Spark OCR

for Data Scientists

Spark-OCR Team, John Snow Labs

Agenda

Main topic	Introduced Concepts
Introduction	Motivation, Overview and General Features.
Tasks & Metrics	Text Detection. OCR. Table Recognition. Form Recognition. VQA.
Basic Image Transformations	Basic Transformers and Pipelines: Image Enhancing, Scaling, Skew Correction. Text Detection Examples.
Optical Character Recognition	ImageToTextV1 vs. ImageToTextV2. Handwriting detection & recognition examples
PDF Processing	PDF Transformers. Pipelines with mixed digital and scanned PDFs.
Sample Notebooks	PDF Processing. Text Recognition. Tables Extraction. Form Extraction. VQA. Chart Extraction. Obfuscation.
Other Topics 🔥	Chart-To-Text. Obfuscation. Dicom-deid.
Questions and Answers	Engage in discussions with the presenter.
Summary and next steps	Next features to be added to Spark OCR.

Presenters today

Alberto

Introduction

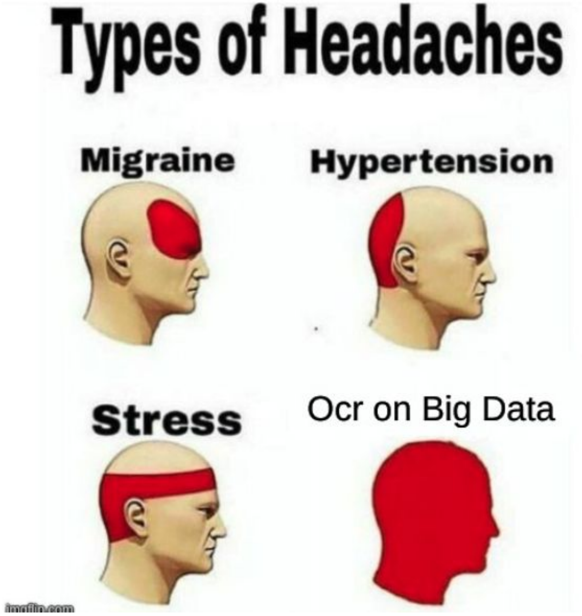
Visual NLP Team, John Snow Labs

Motivation

- Lots of text data locked into document images.
- We identified the strong need for a scalable solution.
- Three sources of stress: big data, big computation, big models.
- The job is not suitable for a single machine: need to scale out.
- Programming in the cluster is challenging.

Motivation

- We identified the strong need for a scalable solution.
- Diversity of input formats.
- Situation is more challenging than NLP.



Motivation

STARBUCKS Store #10208
11302 Euclid Avenue
Cleveland, OH (216) 229-0749

CHK 664290
12/07/2014 06:43 PM
1912003 Drawer: 2 Reg: 2

Vt Pep Mocha 4.95
Sbux Card 4.95
XXXXXXXXXXXX3228

Subtotal $4.95
Total $4.95
Change Due $0.00

------------ Check Closed ------------
12/07/2014 06:43 PM

SBUX Card x3228 New Balance: 37.45
Card is registered

STARBUCKS STORE #10208
11302 EUCLID AVENUE
CLEVELAND, OH (216) 229-0749
CHK 664290
12/07/2014 06:43 PM
1912003 DRAWER: 2. REG: 2
VT PEP MOCHA 4.95
SBUX CARD 4.95
XXXXXXXXXXXX3228
SUBTOTAL $4.95
TOTAL $4.95
CHANGE DUE $0.00
---- CHECK CLOSED
12/07/2014 06:43 PM
SBUX CARD X3228 NEW BALANCE: 37.45
CARD IS REGISTERED

- 111835b(JPEG) vs. 315b, a **355** factor!!.
- Density of information is much lower in OCR than NLP.
- Handling images is challenging.

Motivation

- We provide two flavors of scalability;
  - Strong Scalability: you care about completion time of individual pieces.
  - Weak Scalability: you care about throughput.
- Checkpointing: you want to resume the computation.
- We want to solve all these problems so you don’t have to.

Motivation

John Snow LABS

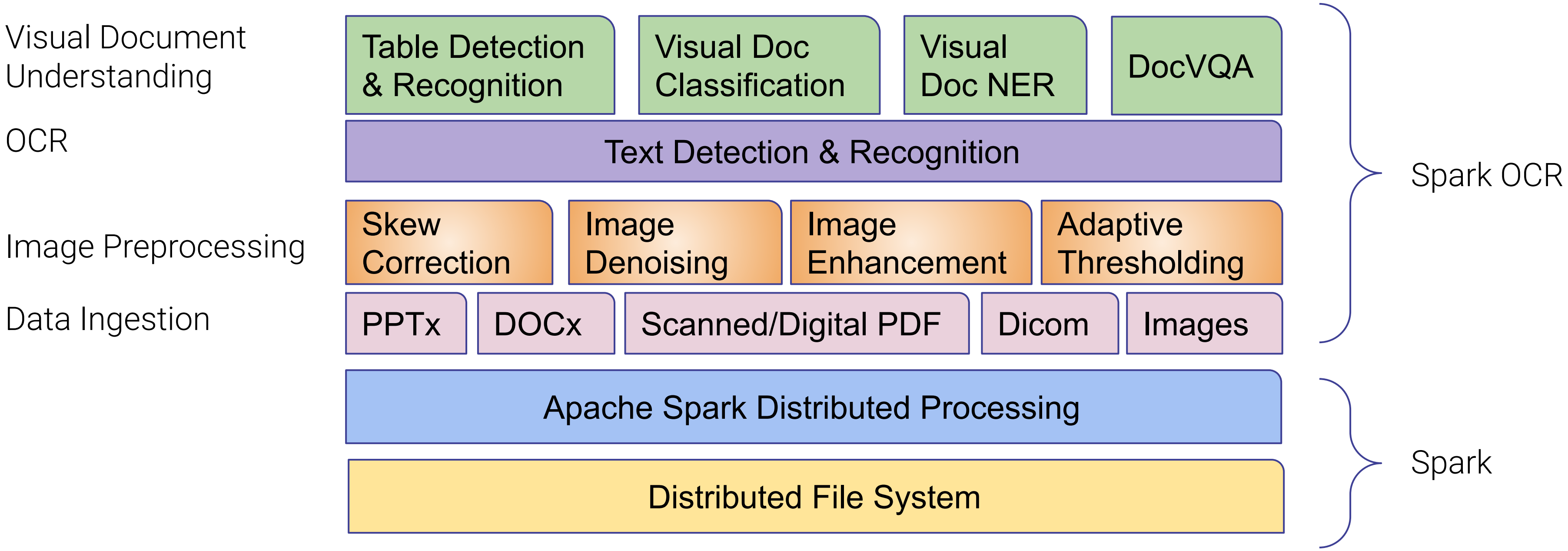
Filter by Event Type...

Event Type	Time	Message
TERMINATING	2022-07-18 19:28:53 -03	Cluster terminated. Reason: Inactivity
RESIZING	2022-07-18 18:47:17 -03	Autoscaling from 3 down to 2 workers.
RESIZING	2022-07-18 18:44:47 -03	Autoscaling from 5 down to 3 workers.
RESIZING	2022-07-18 18:42:17 -03	Autoscaling from 7 down to 5 workers.
RESIZING	2022-07-18 18:39:47 -03	Autoscaling from 11 down to 7 workers.
RESIZING	2022-07-18 18:37:17 -03	Autoscaling from 17 down to 11 workers.
RESIZING	2022-07-18 18:34:47 -03	Autoscaling from 27 down to 17 workers.
RESIZING	2022-07-18 18:32:17 -03	Autoscaling from 44 down to 27 workers.
UPSIZE_COMPLETED	2022-07-18 18:29:49 -03	Cluster upsize to 44 nodes completed.
RESIZING	2022-07-18 18:26:32 -03	Autoscaling from 21 up to 44 workers.
RESIZING	2022-07-18 18:24:12 -03	Autoscaling from 25 down to 21 workers.
UPSIZE_COMPLETED	2022-07-18 18:21:46 -03	Cluster upsize to 25 nodes completed.
RESIZING	2022-07-18 18:18:07 -03	Autoscaling from 2 up to 25 workers.
UPSIZE_COMPLETED	2022-07-18 14:10:08 -03	Cluster upsize to 2 nodes completed.
RESIZING	2022-07-18 14:05:22 -03	Attempting to resize cluster to its target of 2 workers.

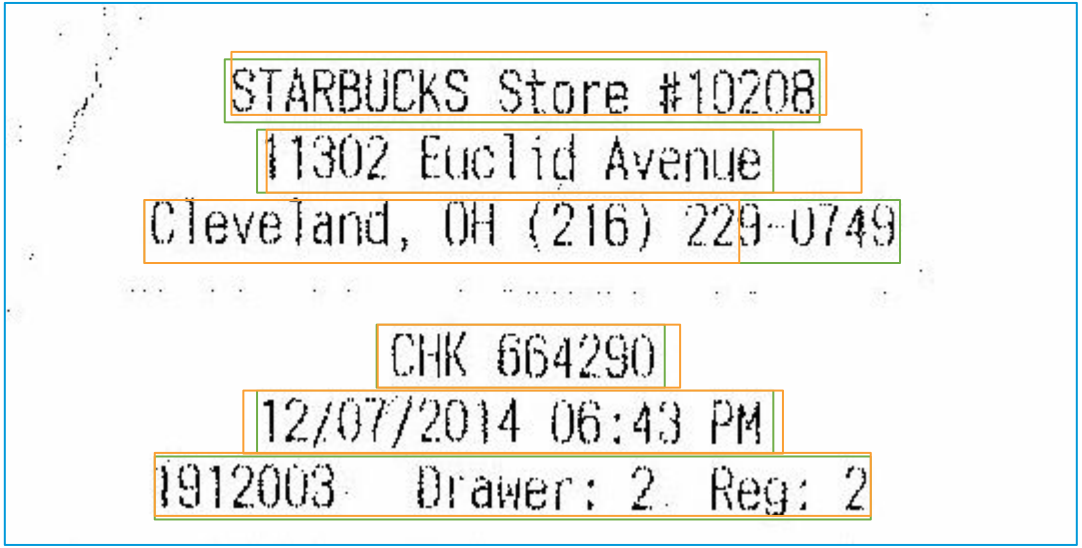
- Different cluster providers with the right maturity

Introduction to Visual NLP

- Visual NLP is an OCR, and Visual Document Understanding library built on top of Apache Spark.
- Curated list of features -> only things that work.
- Optimized for performance and accuracy.
- Created by industry practitioners.
- Actively developed.
- Security minded.


Visual Document Understanding
OCR
Image Preprocessing
Data Ingestion
Table Detection & Recognition
Visual Doc Classification
Visual Doc NER
DocVQA
Text Detection & Recognition
Skew Correction
Image Denoising
Image Enhancement
Adaptive Thresholding
PPTx
DOCx
Scanned/Digital PDF
Dicom
Images
Apache Spark Distributed Processing
Distributed File System
Spark OCR
Spark

Tasks & Metrics: text detection



Intersection over Union (IoU) -> can take various values, .6, .7, .8. etc
Precision.
Recall.
Fscore.

Metrics: TedEval

John Snow LABS

	ICDAR13	ICDAR15	ICDAR17	SROIE	FUNSD	CORD
SPARK (image_text_detector_v2)	R: 0.878 P: 0.907 H: 0.892	R: 0.713 P: 0.902 H: **0.796**	R: 0.774 P: 0.445 H: **0.565**	R: 0.848 P: 0.387 H: 0.532	R: 0.151 P: 0.972 H: 0.261	R: 0.853 P: 0.730 **H: 0.787**
SPARK (image_text_detector_opt)	R: 0.869 P: 0.949 H: **0.907**	R: 0.508 P: 0.914 H: 0.653	R: 0.535 P: 0.430 H: 0.477	R: 0.754 P: 0.385 H: 0.510	R: 0.664 P: 0.965 H: 0.787	R: 0.813 P: 0.721 H: 0.764
SPARK_DIT (image_text_detector_dit)				R: 0.776 P: 0.420 **H: 0.545**	R: 0.844 P: 0.941 **H: 0.890**	R: 0.516 P: 0.374 H: 0.434

Metrics: OCR

blogpost

$$\text{CER} = \frac{S + D + I}{N}$$

Where:

- **S** = Number of **substitutions**
- **D** = Number of **deletions**
- **I** = Number of **insertions**
- **N** = Total number of characters in the **ground truth** text

Ground truth: `"hello"`

Prediction: `"hallo"`

- Substitutions: 1 (`e` → `a`)
- Deletions: 0
- Insertions: 0
- Total ground truth characters: 5

CER= 1/5 = 20%

Metrics: OCR

[blogpost]

John Snow LABS

(iii) Series B Financing

On April 28, 2018, the Company and its subsidiaries entered into the Series B Share Purchase Agreement with the then Series B Preferred Shareholders, pursuant to which the then Series B Preferred Shareholders agreed to subscribe for a maximum of 45,908,818 Series B Preferred Shares in aggregate to be issued by our Company at a subscription price of approximately US$5.66 per share and an aggregate consideration of approximately US$260 million. The Series B Preferred Shares were issued in full on May 8, 2018 as set forth in the table below.

Name of Shareholder	Number of Series B Preferred Shares	Purchase Amount (US$)
WuXi Healthcare Ventures	882,861	4,999,994.99
6 Dimensions Capital, L.P.	3,354,875	18,999,999.08
6 Dimensions Affiliates Fund, L.P.	176,572	999,997.87
Graceful Beauty Limited	4,237,737	23,999,999.73
Tetrad Ventures Pte Ltd	8,828,618	49,999,995.19
Hikeo Biotech L.P.	1,589,151	8,999,997.78
Pure Progress International Limited	1,765,723	9,999,995.64
Kaitai International Funds SPC	882,861	4,999,994.99
Taikang Kaitai (Cayman) Special		

Which text relates to table and which is simply text? How to connect values to rows and columns?

Metrics: OCR

Type	Dataset	Model	Caching	CPU TimeTaken (seconds)	CPU Average Time	CPU CER	GPU TimeTaken (seconds)	GPU Average Time	GPU CER
Printed	FUNSD	ocr_base_printed_v2	No	1709.24	0.01514	0.07333	405.61	0.00359	0.07333
		ocr_base_printed_v2_opt	No	1608.29	0.01425	0.07351	1072.87	0.00951	0.07351
		ocr_large_printed_v2	No	3216.36	0.02850	0.06786	544.25	0.00482	0.06786
		ocr_large_printed_v2_opt	No	3505.75	0.03106	0.06787	1245.23	0.01103	0.06787
		ocr_base_printed_v2	Yes	1559.14	0.01381	0.07333	351.01	0.00311	0.07333
		ocr_base_printed_v2_opt	Yes	1348.17	0.01194	0.07371	732.45	0.00649	0.07371
		ocr_large_printed_v2	Yes	2857.73	0.00771	0.06786	518.96	0.00460	0.06786
		ocr_large_printed_v2_opt	Yes	3111.66	0.02757	0.06773	904.11	0.00801	0.06773
	SROIE	ocr_base_printed_v2	No	3182.52	0.00823	0.01471	867.77	0.00224	0.01471
		ocr_base_printed_v2_opt	No	2839.97	0.00734	0.01489	2410.47	0.00643	0.01489
		ocr_large_printed_v2	No	5440.10	0.01406	0.01428	1093.20	0.00283	0.01428
		ocr_large_printed_v2_opt	No	5063.07	0.01309	0.01463	3111.29	0.00804	0.01463
		ocr_base_printed_v2	Yes	2108.02	0.00545	0.01470	810.26	0.00209	0.01470
		ocr_base_printed_v2_opt	Yes	1907.55	0.00493	0.01488	1608.62	0.00493	0.01488
		ocr_large_printed_v2	Yes	4311.84	0.00330	0.01428	993.43	0.00257	0.01428
		ocr_large_printed_v2_opt	Yes	3864.54	0.00999	0.01434	2035.05	0.00526	0.01434
Handwritten	IAM	ocr_base_handwritten_v2	No	2094.36	0.00372	0.06478	617.20	0.0013	0.06478
		ocr_base_handwritten_v2_opt	No	1885.96	0.00335	0.06548	1588.70	0.0031	0.06548
		ocr_large_handwritten_v2	No	3153.85	0.00560	0.04645	1492.10	0.0027	0.04645
		ocr_large_handwritten_v2_opt	No	3109.64	0.04529	0.04502	2669.50	0.0039	0.04502
		ocr_base_handwritten_v2	Yes	1412.95	0.00251	0.06478	390.72	0.0008	0.06478
		ocr_base_handwritten_v2_opt	Yes	1304.18	0.00232	0.06566	1076.24	0.0023	0.06566
		ocr_large_handwritten_v2	Yes	2281.14	0.00405	0.04645	363.07	0.0006	0.04645
		ocr_large_handwritten_v2_opt	Yes	2137.57	0.00380	0.04487	1325.86	0.0132	0.04487

Metrics: Table recognition and TSR.

John Snow LABS

Preferred Shares in aggregate to be issued by our Company at a subscription price of approximately US$5.66 per share and an aggregate consideration of approximately US$260 million. The Series B Preferred Shares were issued in full on May 8, 2018 as set forth in the table below.

table:0.999985

Name of Shareholder	Number of Series B Preferred Shares	Purchase Amount
		(US$)
WuXi Healthcare Ventures	882,861	4,999,994.99
6 Dimensions Capital, L.P.	3,354,875	18,999,999.08
6 Dimensions Affiliates Fund, L.P.	176,572	999,997.87
Graceful Beauty Limited	4,237,737	23,999,999.73
Tetrad Ventures Pte Ltd	8,828,618	49,999,995.19
Hikeo Biotech L.P.	1,589,151	8,999,997.78
Pure Progress International Limited	1,765,723	9,999,995.64
Kaitai International Funds SPC	882,861	4,999,994.99
Taikang Kaitai (Cayman) Special Opportunity I	2,648,585	14,999,996.29
CJS Medical Investment Limited	3,531,447	19,999,996.94
SCC Growth IV Holdco G, Ltd.	5,297,171	29,999,998.25
YF IV Checkpoint Limited	5,297,171	29,999,998.25
HH CST Holdings Limited	1,765,723	9,999,995.64
ARCH Venture Fund IX, L.P.	441,430	2,499,994.67
ARCH Venture Fund IX Overage, L.P.	1,324,292	7,499,995.32
Terra Magnum CST LLC	353,144	1,999,995.73
3W Partners Fund II, L.P.	882,861	4,999,994.99
Huifu Investments Limited	882,861	4,999,994.99
King Star Med LP	1,765,723	9,999,995.64
Total	**45,908,806**	**259,999,931.98**

On September 23, 2018, the Company and Golden & Longevity Portfolios L.P. entered

Metrics: Table recognition and TSR.

John Snow LABS

	Material	Labor	Total
Surface Facilities			
Buildings and structures	29,380	33,640	63,020
Major equipment	46,350	4,570	50,920
Bulk material	29,040	16,410	45,450
Site development	7,570	4,730	12,300
Shafts and Hoists			
Major equipment	24,500	8,300	32,800
Shafts and lining	58,100	31,400	89,500
Underground Facilities			
Excavations and structures	2,510	4,510	7,020
Major equipment	3,170	220	3,390
Bulk material	1,960	1,470	3,430
Mining			
Major equipment	64,700	---	64,700
Mine construction	582,330	655,640	1,237,970
Backfilling			
Mine backfilling	102,300	116,000	218,300
Shaft sealing	90	110	200
Total Field Costs	952,000	877,000	1,829,000
Architect-Engineer Services			53,000
Owner's Costs			218,000
Contingency			534,000
TOTAL FACILITY COST			2,634,000

+ *generate a list of all adjacency relations between each content cell and its nearest horizontal and vertical neighbours.*
+ *An adjacency relation is a tuple containing the textual content of both cells, the direction and the number of blank cells (if any) in between.*
+ *This 1-D list of adjacency relations can be compared to the ground truth by using precision and recall measures.*
+ *Example: (Material, Labor), (Labor, Total), (Surface Facilities, Building and Structures), etc.*

Metrics: Table Detection evaluation - ICDAR2019 Track A

Model	general_model_table_detection_v2	general_model_table_detection_v2	table_detection_v3	table_detection_v3
Score threshold	0.5	0.8	0.5	0.8
IOU @ 0.6				
precision	0.9339407744874715	0.9399538106235565	0.96636771300448	0.9813953488372
recall	0.9131403118040089	0.9064587973273942	0.95991091314031	0.9398663697104
f1	0.9234234234234234	0.9229024943310656	**0.96312849162011**	0.9601820250284
IOU @ 0.7				
precision	0.9111617312072893	0.9191685912240185	0.96188340807174	0.9767441860465
recall	0.89086859688196	0.8864142538975501	0.95545657015590	0.9354120267260
f1	0.9009009009009009	0.9024943310657596	**0.95865921787709**	0.9556313993174

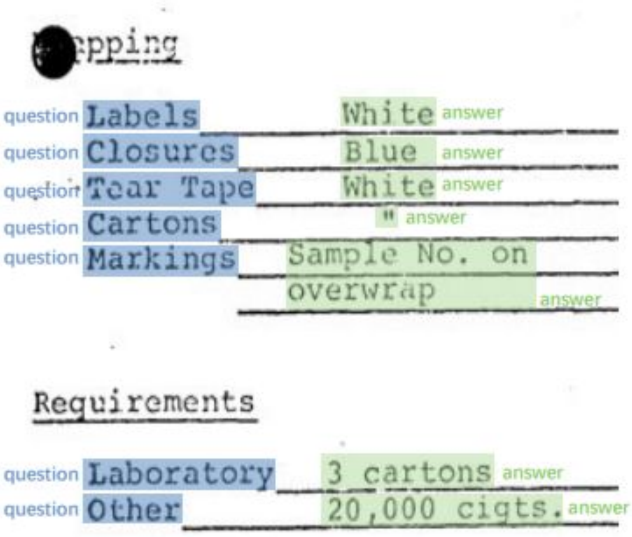
Metrics: TSR Textract vs. JSL (PubMed1 dataset)

John Snow LABS

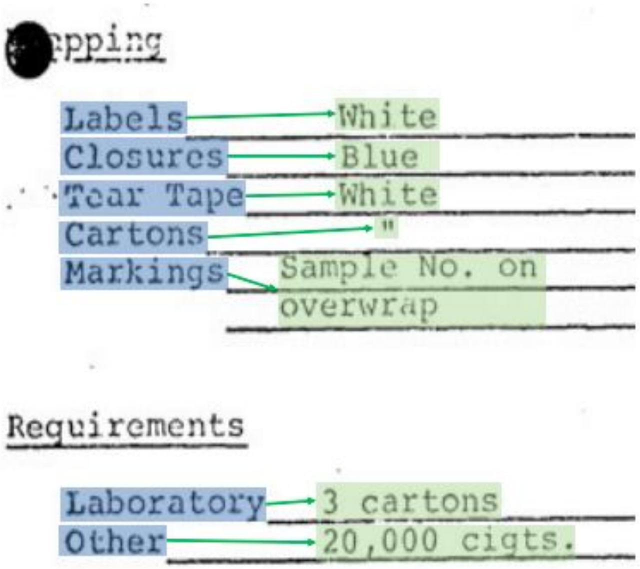
Model	JSL image2text	AWS Textract	JSL image2textV2 with regions merger 5.4.0	JSL image2text 5.4.0
IOU @ 0.6				
precision	0.7435719249478805	0.7728635682158921	0.5064683053040103	0.726122707147375
recall	0.6815286624203821	0.6566878980891719	0.49872611464968153	0.7312101910828025
f1	0.7111997341309405	0.7100550964187328	0.5025673940949936	**0.7286575690257062**
IOU @ 0.7				
precision	0.6414176511466296	0.65244375484872	0.2716688227684347	0.5970904490828589
recall	0.5878980891719745	0.554140127388535	0.267515923566879	0.6012738853503184
f1	**0.6134928547690263**	0.5991735537190084	0.26957637997432604	0.5991748651221834

Metrics: Form Recognition

John Snow LABS



Semantic entity recognition (SER)



Relation extraction (RE)

TP: Correctly predicted entities.

FP: Predicted entities that don't match any reference.

FN: Reference entities missed by the model.

TP: Predicted relation matches a reference relation.

FP: Predicted relation not in reference.

FN: Reference relation not predicted.

Metrics: Form Recognition

[blogpost]

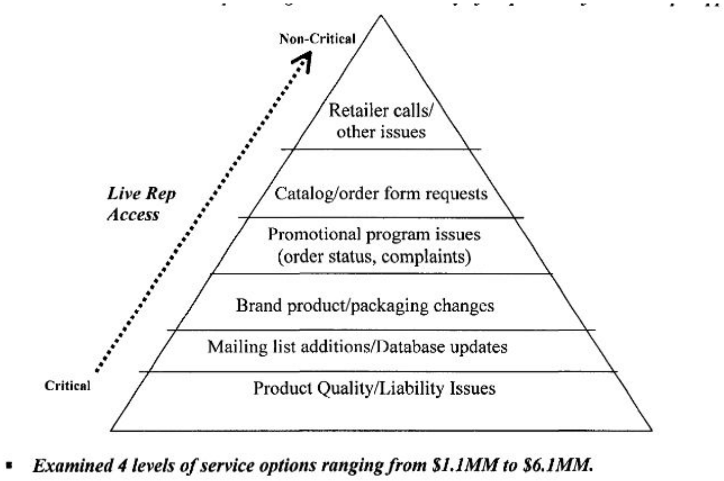
	F1 Score
Amazon Textract	0.48
Azure Form Recognizer	0.57
Google Cloud Document AI	0.54
John Snow Labs Visual NLP	0.93
John Snow Labs Visual NLP*	0.93

Evaluation of the different models on FUNSD dataset

	F1 Score
Amazon Textract	0.42
Azure Form Recognizer	0.36
Google Cloud Document AI	0.38
John Snow Labs Visual NLP	0.53
John Snow Labs Visual NLP*	0.89

Evaluation of the different models on the custom Genetic Reports 0.1 dataset

Metrics: Visual Question Answering



What is the issue at the top of the pyramid? Retailer calls/ other issues

Which is the least critical issue for live rep support? Retailer calls/other issues

Which is the most critical issue for live rep support? Product quality/liability issues

Not only extract and interpret the textual (handwritten, typewritten or printed) content of the document images, but also other visual cues including layout (page structure, forms, tables), non-textual elements (marks, tick boxes, separators, diagrams) and style (font, colours, highlighting).

Metrics: Visual Question Answering

For a prediction-answer pair:

$$\mathrm{ANLS}(p, g) = \begin{cases} 1 - \frac{LD(p,g)}{\max(|p|,|g|)}, & \text{if } \frac{LD(p,g)}{\max(|p|,|g|)} \leq \tau \\ 0, & \text{otherwise} \end{cases}$$

Where:

- p = predicted answer (string)
- g = ground truth answer (string)
- $LD(p, g)$ = Levenshtein Distance between `p` and `g`
- $|p|$ = length of predicted answer
- $|g|$ = length of ground truth answer
- τ = threshold (usually **0.5**)

+ **ANLS** rewards answers that are **close** to the ground truth, even if they're not exact, by using a **normalized edit distance** (Levenshtein distance).
+ It's especially helpful in OCR + NLP tasks where small text deviations (like spacing, OCR noise) can occur.

Model	ANLS Score(DocVQA)
docvqa_donut_base_opt	67.5
docvqa_pix2struct_jsl_opt	72.1

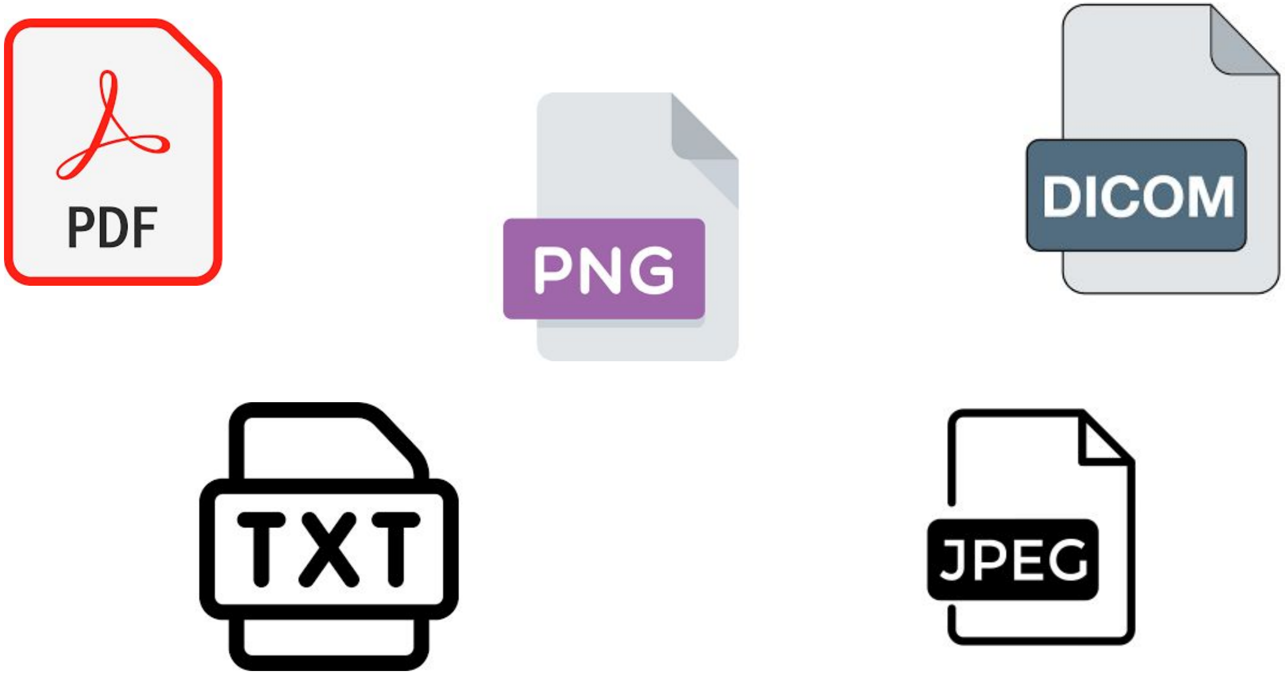
Basic Transformations & Pipelines

for Data Scientists

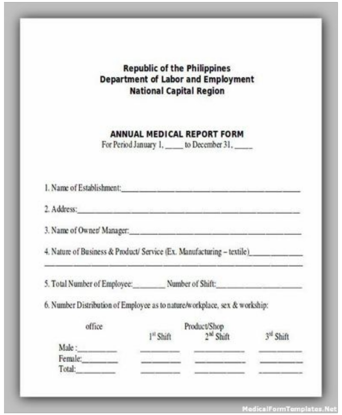
Spark-OCR Team, John Snow Labs

Basic Image Transformation

Documents in real life



Documents in real life



Republic of the Philippines
Department of Labor and Employment
National Capital Region

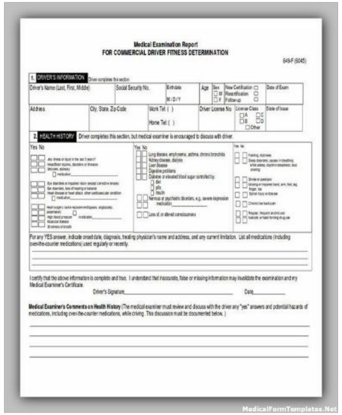
ANNUAL MEDICAL REPORT FORM
For Period January 1, ____ to December 31, ____

1. Name of Establishment:____________________
2. Address:____________________
3. Name of Owner/ Manager:____________________
4. Nature of Business & Product/ Service (Ex. Manufacturing – textile)____________
5. Total Number of Employee:________ Number of Shift:____________
6. Number Distribution of Employee as to nature/workplace, sex & workship:

	office	Product/Shop 1st Shift	2nd Shift	3rd Shift
Male:	________	________	________	________
Female:	________	________	________	________
Total:	________	________	________	________

New Nouveau Brunswick
DEPARTMENT OF PUBLIC SAFETY
LICENSING & RECORDS BRANCH

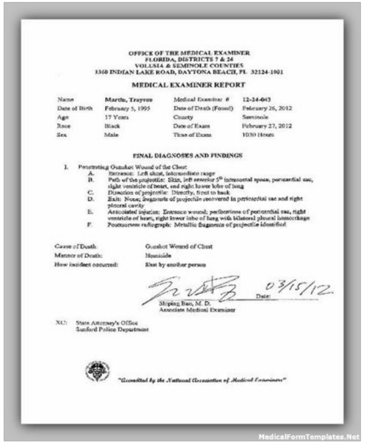
MEDICAL FITNESS REPORT



Medical Examination Report
FOR COMMERCIAL DRIVER FITNESS DETERMINATION

1 DRIVER'S INFORMATION

2 HEALTH HISTORY



OFFICE OF THE MEDICAL EXAMINER
FLORIDA, DISTRICTS 7 & 24
VOLUSIA & SEMINOLE COUNTIES

MEDICAL EXAMINER REPORT

FINAL DIAGNOSES AND FINDINGS

03/15/12

Sample Visual NLP workflow

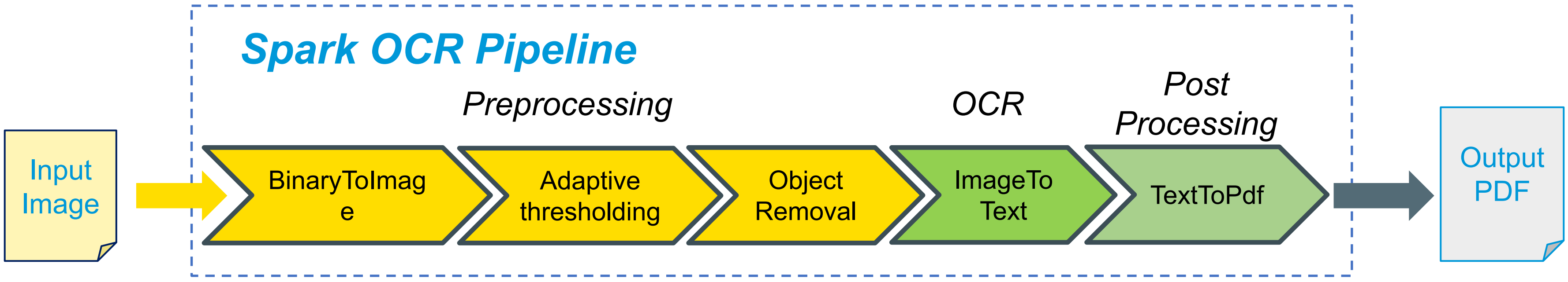


Image Transformations

CPU

GPU

ImageTransformer:

- Erosion
- Dilation
- Scaling
- Otsu Thresholding
- Adaptive Thresholding
- Median Blur
- Blur
- Remove Objects

GPUImageTransformer:

- Erosion
- Dilation
- Scaling
- Otsu Thresholding
- Huang Thresholding

Optical Character Recognition

ImageToText vs ImageToTextV2

ImageToText
ImageToHocr
ImageToTextPdf

- Based on LSTM
- Faster
- End-to-End solution
- Bad accuracy on low quality image

ImageToTextV2

- Based on Transformer architecture.
- Combination of CV and NLP
- Slower in CPU, benefits from GPU
- External Text Detector is required

Pdf processing

Pdf transformers

- **PdfToText** – extract text from selectable PDF
- **PdfToImage** – render each page in the PDF as image
- **ImageToPdf** – store image to PDF format
- **TextToPdf** – render text with positions to PDF format
- **PdfDrawRegions** – draw regions to existing PDF
- **ImageToTextPdf** - recognize text from image and render results on top of the original source image in the PDF.
- **PdfAssembler** - assemble multi page PDF document from single page PDFs.

Sample Notebooks

PDF Processing: SparkOcrProcessMultiplepageScannedPDF.ipynb

Text Recognition: SparkOcrImageToTextV2.ipynb

Tables: SparkOcrImageTableRecognitionWRegionsMerger

Forms: FormRecognitionGeo.ipynb

VQA: VisualQuestionAnsweringOnInvoices.ipynb

(other topics)

Charts: SparkOcrChartToTextTable.ipynb

Obfuscation: SparkOcrImageObfuscation.ipynb

Dicom-deid: webinars/dicom_deid

Questions and Answers

Links

- Workshop
- Documentation
- Annotation Lab
- Spark NLP Medium

Questions and Answers

Summary and Next Steps

- New Dicom Features.
- New Models for OCR.
- New Models for Form Extraction.

Contact Us!

Contact us on Slack!

Emails: alberto@johnsnowlabs.com,
gokhan@johnsnowlabs.com,
mykola@johnsnowlabs.com,
gursev@johnsnowlabs.com,
enes@johnsnowlabs.com

Light Pipelines

Create LightPipeline

Light Pipeline

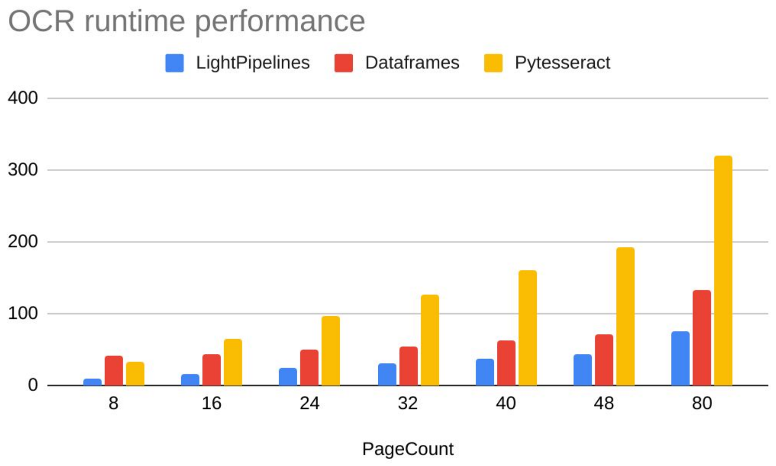
```
from sparkocr.base import LightPipeline
```

```
lp = LightPipeline(pipeline)
```

Spark ML pipeline

```
%%time
lp.fromLocalPath(pdfs_path)
```

```
CPU times: user 4.58 ms, sys: 6.95 ms, total: 11.5 ms
Wall time: 6.24 s
```



See a complete example [here](), and take a look at release notes [here]().