University of Stuttgart
Institute of Industrial Automation
and Software Engineering

Development of a data parser for extracting information from technical documents with text mining methods

Jiayi He

Electrical Engineering

Supervisor: Yuchen Xia

Agenda

- Motivation and Difficulties
- Basis
- Conception
- Implementation
- Evaluation
- Summary and Outlook

[1]

Motivation

Technical Documents in Modern Automation Industry

[2]

[3]

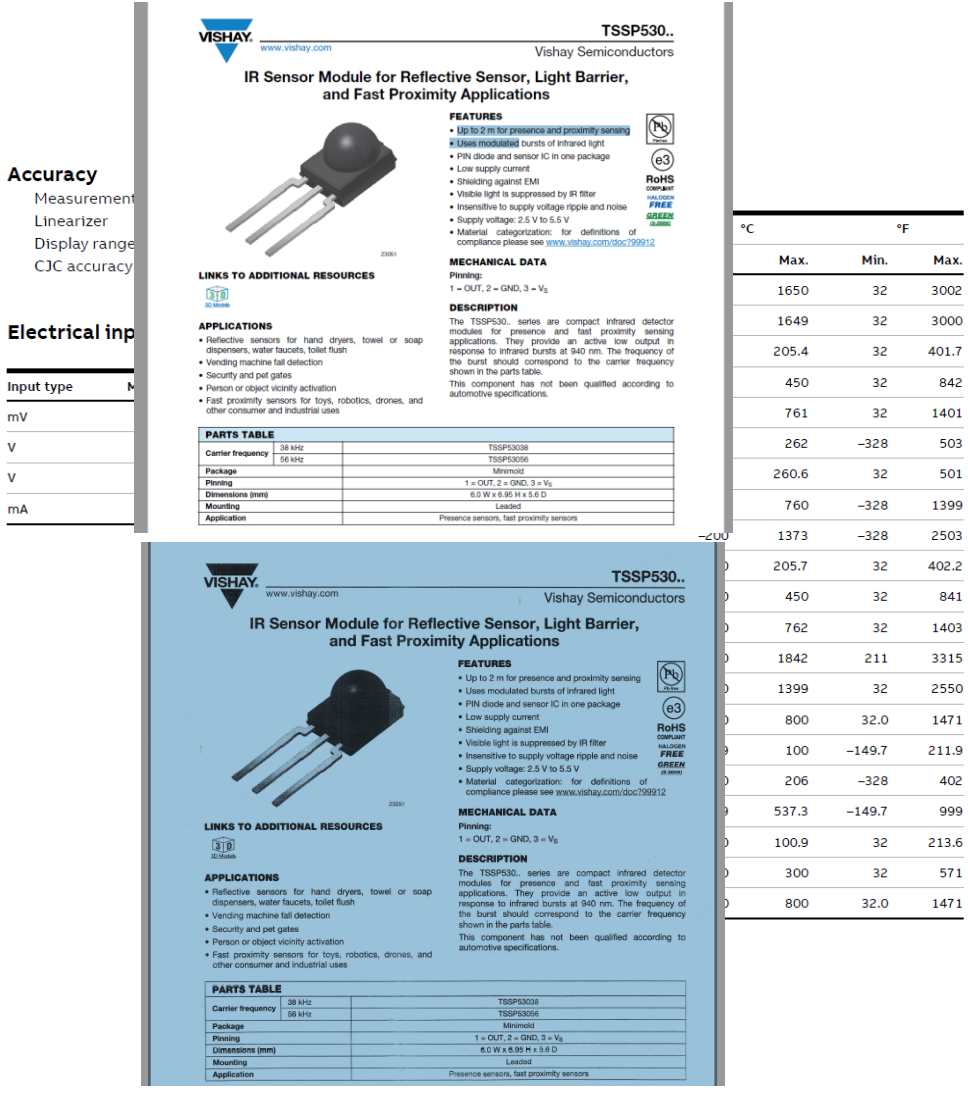
- Equipment maintenance
- Quality control
- Supply chain management
- Regulatory compliance

Defect of Manual Data Extraction & Entry:

- Time-Consuming: large number of documents
- Error-Prone: manual copying and typing
- Inconsistent: inconsistent formats from different staffs in one project

Difficulties

- PDF file format variations: text-based, scanned documents.
- Layout complexity: multi-column, complex tables, embedded images.
- Large size: multiple tables in pages, time-consuming.



Accuracy
Measurement
Linearizer
Display range
CJC accuracy

Electrical inp

Input type	M
mV	
V	
V	
mA	

VISHAY
www.vishay.com
TSSP530..
Vishay Semiconductors

IR Sensor Module for Reflective Sensor, Light Barrier, and Fast Proximity Applications

FEATURES
- Up to 2 m for presence and proximity sensing
- Uses modulated bursts of infrared light
- PIN diode and sensor IC in one package
- Low supply current
- Shielding against EMI
- Visible light is suppressed by IR filter
- Insensitive to supply voltage ripple and noise
- Supply voltage: 2.5 V to 5.5 V
- Material categorization: for definitions of compliance please see www.vishay.com/doc?99912

RoHS COMPLIANT HALOGEN FREE GREEN

LINKS TO ADDITIONAL RESOURCES

APPLICATIONS
- Reflective sensors for hand dryers, towel or soap dispensers, water faucets, toilet flush
- Vending machine fall detection
- Security and pet gates
- Person or object vicinity activation
- Fast proximity sensors for toys, robotics, drones, and other consumer and industrial uses

MECHANICAL DATA
Pinning:
1 = OUT, 2 = GND, 3 = V_S

DESCRIPTION
The TSSP530.. series are compact infrared detector modules for presence and fast proximity sensing applications. They provide an active low output in response to infrared bursts at 940 nm. The frequency of the burst should correspond to the carrier frequency shown in the parts table.
This component has not been qualified according to automotive specifications.

PARTS TABLE

Carrier frequency	38 kHz	TSSP53038
	56 kHz	TSSP53056
Package		Minimold
Pinning		1 = OUT, 2 = GND, 3 = V_S
Dimensions (mm)		6.0 W x 6.95 H x 5.6 D
Mounting		Leaded
Application		Presence sensors, fast proximity sensors

°C		°F	
	Max.	Min.	Max.
	1650	32	3002
	1649	32	3000
	205.4	32	401.7
	450	32	842
	761	32	1401
	262	−328	503
	260.6	32	501
	760	−328	1399
	1373	−328	2503
0	205.7	32	402.2
0	450	32	841
0	762	32	1403
0	1842	211	3315
0	1399	32	2550
0	800	32.0	1471
9	100	−149.7	211.9
0	206	−328	402
9	537.3	−149.7	999
0	100.9	32	213.6
0	300	32	571
0	800	32.0	1471

VISHAY
www.vishay.com
TSSP530..
Vishay Semiconductors

IR Sensor Module for Reflective Sensor, Light Barrier, and Fast Proximity Applications

FEATURES
- Up to 2 m for presence and proximity sensing
- Uses modulated bursts of infrared light
- PIN diode and sensor IC in one package
- Low supply current
- Shielding against EMI
- Visible light is suppressed by IR filter
- Insensitive to supply voltage ripple and noise
- Supply voltage: 2.5 V to 5.5 V
- Material categorization: for definitions of compliance please see www.vishay.com/doc?99912

RoHS COMPLIANT HALOGEN FREE GREEN

LINKS TO ADDITIONAL RESOURCES

APPLICATIONS
- Reflective sensors for hand dryers, towel or soap dispensers, water faucets, toilet flush
- Vending machine fall detection
- Security and pet gates
- Person or object vicinity activation
- Fast proximity sensors for toys, robotics, drones, and other consumer and industrial uses

MECHANICAL DATA
Pinning:
1 = OUT, 2 = GND, 3 = V_S

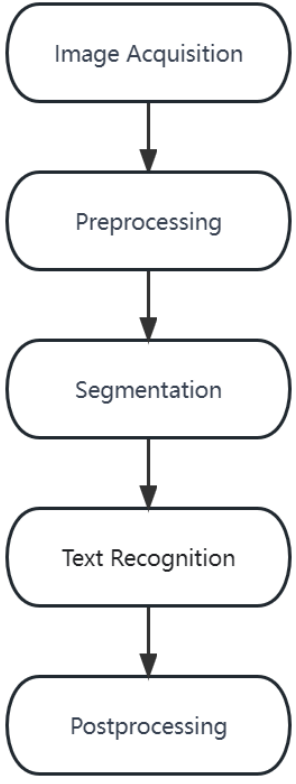
DESCRIPTION
The TSSP530.. series are compact infrared detector modules for presence and fast proximity sensing applications. They provide an active low output in response to infrared bursts at 940 nm. The frequency of the burst should correspond to the carrier frequency shown in the parts table.
This component has not been qualified according to automotive specifications.

PARTS TABLE

Carrier frequency	38 kHz	TSSP53038
	56 kHz	TSSP53056
Package		Minimold
Pinning		1 = OUT, 2 = GND, 3 = V_S
Dimensions (mm)		6.0 W x 6.95 H x 5.6 D
Mounting		Leaded
Application		Presence sensors, fast proximity sensors

Basis

Optical Character Recognition (OCR) Method



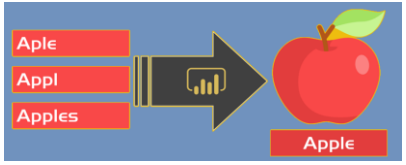
Pros:

- All types of PDF (text-based and image-based).
- High accuracy with pre-trained ML models.
- Time- and effort-saving.

Cons:

- Errors in special symbols.
- Limited supporting languages.
- Costly.

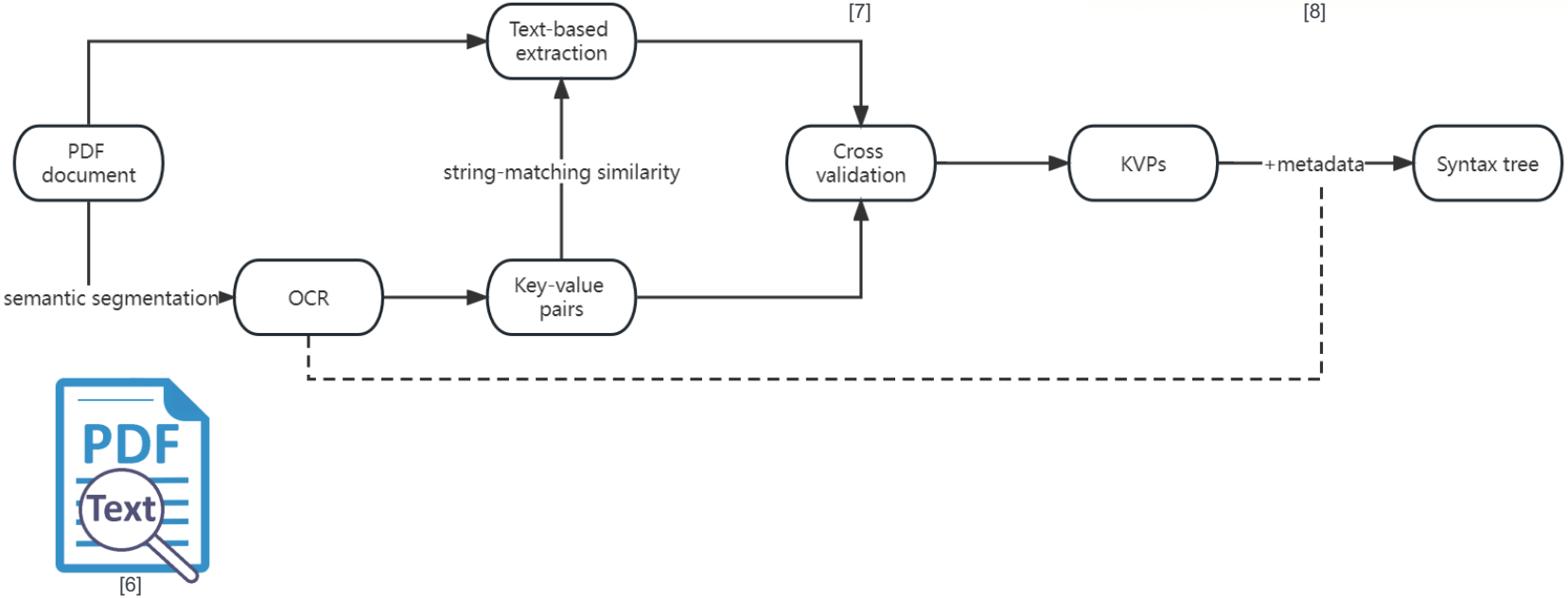
Data Parser Conception



[7]

```
{
  type: 'Program',
  body: [
    {
      type: 'ExpressionStatement',
      expression: { type: 'Literal', value: true, raw: 'true' },
    },
  ],
  sourceType: 'script',
}
```

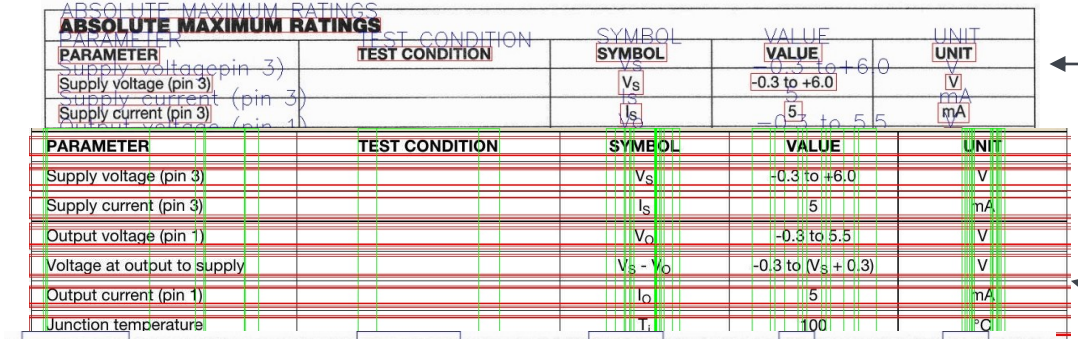
[8]



[6]

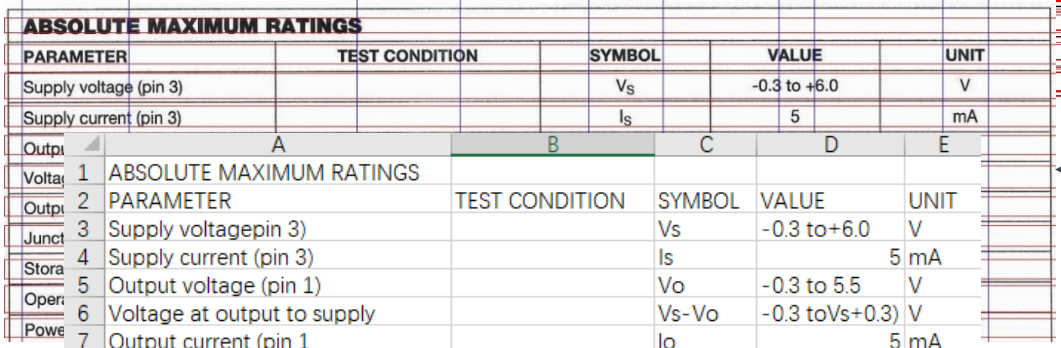
Implementation

OCR with Semantic Segmentation



ABSOLUTE MAXIMUM RATINGS

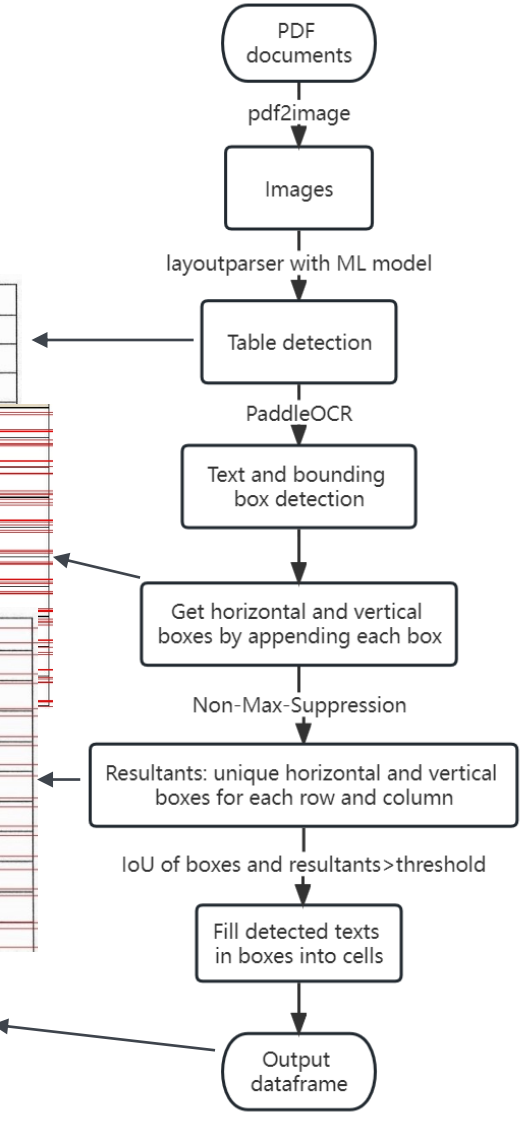
PARAMETER	TEST CONDITION	SYMBOL	VALUE	UNIT
Supply voltage (pin 3)		V_S	-0.3 to +6.0	V
Supply current (pin 3)		I_S	5	mA
Output voltage (pin 1)		V_O	-0.3 to 5.5	V
Voltage at output to supply		V_S - V_O	-0.3 to (V_S + 0.3)	V
Output current (pin 1)		I_O	5	mA
Junction temperature		T_j	100	°C



ABSOLUTE MAXIMUM RATINGS

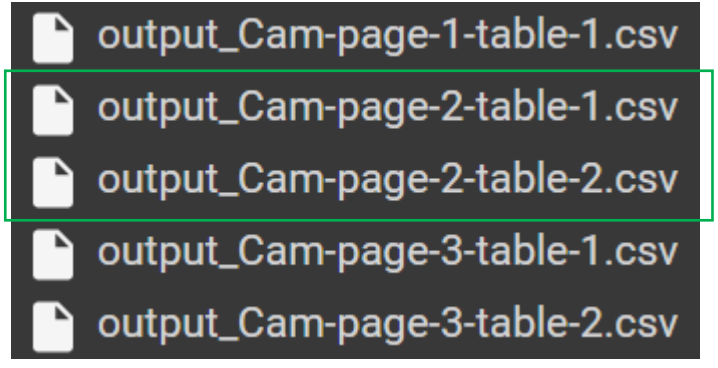
PARAMETER	TEST CONDITION	SYMBOL	VALUE	UNIT
Supply voltage (pin 3)		V_S	-0.3 to +6.0	V
Supply current (pin 3)		I_S	5	mA

	A	B	C	D	E
1	ABSOLUTE MAXIMUM RATINGS				
2	PARAMETER	TEST CONDITION	SYMBOL	VALUE	UNIT
3	Supply voltagepin 3)		Vs	-0.3 to+6.0	V
4	Supply current (pin 3)		Is	5	mA
5	Output voltage (pin 1)		Vo	-0.3 to 5.5	V
6	Voltage at output to supply		Vs-Vo	-0.3 toVs+0.3)	V
7	Output current (pin 1		Io	5	mA
8	Junction temperature		Ti	100	
9	Storage temperature range		Tstg	-25to+85	
10	Operating temperature range		Tamb	-25 to+85	
11	Power consumption	Tamb85C	Ptot	10	mW



Implementation

Text-Based Extraction with Camelot



	A	B	C	D	E
1	Column1	Column2	Column3	Column4	Column5
2					Vishay Semiconductors
3	ABSOLUTE MAXIMUM RATINGS				
4	PARAMETER	TEST CONDITION	SYMBOL	VALUE	UNIT
5	Supply voltage (pin 3)		VS	-0.3 to +6.0	V
6	Supply current (pin 3)		IS	5	mA
7	Output voltage (pin 1)		VO	-0.3 to 5.5	V
8	Voltage at output to supply		VS - VO	-0.3 to (VS + 0.3)	V
9	Output current (pin 1)		IO	5	mA
10	Junction temperature		Tj	100	°C
11	Storage temperature range		Tstg	-25 to +85	°C
12	Operating temperature range		Tamb	-25 to +85	°C
13	Power consumption	Tamb ≤ 85 °C	Ptot	10	mW
14	Note				

Misrecognizes other objects as tables

Performances better in recognizing special symbols

Implementation

Cross Validation using Fuzzy Match

	A
1	0
2	VPF44.50F15
3	
4	VPF54.50F15
5	VPF44.50F25
6	
7	VPF54.50F25
8	VPF44.65F25
9	VPF54.65F25
10	VPF44.65F35
11	VPF54.65F35
12	VPF44.80F35
13	
14	VPF54.80F35
15	VPF44.80F45
16	
17	VPF54.80F45
18	VPF44.100F70
19	
20	VPF54.100F70
21	VPF44.100F90
22	
23	VPF54.100F90
24	
25	
26	

Camelot

	A	B
1		0
2	0	
3	1	
4	2	VPF44.50F15
5	3	
6	4	VPF54.50F15
7	5	VPF44.50F25
8	6	
9	7	VPF54.50F25
10	8	VPF44.65F25
11	9	VPF54.65F25
12	10	VPF44.65F35
13	11	VPF54.65F35
14	12	VPF44.80F35
15	13	
16	14	VPF54.80F35
17	15	VPF44.80F45
18	16	
19	17	VPF54.80F45
20	18	VPF44.100F70
21	19	
22	20	VPF54.100F70
23	21	VPF44.100F90
24	22	
25	23	VPF54.100F9O
26		

OCR

0	fuzzy match	similarity score
VPF44.50F15	VPF44.50F15	100
VPF54.50F15	VPF54.50F15	100
VPF44.50F25	VPF44.50F25	100

	A	B	C
1	OCR_Column4	fuzzy match	similarity score
2	VALUE	VALUE	100
3	-0.3 to+6.0	-0.3 to +6.0	100
4	5	5	100
5	-0.3 to 5.5	5	100
6	-0.3 toVs+0.3)	-0.3 to (VS + 0.3)	94
7	5	5	100
8	100	100	100
9	-25to+85	-25 to +85	93
10	-25 to+85	-25 to +85	100
11	10	10	100

VPF54.100F9O	VPF54.100F90	92

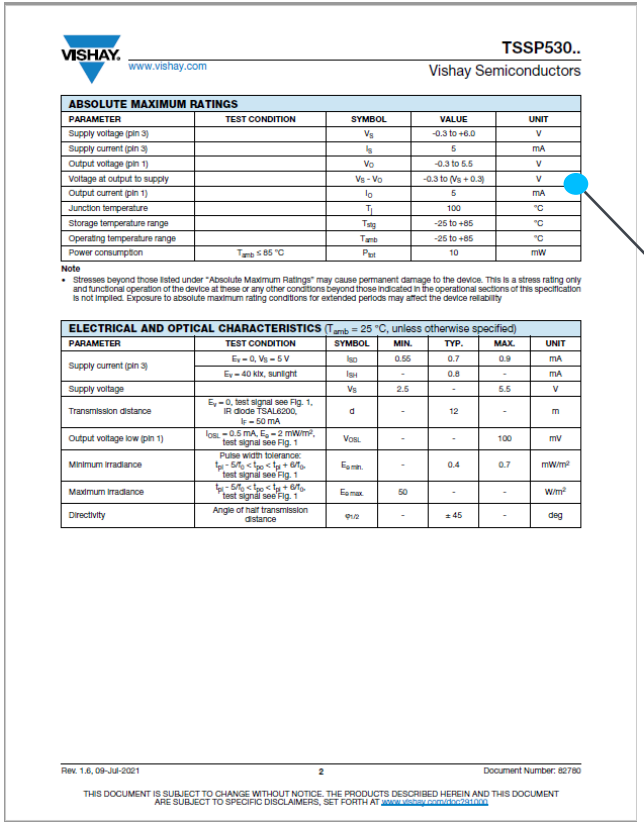
Fuzzy match correction

- ✓ After fuzzy match, ‘O’ in OCR result is corrected to ‘0’.
- ✓ White spaces and some symbols can be corrected.

Implementation

Integration into Syntax Tree

A hierarchical data structure that represents the syntactic structure of text in a tree-like form.



VISHAY. www.vishay.com

TSSP530..

Vishay Semiconductors

ABSOLUTE MAXIMUM RATINGS

PARAMETER	TEST CONDITION	SYMBOL	VALUE	UNIT
Supply voltage (pin 3)		V_S	-0.3 to +6.0	V
Supply current (pin 3)		I_S	5	mA
Output voltage (pin 1)		V_O	-0.3 to 5.5	V
Voltage at output to supply		$V_S - V_O$	-0.3 to (V_S + 0.3)	V
Output current (pin 1)		I_O	5	mA
Junction temperature		T_j	100	°C
Storage temperature range		T_{stg}	-25 to +85	°C
Operating temperature range		T_{amb}	-25 to +85	°C
Power consumption	$T_{amb} \leq 85$ °C	P_{tot}	10	mW

Note

- Stresses beyond those listed under "Absolute Maximum Ratings" may cause permanent damage to the device. This is a stress rating only and functional operation of the device at these or any other conditions beyond those indicated in the operational sections of this specification is not implied. Exposure to absolute maximum rating conditions for extended periods may affect the device reliability

ELECTRICAL AND OPTICAL CHARACTERISTICS (T_{amb} = 25 °C, unless otherwise specified)

PARAMETER	TEST CONDITION	SYMBOL	MIN.	TYP.	MAX.	UNIT
Supply current (pin 3)	E_v = 0, V_S = 5 V	I_{SD}	0.55	0.7	0.9	mA
	E_v = 40 klx, sunlight	I_{SH}	-	0.8	-	mA
Supply voltage		V_S	2.5	-	5.5	V
Transmission distance	E_v = 0, test signal see Fig. 1, IR diode TSAL6200, I_F = 50 mA	d	-	12	-	m
Output voltage low (pin 1)	I_{OSL} = 0.5 mA, E_e = 2 mW/m², test signal see Fig. 1	V_{OSL}	-	-	100	mV
Minimum irradiance	Pulse width tolerance: $t_{pi} - 5/f_0 < t_{po} < t_{pi} + 6/f_0$, test signal see Fig. 1	$E_{e\,min.}$	-	0.4	0.7	mW/m²
Maximum irradiance	$t_{pi} - 5/f_0 < t_{po} < t_{pi} + 6/f_0$, test signal see Fig. 1	$E_{e\,max.}$	50	-	-	W/m²
Directivity	Angle of half transmission distance	$\varphi_{1/2}$	-	± 45	-	deg

Rev. 1.6, 09-Jul-2021 2 Document Number: 82780

THIS DOCUMENT IS SUBJECT TO CHANGE WITHOUT NOTICE. THE PRODUCTS DESCRIBED HEREIN AND THIS DOCUMENT ARE SUBJECT TO SPECIFIC DISCLAIMERS, SET FORTH AT www.vishay.com/doc?91000


```
"3": {
    "0": "VALUE",
    "1": "-0.3 to +6.0",
    "2": "5",
    "3": "-0.3 to 5.5",
    "4": "-0.3 to (Vs+0.3)",
    "5": "5",
    "6": "100",
    "7": "-25 to +85",
    "8": "-25 to +85",
    "9": "10"
},
"4": {
    "0": "UNIT",
    "1": "V",
    "2": "mA",
    "3": "V",
    "4": "V",
    "5": "mA",
    "6": "°C",
    "7": "°C",
    "8": "°C",
    "9": "mW"
},
"location": {
    "0": "0.22",
    "1": 1,
    "2": null,
    "3": null,
    "4": null,
    "5": null,
    "6": null,
    "7": null,
    "8": null,
    "9": null
}
```

Data frame format: columns

```
"4": {
    "0": "Output voltage (pin 1)",
    "1": "",
    "2": "Vo",
    "3": "-0.3 to 5.5",
    "4": "V",
    "location": null
},
```

Data frame format: index

Evaluation

Table Detection

	Number of detected tables	Number of real tables	Extraction rate φ	Completeness χ
Sensor				
Document 1 (Temperature)	4	4	1.00	0.94
Document 2 (Reflective)	2	3	0.67	0.97
Actuator				
Document 3 (Camera)	1	1	1.00	1.00
Document 4 (Motor)	0	1	0.00	0.00
Controller				
Document 5 (Nebula)	1	1	1.00	1.00
Document 6 (DIN)	2	2	1.00	1.00

- Most tables are detected with high completeness.
- Some missing lines is tolerable.

Failed cases:

- No ruling lines at all.
- Objects around the table have similar structure.

Evaluation

Table Structure

	Total adjacency relations	Detected adjacency relations	Correct adjacency relations	Recall β	Precision α	F1-score
Document 1						
Table 1	94	92	68	0.723	0.739	0.731
Table 2	52	52	52	1.000	1.000	1.000
Table 3	31	39	26	0.839	0.667	0.743
Table 4	108	110	100	0.926	0.909	0.918
Document 2						
Table 1	-	-	-	-	-	-
Table 2	85	85	85	1.000	1.000	1.000
Table 3	109	127	94	0.862	0.740	0.797
Document 3						
Table 1	58	58	58	1.000	1.000	1.000
Document 4						
-	-	-	-	-	-	-
Document 5						
Table 1	31	33	29	0.935	0.879	0.906
Document 6						
Table 1	186	167	137	0.737	0.820	0.776
Table 2	40	40	40	1.000	1.000	1.000

tp0rts(rear
ont,local management only)
nection of a mouse or keyboard or combination of both
ge,4 opto-isolated24V sourced) user assignable
A,7user assignable
via optionalQS873device)
creen and traditional tactile navigation buttons
EU;EN55032,(CISPR22)Class B,EN55035(CISPR24)

ns in Predicted Table

- Almost no error with a neat structure.

Failed cases:

- Merging rows/columns.
- Multiple lines in one cell.

Supply current (pin 3)	E_v = 0, V_S = 5 V
	E_v = 40 klx, sunlight
Supply voltage	
Transmission distance	E_v = 0, test signal see Fig. 1, IR diode TSAL6200, I_F = 50 mA

Evaluation

Table Content

	Levenshtein distance	Number of true characters	Error rate	Number of corrected strings	Error rate after correction
Sensor					
Document 1 (Temperature)	74	1488	5.0%	11	4.2%
Document 2 (Reflective)	71	852	8.3%	9	7.3%
Actuator					
Document 3 (Camera)	103	720	14.3%	10	12.9%
Document 4 (Motor)	-	-	-	-	-
Controller					
Document 5 (Nebula)	5	637	0.8%	2	0.3%
Document 6 (DIN)	3	449	0.7%	2	0.2%

Main Errors:

- Missing information.
- Special symbols such as ° , ≤.

Maximum irradiance	$t_{pi} - 5/f_0 < t_{po} < t_{pi} + 6/f_0$, test signal see Fig. 1

Maximum irradiance	tpi-5/fo<tpo<tpi+6/fo

Summary and Outlook

Functions:

- Data parsing from text-based and image-based PDFs
- Automatically key-value-pairs extraction from tables
- Loop through each page of PDF
- Well-structured multiple output formats available

Deficiencies:

- Relatively low accuracy for complex table
- Manual error correction
- Limited metadata provided

Outlook:

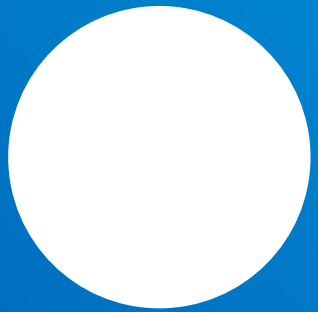
- NER and sematic analysis
- Store in database for user's query

Quelle

1. https://www.tibco.com/reference-center/what-is-data-parsing
2. https://www.boschmanufacturingsolutions.com/
3. https://www.dvz.de/rubriken/logistik/detail/news/industrie-40-bosch-schafft-einheit-fuer-software-und-services.html
4. G. Endignoux, O. Levillain, J.Y. Migeon, “Caradoc: a pragmatic approach to PDF parsing and validation” in IEEE Security and Privacy Workshops, 2016, pp.126-139.
5. https://products.aspose.app/ocr/de/pdf-ocr
6. https://www.imranabdullah.com/2021-09-17/Fuzzy-word-replace-from-string-in-Python
7. https://www.twilio.com/blog/abstract-syntax-trees

University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!



Jiayi He

e-mail st176731@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart

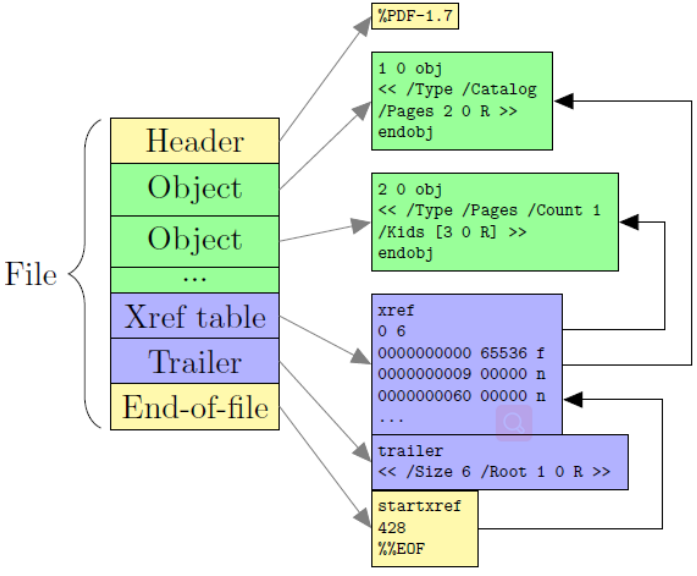
Basis

Text Mining

- Preprocessing: filtering, stemming, lemmatizing.
- Text mining methods: NLP, Named Entity Recognition, information extraction.
- Text analysis: semantic analysis.
- Discovery of knowledge: stored in knowledge database.

PDF Syntax-Postscript

A page description language, provides internal structure of documents.



[5]

Pros:

- Preserves the file formatting.
- Metadata accessible.

Cons:

- Requires technical expertise to work with.
- Limited supporting tools and researches.

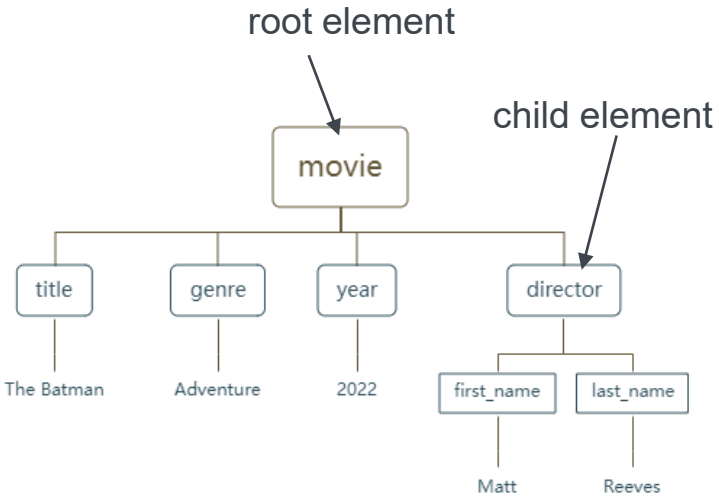
Makeup Language Conversion

Pros:

- Structured data, helps to identify different elements of the PDF document.
- Use tags and attributes to define elements, accurately extract symbols and units.

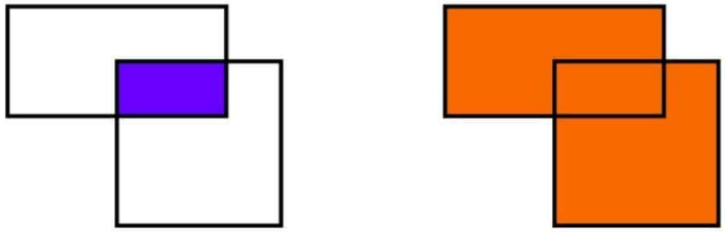
Cons:

- Conversion can be difficult. (especially for image-based PDF)
- Increased system complexity.
- Data loss.




```
<movie lang="English">
        <title>The Batman</title>
        <genre>Adventure</genre>
        ...
</movie>
```

NMS



- Firstly, the boxes whose confidence are smaller than a threshold should be removed.
- Secondly, the box with a lower confidence among the boxes which overlap too much with each other (IoU > 0.1) should also be removed.
- At last, the process is repeated until only one unique box is left for each class, which represents the final prediction.