Documentation for RAG Model for QA Bot on P&L Data

Model Architecture

The Retrieval-Augmented Generation (RAG) model developed for the QA bot comprises the following components:

1. Data Extraction

- **Tool Used**: Camelot
- **Process**: Extract P&L tables from PDF documents.
  - Pages of interest are specified, and Camelot's `stream` flavor is used for parsing.
  - Extracted tables are combined into a unified DataFrame.

2. Preprocessing

- **Column Naming**: Duplicate column names in the extracted tables are made unique using a helper function.
- **Header Adjustment**: The first row of the table is set as the header, and redundant rows are removed.
- **Index Reset**: Ensures the DataFrame is clean and ready for further processing.

3. Embedding Generation

- **Tool Used**: OpenAI's `text-embedding-ada-002` model.
- **Process**:
  - Each row of the table is converted into a structured text string of key-value pairs.
  - OpenAI's API generates embeddings for these text strings.

4. Embedding Storage

- **Tool Used**: Pinecone
- **Process**:
  - A vector database stores embeddings for efficient retrieval.
  - Index is created with cosine similarity as the metric.
  - Each embedding is stored with metadata (original text) for context retrieval.

5. Query Processing and Response Generation

- **Embedding Query**:
  - OpenAI generates an embedding for the user's query.
  - Pinecone searches for the top-k similar embeddings.
- **Response Generation**:
  - Retrieved text data is passed to OpenAI's `gpt-3.5-turbo` for response generation.
  - The model generates a detailed and contextually accurate response based on retrieved data.

Approach to Data Extraction and Preprocessing

Data Extraction

- Extracted tables are parsed using Camelot's `read_pdf` function.
- Multiple pages can be processed sequentially, combining tables into one unified dataset.

Preprocessing

- Ensures that all columns in the DataFrame are unique.
- Converts the DataFrame into a clean and structured format for embedding generation.
- Handles edge cases where tables might be empty or improperly extracted.

Generative Response Creation

Workflow

1. User query is embedded using OpenAI's embedding model.
2. Pinecone retrieves the top-k similar embeddings based on cosine similarity.
3. Retrieved embeddings are converted back into a context string.
4. A ChatCompletion API call uses the context and query to generate a coherent response.

Challenges and Solutions

1. Challenge: Inconsistent Table Extraction

Solution: Applied preprocessing techniques to handle misaligned or duplicate headers.

2. Challenge: Embedding Size Management

Solution: Limited the number of rows processed in a single batch to avoid API limitations.

3. Challenge: Ensuring Query Relevance

Solution: Used Pinecone's ranking to focus on the top-k matches, ensuring only the most relevant data is used for generating responses.

4. Challenge: Handling Empty Tables

Solution: Incorporated validation checks to skip pages without extractable tables.

Example Queries and Outputs

Query 1: "What is the total value of non-current assets as of March 31, 2024?"

Response: To calculate the total value of non-current assets as of March 31, 2024, you need to sum the total non-current assets and other non-current assets for that particular date.

Total non-current assets as of March 31, 2024: $48,382 (from .1) + $54,935 (from .2) = $103,317

Other non-current assets as of March 31, 2024: $2,121 (from 2.9.1) + $2,318 (from 2.9.2) = $4,439

Therefore, the total value of non-current assets as of March 31, 2024 is: $103,317 (Total non-current assets) + $4,439 (Other non-current assets) = $107,756

Query 2: "What are the non-current assets for March 31, 2023?"

Response: The non-current assets for March 31, 2023 are $54,935 million.

Query 3: "What is the difference in total non-current assets between March 31, 2023, and March 31, 2024?"

Response: To find the difference in total non-current assets between March 31, 2023, and March 31, 2024, you can subtract the total non-current assets for March 31, 2023, from the total non-current assets for March 31, 2024.

Total non-current assets on March 31, 2024 = $54,935 Total non-current assets on March 31, 2023 = $48,382

Difference = $54,935 - $48,382 = $6,553

Therefore, the difference in total non-current assets between March 31, 2023, and March 31, 2024 is $6,553.