# untitled238

April 29, 2024

The `input` to the `extractor function` is the `list of PDF Paths`.

Then the function takes 3 `approaches`

- **Unstructured** -
- – First the `PDF` is converted to `Images`, lets say we have a `PDF` of 5 pages, then we will first get 5 images, this is like taking screen-shot of each image.
- – After this we use the `Facebook/Detectron2` model family
- – In the model family we focus on `Faster RCNN R 50 FPN 3x` model to detect `different bounding boxes`, majorly based on `Title/Heading/Bullet Points`
- – After this the Image in the bounding box is sent to `OCR Model`
- – The `OCR Model` is based on `CRAFT Architechture` by `Keras`
- – The text is then segregated and used to make chunks
- **Full Text**
- – In this we simply read the pdf character by character and use the `RecursiveCharacterTextSplitter` by `Langchain`
- **Heading and Content**
- – First the `PDF` is converted to `Images`.
- – The Images are first converted to `High Contrast`
- – Further they are converted to `Gray-Scale`
- – After this a `Bi-Linear Function` is applied
- – Further a `Bi-Cubic Function` is applied
- – After this Image is sent to `OCR`
- – After this `Rectangles are created for bounding boxes`
- – `Overlapping rectangles are merged`
- – `Images in bounding boxes are sent of ocr`
- – `Text is retrieved` and `chunks are made`

[ ]: