NATURAL LANGUAGE PROCESSING

AUTOMATIC DOCUMENT FORMAT AND CONTENT RECOGNITION FOR ACADEMIC PAPERS

Pranjul Mishra
Nazira Tukeyeva
Saurabh Singh

RESEARCH QUESTIONS

The project focuses on addressing the following **key research questions**:

1. How can we design a method to automatically detect and extract essential components from academic documents (e.g., title, author(s), abstract, and section content including non textual elements like images, tables etc)?
2. Which NLP techniques are most effective for segmenting document content by headings and sub-headings and in hierarchical way?

BACKGROUND WORK

- **Document Structure Recognition:**
  - Evolution from rule-based methods (e.g., CERMINE) [1] to transformer-based model like BERT [2].
- **Content Segmentation:**
  - Shift from traditional template-based approaches to applying RNNs to segment text by learning contextual patterns within document content [3].
- **Non-Textual Element Extraction:**
  - Tools like DeepDeSRT, Camelot, and Tabula enhance detection of tables and figures [4].
- **Challenges in Document Analysis:**
  - Diversity in document layouts (formatting styles, heading structures), and limited annotated datasets for training document analysis models [5].

METHODOLOGY

Datasets:

- PubMed Central Open Access Subset (PMC-OAS)
- arXiv Dataset
- ICDAR Competition Datasets
- GROBID

Data Processing:

- Text Extraction: PDFMiner (extracts raw text from PDF)
- Tokenization and Segmentation
- Noise Removal

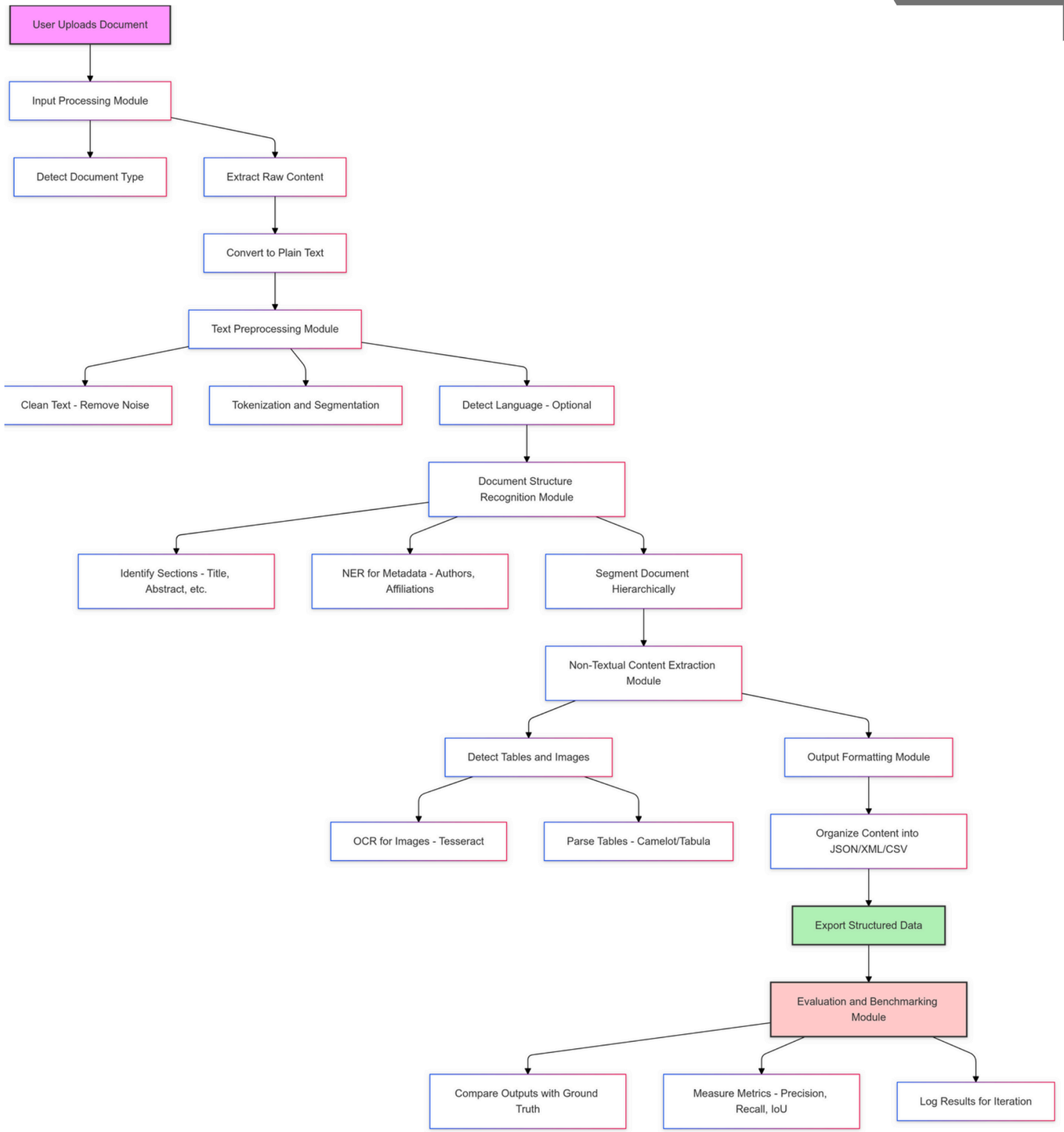
Document Segmentation and Content Extraction:

- NLP Models (eg. BERT): understand and segment the document based on contextual relationships within the text hierarchically.
- NER: identify entities
- Regular Expressions: identify commonly formatted sections, such as references or bibliography

Non-Textual Element Extraction:

- OCR for Embedded Text: extract text from images
- Table Extraction Tools: detect and extract tabular data from PDF files (eg. Camelot, Tabula)
- Layout Processing with OpenCV: analyze layout structure

ARCHITECTURE



FUTURE SCOPES

1. The output format which we are focusing is in Json/XML/CSV which further can be converted into set of triples using set of rules and thus a knowledge graph can be created .
2. Upon integration with LLM models (ex. Llama 3.2 -3B instruct etc) it can be converted into a RAG based system which can be used to query regarding academic documents.

REFERENCES

[1] Tkaczyk, Dominika, et al. ”CERMINE: automatic extraction of structured metadata from scientific literature.” International Journal on Document Analysis and Recognition (IJDAR) 18.4 (2015).

[2] Devlin, Jacob, et al. ”BERT: Pre-training of deep bidirectional transformers for language understanding.” arXiv:1810.04805 (2019).

[3] Yang, Zhilin, et al. ”Neural machine translation with recurrent attention modeling.” arXiv:1703.04675 (2017).

[4] Schreiber, Sebastian, et al. ”Deepdesrt: Deep learning for detection and structure recognition of tables in document images.” 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Vol. 1. IEEE, 2017.

[5] Gao, Liangcai, et al. ”ICDAR 2019 competition on table detection and recognition (cTDaR).” International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019.

THANK YOU!