Repo Link: Results

Input Data

Document: Monthly Economic Review April 2021 (MER April 2021_0.pdf)

Document Type: Semi-Structured

Content Breakdown: Text: Headings and subpoints structured numerically.

Tables: Includes small, large coloured tables, 8 tables are present.

Figures: Contains approximately 27 coloured visualizations embedded within the text.

Results

Image Extraction:

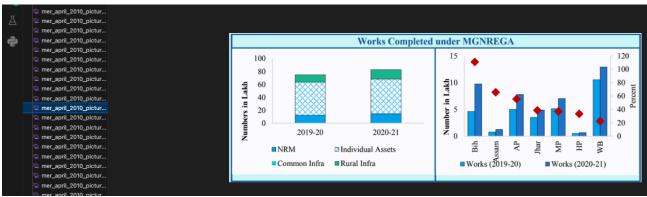
Extracted all 28 figures with exact accuracy, preserving their original format.

Retained figure headings along with the images.

Extracted all the 28 images.

Input: pdf document

Image Resolution set to = image_resolution_scale = 2.0



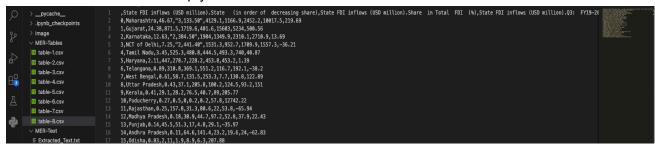
Find more image results in the repository and the code (image.py)

Tables Extraction:

Extracted all the 8 tables and converted in csv format.

Issue 1 - Headers and subheaders are mixed.

Issue 2 - Some rows are empty.

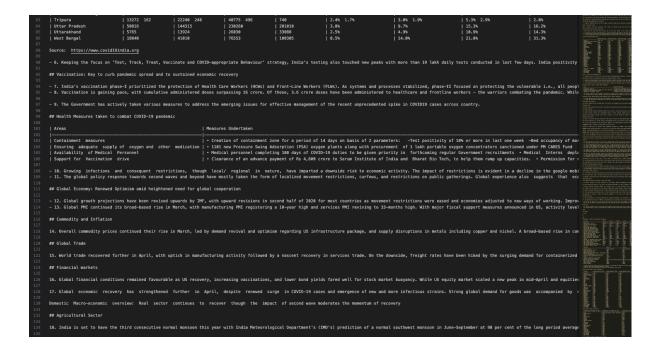


Find more results for table extraction in the repository and the code (tables.py)

Text Extraction

Extracted all the headings , subheadings , bulleted or numbered points correctly , and text present in the boxed section.

Maintained the document structure, also extracted the tables text in tabular format. Find extracted text file (MER-Extracted_Text.txt) in the repository and the code (text.py)



Processing Time

Takes approximately 15-20 minutes for documents ranging between 17-30 pages.