Project Report on

Problem statement 4

Convert Enterprise PDFs into Searchable Knowledge

College Name
Nutan College of Engineering and Research, Talegaon Dhabade, Pune - 410507

Submitted By:

1. **Mithilesh Yogesh Kolhapurkar**
2. **Ankita Shubash Patil**
3. **Vedant Machindra Thorat**

Under the Guidance of :
Prof. Priyanka Vyas

Intel Technology India Private Limited

2025 - 26

Index

Intel Nexus: Enterprise Document Analyzer

1. Introduction

In today's digital world, organizations handle a massive amount of unstructured data such as PDFs, scanned documents, reports, and handwritten files. Extracting meaningful information from these documents manually is time-consuming and error-prone.

Intel Nexus is an intelligent document processing system designed to automatically extract, organize, search, and generate answers from enterprise documents using modern Artificial Intelligence techniques such as OCR, Semantic Search, Vector Databases, and Retrieval-Augmented Generation (RAG).

2. Problem Statement

Companies have thousands of PDF documents like reports, manuals, and policies. Searching inside these documents is hard because the information is unstructured. Your task is to build a tool that converts PDFs into a structured format so that information can be easily searched and retrieved.

The tool should:

- Read PDFs properly (including scanned ones using OCR), keeping the Table of Contents and document layout in mind.
- Break text into meaningful sections (without cutting across chapters) and store them in a searchable database. Extract tables and store them in a NoSQL database for accurate retrieval. Store the text in a Store text in a Vector Database for better search.
- Handle tables by extracting them and storing them in a way that allows precise queries.
- Handle images and charts by adding short descriptions so they can also be searched.

3. Objective of the Project

The main objectives of this project are:

- To automate extraction of information from unstructured documents
- To enable semantic (meaning-based) search instead of keyword search
- To provide accurate answers grounded in document content
- To reduce manual effort and improve information accessibility
- To build an enterprise-ready intelligent document query system

4. System Architecture Overview

The Intel Nexus system follows a modular and layered architecture consisting of:

- Document Ingestion Layer
- Preprocessing and OCR Layer
- Knowledge Extraction Layer
- Vector Database Layer
- Retrieval-Augmented Generation (RAG) Layer
- User Interface Layer

Each layer performs a specific role to ensure accurate, scalable, and efficient document understanding.

5. Flow of the System (Step-by-Step)

Step 1: Document Upload

Users upload documents such as PDFs (digital or scanned) through the web interface.

Step 2: Preprocessing

The system checks document type:

- If the document is scanned → OCR is applied
- If the document is digital → text is extracted directly

Step 3: OCR and Content Extraction

- Text is extracted from pages
- Tables and images are identified
- Content is cleaned and structured

Step 4: Text Chunking and Embedding

- Extracted text is divided into smaller chunks
- Each chunk is converted into numerical vectors (embeddings)

Step 5: Vector Database Storage

- All embeddings are stored in a vector database
- Enables fast semantic similarity search

Step 6: User Query

The user enters a natural language question.

Step 7: Semantic Retrieval

- System searches the vector database
- Retrieves the most relevant document chunks

Step 8: RAG-Based Answer Generation

- Retrieved content is sent to the language model
- The model generates an accurate answer grounded in document data

Step 9: Result Display

- Final answer is displayed to the user
- Supporting document context can also be shown

6. Key Technologies Used

Technology	**Purpose**
OCR	Converts scanned images to text
Vector Embeddings	Represents text meaning numerically
Vector Database	Stores and retrieves embeddings
Semantic Search	Meaning-based information retrieval
RAG	Combines retrieval and generation

FastAPI	Backend API services
Streamlit	Interactive user interface

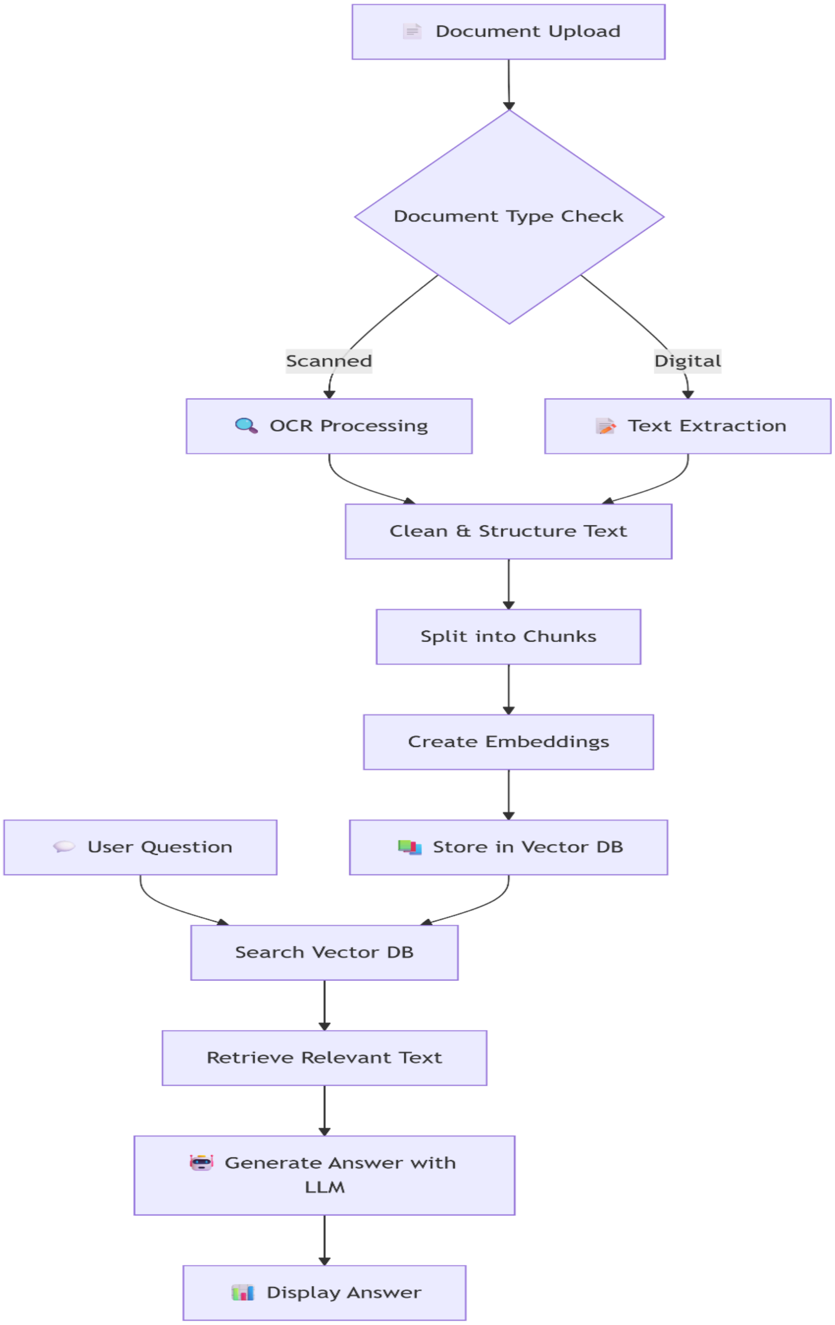

Fig 1. Workflow

8. Advantages of the System

- Handles large volumes of documents efficiently
- Supports scanned and handwritten documents
- Provides accurate and context-aware answers
- Reduces manual document analysis effort
- Scalable and enterprise-ready architecture

9. Applications

- Enterprise document management
- Legal and contract analysis
- Academic research document querying
- Financial and audit document processing
- Healthcare and policy document analysis

10. Conclusion

The Intel Nexus Document Intelligence System demonstrates how modern AI techniques can transform unstructured documents into valuable knowledge. By integrating OCR, vector databases, semantic search, and Retrieval-Augmented Generation, the system provides an efficient, scalable, and intelligent solution for document understanding and query answering.