CONVERT ENTERPRISE PDFs INTO SEARCHABALE KNOWLEDGE

Computer Science & Engineering (DS)
of
Maulana Abul Kalam Azad University of Technology
(Formerly known as West Bengal University of Technology)

Intel Unnati Industrial Training 2025 Report

Submitted by

Name of Student(s)	University Roll No
MAYUKH ADHIKARI	11600322028
SALMALI SAMANATA	11600322048
ROUNAK KUMAR SINGH	11600322045

Conducted at

INTEL®

Department of Computer Science & Engineering,
MCKV Institute of Engineering 243, G.T. Road(N)Liluah,
Howrah - 711204

Project Report

Converting Enterprise PDFs into Searchable Knowledge Using Retrieval-Augmented Generation (RAG)

Introduction

Enterprises rely heavily on PDF documents such as manuals, policies, research papers, and reports to store critical knowledge. Although PDFs preserve formatting, they are inherently unstructured, making it difficult for machines to interpret their content. As organizations grow, retrieving relevant information from large collections of PDFs becomes inefficient and time-consuming. This project addresses this challenge by leveraging Retrieval-Augmented Generation (RAG) to transform enterprise PDFs into searchable and structured knowledge.

Problem Statement

Traditional PDF documents lack semantic structure, which limits the effectiveness of keyword-based search systems. Scanned documents further complicate retrieval as they require Optical Character Recognition (OCR). Additionally, tables, charts, and images embedded within PDFs are not directly searchable. Multilingual documents introduce further complexity, leading to poor information retrieval and reduced productivity.

Project Objectives

The primary objective of this project is to extract text from both digital and scanned PDFs and preprocess it for analysis. The system segments documents into meaningful semantic chunks and generates vector embeddings to enable semantic search. Using RAG, the system provides accurate and context-aware answers while supporting multilingual queries.

Proposed Solution

The proposed system integrates document retrieval with controlled language generation. OCR is used to extract text from scanned PDFs, followed by cleaning and semantic chunking. Vector embeddings capture the meaning of document segments, enabling similarity-based retrieval. The RAG framework ensures responses are grounded in retrieved documents, minimizing hallucinations and improving accuracy.

System Architecture

The system follows a modular architecture consisting of PDF ingestion, OCR and text extraction, preprocessing, embedding generation, a vector database, a RAG engine, and a user interface. This design supports scalability and makes the solution suitable for enterprise environments.

Methodology

The methodology begins with ingesting enterprise PDFs and applying OCR where necessary. Extracted text is cleaned, normalized, and divided into semantic chunks. Embeddings are generated and stored in a vector database. User queries are processed by retrieving relevant chunks and generating grounded responses using RAG.

Business and Engineering Impact

The system significantly reduces the time required to search enterprise documents and improves decision-making. It is applicable across departments such as HR, compliance, engineering, and research, providing both business and engineering value.

Relevance to Intel

This project addresses large-scale enterprise documentation challenges and aligns with responsible AI principles. It demonstrates end-to-end system design and is relevant to roles in AI/ML, data engineering, software development, and research.

Conclusion

This project demonstrates a practical and scalable enterprise AI solution for transforming PDFs into searchable knowledge. By combining OCR, semantic search, and Retrieval-Augmented Generation, it overcomes the limitations of traditional document search and delivers reliable and accurate information retrieval.