GenAI Data Scientist Coding Assignment: EDA Design for RAG

Task Definition:

Our customer, a technology company, aims to create an internal assistant capable of handling complex queries related to their various policies. The assistant should function similarly to systems that interpret and respond based on provided policy documents. Your task is to develop an Exploratory Data Analysis (EDA) framework that will guide the development of a Retrieval-Augmented Generation (RAG) model for this purpose. This model will need to retrieve and utilize relevant policy information efficiently to answer queries accurately and contextually.

Initial Project Phase and Data Considerations:

Given that we are in the initial phase of this imaginary project, you have received an initial dataset comprising examples of the company's policies. However, to enhance your analysis and the eventual utility of the RAG model, you should also consider what additional data might be beneficial.

Assignment Details:

Dataset: You will be provided with a dataset containing exemplary policy documents. These documents represent the type of content the internal assistant will need to understand and retrieve.

Task Requirements:

- **Framework Development**: Construct a detailed EDA roadmap that guides your analytical strategy. Presume the availability of additional information as needed, and clearly identify which extra data could enhance your analysis. This roadmap should delineate each phase of your EDA effort, outlining the specific concepts and tools employed at every step and providing justifications for each decision made. The framework should encourage exploratory analysis by suggesting a variety of methods to analyze the data, allowing for flexibility in approach and technique. Insights extracted by following this framework should be instrumental in designing the RAG in subsequent phases of the project, ensuring that the analysis directly informs and enhances the development process. During the coding phase, while using the exemplary data, you might choose to implement only a selection of the methods suggested in the framework, focusing on those that provide the most immediate and relevant insights for the project at hand. You can use any open source library.
- **Quantitative Analysis and Visualization**:
    - Perform statistical summaries and correlation analyses to identify patterns and metrics critical to the development of the RAG model.
    - Develop visualizations like document distribution graphs, heatmaps, and word clouds to elucidate the document characteristics and predominant themes.
- **Insightful Observations**: Provide insights into potential challenges and opportunities within the dataset that could affect the RAG model's performance. Predict and discuss possible complications with unseen data and suggest how the model might address these challenges.

- **Discussion on Framework Adaptability**: Your report should include a discussion on the EDA methodologies, challenges, and opportunities specific to RAG systems for internal policy assistants. Explain how the EDA findings can influence the design and effectiveness of the RAG solution. Also, consider how the framework might adapt to include external resources like GPT models if access to such APIs becomes available.

Deliverables:

- A Jupyter Notebook consistent with the EDA framework.
- A presentation including the EDA framework for RAGs, EDA of the imaginary project (walking through code), results, recommendations, and conclusion.

Submission Guidelines:

- Submit all components as a ZIP file.
- Ensure your code is well-commented, and your report is clear and professionally structured.