Tesseract TSV format

Tom Rochette <tom.rochette@coreteks.org>

November 2, 2024 — 36c8eb68

Tesseract (an open source OCR engine) supports a TSV format as output. I looked online for some documentation about the columns but couldn't find anything, so I looked at the source code.

Here is a summary description of each column, what they represent, and the range of valid values they can have.

- **level:** hierarchical layout (a word is in a line, which is in a paragraph, which is in a block, which is in a page), a value from 1 to 5
  - 1: page
  - 2: block
  - 3: paragraph
  - 4: line
  - 5: word
- **page_num:** when provided with a list of images, indicates the number of the file, when provided with a multi-pages document, indicates the page number, starting from 1
- **block_num:** block number within the page, starting from 0
- **par_num:** paragraph number within the block, starting from 0
- **line_num:** line number within the paragraph, starting from 0
- **word_num:** word number within the line, starting from 0
- **left:** x coordinate in pixels of the text bounding box top left corner, starting from the left of the image
- **top:** y coordinate in pixels of the text bounding box top left corner, starting from the top of the image
- **width:** width of the text bounding box in pixels
- **height:** height of the text bounding box in pixels
- **conf:** confidence value, from 0 (no confidence) to 100 (maximum confidence), -1 for all level except 5
- **text:** detected text, empty for all levels except 5

Here is an example of the TSV format output, for reference.

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
1	1	0	0	0	0	0	0	1024	800	-1	
2	1	1	0	0	0	98	66	821	596	-1	
3	1	1	1	0	0	98	66	821	596	-1	
4	1	1	1	1	0	105	66	719	48	-1	
5	1	1	1	1	1	105	66	74	32	90	The
5	1	1	1	1	2	205	67	143	40	87	(quick)
5	1	1	1	1	3	376	69	153	41	89	[brown]
5	1	1	1	1	4	559	71	105	40	89	{fox}
5	1	1	1	1	5	687	73	137	41	89	jumps!
4	1	1	1	2	0	104	115	784	51	-	
5	1	1	1	2	1	104	115	96	33	91	Over
5	1	1	1	2	2	224	117	60	32	89	the
5	1	1	1	2	3	310	117	224	39	88	$43,456.78
5	1	1	1	2	4	561	121	136	42	92	<lazy>

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
5	1	1	1	2	5	722	123	70	32	92	#90
5	1	1	1	2	6	818	125	70	41	89	dog

1 References

- https://tesseract-ocr.github.io/tessdoc/Command-Line-Usage.html