Document Layout Analysis – Title Detection

The purpose of this exercise is to test the candidate's ability to handle textual data, which is a core part of what we do at Semantic Evolution. 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:
    - Text: the raw text of the section as interpreted by the OCR;
    - IsBold: is the section bold or not;
    - IsItalic: is the section italic or not;
    - IsUnderlined: is the section underlined or not;
    - Left: the left coordinate on the page;
    - Right: the right coordinate on the page;
    - Top: the top coordinate on the page;
    - Bottom: the bottom coordinate on the page; and
    - 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 non-titles. 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:
    - The accuracy achieved against the test set.
    - A brief explanation of your solution including reasons for your model of choice.
    - A description of any improvements you would have made given extra time.
    - 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.