



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



Gokham



Gursey



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.

Types of Headaches

Migraine



Hypertension



Stress



Ocr on Big Data



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 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.

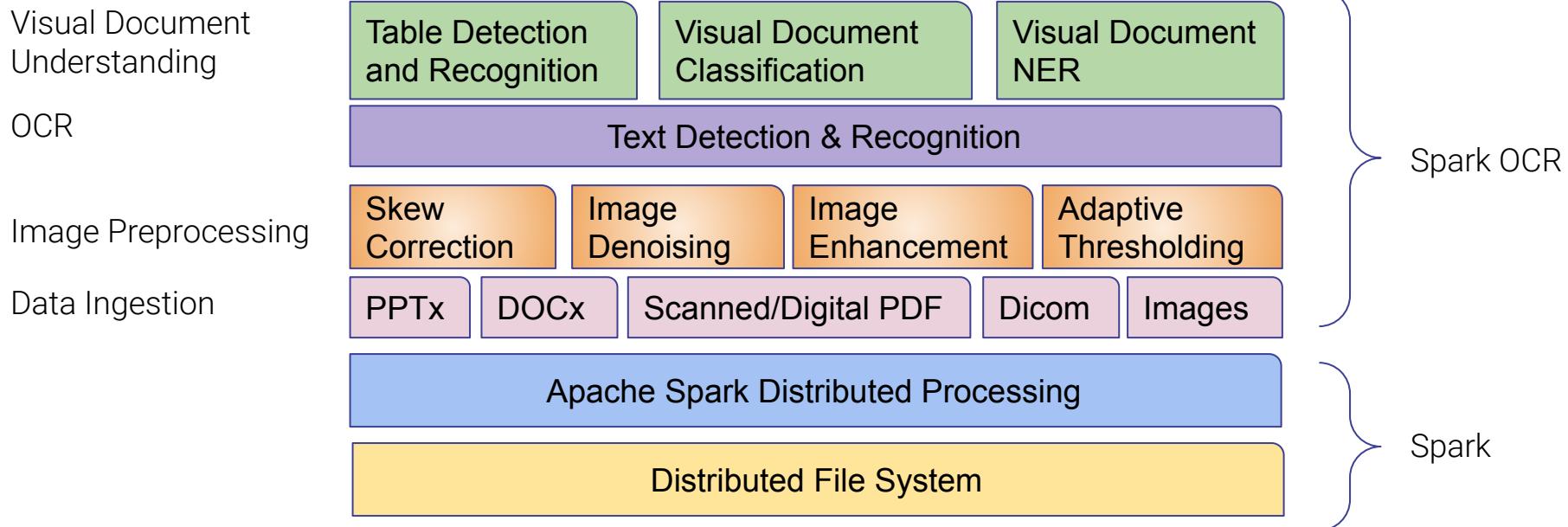
Motivation

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 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.





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

<p align="center">OFFICE OF THE MEDICAL EXAMINER 210 ORLANDA, DISTRICT 5 & 24 ORLANDA, FLORIDA 32801-3225 1300 INDIAN LAKE ROAD, DAYTONA BEACH, FL 32114-1061</p> <p align="center">MEDICAL EXAMINER REPORT</p> <p>Name: Marie, Trayvon Medical Examiner # 12-14-043</p> <p>Date of Birth: February 5, 1995 Date of Death (if known): February 26, 2012</p> <p>Age: 17 Sex: Male Date of Exam: February 27, 2012</p> <p>Sex: Male Time of Death: 10:00 Hours</p> <p align="center">FINAL DIAGNOSES AND FINDINGS</p> <p>1. Petechial Gunshot Wound of the Chest A. Wounds of the chest wall, right side, multiple, irregular, non-healing, necrotic, tan/yellow. B. Path of the projectile, skin, skin, left sternocostal junction, superficial, petechial, tan, irregular, non-healing, necrotic, tan/yellow. C. Dimensions of projectile, diameter, front to back, 1 cm. D. Dimensions of projectile, depth, front to back, 1 cm. E. Irregular, non-healing, necrotic, tan/yellow, petechial area, right sternocostal junction. F. Right lung, pleural cavity, external surface, petechial, petechial area, right sternocostal junction, right lower lobe of lung with bilateral pleural hemorrhage. G. Postmortem refrigeration. Metallic fragments of projectile identified.</p> <p>Cause of Death: Gunshot Wound of Chest Manner of Death: Homicide How death occurred: Killed by person</p> <p align="right"> John S. St. John, M.D. Board Certified American Academy of Forensic Medicine</p> <p align="right">03/05/12</p> <p>NIC: State Attorney's Office Lead/Police Department</p> <p align="center"></p> <p align="center"><i>"Certified by the National Commission of Medical Examiners"</i></p>

Ocr workflow

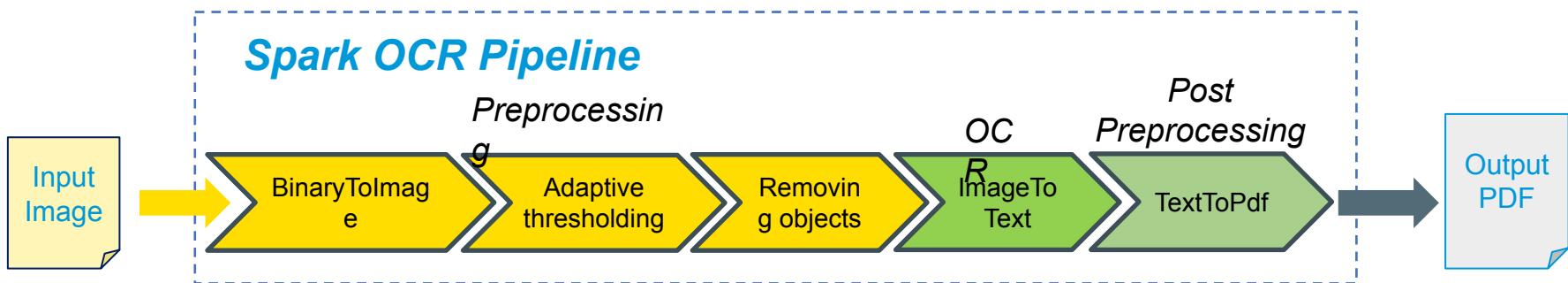


Image Transformations

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

Pdf processing

Pdf transformers



- **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



- [Workshop](#)
- [Documentation](#)
- [Annotation Lab](#)
- [Spark NLP Medium](#)

Plain OCR for Table Recognition

Why can't we use it?

Plain OCR

(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.

Number of Series B Purchase Name of Shareholder (US\$)	Preferred Shares	Amount
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 10.999985

Name of Shareholder	Number of Series B Preferred Shares	Purchase Amount (USS)
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
Hiketo 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

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

Name of Shareholder	Number of Preferred Shares		Purchase (US\$)
	Series B	Purchase	
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

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 Preferred Agreement Shareholders agreed to subscribe for a maximum of pursuant 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. 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 B million. In Preferred Shares at the Golden& aggregate purchase Portfolios price L.P. equivalent to other approximately be US\$1.88bound the terms 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.

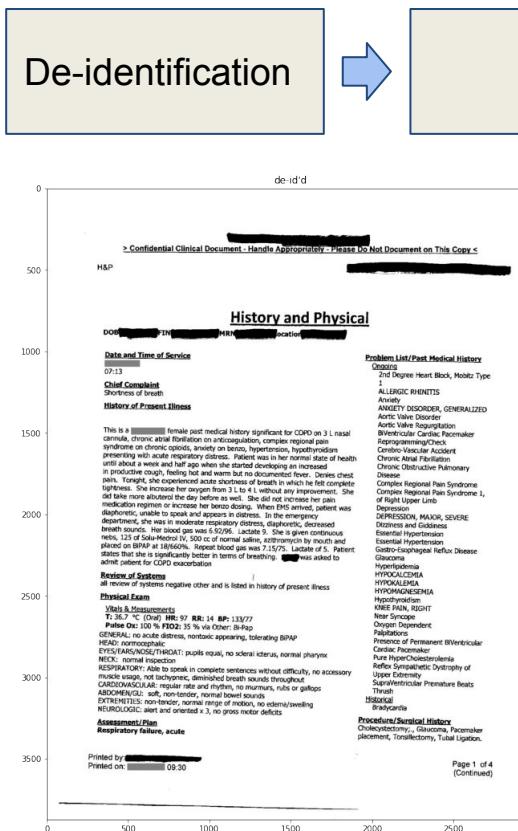
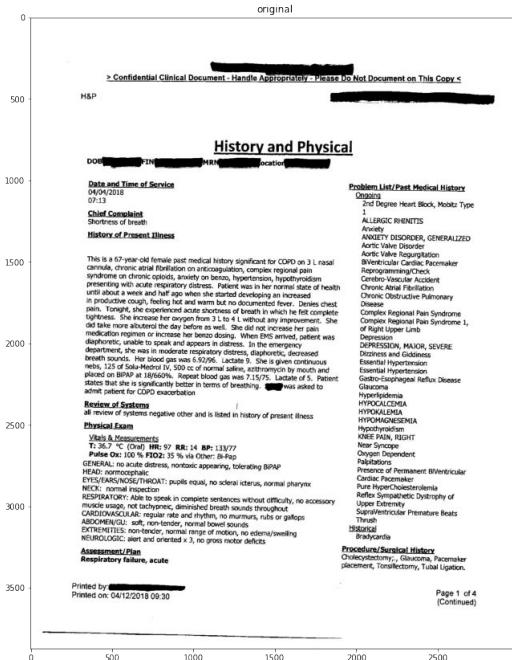
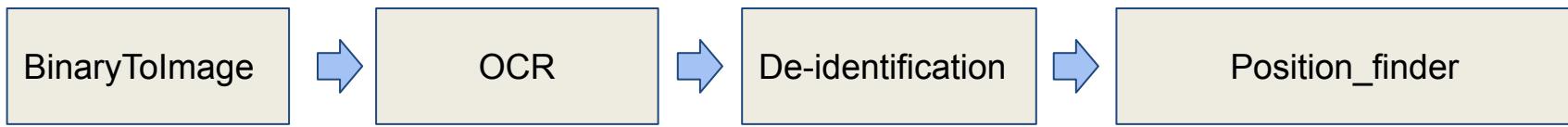
De-identification (1)

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



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)

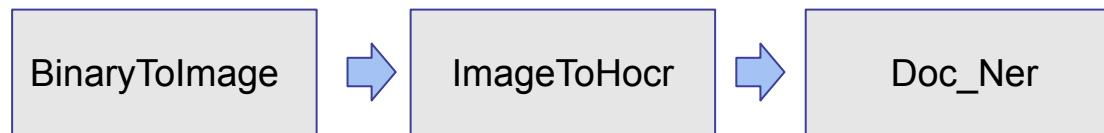


(d)



(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

Version: 11
Study ID: 56

Institution Name

Institution Address

Institution Address Line #2

Telephone & email

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

Summary Statements

This was essentially a normal study. A two-dimensional transthoracic echocardiogram was performed. The study was technically limited.

There is no thrombus.
preliminary test report.
amended.

This was essentially a normal study.
The left ventricle is grossly normal size.
The right atrium is moderately dilated.

213
321
321
231
231
3
21421
yeyeyayaya

Left Ventricle

The left ventricle is grossly normal size. There is no thrombus. There is global thinning of the left ventricular walls.

Atria

The left atrial size is normal. Right atrium is small. The right atrium is moderately dilated.

MMode/2D Measurements & Calculations

Version: 11
Study ID: 56

Institution Name

Institution Address

Institution Address Line #2

Telephone & email

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

Summary Statements

This was essentially a normal study. A two-dimensional transthoracic echocardiogram was performed. The study was technically limited.

There is no thrombus.
preliminary test report.
amended.

This was essentially a normal study.
The left ventricle is grossly normal size.
The right atrium is moderately dilated.

213
321
321
231
231
3
21421
yeyeyayaya

Left Ventricle

The left ventricle is grossly normal size. There is no thrombus. There is global thinning of the left ventricular walls.

Atria

The left atrial size is normal. Right atrium is small. The right atrium is moderately dilated.

MMode/2D Measurements & Calculations

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.

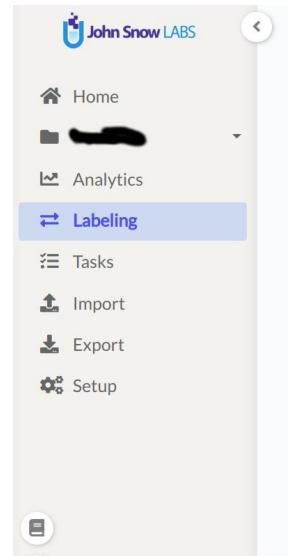
PHYSICAL EXAMINATION: On exam, he was moderately ill appearing and lethargic. HEENT: Atraumatic, normocephalic. Pupils are equal, round, and reactive to light; Tympanic membranes were red and yellow, and opaque bilaterally. Nares were patent. Oropharynx was slightly moist and pink. Neck was soft and 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/>



Hello,
 We are writing to you from [REDACTED] a consumer reporting agency that performs employment related background checks including verification of education.
 On behalf of our client, we are requesting verification of their applicant's education history at your institution. A consent for release of information has been signed by the subject of this request.
 Please note that this request is time sensitive, as our client needs to fill the open position as soon as possible. Any efforts you can make to expedite the transmission of this information are greatly appreciated.

Please complete the following information and return it by replying directly to this email.

Applicant Information/ Institute Name: Test

Student's Name: [REDACTED]

Student's DOB: [REDACTED]

Student's SSN: [REDACTED]

Maiden Name (If Applicable):

Degree/Diploma/Certificate: [REDACTED]

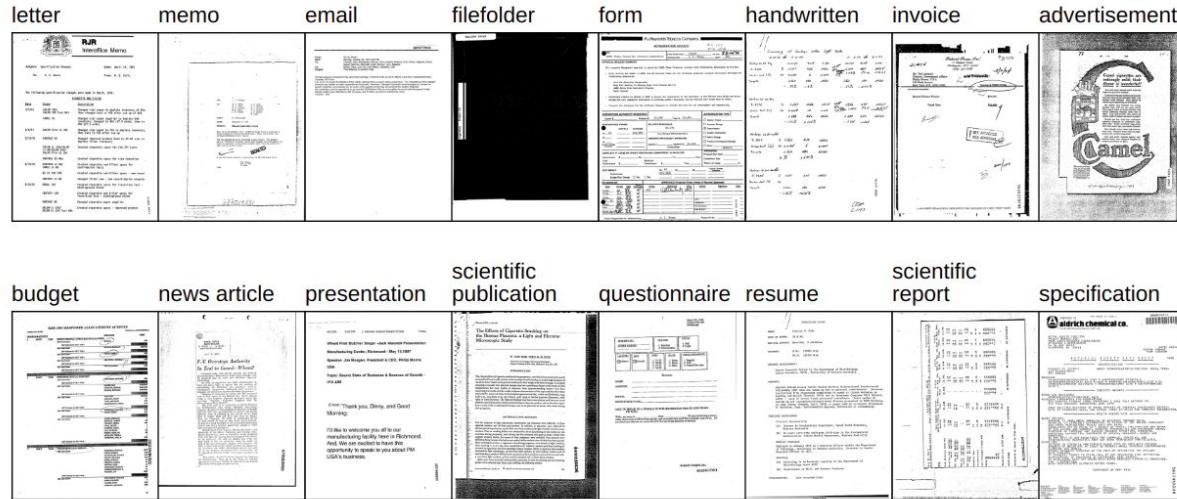
Would you please confirm the following information	Information provided by applicant	Does this match what's on record?
Name of your institution	Columbia Southern University Test	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No (if no please comment) <input type="checkbox"/> Cannot disclose
Type of institution	College/University	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No (if no please comment) <input type="checkbox"/> Cannot disclose
Name of Education Qualification for [REDACTED]	Bachelor's -Business Management Master's-NA	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No (if no please comment) <input type="checkbox"/> Cannot disclose
Education Qualification Type	Bachelor's Master's	<input checked="" type="checkbox"/> Yes

Questions and Answers



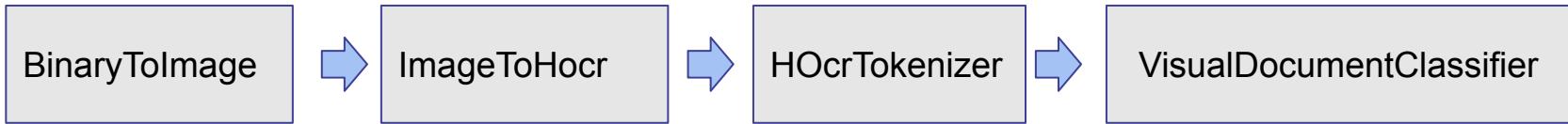
Multimodal Document Classification

Document Classification



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

Document Classification Pipeline



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: \$7.45
Card is registered.

HOcr(XML)

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

chk:0 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:0 closed:0
Check Closed
12/07/2014 06:43 PM

sbusx:0 card:0 \$3228.0 new:0 balance:0 \$7.45
Sbux Card x3228 New Balance: \$7.45
Card is registered.

Receipt

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%	

Spark OCR Streaming

Spark Structured Streaming

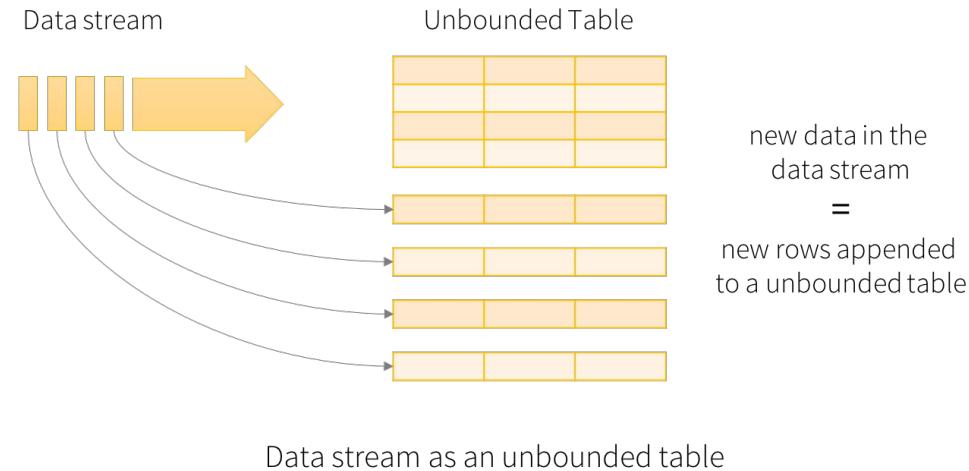
"Structured Streaming is a scalable and fault-tolerant stream processing engine built on the Spark SQL engine.

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"

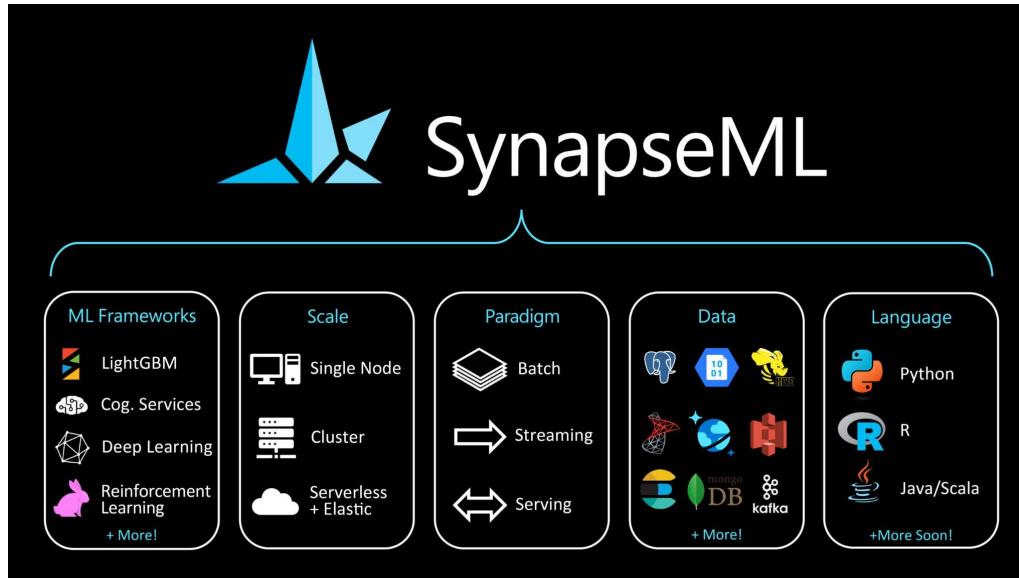
[Complete Spark Structured Streaming Guide here.](#)

Basic concepts



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

Serving with Synapse



As defined in the documentation page,

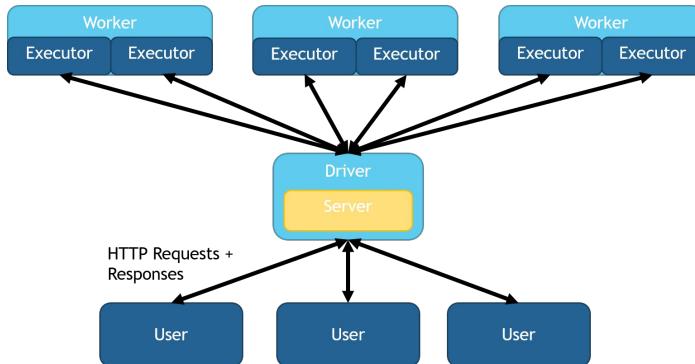
“An ecosystem of tools aimed towards expanding the distributed computing framework [Apache Spark](#) in several new directions.”

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/>

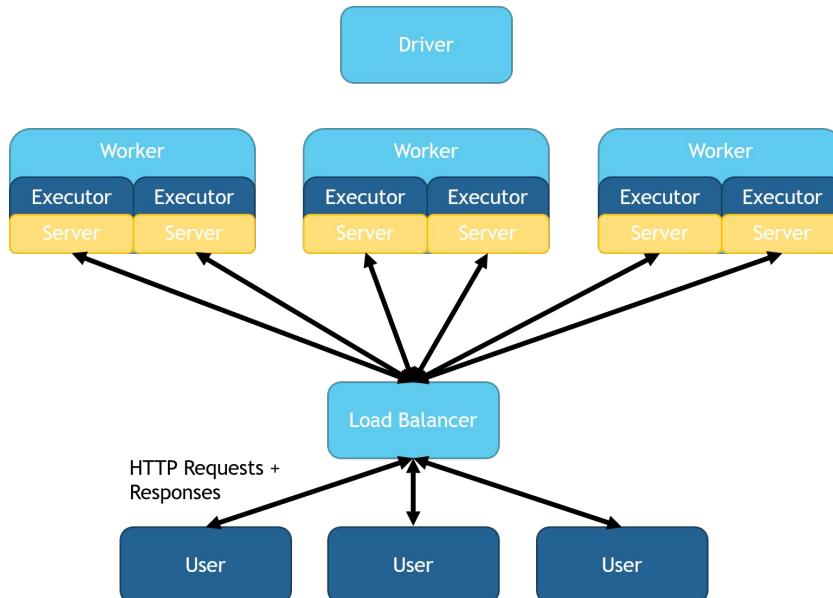
Head node load balanced



- 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