PaperIQ AI Powered Research Insight Analyzer

By: Ishan Patil

Overview

- Introduction 01
- Objectives 02
- The Research Bottleneck 03
- Technical Architecture 04
- Powerful Feature Set 05
- Future Roadmap 06
- Conclusion and Impact 07

Introduction

With the growing volume of research papers, reports, and documents, extracting relevant information has become tedious and time-consuming. Traditional keyword search often misses contextual meaning, making precise retrieval difficult.

The Document Analyzer Chatbot addresses this by processing PDF/DOCX files and answering queries in a context-aware manner. It leverages lightweight NLP techniques (tokenization, lemmatization, stopword removal, topic extraction, TF-IDF) combined with regex-based rule matching to deliver accurate and efficient information retrieval.

Objectives

Objective 01

Extract and structure text from PDF and DOCX documents for analysis.

Objective 02

Understand user queries and identify key topics for retrieval.

Objective 03

Retrieve relevant content using regex-based pattern matching and TF-IDF cosine similarity.

Objective 04

Generate natural, context-aware responses tailored to query types (definition, features, how/why).

Objective 05

Provide document summarization and an interactive chat interface for easy access to information.

The Research Bottleneck

Time-Intensive Reading

Researchers spend 40% of their time manually reading lengthy academic papers and reports

Manual Extraction Burden

Extracting definitions, key features, processes, and summaries requires painstaking manual effort across multiple documents

No Interactive Q&A

Traditional PDF readers lack conversational capabilities - researchers can't ask questions and get instant answers from their documents

Technical Architecture

01

Document Processing

- Extract text from PDF (PyMuPDF) and DOCX (python-docx) files
- Maintain paragraph and section-level structure for downstream analysis

02

Text preprocessing

- Tokenization, lemmatization, and stopword removal using NLTK
- Clean text and normalize content for accurate retrieval

03

Query Understanding

- Identify headings and split text into structured sections
- Break sections into manageable paragraph chunks (about 200 chars)

04

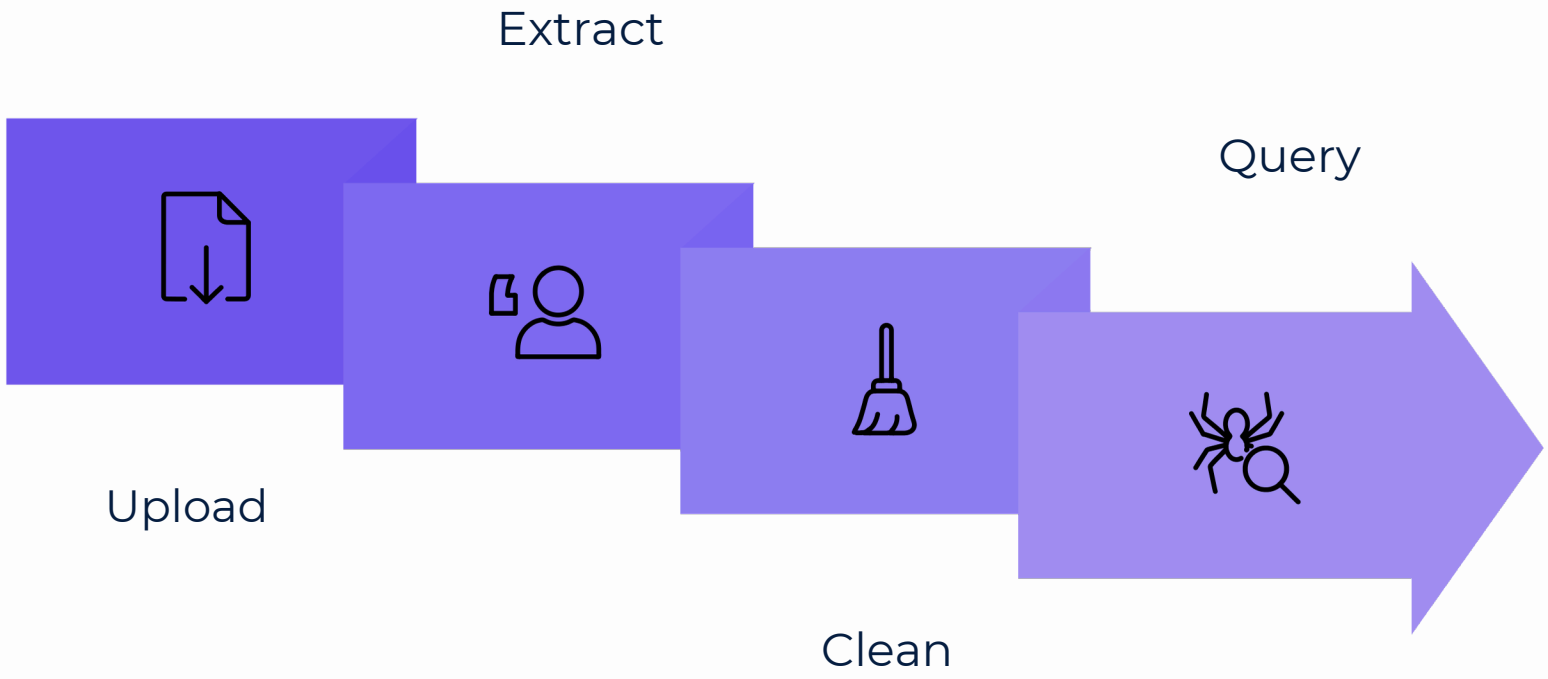
Smart Retrieval

- Regex-based and NLP-assisted extraction of query topics
- Classify queries into types: definition, features, working, literature review
- Display page number for the reference

05

Response Generation

- Regex + TF-IDF similarity to find relevant paragraphs
- Rule-based NER
- Context-aware natural responses formatted for readability


Extract
Query
Upload
Clean

Tech Stack

Frontend

- HTML
- CSS
- JavaScript

Backend

- Python
- NLTK (Natural Language Toolkit)
- scikit-learn
- PyMuPDF (fitz)
- python-docx
- NumPy

Other libraries used

- re
- math
- io (BytesIO)

Future Roadmap

1. **Multi-Document Support**

   Processand cross-reference multiple research papers simultaneously for comprehensive literature reviews

2. **Advanced LLM Integration**

   Enhanced natural language understanding with state-of-the-art language models for deeper insights

3. **Export Capabilities**

   Save answers and summaries directly to Word/PDF formats for seamless integration with research workflows

4. **Voice-Based Q&A**

   Hands-free interaction allowing researchers to ask questions while reviewing other materials

Conclusion & Impact

The Document Analyzer Chatbot provides an effective solution for extracting and retrieving relevant information from large PDF/DOCX documents. By combining preprocessing, topic extraction, regex-based rules, and TF-IDF similarity, it delivers accurate, context-aware responses to user queries. The system reduces manual effort, improves efficiency, and demonstrates how lightweight NLP can be used for practical document analysis. With further enhancements, it can evolve into a robust research and enterprise assistant.

Immediate Benefits

PaperIQ delivers instant insights from long documents, reducing manual effort and accelerating research timelines

Productivity Enhancement

Transforms research review from hoursto minutes, enabling teams to focus on analysis rather than information gathering

THANK YOU!