PDF Extraction Database Report

*Report generated on: 2025-05-19 09:18:38*

# PDF Extraction Progress

Total PDFs Processed: 1696 documents

Total Text Chunks Extracted: 108948 chunks

Average Chunks per PDF: 64.24

Processing Rate: 101.24 PDFs per hour

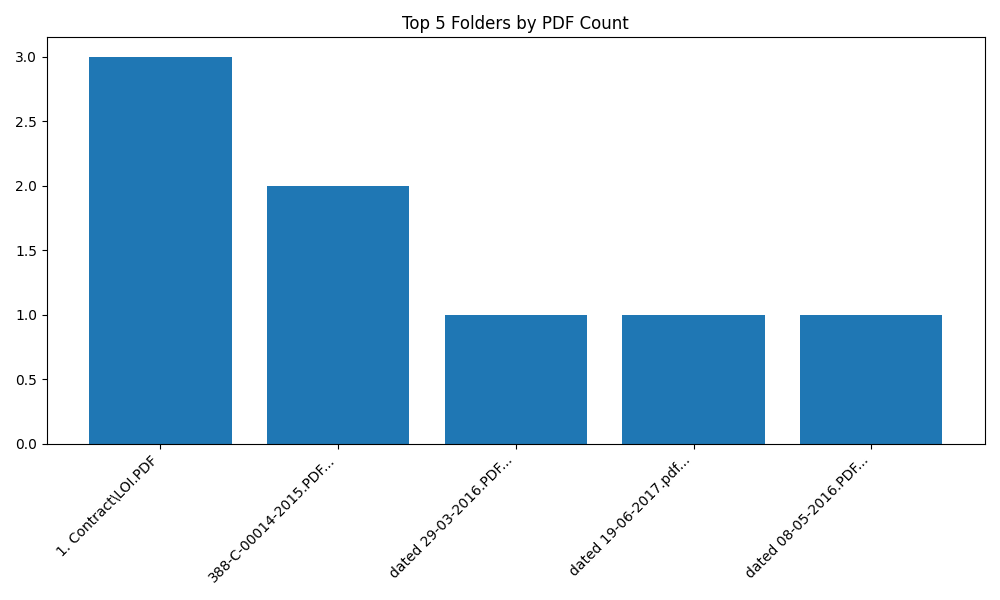
Processing Duration: 16.75 hours

Started on: 2025-05-18 16:32:13

Last PDF Processed: "عقد مصنعيات اعمال محارة-سيفك.pdf" at 2025-05-19 09:17:19

*The extraction process is handling both English and Arabic documents successfully, as evidenced by the Arabic filename in the last processed PDF.*

## Top 5 Folders by PDF Count



# Database Structure

## po.pdfs table:

• Stores basic PDF information (path, filename, hash)

• Includes the full extracted text

• Tracks processing timestamp

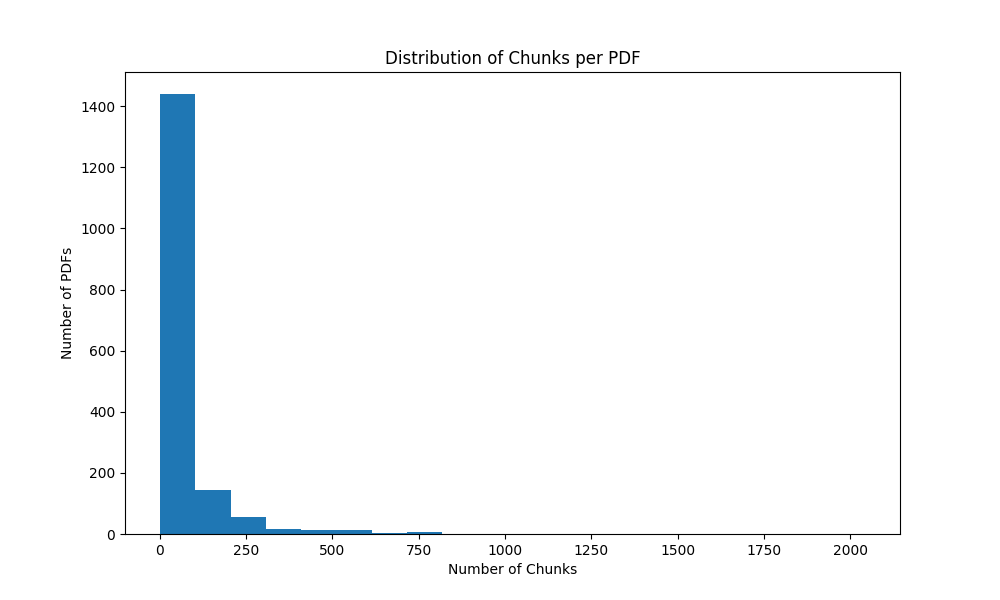
## po.pdf\_chunks table:

• Stores individual text chunks

• Links back to the parent PDF via pdf\_id

• Contains 108948 text segments from the processed PDFs

## Distribution of Chunks per PDF



# Interesting Statistics

• The PDF with the most chunks ("Beni swif Aresco signd.PDF") contains 2042 text segments

• The PDF with the fewest chunks ("قيمة امر توريد - شركة البركة.PDF") contains only 2 text segments

• The resume capability is working well, as the system continues processing new files without duplicating already processed ones.

# Conclusion

The PDF extraction process is running successfully and has made significant progress. The system is handling both English and Arabic documents effectively, as evidenced by the variety of filenames in the processed PDFs.

The extraction process is efficiently splitting the text into meaningful chunks, with an average of 64.24 chunks per document. This granular approach will make it easier to analyze and search through the extracted content.

The resume capability is functioning as designed, allowing the process to be stopped and restarted without duplicating work. This is particularly important for processing large collections of PDFs over extended periods.