PDF ANALYZER PROJECT – REPORT

PROBLEM STATEMENT

Organizations and businesses deal with a huge number of PDF documents such as invoices, research papers, contracts, reports, and forms. Manual processing of these documents is:

Time-intensive – extracting and organizing information takes hours.

Error-prone – manual data entry often leads to mistakes.

Not scalable – difficult to manage when document volume grows.

There is a strong need for an **automated PDF Analyzer** that can read, extract, and interpret information from PDFs accurately, efficiently, and at scale.

TECHNOLOGIES USED

Programming Language: Python

Frontend: Streamlit (interactive user interface)

AI/LLMs:

IBM Watson (for NLP and enterprise AI integration)

Hugging Face Models (IBM Granite, Mistral, etc.)

Libraries/Frameworks:

PyMuPDF (PDF parsing and text extraction)

FAISS (vector database for semantic search and document similarity)

Deployment: Cloud-ready (Streamlit apps can be hosted for end-users)

FLOW DIAGRAM

Workflow of PDF Analyzer Project

Upload PDF Document → **Parse & Extract Text/Tables** (PyMuPDF) → **Vectorize Content** (FAISS for indexing/search) → **Process with LLMs** (IBM Watson, Mistral,

Hugging Face Granite) → **Generate Insights** (Summaries, Key Fields, Classifications) → **Display Output** (via Streamlit UI)

(Can be shown visually in slides as a block flow diagram)

SOLUTION

The **PDF Analyzer** automates document processing with the following approach:

Extraction: PyMuPDF extracts raw text, metadata, and tables from uploaded PDFs.

Indexing: Extracted data is embedded and stored using FAISS for quick semantic search and retrieval.

Analysis: Large Language Models (IBM Watson, Mistral, Granite via Hugging Face) interpret the data, summarize content, classify sections, and highlight key fields.

User Interface: Streamlit provides an interactive platform for users to upload files, view results, and download structured outputs.

This pipeline ensures **fast, intelligent, and scalable** PDF analysis.

REAL-TIME USES / APPLICATIONS

Finance: Automated invoice extraction, billing summaries, receipt tracking.

Healthcare: Patient report analysis, extracting lab values, digitizing prescriptions.

Education & Research: Summarizing research articles, extracting references, study note creation.

Business & Legal: Analyzing contracts, compliance verification, and clause detection.

Logistics: Processing shipment forms, delivery notes, and customs documents.
