DATA+AI SUMMIT 2022

State-of-the-art Natural Language Processing with **Spark NLP 4**

David Talby
CTO, John Snow Labs

ORGANIZED BY databricks

Agenda

- Why Spark NLP?
- What's New in 4.0?
- 10 Hidden Gems

John Snow LABS

John Snow Labs is the team behind Spark NLP

Most Popular

NLP library in the enterprise, for 4 years

O'Reilly Media

40x Growth

In weekly downloads since 2020

PyPI Download Stats

31% Share

in data science teams with NLP in production

Gradient Flow

John Snow LABS

Spark NLP

State-of-the-art models

Enterprise grade

Natively Scalable

Python, Java, and Scala

Trainable & Tunable

Releases every 2 weeks

Entity Recognition

I love Lucy PERSON

Information Extraction

They met Last week DATE -> 29-04-2020

Spelling & Grammar

She become the first... ->
She became the first

Text Classification

Translation

[je t'aime -> **i love you**]

Summarization

Question Answering

Q&A

Emotion Detection

Split Text

- Sentence Detector
- Tokenizer
- Normalizer
- nGram Generator
- Word Segmentation

Clean Text

- Spell Checker
- Grammer Checker
- Writing Style Checker
- Stopword Cleaner
- Summarization

6,000+

Pre-trained Pipelines, Models & Transformers

BERT | ELMO | GloVe
ALBERT | DeBERTa | USE
Longformer | ELECTRA
T5 | NMT | LaBSE
DistilBERT | RoBERTa
XLM-RoBERTa
S-BERT | XLNet

250+

Languages

Understand Grammar

- Stemmer
- Lemmatizer
- Part of Speech Tagger
- Dependency Parser
- Translation

Find in Text

- Text Matcher
- Regex Matcher
- Date Matcher
- Chunker
- Question Answering

Trainable & Tunable | **Scalable to a Cluster** | **Fast Inference** | **Hardware Optimized** | **Community**

Apache Spark ML Pipelines

LightPipeline

intel NVIDIA

NLP SUMMIT

Spark NLP

- 99 total releases
- Release every two weeks for the past 4 years
- A single unified library for all your NLP/NLU need
- Active community on Slack and GitHub

NLP Feature	Spark NLP	spaCy	NLTK	CoreNLP	Hugging Face
Tokenization	Yes	Yes	Yes	Yes	Yes
Sentence segmentation	Yes	Yes	Yes	Yes	No
Stemming	Yes	Yes	Yes	Yes	No
Lemmatization	Yes	Yes	Yes	Yes	No
POS tagging	Yes	Yes	Yes	Yes	Yes
Entity recognition	Yes	Yes	Yes	Yes	Yes
Dependency parser	Yes	Yes	Yes	Yes	No
Text matcher	Yes	Yes	No	No	No
Date matcher	Yes	No	No	No	No
Sentiment detector	Yes	No	Yes	Yes	Yes
Text classification	Yes	Yes	Yes	No	Yes
Spell checker	Yes	No	No	No	No
Language detector	Yes	No	No	No	No
Keyword extraction	Yes	No	No	No	No
Pretrained models	Yes	Yes	Yes	Yes	Yes
Trainable models	Yes	Yes	Yes	Yes	Yes
Question Answering	Yes	Yes	No	No	Yes
Text Style Transfer	Yes	No	No	No	Yes
Translation	Yes	No	No	No	Yes
Summarization	Yes	No	No	No	Yes
250+ Languages supported	Yes	No	No	No	Yes
Text Generation (GPT, T5)	Yes	No	No	No	Yes

Text Classification with Word & Sentence Embeddings & Transformers in Spark NLP

Glove
ELMO
BERT
Electra
Longformer
Universal Sentence Encoder
Albert
XLNet
RoBerta
DistilBert
XlmRoBerta
Bert Sentence Embeddings
Word2Vec
Bert Token Classifier
DistilBert Token Classifier
Albert Token Classifier
RoBerta Token Classifier
RoBerta Token Classifier
DeBERTa
XlmRoBerta Token Classifier
XLNet Token Classifier
LongFormer Token Classifier
Bert Sequence Classifier
DistilBert Sequence Classifier
DistilBert Sequence Classifier
Albert Sequence Classifier
RoBerta Sequence Classifier
XlmRoBerta Sequence Classifier
XLNet Sequence Classifier
Longformer Sequence Classifier
RoBerta Sentence Embeddings
XlmRoBerta Sentence Embeddings

- Non Transformer Embedding
- Transformer Embedding
- Transformer Sequence Classifier
- Transformer Token Classifier

	Sentence	label
0	a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s	1
1	apparently reassembled from the cutting room floor of any given daytime soap	0
…	…	…
1,999	the movie is undone by a filmmaking methodology that 's just experimental enough	1

DistilBERT

Already (pre-)trained

	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1	-0.172	-0.144	…	0.371	0
…	…	…	…	…	…
1,999	0.124	0.014	…	0.274	1

Spark NLP

- ClassifierDL
- SentimentDL
- MultiClassifierDL
- Sequence Classifier
- Token Classifier

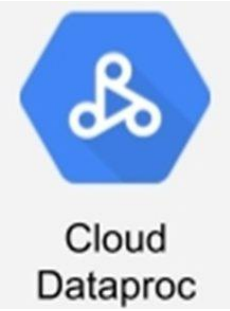
The Production-Grade NLP Models Hub

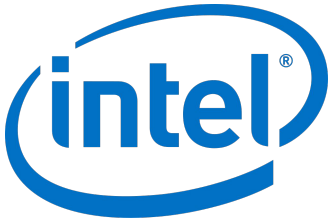
+ Add:	– Remove:	X Unlock:
John Snow Labs Models	Fails Basic Tests	Mix & Match in Pipelines
Hugging Face	Fails Security Scans	Scale to a Cluster
TensorFlow Hub	Test or Play Models	Enterprise Reliability
PyTorch Hub	Undocumented	Run on the JVM
spaCy	Duplicate	Hardware Optimizations
FLAIR	Stale	Free API with NLP Server
FastText	Superseded	No-Code Annotation Lab

https://nlp.johnsnowlabs.com/models

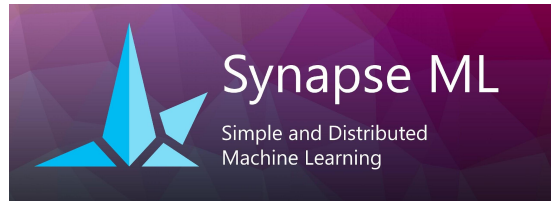
Optimized, Tested, Supported Integrations

kubernetes





databricks



NLU: The Simplicity of Python, the Power of Spark NLP

Powerful One-Liners

Hundreds of NLP models in tens of languages are at your fingertips with just one line of code

Elegant Python

Directly read and write pandas dataframes for frictionless integration with other libraries and existing ML pipelines

100% Open Source

Including pre-trained models & pipelines

```
nlu.load("ner").predict("Angela Market is from Germany")

+---------------+---------+
| entity        |  tag    |
+---------------+---------+
| Angela Merkel | PER     |
| Germany       | LOC     |


nlu.load('en.classify.fakenews').predict('Unicorns sighted on Mars!')

+-----------+-------------+
| fake      | confidence  |
+-----------+-------------+
| FAKE      | 1.000000    |
```

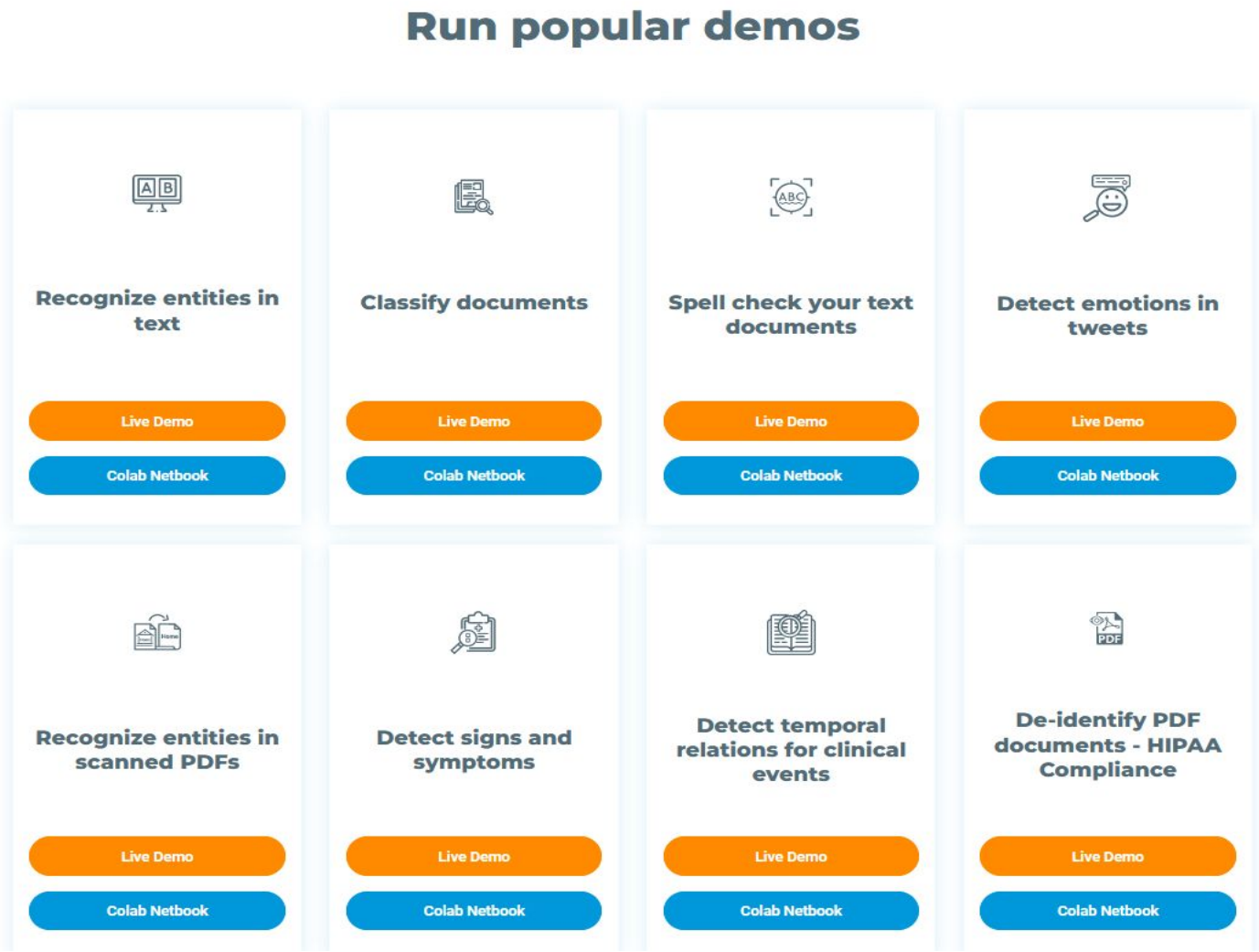
The Demos Hubs: 300+ Notebooks

Spark NLP: English FREE

- Infer Meaning & Intent
- Classify Documents
- Recognize Entities
- Detect Sentiment & Emotion
- Analyze Spelling & Grammar

Spark NLP: World Languages FREE

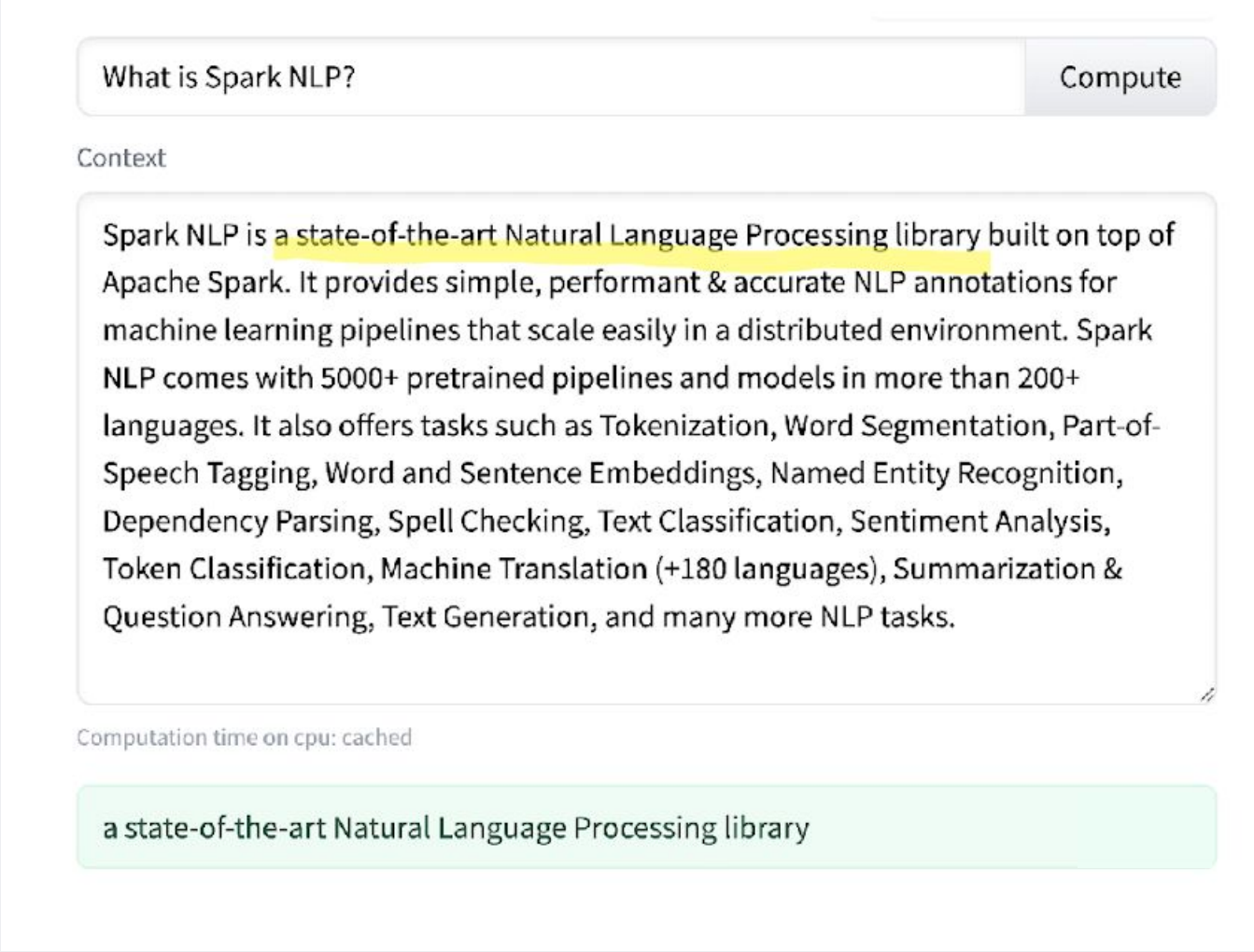
- Identify & Translate Languages
- European Languages
- East Asian Languages
- Languages of India
- Middle Eastern Languages
- African Languages



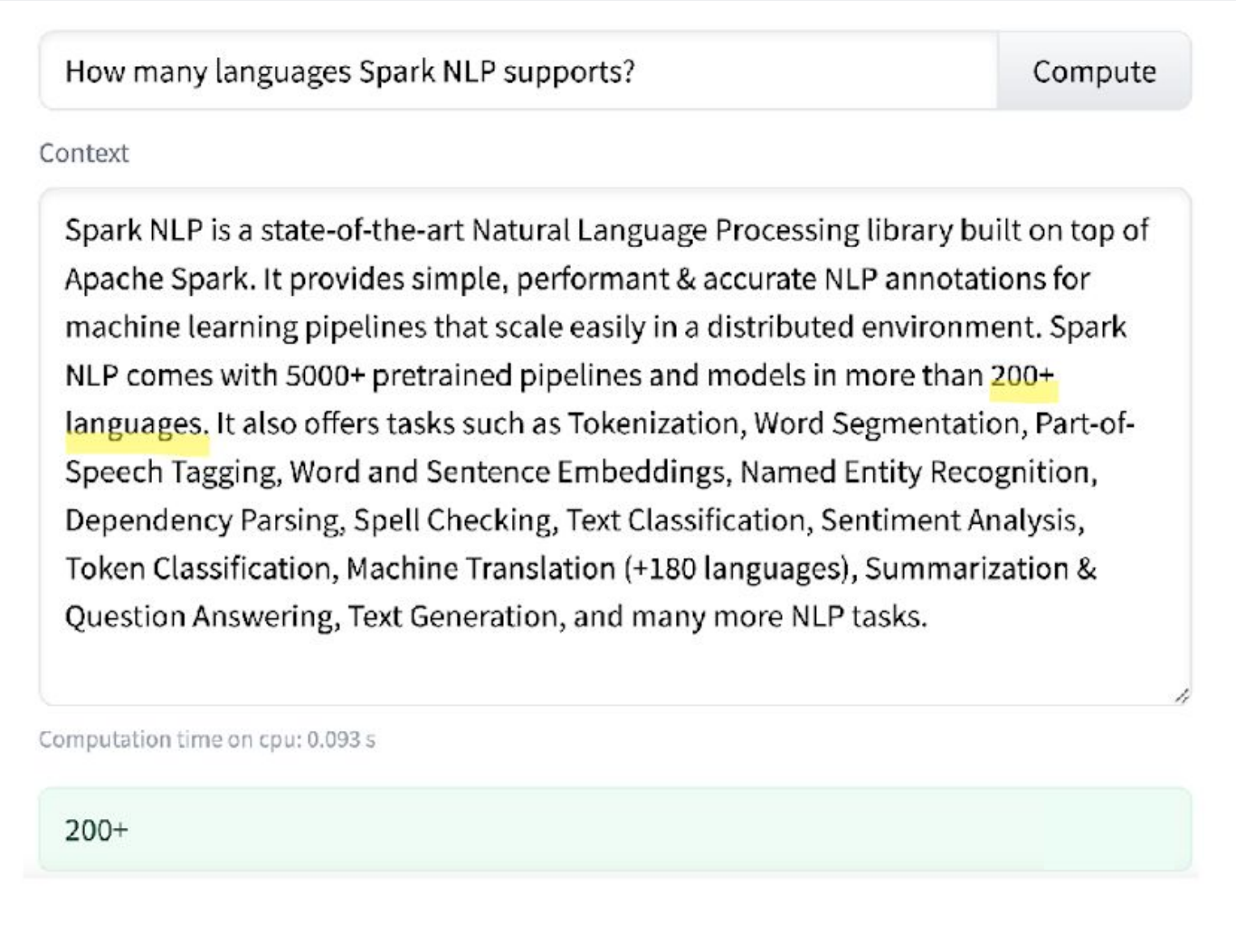
https://nlp.johnsnowlabs.com/demos

What’s new in Spark NLP 4.0?

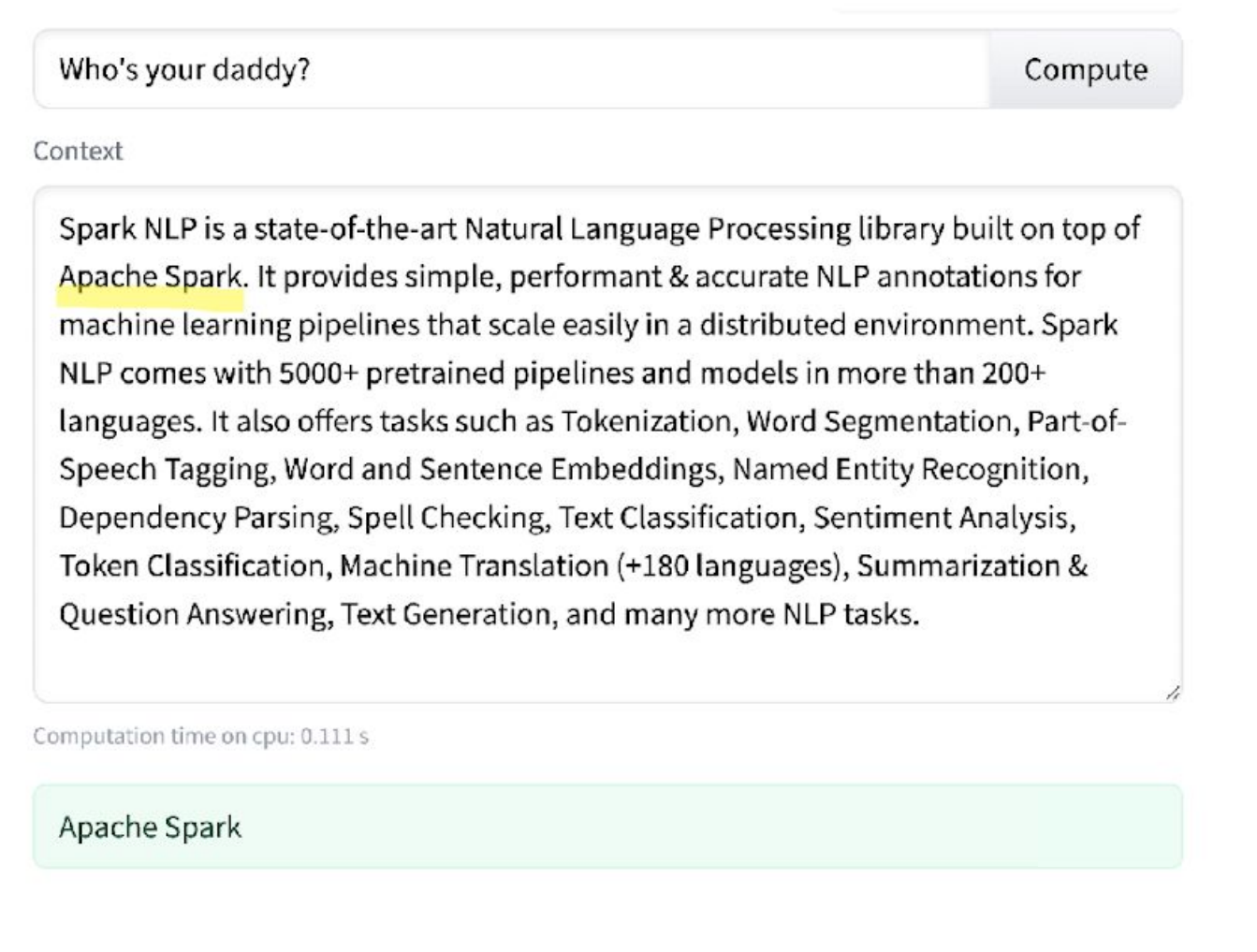
Extractive transformer-based Question answering



Extractive transformer-based Question answering

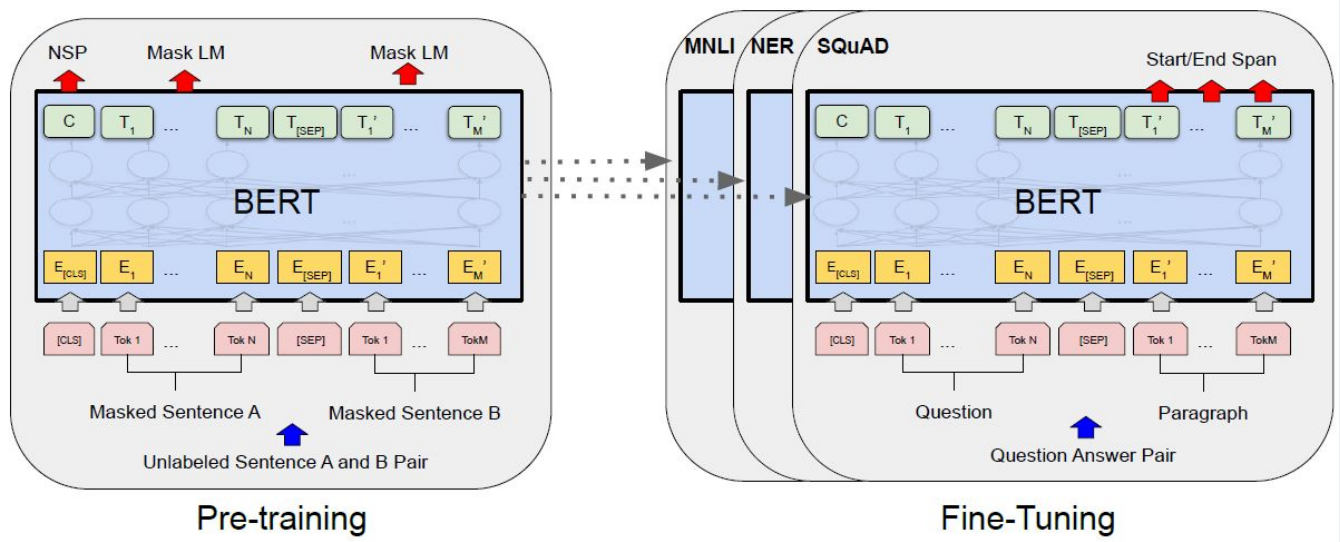


Extractive transformer-based Question answering



1,000+ Models for Modern extractive transformer-based Question answering (QA)

- BERT
- ELECTRA
- DistilBERT
- RoBERTa
- DeBERTa
- XLM-RoBERTa
- Longformer



More Accurate: NER

Highest F1 for an English Named Entity Recognition Model based on CoNLL03

- Spark NLP v3 achieved F1 95% on CoNLL03 Dev dataset by using BERT Large
- Spark NLP v4 achieves F1 96% on CoNLL03 Dev dataset by using DeBERTa Large

SYSTEM	YEAR	LANGUAGE	CONLL '03
Spark NLP v4	2022	Python/Scala/Java/R	**93 (test F1)** **96 (dev F1)**
Spark NLP v3	2021	Python/Scala/Java/R	**93 (test F1)** **95 (dev F1)**
spaCy v3	2021	Python	91.6
Stanza (StanfordNLP)	2020	Python	92.1
Flair	2018	Python	93.1
CoreNLP	2015	Java	89.6

SYSTEM	YEAR	LANGUAGE	ONTONOTES
Spark NLP v3	2021	Python/Scala/Java/R	**90.0 (test F1)** **92.5 (dev F1)**
spaCy RoBERTa	2020	Python	89.8 (dev F1)
Stanza (StanfordNLP)	2020	Python	88.8 (dev F1)
Flair	2018	Python	89.7

More Accurate: Coreference Resolution

John told **Sally** that **she** should come watch **him** play the **violin**.

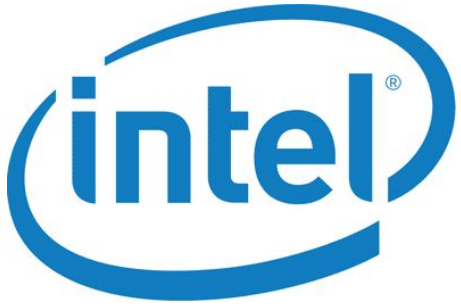
John told **Sally** that **she** should come watch **him** play the **violin**.

BERT for Coreference Resolution: Baselines and Analysis

Mandar Joshi, Omer Levy, Daniel S. Weld, Luke Zettlemoyer

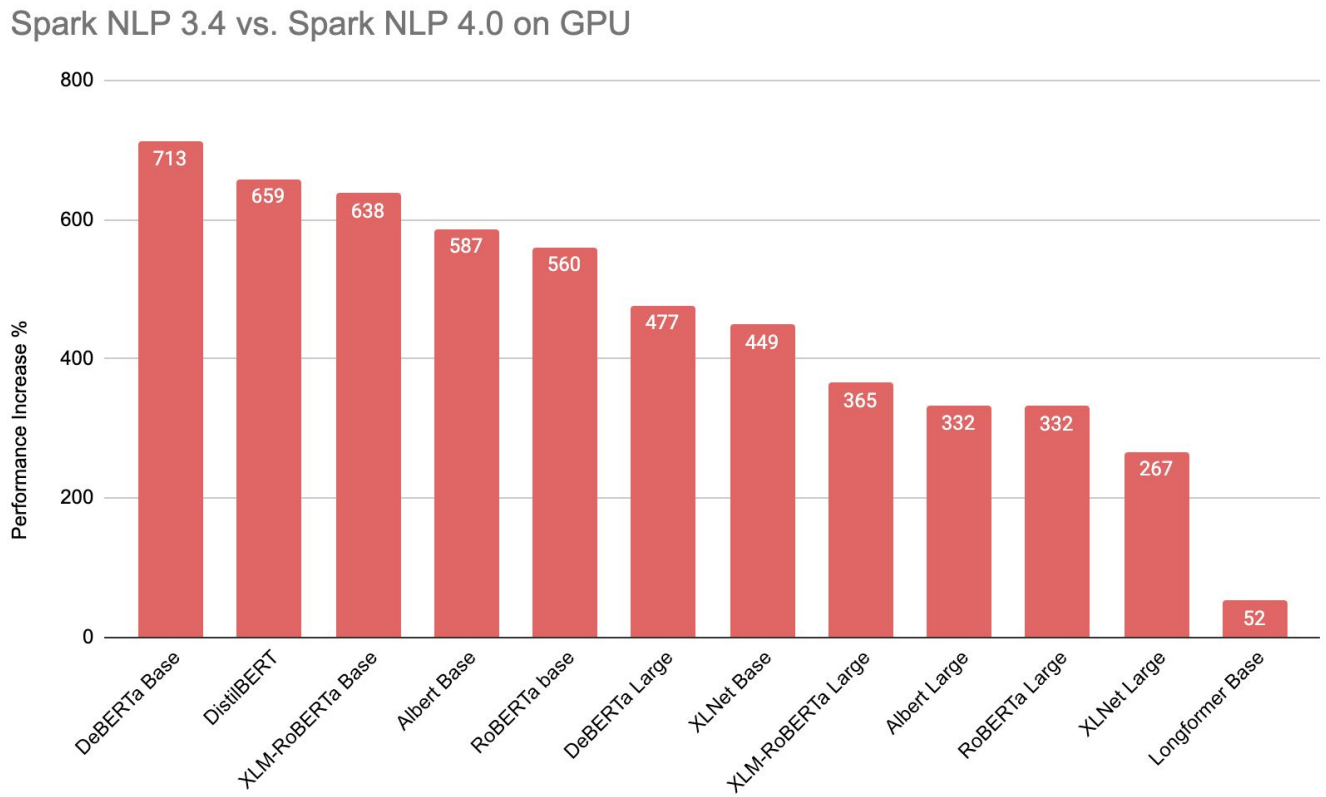
We apply BERT to coreference resolution, achieving strong improvements on the OntoNotes (+3.9 F1) and GAP (+11.5 F1) benchmarks. A qualitative analysis of model predictions indicates that, compared to ELMo and BERT-base, BERT-large is particularly better at distinguishing between related but distinct entities (e.g., President and CEO). However, there is still room for improvement in modeling document-level context, conversations, and mention paraphrasing. Our code and models are publicly available.

Faster! Optimized for the Latest Hardware



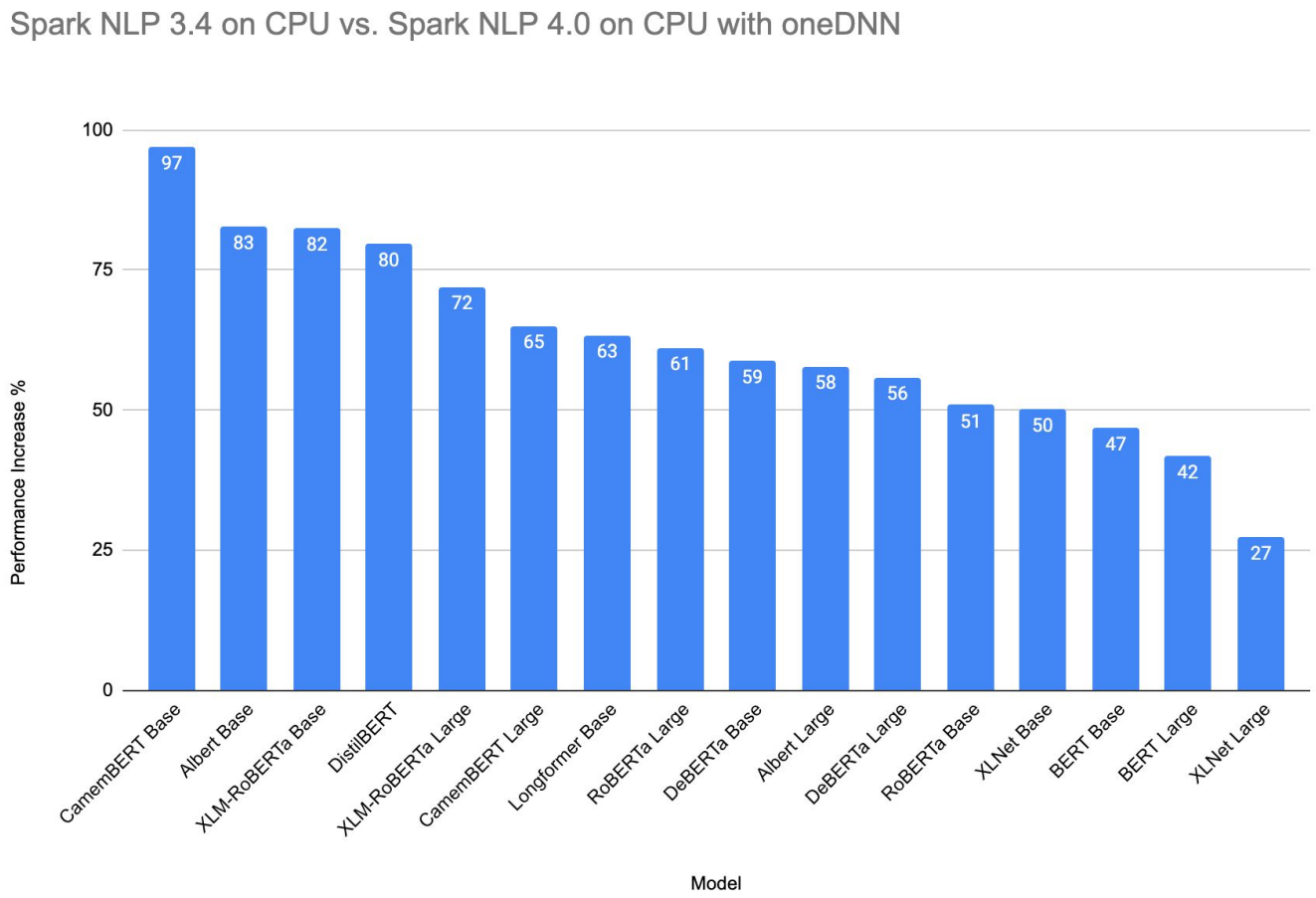
Performance Improvements on GPU Devices

Up to **8X performance** improvements by optimizing batch processing of rows on **GPU** devices

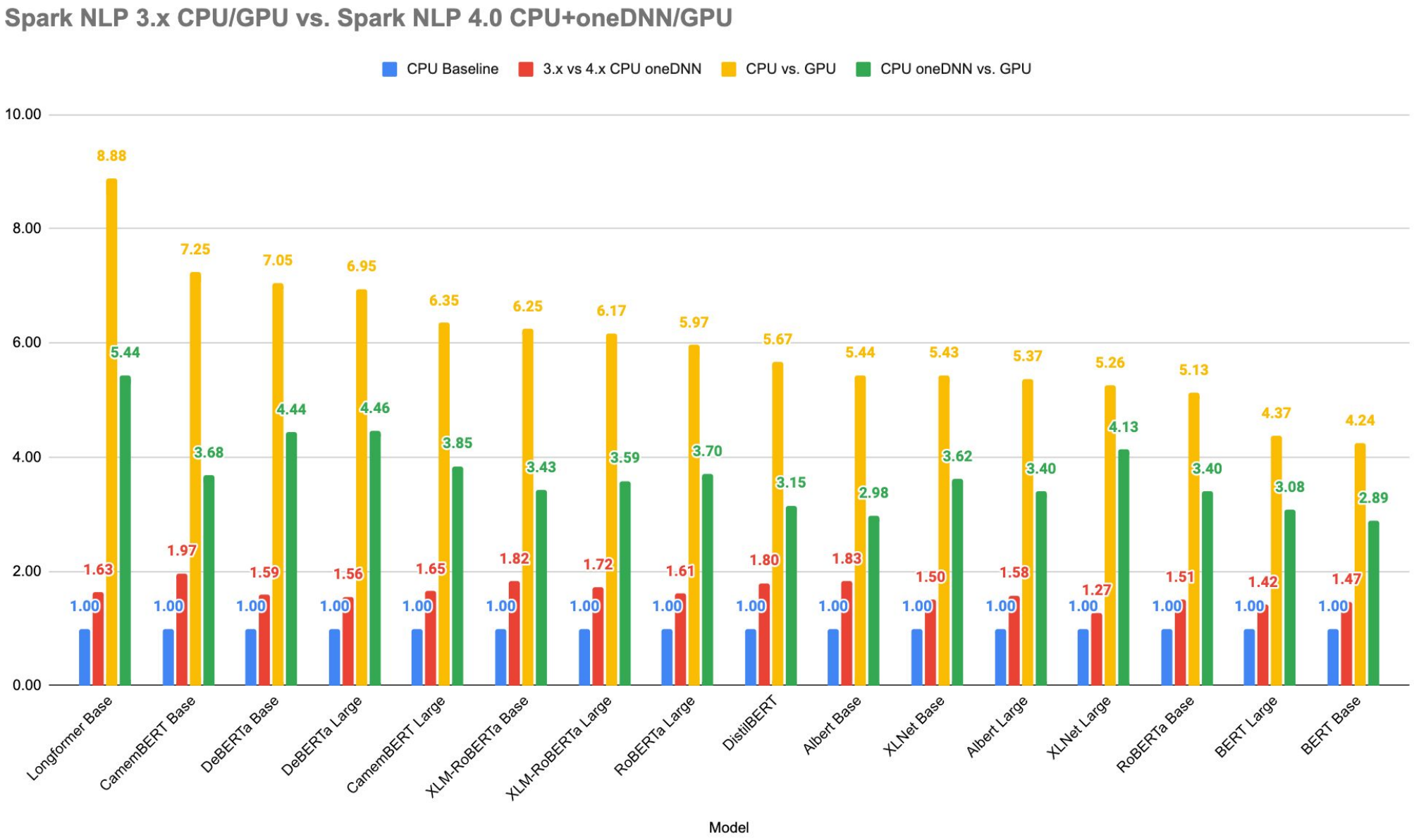


Performance Improvements on CPU Devices

Up to **1.97X** performance improvements on **Intel CPU** by using **oneAPI** Deep Neural Network Library (**oneDNN**)



CPU vs. CPU+oneDNN
GPU vs. CPU+oneDNN
CPU vs. GPU



Newly Supported Platforms

databricks

10.4 +LTS +ML +GPU
10.5 +LTS +ML +GPU

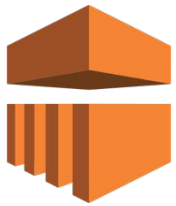
Spark 3.2.x
PySpark 3.2.x

TensorFlow

TensorFlow 2.7.1

kubernetes

Clusters without
HDFS/DBFS/S3 Storage



amazon
EMR

EMR 6.6.0

Scala 2.12.15

10 Hidden Gems

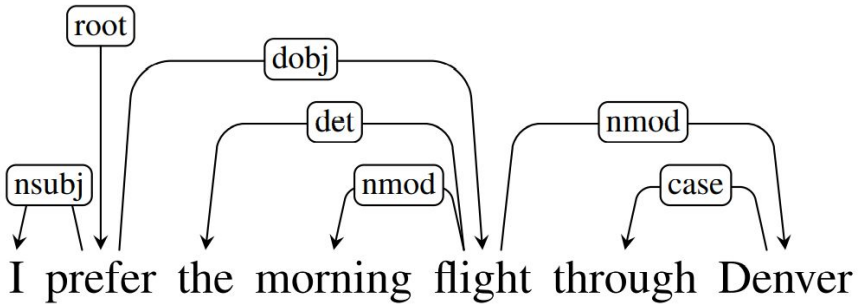
1. Edit Sentences for Style

```
nlu.load("en.t5.informal_to_formal_styletransfer")
   .predict("Who gives a crap?")
>> Who cares?
```

Spark NLP also comes with a Slang normalizer:

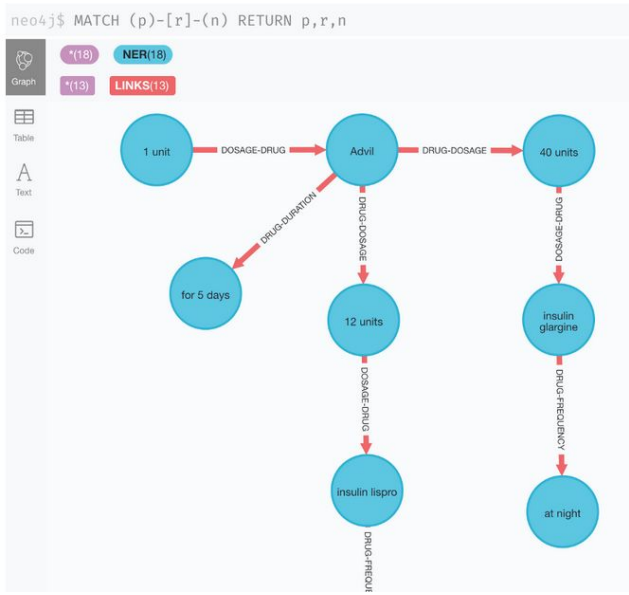
2. Build Knowledge Graphs

Automatically find relationships in free text, using dependency parsing or semantic relation extraction:



Then, use the GraphExtraction annotator to transform them into triplets ready to be loaded into a graph database:

```
pipeline = Pipeline().setStages([
    documentAssembler,
    sentence,
    tokenizer,
    embeddings,
    nerTagger,
    posTagger,
    dependencyParser,
    typedDependencyParser,
    graph_extraction
])
```



3. Train in one language, predict in 100+ languages

100+ Languages supported by
Language-agnostic BERT Sentence Embedding (LABSE) and **XLM-RoBERTa**

```
# Binary Class Classifier, 2 classes
nlu.load('xx.embed_sentence.labse train.sentiment').fit(train_df).predict(test_df)

# Multi Class Classifier, N classes
nlu.load('xx.embed_sentence.labse train.classifier').fit(train_df).predict(test_df)

# Multi Class Classifier with multiple labels example (i.e. Hashtags)
# N classes, where one row can be assigned up to N labels
nlu.load('xx.embed_sentence.labse train.multi_classifier').fit(train_df).predict(test_df)
```

ISO	NAME	ISO	NAME	ISO	NAME
af	AFRIKAANS	ht	HAITIAN_CREOLE	pt	PORTUGUESE
am	AMHARIC	hu	HUNGARIAN	ro	ROMANIAN
ar	ARABIC	hy	ARMENIAN	ru	RUSSIAN
as	ASSAMESE	id	INDONESIAN	rw	KINYARWANDA
az	AZERBAIJANI	ig	IGBO	si	SINHALESE
be	BELARUSIAN	is	ICELANDIC	sk	SLOVAK
bg	BULGARIAN	it	ITALIAN	sl	SLOVENIAN
bn	BENGALI	ja	Japanese	sm	SAMOAN
bo	TIBETAN	jv	JAVANESE	sn	SHONA
bs	BOSNIAN	ka	GEORGIAN	so	SOMALI
ca	CATALAN	kk	KAZAKH	sq	ALBANIAN
ceb	CEBUANO	km	KHMER	sr	SERBIAN
co	CORSICAN	kn	KANNADA	st	SESOTHO
cs	CZECH	ko	KOREAN	su	SUNDANESE
cy	WELSH	ku	KURDISH	sv	SWEDISH
da	DANISH	ky	KYRGYZ	sw	SWAHILI
de	GERMAN	la	LATIN	ta	TAMIL
el	GREEK	lb	LUXEMBOURGISH	te	TELUGU
en	ENGLISH	lo	LAOTHIAN	tg	TAJIK
eo	ESPERANTO	lt	LITHUANIAN	th	THAI
es	SPANISH	lv	LATVIAN	tk	TURKMEN
et	ESTONIAN	mg	MALAGASY	tl	TAGALOG
eu	BASQUE	mi	MAORI	tr	TURKISH
fa	PERSIAN	mk	MACEDONIAN	tt	TATAR
fi	FINNISH	ml	MALAYALAM	ug	UIGHUR
fr	FRENCH	mn	MONGOLIAN	uk	UKRAINIAN
fy	FRISIAN	mr	MARATHI	ur	URDU
ga	IRISH	ms	MALAY	uz	UZBEK
gd	SCOTS_GAELIC	mt	MALTESE	vi	VIETNAMESE
gl	GALICIAN	my	BURMESE	wo	WOLOF
gu	GUJARATI	ne	NEPALI	xh	XHOSA
ha	HAUSA	nl	DUTCH	yi	YIDDISH
haw	HAWAIIAN	no	NORWEGIAN	yo	YORUBA
he	HEBREW	ny	NYANJA	zh	Chinese
hi	HINDI	or	ORIYA	zu	ZULU
hmn	HMONG	pa	PUNJABI		
hr	CROATIAN	pl	POLISH		

4. Translate between 200+ languages


```
# Use ISO standards for the languages
nlu.load('<start_language>.translate_to.<target_language>')

#Translate Turkish to English:
nlu.load('tr.translate_to.en')

#Translate English to French:
nlu.load('en.translate_to.fr')

#Translate French to Hebrew
nlu.load('fr.translate_to.he')`

#Translate English to German
nlu.load('en.translate_to.de')`
```

5. Import and Scale models from HF and TFHub

TF Hub to Spark NLP

Spark NLP	TF Hub Notebooks	Colab
BertEmbeddings	TF Hub in Spark NLP - BERT	Open in Colab
BertSentenceEmbeddings	TF Hub in Spark NLP - BERT Sentence	Open in Colab
AlbertEmbeddings	TF Hub in Spark NLP - ALBERT	Open in Colab

HuggingFace to Spark NLP

Spark NLP	HuggingFace Notebooks	Colab
BertEmbeddings	HuggingFace in Spark NLP - BERT	Open in Colab
BertSentenceEmbeddings	HuggingFace in Spark NLP - BERT Sentence	Open in Colab
DistilBertEmbeddings	HuggingFace in Spark NLP - DistilBERT	Open in Colab
RoBertaEmbeddings	HuggingFace in Spark NLP - RoBERTa	Open in Colab
XlmRoBertaEmbeddings	HuggingFace in Spark NLP - XLM-RoBERTa	Open in Colab
AlbertEmbeddings	HuggingFace in Spark NLP - ALBERT	Open in Colab
XlnetEmbeddings	HuggingFace in Spark NLP - XLNet	Open in Colab
LongformerEmbeddings	HuggingFace in Spark NLP - Longformer	Open in Colab
BertForTokenClassification	HuggingFace in Spark NLP - BertForTokenClassification	Open in Colab
DistilBertForTokenClassification	HuggingFace in Spark NLP - DistilBertForTokenClassification	Open in Colab
AlbertForTokenClassification	HuggingFace in Spark NLP - AlbertForTokenClassification	Open in Colab
RoBertaForTokenClassification	HuggingFace in Spark NLP - RoBertaForTokenClassification	Open in Colab
XlmRoBertaForTokenClassification	HuggingFace in Spark NLP - XlmRoBertaForTokenClassification	Open in Colab
BertForSequenceClassification	HuggingFace in Spark NLP - BertForSequenceClassification	Open in Colab
DistilBertForSequenceClassification	HuggingFace in Spark NLP - DistilBertForSequenceClassification	Open in Colab
AlbertForSequenceClassification	HuggingFace in Spark NLP - AlbertForSequenceClassification	Open in Colab
RoBertaForSequenceClassification	HuggingFace in Spark NLP - RoBertaForSequenceClassification	Open in Colab
XlmRoBertaForSequenceClassification	HuggingFace in Spark NLP - XlmRoBertaForSequenceClassification	Open in Colab
XlnetForSequenceClassification	HuggingFace in Spark NLP - XlnetForSequenceClassification	Open in Colab

https://nlp.johnsnowlabs.com/docs/en/transformers#example-notebooks

6. Unsupervised Keyword Extraction

YAKE Is Yet Another Keyword Extraction Algorithm that can extract keywords without any by leveraging statistical properties of ngrams

```
nlu.load('yake').predict("NLU is a Python Library for beginners and experts in NLP")
```

keywords_score_confidence	keywords	sentence
0.454232	[nlu, nlp, python library]	NLU is a Python Library for beginners and expe...

7. Check Spelling Based on Context

```
# check for the different occurrences of the word "siter"
example1 = ["I will call my siter.",\
    "Due to bad weather, we had to move to a different siter.",\
    "We travelled to three siter in the summer."]
beautify(lp.annotate(example1))
```

```
['I  will  call  my  sister  .\n',
 'Due  to  bad  weather  ,  we  had  to  move  to  a  different  site  .\n',
 'We  travelled  to  three  sites  in  the  summer  .\n']
```

Word error rate of 8.09% versus 20.24% with the JamSpell library

8. Deep-Learning-Based Sentence Splitter

How do you correctly split text to sentences when there is no punctuation, bad punctuation, or unexpected line breaks? (common in OCR and ASR outputs)

```
text = """John loves Mary.Mary loves Peter
Peter loves Helen .Helen loves John;
Total: four people involved."""

for anno in sd_model.fullAnnotate(text)[0]["sentences"]:
    print("{}\t{}\t{}\t{}".format(
        anno.metadata["sentence"], anno.begin, anno.end, anno.result))
```

```
0       0       15      John loves Mary.
1       16      32      Mary loves Peter
2       33      51      Peter loves Helen .
3       52      68      Helen loves John;
4       71      98      Total: four people involved.
```

Trainable. Multilingual. Compared accuracy with spacy's nlp.sents()

9. Identify 10+ Kinds of Toxic Content

Identify Fake news

Live Demo

Colab Netbook

Detect Spam messages

Live Demo

Colab Netbook

Detect toxic content in comments

Live Demo

Colab Netbook

Identify Antisemitic Texts

Live Demo

Colab Netbook

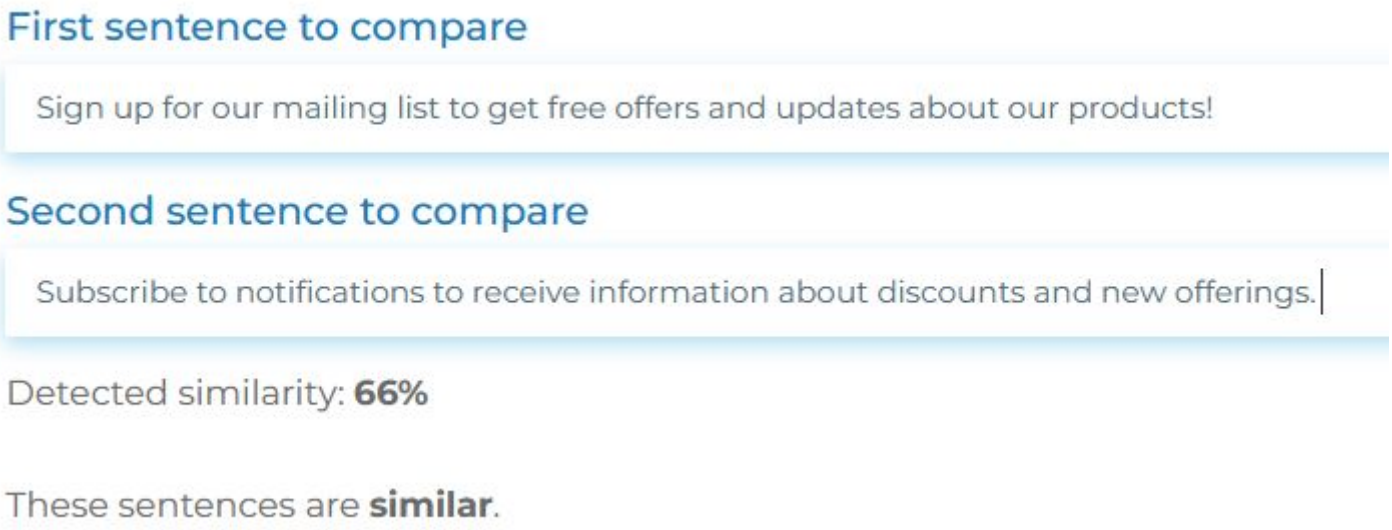
Identify Hate Speech Texts

Live Demo

Colab Netbook

```
nlu.load('en.classify.toxic').predict('You are to stupid')
```

10. Measure Semantic Similarity



Useful for searching in knowledge bases, filtering content, or for unsupervised topic modelling.

DATA+AI
SUMMIT 2022

Thank you!

In/davidtalby

@davidtalby

nlp.johnsnowlabs.com

github.com/johnsnowlabs

John Snow LABS