Design Doc

Initialization: The program takes the path of a file or a directory containing files. It checks if the path is valid and the files are images. If there is an error, it prints the error and exits.

Opening: If the path is correct, the program opens the images using the CV2 library of python.

Parsing: The program uses the Easyocr library for parsing the images. Easyocr returns the text and its position.

Extracting Data: After the image is parsed, we need to extract the title, authors, publishers, and ISBN code of the image. Here is how it is done:

Extracting Title: The title will have the maximum font size among all the data. I am calculating the height of every text and calculating the maximum height. After this, I am taking all the words whose relative difference with maximum height is not more than 0.1.

Extracting ISBN: ISBN code will be a 10 or 13-digit number consisting of numbers from 0 to 9 and ‘-’. It will also be preceded by ‘ISBN’. I am using keyword search and regex to extract ISBN code.

Extracting Authors: To extract the authors, I am using the NLP library. We will know if the given token is a person's name in the image. If there are multiple person names, I am using a size heuristic to get the author's name.

Extracting Publisher: This is done similarly to authors. When NLP returns ‘ORG’, I am assuming the token to be a publisher

Testing: Wrote unit tests for the code and got coverage of 94%.

Module	statements	missing	excluded	coverage
extract_data.py	118	3	0	97%
main.py	64	7	0	89%
parse_image.py	9	0	0	100%
test_main.py	52	0	0	100%
validator.py	45	9	0	80%
write_csv.py	7	0	0	100%
Total	**295**	**19**	**0**	**94%**

coverage.py v6.3.2, created at 2022-04-15 23:31 +0530