_de_datos_API.ipynb**: Complete.
- **02_Pipeline_datos.ipynb**: Complete.
- **03_Data_Analytics.ipynb**: In progress.
- **04_Chatbot.ipynb**: In progress.

Use of AGN Website Data

This project uses data obtained from the website of the Auditoría General de la Nación (AGN). The guidelines provided in the /robots.txt file of the site have been respected, allowing access and analysis of its content.

It is important to emphasize that the use of this data is solely for educational and demonstration purposes for my personal portfolio and has no commercial purposes. All site usage and privacy policies have been followed, and proper credit is given to AGN as the source of the data.

This project is not licensed for use in other projects, and anyone interested in using AGN data should go directly to the site and comply with its policies and licenses.

Notebook Details

Here are descriptions of the Jupyter notebooks included in the project:

00_Análisis_API.ipynb

Objective:

This notebook is dedicated to a detailed analysis of the Auditoría General de la Nación (AGN) API, focusing on the audits available on its website. The main objective is to understand the structure and functionality of the API to facilitate the extraction of audit reports.

Used Libraries:

- **requests**: For making HTTP requests to the API.
- **json**: For encoding and decoding data in JSON format.

API Content and Structure:

Various URLs associated with the AGN API are explored, identifying the use of the JSON:API format. The importance of considering aspects such as request limitations, pagination, resource depth, specific fields, validation errors, and API updates is highlighted.

Analyzed URLs:

A critical selection of 10 essential URLs for analysis and subsequent database creation is made, providing relevant information about reports, infographics, files, and descriptive categories used as filters on the AGN website.

Key API Components:

The "attributes" and "relationships" components within the JSON structure of the URLs are analyzed, providing details about reports, files, and descriptive categories.

Special Considerations:

- The need to manage multiple requests to obtain all data due to pagination is mentioned.
- The importance of considering possible limitations and errors when interacting with the API is highlighted.

01_DB_Extracción_de_datos_API.ipynb

Objective:

This notebook focuses on data extraction from the Auditoría General de la Nación (AGN) API and its subsequent storage in a SQLite database. It is part of a larger project that encompasses data acquisition to the creation of a chatbot prototype.

Used Libraries:

- **requests**: For making HTTP requests to the API.
- **pandas**: For data manipulation and analysis.
- **sqlite3**: For working with the SQLite database.
- Other specific libraries for file and text processing.

Description of Content and Structure:

This notebook addresses the following key aspects:

- Function to Suppress Warnings: A function is presented to suppress warnings, which is useful when making requests to the API.
- Function to Make API Requests: Defines a function for making HTTP requests to the AGN API, with handling of retries and timeouts.
- Function to Style and Display DataFrames: Allows displaying DataFrames in a styled manner in Jupyter environments.
- Function to Save DataFrames to SQLite: Presents a function for saving DataFrames to the SQLite database, replacing existing data in specific tables.
- SQLite Database: Describes the choice of SQLite as the database for this project and highlights its advantages.
- Tables in the Database: Mentions the tables created in the database, such as "detalle_Informes," "codigo_archivos," "codigo_tid," and "codigo_infografias," along with their respective descriptions.
- Obtaining Tables in the Database: Provides functions to obtain information about tables and columns in the database.
- Data Visualization in the Database: Presents a function to display data from tables in the database, either randomly or based on the "codigo_nid."

This notebook lays the foundation for data extraction from the AGN API and its storage in a SQLite database, facilitating analysis and the creation of the chatbot based on this data.

02_Pipeline_datos.ipynb

Objective:

The Jupyter notebook 02_Pipeline_datos.ipynb focuses on the analysis and modeling of data related to the resolutions and reports of the Auditoría General de la Nación (AGN). The main objective of this notebook is to automate and standardize the process of data acquisition, cleaning, transformation, and modeling to obtain valuable insights and possibly develop a predictive or classification model.

Used Libraries:

- **sqlite3**: For working with SQLite databases and accessing data stored in the "informes_agn.db" database.
- **IPython.display**: This tool is used to display and format output in Jupyter environments.
- **pandas**: The pandas library is used for data manipulation and analysis, especially for tabular data structures.

Description of the Dataset:

The dataset primarily used comes from the "detalle_Informes" table of the "informes_agn.db" database. Additionally, it was supplemented with data from other tables in the database, resulting in a more comprehensive dataset suitable for detailed analysis.

Database:

- Name: "informes_agn.db"

Used Tables:

1. **detalle_Informes**: This table is the primary source of most of the data used in the analysis.
2. **codigo_archivos**: Contains information about related files.
3. **codigo_tid**: Provides codes and descriptions used to enrich the information.
4. **codigo_infografias**: Provides information about related infographics.
5. **resolucion_informe_texto**: Contains the text of resolutions and reports.

Dataset - Main Columns:

The dataset contains various columns that include relevant information about AGN reports, such as title, status, responsible management, audit type, audited entity, performance, year of approval, work windows, and more. Additionally, code transformations and substitutions for descriptions were made to enrich the report's keywords.

Analysis and Modeling Process:

The Jupyter notebook details step by step the process of data analysis and modeling, including data loading from the SQLite database, column transformation, text situation detection, and the creation of a new dataset called "df_DA_informes."

Save Results:

The final result of the analysis is saved in a CSV file called "df_DA_informes.csv," which contains the processed dataset ready for further analysis or modeling.

This notebook is a fundamental part of the larger Big Data Analysis and Chatbot Prototype project and significantly contributes to data processing and preparation for use in later stages of the project.

Important Note: This "readme" file is currently in the process of improvement regarding its wording and formatting. However, it is essential to note that within each notebook, the code is properly explained, and the processes are detailed.

Furthermore, it should be noted that the information is analyzed in Spanish for use in Argentina, so the descriptions and data in the notebooks are maintained in that language.

We appreciate your understanding as we work on continuous improvement of this "readme" file to provide a clearer and more effective experience for our users. If you have any questions or need additional information, please do not hesitate to contact us.