LaTeX Rainbow:

Open-Source LaTeX to PDF Document Semantic & Layout Annotation Framework

Changxu Duan, Sabine Bartsch
Technische Universität Darmstadt
Institute of Linguistics and Literary Studies
Residenzschloss Darmstadt, 64283 Darmstadt, Germany

Introduction

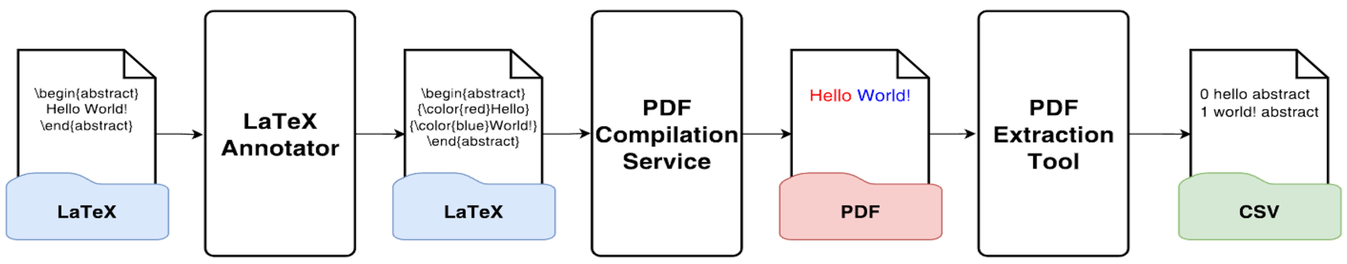
Document Layout Analysis (DLA) systems are instrumental in semantically labelling diverse document elements such as text blocks, figures, and tables.

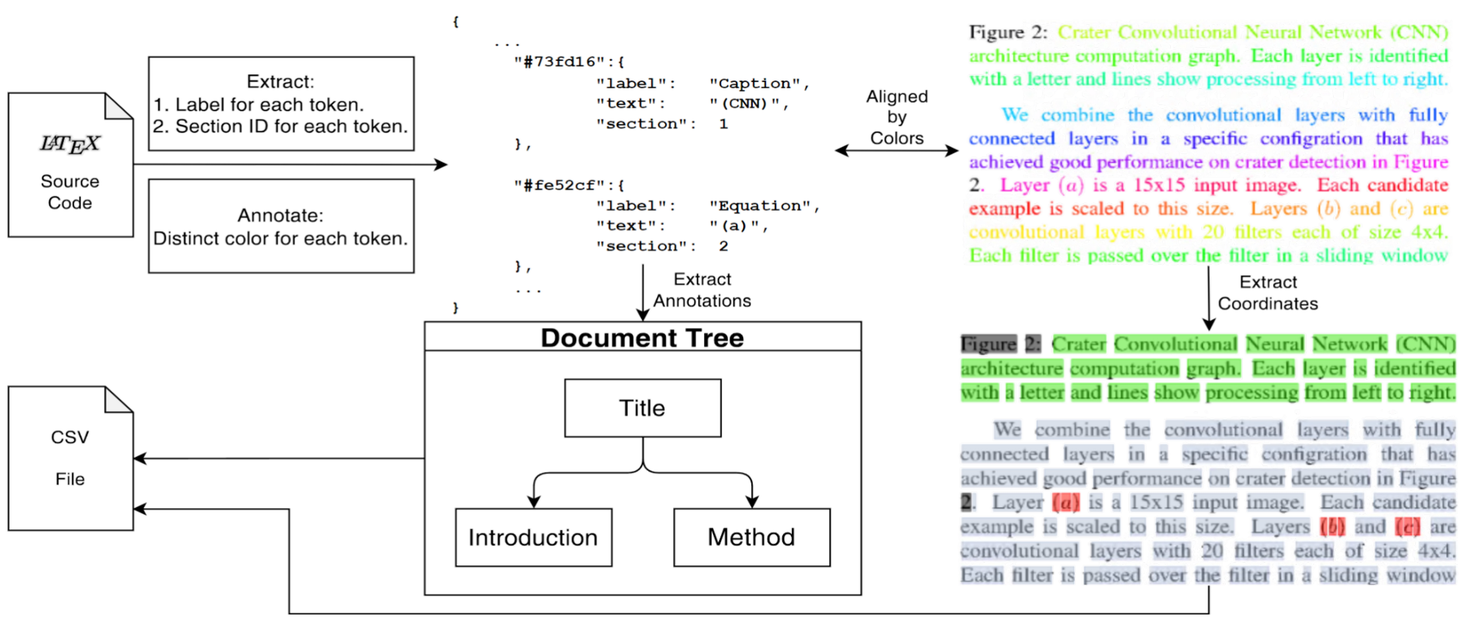
Presently, annotations for DLA datasets are confined to single pages, neglecting the continuity of elements across pages. This oversight can lead to misinterpretations, affecting the coherence of extracted information. Another issue is the inclusion of irrelevant elements like watermarks and headers, introducing noise into the data.

Furthermore, the increasing volume of scientific publications and the absence of code for dataset compilation pose challenges in dataset currency and reproducibility.

We introduce a novel framework for automatic PDF annotation, producing document-oriented, fine-grained, and reading-ordered annotations.

Annotation Process





Methodology

LaTeX Rainbow framework is built around three modules.

LaTeX Annotator

- We parse LaTeX asource code into segments: ***macro***, ***environment***, ***body***, and ***comments***.
- Special attention is given to labeling specific macros like \title{} and \author{}, and environments enclosed within \begin{} and \end{} markers. Texts and figures are tokenized and annotated with corresponding color-coded tags.

PDF Compilation Service

TeXLive 2023 Docker container, Plus Perl-based AutoTeX and a Python-based API for LaTeX compilation. Offers a streamlined, reliable LaTeX compilation service accessible via HTTP.

PDF Extraction Tool

Starting by identifying the position of each rectangle and matching it to corresponding figures in the annotated source code using fill colors. It then discerns the color and position of each letter, aligning them with their respective annotations.

From and To the Open-source Community

All the main features of the LaTeX Rainbow Framework are assembled from and rely on many open-source projects:

- **pylatexenc**[1] 3.0alpha is used to identify and traverse the LaTeX code.
- Improved LaTeX parsing rules rule from **TeX-Workshop**[2] and **TeXstudio**[3].
- The PDF compilation service is essentially inherited from **texcompile**[4].
- **pdfplumber**[5] is used to extract shapes and texts from PDF files.

LaTeX Rainbow framework itself is based on Apache 2.0 protocol is completely open source.

There are many papers that cannot yet be parsed correctly. So, we greatly welcome and depend on the open-source community to contribute the detailed parsing rules for each publisher's template.

Conclusion

LaTeX offers unique advantages for Document Layout Analysis due to its explicit markup, clearly defining document elements and author intent.

Our framework leverages this structure, enhancing Document Layout Analysis by simplifying the identification and annotation of document components. Aimed to be fine-grained and scalable, and to view it as a universal tool, fostering innovative applications across multiple disciplines within the open-source community.

References

[1] https://github.com/phfaist/pylatexenc
[2] https://github.com/James-Yu/LaTeX-Workshop
[3] https://github.com/texstudio-org/texstudio
[4] https://github.com/andrewhead/texcompile
[5] https://github.com/jsvine/pdfplumber

This project was conducted within the research project InsightsNet which is funded by the Federal Ministry of Education and Research (BMBF).