Look, i can read now !

Koen Decorte

CD Invest

CD-Invest - Introduction

- International IBM i ISV and IBM business partner located in Antwerp, Belgium and Madrid, Spain
- Working with IBM i and its predecessors for more than 40 years
- Applications : CDQuery, CDSecure, CDView, CDLightning, CDReport, CDAccount and CDERP
- Expertise in RPG, SQL, Python, PHP, HTML, Unity, nodejs, linux, …
- Website : www.cdinvest.eu
- Member of CEAC since 2018
- 6 IBM Champions in the company
- What others talk about, we do.

CD INVEST
BI · AI · IBMi

CD-Invest - Some of our customers

Optimco
verzekeren op mensenmaat

Santander

winsol

Diners Club INTERNATIONAL®

KBC

CRAS WOODGROUP

JORI® 1963

DL CHEMICALS
Since 1936
Manufacturer of Sealants

CD INVEST
BI · AI · IBMi

CD-Invest - IBM i Client Stories

Deknudt Frames
Building the framework for a thriving e-commerce operation with IBM i

JORI
Increasing Manufacturing Efficiency During COVID-19 With IBM I and advanced 3D-configurator

Wijnen Van Maele
Tracking wine production with blockchain on IBM i

ID-Logistics
Meeting the Challenges of a Pandemic with IBM i in the Cloud

Diners Club Spain
Streamlining Customer Support with a Hybrid Cloud Application and IBM i

Optimco
Introducing AI and a new customer experience in the car insurance industry on IBM i

CD-Invest - IBM i Client Stories

Fibrocit
Providing a comfortable seat with IBM i

Cras Woodgroup
Modernizing the wood industry with IBM i

Oris
Making vacations easier with IBM i

Steffimmo
Moving to IBM i on POWER9 in the cloud for growth

Stonetales properties
Upgrading and Centralizing on the Cloud with IBM i

Winsol
Digitizing manufacturing on IBM i

CD-Invest - IBM i Client Stories

CSM
Empower more small businesses to access global trade

CSMe
Empowering real-estate since 1978

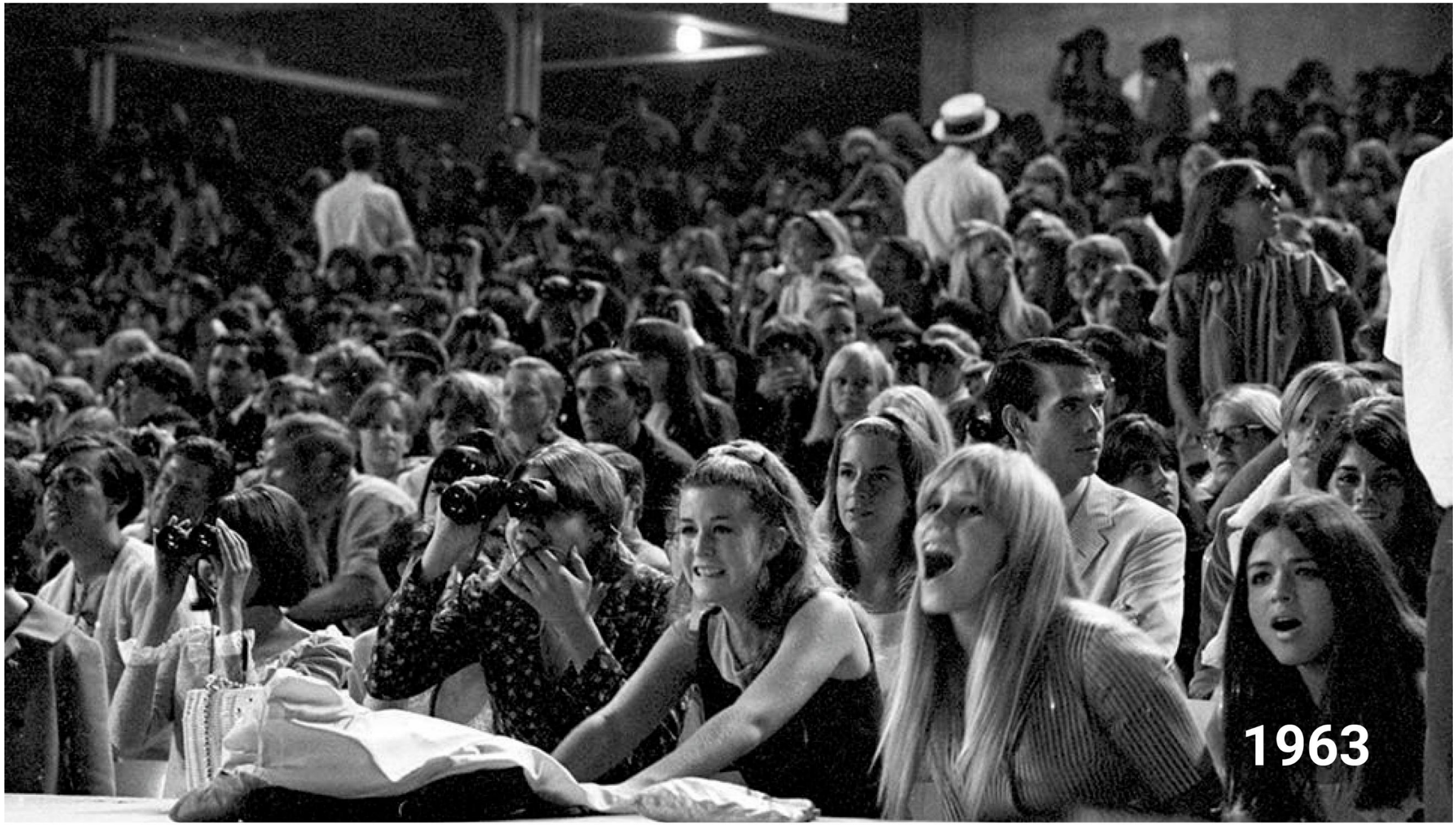
Bonehill
Adapting IBM i to the modern web

Immo Bonehill

Read more on on https://www.ibm.com/it-infrastructure/us-en/resources/power/ibm-i-customer-stories/

CD INVEST
BI · AI · IBMi

Introduction


1963


2015

Importance of pictures

- 3.2 billion pictures are shared online each day
- They contain brands, products, services, ads ...

  Interesting to find out if they are yours.
- 80% of those images don't mention brand or product in the accompanying text.

→ Insight missing and what about reading or understanding the text ??

But first things first

What is AI ?

Artificial Intelligence (AI)

= science of making computers do things that require intelligence when done by humans.

Above intelligent or Abysmal idiot ?

What is OCR?

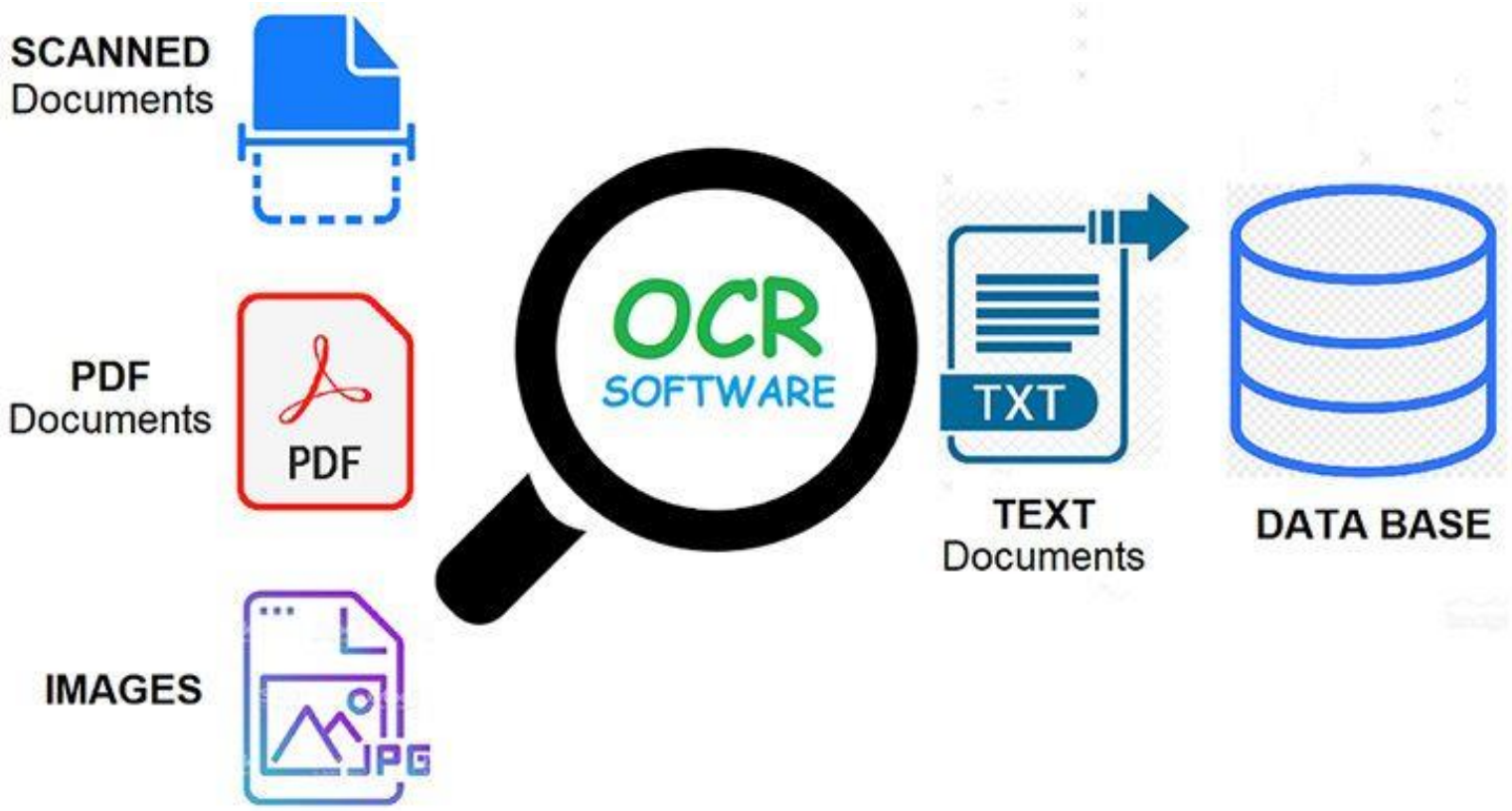
- Stands for Optical Character Recognition
- Extracts the text from a given image

My invention relates to statistical machines
of the type in which successive comparisons
are made between a character and a charac-

OCR

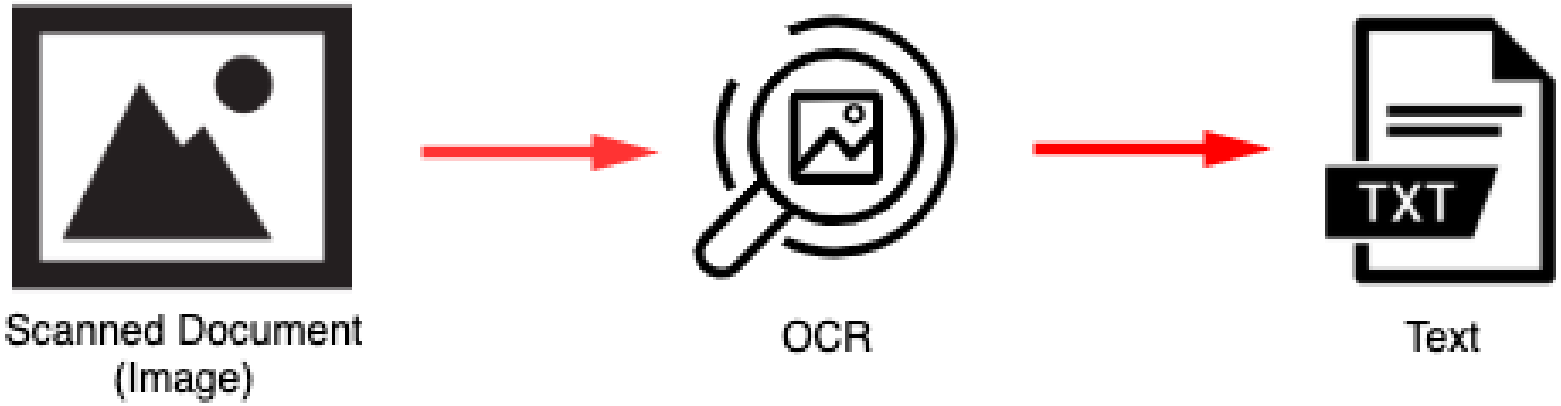
My invention relates to statistical machines
of the type in which successive comparisons
are made between a character and a charac-

What is OCR ?


SCANNED
Documents
PDF
Documents
PDF
IMAGES
JPG
OCR
SOFTWARE
TXT
TEXT
Documents
DATA BASE

What is OCR ?

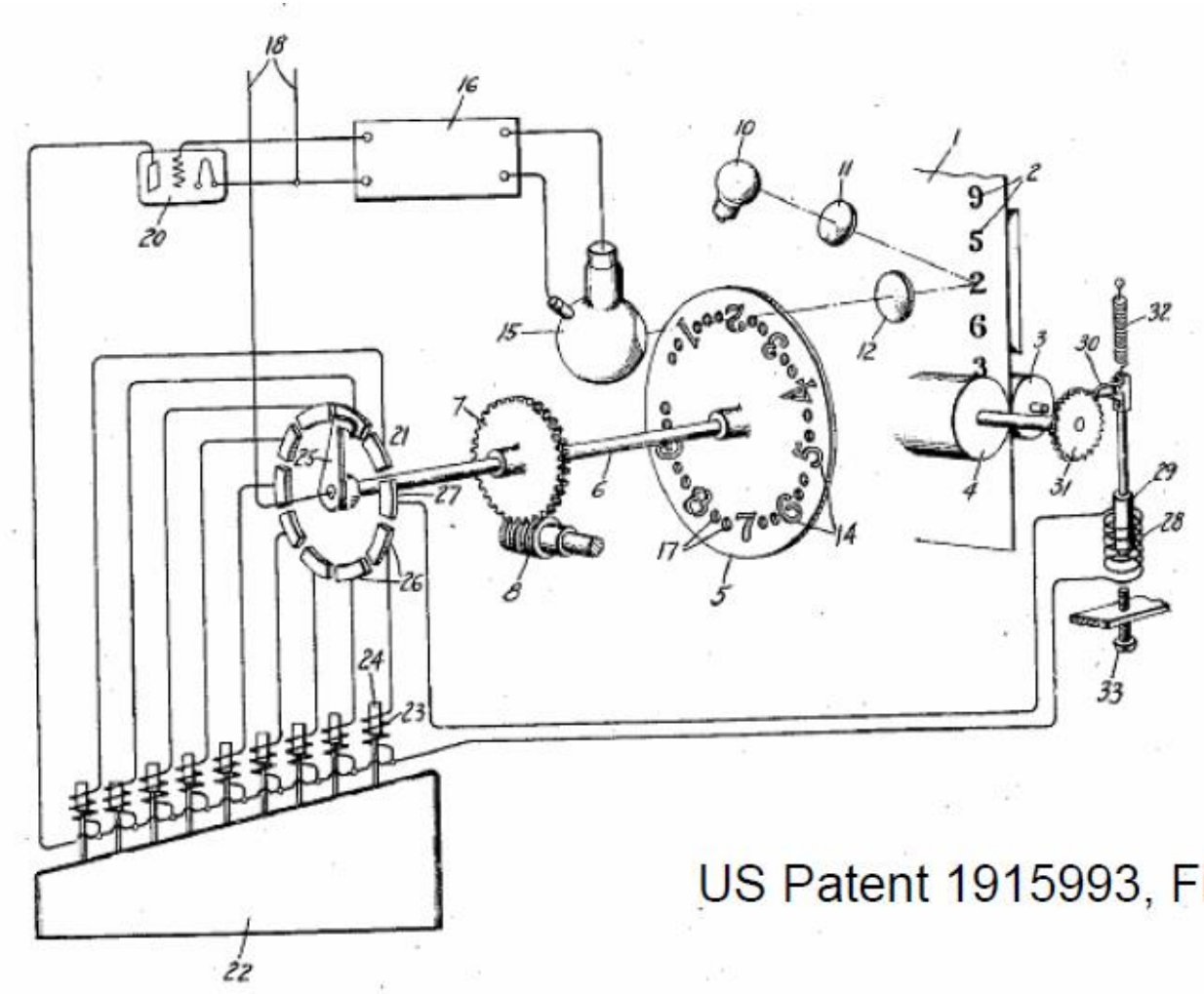
OCR predates electronic computers !!



What is OCR?

- Invented by Gustav Tauschek
- Tauschek obtained a patent on OCR
  - 1929 in Germany
  - 1935 in USA
- Tauschek's machine
  - Was a mechanical device
  - Uses templates, light and photodetector
  - When a light was directed towards the templates no light reach the photodetector

What is OCR?



US Patent 1915993, Filed Apr 27, 1931

OCR history

- 1929 – Digit recognition machine
- 1953 – Alphanumeric recognition machine
- 1965 – US mail sorting
- 1965 – British banking system
- 1976 – Kurzweil reading machine
- 1985 – Hardware assisted PC software
- 1988 – Software-only PC software
- 1994-2000 – Industry consolidation

What is OCR ?

- OCR Subprocesses
    - Preprocessing of the Image
    - Text Localization
    - Character Segmentation
    - Character Recognition
    - Post Processing

OCR use cases

- building license plate readers
- digitizing invoices
- reading container numbers
- digitizing ID cards
- digitizing letters, insurance documents, …

Project Tesseract

Project Tesseract

- History of Tesseract
  - Open source OCR engine
  - Developed by HP between 1985 and 1995
  - Never used in an HP product
  - Rated highly at The Fourth Annual Test of OCR Accuracy in 1995
  - In 2005 HP transferred Tesseract to the ISRI and released it as open source
    - ISRI == Information Science Research Institute
  - The development is currently led by Google (since 2006 !)

Project Tesseract

- Tesseract is an OCR Engine and is NOT a complete OCR program
  - Originally intended to serve as a component part of other programs
  - Works from the command line
  - Has no GUI
  - Integration to many programming languages
  - Runs on IBM i

Google Trends | Compare

Tesseract OCR
Search term

OCRopus
Search term

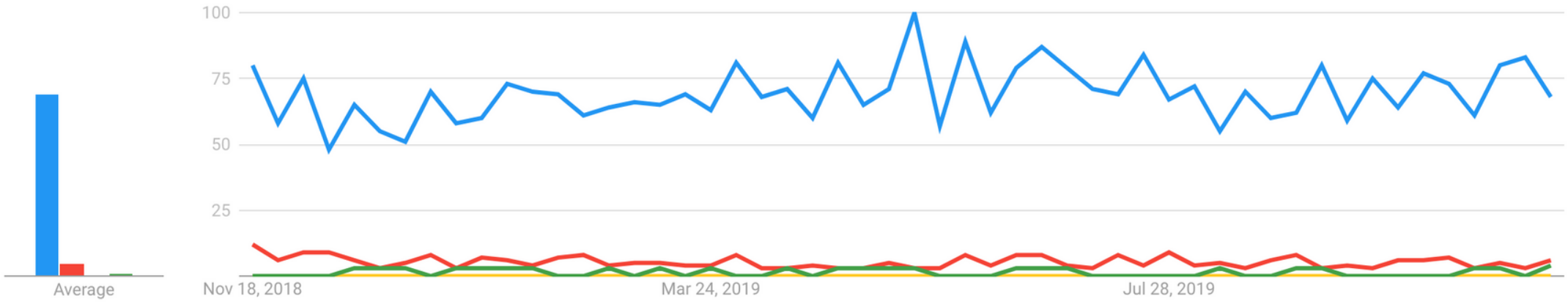
Ocular OCR
Search term

SwiftOCR
Search term

+

Worldwide | Past 12 months | All categories | Web Search

Interest over time



Install Tesseract

Install tesseract

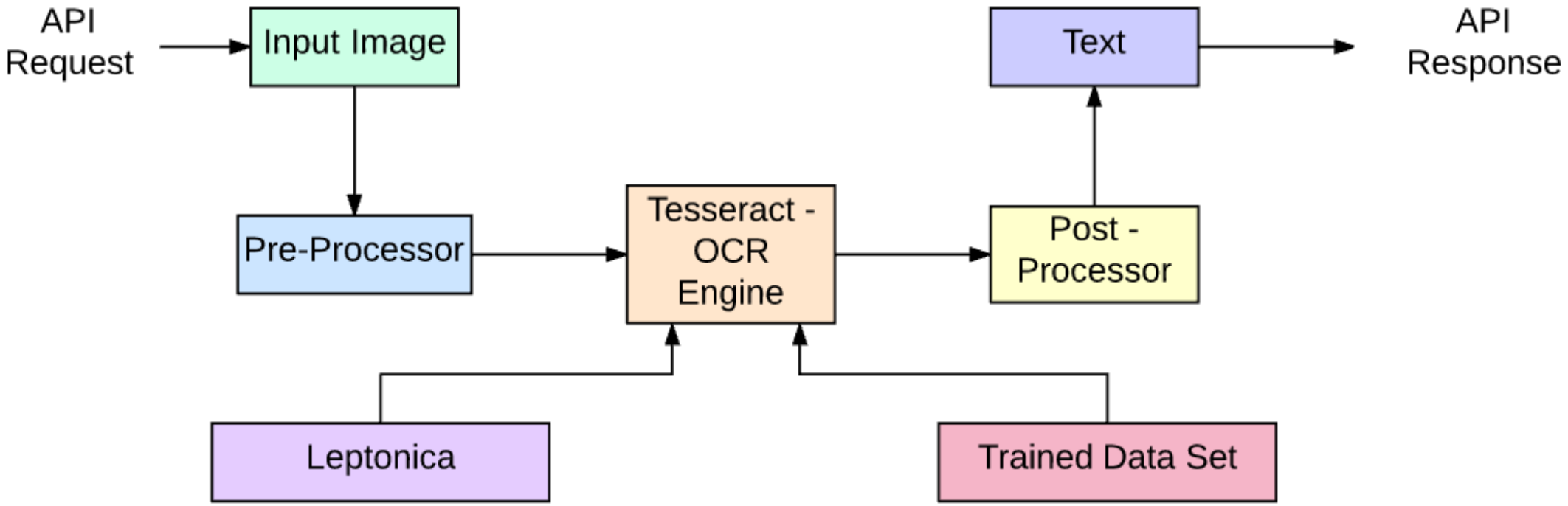
- Install from open source package manager (if needed I can provide compile scripts since I've given mine to IBM).

leptonica-devel	1.80.0-2	@ibmi-base
leptonica-tools	1.80.0-2	@ibmi-base

tesseract-devel	4.1.1-1	@ibmi-base
tesseract-tessdata	4.1.1-1	@ibmi-base
tesseract-tools	4.1.1-1	@ibmi-base

How does Tesseract work ?

OCR Process Flow



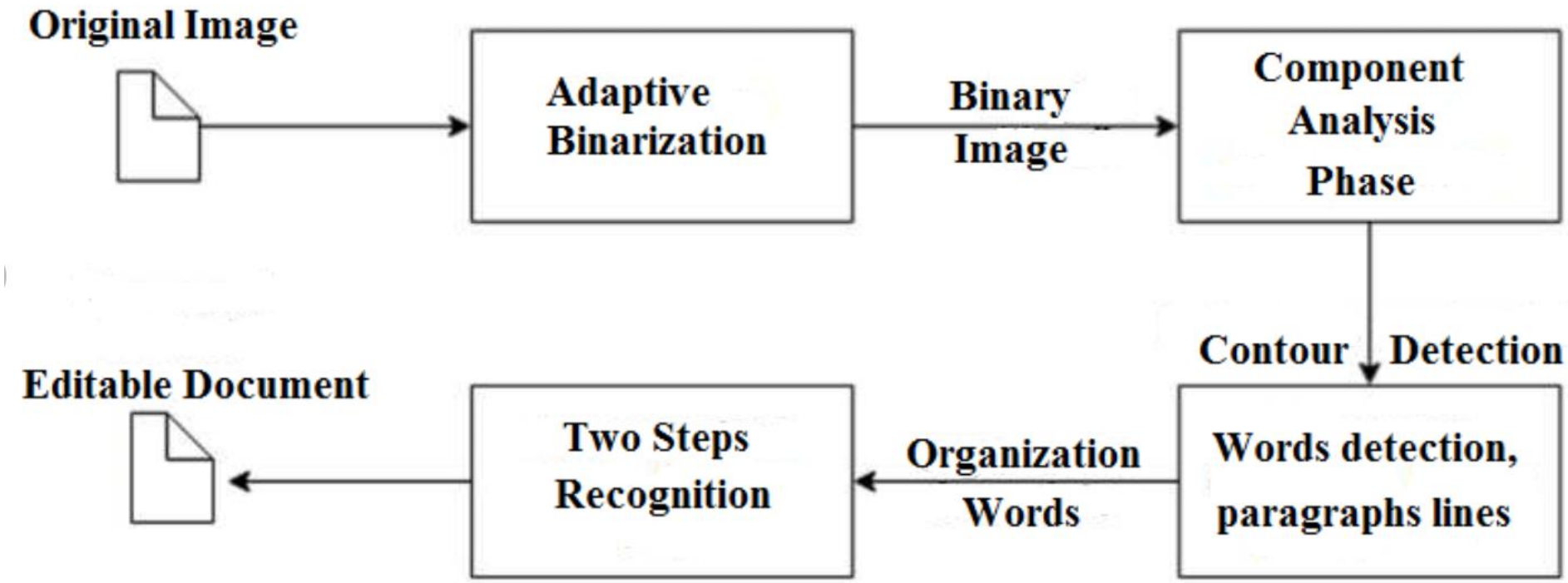
How does Tesseract work ?

- Tesseract 4.00 includes a new neural network subsystem configured as a text line recognizer.

- It has its origins in OCRopus Python LSTM implementation but has been redesigned for **Tesseract in C++. The neural network system in Tesseract pre-dates TensorFlow.**

- To recognize an image containing a single character, we typically use a Convolutional Neural Network (CNN). Text of arbitrary length is a sequence of characters, and such problems are solved using RNNs and LSTM (Long short-term memory) is a popular form of RNN (recurrent neural network).

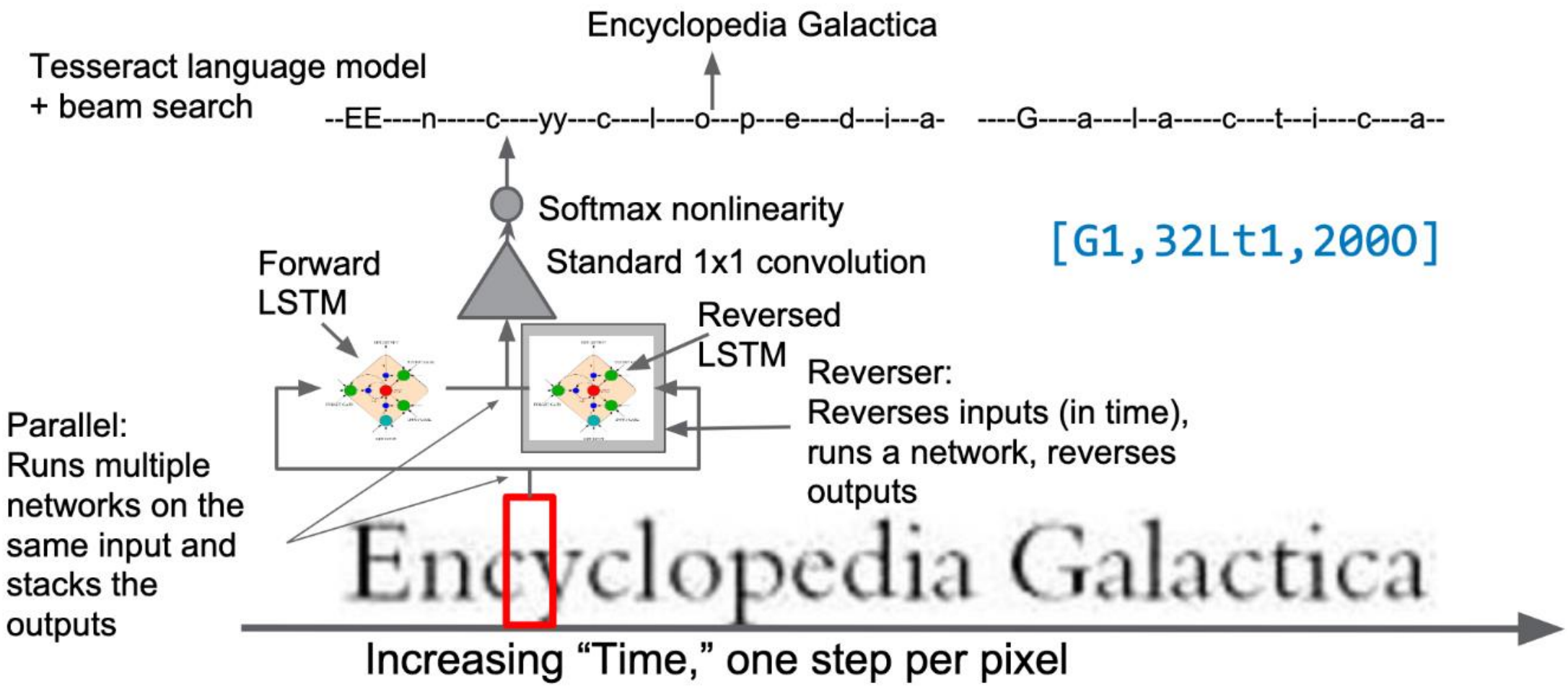
How does tesseract work ?



How Tesseract Works? (Old way and many traditional OCR

1. Adaptive thresholding on the input image
2. Analyze connected components in the binary image
3. Find text lines and words
4. First pass of recognition process
   - Attempts to recognize each word in turn
5. Satisfactory words are passed to adaptive trainer
6. Lessons learned are employed in a second pass
   - Used for words not satisfactory recognized
7. Producing the output text

How Tesseract uses LSTMs...



[G1,32Lt1,200O]

Running Tesseract ?

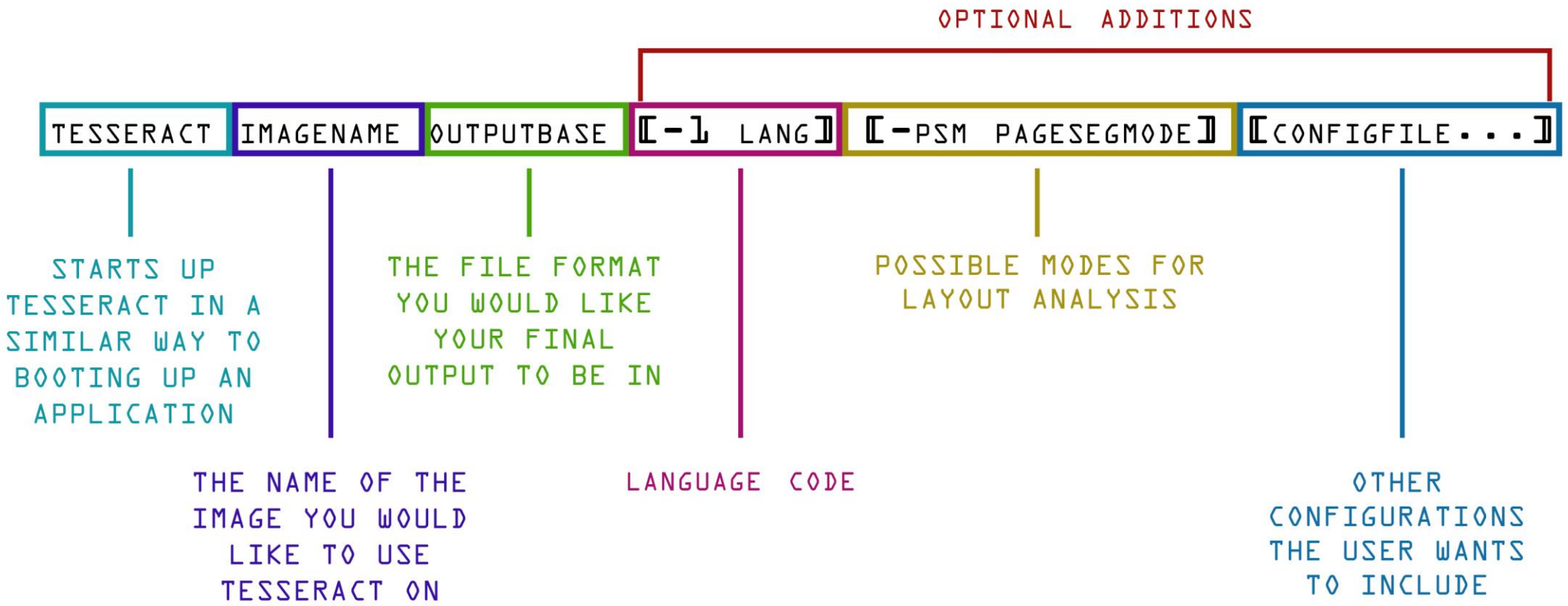
Running tesseract

```
$
> tesseract
Usage:
  /QOpenSys/pkgs/bin/tesseract --help | --help-extra | --version
  /QOpenSys/pkgs/bin/tesseract --list-langs
  /QOpenSys/pkgs/bin/tesseract imagename outputbase [options...] [configfile...]

OCR options:
  -l LANG[+LANG]        Specify language(s) used for OCR.
NOTE: These options must occur before any configfile.

Single options:
  --help                Show this help message.
  --help-extra          Show extra help for advanced users.
  --version             Show version information.
  --list-langs          List available languages for tesseract engine.
$
```

Running tesseract

```
> tesseract --version
  tesseract 4.1.1
   leptonica-1.80.0
    libgif 5.1.4 : libpng 1.6.37 : libtiff 4.0.9 : zlib 1.2.13 : libwebp 1.0.2
  $
```

Running Tesseract



Running Tesseract

$ tesseract image_path stdout

To write the output text in a file:

$ tesseract image_path text_result.txt

To specify the language model name, write language shortcut after -l flag, by default it takes English language:

$ tesseract image_path text_result.txt -l eng

Running tesseract

receipt-home-depot.jpg
receipt-mels.jpg
receipt-sunshine.jpg
receipt-tadish.jpg

Welcome to Mel's

Check #: 0001	12/20/11
Server: Josh F	4:38 PM
Table: 7/1	Guests: 2

---

2 Beef Burgr (@9.95/ea)	19.90
SIDE: Fries	
1 Bud Light	3.79
1 Bud	4.50

Sub-total	28.19
Sales Tax	2.50
TOTAL	30.69

Balance Due	30.69

Thank you for your patronage!

1425788830...depot.jpg.txt
1425788831...t-mels.jpg.txt
1425788834...shine.jpg.txt
1425788834...tadish.jpg.txt

```
welcome to MeT's

Check #: 0001 12/20/11
Server: Josh F 4:38 PM
Table: 7/1 Guests: 2
2 Beef Burgr (09.95/ea) 19.90
SIDE: Fries
1 Bud Light 3.79
1 Bud 4.50
Sub-totaT 28.19
SaTes Tax _____ __mg.§g
TOTAL 30.69
eai;r};;"13;; """""""""""""""""""""""""""""
"Tédféé

Thank you for your patronage!
```

Running Tesseract

```
> tesseract --list-langs
  List of available languages (6):
  bel
  deu
  eng
  fra
  ita
  spa
  $
```

Running Tesseract

```
$ tesseract image_path text_result.txt -l eng --psm 6
```

There is also one more important argument, OCR engine mode (oem). Tesseract 4 has two OCR engines — Legacy Tesseract engine and LSTM engine. There are four modes of operation chosen using the --oem option.

- 0 Legacy engine only.
- 1 Neural nets LSTM engine only.
- 2 Legacy + LSTM engines.
- 3 Default, based on what is available.

Running Tesseract

- Beware of memory / CPU limitations using LSTM ! It is CPU intensive.

- In addition, LSTM networks have pre-trained models that can be used in OCR enhancement. You can also adapt these models to the specific requirements of OCR with little effort.

- In practice, LSTM networks are mainly used to recognize handwritten texts and old documents with high accuracy - where OCR algorithms have problems.

Running Tesseract – supported files

- JPG
- PNG
- GIF
- PNM
- TIFF
- Unfortunately, Tesseract engine can’t read PDF file. (Use ghostscript or imagemagick to convert PDF to images).

Running Tesseract - Programming Language Package

- Gosseract (Golang)
- Tess4J (Java)
- RTesseract (Ruby)
- PyTesseract (Python)
- Tesseract.js (Javascript)
- Thiagoalessio TesseractOCR (PHP)
- etc…

Running Tesseract - Tesseract hOCR output format

- either text (txt) or hOCR, an html-format with embedded segmentation info
- bounding box: x0y0 x1y1
- <span class='ocrx_word' id='word_1_33' title='bbox 1584 1199 1997 1284; \
- x_wconf 87' lang='deu-frak' dir='ltr'>Verhältnisse.
- hOCR has word tokens (separated by white space) as smallest unit
- Ben Kiessling (Nidaba project), Kay Würzner have achieved character xml output

Running Tesseract


```
 <meta name='ocr-system' content='tesseract 4.1.1' />
 <meta name='ocr-capabilities' content='ocr_page ocr_carea ocr_par ocr_line ocrx_word ocrp_wconf'/>
</head>
<body>
 <div class='ocr_page' id='page_1' title='image "invoice-1.jpg"; bbox 0 0 595 842; ppageno 0'>
  <div class='ocr_carea' id='block_1_1' title="bbox 26 25 145 56">
   <p class='ocr_par' id='par_1_1' lang='fra' title="bbox 26 25 145 56">
    <span class='ocr_line' id='line_1_1' title="bbox 27 25 145 40; baseline 0.008 -1; x_size 19.26087; x_descenders 5.2608695; x_ascenders 3">
     <span class='ocrx_word' id='word_1_1' title='bbox 27 25 145 40; x_wconf 92'>Prestatiestaat</span>
    </span>
    <span class='ocr_line' id='line_1_2' title="bbox 26 47 63 56; baseline 0 0; x_size 22.5; x_descenders 5.5; x_ascenders 5.5">
     <span class='ocrx_word' id='word_1_2' title='bbox 26 47 63 56; x_wconf 54'>core</span>
    </span>
   </p>
  </div>
```

Running Tesseract – tsv ouput

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	595	842	-1	
2	1	1	0	0	0	26	25	119	31	-1	
3	1	1	1	0	0	26	25	119	31	-1	
4	1	1	1	1	0	27	25	118	15	-1	
5	1	1	1	1	1	27	25	118	15	92	Prestatiestaat
4	1	1	1	2	0	26	47	37	9	-1	
5	1	1	1	2	1	26	47	37	9	54	core
2	1	2	0	0	0	26	71	76	11	-1	
3	1	2	1	0	0	26	71	76	11	-1	
4	1	2	1	1	0	26	71	76	11	-1	
5	1	2	1	1	1	29	61	33	28	47	ringe
5	1	2	1	1	2	66	61	36	28	83	0172001
2	1	3	0	0	0	22	88	551	5	-1	
3	1	3	1	0	0	22	88	551	5	-1	
4	1	3	1	1	0	22	88	551	5	-1	
5	1	3	1	1	1	22	88	551	5	95	
2	1	4	0	0	0	31	103	125	8	-1	
3	1	4	1	0	0	31	103	125	8	-1	
4	1	4	1	1	0	31	103	125	8	-1	
5	1	4	1	1	1	31	103	52	8	25	Nedocunet:
5	1	4	1	1	2	99	103	57	8	42	2021010030
2	1	5	0	0	0	22	120	551	5	-1	
3	1	5	1	0	0	22	120	551	5	-1	
4	1	5	1	1	0	22	120	551	5	-1	
5	1	5	1	1	1	22	120	551	5	95	
2	1	6	0	0	0	22	133	551	5	-1	
3	1	6	1	0	0	22	133	551	5	-1	

Tesseract Training

- enable recognition of a new "language"
- rather, what is trained are glyph shapes:
  - new alphabet (Latin, Cyrillic, Greek etc.)
  - new typeface (Antiqua, Fraktur)
  - new font (special instance of a typeface, e.g. 12 pt Caslon italic)
- optionally add language data (wordlists)

Tesseract Training

- training data are in files of the form LANG.traineddata
- glyph shape training data and language support data (wordlists) are tied up in the same file
- language data can be exchanged without retraining (better and larger wordlists)

Tesseract accuracy

268702 Characters

5954 Errors

97.78% Accuracy

Errors	Marked	Correct-Generated
372	0	{}-{-}
246	0	{ü}-{ii}
228	0	{}-{ }
215	0	{I}-{J}
211	0	{v}-{V}
208	0	{}-{<\n>}
175	0	{u}-{n}
140	0	{n}-{u}
136	0	{c}-{e}

Tesseract guidelines

- Tesseract performs well when document images follow the next guidelines:

- Clean segmentation of the foreground text from background
- Horizontally aligned and scaled appropriately
- High-quality image without blurriness and noise

Tesseract guidelines

- The latest release of Tesseract 4.0 supports deep learning based OCR that is significantly more accurate. The OCR engine itself is built on a Long Short-Term Memory (LSTM) network, a kind of Recurrent Neural Network (RNN).

- Tesseract is perfect for scanning clean documents and comes with pretty high accuracy and font variability since its training was comprehensive. I would say that Tesseract is a go-to tool if your task is scanning of books, documents and printed text on a clean white background.

Application examples using Tesseract ?

Tesseract examples

- **Healthcare industry**

As a rule, hospitals and doctors' offices keep medical records in written form. In large quantities, these are therefore difficult to search. Tesseract can digitize these records, organize them - and thus make them easily searchable. Doctors and nurses can thus automatically analyze large volumes of medical records and extract important information. This leads to more efficient diagnosis and treatment of patients.

- **Finance**

Financial documents such as bank statements, Invoices and tax returns are still often created in writing. Searching these is therefore time-consuming. Tesseract can index and categorize these documents quickly and automatically. Banks can thus automatically read in checks, for example, and thus significantly reduce the manual workload.

Tesseract examples

- **Logistics**

In the logistics industry, it is important to be able to quickly access information such as package numbers, inventory figures and shipping addresses. Tesseract enables automatic recognition of product labels and Barcodes. This leads to faster and more accurate recording of inventories. In this way, companies in logistics can increase their efficiency and avoid bottlenecks in inventory management.

- **Mobile applications**

Tesseract can be embedded as a component in mobile apps to recognize text within images on mobile devices. This is particularly useful for applications such as translation and text recognition apps.

A - CDINVEST

File Edit View Communication Actions Window Help

```
Browse : /invoice2data/invoice2data.sh
Record :      18    of       46 by  18                    Column :    1     204 by 131
Control :
....+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....8....+....9....+....0....+....1....+....2....+....3.
    #Execute invoice2data inside chroot container

    chroot /QOpenSys/invoice2data invoice2data --input-reader pdftotext --template-folder ocr/templates ocr/pdfs/$invoice --output-f

    #Copy the result back to ifs

    cp /QOpenSys/invoice2data/ocr/json/$json /invoice2data/json/$json 2>>/invoice2data/log.txt

    #move the invoice to the done folder

    mv /invoice2data/invoices/$invoice /invoice2data/done/$invoice 2>>/invoice2data/log.txt

    #delete files inside de chroot container

    rm /QOpenSys/invoice2data/ocr/pdfs/$invoice 2>>/invoice2data/log.txt
    rm /QOpenSys/invoice2data/ocr/json/$json 2>>/invoice2data/log.txt

    echo "Done Processing ------------------------------------------------------------------------" >>log.txt
}

 F3=Exit   F10=Display Hex   F12=Cancel   F15=Services   F16=Repeat find   F19=Left   F20=Right
 1 records folded.
MA    A                  MW                                        03/012
```

192.168.2.41:23

Type here to search

19°C ENG 21:22 11/09/2023

A - CDINVEST

File Edit View Communication Actions Window Help

```
Browse : /invoice2data/json/coenen.json
Record :       1   of      13 by  18                    Column :   1     59 by 131
Control :
....+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....8....+....9....+....0....+....1....+....2....+....3.
 **********Beginning of data**************
[
    {
        "account": 70202,
        "amount": 2736.35,
        "baseamount": "2.299,45",
        "currency": "EUR",
        "date": "2022-02-24",
        "desc": "Invoice from Coenen Neuss GmbH & Co. KG",
        "invoice_number": "40201083/ 1",
        "issuer": "Coenen Neuss GmbH & Co. KG",
        "vat": "436,90"
    }
]
 **********End of Data********************

 F3=Exit   F10=Display Hex   F12=Cancel   F15=Services   F16=Repeat find   F19=Left   F20=Right
 1 records folded.
MA    A                MW                                        03/012
```

192.168.2.41:23

Type here to search

19°C ENG 21:22 11/09/2023

A - CDINVEST

File Edit View Communication Actions Window Help

```
Browse : /invoice2data/json/Kylma_2.json
Record :       1   of      30 by  18                    Column :   1     59 by 131
Control :
....+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....8....+....9....+....0....+....1....+....2....+....3.
 **********Beginning of data**************
[
    {
        "amount": 1032.62,
        "currency": "EUR",
        "date": "2023-03-14",
        "desc": "Invoice from Kylma AB",
        "invoice_number": "3213036-1",
        "issuer": "Kylma AB",
        "line": [
            [
                "916868",
                "2023-03-28",
                "1,00",
                "52,54"
            ],
            [
                "590240",
                "2023-03-28",

 F3=Exit   F10=Display Hex   F12=Cancel   F15=Services   F16=Repeat find   F19=Left   F20=Right
 1 records folded.
MA    A                MW                                        03/012
```

192.168.2.41:23

Type here to search

19°C ENG 21:23 11/09/2023

Dashboard | Aankoopfacturen

Ingave aankoopfacturen

Einde ingave | Bevestigen | PDF | Convert Pdf to Jso

Datum ingave: 11.09.2023

Journaal: AF - AANKOOPFACTUREN

DocumentNr: 2300202

Leverancier:

Factuurnummer:

Datum:

Vervaldatum:

Blokkeren: ☑ Betaald: ☐

Valutadatum: 11.09.2023

Te betalen bedrag: EUR

Btw: 4 - 21%

Soort:

Land: BE

Totaal zonder btw: 0

Btw aftrekbaar: 0

Btw niet aftrekbaar: 0

Rekening:

Te betalen btw: 0

Investering: 0

Disconto: 0

Vervaldatum:

Buiten maatstaf: 0

Kredietbeperking: 0

Te betalen leeggoed: 0

Debet: **EUR 0,00** Credit: **EUR 0,00**

Te betalen bedrag niet gelijk aan de som van de uitsplitsing

	K/L/R	Rekening	Rekening omschrijving	Kostenplaats	Kostenplaats omschrijving	Bedrag EUR	D/C	Btw vak	Omschrijving	Personeel	Personeel omschrijving
	L	440000				0,00	C				
✕	R					0,00	D				

Ingave aankoopfacturen

Einde ingave | Bevestigen | PDF

Datum ingave: 11.09.2023

Journaal: AF - AANKOOPFACTUREN DocumentNr: 2300202

Leverancier:

Factuurnummer:

Blokkeren:

Te betalen bedrag:

Totaal zonder btw:

Btw aftrekbaar:

Btw niet aftrekbaar:

Te betalen btw:

Investering:

Disconto:

Buiten maatstaf:

Kredietbeperking:

Te betalen leeggoed:

/accounting/scan/

	Filename	Creation date	Creation time
☐	20230824153541.pdf	24.08.2023	15.33.03
☐	20230906130652.pdf	06.09.2023	13.23.18
☑	820248_2023-08-15_FR_E-INVOICE_23-631304081-002.pdf	24.08.2023	15.35.37
☐	ZZZ_0000001.pdf	24.08.2023	15.24.40
☐	ZZZ_0000002.pdf	24.08.2023	15.26.26

Confirm | Cancel

etalen bedrag niet gelijk aan de s

K/L/R	Rekening	Rekening omschrijving	Kostenplaats	Kostenplaats omschrijving	Bedrag EUR	D/C	Btw vak	Omschrijving	Personeel
L	440000				0,00	C			
R					0,00	D			

Dashboard | Aankoopfacturen

Ingave aankoopfacturen

Einde ingave | Bevestigen | PDF | Convert Pdf to Json

Datum ingave: 11.09.2023

Journaal: AF - AANKOOPFACTUREN DocumentNr: 2300202

Leverancier: 48348 - DKV NIEDERLASSUNG DUESSELDORF - DUESSELDORF

Factuurnummer: 1234567 Datum: 11.09.2023 Vervaldatum: 11.09.2023

Blokkeren: ☑ Betaald: ☐ Valutadatum: 11.09.2023

Te betalen bedrag: 2100 EUR Btw: 4 - 20% Land: FR

Totaal zonder btw: 2100

Btw niet aftrekbaar: 420 Rekening: 411020 - TERUG TE VORDEREN BTW FRANKRIJK

Disconto: 0 Vervaldatum: 11.09.2023

Buiten maatstaf: 0

Kredietbeperking: 0

Te betalen leeggoed: 0

1 of 1 Automatic Zoom

DKV

E-FACTURE / E-REKENING

Pour livraisons et prestations en
Voor dienstverlening en leveringen in
France

Adresse du client/adresse de facturation
D.I.C. BVBA
Deknudt Invest en Consult
Charline Bourgois
Breestraat 31A
8540 DEERLIJK
BELGIEN

Numéro client / Klantnummer: 0000820248
numéro fiscal EU client/ No. fiscal national / Kl.-btw-id./Nat. bel.-nr.: BE0894259628
Numéro de facture / Factuurnummer: 23/631304081/002
Date de facturation / Factuurdatum: 15.08.2023
Page / Pagina: 1 / 1

TVA (%) : 20,00
Monnaie: EUR

Date de livraison / Leveringsdatum	nom / naam	Station service / lieu / Servicestation / plaats	numéro / nummer	Numéro de la transaction / Transactie-nummer	Heure / Tijd	Kilo-métrage / Kilometer-stand	Produit / Product	Unité / Eenheid	Quantité / Hoeveelheid	Prix/unité HT Prix/Eenheid brutto	Prix/unité HT Prix/Eenheid netto	Valeur de référence HT Referentie netto	Frais de service HT Servicelire netto	Valeur totale HT Totaalbedrag netto	TVA Btw.	Valeur totale TTC Totaalbedrag brutto
VOORTUIG: 1 GDH 693 CARD NO.: 7043110 0105208732																
10.08.2023 BÉGEN NORMANDIE	ESSO	BOLLEVILLE	1303647	69560119 16:47	2	EURO 95 (SUPER)	0036	LTR	24,990	2,0440	1,7035	42,57	1,17	43,74	8,75	52,49
•						TOTAL:			24,990			42,57	1,17	43,74	8,75	52,49
Bilan des sommes totales par type de prestation avec 20,00% TVA.																
EURO 95 (SUPER)							0036	LTR	24,990			42,57	1,17	43,74	8,75	52,49
•												42,57	1,17	43,74	8,75	52,49

Pas d'escompte en cas de règlement anticipé. Toute somme non réglée à échéance engendre des pénalités de retard à un taux égal au taux d'intérêt appliqué par la banque centrale européenne à son opération de financement la plus récente majorée de 8 points de pourcentage et le paiement d'une indemnité forfaitaire pour frais de recouvrement d'un montant de 40€ (article L441-6 Code de Commerce).

Statistiques TVA :			
20,00 %	43,74	8,75	52,49
•	43,74	8,75	52,49

Numéro de TVA DKV : FR754275448125

Debet: **EUR 840,00** Credit: **EUR 2.100,00** **Totaal debet niet gelijk aan totaal credit**

	K/L/R	Rekening	Rekening omschrijving	Kostenplaats	Kostenplaats omschrijving	Bedrag EUR	D/C	Btw vak	Omschrijving	Personeel	Personeel omschrijving
	L	440000	Leveranciers			2.100,00	C				
	R	411020	Terug te vorderen btw Frankrijk			420,00	D		DKV NIEDERLASSUNG DUESSELDORF		
	R	411003	Terug te vorderen btw België medecontractant			420,00	D		DKV NIEDERLASSUNG DUESSELDORF		

A - CDINVEST

File Edit View Communication Actions Window Help

```
 Browse : /tesseract-master/test/0200307048.txt
 Record :      36    of       63 by  18                    Column :    1     121 by 131
 Control :
....+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....8....+....9....+....0....+....1....+....2....+....3.
Art Code Art. Omschrijving Hoeveelheid Eenh  Prijs Korting Bedrag Verz.dat
Aluk BE is op 2/4, 1/5, 10/5, 11/5 en 21/5/2018 gesloten.
Ordernummer 118014973
Uw ordernummer 4.75442
Uw referentie Isabel Soete
A-PH-2710-ST100 Pershoek 5 ST100 0,320 160,00 27/03/2018
A-05-115-ST10-BLACK Eindstuk voor P-05-110 2 ST10 0,860 9,23 27/03/2018
Korting klant 46,34 %
A-GS-903-ST1 Isolatie Kolos (2m) 56 ST1 37,570 1.128,97 27/03/2018
Korting klant 46,34 %
BTW % Goed.Diensten Korting cnt % Bedrag BTW Basis BTW Bedrag Totaal
21,0000 1.298,20 3,0000 38,95 1.259,25 264,44 1.562,64
Netto te betalen EUR 1.562,64 voor 30/05/2018
Contante betaling EUR 1.523,69 voor 7/04/2018
Met vermelding van +++701/8005/83459+++

DEZE FACTUUR IS WEGENS CESSIE UITSLUITEND BEVRIJDEND BETAALBAAR AAN: BELFIUS COMMERCIAL FINANCE NV/SA
PLACE ROGIERPLEIN 11, B 1210 BRUSSELS, BELGIUM, TEL. 32 2.285 26.38, FAX . 32.2.285 26 99, BANK TRANSFER TO' BELFIUS BANK
BELFIUS - IBAN BE52 5645 1427 6508 en BIC GKCCBEBB

 F3=Exit   F10=Display Hex   F12=Cancel   F15=Services   F16=Repeat find   F19=Left   F20=Right
 1 records folded.
MA*   A                 MW                                        03/012
```

192.168.2.40:23

Type here to search

NLD BEP 19:19 13/06/2020

Import_Winsol_Invoices_202068145_5 - Kofax Transformation Modules - Validation

Batch Edit View Folder Document Page Field Options Help

Edit Batch

Batch Content

- Import_Winsol_Invoice
  - 1: Winsol Invoice
    - Page 1
    - Page 2
  - 2: Winsol Invoice
    - Page 1
    - Page 2
  - 3: Winsol Invoice
    - Page 1
    - Page 2
  - 4: Winsol Invoice
    - Page 1
    - Page 2
  - 5: Winsol Invoice
    - Page 1
    - Page 2
  - 6: Winsol Invoice
    - Page 1
    - Page 2
  - 7: Winsol Invoice
    - Page 1
    - Page 2
  - 8: Winsol Invoice
    - Page 1
    - Page 2
  - 9: Winsol Invoice
    - Page 1
    - Page 2
  - 10: Winsol Invoice
    - Page 1
    - Page 2
  - 11: Winsol Invoice
    - Page 1
    - Page 2
  - 12: Winsol Invoice
    - Page 1
    - Page 2
  - 13: Winsol Invoice
    - Page 1
    - Page 2
  - 14: Winsol Invoice
    - Page 1
    - Page 2

Classification Result

The document was classified to the following class: Invoice

40 fields valid, 1 field invalid (24 fields invisible, 2 fields read-only)

Pagina 1 (1 invalid field) | Pagina 2 (0 invalid fields)

Streepjescode groep [v. 5.0.4]

Streepjescode:	0200600710

Firma groep

Firma identificatie:	AC
Firmanaam:	AC - WINSOL ACTUELL N.V.
BTW-nummer:	BE0448902241

Leveranciersgroep

Leveranciersidentificatie:	00000600
Leveranciersnaam:	HYDRO BUILDING SYSTEMS BELGI
BTW-nummer:	BE0428769494
Rekeningnummer:	
IBAN nummer:	BE31380080935255
ACCOUNT	600005
ANALYTICALACCOUNT1	7000

Algemene groep

Factuurtype:		I
Factuurnummer:	H12248050	H12248050
Factuurdatum:	Datum:04-06-2020	04/06/2020

InPlace Editor - Field "INVOICENUMBER"

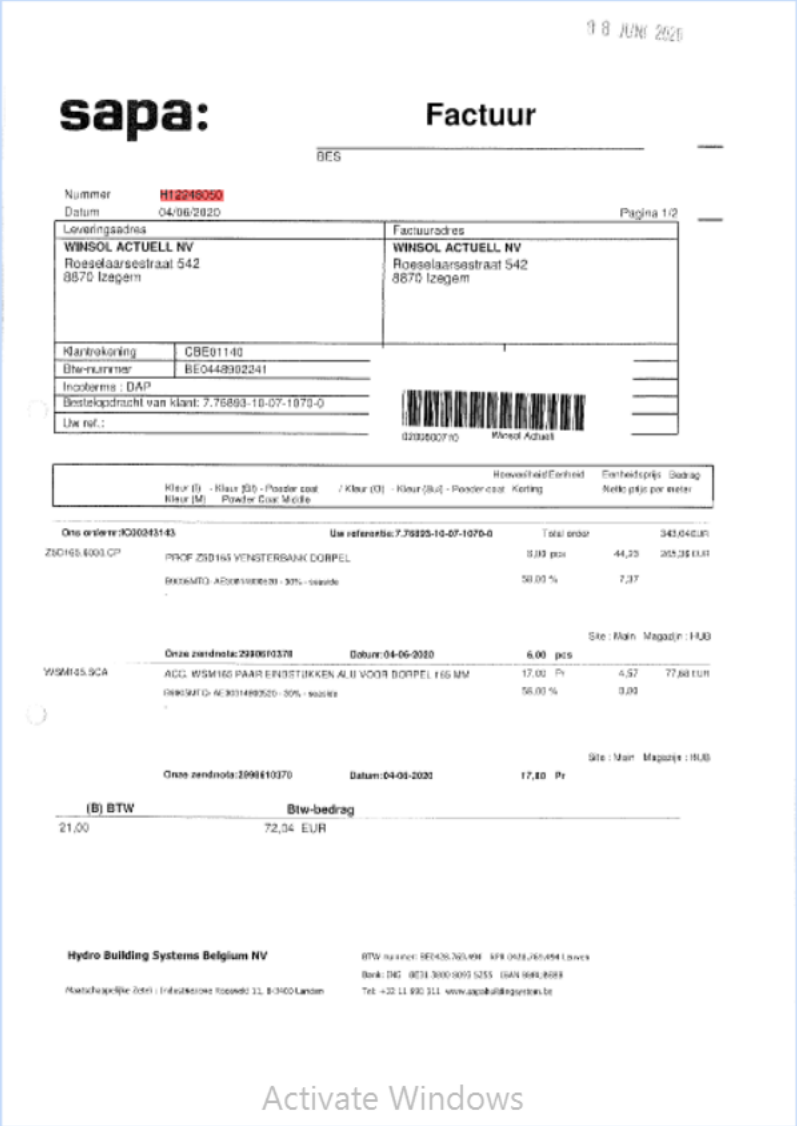
H12248050

INVOICENUMBER

H12248050

Current Error

Please confirm this field content.



sapa:

Factuur

Nummer H12248050

Datum 04/06/2020

Pagina 1/2

Leveringsadres	Factuuradres
WINSOL ACTUELL NV Roeselaarsestraat 542 8870 Izegem	WINSOL ACTUELL NV Roeselaarsestraat 542 8870 Izegem

Klantrekening CBE01140

Btw-nummer BE0448902241

Incoterms : DAP

Bestelopdracht van klant: 7.76893-10-07-1070-0

Uw ref.:

(B) BTW	Btw-bedrag
21,00	72,04 EUR

Hydro Building Systems Belgium NV

Activate Windows

Go to System in Control Panel to activate Windows.

Ready | OVR | NUM | CAPS | Doc: 1/23 | Page: 1/2 | Invalid Docs: 23

NLD 5:03 PM 6/8/2020

Import_Winsol_Invoices_202069140_5 - Kofax Transformation Modules - Validation

Batch Edit View Folder Document Page Field Options Help

Edit Batch

Batch Content

- Import_Winsol_Invoice
  - 1: Winsol Invoice
    - Page 1
    - Page 2
  - 2: Winsol Invoice
    - Page 1
    - Page 2
  - 3: Winsol Invoice
    - Page 1
    - Page 2
  - 4: Winsol Invoice
    - Page 1
    - Page 2
  - 5: Winsol Invoice
    - Page 1
    - Page 2
    - Page 3
    - Page 4
  - 6: Winsol Invoice
    - Page 1
    - Page 2
    - Page 3
    - Page 4
  - 7: Winsol Invoice
    - Page 1
    - Page 2
  - 8: Winsol Invoice
    - Page 1
    - Page 2
    - Page 3
    - Page 4
  - 9: Winsol Invoice
    - Page 1
    - Page 2
  - 10: Winsol Invoice
    - Page 1
    - Page 2
  - 11: Winsol Invoice
    - Page 1
    - Page 2
  - 12: Winsol Invoice
    - Page 1
    - Page 2

Classification Result

The document was classified to the following class: Invoice

40 fields valid, 1 field invalid (24 fields invisible, 2 fields read-only)

Pagina 1 (1 invalid field) | Pagina 2 (0 invalid fields)

Streepjescode groep [v. 5.0.4]

Streepjescode:	0200601110

Firma groep

Firma identificatie:	AC
Firmanaam:	AC - WINSOL ACTUELL N.V.
BTW-nummer:	BE0448902241

Leveranciersgroep

Leveranciersidentificatie:	00002083
Leveranciersnaam:	ALUK BELGIUM
BTW-nummer:	BE0888805159
Rekeningnummer:	
IBAN nummer:	BE29001853967464
ACCOUNT	600011
ANALYTICALACCOUNT1	4000

Algemene groep

Factuurtype:		I
Factuurnummer:	7020006494	7020006494
Factuurdatum:	8/06/2020	08/06/2020

InPlace Editor - Field "INVOICENUMBER"

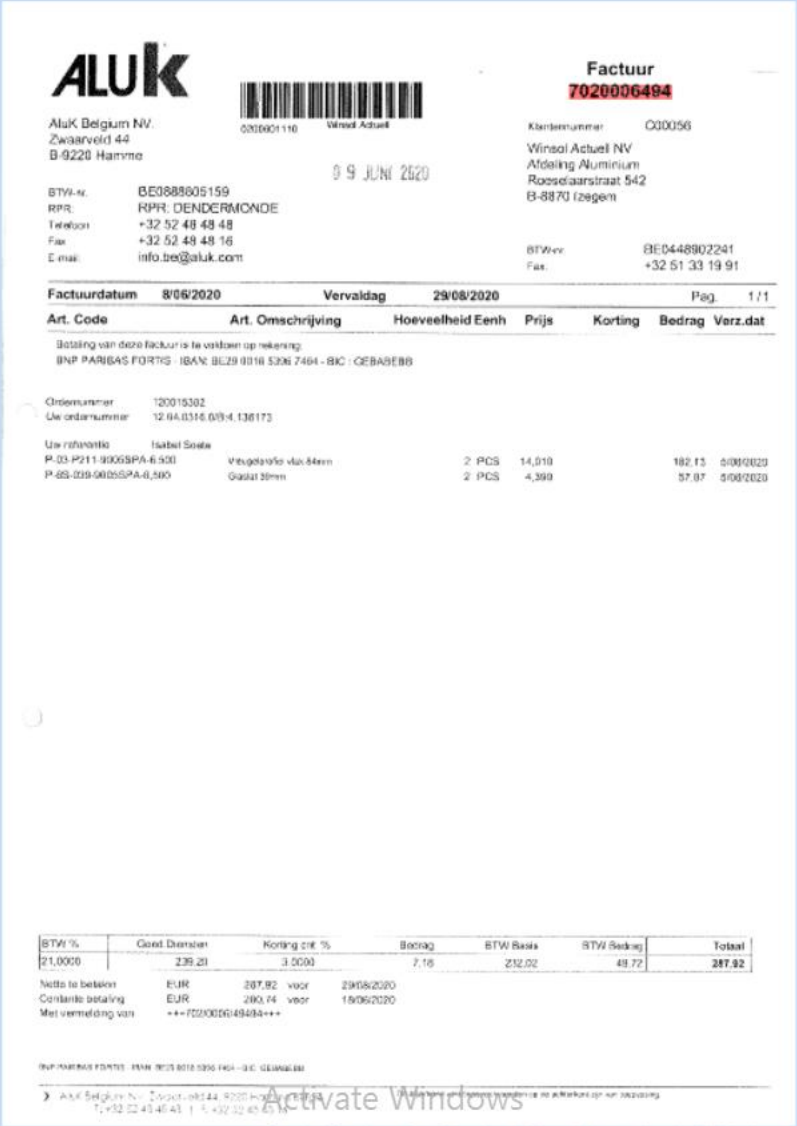
7020006494

INVOICENUMBER

7020006494

Current Error

Please confirm this field content.



Activate Windows
Go to System in Control Panel to activate Windows.

Ready | OVR | NUM | CAPS | Doc: 1/19 | Page: 1/2 | Invalid Docs: 19

NLD 6:27 PM 6/9/2020

Mail - Inbox - IBM Notes

File Edit View Create Actions Tools Window Help

Open

Discover | Bart Van Keer - Mail

Bart Van Keer
sds11

Inbox (11)
Drafts
Sent
Follow Up
All Documents
Junk
Trash
Chat History
Views
Folders
AANVRAAG-VERLOF (150)
ACCOUNTING (6)
1DB-A58 (1)
ADFINITY (2)
Approval CN
BTW Matirx
Bulk Invoices (4)
D.S.O.
FINDAS (5)
I70/I80
Payments List (14)
Review Invoices
APICS2
APM INLAND SERVICES
MAASVLAKTE
AS2 - Menten
ASIST
AVAYA
BOTS

New | Reply | Reply to All | Forward

Who	Subject
herfurthgroup	Flash 52/2020 : Herfurth Group C

Dropzone - Herfurth

Herfurth Dropzone

Please provide some extra information
about *B1_312620_20200511164621772000.png*

DropZone Dox4Us Documents

Filenumber
1834211 Search

031/622/1834211 (HERBA INGREDIENTS BV (NL BTW), 85

Source
Other Documents
Other Documents
Customs Documents
Shipping Documents

Archive

Status: Form is validated correctly and ready for archiving

Y:\CD-INVEST\DropZone

File Home Share View Manage Picture Tools

This PC > HomeDrive (Y:) > CD-INVEST > DropZone

Quick access

Name	Date modified	Type	Size
Test.docx	27/04/2020 14:52	Microsoft Word D...	89 KB
Analyse_Shipping Archive.docx	12/07/2019 10:51	Microsoft Word D...	265 KB
Analyse_Dropzone_Oproepen.pdf	10/05/2019 12:11	Foxit Reader PDF ...	265 KB
Analyse_Dropzone_Oproepen.docx	7/05/2019 11:48	Microsoft Word D...	19 KB
Integration_record.xlsx	30/04/2019 15:15	Microsoft Excel W...	12 KB
Analyse_Dropzone.docx	26/03/2019 8:35	Microsoft Word D...	1.349 KB
LeftMenu.bmp	19/03/2019 15:34	Bitmap image	4.572 KB
DropZone_Testverslag_20190313.docx	13/03/2019 9:31	Microsoft Word D...	1.057 KB
DropAnother.bmp	13/03/2019 9:26	Bitmap image	0 KB
DropZone First test.docx	12/03/2019 10:55	Microsoft Word D...	0 KB
Dropzone_record.xlsx	11/02/2019 13:18	Microsoft Excel W...	10 KB
B2.png	31/01/2019 15:40	PNG File	77 KB
B1.png	31/01/2019 15:40	PNG File	117 KB
A1.png	31/01/2019 15:39	PNG File	19 KB
A2.png	31/01/2019 15:37	PNG File	99 KB
DropZone_Test_Eucon.xlsx	7/01/2019 11:33	Microsoft Excel W...	10 KB
DropZone_DB_schema.docx	17/12/2018 17:02	Microsoft Word D...	200 KB
Dropzone_scherm_DOX4US4ALL_3c.bmp	22/11/2018 14:28	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL_3b.bmp	22/11/2018 14:28	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL_3a.bmp	22/11/2018 14:27	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL_2d.bmp	22/11/2018 14:25	Bitmap image	1.292 KB
Dropzone_scherm_DOX4US4ALL_2c.bmp	22/11/2018 14:24	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL_2b.bmp	22/11/2018 14:22	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL_2a.bmp	22/11/2018 14:19	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL.bmp	22/11/2018 14:19	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL_4.bmp	22/11/2018 13:40	Bitmap image	1.593 KB
Dropzone_scherm_DOX4US4ALL_5.bmp	22/11/2018 13:38	Bitmap image	1.292 KB
DorpZone_scherm_select.xlsx	22/11/2018 11:54	Microsoft Excel W...	9 KB
GYDO_DBmodel.png	22/11/2018 10:58	PNG File	39 KB
SDSFORM_DBmodel.png	22/11/2018 10:41	PNG File	67 KB
DOX4US4ALL_FormDB.png	22/11/2018 10:27	PNG File	76 KB
DropArea_CdInvest.docx	4/10/2018 9:41	Microsoft Word D...	363 KB
DropWidget Dox4Us - SDS getting started.docx	14/07/2015 9:16	Microsoft Word D...	763 KB

Meetings
My Widgets
Day-At-A-Glance

Online

NLD 16:47 11/05/2020

Document Preview

Mail2Me Download PDF Print PDF

HERFURTH LOGISTICS

Operations
Walter Mulle
mulle.walter@pelsto.com
+32 3 271 2817

Administration
ACCOUNTING
accounting@pelsto.com
+32 3 271 2319

HERFURTH LOGISTICS NV
Registered Office: CASSIERSSTRAAT 19
2000 ANTWERPEN 6, BELGIUM

VAT BE0445364909
R.P.R. 0445.364.909, ANTWERPEN, AFDELING ANTWERPEN
Registration nr. customs representative 2005

Invoice Nr. 027/908601 Date 09/12/2019

STEPAN DEUTSCHLAND GMBH
RODENKIRCHENER STR. 400
50389 WESSELING
GERMANY

ORIGINAL

Please mention on payment **027/908601**
Due date 09/12/2019

Your customer nr. S16359
Your VAT nr. DE811982539

Bank	IBAN	BIC/Swift	
KBC	BE84410064328197	KREDBEBB	EUR
BNP PARIBAS FORTIS	BE63220069693008	GEBABEBB	EUR

Invoice total amount

Base VAT excl.	725,00 EUR	
VAT 0%	0,00 EUR	(on 725,00 EUR)(1)
In our favour	**725,00 EUR**	

(1): Transport exemption from VAT: Art 146, e) from Eur. VAT Council Directive

Vessel/Voyage	Monte Alegre/ 948S	Our File	031/610/832411
Port of Loading	Antwerpen, BE	Total Packages	1
Port of Discharge	El Iskandariya (= Alexandria), EG	Total Weight	286,00 kg
E.T.S.	01/12/2019		
E.T.A.	14/12/2019		

Goods harmless chemicals

Description	Quantity	Base	Amount	Curr.	Rate of Exchange	VAT%	Amount in EUR	VAT Class
Compleum cost Fca to Fob loadport	1,00	535,00	535,00+	EUR	1,000000	0%	535,00+	(1)
Our handling fee	1,00	65,00	65,00+	EUR	1,000000	0%	65,00+	(1)
Courier charges	1,00	35,00	35,00+	EUR	1,000000	0%	35,00+	(1)
Insurance premium and tax	1,00	65,00	65,00+	EUR	1,000000	0%	65,00+	(1)
Insurance administration charges	1,00	25,00	25,00+	EUR	1,000000	0%	25,00+	(1)
Totals per currency	725,00 EUR						725,00 EUR	

10936349

- All our operations are subject to the Belgian Freight Forwarding Standard Trading Conditions 2005 (Annex M.B. 24/06/2005 nr. 05/06/231). Text will be sent free of charge upon first request. National and international trucking, organized by us, is subject to the stipulations of the Belgian Law dated 4/8/1982 enacting the international convention of the contract for the international carriage of goods by road (CMR). Our warehouse activities are governed by the General LogisticConditions, lodged October 9th, 2015 at the Office of the Chamber of Commerce and Industry of Antwerp and Waasland.
- In case of delay in the settlement of the invoice, the costs for recovery shall be due, as well as a yearly interest of 10% and an increase of 10% as a fixed compensation.
- Any legal proceedings shall be within the exclusive jurisdiction of the Antwerp Courts.

If you would like to receive your invoices by email in future, please contact our administration. Contact details in the right top corner of this invoice.

Invoice 027/908601 File 031/610/832411 Page 1 of 1

Document details

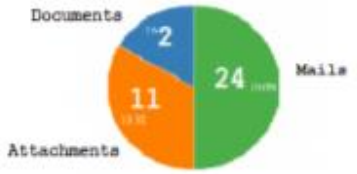
Type : Outgoing invoice
Invoice : 027/908601

Company : 031
Dept. : 610
File : 1832411

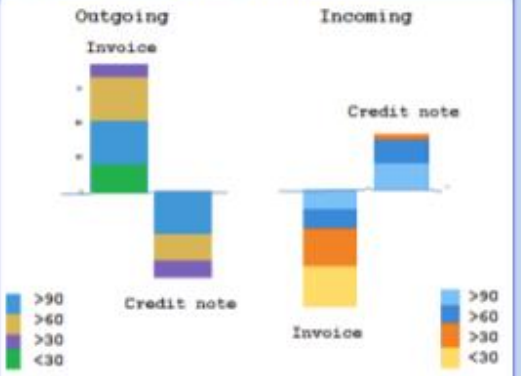
Customer: S16359
STEPAN DEUTSCHLAND
RODENKIRCHENER STR. 400
50389 WESSELING
DEU GERMANY

AMOUNT : 725,00 EURO

Related documents

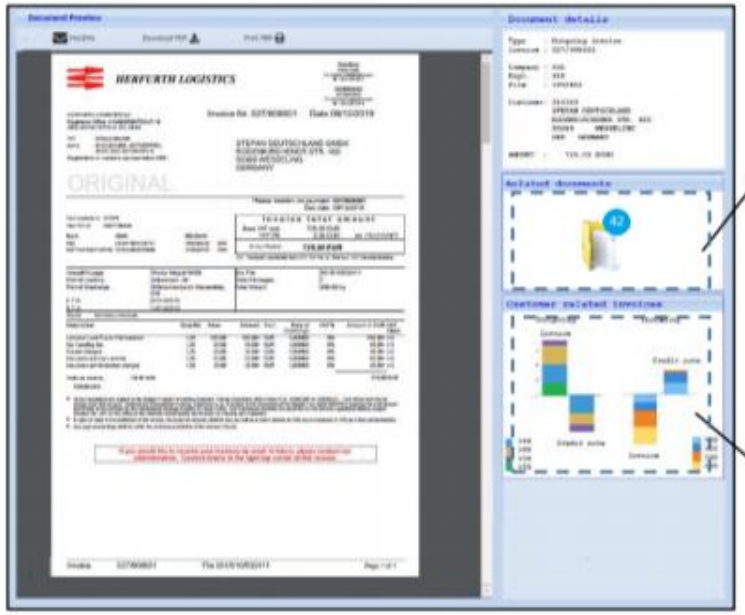


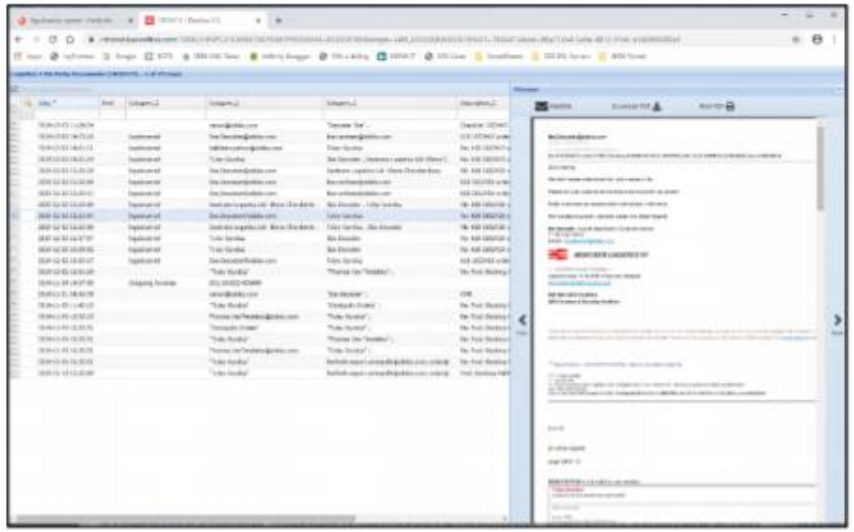
Customer related invoices

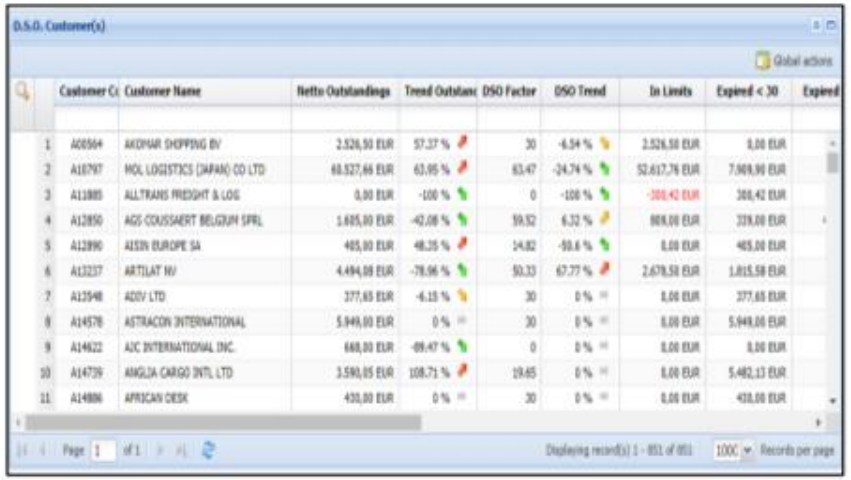


Flow of functionality.

= Clickable area







Recognize text

Results

Words	Score
ibm	0.672
code	0.978
like	0.982
a	0.813
startup	0.989
ibm	0.947

Recognize tekst - Watson


```
"images": [
    {
        "image": "img_0235.jpg",
        "text": "62 [09]\n22",
        "words": [
            {
                "line_number": 0,
                "location": {
                    "height": 48,
                    "left": 322,
                    "top": 227,
                    "width": 49
                },
                "score": 0.5176,
                "word": "62"
            },
            {
                "line_number": 0,
```

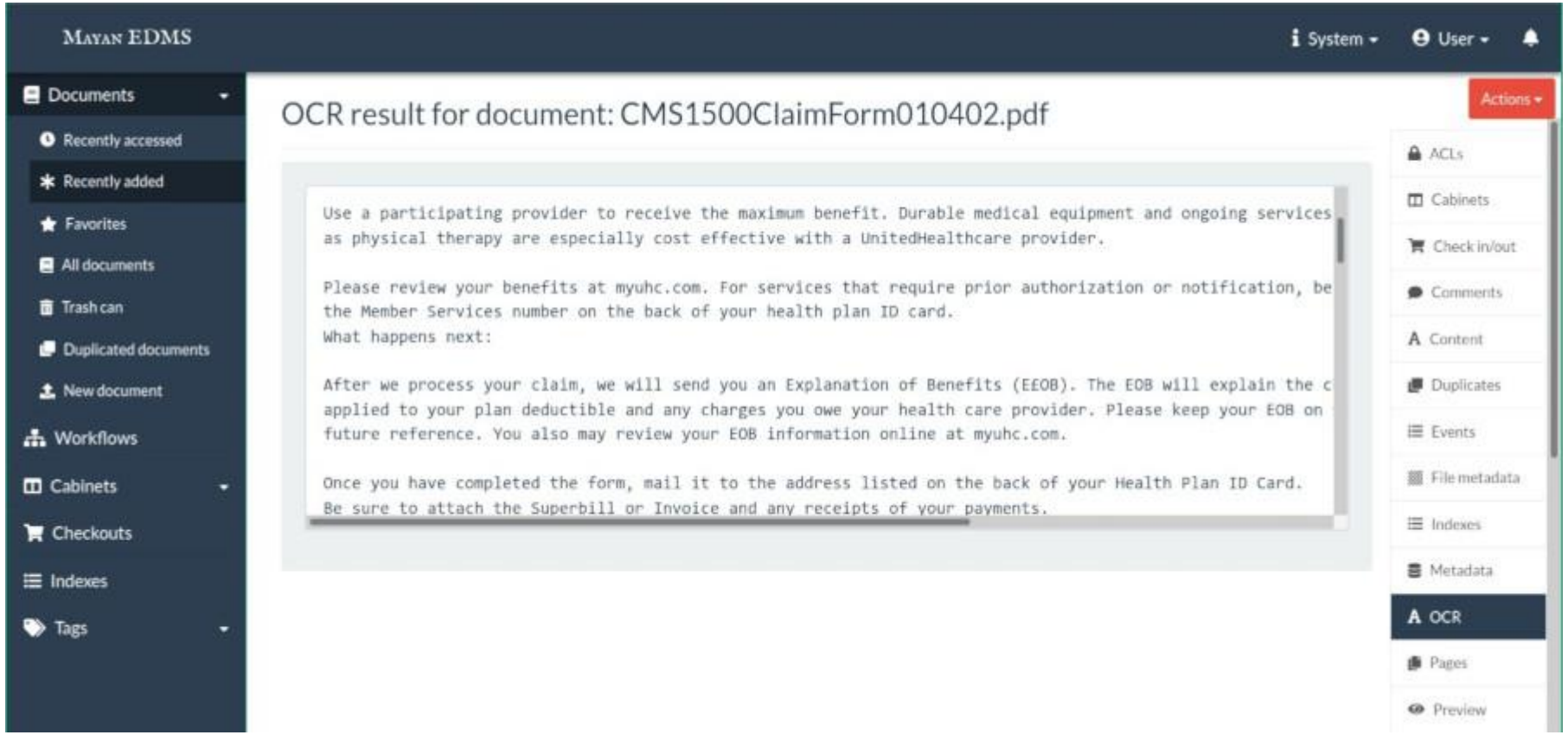
Recognize tekst – Tesseract


```
"textAnnotations": [
  {
    "locale": "en",
    "description": "YMLU 621093 9\n22U1\nCAUTION\n07/08/2017\n",
    "boundingPoly": {
      "vertices": [
```

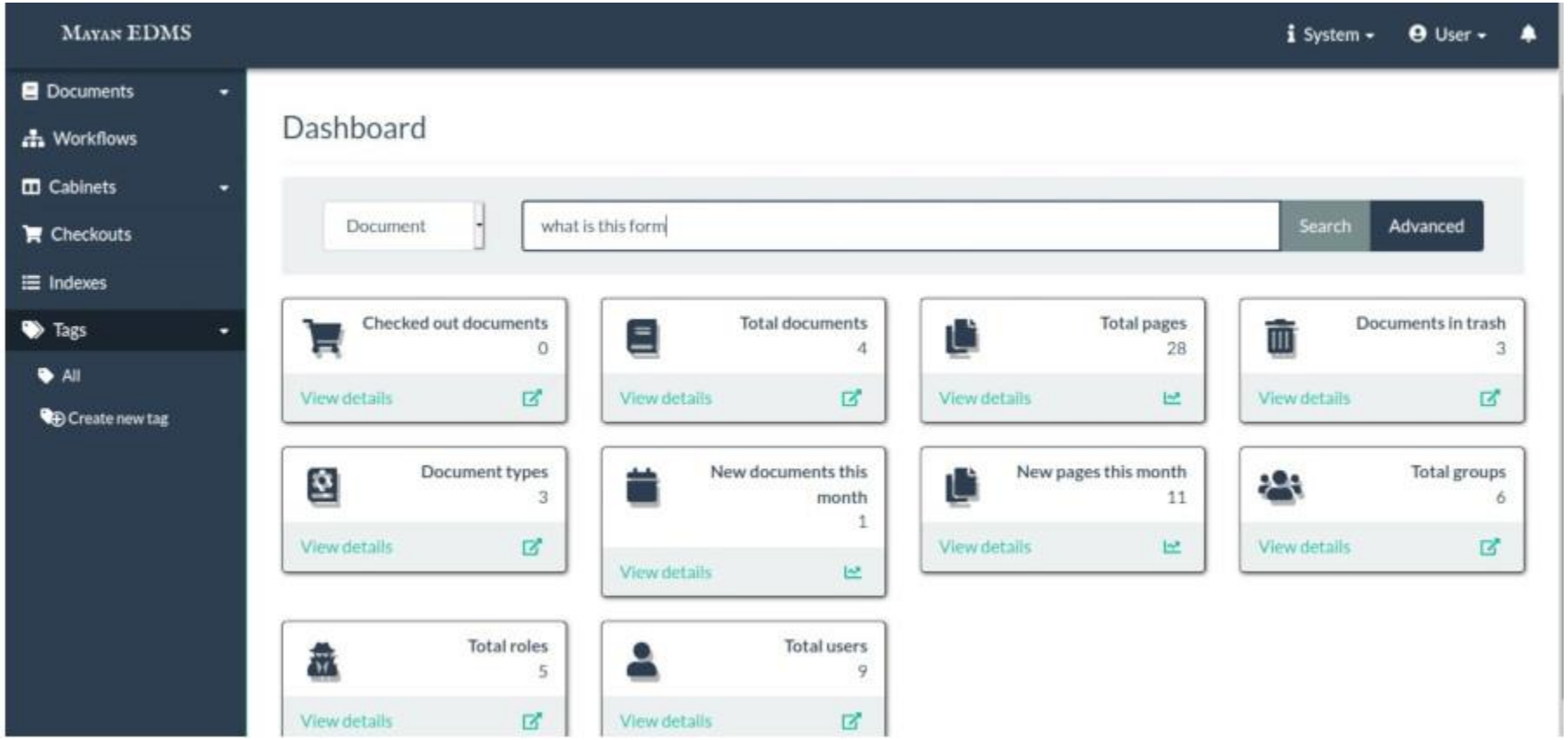
Mayan EDMS


Mayan EDMS
System
User
Documents
Workflows
Cabinets
Checkouts
Indexes
Tags
All
Create new tag
Preview of: Claims Workflow
Actions
ACLs
Document types
Events
Preview
States
Subscriptions
Transitions
1. Received
To Claim Analyst
2. Claim Review
On exit
Rename
To Expert
Reject
3. Expert
On entry
On exit
Claim Review Entry
Claim Review Exit
Archive
4. Vault
On entry
On exit
Expert Entry
Expert Exit

Mayan EDMS


MAYAN EDMS
System
User
Documents
Recently accessed
Recently added
Favorites
All documents
Trash can
Duplicated documents
New document
Workflows
Cabinets
Checkouts
Indexes
Tags
OCR result for document: CMS1500ClaimForm010402.pdf
Actions
Use a participating provider to receive the maximum benefit. Durable medical equipment and ongoing services as physical therapy are especially cost effective with a UnitedHealthcare provider.
Please review your benefits at myuhc.com. For services that require prior authorization or notification, be the Member Services number on the back of your health plan ID card.
What happens next:
After we process your claim, we will send you an Explanation of Benefits (EEOB). The EOB will explain the c applied to your plan deductible and any charges you owe your health care provider. Please keep your EOB on future reference. You also may review your EOB information online at myuhc.com.
Once you have completed the form, mail it to the address listed on the back of your Health Plan ID Card. Be sure to attach the Superbill or Invoice and any receipts of your payments.
ACLs
Cabinets
Check in/out
Comments
Content
Duplicates
Events
File metadata
Indexes
Metadata
OCR
Pages
Preview

Mayan Full Text Search



Tesseract best practices

- Provide input file with high quality

If possible, provide the software with a high quality input file. Poor image or document quality may prevent Tesseract from recognizing the text correctly. This is also true when processing documents with complex structures. Tesseract has problems recognizing complex structures such as tables and mixed text-image documents.

- Perform preprocessing

Perform appropriate preprocessing of the image data, such as contrast adjustment, noise reduction, and sharpening, to improve Tesseract's text recognition performance.

Tesseract best practices

- Define Region of Interest (ROI)

Define a region of interest around the relevant text area to increase recognition accuracy and reduce processing time.

- Make language selection

Make sure that the language setting of Tesseract matches the detected language in the image for best results.

Tesseract best practices

- Perform model training

If needed, you can improve Tesseract OCR by training a custom model for specific text types or fonts. This enables more accurate text recognition in specific scenarios.

- Perform validation and error correction

Review and correct recognized text results. Use validation tools and implementations for automatic error correction to improve the quality of recognized texts.

Tesseract resources

- The official documentation for Tesseract
- https://tesseract-ocr.github.io/
- https://github.com/invoice-x/invoice2data
- https://regex101.com/

Conclusion

Conclusion

After having fun with Tesseract OCR, I can say that the engine is amazing!! Here the list of interesting point from Tesseract in my opinion:

- Cost efficient - no expensive licenses
- Open Source.
- Easy to use.
- Good extract result.
- Support multi language (Latin & Non-Latin).