

TEAM - 3

JIBIN C BABY (TEAM LEAD)

ANURUDH DEEPAN

BADUSH

FIDHA P HARIS

JESWIN JAISON

AUTOMATIC TITLE AND AUTHOR EXTRACTION SYSTEM FOR SCIENTIFIC DOCUMENTS

PROBLEM STATEMENT

Develop an automated system to extract title and author information from scientific documents and propose alternative titles based on the document content.




INTRODUCTION

The system automates title and author extraction from scientific documents using OCR and NLP techniques, while also generating alternative titles based on content analysis. It evaluates the accuracy of extraction and relevance of alternative titles to ensure effectiveness in conveying the document's essence.

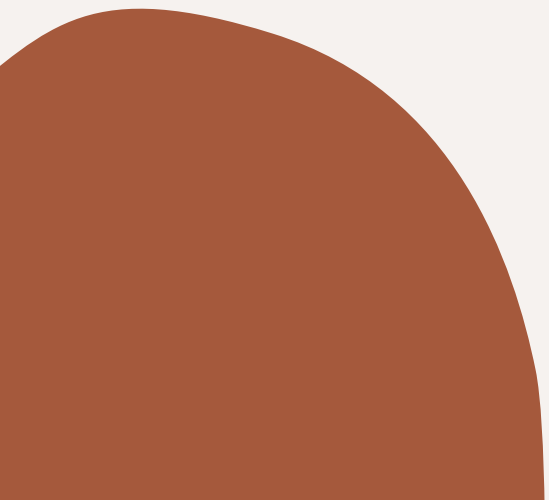


PREPROCESSING STEPS

- Dataset provided as scanned/photographed images of scientific documents.
 - OCR (Optical Character Recognition): Convert printed document images into machine-readable text.
 - Store the preprocessed text data in a suitable format for further classifications and analysis.
- 

TITLE AND AUTHOR EXTRACTION

- Title Identification: Utilize NLP techniques such as pattern matching and keyword extraction to identify document titles.
- Author Extraction: Employ Named Entity Recognition (NER) algorithms to detect author names from the document text.



CONTENT ANALYSIS

- Text Summarization: Generate a concise summary of the document content using techniques such as LexRank or TextRank.
- Keyword Extraction: Identify key concepts and terms within the document using TF-IDF or RAKE algorithms.



ALTERNATIVE TITLE GENERATION

- **Summarization-Based:** Generate alternative titles by summarizing the document content using NLP models like BERT or GPT.
- **Keyword-Based:** Extract significant keywords from the document and construct alternative titles based on their relevance and importance.

IMPLEMENTATION DETAILS

- Programming Language: Python
- Libraries:
 - pytesseract - OCR
 - transformers - Advanced NLP tasks
 - spacy - NLP tasks
 - sumy - Text summarization
 - openpyxl - Write to an Excel File

WORKFLOW

- The system follows a sequential workflow:
 - OCR Engine extracts text from printed documents.
 - Title and Author Extraction Module identifies document titles and authors from the extracted text.
 - Content Analysis Module summarizes the document's content and extracts key concepts.
 - Alternative Title Generation Module uses the summarized content and key concepts to generate alternative titles.
 - Write Author name, Title and Alternative Title into a Excel File.

CONCLUSION

- The automated system demonstrates high efficiency and accuracy in extracting title and author information from scientific documents.
- The system provides valuable insights for researchers and academics by generating alternative titles that effectively capture the essence of the document's content.