Spark OCR

for Data Scientists

Spark-OCR Team, John Snow Labs

Agenda

Main topic	Introduced Concepts
Introduction	Motivation, Overview and General Features.
Basic Transformations & Pipelines	Basic Transformers and Pipelines: Image Enhancing, Scaling, SkewCorrection. ImageToTextV1 vs. ImageToTextV2. Handwriting detection & recognition examples. Text Detection Examples.
PDF Processing	Pipelines with mixed digital and scanned PDFs. Table detection and extraction; from scanned PDFs and from digital PDFs.
Document Deidentification	Basic Deidentification Pipelines
Visual Document NER	Introduction to the task. Relation Extraction.
Improving Pretrained Models	Visual NER Fine Tuning. Data Labeling on Alab for Visual NER.
Visual Document Classification	Different Visual Document Classification models.
Spark OCR Streaming	Basic Spark OCR Streaming. Rest APIs with Synapse.
Summary and next steps	Next features to be added to Spark OCR.

Presenters today

Alberto

Mykola

Aymane

Introduction

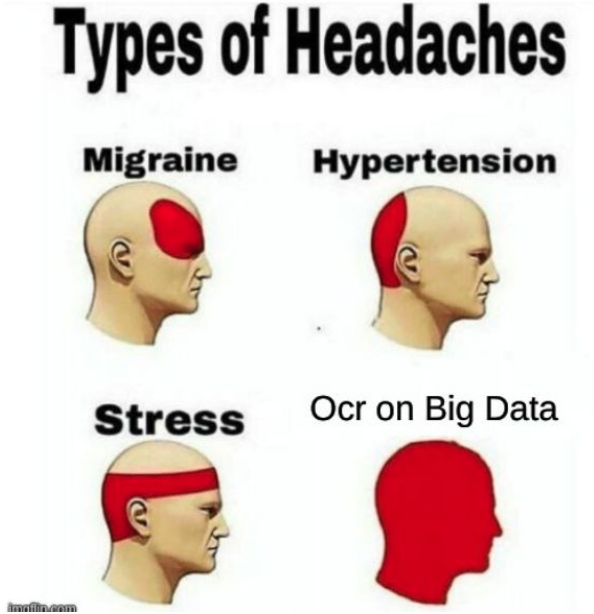
Spark-OCR Team, John Snow Labs

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.
- 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

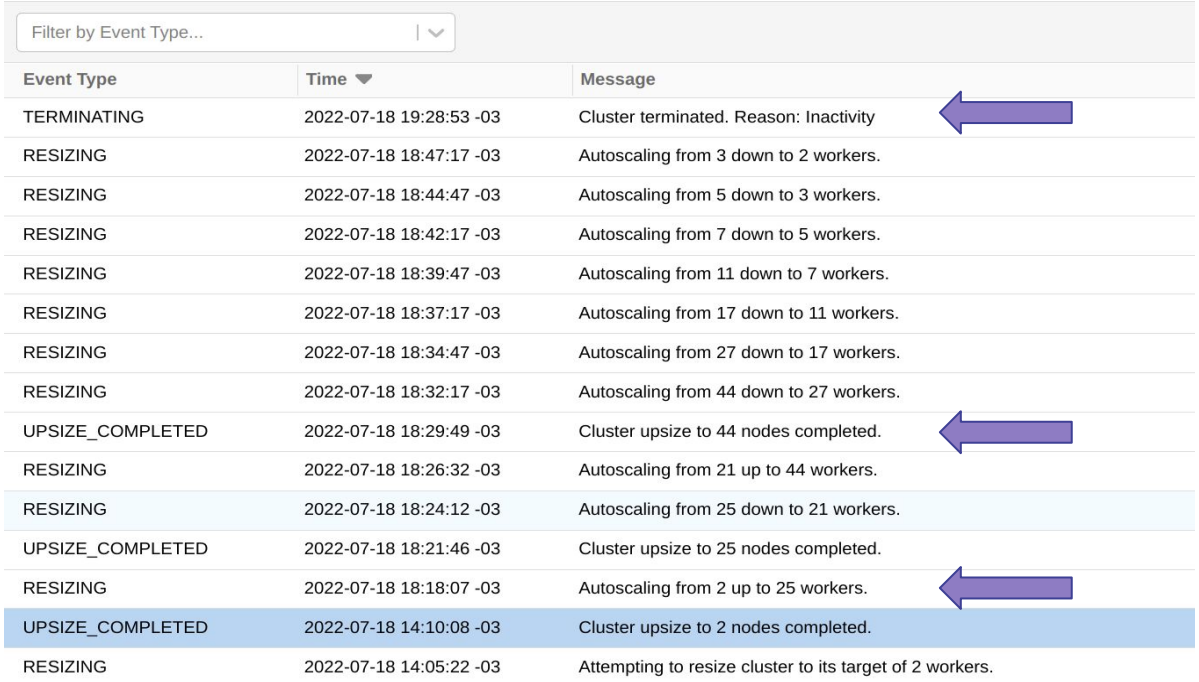
- 111835b vs. 315b, a **355** factor!!.
- Density of information is much lower in OCR than NLP.
- Handling images is challenging.

Motivation

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

Motivation

John Snow LABS

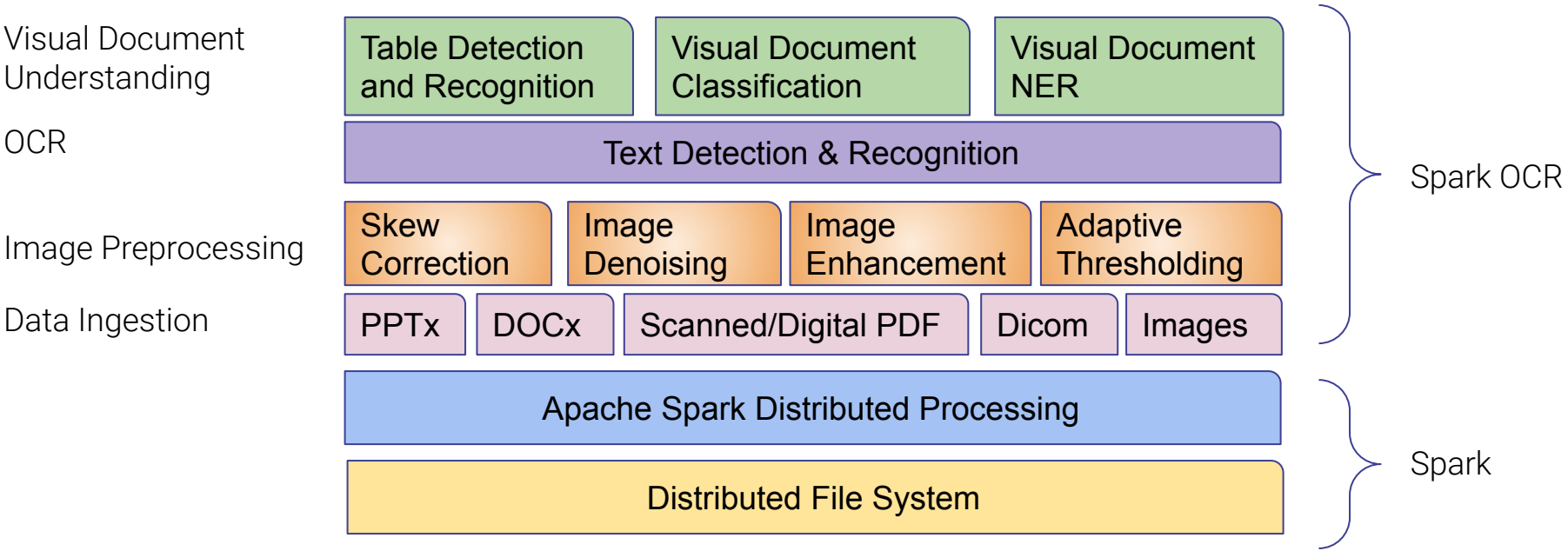


- Different cluster providers with the right maturity

John Snow LABS

Introduction to Spark-OCR

- Spark-OCR 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 and Recognition
Visual Document Classification
Visual Document NER
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

Basic Transformations & Pipelines

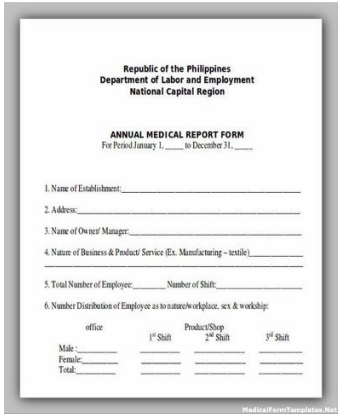
for Data Scientists

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

Name of applicant: _____ Date of Birth: _____
Address: _____
License Number _____ Class of license applied for: _____

Driver Examiner's Use only
VISION SCREENING
AUDIOGRAMS
APPLIANCES
EXAMINER
AUTHORIZED TRAINING
FOR CLASS
UNITS
DD.MM.YY
EXAMINER
DATE
VALID FOR M.S., P.C., H.S.

Medical Examination Report
FOR COMMERCIAL DRIVER FITNESS DETERMINATION

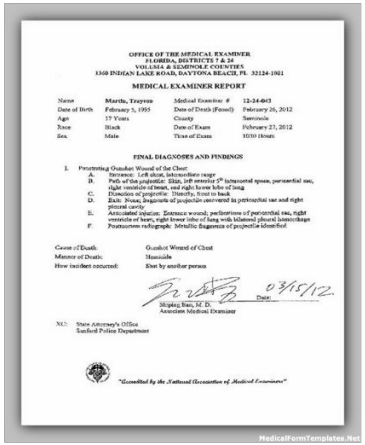
1 DRIVER'S INFORMATION Driver completes this section

2 HEALTH HISTORY Driver completes this section, but medical examiner is encouraged to discuss with driver

Medical Examiner's Comments on Health History (The medical examiner must review and discuss with the driver any "yes" answers and potential hazards of medications, including over-the-counter medications, while driving. This discussion must be documented below.)

I certify that the above information is complete and true. I understand that inaccurate, false or missing information may invalidate the examination and any Medical Examiner's Certificate.

Driver's Signature _____ Date _____



OFFICE OF THE MEDICAL EXAMINER
FLORIDA, DISTRICTS 7 & 24
VOLUSIA & SEMINOLE COUNTIES
1360 INDIAN LAKE ROAD, DAYTONA BEACH, FL 32124-1001

MEDICAL EXAMINER REPORT

Name	Martin, Trayvon	Medical Examiner #	12-24-0413
Date of Birth	February 5, 1995	Date of Death (Found)	February 26, 2012
Age	17 Years	County	Seminole
Race	Black	Date of Exam	February 27, 2012
Sex	Male	Time of Exam	1030 Hours

FINAL DIAGNOSES AND FINDINGS

I. Penetrating Gunshot Wound of the Chest
A. Entrance: Left distal, intermediate range
B. Path: perforation of 4th anterior 5th intercostal space, pericardial sac, right ventricle of heart, and right lower lobe of lung
Direction of projectile: Directly, front to back
C. Exit: None, fragments of projectile recovered in pericardial sac and right pleural cavity
D. Associated injuries: Entrance wound; perforations of pericardial sac, right ventricle of heart, right lower lobe of lung with ipsilateral pleural hemorrhage
Postmortem radiograph: Metallic fragments of projectile identified

Cause of Death: Gunshot Wound of Chest
Manner of Death: Homicide
How incident occurred: Shot by another person

Shipley Rao, M.D.
Associate Medical Examiner
Dated: 03/15/12

NCI: State Attorney's Office
Sanford Police Department

"Accredited by the National Association of Medical Examiners"

Ocr workflow

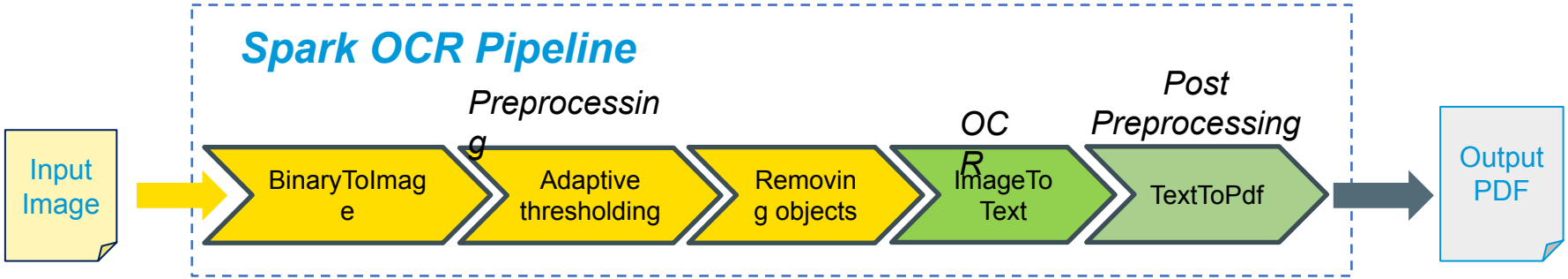


Image Transformations

CPU

ImageTransformer:

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

GPU

GPUImageTransformer:

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

Optical Character Recognition

ImageToText vs ImageToTextV2

John Snow LABS

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
- Work on the line level, for more complex document need to run text detection step

Pdf processing

Pdf transformers

John Snow LABS

- **PdfToText** – extract text from selectable PDF
- **PdfToImage** – render each page 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 to the single page Pdf document.
- **PdfAssembler** - assemble multi page PDF document from single page documents

Questions and Answers

Links

John Snow LABS

- Workshop
- Documentation
- Annotation Lab
- Spark NLP Medium

Plain OCR for Table Recognition

Why can't we use it?

Plain OCR

John Snow LABS

(iii) Series B Financing

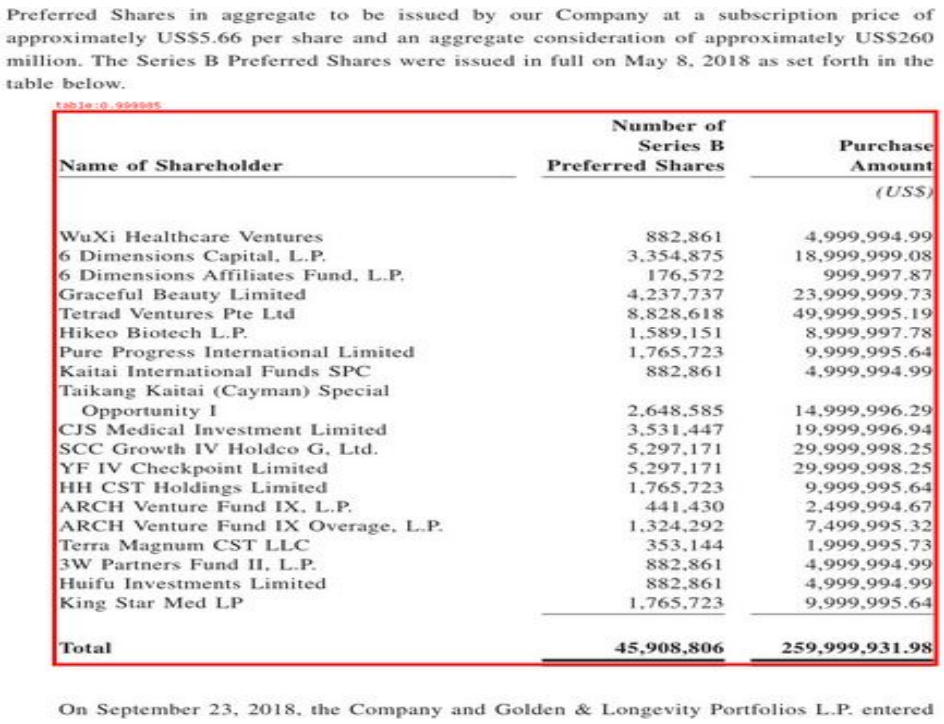
On April 28, 2018, the Company and its subsidiaries entered into the Series B Share Purchase Agreement with the then Series B Preferred Shareholders, pursuant to which the then Series B Preferred Shareholders agreed to subscribe for a maximum of 45,908,818 Series B Preferred Shares in aggregate to be 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?

Table Recognition

Table Region Detection



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

Table Cells Recognition

John Snow LABS

Name of Shareholder	Preferred Shares	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

Data Extraction

John Snow LABS

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 AS48375	18,999.999.08
6 Dimensions Affiliates Fund, L.P.	176.572	999 997.87
Graceful Beauty Limited	4.937.737	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		

Free Text Extraction

1 entry Filter

index	text
0	HISTORY, DEVELOPMENT AND CORPORATE STRUCTURE (iii) Series B Financing On 28, 2018, the and its subsidiaries entered into the Series B Share Purchase April with the thenCompany Series B Preferred Shareholders, to which the then Series B PreferredAgreement Shareholders agreed to subscribe for a maximum of pursuant45,908,818 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. On September 23, 2018, the Company and Golden & Longevity Portfolios L.P. entered into a purchase agreement, pursuant to which Golden & Longevity Portfolios L.P. agreed to purchase 332,165 Series Bmillion. In Preferred Sharesat the Golden& aggregate purchase Portfoliosprice L.P. equivalent to otherapproximately be US$1.88bound theterms addition, and conditions under Longevity the Series B Share agreed to, Purc among and the things, Shareholders by Agreement Agreement. -157-

Show 25 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

De-identification (1)

John Snow LABS

What is De-identification?

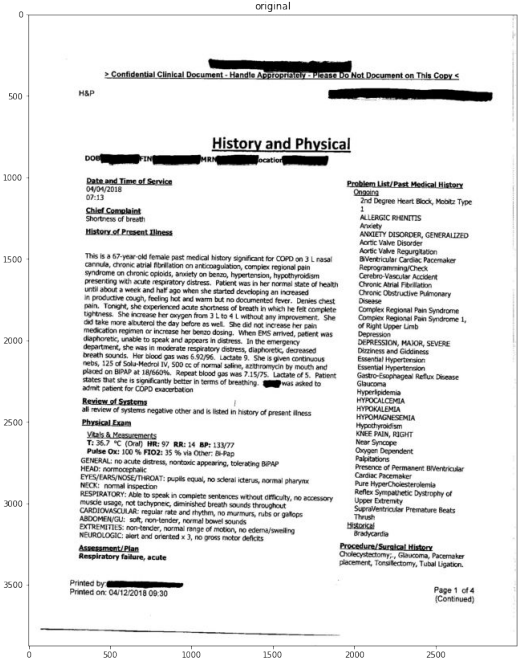
- Simple process & setup
- Automatically de-identify structured data, unstructured data, documents, PDF files, and images in compliance with HIPAA, GDPR, or custom needs
- >99% Accuracy on real-world documents
- Trusted by 5 of 8 Top Pharma Companies

De-identification pipeline

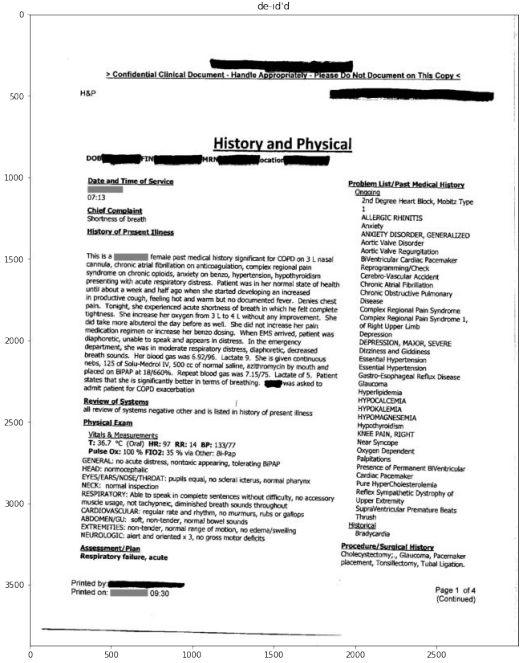
John Snow LABS

BinaryToImage → OCR → De-identification → Position_finder

original



de-id'd



De-identification (3)

```
|ner_chunk
|
+--------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------+
|[[chunk, 193, 202, 04/04/2018, [entity -> DATE, sentence -> 1, chunk -> 0], []], [chunk, 435, 445, 67-year-old, [entity -> AGE, sentence
-> 2, chunk -> 1], []], [chunk, 3367, 3373, Qi neem, [entity -> NAME, sentence -> 19, chunk -> 2], []], [chunk, 3388, 3397, 04/12/2018, [e
ntity -> DATE, sentence -> 20, chunk -> 3], []]]|
+--------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------+
```

```
+-----------------------------------------------------------------------------------------------------------------------------------------------------+
|coordinates                                                                                                                                          |
+-----------------------------------------------------------------------------------------------------------------------------------------------------+
|[[0, 0, 356.0, 1053.0, 217.0, 41.0], [1, 0, 518.0, 1467.0, 210.0, 43.0], [3, 0, 495.0, 3527.0, 231.0, 43.0]]|
+-----------------------------------------------------------------------------------------------------------------------------------------------------+
```

Visual NER

- Named-entity recognition
- Use text and layout data
- SROIE dataset

(a)

(b)

(c)

(d)

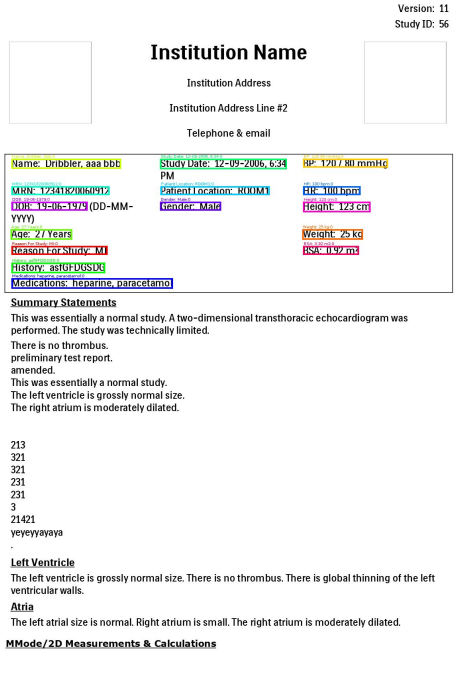
(e)

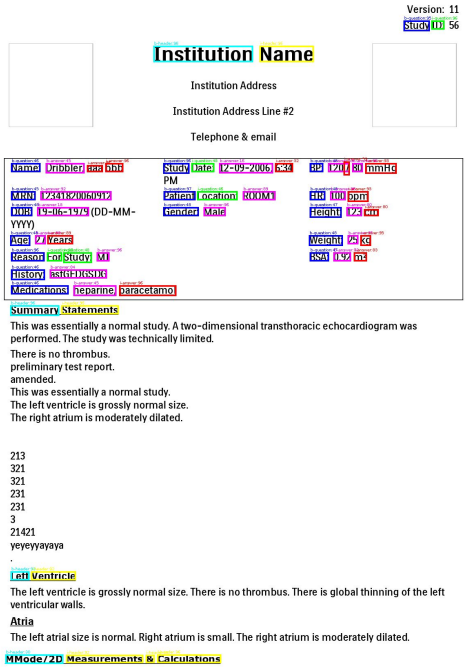
Visual NER - situation #1 - entities alone

IMPRESSION: At this time is refractory anemia, which is transfusion dependent. He is on B12, iron, folic acid, and Procrit. There are no sign or symptom of blood loss and a recent esophagogastroduodenoscopy, which was negative. His creatinine was 1. My impression at this time is that he probably has an underlying myelodysplastic syndrome or bone marrow failure. His creatinine on this hospitalization was up slightly to 1.6 and this may contribute to his anemia.

Visual NER - situation #2 Forms

key	value
Name:	Dribbler, bbb
Study Date:	12-09-2006, 6:34
BP:	120 80 mmHg
MRN:	12341820060912
Patient Location:	ROOM1
HR:	100 bpm
DOB:	19-06-1979
Gender:	Male
Height:	123 cm
Age:	27 Years
Weight:	25 kg
Reason For Study:	MI
BSA:	0.92 m2
History:	asfGFDGSDG
Medications:	heparine, paracetamol





Visual NER Fine-Tuning

Otitis Media - Discharge Summary

Description: Fever, otitis media, and possible sepsis. (Medical Transcription Sample Report)

ADMITTING DIAGNOSES:
1. Fever.
2. Otitis media.
3. Possible sepsis.

HISTORY OF PRESENT ILLNESS: The patient is a 10-month-old male who was seen in the office 1 day prior to admission. He has had a 2-day history of fever that has gone up to as high as 103.6 degrees F. He has also had intermittent cough, nasal congestion, and rhinorrhea and no history of rashes. He has been taking Tylenol and Advil to help decrease the fevers, but the fever has continued to rise. He was noted to have some increased workup of breathing and parents returned to the office on the day of admission.

PAST MEDICAL HISTORY: Significant for being born at 33 weeks' gestation with a birth weight of 5 pounds and 1 ounce.

PHYSICAL EXAMINATION: On exam, he was moderately ill appearing and lethargic. HEENT: Atraumatic, normocephalic. Pupils are equal, round, and reactive to light. Tympanic membranes were red and yellow, and opaque bilaterally. Nares were patent. Oropharynx was slightly moist and pink. Neck was soft and supple without masses. Heart is regular rate and rhythm without murmurs. Lungs showed increased workup of breathing, moderate tachypnea. No rales, rhonchi or wheezes were noted. Abdomen: Soft, nontender, nondistended. Active bowel sounds. Neurologic exam showed good muscle strength, normal tone. Cranial nerves II through XII are grossly intact.

LABORATORY FINDINGS: He had electrolytes, BUN and creatinine, and glucose all of which were within normal limits. White blood cell count was 8.6 with 61% neutrophils, 21% lymphocytes, 17% monocytes, suggestive of a viral infection. Urinalysis was completely unremarkable. Chest x-ray showed a suboptimal

Otitis Media - Discharge Summary

Description: Fever, otitis media, and possible sepsis. (Medical Transcription Sample Report)

ADMITTING DIAGNOSES:
1. Fever.
2. Otitis media.
3. Possible sepsis.

HISTORY OF PRESENT ILLNESS: The patient is a 10-month-old male who was seen in the office 1 day prior to admission. He has had a 2-day history of fever that has gone up to as high as 103.6 degrees F. He has also had intermittent cough, nasal congestion, and rhinorrhea and no history of rashes. He has been taking Tylenol and Advil to help decrease the fevers, but the fever has continued to rise. He was noted to have some increased workup of breathing and parents returned to the office on the day of admission.

PAST MEDICAL HISTORY: Significant for being born at 33 weeks' gestation with a birth weight of 5 pounds and 1 ounce.

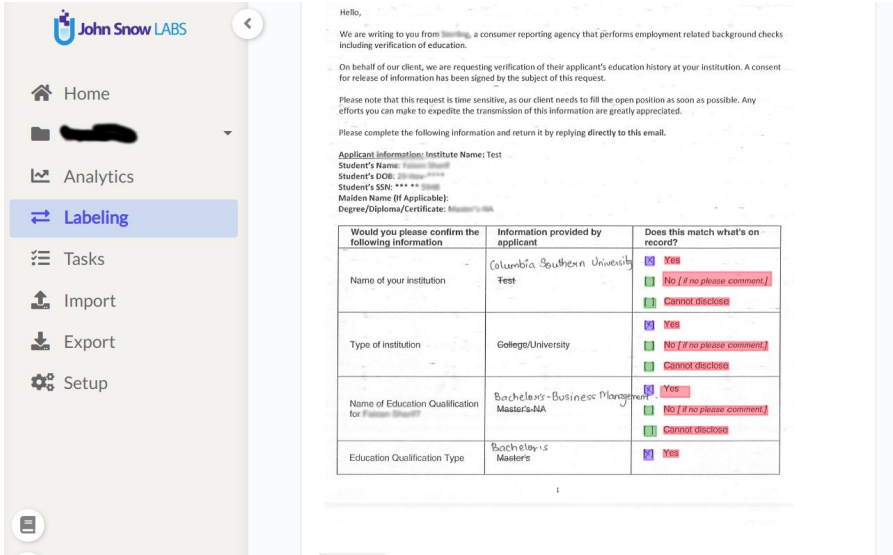
PHYSICAL EXAMINATION: On exam, he was moderately ill appearing and lethargic. HEENT: Atraumatic, normocephalic. Pupils are equal, round, and reactive to light. Tympanic membranes were red and yellow, and opaque bilaterally. Nares were patent. Oropharynx was slightly moist and pink. Neck was soft and supple without masses. Heart is regular rate and rhythm without murmurs. Lungs showed increased workup of breathing, moderate tachypnea. No rales, rhonchi or wheezes were noted. Abdomen: Soft, nontender, nondistended. Active bowel sounds. Neurologic exam showed good muscle strength, normal tone. Cranial nerves II through XII are grossly intact.

LABORATORY FINDINGS: He had electrolytes, BUN and creatinine, and glucose all of which were within normal limits. White blood cell count was 8.6 with 61% neutrophils, 21% lymphocytes, 17% monocytes, suggestive of a viral infection. Urinalysis was completely unremarkable. Chest x-ray showed a suboptimal

Improving Pretrained models: Data Labeling on Alab for visual NER.

Deploy on AWS, Azure or locally on your linux VMs,

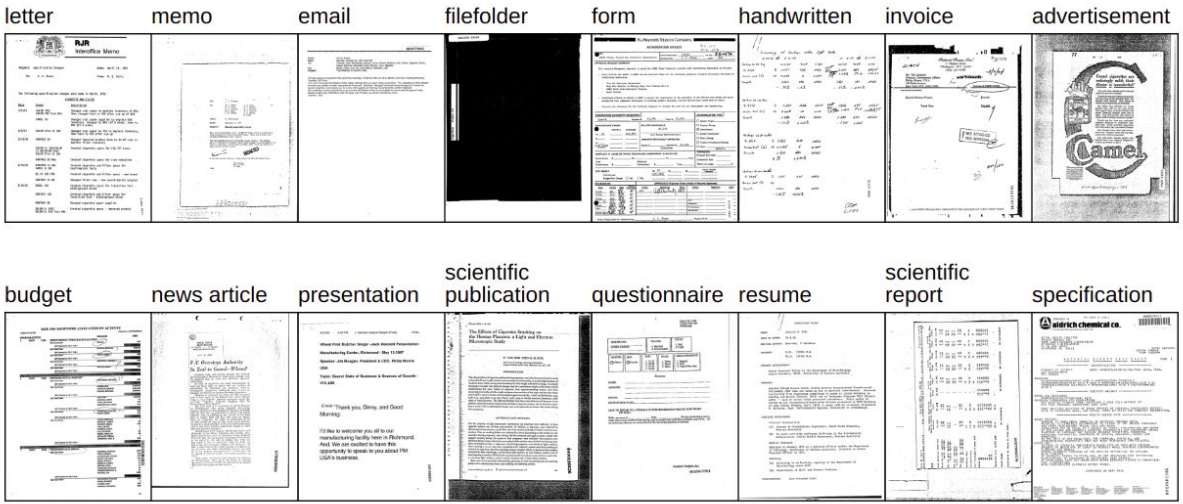
https://www.johnsnowlabs.com/install/



Questions and Answers

Multimodal Document Classification

Document Classification

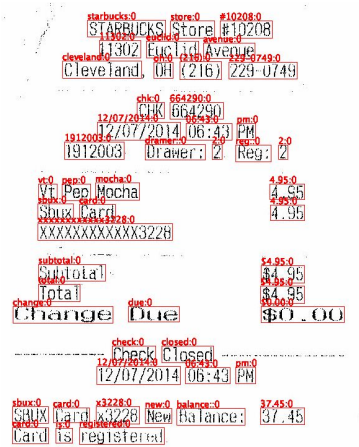


General Idea: provide different models according to whether they use *text, layout, image* or a combination of them.

Document Classification Pipeline

BinaryToImage → ImageToHocr → HOcrTokenizer → VisualDocumentClassifier

HOcr(XML)



Receipt

John Snow LABS

Document Classification in Spark OCR

Annotator	How it works	Acc %	Dataset
VisualDocumentClassifierV1	Use text, and layout information to make a decision.	92.12%	Tobacco3482: 10 Categories including: letter, form, image, resume, and memo.
VisualDocumentClassifierV2	Use image, text, and layout information to make a decision.	88%	RVL-CDIP: 16 categories.
VisualDocumentClassifierV3	Uses only image information to make decisions	92.3%	

John Snow LABS

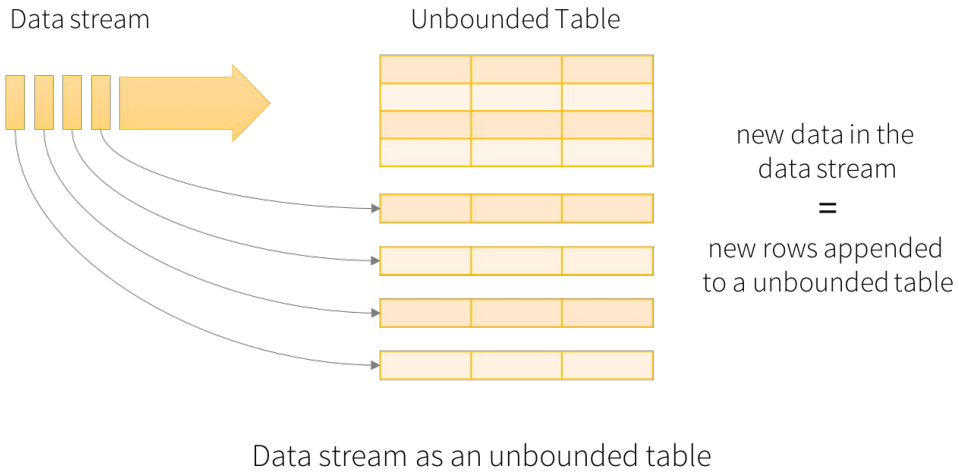
Spark OCR Streaming

Spark Structured Streaming

"Structured Streaming is a scalable and fault-tolerant stream processing engine built on the Spark SQL engine.
You can express your streaming computation the same way you would express a batch computation on static data.
The Spark SQL engine will take care of running it incrementally and continuously and updating the final result as streaming data continues to arrive"

Complete Spark Structured Streaming Guide here.

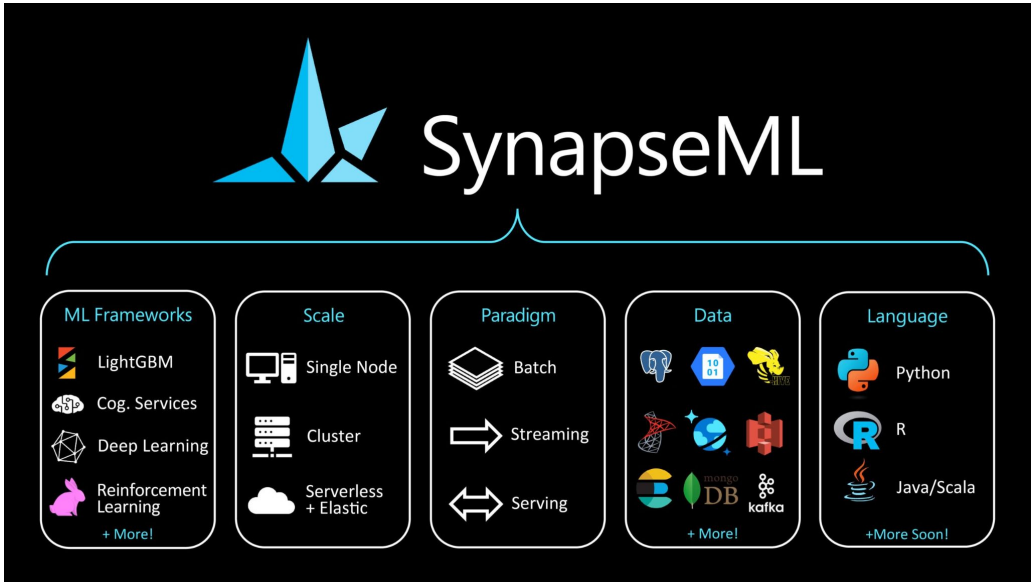
Basic concepts



Data stream as an unbounded table

- Sources and Sinks.
- Streaming Dataframe.
- Unbounded Table.

Serving with Synapse



As defined in the documentation page,

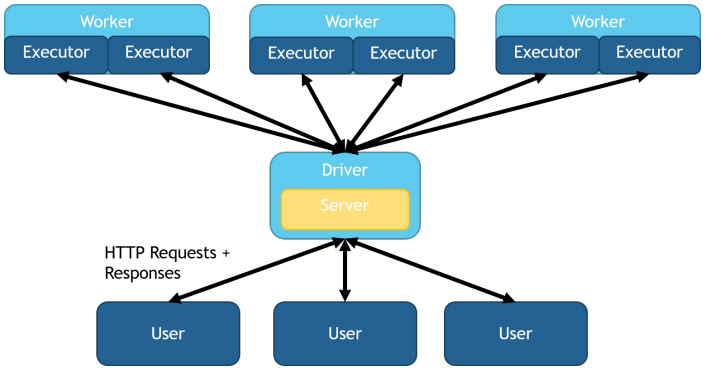
"An ecosystem of tools aimed towards expanding the distributed computing framework Apache Spark in several new directions."

Serving with Synapse

- Ready-to-use server
- Includes a Load Balancer
- Distributes the work over a Spark Cluster
- Can be used for both Spark NLP and Spark OCR

https://www.johnsnowlabs.com/serving-spark-nlp-via-api-1-3-microsofts-synapse-ml/

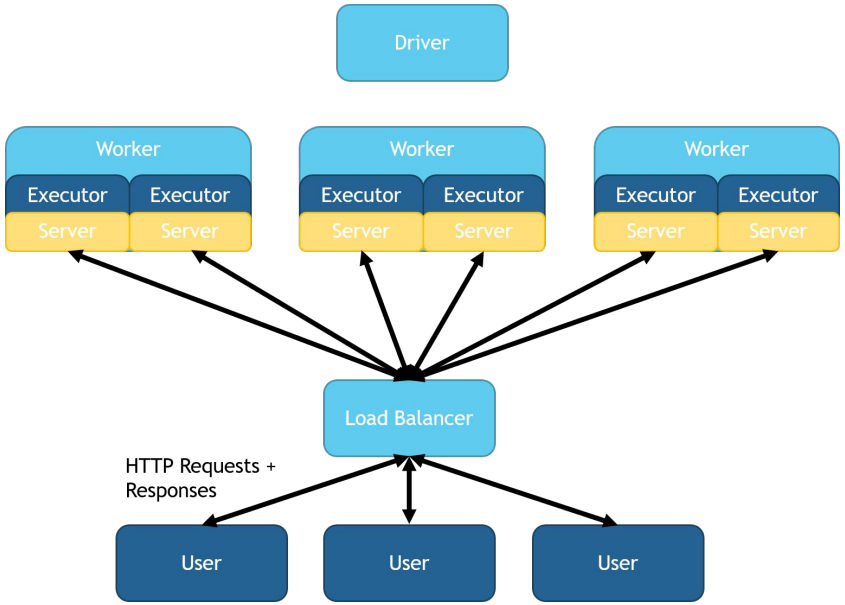
Head node load balanced



- Spins up a queue on the head node, distributes work across partitions, then collects response data back to the head node.
- Allows for more complex windowing, repartitioning, and SQL operations.
- Ideal for rapid setup and testing, no additional load balancing or network switches.

Read more: https://microsoft.github.io/SynapseML/docs/features/spark_serving/about/

Fully Distributed



- This mode spins up servers on each executor JVM
- Each server will feed its executor's partitions in parallel.
- This mode is key for high throughput and low latency as data does not need to be transferred to and from the head node

Summary and Future Roadmap

- Make everything trainable.
- Integration with other cluster providers, and other interesting frameworks(e.g., mlflow).
- Better Integration with Annotation Lab.
- Improve table recognition.
- Continue to profile the library.
- New models: image quality ranking, visual question answering, chart recognition.
- Continue to leverage our experience and materialize it as best practices: pretrained pipelines.

Questions and Answers

Contact Us!

Contact us on Slack!

Emails: alberto@johnsnowlabs.com,
aymane@johnsnowlabs.com,
mykola@johnsnowlabs.com,
enes@johnsnowlabs.com