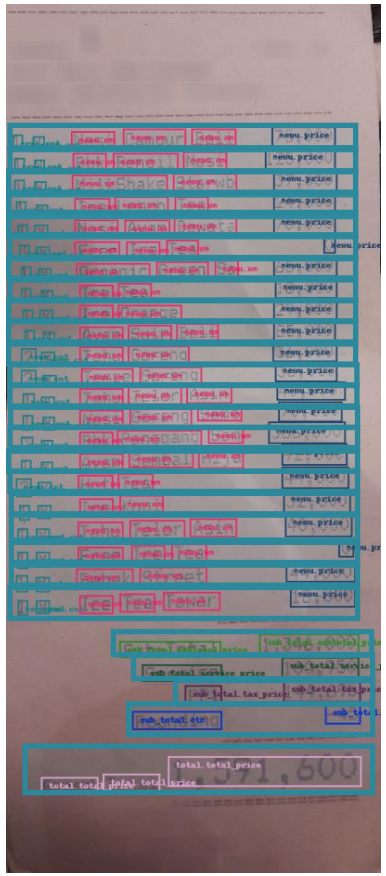
Ein Open Source Package
Analyse von komplex strukturierten Dokumenten

01.03.2022 – Dr. Janis Meyer

Beleg

Extraktion aus Rechnungen

Entitätenerkennung, Gruppierung von Positionen



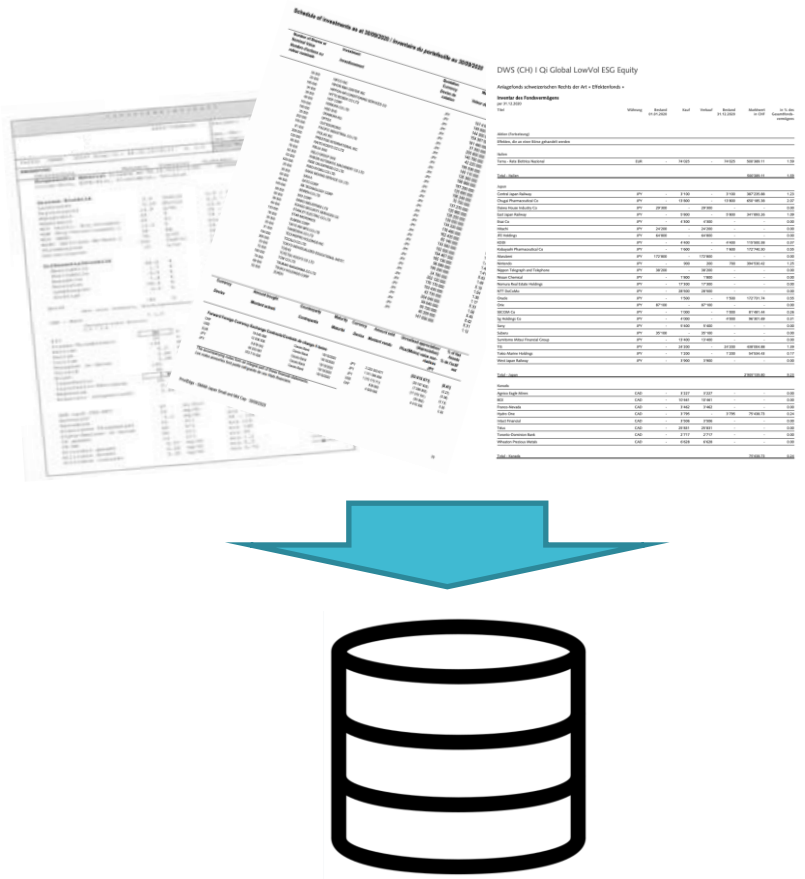
```
results: {
 line_items:
  [
   { quantity: 2,
     menu_item: "burger",
     price: 13.50
   }, ...
  ],
   sub_total_net: 150,
   service: 5,
   tax: 28,
   total: 183
  }
```

deepdoctection

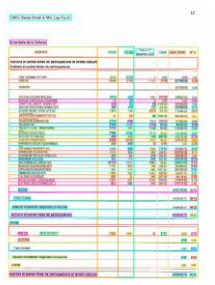
Berichte

Positionen aus Fondsberichten

Document to Database: Extraktion von Fondsdaten aus Jahresberichten



- Document Layout Analysis
  - Tabellenerkennung
  - Textextraktion
  - Dokument parsen



- Tabellenklassifikation
- Tabellennormalisierung

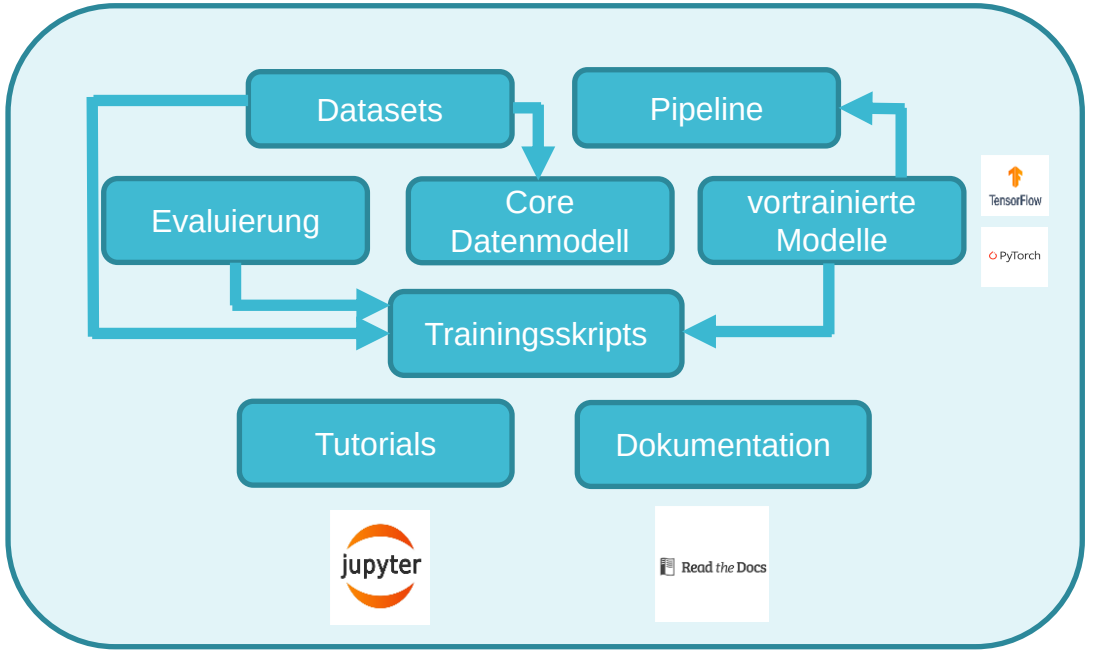
Datum	Fonds	Asset	Marktwert
31.12.21	HV Funds	VW AG	100

deepdoctection

Open Source Tool-Box

Ein Python Package für Document Layout Analysis und Anwendungen

deepdoctection



Pipelines

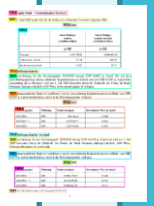
```
doc_path = "/path/to/dir/anr-5181_00_01-20201023-de.pdf"
```

```
df = analyzer.analyze(path=doc_path)
df.reset_state()
doc = iter(df)
```

```
page = next(doc)
```

```
processing anr-5181_00_01-20201023-de_22.pdf
```

PDF-> numpy



Globales Layout

Tabellenstruktur

{text: “USC“, box: [50,70,20,34], text: “Equity“, box: [70,70,45,34],…}

Text/ Zuordnung zum Layout

Lesereihenfolge

{text: “USC“, position: 1, text: “Equity“, position: 2,…}

Ausgabe

{page: 123
items: [{type: TABLE, box:…},
{type: CELL, relation: …, text: …}

Pipeline

Loader

Service layout

Service sub-layout

Service ocr

Service reading order

Output

np.array

bboxes

np.array

bboxes

Tensorpack

Detectron2

TensorFlow

PyTorch

layout predictor

OCR

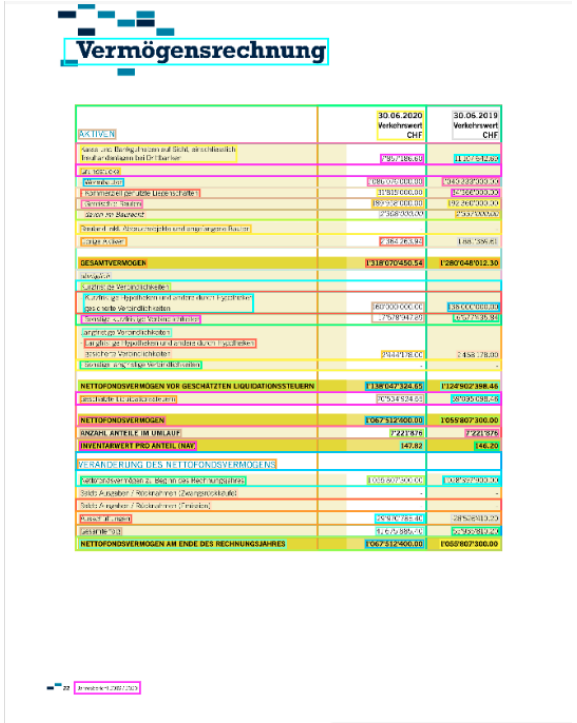
Tesseract OCR

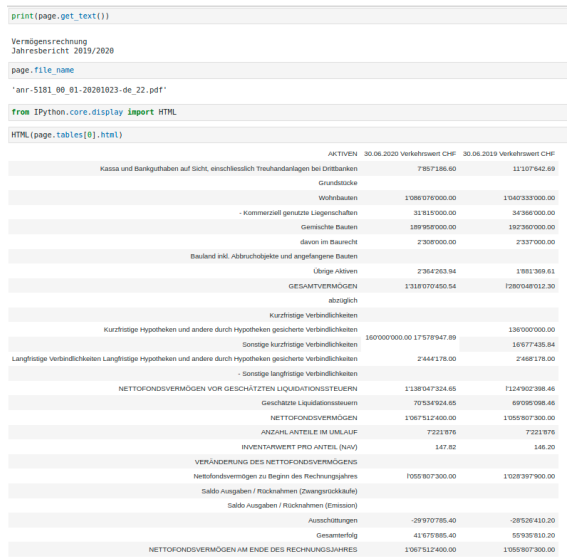
deepdoctection

Fine-tuning

Fine-Tuning auf spezifischen Use-Case essentiell

deepdoctection's pre-trained model





```
print(page.get_text())
```

```
Vermögensrechnung
Jahresbericht 2019/2020
```

```
page.file_name
```

```
'anr-5181_00_01-20201023-de_22.pdf'
```

```
from IPython.core.display import HTML
```

```
HTML(page.tables[0].html)
```

AKTIVEN	30.06.2020 Verkehrswert CHF	30.06.2019 Verkehrswert CHF
Kassa und Bankguthaben auf Sicht, einschliesslich Treuhandanlagen bei Drittbanken	7'857'186.60	11'107'642.69
Grundstücke		
Wohnbauten	1'086'076'000.00	1'040'333'000.00
- Kommerziell genutzte Liegenschaften	31'815'000.00	34'366'000.00
Gemischte Bauten	189'958'000.00	192'360'000.00
davon im Baurecht	2'308'000.00	2'337'000.00
Bauland inkl. Abbruchobjekte und angefangene Bauten		
Übrige Aktiven	2'364'263.94	1'881'369.61
GESAMTVERMÖGEN	1'318'070'450.54	1'280'648'012.30
abzüglich		
Kurzfristige Verbindlichkeiten		
Kurzfristige Hypotheken und andere durch Hypotheken gesicherte Verbindlichkeiten	160'000'000.00	136'000'000.00
Sonstige kurzfristige Verbindlichkeiten	17'578'947.89	16'677'435.84
Langfristige Verbindlichkeiten Langfristige Hypotheken und andere durch Hypotheken gesicherte Verbindlichkeiten	2'444'178.00	2'468'178.00
- Sonstige langfristige Verbindlichkeiten		
NETTOFONDSVERMÖGEN VOR GESCHÄTZTEN LIQUIDATIONSSTEUERN	1'138'047'324.65	1'124'902'398.46
Geschätzte Liquidationssteuern	70'534'924.65	69'095'098.46
NETTOFONDSVERMÖGEN	1'067'512'400.00	1'055'807'300.00
ANZAHL ANTEILE IM UMLAUF	7'221'876	7'221'876
INVENTARWERT PRO ANTEIL (NAV)	147.82	146.20
VERÄNDERUNG DES NETTOFONDSVERMÖGENS		
Nettofondsvermögen zu Beginn des Rechnungsjahres	1'055'807'300.00	1'028'397'900.00
Saldo Ausgaben / Rücknahmen (Zwangsrückkäufe)		
Saldo Ausgaben / Rücknahmen (Emission)		
Ausschüttungen	-29'970'785.40	-28'526'410.20
Gesamterfolg	41'675'885.40	55'935'810.20
NETTOFONDSVERMÖGEN AM ENDE DES RECHNUNGSJAHRES	1'067'512'400.00	1'055'807'300.00

Nach Fine-tuning mit Domain-Dataset (ca. 400 Seiten)

deepdoctection

Ausblick

Planung und Ideen

Schnittstellen zu NLP-Packages (z.B. SpaCy)

PDF- Miner

Weitere OCR Tools

Pre-Training/ Integration neuer Modelle

Nützliche Features, mehr Tutorials, …

Anregungen und PRs willkommen !

https://github.com/deepdoctection/deepdoctection