



# Applied Generative AI for Data Scientists

---

July 2024

# Contents

- Introduction to the John Snow Labs ecosystem
- Open-Source language models
- Pipelines
- Multi-modal pipelines
- Domain-Specific language models
- API endpoints

# Spark NLP Pipelines

# Spark NLP Pipelines

## Pipelines

- Custom stages
- Can process huge amounts of data (Spark Data Frames)
- End-to-end solutions: data processing, model training or inference

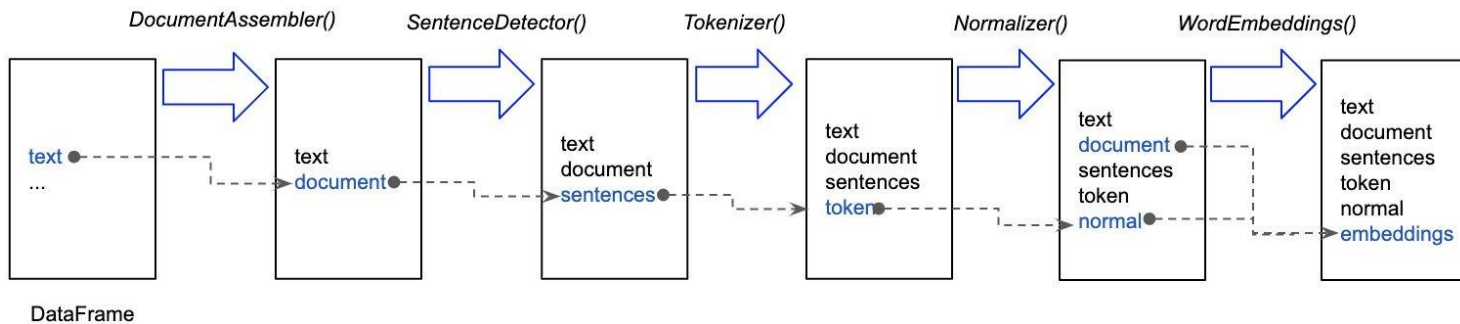
## Light Pipelines

- Fast inference
- Not run in Spark (single multithreaded machine)
- Use (list of) strings directly (no Spark Data Frame)

## Pretrained Pipelines

- Fixed components or stages
- Built to solve specific problems
- Easy to use

# Spark NLP Pipelines



```
from pyspark.ml import Pipeline

document_assembler = DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")

sentence_detector = SentenceDetector()\
    .setInputCols(["document"])\
    .setOutputCol("sentences")

tokenizer = Tokenizer()\
    .setInputCols(["sentences"])\
    .setOutputCol("token")

normalizer = Normalizer()\
    .setInputCols(["token"])\
    .setOutputCol("normal")

word_embeddings = WordEmbeddingsModel.pretrained()\
    .setInputCols(["document", "normal"])\
    .setOutputCol("embeddings")

nlpPipeline = Pipeline(stages=[
    document_assembler,
    sentence_detector,
    tokenizer,
    normalizer,
    word_embeddings,
])

nlpPipeline.fit(df).transform(df)
```

# LLM stages in Pipelines

Spark NLP supports the following NLP tasks using LLMs

- Text Summarization
- Question Answering
- Text Generation
- Neural Machine Translation
- Text Classification

## Encoder only\*

ALBERT, BERT, CamemBERT, DeBERTa, DistilBERT, Longformer, MPNet, RoBERTa, XlmRoBERTa

---

## Encoder-Decoder

BART, Flan T5, MarianMT, M2M100, T5, XLNet

---

## Decoder only

GPT 2, Llama 2, Mistral, Phi 2

\*Heads can be applied to perform, e.g., classification or Question-Answering

# Other LLM Features

## Utilities and integration to third-party software

### LLM

- OpenAI Chat Completions
- OpenAI Embeddings
- LLMLoader (gguf)

Make API calls to OpenAI models directly from Spark NLP Pipelines and load GGUF models.

[Notebook with OpenAI integration](#)

### RAG

- Document Splitting
- LangChain
- Haystack

Build RAG applications using Spark NLP models on LangChain or Haystack.

[Notebook with Langchain integration](#)

[Notebook with Haystack integration](#)

# Annotator Classes

## Transformers LLMs

Albert

Bert

CamemBert

DeBerta

DistilBert

Longformer

MPNet

RoBerta

XlmRoBerta

Xlnet

Embeddings

SentenceEm  
beddings

ForTokenCla  
ssification

ForSequence  
Classification

ForQuestionA  
nswering

ForZeroShot  
Classification

## Embeddings Special Cases

BGEEembeddings

E5Embeddings

ElmoEmbeddings

InstructorEmbedd  
ings

UAEmbeddings

UniversalSentenc  
eEncoder

## Seq2Seq

BartTransformer

GPT2Transformer

LLAMA2Transformer

M2M100Transformer

MarianTransformer

MistralTransformer

OpenAICompletion

Phi2Transformer

T5Transformer

## Healthcare LLM

JSL\_MedSNer\_ZS

JSL\_MedM

JSL\_MedS

MedicalBertForSequenc  
eClassification

MedicalBertForTokenCl  
assifier

MedicalDistilBertForSe  
quenceClassification

MedicalSummarizer

MedicalTextGenerator

Text2SQL



# Coding time!

---

Pipelines - LLM

# Multimodal Pipelines

# Multimodal Pipelines

## Tabular

- Question-Answering on tables
  - TapasForQuestionAnswering

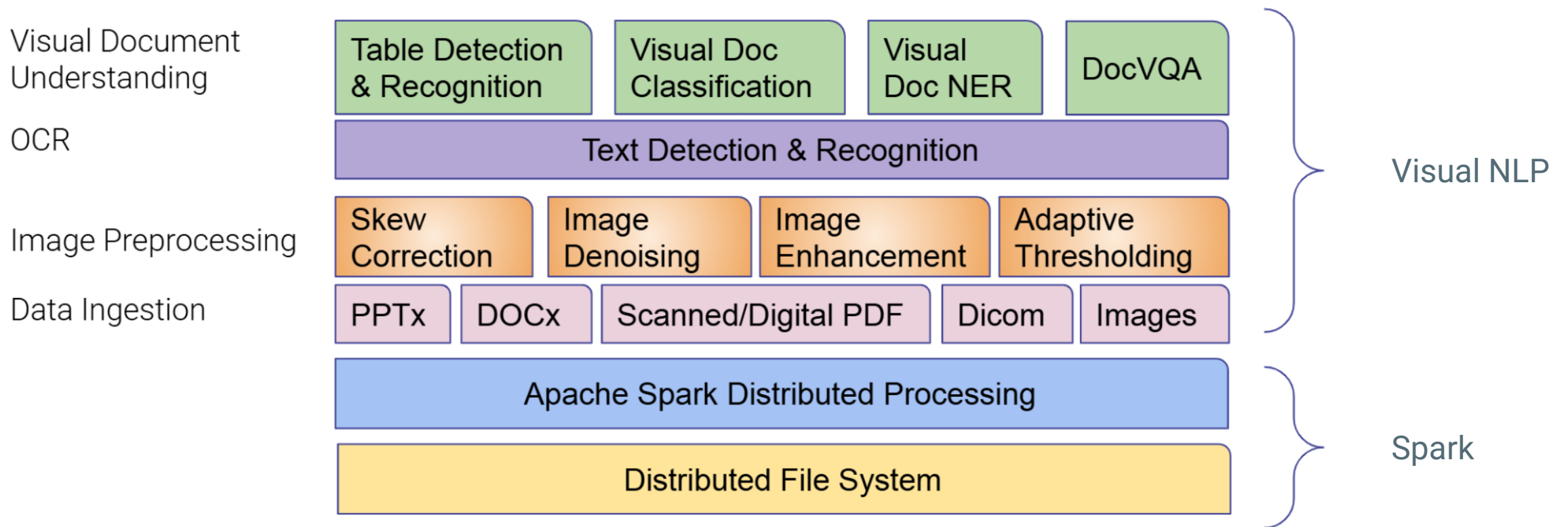
## Audio

- Automatic Speech Recognition (ASR)
  - HubertForCTC
  - Wav2Vec2ForCTC
  - WhisperForCTC

## Image

- Image Classification
  - CLIPForZeroShotClassification
  - ConvNextForImageClassification
  - SwinForImageClassification
  - ViTForImageClassification
- Image Captioning
  - VisionEncoderDecoderForImageCaptioning
- **Visual NLP:** licensed library for Computer Vision tasks

# Visual NLP



# Coding time!

---

Multimodal Pipelines - ASR

Multimodal Pipelines - Visual NLP