Brief Report on Information Extraction and Classification for ICDAR 2024 Papers. Captonomy challenge.

Theodoros Anagnos

October 8, 2024

1 Introduction

The primary goal of this project was to extract key information from research papers, including the title, authors, abstract, and introduction, and to classify each document into predefined categories such as Tables, Classification, Key Information Extraction, Optical Character Recognition, Datasets, Document Layout Understanding, and Others. To achieve this, we utilized various NLP techniques and deep learning models, including `spaCy` and `BERT`-like models for zero-shot classification.

2 Methodology

We followed a multi-step approach to accomplish the task:

- **PDF Text Extraction**: We utilized the `PyMuPDF` library to extract textual content from PDF files. Text blocks and font sizes were used to identify titles, authors, and other key sections.
- **Title and Author Extraction**: For title and author extraction, we used a combination of heuristics based on font size and regular expressions. We also leveraged `spaCy`'s Named Entity Recognition (NER) capabilities to identify people (authors) and organizations.
- **Zero-shot Classification**: For document classification, we employed a zero-shot learning approach using a `BERT`-like model. The model was tasked with categorizing papers into one of the predefined categories based on the text in the abstract or introduction.
- **CSV Data Cross-Check**: We utilized a reference CSV file that provided metadata (ID, title, authors) for the papers, which helped in validating the extracted information.

3 Challenges

Several challenges were encountered during the project:

- **Title and Author Recognition**: While we achieved a recognition rate of approximately 70%, there were instances where the authors' names or titles were not correctly extracted, largely due to the variability in PDF formatting.
- **Zero-Shot Classification Performance**: Some categories like Tables, Classification, Key Information Extraction, Optical Character Recognition, and Datasets were not assigned at all by the zero-shot classifier. This raised questions about the reliability of using zero-shot methods without fine-tuning in such domain-specific tasks.
- **Handling Noisy PDF Data**: PDF files can contain a lot of noise, including headers, footers, and artifacts that interfere with text extraction. Additionally, some papers had inconsistent formatting, making it harder to consistently extract sections like the abstract or introduction.

To address these issues, we employed a combination of regex-based cleaning and entity recognition. However, certain inaccuracies in extraction and classification remained.

4 Conclusion and Future Work

Our approach demonstrated moderate success in extracting titles and authors from research papers, with a recognition accuracy of around 70%. However, several challenges were identified, especially regarding the zero-shot classification approach, which failed to assign certain categories.

Future Improvements: To enhance the performance, especially for title and author extraction, we propose fine-tuning the `spaCy` model using the data from the CSV file. Additionally, to improve document classification, a promising avenue would be to label a subset of papers using large language models (LLMs) and fine-tune the `BERT`-like model specifically for this task. This fine-tuning step could significantly improve the classification performance across all predefined categories.

Final Thoughts: This project highlighted the complexities involved in information extraction from heterogeneous PDF documents. Leveraging domain-specific fine-tuning and applying more sophisticated models for text extraction and classification are key directions for future development.