|  |  |  |
| --- | --- | --- |
|  | Function Name | Logic |
|  | process\_document | • Entry point for the entire process.  • Takes inputs like document path, field names, and OCR data(engine data).  • Coordinates all other functions in the proper sequence.  • Handles both PDF and image files differently |
|  | load\_ocr\_data | • Loads pre-processed text recognition data from JSON file(engine data).  • Contains information about all text found in the document.  • Includes position, size, and content of each text element.  • Handles potential encoding issues when reading JSON |
|  | extract\_text\_regions | • Uses the engine data to locate specific fields of interest.  • Searches through the text blocks to find matching field names.  • Records the position and size of these found text regions.  • Prepares for checkbox detection near these regions |
|  | convert\_ocr\_blocks\_to\_structured\_data | • Transforms raw OCR blocks into organized data.  • Groups text by lines for better context understanding.  • Prioritizes LINE blocks over individual WORD blocks.  • Filters out common checkbox answers to avoid confusion |
|  | detect\_checkboxes | • Analyzes the document image using computer vision.  • Applies thresholding to highlight potential checkbox shapes.  • Uses the text region heights to determine appropriate checkbox size.  • Filters candidates based on height relative to text fields, and only after the text fields ends.  • Applies convex and solidity principles to filter the checkboxes further more. |
|  | calculate\_iou | • Measures the Intersection over Union between box regions.  • Used within checkbox detection to evaluate overlapping shapes.  • Converts box coordinates to calculate precise overlap area.  • Returns a value from 0 (no overlap) to 1 (perfect overlap) |
|  | remove\_overlapping\_checkboxes | • Takes the list of detected checkbox candidates  • Sorts them by size (larger boxes prioritized)  • Compares each box against already-kept boxes using IoU  • Removes duplicates that significantly overlap with existing selections |
|  | is\_checkbox\_checked | • Examines the pixel content inside each detected checkbox.  • Applies morphological operations to clean up noise.  • Calculates the ratio of filled pixels to total pixels.  • Focuses on the center region where check marks typically appear |
|  | visualize\_text\_and\_checkboxes | • Creates a visual output showing the detection results.  • Draws blue rectangles around identified text fields.  • Marks unchecked boxes in yellow with an "X".  • Highlights checked boxes in green with a checkmark symbol |
|  | update\_json\_with\_results | • Final step that updates the JSON data structure.  • Records which checkboxes were checked for each field  • Includes precise position information for each result.  • Formats everything for downstream processing or storage |
|  |  |  |