NLU for Hackers at Python Web Conf 2021

pip install nlu

2021

Introducing Spark NLP

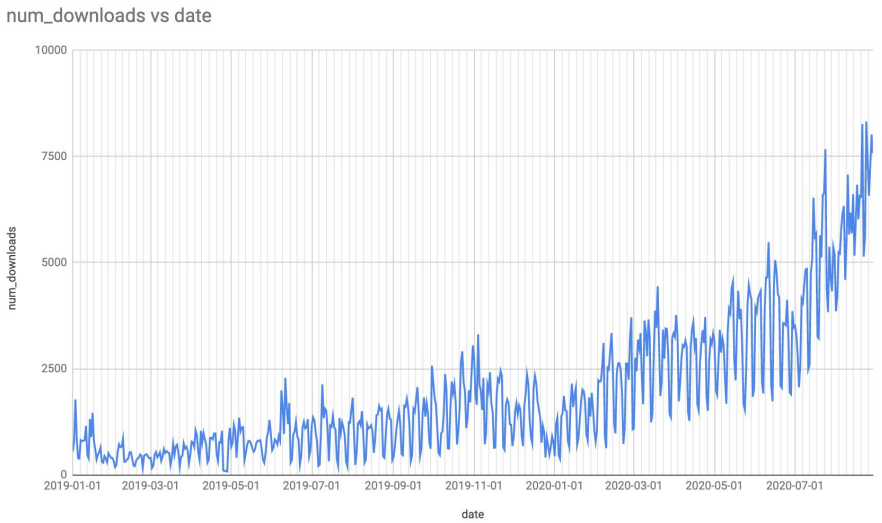
State of the art NLP:

1. **Accuracy**
2. **Speed**
3. **Scalability**

Open-Source Python, Java & Scala Libraries
200+ Pre-Trained Models & Pipelines
Vibrant: 26 new releases in 2018, 28 in 2019

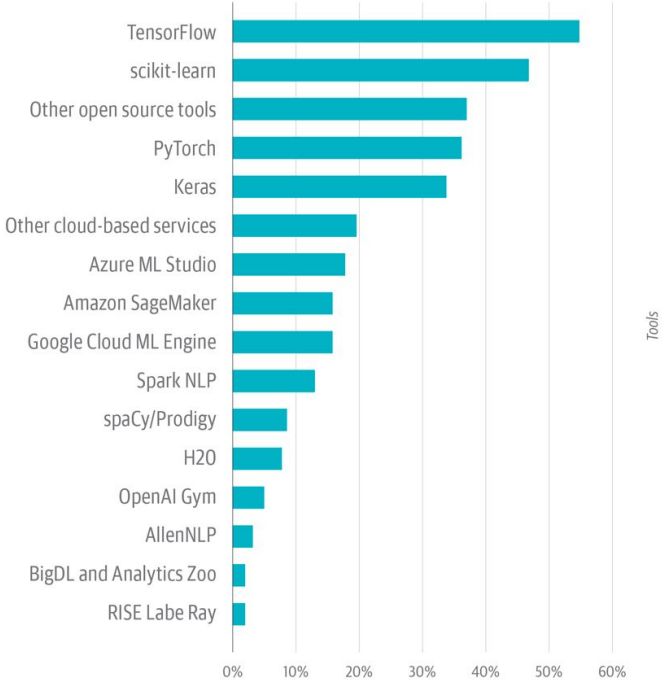
Daily ~ 20K
Monthly ~ 600K

PyPI link	https://pypi.org/project/spark-nlp
Total downloads	3,976,595
Total downloads - 30 days	656,474
Total downloads - 7 days	152,742

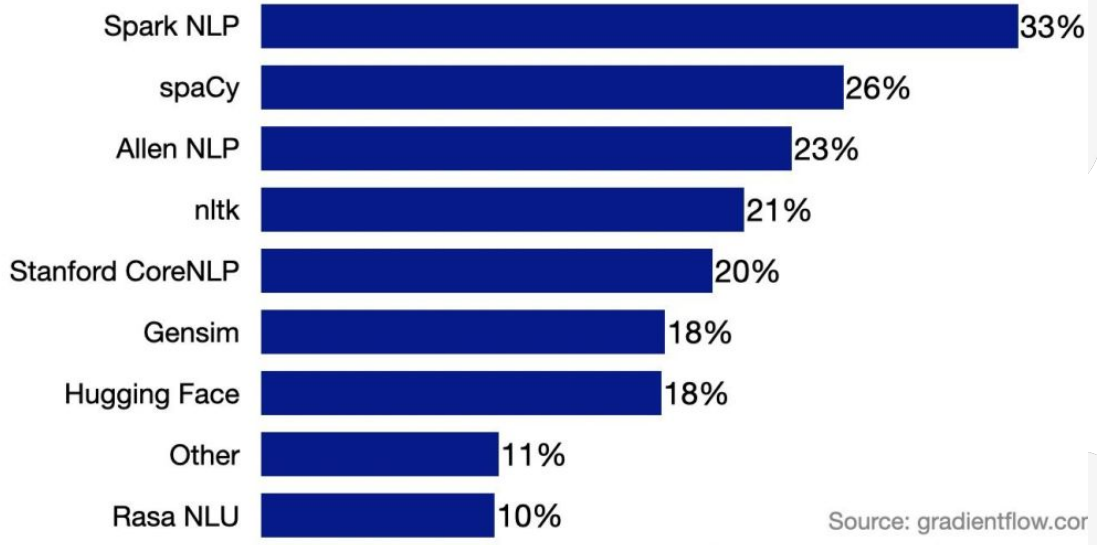


Spark NLP in Industry

Which of the following AI tools do you use?



Which NLP libraries does your organization use?



NLP Industry Survey by Gradient Flow,
an independent data science research & insights company, September 2020

TRUSTED BY


Microsoft


Google


amazon


intel


IBM


databricks


verizon


indeed


CapitalOne


viacom


McKESSON


MERCK


Roche


selectdata


UiPath
Reboot Work.


ASCO
CANCER·LINQ
DISCOVERY


cnrs


Imperial College
London


Georgia
Tech


STANFORD
UNIVERSITY

AI
ARTIFICIAL INTELLIGENCE
EXCELLENCE
AWARD
2020

"John Snow Labs wins our best AI product or service award thanks to **exceptional success turning AI research into real & dependable systems for a global community**."

"An open source project, tool, or contribution that **significantly advances the state of data science** is recognized with this award."

HEALTHCARE ANALYTICS
PROVIDER OF THE YEAR 2020
Healthcare
Tech OUTLOOK

"By all accounts, John Snow Labs has created **the most accurate software in history** to extract facts from unstructured text."

Powerful NLU 1 Liners you learn today applicable in 200+ Languages

1. Spell Checking
2. Binary Sentiment Classification
3. Multi Class Emotion Classification
4. Parts of Speech (POS)
5. Named Entity Recognition (NER)
6. Unsupervised Keyword Extraction (YAKE!)
7. Question Answering
8. Translation between 200+ languages
9. Train a Multilingual Classifier for 100+ languages from just 1 input language

INTRODUCING THE NLU LIBRARY

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

How does it work?

```
model= nlu.load(model)
```

- **Returns a nlu pipeline object**

```
model.predict(data)
```

- Returns a pandas DF

How does it work?

```
model = nlu.load('emotion')
```

- **Returns a nlu pipeline object**

```
model.predict('I love NLU!')
```

- Returns a pandas DF

EMOTION DETECTION

```
nlu.load('emotion').predict('I love NLU!')
```

sentence_embeddings	category_sentence	category_surprise	category_sadness	category_joy	category_fear	sentence	category	id
[0.027570432052016258, -0.052647676318883896, ...]	0	0.012899903	0.0015578865	0.9760173	0.0095249	I love NLU!	joy	1

TOKENIZATION & SPELL CHECKING

```
nlu.load('spell').predict('I liek pentut butr and jelli')
```

token	checked	id
I	I	1
liek	like	1
peantut	peanut	1
buttr	butter	1
and	and	1
jelli	jelly	1

FAKE NEW CLASSIFICATION

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

sentence_embeddings	category_confidence	sentence	category	id
[-0.01756167598068714, 0.015006818808615208, -...]	1.000000	Unicorns have been sighted on Mars!	FAKE	1

NAMED ENTITY RECOGNITION

```
nlu.load('ner').predict('Angela Merkel from Germany and the American Donald Trump dont share many opinions')
```

embeddings	ner_tag	entities
[[-0.563759982585907, 0.26958999037742615, 0.3...	PER	Angela Merkel
[[-0.563759982585907, 0.26958999037742615, 0.3...	LOC	Germany
[[-0.563759982585907, 0.26958999037742615, 0.3...	MISC	American
[[-0.563759982585907, 0.26958999037742615, 0.3...	PER	Donald Trump

CALCULATING EMBEDDINGS

```
#watch out for your RAM, this could kill your machine

nlu.load('bert elmo albert xlnet use glove').predict('Get all of them at once! Watch your RAM tough!')
```

token	glove_embeddings	albert_embeddings	xlnet_embeddings	bert_embeddings	elmo_embeddings	use_embeddings	id
Get	[0.1443299949169159, 0.4395099878311157, 0.583...]	[-0.41224443912506104, -0.4611411392688751, 1...]	[-0.003953204490244389, -1.5821468830108643, -...]	[-0.7420049905776978, -0.8647691011428833, 0.1...]	[0.04002974182367325, -0.43536433577537537, -0...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
all	[-0.2182299941778183, 0.6919900178909302, 0.70...]	[1.1014549732208252, -0.43204769492149353, -0...]	[0.31148090958595276, -1.0986182689666748, 0.3...]	[-0.8933112025260925, 0.44822725653648376, -0...]	[0.17885173857212067, 0.045830272138118744, -0...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
of	[-0.15289999544620514, -0.24278999865055084, 0...]	[1.1535910367965698, 0.28440719842910767, 0.60...]	[-1.403516411781311, 0.3108177185058594, -0.32...]	[-0.5550722479820251, 0.2702311873435974, 0.04...]	[0.24783466756343842, -0.248960942029953, 0.02...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
them	[-0.10130999982357025, 0.10941000282764435, 0...]	[0.5475010871887207, 0.8660883903503418, 2.817...]	[-0.7559828758239746, -0.4712887704372406, -1...]	[-0.2922026813030243, -0.1301671266555786, -0...]	[-0.24157099425792694, -0.8055092692375183, -0...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
at	[0.17659999430179596, 0.0938510000705719, 0.24...]	[-0.5005946159362793, -0.4600788354873657, 0.5...]	[0.04092511534690857, -1.0951932668685913, -1...]	[-0.5613634586334229, -0.00903533399105072, -0...]	[-0.11999595910310745, 0.012994140386581421, -...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
once	[-0.23837999999523163, 0.221670001745224, 0.35...]	[-0.39100387692451477, -0.8297092914581299, 2...]	[-0.46001458168029785, -1.2062749862670898, 0...]	[0.2988640069961548, 0.3360409140586853, -0.37...]	[0.6701997518539429, 1.1368376016616821, 0.244...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
!	[0.38471999764442444, 0.49351000785827637, 0.4...]	[0.007945209741592407, -0.27733859419822693, 0...]	[-1.5816600322723389, -0.992130696773529, -0.1...]	[0.7550013065338135, -0.5257778167724609, -0.4...]	[-1.3351283073425293, 0.6296550035476685, -1.4...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
Watch	[-0.38264000415802, -0.08968199789524078, 0.02...]	[-0.10218311846256256, -0.4334276020526886, 0...]	[-1.3921688795089722, 0.6997514963150024, -0.8...]	[-0.24852752685546875, 1.222611427307129, -0.1...]	[0.04002974182367325, -0.43536433577537537, -0...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
your	[-0.5718399882316589, 0.046348001807928085, 0...]	[-0.4086211323738098, 1.0755341053009033, 1.78...]	[-0.8588163256645203, -2.3702170848846436, 0.0...]	[-0.035358428955078125, 0.7711482048034668, 0...]	[0.17885173857212067, 0.045830272138118744, -0...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
RAM	[-1.875599980354309, -0.40814998745918274, -0...]	[-0.09772858023643494, 0.3632940351963043, -0...]	[1.1277621984481812, -1.689896583557129, -0.19...]	[0.4528151750564575, -0.36768051981925964, -0...]	[0.24783466756343842, -0.248960942029953, 0.02...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
tough	[-0.5099300146102905, -0.1428000032901764, 0.5...]	[-0.22261293232440948, 0.21325691044330597, 0...]	[-1.3547197580337524, 0.43423181772232056, -1...]	[0.460737079381942 75, 0.05694812536239624, 0.5...]	[-0.24157099425792694, -0.8055092692375183, -0...]	[[-0.0019260947592556477, 0.009215019643306732...]	1
!	[0.38471999764442444, 0.49351000785827637, 0.4...]	[0.21658605337142944, -0.0493735149502754 2, 0...]	[-1.5816600322723389, -0.992130696773529, -0.1...]	[0.6830563545227051, -0.5751053094863892, -0.6...]	[-0.11999595910310745, 0.012994140386581421, -...]	[[-0.0019260947592556477, 0.009215019643306732...]	1

NLU WORKS DIRECTLY ON TYPICAL PYTHON DATASETS

Strings

```
import nlu
nlu.load('sentiment').predict('This is just one string')
```

Lists

```
import nlu
nlu.load('sentiment').predict(['This is an arrray', ' Of strings!'])
```

Pandas data frame

```
import nlu
import pandas as pd
data = {"text": ['This day sucks', 'I love this day', 'I dont like Sami']}
text_df = pd.DataFrame(data)
nlu.load('sentiment').predict(text_df)
```

Pandas series

```
import nlu
import pandas as pd
data = {"text": ['This day sucks', 'I love this day', 'I dont like Sami']}
text_df = pd.DataFrame(data)
nlu.load('sentiment').predict(text_df['text'])
```

Spark
Data Frame

Ray
Data Frame

Dask
Data Frame

NLU

- A single **unified** library for all your **NLP/NLU** needs
- 1000+ Models,
- 200+ Languages
- 1 Line of code
- Active **community** on **Slack** and **GitHub**

NLP Feature	NLU	spaCy	NLTK	CoreNLP	Hugging Face
Tokenization	**Yes**	Yes	Yes	Yes	Yes
Sentence segmentation	**Yes**	Yes	Yes	Yes	**No**
Steeming	**Yes**	Yes	Yes	Yes	**No**
Lemmatization	**Yes**	Yes	Yes	Yes	**No**
POS tagging	**Yes**	Yes	Yes	Yes	**No**
Entity recognition	**Yes**	Yes	Yes	Yes	Yes
Dep 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

NLU : Apache License 2.0

```
# Multiple binary sentiment classifiers trained on various datasets
nlu.load('classify.sentiment').predict('I love NLU and Python WebDev Conf 2021!')
nlu.load('classify.sentiment.imdb').predict('The Matrix was a pretty good movie')
nlu.load('classify.sentiment.twitter').predict('@elonmusk Tesla stock price is too high imo')

# Translate between 200 languages
nlu.load('en.translate_to.zh').predict('NLU can translate between 200 languages!')

# Spellchecking
nlu.load('spell').predict('I liek to live dangertus!')

# Extract Named Entities
nlu.load('ner').predict('Donald Trump and John Biden dont share many oppinions')

# Unsupervised Keyword Extraction
nlu.load('yake').predict('Weights extract keywords withouth requiring weights!')

# Over 50+ classifiers on various problems
nlu.load('classify.emotion').predict('He was suprised by the diversity of NLU')
nlu.load('classify.spam').predict('Hello you are the heir to a 100 Million fortune!')
nlu.load('classify.fakenews').predict('Unicorns landed on mars!')
nlu.load('classify.sarcasm').predict('love the teachers who give exams the day after halloween')
nlu.load('en.classify.question').predict('How expensive is the Watch?')
nlu.load('en.classify.toxic').predict('You are to stupid')
nlu.load('classify.cyberbullying').predict('Women belong in the kitchen!') #sorry

# Get BERTology and Transformer Embeddings for Sentences and Words
nlu.load('bert').predict('BERTolgy Word embeddings!')
nlu.load('bert elmo albert glove').predict('Multiple BERTolgy Word embeddings!')
nlu.load('embed_sentence.bert ').predict('BERTolgy Sentence embeddings!')

# Text cleaning and Pre-Processing
nlu.load('lemmatize').predict('Get me the lemmatized version of a string')
nlu.load('normalize').predict('Get me the lemmatized version of a string')
nlu.load('clean').predict('Get me the lemmatized version of a string')

# Grammatical Parts of Speech
nlu.load('pos').predict('Extract Parts of Speech')
```

- Tokenization
- Sentence Detector
- Stop Words Removal
- Normalizer
- Stemmer
- Lemmatizer
- NGrams
- Regex Matching
- Text Matching
- Chunking
- Date Matcher
- Part-of-speech tagging
- Dependency parsing
- Sentiment Detection (ML models)
- Spell Checker (ML and DL models)
- Word Embeddings

- BERT Embeddings
- ELMO Embeddings
- ALBERT Embeddings
- XLNet Embeddings
- Universal Sentence Encoder
- BERT Sentence Embeddings
- Sentence Embeddings
- Chunk Embeddings
- Unsupervised keywords extraction
- Language Detection & Identification
- Multi-class Text Classification
- Multi-label Text Classification
- Multi-class Sentiment Analysis
- Named entity recognition
- Easy TensorFlow integration
- Full integration with Spark ML functions
- +250 pre-trained models in 46 languages!
- +90 pre-trained pipelines in 13 languages!

109 Languages supported by **Language-agnostic BERT Sentence Embedding (LABSE)**

Train in **1 Language**, classify in **100 different languages** correct

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		

```
# 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)
```

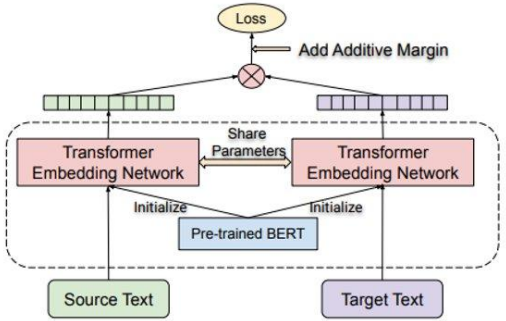
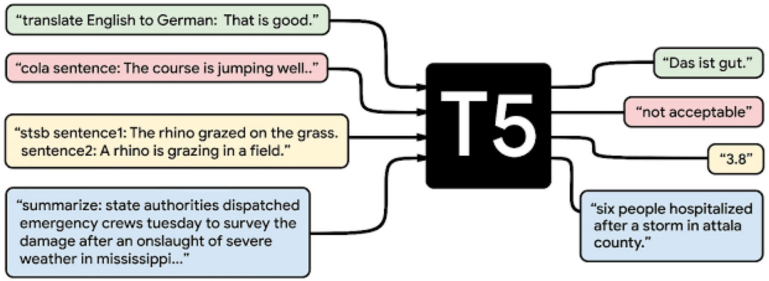


Figure 1: Dual encoder model with BERT based encoding modules.




```
# Closed book Question Answering
nlu.load('en.t5').predict('what is the capital of Germany?') # >>> Berlin
# Open Book Question answering
nlu.load('en.t5').predict('Who is president of Nigeria?') # >>> Muhammadu Buhari


# Open book Question Answering
context   = 'Peters last week was terrible! He had an accident and broke his leg while skiing!'
question1  = 'Why was peters week so bad?'
question2  = 'How did peter broke his leg?'
nlu.load('answer_question').predict(question1 + context) # >>> broke his leg
nlu.load('answer_question').predict(question2 + context) # >>> skiing



# Big T5 model for Summarization, Sentiment, Text Similarity and other SQUAD/GLUE tasks
pipe = nlu.load('t5')
pipe['t5'].settask('summarize')
pipe.predict(long_text)
```

[Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer]

1. Text summarization
2. Question answering
3. Translation
4. Sentiment analysis
5. Natural Language inference
6. Coreference resolution
7. Sentence Completion
8. Word sense disambiguation

Google

Every T5 Task with explanation:

Task Name	Explanation
1.CoLA	Classify if a sentence is grammatically correct
2.RTE	Classify whether a statement can be deducted from a sentence
3.MNLI	Classify for a hypothesis and premise whether they contradict or contradict each other or neither of both (3 class).
4.MRPC	Classify whether a pair of sentences is a re-phrasing of each other (semantically equivalent)
5.QNLI	Classify whether the answer to a question can be deducted from an answer candidate.
6.QQP	Classify whether a pair of questions is a re-phrasing of each other (semantically equivalent)
7.SST2	Classify the sentiment of a sentence as positive or negative
8.STSB	Classify the sentiment of a sentence on a scale from 1 to 5 (21 Sentiment classes)
9.CB	Classify for a premise and a hypothesis whether they contradict each other or not (binary).
10.COPA	Classify for a question, premise, and 2 choices which choice the correct choice is (binary).
11.MultiRc	Classify for a question, a paragraph of text, and an answer candidate, if the answer is correct (binary),
12.WiC	Classify for a pair of sentences and a disambigous word if the word has the same meaning in both sentences.
13.WSC/DPR	Predict for an ambiguous pronoun in a sentence what it is referring to.
14.Summarization	Summarize text into a shorter representation.
15.SQuAD	Answer a question for a given context.
16.WMT1.	Translate English to German
17.WMT2.	Translate English to French
18.WMT3.	Translate English to Romanian

Microsoft

Translate between **200+ Languages**

With Marian: Fast Neural Machine Translation in C++

MARIANNMT

Fast Neural Machine Translation in C++


```
# 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')`
```

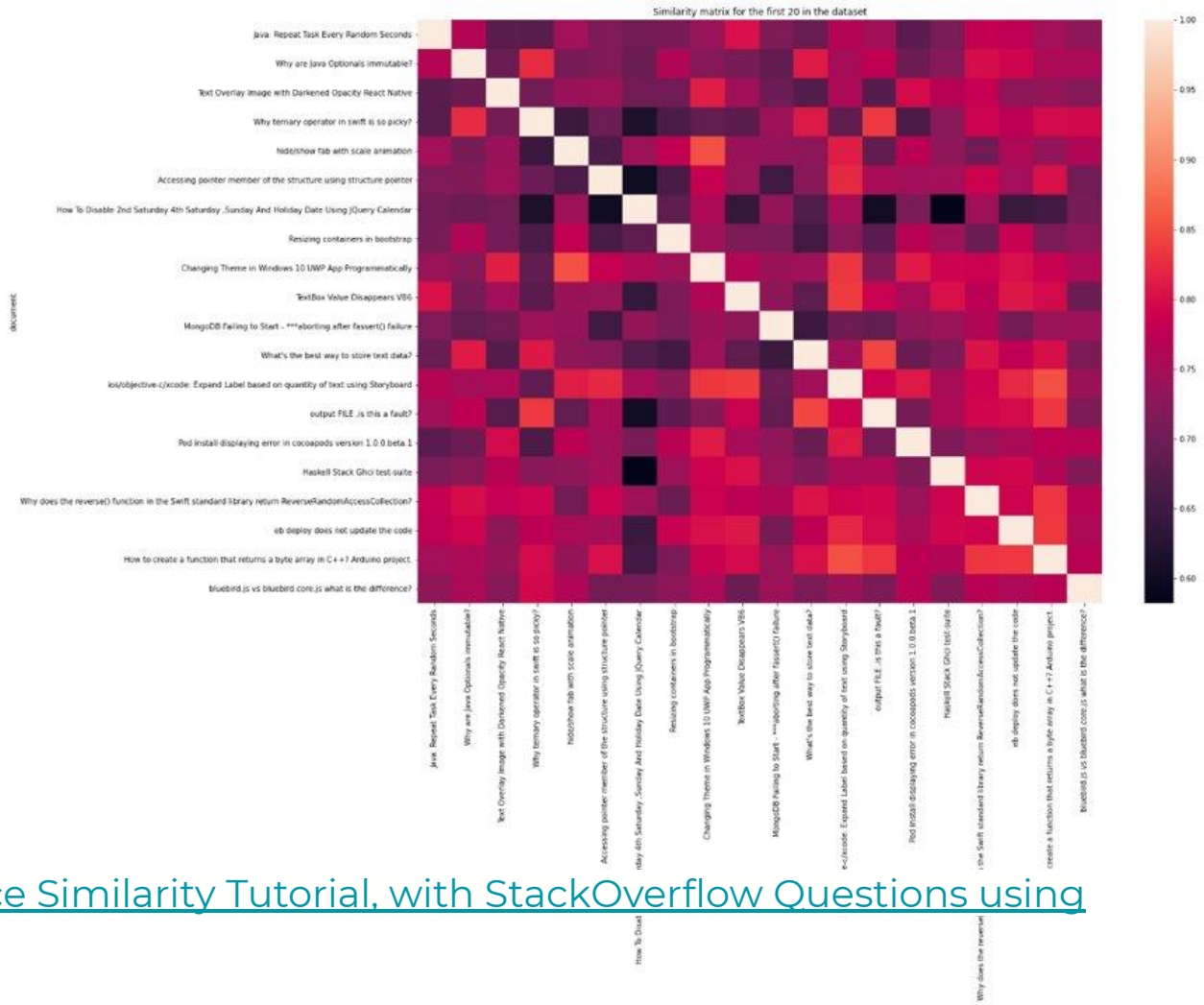
Demo Time

What does NLU include?

100% Open Source	1000+ pre-trained models	**100+** of the latest NLP word embeddings (BERT, ELMO, ALBERT, XLNET, GLOVE, BIOBERT, ELECTRA, COVIDBERT) and different variations of them	**50+** of the latest NLP sentence embeddings (BERT, ELECTRA, USE) and different variations of them	Multilingal Sentence and Word embeddings
50+ Classifiers	200 + Supported Languages	Labeled and Unlabeled Dependency parsing	Spell Checking	All in one line of code!
Unsupervised Keyword Extraction with YAKE	Various text-preprocessing and cleaning methods	Summarize	Answer questions	SQUAD/GLUE/SUPERGLUE
Train Classifiers	State of the art	Aspect NER	Aspect Sentiment	Intent Classification

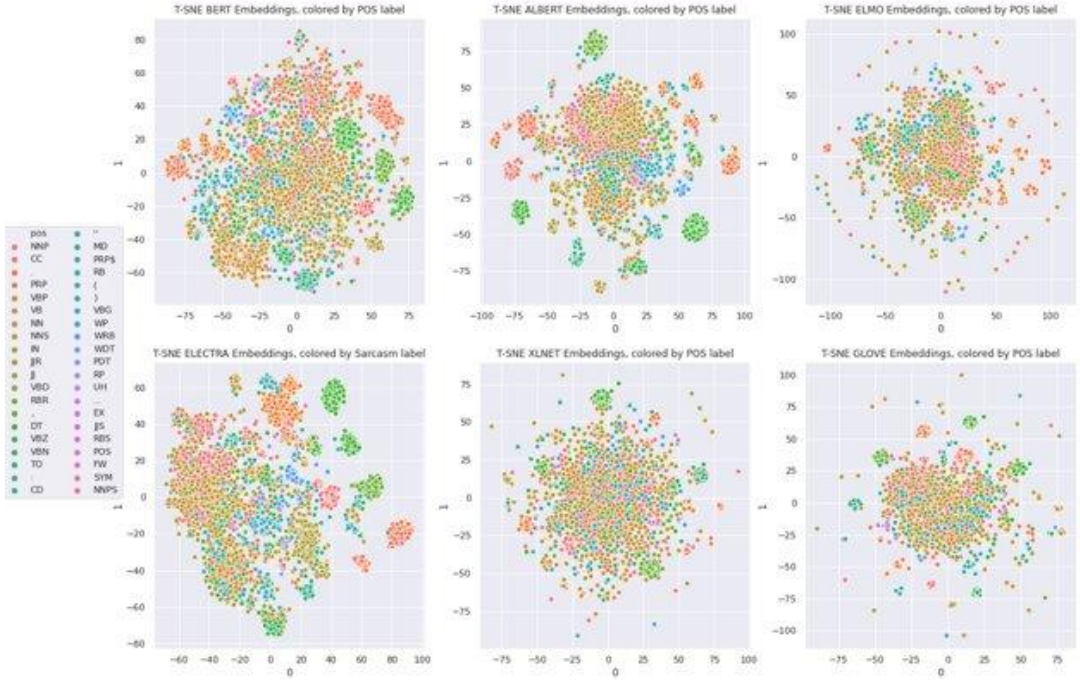
200+ Supported Languages

Sentence Similarity With BERTology Embeds or T5



[Indepth and Easy Sentence Similarity Tutorial, with StackOverflow Questions using BERTology embeddings]

t-SNE Visualizations with NLU



[1 line of Python code for BERT, ALBERT, ELMO, ELECTRA, XLNET, GLOVE, Part of Speech with NLU and t-SNE]

NLU RESSOURCES

- Join our Slack
- NLU Website
- NLU Github
- Many more NLU example tutorials
- Overview of every powerful nlu 1-liner
- Checkout the Modelshub for an overview of all models
- Checkout the NLU Namespace where you can find every model as a tabel
- Intro to NLU article
- Indepth and Easy Sentence Similarity Tutorial, with StackOverflow Questions using BERTology embeddings
- 1 line of Python code for BERT, ALBERT, ELMO, ELECTRA, XLNET, GLOVE, Part of Speech with NLU and t-SNE

Thank you.


christian@JohnSnowLabs.com


@ckl_it


In/Christian-Kasim-Loan

Medium.com/@Christian.Kasim.Loan