John Snow LABS

Applied Generative AI for Data Scientists

April 2025

Pipelines, combining LLM & NLP models

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

Multimodal Pipelines

Check out the ***Visual NLP*** session

Tabular

- Question-Answering on tables
  - TapasForQuestionAnswering

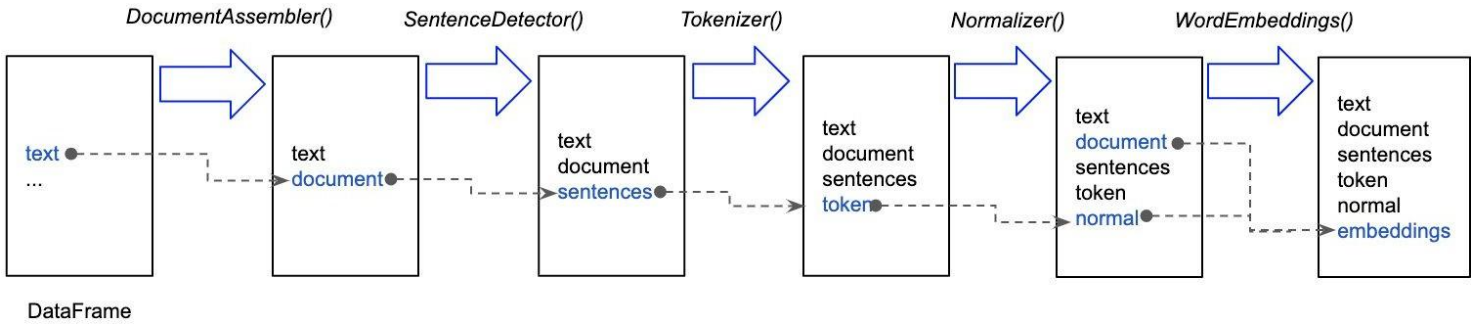
Audio

- 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

Spark NLP Pipelines




```
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
- Text Classification

Encoder only*	ALBERT, BERT, CamemBERT, DeBERTa, DistilBERT, Longformer, MPNet, RoBERTa , XlmRoBERTa
Encoder-Decoder	BART, Flan T5, MarianMT, M2M100, NLLB , T5, XLNet
Decoder only	GPT 2, Llama 2/3, Mistral, Phi 2/3, Qwen-2

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

Other LLM Features

Utilities and integration to third-party software

Through library ***johnsnowlabs****

LLM
- OpenAI Chat Completions
- OpenAI Embeddings
- Auto GGUF with Llama.cpp
- Document Ranking

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

[Notebook with OpenAI integration]

[Notebook for Auto GGUF]

[Notebook for Document Similarity Ranker]

RAG
- Document Splitting
- LangChain
- Haystack

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

[Notebook with Langchain integration]

[Notebook with Haystack integration]

*[johnsnowlabs documentation]

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

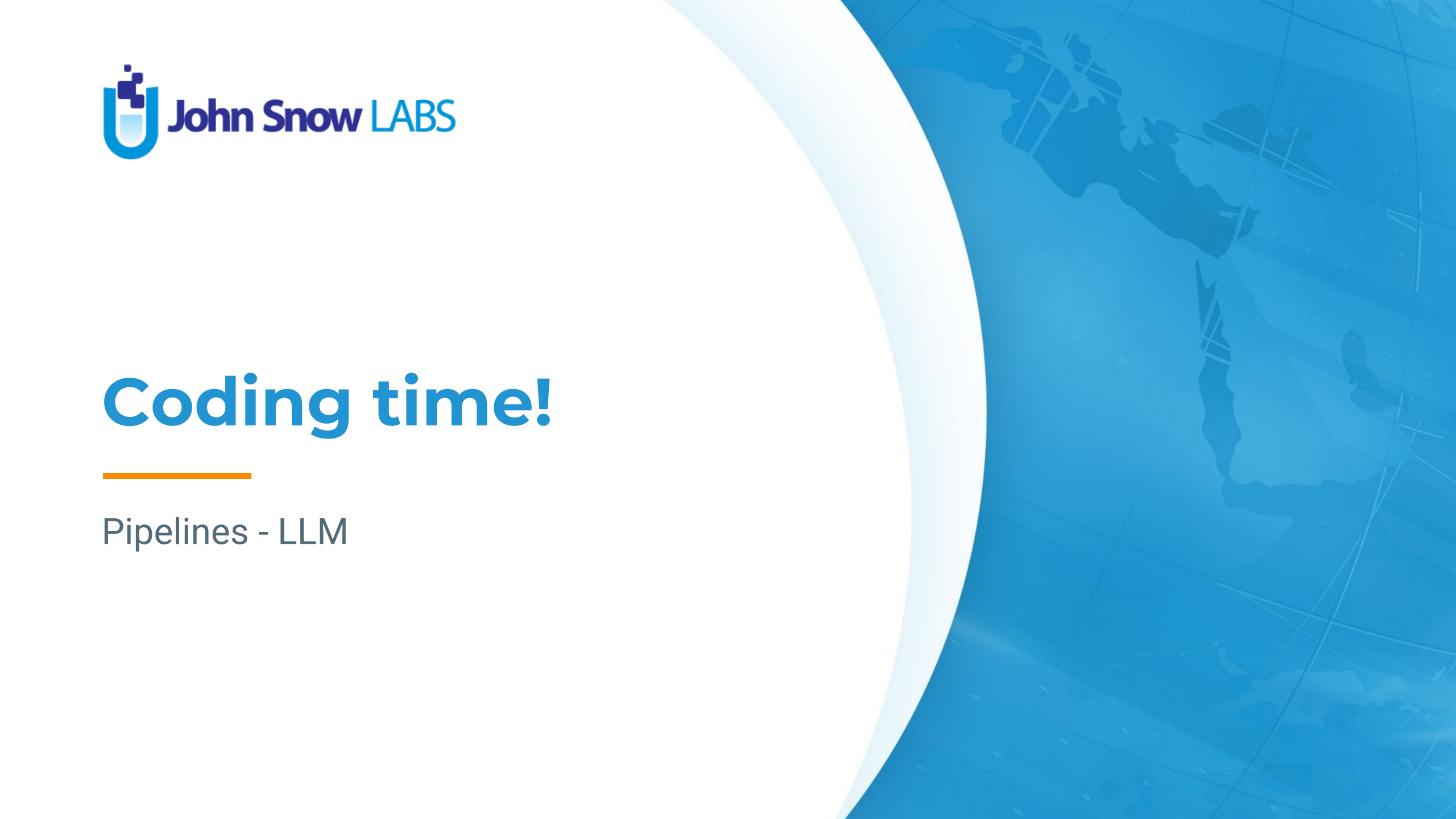
- BGEEmbeddings
- E5Embeddings
- ElmoEmbeddings
- InstructorEmbeddings
- UAEEmbeddings
- UniversalSentenceEncoder

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**



Coding time!

Pipelines - LLM