Multi-Modal Network Support RAG System - Cisco & Mikrotik

Emmanuel Chibua, NU ID: 002799484
Gunjit Arora, NU ID: 002679282
Abdulafeez Abobarin, NU ID: 002922336

August, 2024

Contents

Chapter 1

Introduction

The rapid evolution of telecommunication technologies has led to a complex landscape where network administrators are required to manage, troubleshoot, and optimize diverse networking equipment, particularly from leading vendors like Cisco and Mikrotik. The need for efficient, accurate, and timely support for these systems has never been greater. To address this, our project focuses on developing a Retrieval-Augmented Generation (RAG) system using advanced Large Language Models (LLMs) and the LangChain framework.

Our system aims to assist network professionals by providing expert-level answers to queries related to Cisco and Mikrotik devices. By leveraging the capabilities of LLMs such as GPT-4 and LLaMA, combined with an extensive knowledge base extracted from relevant documents, this system promises to revolutionize how network support is delivered.

Chapter 2

Project Objectives

The primary objectives of this project include:

- **Enhanced Network Support:** Provide accurate and contextualized answers to complex network-related queries.
- **Leveraging Advanced AI:** Utilize state-of-the-art generative AI models to generate human-like responses.
- **Document and Image Parsing:** Extract valuable information from both PDFs and images, making it accessible for the RAG system.
- **Seamless User Interaction:** Offer an intuitive interface for network professionals to interact with the system and retrieve expert advice.

Chapter 3

Use Case Explanation

The use case revolves around a system designed to assist network engineers and administrators who manage Cisco and Mikrotik devices. In their day-to-day operations, they may face complex issues that require detailed troubleshooting and precise configurations. Our system is designed to provide on-demand support by answering their questions based on context extracted from documentation.

3.1 Generative AI & RAG

Generative AI: We utilize GPT-4 and LLaMA models, which are at the forefront of generative AI, capable of understanding context and generating coherent and contextually relevant responses.

Retrieval-Augmented Generation (RAG): The system implements RAG by combining the strengths of retrieval (searching through a vast knowledge base) with generative capabilities (creating human-like responses). The workflow involves:

1. **Document Parsing:** Using LangChain's PyPDFLoader, the system extracts and splits PDF documents into manageable pages.
2. **Vector Store Creation:** Text chunks from these documents are embedded into a vector space using OpenAI embeddings and stored in a FAISS vector store for efficient similarity search.
3. **Conversational Chain:** Depending on the selected model (GPT-4 or LLaMA), the system uses a custom prompt template to generate responses based on the retrieved context.

3.2 LangChain Framework

LangChain plays a pivotal role in this project by enabling seamless integration between document parsing, text processing, vectorization, and interaction with LLMs. It allows for efficient handling of large text chunks and supports the development of a sophisticated question-answering system.

Chapter 4

Key Features and Functionalities

The RAG system is packed with features designed to cater to the needs of network professionals:

- **PDF Document Parsing:** Extracts and processes text from PDF documents, splitting them into smaller, more manageable sections for efficient querying.
- **Image to Text Conversion:** Utilizes Optical Character Recognition (OCR) via pytesseract to extract text from images, making even non-digital documents accessible.
- **Vector Store Management:** Employs FAISS for efficient storage and retrieval of document vectors, allowing rapid access to relevant information.
- **Question-Answering Module:** Supports conversational queries, where users can ask detailed questions about network issues and receive expert-level responses based on the extracted knowledge.
- **Model Selection:** Allows users to choose between different LLMs (GPT-4 or LLaMA), depending on their preference or specific use case.
- **User-Friendly Interface:** Built with Streamlit, providing an intuitive and interactive platform for users to upload documents, ask questions, and retrieve answers.

Chapter 5

Challenges and Solutions

During the development of the RAG system, several challenges were encountered:

- **Document Parsing Accuracy:** Some PDF documents were not easily parsed due to formatting issues. This was mitigated by implementing error-handling mechanisms and providing feedback to the user.
- **Efficient Vectorization:** Managing large amounts of text data while ensuring quick retrieval posed a challenge. This was addressed by optimizing the vector store using FAISS and fine-tuning the embedding process.
- **Latency in Response Generation:** Given the computational complexity of LLMs, generating responses in real-time could lead to delays. By optimizing the query process and allowing users to select models based on their performance, this issue was minimized.
- **Model Integration:** Integrating different LLMs required careful consideration of their respective APIs and capabilities. This was successfully handled by creating flexible prompts and chains that could adapt to the chosen model.

Chapter 6

Conclusion and Future Scope

The Network Support RAG System represents a significant advancement in the use of AI for network management and support. By combining cutting-edge generative AI with retrieval-augmented techniques, the system offers an unprecedented level of assistance to network professionals.

6.1 Future Scope

- **Expanded Knowledge Base:** Integrate additional documents and resources to cover a wider range of networking topics.
- **Multilingual Support:** Enable the system to process and respond to queries in multiple languages, making it accessible to a global audience.
- **Real-Time Collaboration:** Incorporate features that allow multiple users to interact with the system simultaneously, facilitating collaborative troubleshooting.
- **Fine-tuned Model:** As this tool grows into a full fledged product, we hope to incorporate a fine-tuned LLM with curated dataset to better enhance user experience and service delivery.
- **Enhanced Security:** Implement robust security measures to protect sensitive data and ensure compliance with industry standards.

Chapter 7

Appendix

This report includes contributions from the following team members:

- **Gunjit Arora, NU ID: 002679282**
- **Emmanuel Chibua, NU ID: 002799484**
- **Abdulafeez Abobarin, NU ID: 002922336**

We have included a detailed breakdown of each component of the project and its implementation. The system is not only a tool but a testament to the power of AI in revolutionizing network management.