

PDF Title Extraction

The task is to identify important sections (titles) from styled pdf documents, as they represent anchor points for the extraction of various other important information. In the attached CSV files (split into training and testing), you will have access to a limited set of features from sample documents.

The structure of these CSVs is as follows:

- Each line represents information about one particular block of text from the document. This information contains:

- o Text: the raw text of the section as interpreted by the OCR;
- o IsBold: is the section bold or not; o IsItalic: is the section italic or not;
- o IsUnderlined: is the section underlined or not;
- o Left: the left coordinate on the page;
- o Right: the right coordinate on the page;
- o Top: the top coordinate on the page;
- o Bottom: the bottom coordinate on the page; and
- o Label: a label to state is the section represent a title (1) or normal text (0).

You will also find some sample pdfs files to give you a better idea of the documents you are dealing with (titles are highlighted by red rectangles).

Your task is to build a machine learning model that is capable of classifying sections as titles and nontitles. You are free to use any machine learning model you are comfortable with (deep learning models are preferable but not mandatory).

We expect you to submit the following as part of your solution:

- A fully working python project to achieve the above task (with a setup script).
- A report stating the following:
 - o The accuracy achieved against the test set.
 - o A brief explanation of your solution including reasons for your model of choice.
 - o A description of any improvements you would have made given extra time.
 - o Please also specify the amount of time you have spent designing your solution. We intend to run your solution, so please give us an idea of setup, train and run time

Implementing a solution

Dataset

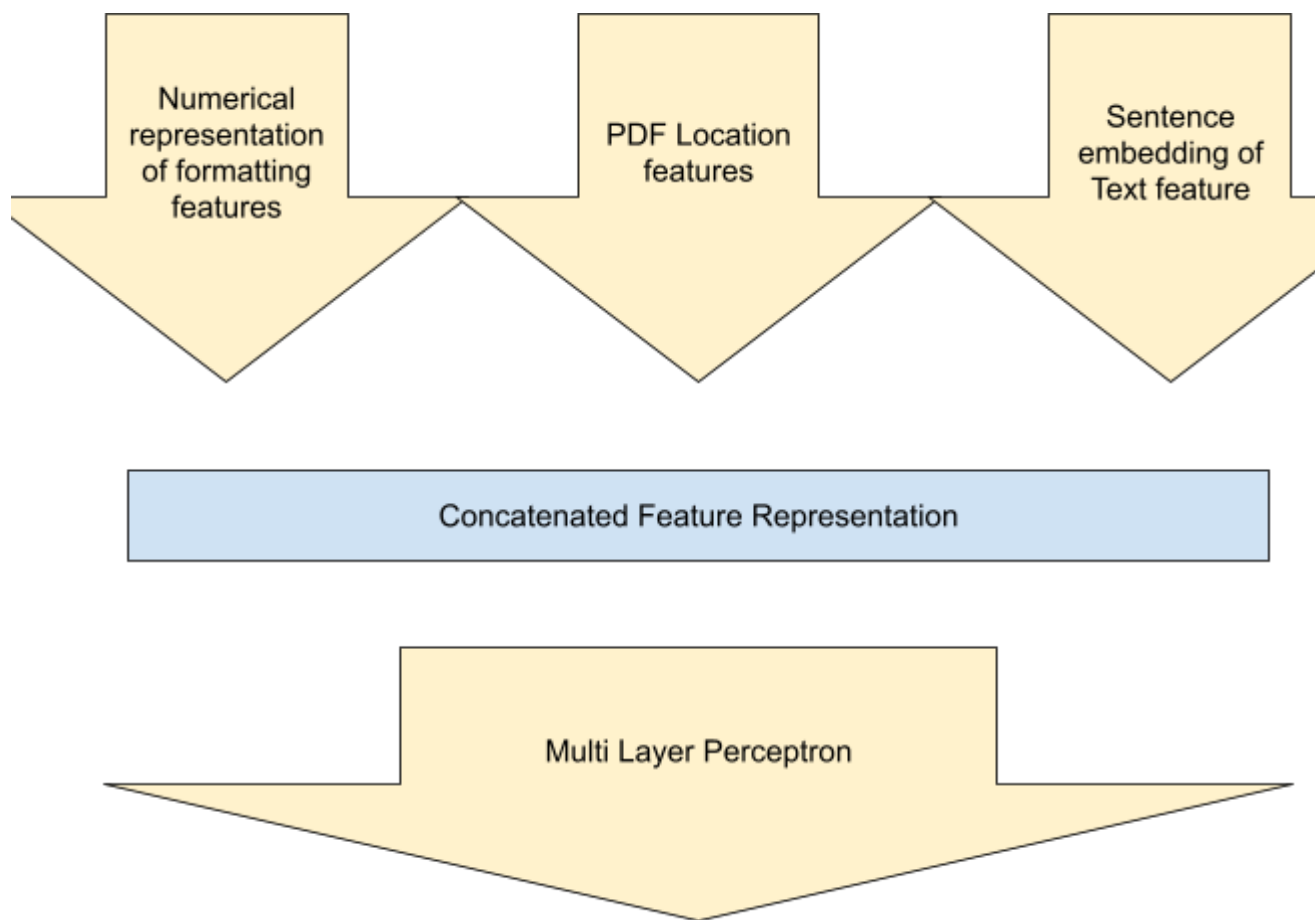
Two datasets have been provided. train_sections_data.csv and test_sections_data.csv. Here is some records from the test dataset.

Text	IsBold	IsItalic	IsUnderlined	Left	Right	Top	Bottom	FontType	Label
Employee Involvement	TRUE	FALSE	FALSE	77.9	175.6	425.3	434.2	New Times Roman	1
Income Taxes	TRUE	TRUE	FALSE	33.1	79.6	284.8	290.6	New Times Roman	1
•	FALSE	FALSE	FALSE	77.9	81.3	346.1	355	New Times Roman	0
the Audit & Remuneration Committee	FALSE	TRUE	FALSE	124.8	247.3	380.3	385.9	New Times Roman	0

Model Architecture

o A brief explanation of your solution including reasons for your model of choice.

I went for an ensemble approach combining three different sets of features.



Formatting Features

The dataset contained columns related to the formatting used in the sample text. These were boolean columns stating whether the formatting had been applied or not. I converted these features into numerical representations. 1 being True, 0 being False.

Location Features

The dataset contained columns related to the location of sample text on the pdf it is a part of. I took these features as they were and concatenated them to the rest of the numerical features.

Sentence Embeddings

Sentence embeddings capture the semantic meaning of text. This is crucial for this task as we are dealing with titles in PDF documents. Titles often convey specific semantic importance, and sentence embeddings help in understanding the contextual meaning of these titles.

I used the 'multi-qa-MiniLM-L6-cos-v1' model from sentence-transformers as my text embedder. It maps sentences & paragraphs to a 384 dimensional dense vector space and was designed for semantic search. It has been trained on 215M (question, answer) pairs from diverse sources. This

means it has learned to generate robust representations for a wide range of textual content. This robustness is valuable when dealing with the varied writing styles and formats that might be present in a PDF document.

Training Results

The F1 score and accuracy achieved against the test set.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1174
1	0.97	0.90	0.93	405
accuracy			0.97	1579
macro avg	0.97	0.95	0.96	1579
weighted avg	0.97	0.97	0.97	1579

Future Improvements

Future improvements:

- Sentence embedding improvements
 - Using a Larger embedding model may increase performance
 - More expressive due to more parameters
 - Experimenting with the max_length of embeddings
- Location feature normalisation
- Efficiency improvements
 - I could have encoded all of the text features records at one time using sentence transformers, instead of record by record. This would save a lot of time and compute

Running the project

If you do not already have the project directory saved to your development machine you can clone it via this link: <https://github.com/ajbanks/pdf-title-extraction.git>.

Run all commands from the root directory. Assumes you are using Python 3.11.5.

1. Set PYTHONPATH

```
set PYTHONPATH=src
```

for linux systems use:

```
export PYTHONPATH=src
```

2. Install dependencies

```
pip install -r requirements.txt
```

3. Run training and evaluation, then print results to cmd line

```
python3 src\title_detection_model.py
```