Spark for NLP

By Avinash Ramesh

What is NLP?

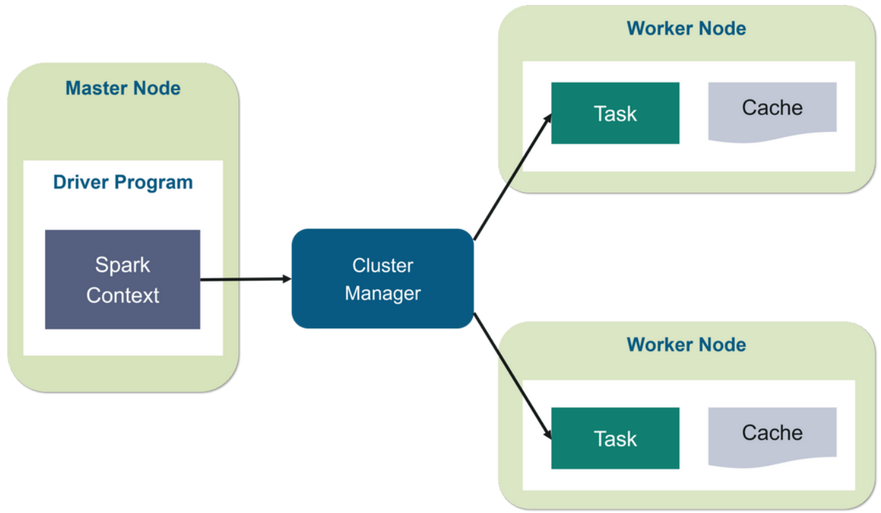
Natural language processing strives to build machines that understand and respond to text or voice data—and respond with text or speech of their own—in much the same way humans do.

Usecases for NLP

Sentiment Analysis
Information Retrieval Systems
Paraphrasing or Summarising
QA Systems/Chatbots
Spam Filters
cognitive assistant
Predictions etc.,

What is Spark

Apache Spark is a lightning-fast unified analytics engine for big data and machine learning. It was originally developed at UC Berkeley in 2009



Spark for NLP

- Spark NLP is an Apache Spark ML-based Natural Language Processing (NLP) library.
- It provides easy-to-scale, performant, and accurate NLP annotations for machine learning pipelines in a distributed environment.
- Spark NLP includes over **1100** pre-trained pipelines and models in over 192 languages.
- It supports nearly all of the NLP tasks and modules that can be used in a cluster seamlessly.

John Snow Labs

https://nlp.johnsnowlabs.com

Why Spark NLP?

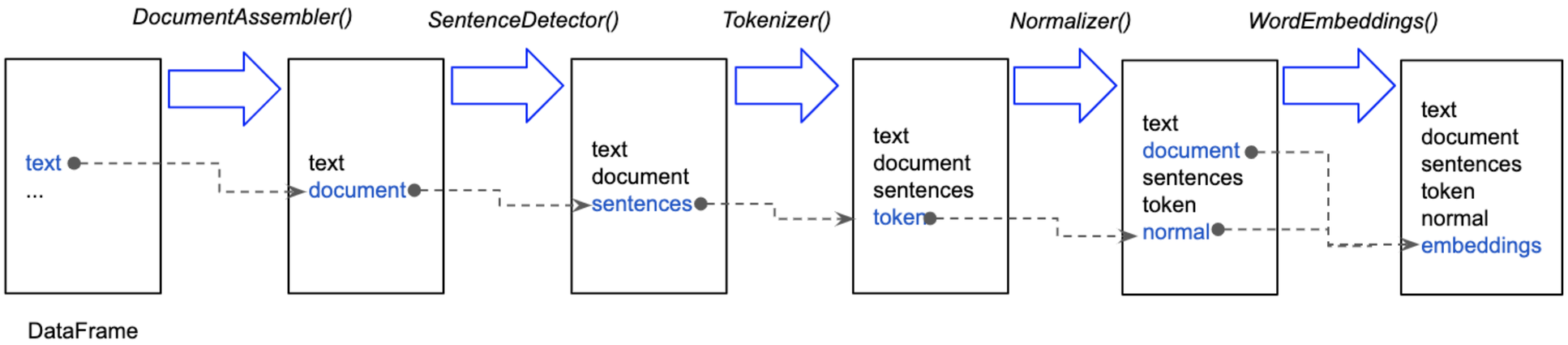
Spark NLP is the only open-source NLP library in **production** that offers state-of-the-art transformers such as BERT, ALBERT, ELECTRA, XLNet, DistilBERT, RoBERTa, XLM-RoBERTa, Longformer, ELMO, Universal Sentence Encoder, Google T5, and MarianMT not only to Python and R, but also to JVM ecosystem (Java, Scala, and Kotlin) at scale by extending Apache Spark natively.

Spark NLP Features

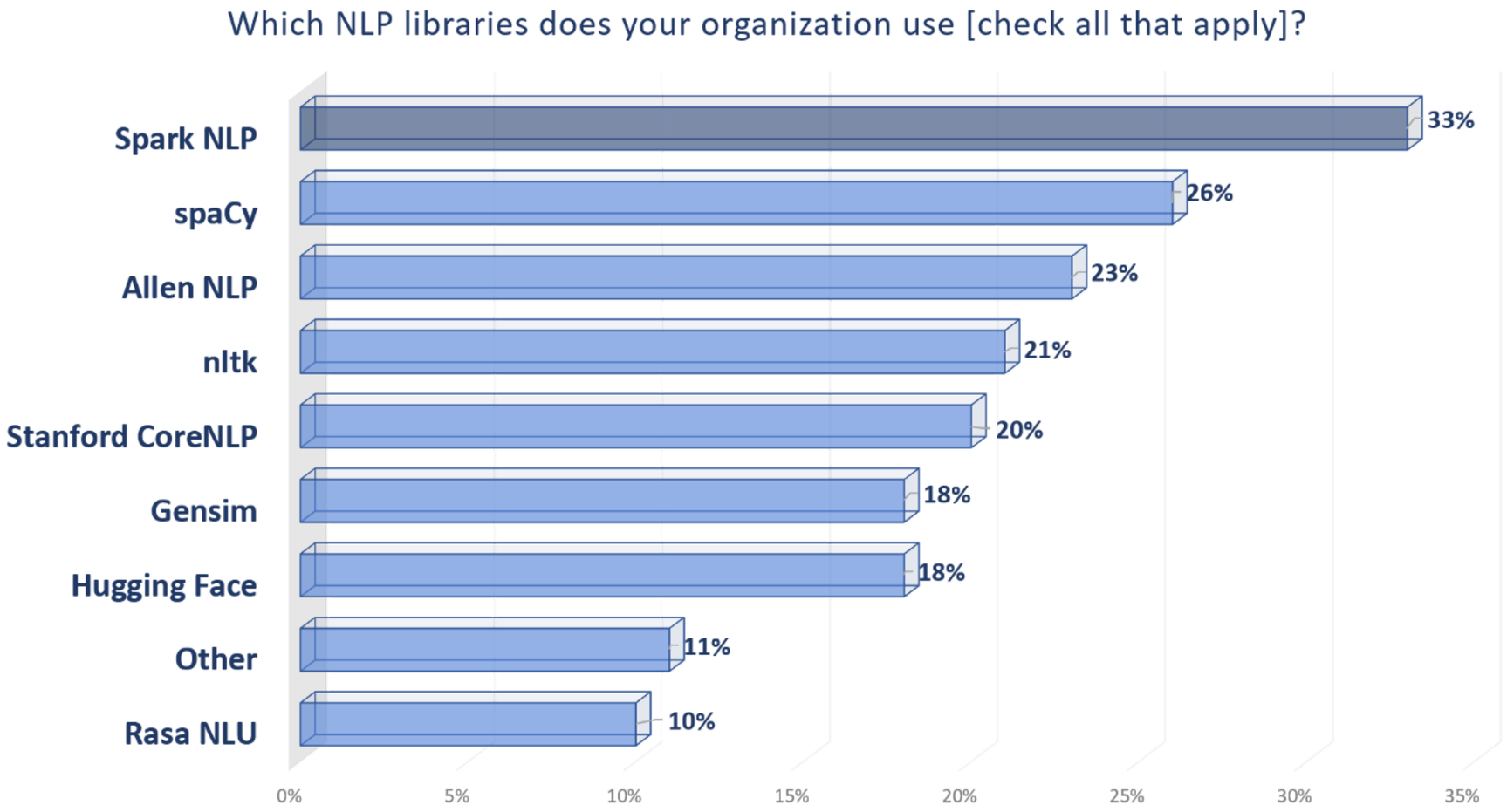
NLP Features

- Tokenization
- Word Segmentation
- Stop Words Removal
- Normalizer
- Stemmer
- Lemmatizer
- NGrams
- Regex Matching
- Text Matching
- Chunking
- Date Matcher
- **Part-of-speech** tagging
- Sentence Detector (DL models)
- **Dependency** parsing
- **Sentiment** Detection (ML models)
- **Spell** Checker (ML & DL models)
- Word Embeddings (**GloVe** & **Word2Vec**)
- Doc2Vec Embeddings (**Word2Vec**)
- **BERT** Embeddings
- **DistilBERT** Embeddings
- **RoBERTa** Embeddings
- **XLM-RoBERTa** Embeddings
- **Longformer** Embeddings
- **ALBERT** Embeddings
- **XLNet** Embeddings
- **ELMO** Embeddings
- **Universal Sentence** Encoder
- **Sentence** Embeddings
- **Chunk** Embeddings
- Neural **Machine Translation** (MarianMT)
- **Text-To-Text** Transfer Transformer **(Google T5)**
- Unsupervised **keywords extraction**
- Language **Detection** & **Identification** (up to 375 languages)
- Multi-class Text **Classification** (DL model)
- Multi-label Text **Classification** (DL model)
- Multi-class **Sentiment Analysis** (DL model)
- BERT for **Sequence Classification**
- DistilBERT for **Sequence Classification**
- BERT for **Token Classification**
- DistilBERT for **Token Classification**
- ALBERT for **Token Classification**
- RoBERTa for **Token Classification**
- XLM-RoBERTa for **Token Classification**
- XLNet for **Token Classification**
- Longformer for **Token Classification**
- **Named entity** recognition (DL model)
- Easy **TensorFlow** integration
- **GPU** Support
- Full integration with **Spark ML** functions
- **2000+** pre-trained **models** in **200+ languages!**
- **1700+** pre-trained **pipelines** in **200+ languages!**

Spark NLP architecture



Spark NLP is currently the most widely used NLP library in the enterprise



What's Next?

Hugging Face is the currently leading NLP startup with more than a thousand companies using their library(mostly around the Transformers library) in production.

Spark NLP ALLOWS IMPORTING HUGGING FACE MODELS

Thank you!