Optical Character Recognition (OCR) for Printed Text

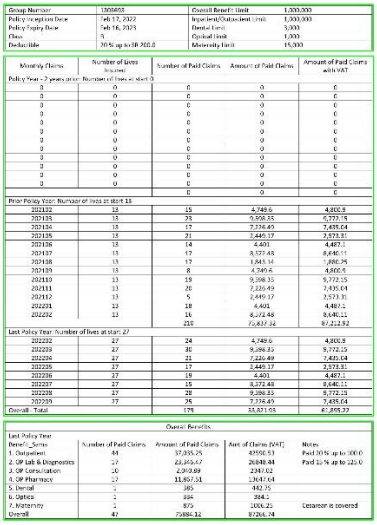
I will use **OpenCV** and **PaddleOCR** to extract a table from an image. The code comprises 3 classes for the 3 stages of the process.

Stage 1: Detecting the table

Finding the areas that are most likely the tables. This is done by simply looking for the biggest number of “boxes” in the image depending on the parameter that the user enter ”tab_num”**.**

This step is composed of 4 sub-steps:

- **Grey-scaling:** it makes processing faster.
- **Thresholding:** Convert the image to just “black” or “white” pixels
- **Inverting:**
- **Dilating:** This will help us to correctly identify the “contours” and hopefully the “contour” that makes up the largest box.



The resulted tables:

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

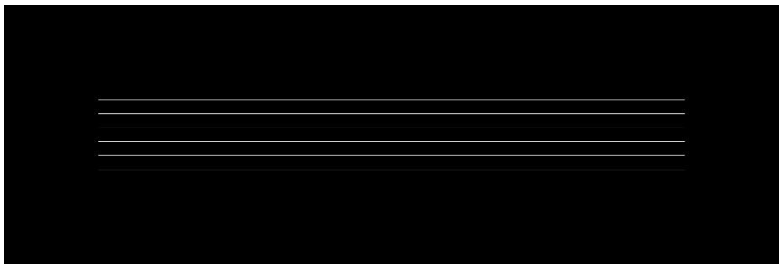
Monthly Claims	Number of Lives Insured	Number of Paid Claims	Amount of Paid Claims	Amount of Paid Claims with VAT
Policy Year - 2 years prior: Number of lives at start 0				
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
		0	0	0
Prior Policy Year: Number of lives at start 13				
202102	13	15	4,749.6	4,800.9
202103	13	23	9,598.35	9,772.15
202104	13	17	7,226.49	7,435.04
202105	13	21	2,449.17	2,573.31
202106	13	14	4,401	4,487.1
202107	13	17	8,572.48	8,640.11
202108	13	17	1,843.14	1,880.25
202109	13	8	4,749.6	4,800.9
202110	13	19	9,598.35	9,772.15
202111	13	20	7,226.49	7,435.04
202112	13	5	2,449.17	2,573.31
202201	13	18	4,401	4,487.1
202202	13	16	8,572.48	8,640.11
		210	75,837.32	87,212.92
Last Policy Year: Number of lives at start 27				
202202	27	24	4,749.6	4,800.9
202203	27	30	9,598.35	9,772.15
202204	27	21	7,226.49	7,435.04
202205	27	17	2,449.17	2,573.31
202206	27	19	4,401	4,487.1
202207	27	15	8,572.48	8,640.11
202208	27	28	9,598.35	9,772.15
202209	27	25	7,226.49	7,435.04
Overall - Total		179	53,821.93	61,895.22

Overall Benefits				
Last Policy Year				
Benefit_Sama	Number of Paid Claims	Amount of Paid Claims	Amt of Claims (VAT)	Notes
1. Outpatient	44	37,035.25	42590.53	Paid 20 % up to 100.0
2. OP Lab & Diagnostics	17	23,346.47	26848.44	Paid 15 % up to 125.0
3. OP Consultation	10	2,040.89	2347.02	
4. OP Pharmacy	17	11,867.51	13647.64	
5. Dental	1	385	442.75	
6. Optical	1	334	384.1	
7. Maternity	1	875	1006.25	Cesarean is covered
Overall	47	75884.12	87266.74	

Stage 2: Removing The Lines

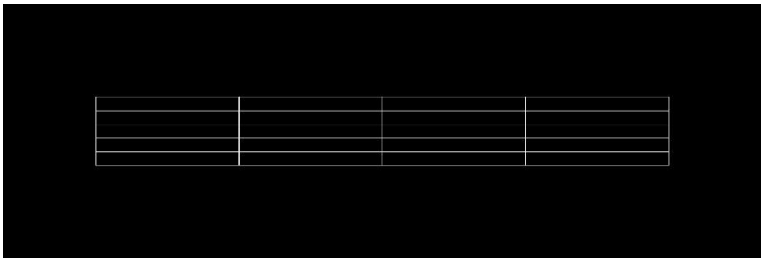
In this stage we'll handle the lines of the table. This will help us to have a clear image for the OCR processing. The only things left of the image, in the end, will be the text in the table cells.

- Eroding Vertical Lines



- Eroding Horizontal Lines

- Combine Horizontal & Vertical Lines



- Subtract and get an image without the lines.

Stage 3: Detect cells in tables and extract text

- Use dilation to convert the words into blobs

- find the contours of the blobs

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

- convert the blobs into bounding boxes

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

- Extracting The Text From The Bounding Boxes Using PaddleOCR
- Generate the CSV file for each table

Manipulate the data to create the required Data Frames

Claims_data:

	Monthly Claims	Number of Lives Insured	Number of Claims	Amount of Paid Claims	Amount of Claims VAT	Policy Year	End Date	Class	Overall Limit
1	0	0	0	0.0	0.0	Policy Year-2 years prior	2023-02-16	B	1000000
2	0	0	0	0.0	0.0	Policy Year-2 years prior	2023-02-16	B	1000000
3	0	0	0	0.0	0.0	Policy Year-2 years prior	2023-02-16	B	1000000
4	0	0	0	0.0	0.0	Policy Year-2 years prior	2023-02-16	B	1000000
5	0	0	0	0.0	0.0	Policy Year-2 years prior	2023-02-16	B	1000000

Benefits_data:

	Benefit	Number of Claims	Amount of Paid Claims	Amount of Claims VAT	Notes	Policy Year	End Date	Class	Overall Limit
0	1.Outpatient	44	37035.25	42590.53	20	Last Policy Year	2023-02-16	B	1000000
1	2.OP Lab & Diagnostics	17	23346.47	26848.44	15	Last Policy Year	2023-02-16	B	1000000
2	3.OP Consultation	10	2040.89	2347.02	No info	Last Policy Year	2023-02-16	B	1000000
3	4.OP Pharmacy	17	11867.51	13647.64	No info	Last Policy Year	2023-02-16	B	1000000
4	5.Dental	1	385	442.75	No info	Last Policy Year	2023-02-16	B	1000000
5	6.Optical	1	334	384.1	No info	Last Policy Year	2023-02-16	B	1000000
6	7.Maternity	1	875	1006.25	Yes	Last Policy Year	2023-02-16	B	1000000
7	Overall	47	75884.12	87266.74	No info	Last Policy Year	2023-02-16	B	1000000