Docx-DRS: Advanced Document Retrieval System

# 1. Project Overview

Docx-DRS is an advanced document search and retrieval system built using Python and Streamlit, designed to process and search through multiple document types with intelligent text extraction and search capabilities.

# 2. Technical Architecture

## 2.1 System Components

**The system consists of two primary Python modules:**

• document\_processor.py: Core text extraction and search logic

• app.py: Streamlit user interface and application workflow

# 3. DocumentProcessor Class

## 3.1 Text Extraction Method

The *extract\_text()* method supports multiple file formats:

• PDF: Uses PyPDF2 for text extraction

• DOCX: Utilizes python-docx to extract paragraph text

• XLSX: Converts Excel data to string representation

• PNG/JPEG: Leverages Tesseract OCR for image text recognition

## 3.2 Suggestion Generation

The *generate\_suggestions()* method creates meaningful search suggestions by:

• Converting text to lowercase

• Removing non-alphabetic characters

• Filtering out common stop words

• Identifying top 5 most frequent meaningful words

## 3.3 Advanced Search Mechanism

The *advanced\_search()* method provides intelligent document searching through:

• Case-insensitive search functionality

• Removal of non-alphabetic characters

• Extraction of relevant text passages

• TF-IDF based search algorithm

# 4. Streamlit User Interface

## 4.1 Key Features

• Upload up to 3 documents simultaneously

• Support for PDF, DOCX, XLSX, PNG, JPG formats

• Intelligent document suggestions

• Flexible search across documents

• Direct document download functionality

# 5. Technical Dependencies

• Streamlit

• PyPDF2

• python-docx

• pandas

• pytesseract

• Pillow (PIL)

• scikit-learn

# 6. System Limitations

• Limited to 3 simultaneous document uploads

• Basic OCR capabilities

• Simple TF-IDF search algorithm

# 7. Future Enhancement Roadmap

• Advanced NLP search techniques

• Expanded file type support

• Improved OCR accuracy

• Machine learning-based document similarity

# 8. Conclusion

Docx-DRS represents an innovative approach to document retrieval, combining multiple technologies to provide a user-friendly, versatile document search experience.

# 9. About

Created with ❤️ by Renix