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

John Snow LABS

Pipelines

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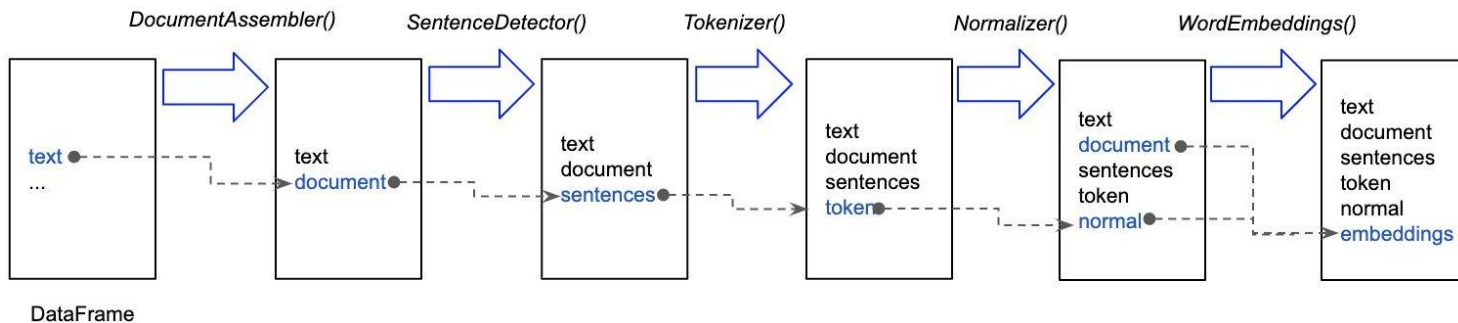- End-to-end solutions: data processing, model training or inference

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

John Snow LABS


```python
from pyspark.ml import Pipeline

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

sentenceDetector = 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,
  sentenceDetector,
  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

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)

RAG
- Document Splitting
- LangChain
- Haystack

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

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

Notebook with OpenAI integration

Notebook with Langcahin integration

Notebook with Haystack integration

John Snow LABS

Annotator Classes

John Snow LABS

Transformers LLMs

- Albert
- Bert
- CamemBert
- DeBerta
- DistilBert
- Longformer
- MPNet
- RoBerta
- XlmRoBerta
- Xlnet

- Embeddings
- SentenceEmbeddings
- ForTokenClassification
- ForSequenceClassification
- ForQuestionAnswering
- ForZeroShotClassification

Embeddings Special Cases

- BGEEmbeddings
- E5Embeddings
- ElmoEmbeddings
- InstructorEmbeddings
- UAEEmbeddings
- UniversalSentenceEncoder

Seq2Seq

- BartTransformer
- GPT2Transformer
- LLAMA2Transformer
- M2M100Transformer
- MarianTransformer
- MistralTransformer
- OpenAICompletion
- Phi2Transformer
- T5Transformer

Healthcare LLM

- **JSL_MedSNer_ZS**
- **JSL_MedM**
- **JSL_MedS**
- **MedicalBertForSequenceClassification**
- **MedicalBertForTokenClassifier**
- **MedicalDistilBertForSequenceClassification**
- **MedicalSummarizer**
- **MedicalTextGenerator**
- **Text2SQL**

Multimodal Pipelines

Multimodal Pipelines

John Snow LABS

Tabular

- Question-Answering on tables
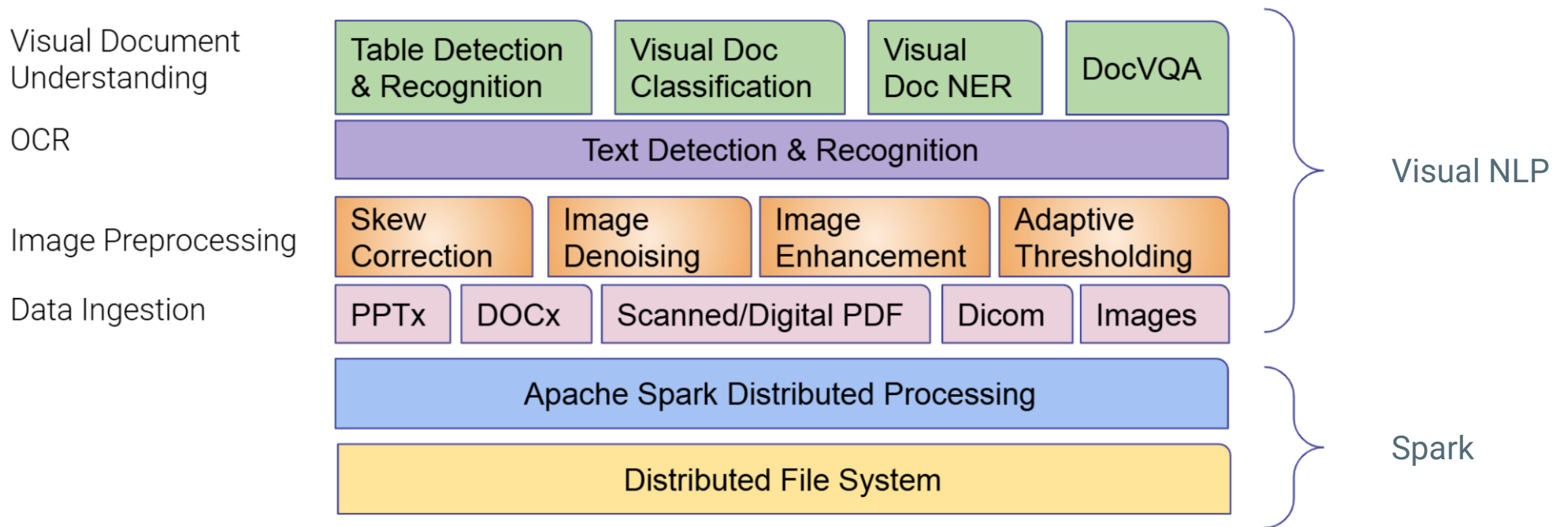 - ○ TapasForQuestionAnswering

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

John Snow LABS

Visual Document Understanding	Table Detection & Recognition	Visual Doc Classification	Visual Doc NER	DocVQA

OCR

Text Detection & Recognition

Image Preprocessing	Skew Correction	Image Denoising	Image Enhancement	Adaptive Thresholding

Data Ingestion	PPTx	DOCx	Scanned/Digital PDF	Dicom	Images

Visual NLP

Apache Spark Distributed Processing

Distributed File System

Spark