



Spark NLP for Healthcare Data Scientists

October 15-16, 2020

Veysel Kocaman

Sr. Data Scientist

veysel@johnsnowlabs.com

Welcome

Day-1	60 min	<ul style="list-style-type: none">- Intro to John Snow Labs and Spark NLP- Intro to NLP and Clinical NLP Modules in Spark NLP
	10 min	Break
	50 min	<ul style="list-style-type: none">- Pretrained Clinical Pipelines- Clinical Named Entity Recognition
	10 min	Break
	50 min	<ul style="list-style-type: none">- Clinical Assertion Status Model- Clinical Relation Extraction Model
Day-2	50 min	<ul style="list-style-type: none">- Clinical Entity Resolvers
	10 min	Break
	60 min	<ul style="list-style-type: none">- Deidentification and Obfuscation of PHI- Clinical Classifier (Adverse Drug Events)
	10 min	Break
	50 min	<ul style="list-style-type: none">- Spark OCR

Setup

 Open in Colab

RUNNING CODE:

https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/Certification_Trainings/Healthcare

[How to set up Google Colab]

BOOKMARK:

<https://nlp.johnsnowlabs.com/models>

<https://nlp.johnsnowlabs.com/docs/en/quickstart>
spark-nlp.slack.com

<https://www.johnsnowlabs.com/spark-nlp-in-action/>

master		spark-nlp-workshop / tutorials / Certification_Trainings / Healthcare /	Go to file	Add file ▾
 vkocaman	tf graph tip	✓ 4884270 8 days ago		
..				
 clinical_text_classification	ade classifier updated with bert	11 days ago		
 data	ADE dataset added	11 days ago		
 databricks_notebooks	clear cell output	2 months ago		
 generic_classifier_graph	clinical REE model	2 months ago		
 nerdl_graph	ADE classifier updated	last month		
 1.2.Contextual_Parser_Rule_Based_NER.ipynb	conll parser is added	21 days ago		
 1.3.prepare_CoNLL_from_annotations_for_NER.ipynb	1hr workshop notebook added	21 days ago		
 1.Clinical_Named_Entity_Recognition_Model.ipynb	tf graph tip	8 days ago		
 10.Clinical_Relation_Extraction.ipynb	1hr workshop notebook added	21 days ago		
 11.Pretrained_Clinical_Pipelines.ipynb	ade pipeline updated	10 days ago		
 12.Named_Entity_Disambiguation.ipynb	ocr lines removed	last month		
 13.Snomed_Entity_Resolver_Model_Training.ipynb	ocr lines removed	last month		
 14.German_Healthcare_Models.ipynb	German models splitted	last month		
 15.German_Legal_Model.ipynb	German models splitted	last month		
 16.Adverse_Drug_Event_ADE_NER_and_Classifier.ipynb	ade pipeline updated	10 days ago		
 17.Graph_builder_for_DL_models.ipynb	ADE dataset added	11 days ago		
 2.1.Clinical_Assertion_Graph_Generation.ipynb	conll parser is added	21 days ago		
 2.Clinical_Assertion_Model.ipynb	ocr lines removed	last month		
 3.Clinical_Entity_Resolvers.ipynb	typo fixed	9 days ago		
 4.Clinical_Deidentification.ipynb	ocr lines removed	last month		
 5.Spark_OCR.ipynb	OCR keys updated	last month		
 6.Clinical_Context_Spell_Checker.ipynb	ocr lines removed	last month		
 7.Clinical_NER_Chunk_Merger.ipynb	ocr lines removed	last month		
 8.Generic_Classifier.ipynb	ocr lines removed	last month		

Part - I

- ❖ Overview and key concepts in Spark NLP
- ❖ NLP basics & review
- ❖ Common medical NLP use cases
- ❖ Clinical named entity recognition



"John Snow Labs enables healthcare organizations to deploy state-of-the-art artificial intelligence (AI) platforms, models and data in production today."

JOHN SNOW LABS



"John Snow Labs wows in both proven customer success and verifiable state-of-the-art technology – making it a natural winner of the highly competitive 2019

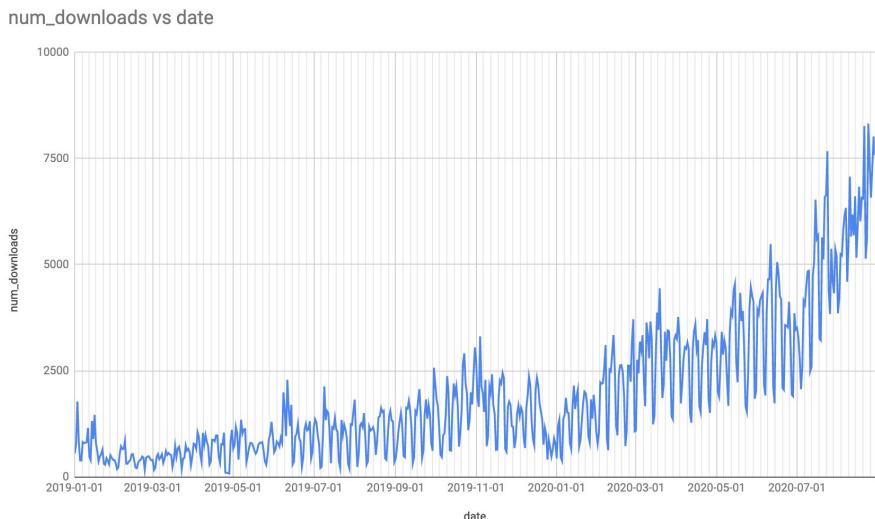
AI Platform of the Year Award."



"Keep an eye on this company – as it represents where the industry and data science are headed."

Introducing Spark NLP

Daily ~ 10K	
Monthly ~ 250K	
PyPI link	https://pypi.org/project/spark-nlp
Total downloads	1,471,785
Total downloads - 30 days	245,960
Total downloads - 7 days	76,762



- Spark NLP is an open-source natural language processing library, built on top of Apache Spark and Spark ML. (initial release: Oct 2017)
 - A single unified solution for all your NLP needs
 - Take advantage of transfer learning and implementing the latest and greatest SOTA algorithms and models in NLP research
 - The most widely used NLP library in industry (3 yrs in a row)
 - Delivering a mission-critical, enterprise grade NLP library (used by multiple Fortune 500)
 - Full-time development team (26 new releases in 2018. 30 new releases in 2019.)

Spark NLP

- 73 total releases
- Release every two weeks for the past 3 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
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

TRUSTED BY



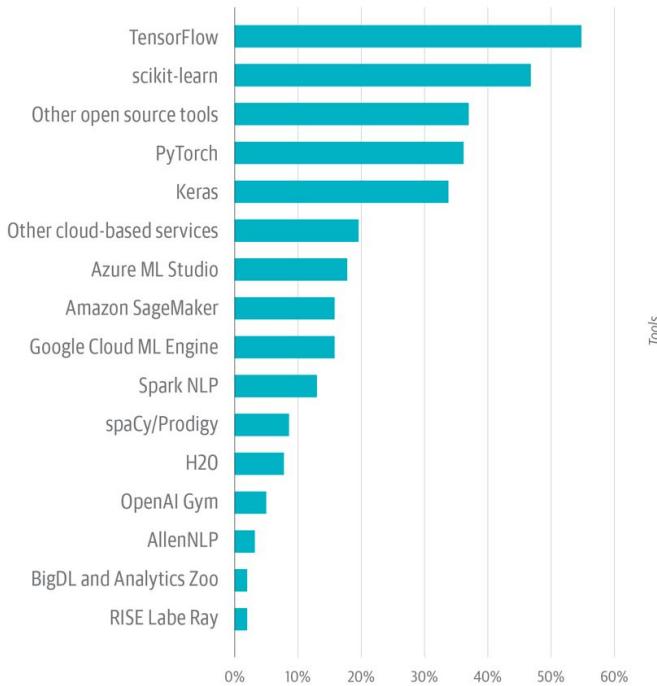
Imperial College
London



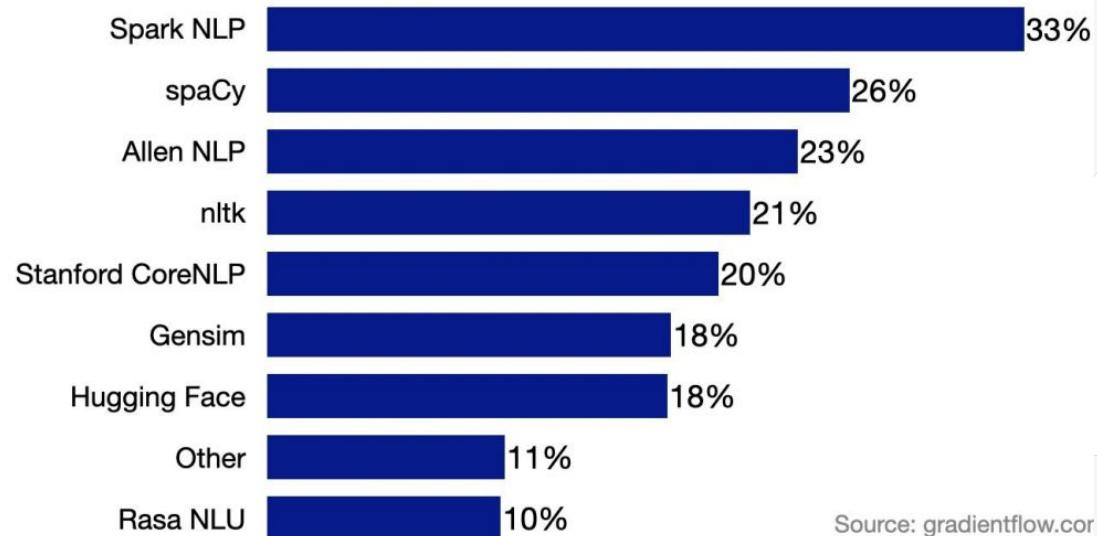
STANFORD
UNIVERSITY

Spark NLP in Industry

Which of the following AI tools do you use?



Which NLP libraries does your organization use?



Source: gradientflow.co

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

OFFICIALLY SUPPORTED RUNTIMES



databricks

CLOUDERA



Azure



Spark NLP: Apache License 2.0

```
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)
```

- 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!

Spark NLP Modules (Enterprise and Public)

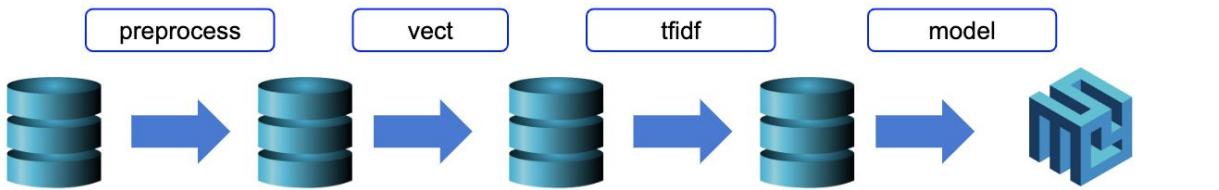
Clinical Entity Recognition	Clinical Entity Linking	Assertion Status	De-Identification						
40 units: DOSAGE of Insulin glargine DRUG at night FREQUENCY	Suspect diabetes: SNOMED-CT: 473127005 Lisinopril 10 MG RxNorm: 316151 Pyponatremia ICD-10: E87.1	Fever and sore throat → PRESENT No stomach pain → ABSENT Father with Alzheimer → FAMILY	Ora [NAME] a 25 [AGE] yo cashier [PROFESSION] from Morocco [LOCATION]						
Algorithms		Content							
Extract Knowledge <ul style="list-style-type: none"> Entity Linker Entity Disambiguator Document Classifier Contextual Parser 	De-Identity Text <ul style="list-style-type: none"> Structured Data Unstructured Text Obfuscator Generalizer 	Medical Transformers JSL-BERT-Clinical BioBERT GloVe-Med GloVe-ICD-O	Linked Medical Terminologies SNOMED-CT CPT ICD-10-CM RxNorm ICD-10-PCS ICD-O						
Split Text <ul style="list-style-type: none"> Sentence Detector Deep Sentence Detector Tokenizer nGram Generator 	Clean Medical Text <ul style="list-style-type: none"> Spell Checking Spell Correction Normalizer Stopword Cleaner 	50+ Pretrained Models <table border="1"> <tr> <td>Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs</td> <td>Anatomy: Organ, Subdivision, Cell, Structure</td> </tr> <tr> <td>Biological: Organism, Tissue, Gene, Chemical</td> <td>Demographics: Age, Gender, Vital Signs, Smoking Indicators</td> </tr> <tr> <td>Drugs: Name, Dosage, Strength, Route, Duration, Frequency</td> <td>Sensitive Data: Patient Name, Address, Dates, Providers, Identifiers</td> </tr> </table>		Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs	Anatomy: Organ, Subdivision, Cell, Structure	Biological: Organism, Tissue, Gene, Chemical	Demographics: Age, Gender, Vital Signs, Smoking Indicators	Drugs: Name, Dosage, Strength, Route, Duration, Frequency	Sensitive Data: Patient Name, Address, Dates, Providers, Identifiers
Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs	Anatomy: Organ, Subdivision, Cell, Structure								
Biological: Organism, Tissue, Gene, Chemical	Demographics: Age, Gender, Vital Signs, Smoking Indicators								
Drugs: Name, Dosage, Strength, Route, Duration, Frequency	Sensitive Data: Patient Name, Address, Dates, Providers, Identifiers								
Clinical Grammar <ul style="list-style-type: none"> Stemmer Lemmatizer Part of Speech Tagger Dependency Parser 	Find in Text <ul style="list-style-type: none"> Text Matcher Regex Matcher Date Matcher Chunker 								

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

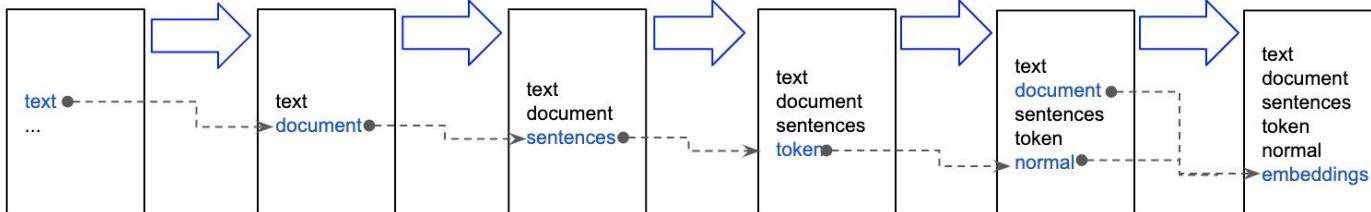
Entity Recognition	Information Extraction	Sentiment Analysis	Document Classification
Algorithms		Content	
Split Text <ul style="list-style-type: none"> Sentence Detector Deep Sentence Detector Tokenizer nGram Generator 	Clean Text <ul style="list-style-type: none"> Spell Checking Spell Correction Normalizer Stopword Cleaner 	Transformers GloVe ELMO BERT ALBERT XLNet	Languages Bulgarian Czech Dutch English French German Greek Hungarian Italian Finnish Norwegian Polish Portuguese Spanish Romanian Russian Swedish Turkish Ukrainian
Understand Grammar <ul style="list-style-type: none"> Stemmer Lemmatizer Part of Speech Tagger Dependency Parser 	Find in Text <ul style="list-style-type: none"> Text Matcher Regex Matcher Date Matcher Chunker 	Models 90+ Pretrained	Pipelines 70+ Pretrained
Trainable & Tunable 	Scalable to a Cluster 	Fast Inference 	Hardware Optimized  
Community 			

Introducing Spark NLP

Pipeline of annotators



DocumentAssembler() SentenceDetector() Tokenizer() Normalizer() WordEmbeddings()



```
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)
```

Introducing Spark NLP



Faster inference

```
from sparknlp.base import LightPipeline  
LightPipeline(someTrainedPipeline).annotate(someStringOrArray)
```

Spark is like a [locomotive](#) racing a [bicycle](#). The [bike](#) will win if the load is light, it is quicker to accelerate and more agile, but with a heavy load the [locomotive](#) might take a while to get up to speed, but [it's](#) going to be faster in the end.

LightPipelines are Spark ML pipelines converted into a single machine but multithreaded task, becoming more than 10x times faster for smaller amounts of data (small is relative, but 50k sentences is roughly a good maximum).

Natural Language Processing

Information Retrieval

Doc A



Doc 1

Doc 2

Doc 3

Sentiment Analysis



Information Extraction



Machine Translation



Question Answering



Human: When was Apollo sent to space?

Machine: First flight -
AS-201,
February 26,
1966

NLP Basics

LEMMATIZATION

Find the **lemma** of each word:

- How does it show in the dictionary?

Uses a lookup from a full dictionary.

am, are, is → be

liver → liver

lives → live

STEMMING

Find the **stem** of each word.

Uses rules: e.g, remove common suffixes.

Form	Suffix	Stem
studies	-es	studi
study ing	- ing	study
niñ as	- as	niñ
niñ ez	- ez	niñ

- The goal of both **stemming** and **lemmatization** is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form for normalization purposes.
- Lemmatization always returns real words, **stemming** doesn't.

NLP Basics

Remove stop words and apply stemming

it was a bright cold day in april
and the clocks **were** striking
thirteen winston smith **his** chin
nuzzled **into his** breast in an
effort **to** escape **the** vile wind
slipped quickly **through the** glass
doors **of** victory mansions though
not quickly enough **to** prevent a
swirl **of** gritty dust **from** entering
along **with him**



bright cold day april clocks
striking thirteen winston smith
chin nuzzled breast effort
escape vile wind slipped quickly
glass doors victory mansions
though quickly enough prevent
swirl gritty dust entering along

- For tasks like text classification, where the text is to be classified into different categories, **stopwords** are **removed** or excluded from the given text so that more focus can be given to those words which define the meaning of the text.

Stopwords

a
able
about
above
according
accordingly
across
actually
after
afterwards
again
against
ain
all
allow
allows
almost
alone
along
already
also

(520 stopwords)

Spell Checking & Correction



```
val pipeline = PretrainedPipeline("spell_check_ml", "en")
val result = pipeline.annotate("Harry Potter is a graet muvie")

println(result("spell"))
/* will print Seq[String](..., "is", "a", "great", "movie") */
```

- 3 trainable approaches
- **Norvig Approach:**
 - Retrieves tokens and auto-corrects based on a given dictionary
- **Symmetric Delete:**
 - Uses distance metrics to find possible words
- **Context Aware:**
 - Most accurate: Judges words in context
 - Deep learning based

Context Spell Checker

The Spell Checker can leverage the context of words for ranking different correction sequences. Let's take a look at some examples,

```
# 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']
```

```
# check for the different occurrences of the word "ueather"
example2 = ["During the summer we have the best ueather.", \
            "I have a black ueather jacket, so nice.", \
            "I introduce you to my sister, she is called ueather."]
beautify(lp.annotate(example2))
```

```
['During the summer we have the best weather .\n',
 'I have a black leather jacket , so nice .\n',
 'I introduce you to my sister , she is called Heather .\n']
```

Notice that in the first example, 'siter' is indeed a valid English word,

<https://www.merriam-webster.com/dictionary/siter>

NORMALIZATION

Remove or replace undesirable characters or regular expressions:

from: @Have a\$ #2great birth) day>!
to: Have a great birth day!

Spark NLP also comes with a Slang normalizer:

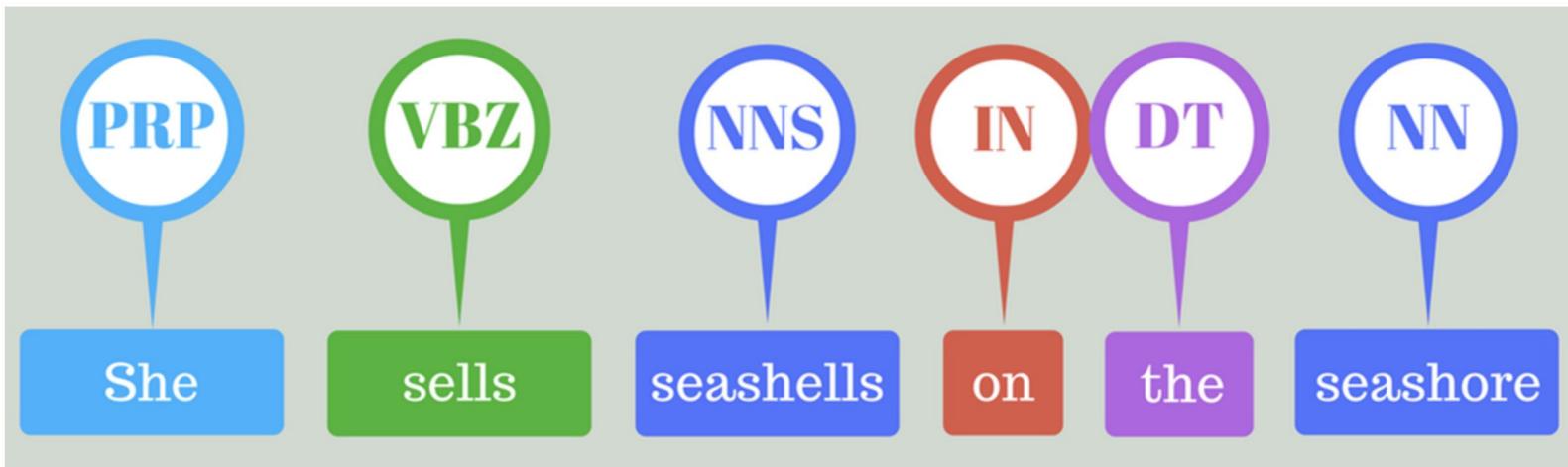
Original tweet
@USER, r u cuming 2 MidCorner dis Sunday?

Normalized tweet

@USER, are you coming to MidCorner this Sunday?

PART OF SPEECH TAGGING

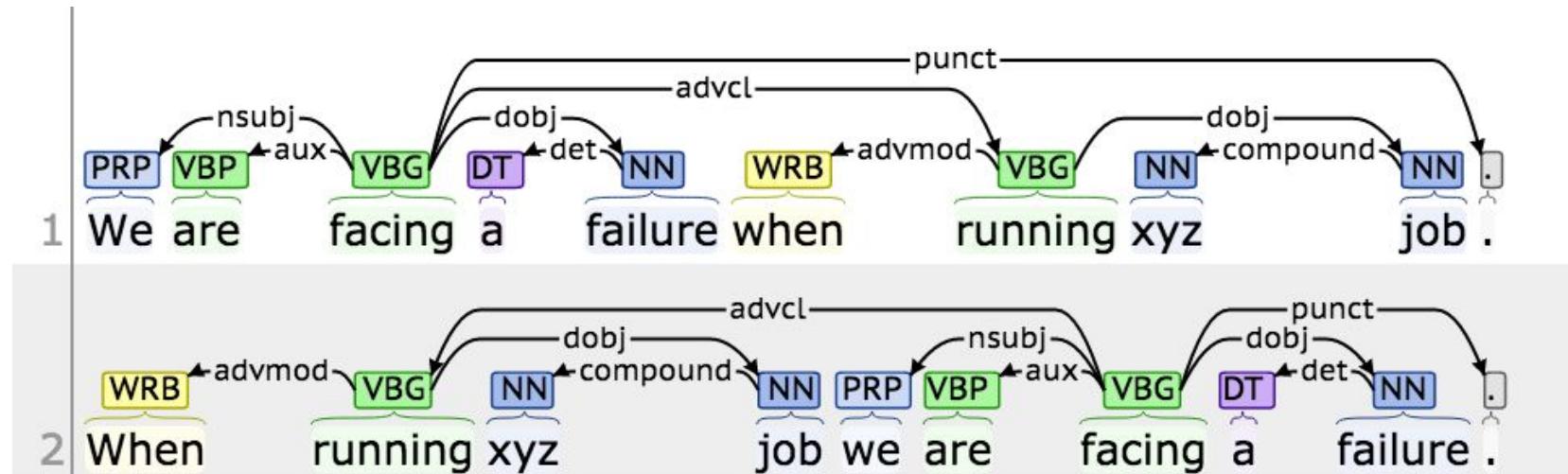
Often useful for recognizing named entities or word relationships.



A **POS tag** (or **part-of-speech tag**) is a special label assigned to each token (word) in a text corpus to indicate the **part of speech** and often also other grammatical categories such as tense, number (plural/singular), case etc.

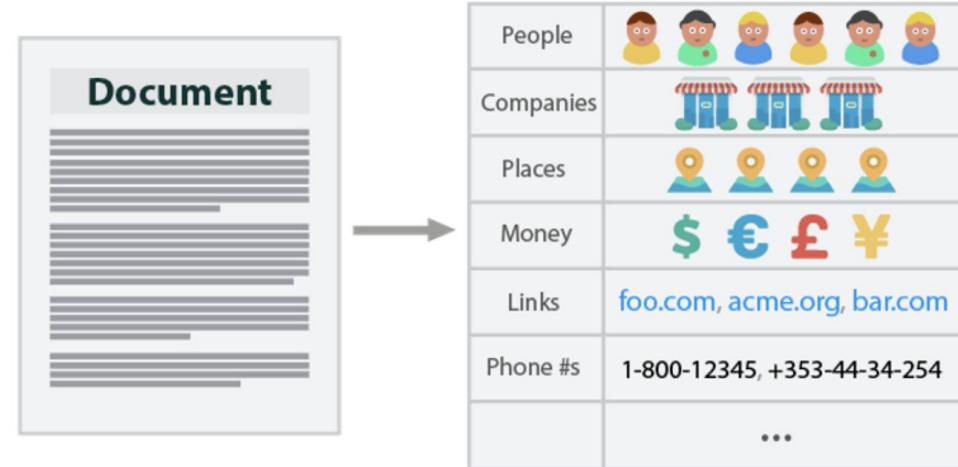
DEPENDENCY PARSING

Useful for extracting relationships (i.e. building knowledge graphs):



Named Entity Recognition (NER)

NER is a subtask of information extraction that seeks to **locate and classify named entity** mentioned in unstructured text into pre-defined categories such as **person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.**



But Google **ORG** is starting from behind. The company made a late push into hardware, and Apple **ORG**'s Siri **PRODUCT**, available on iPhones **PRODUCT**, and Amazon **ORG**'s Alexa **PRODUCT** software, which runs on its Echo **PRODUCT** and Dot **PRODUCT** devices, have clear leads in consumer adoption.

Word & Sentence Embeddings

Vocabulary

index: Word:

0 aardvark
1 able
...

2409 black
2410 bling
...

3202 candid

3203 cast

3204 cat

...

5281 is

5282 island

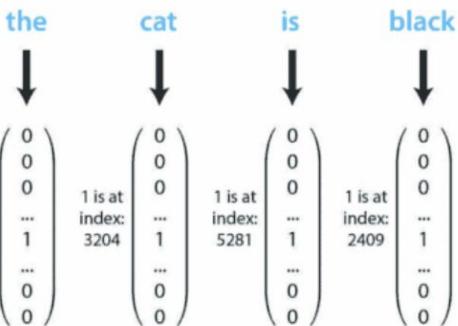
...

8676 the

8677 thing

...

9999 zombie



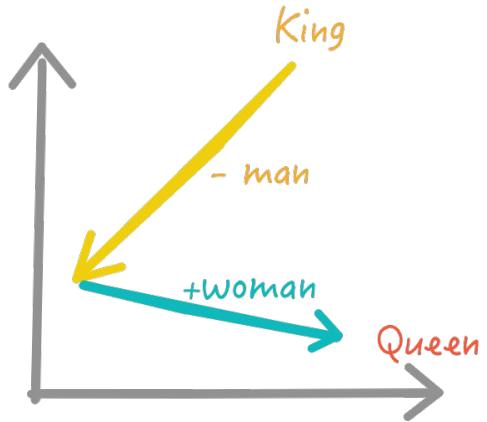
One-hot vector encoding for words in input sentence complete.

In [9]: doc[3].vector

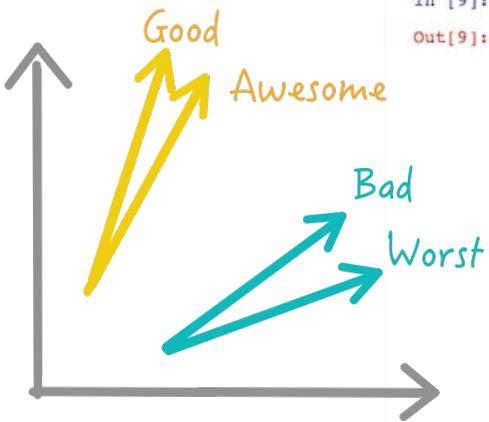
```
Out[9]: array([ 0.037103 , -0.31259 , -0.17857 ,  0.30001 ,  0.078154 ,
 0.17958 ,  0.12048 , -0.11879 , -0.20601 ,  1.2849 ,
-0.20409 ,  0.80613 ,  0.34344 , -0.19191 , -0.084511 ,
 0.17339 ,  0.042483 ,  2.0282 , -0.16278 , -0.60306 ,
-0.53766 ,  0.35711 ,  0.22882 ,  0.1171 ,  0.42983 ,
 0.16165 ,  0.407 ,  0.036476 ,  0.52636 , -0.13524 ,
-0.016897 ,  0.029259 , -0.079115 , -0.32305 ,  0.052255 ,
-0.3617 , -0.18355 , -0.34717 , -0.3691 ,  0.16881 ,
 0.21018 , -0.38376 , -0.096909 , -0.36296 , -0.37319 ,
 0.00211152,  0.32512 ,  0.063977 ,  0.36249 , -0.26935 ,
-0.59341 , -0.13625 ,  0.016425 , -0.2474 , -0.07498 ,
 0.034708 , -0.01476 , -0.11648 ,  0.25559 , -0.35002 ,
-0.52707 ,  0.21221 ,  0.062456 ,  0.26184 ,  0.53149 ,
 0.34957 , -0.22692 ,  0.44076 ,  0.4438 ,  0.6335 ,
-0.049757 , -0.08134 ,  0.65618 , -0.4716 ,  0.090675 ,
-0.084873 ,  0.31455 , -0.38495 , -0.19247 ,  0.48064 ,
 0.26688 ,  0.095743 ,  0.13024 ,  0.37023 ,  0.46269 ,
-0.32844 ,  0.17375 , -0.36325 ,  0.30672 , -0.075042 ,
-0.64684 , -0.49822 ,  0.12372 , -0.28547 ,  0.61811 ,
-0.19228 ,  0.0040473 ,  0.1774 ,  0.033154 , -0.54862 ,
 0.34695 , -0.53506 , -0.013381 ,  0.085712 , -0.054447 ,
-0.64673 ,  0.016749 ,  0.47676 ,  0.037803 , -0.10066 ,
-0.4165 , -0.20252 ,  0.2794 ,  0.10852 , -0.40154 ])
```

- Words that are used in similar contexts will be given similar representations. That is, words that are used in similar ways will be placed close together within the high-dimensional semantic space—these points will cluster together, and their distance to each other will be low.

Word & Sentence Embeddings



a) Learns Analogy



b) Similar Words have same angles

```
In [9]: doc[3].vector
```

```
Out[9]: array([ 0.037103 , -0.31259 , -0.17857 ,  0.30001 ,  0.078154 ,  
  0.17958 ,  0.12048 , -0.11879 , -0.20601 ,  1.2849 ,  
 -0.20409 ,  0.80613 ,  0.34344 , -0.19191 , -0.084511 ,  
  0.17339 ,  0.042483 ,  2.0282 , -0.16278 , -0.60306 ,  
 -0.53766 ,  0.35711 ,  0.22882 ,  0.1171 ,  0.42983 ,  
  0.16165 ,  0.407 ,  0.036476 ,  0.52636 , -0.13524 ,  
 -0.016897 ,  0.029259 , -0.079115 , -0.32305 ,  0.052255 ,  
 -0.3617 , -0.18355 , -0.34717 , -0.3691 ,  0.16881 ,  
  0.21018 , -0.38376 , -0.096909 , -0.36296 , -0.37319 ,  
  0.0021152,  0.32512 ,  0.063977 ,  0.36249 , -0.26935 ,  
 -0.59341 , -0.13625 ,  0.016425 , -0.2474 , -0.07498 ,  
  0.034708 , -0.01476 , -0.11648 ,  0.25559 , -0.35002 ,  
 -0.52707 ,  0.21221 ,  0.062456 ,  0.26184 ,  0.53149 ,  
  0.34957 , -0.22692 ,  0.44076 ,  0.4438 ,  0.6335 ,  
 -0.049757 , -0.08134 ,  0.65618 , -0.4716 ,  0.090675 ,  
 -0.084873 ,  0.31455 , -0.38495 , -0.19247 ,  0.48064 ,  
  0.26688 ,  0.095743 ,  0.13024 ,  0.37023 ,  0.46269 ,  
 -0.32844 ,  0.17375 , -0.36325 ,  0.30672 , -0.075042 ,  
 -0.64684 , -0.49822 ,  0.12372 , -0.28547 ,  0.61811 ,  
 -0.19228 ,  0.0040473 ,  0.1774 ,  0.033154 , -0.54862 ,  
  0.34695 , -0.53506 , -0.013381 ,  0.085712 , -0.054447 ,  
 -0.64673 ,  0.016749 ,  0.47676 ,  0.037803 , -0.10066 ,  
 -0.4165 , -0.20252 ,  0.2794 ,  0.10852 , -0.40154 ])
```

- Deep-Learning-based natural language processing systems.
- They encode **words** and **sentences** in fixed-length dense vectors to drastically improve the processing of textual data.
- Based on **The Distributional Hypothesis**: Words that occur in the same contexts tend to have similar meanings.

Word & Sentence Embeddings

Glove
(100, 200, 300)

ELMO
(512, 1024)

BERT
(768d)

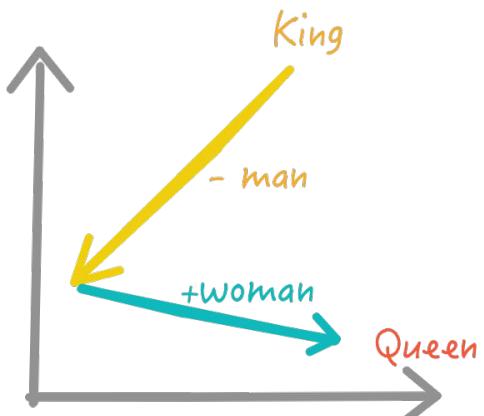
Universal Sentence Encoders
(512)

Albert
(768, 1024, 2048, 4096)

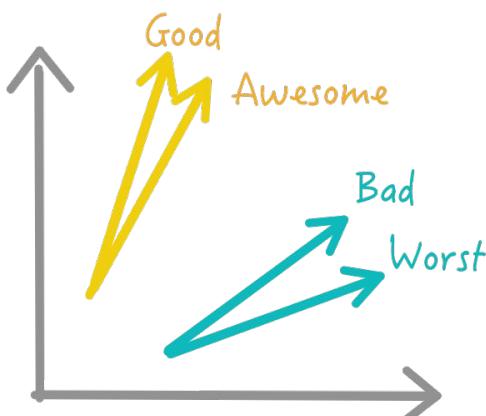
XLNet
(768, 1024)

Electra
(768)

Bert Sentence Embeddings
(768)



a) Learns Analogy



b) Similar Words have same angles

- Deep-Learning-based natural language processing systems.
- They encode **words** and **sentences** in fixed-length dense vectors to drastically improve the processing of textual data.
- Based on **The Distributional Hypothesis**: Words that occur in the same contexts tend to have similar meanings.
- Elmo and Bert-family embeddings are context-aware.

Text Classification with Word & Sentence Embeddings

Glove
(100, 200, 300)

ELMO
(512, 1024)

BERT
(768d)

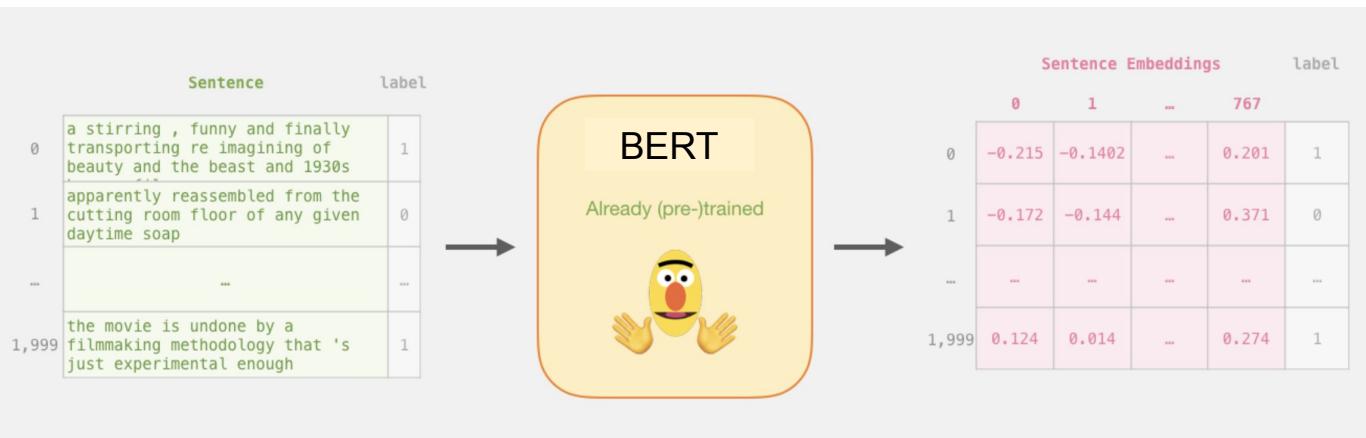
Universal Sentence Encoders
(512)

Albert
(768, 1024, 2048, 4096)

XLNet
(768, 1024)

Electra
(768)

Bert Sentence Embeddings
(768)



Model	Name
WordEmbeddingsModel (GloVe)	glove_100d
BertEmbeddings	electra_small_uncased
BertEmbeddings	electra_base_uncased
BertEmbeddings	electra_large_uncased
BertEmbeddings	bert_base_uncased
BertEmbeddings	bert_base_cased
BertEmbeddings	bert_large_uncased
BertEmbeddings	bert_large_cased
BertEmbeddings	biobert_pubmed_base_cased
BertEmbeddings	biobert_pubmed_large_cased
BertEmbeddings	biobert_pubmed_pmc_base_cased
BertEmbeddings	biobert_clinical_base_cased
BertEmbeddings	biobert_discharge_base_cased
BertEmbeddings	covidbert_large_uncased
BertEmbeddings	small_bert_L2_128
BertEmbeddings	small_bert_L4_128
BertEmbeddings	small_bert_L6_128
BertEmbeddings	small_bert_L8_128

WordEmbeddingsModel	embeddings_clinical	2.4.0	
WordEmbeddingsModel	embeddings_healthcare_100d	2.5.0	
WordEmbeddingsModel	embeddings_healthcare	2.4.4	
BertEmbeddings	biobert_pubmed_base_cased	2.6.0	en
BertEmbeddings	biobert_pubmed_large_cased	2.6.0	en
BertEmbeddings	biobert_pubmed_pmc_base_cased	2.6.0	en
BertEmbeddings	biobert_clinical_base_cased	2.6.0	en
BertEmbeddings	biobert_discharge_base_cased	2.6.0	en
BertEmbeddings	covidbert_large_uncased	2.6.0	en
BertSentenceEmbeddings	sent_bert_large_cased	2.6.0	en
BertSentenceEmbeddings	sent_biobert_pubmed_base_cased	2.6.0	en
BertSentenceEmbeddings	sent_biobert_pubmed_large_cased	2.6.0	en
BertSentenceEmbeddings	sent_biobert_pubmed_pmc_base_cased	2.6.0	en
BertSentenceEmbeddings	sent_biobert_clinical_base_cased	2.6.0	en
BertSentenceEmbeddings	sent_biobert_discharge_base_cased	2.6.0	en
BertSentenceEmbeddings	sent_covidbert_large_uncased	2.6.0	en

Clinical Word Embeddings

Clinical Glove
(200d)

PubMed + PMC

ICDO Glove
(200d)

PubMed + ICD10
UMLS + MIMIC III

Bio BERT

Pubmed + PMC

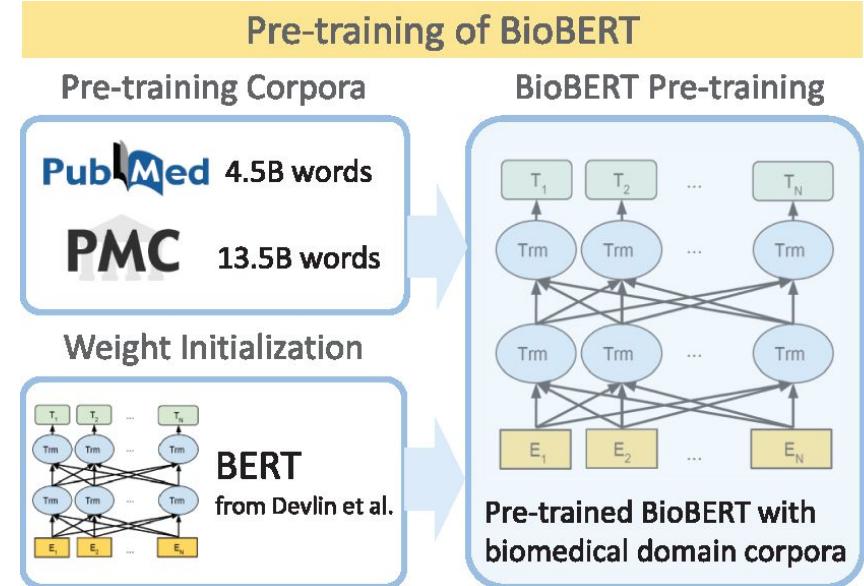
Clinical BERT

Fine tuned Pubmed + PMC + Discharge summaries



PubMed abstracts and PMC full-text articles

<https://www.nlm.nih.gov/bsd/difference.html>



Spark NLP in Healthcare

Entity Recognition & Data Normalization

500mg

Dosage: 500
Unit: mg

Sentiment Analysis

nasty

Sentiment: Negative

Data normalization, Standard Coding

SOB

SNOMED-CT: 267036007
Preferred Name: Dyspnea

Prescribing **500mg azithromycin** for **nasty pneumonia** w/o **SOB**.

POS tagging
Prescribing
Verb: to prescribe

Normalization for clinical drugs
azithromycin
Drug: azithromycin
RxNorm: C0732484

Spell checker
pneumonia
Suggested spelling:
pneumonia

Negation
w/o
Scope:
Negative

Spark NLP in Healthcare

Clean & structured data



Raw & unstructured data



Healthcare data



- Less than **50% of the structured data** and less than **1% of the unstructured data** is being leveraged for decision making in companies (HBR). This is even worse in healthcare.
- NLP is ultra domain specific, so train your own models.

Why is language understanding hard?

Human Language is:

- Nuanced
- Fuzzy
- Contextual
- Medium specific
- Domain specific

Healthcare specific needs:

1. Core Annotators

Part of speech, spell checking, ...

2. Vocabulary

Ontologies, relationships, word embeddings, ...

3. ML & DL Models

Named entity recognition, entity resolution, ...

ED Triage Notes
states started last night, upper abd, took alka seltzer approx 0500, no relief. nausea no vomiting
Since yesterday 10/10 "constant Tylenol 1 hr ago. +nausea. diaphoretic. Mid abd radiates to back
Generalized abd radiating to lower x 3 days accompanied by dark stools. Now with bloody stool this am. Denies dizzy, sob, fatigue. Visiting from Japan on business."



Features	
Type of Pain	Symptoms
Intensity of Pain	Onset of symptoms
Body part of region	Attempted home remedy

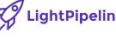
Healthcare extensions

NLP Library / Feature	State of the Art (SOTA) Research
Named Entity Recognition	"Entity Recognition from Clinical Texts via Recurrent Neural Network". <i>Liu et al., BMC Medical Informatics & Decision Making, July 2017.</i>
Word Embeddings	<ul style="list-style-type: none">- "How to Train Good Word Embeddings for Biomedical NLP". <i>Chiu et al., In Proceedings of BioNLP'16, August 2016.</i>- "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". <i>Devlin et. al. (Google Research), October 2018.</i>
Assertion Status Detection	<ul style="list-style-type: none">- "Improving Classification of Medical Assertions in Clinical Notes". <i>Kim et al., In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.</i>- "Neural Networks For Negation Scope Detection" <i>Fancellu et al., In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.</i>
Entity Resolution	"CNN-based ranking for biomedical entity normalization". <i>Li et al., BMC Bioinformatics, October 2017.</i>

Clinical Named Entity Recognition (NER)

- Extract structured data from free text
- Automate record keeping & abstraction process
- Feeding downstream tasks
- Features for ML models

Clinical Entity Recognition	Clinical Entity Linking	Assertion Status	Relation Extraction
40 units DOSAGE of insulin glargine DRUG at night FREQUENCY	Suspect diabetes SNOMED-CT: 473127005 Lisinopril 10 MG RxNorm: 316151 Hyponatremia ICD-10: E87.1	Fever and sore throat → PRESENT No stomach pain → ABSENT Father with Alzheimer → FAMILY	AFTER Admitted for nausea due to chemo Occurrence Symptom Treatment CAUSED BY
Algorithms		Content	
Extract Knowledge <ul style="list-style-type: none">Entity LinkerEntity DisambiguatorDocument ClassifierContextual Parser	De-identify text <ul style="list-style-type: none">Structured DataUnstructured TextObfuscatorGeneralizer	Medical Transformers <ul style="list-style-type: none">JSL-BERT-ClinicalBioBERTClinicalBERTGloVe-MedGloVe-ICD-O	Linked Medical Terminologies <ul style="list-style-type: none">SNOMED-CTCPTICD-10-CMICD-OICD-10-PCSRxNormLOINC
Split Text <ul style="list-style-type: none">Sentence DetectorDeep Sentence DetectorTokenizernGram Generator	Clean Medical Text <ul style="list-style-type: none">Spell CheckingSpell CorrectionNormalizerStopword Cleaner	50+ Pretrained Models	
Clinical Grammar <ul style="list-style-type: none">StemmerLemmatizerPart of Speech TaggerDependency Parser	Find in Text <ul style="list-style-type: none">Text MatcherRegex MatcherDate MatcherChunker	Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections	Anatomy: Organ, Subdivision, Cell, Structure, Organism, Tissue, Gene, Chemical
		Drugs: Name, Dosage, Strength, Route, Duration, Frequency	Demographics: Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs
		Risk Factors: Smoking, Obesity, Diabetes, Hypertension, Substance Abuse	Sensitive Data: Patient Name, Address, Phone, Email, Dates, Providers, Identifiers

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

Clinical Named Entity Recognition

Pretrained NER Models in Spark NLP

PROBLEM	X	TEST	X	TREATMENT	X	DURATION	X	EVIDENTIAL	X	TREATMENT	X	FREQUENCY	X
OCCURRENCE	X	TEST	X	TIME	X	PROBLEM	X	DATE	X	CLINICAL_DEPT	X	Disease	X
DrugChem	X	Immaterial_anatomic...	X	Developing_anatomic...	X	PathologicalFormation	X	Organ	X				
Organism_subdivision	X	Cellular_component	X	Multi	X	Tissue	X	Anatomical_system	X				
Organism_substance	X	Cell	X	DURATION	X	ROUTE	X	FREQUENCY	X	DOSAGE	X	DRUG	X
FORM	X	STRENGTH	X	Immaterial_anatomic...	X	Developing_anatomic...	X	PathologicalFormation	X				
Cancer	X	Organism	X	Organ	X	Organism_subdivision	X	Cellular_component	X	Amino_acid	X		
Multi	X	Tissue	X	Anatomical_system	X	Gene_or_gene_product	X	Simple_chemical	X				
Organism_substance	X	Cell	X	Weight	X	Drug_Name	X	Negation	X	Procedure	X		
Causative_Agents_(Vi...	X	O2_Saturation	X	Route	X	Temperature	X	Procedure_Name	X				
Substance_Name	X	Symptom_Name	X	Respiratory_Rate	X	Dosage	X	Name	X	Gender	X		
Pulse_Rate	X	Lab_Result	X	Lab_Name	X	Maybe	X	Allergenic_substance	X	Age	X	Frequency	X
Diagnosis	X	Modifier	X	Section_Name	X	Blood_Pressure	X	MEDICATION	X	CAD	X		
HYPERLIPIDEMIA	X	FAMILY_HIST	X	DIABETES	X	SMOKER	X	OBSE	X	PHI	X		
HYPERTENSION	X	RNA	X	cell_type	X	protein	X	cell_line	X	DNA	X	CHEM	X
Organization	X	Body_System	X	Professional_or_Occu...	X	Clinical_Attribute	X	Indicator_Reagent	...				
Organic_Chemical	X	Anatomical_Structure	X	Organism_Attribute	X	Food	X	Body_Part,_Organ,_or...	X				
Biologic_Function	X	Medical_Device	X	Tissue	X	Disease_or_Syndrome	X	Chemical	X				
Neoplastic_Process	X	Health_Care_Activity	X	Body_Location_or_Re...	X	Qualitative_Concept	X						
Injury_or_Poisoning	X	Population_Group	X	Geographic_Area	X	Manufactured_Object	X	Mental_Process	X				
Group	X	Daily_or_Recreational...	X	Therapeutic_or_Preve...	X	Research_Activity	X	Cell	X				
Pathologic_Function	X	Mammal	X	Quantitative_Concept	X	Spatial_Concept	X	Pharmacologic_Subst...	X				
Diagnostic_Procedure	X	Eukaryote	X	Cell_Component	X	Prokaryote	X	Molecular_Biology_R...	X				
Substance	X	Mental_or_Behavioral...	X	Molecular_Function	X	Fungus	X	Virus	X				
Laboratory_Procedure	X	Nucleotide_Sequence	X	Body_Substance	X	Plant	X	Amino_Acid,_Peptide...	X				
Genetic_Function	X	Nucleic_Acid_Nucle...	X	Biomedical_or_Denta...	X	Gene_or_Genome	X						
Sign_or_Symptom	X	HP	X	GO	X	HP	X	GENE	X				

The patient was prescribed 1 capsule of Advil for 5 days. He was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely fro 3 months.

Color codes:FREQUENCY, DOSAGE, DURATION, DRUG, FORM, STRENGTH, **Posology NER**

No findings in urinary system , skin color is normal , brain CT and cranial checks are clear . Swollen fingers and eyes . Extensive stage small cell lung cancer . Chemotherapy with carboplatin and etoposide . Left scapular pain status post CT scan of the thorax .

Color codes:Organ, Organism_subdivision, Organism_substance, PathologicalFormation, Anatomical_system, **Anatomy NER**

A . Record date : 2093-01-13 , David Hale, M.D . , Name : Hendrickson , Ora MR . # 7194334 Date : 01/13/93 PCP : Oliveira , 25 years-old , Record date : 2079-11-09 . Cocke County Baptist Hospital . 0295 Keats Street

Color codes:STREET, DOCTOR, AGE, HOSPITAL, PATIENT, DATE, MEDICALRECORD, **PHI NER**

Clinical Named Entity Recognition

- ❖ ner_anatomy
- ❖ ner_bionlp
- ❖ ner_cellular
- ❖ ner_clinical
- ❖ ner_deid
- ❖ ner_diseases
- ❖ ner_drugs
- ❖ ner_events
- ❖ ner_jsl
- ❖ ner_medmentions
- ❖ ner_posology
- ❖ ner_risk_factors
- ❖ ner_human_phenotype_go
- ❖ ner_human_phenotype_gene
- ❖ ner_chemprot
- ❖ ner_ade
- ❖ ner_ade_healthcare

The patient was prescribed 1 capsule of Advil for 5 days. He was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely fro 3 months.

Color codes:FREQUENCY, DOSAGE, DURATION, DRUG, FORM, STRENGTH, **Posology NER**

No findings in urinary system , skin color is normal , brain CT and cranial checks are clear . Swollen fingers and eyes . Extensive stage small cell lung cancer . Chemotherapy with carboplatin and etoposide . Left scapular pain status post CT scan of the thorax .

Color codes:Organ, Organism_subdivision, Organism_substance, PathologicalFormation, Anatomical_system, **Anatomy NER**

A . Record date : 2093-01-13 , David Hale, M.D . , Name : Hendrickson , Ora MR . # 7194334 Date : 01/13/93 PCP : Oliveira , 25 years-old , Record date : 2079-11-09 . Cocke County Baptist Hospital . 0295 Keats Street

Color codes:STREET, DOCTOR, AGE, HOSPITAL, PATIENT, DATE, MEDICALRECORD, **PHI NER**

NER in Healthcare

- ❖ ner_anatomy
- ❖ ner_bionlp
- ❖ ner_cellular
- ❖ ner_clinical
- ❖ ner_deid
- ❖ ner_diseases
- ❖ ner_drugs
- ❖ ner_events
- ❖ ner_jsl
- ❖ ner_medmentions
- ❖ ner_posology
- ❖ ner_risk_factors
- ❖ ner_human_phenotype_go
- ❖ ner_human_phenotype_gene
- ❖ ner_chemprot
- ❖ ner_ade
- ❖ ner_ade_healthcare

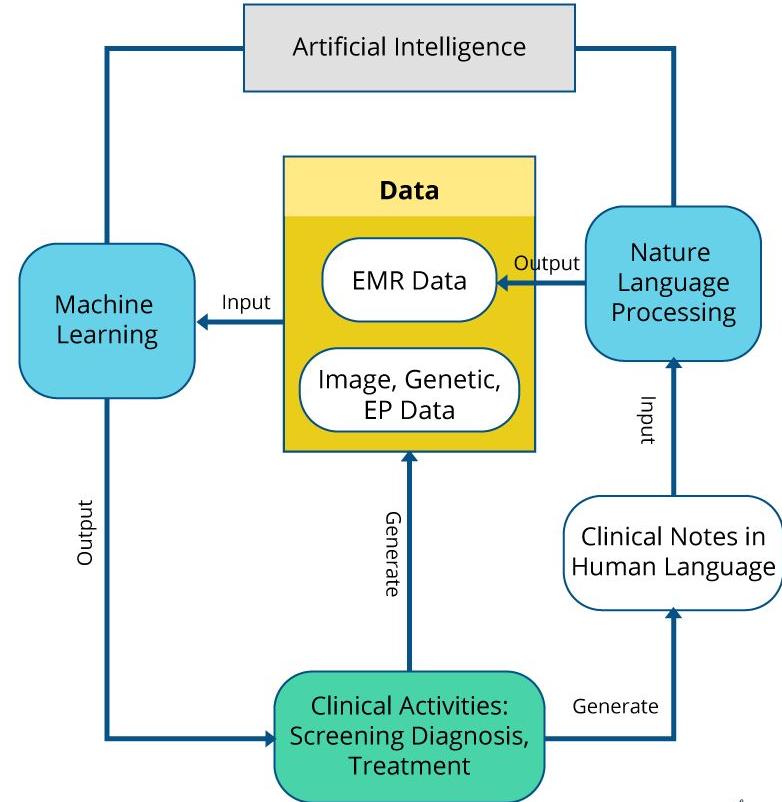




Fig. 3 | Making predictions using EHRs. **a**, Unstructured EHR data. Medical records are stored in idiosyncratic data structures and formats such that models built on a given hospital's record do not necessarily work with data from a different hospital. **b**, Data standardization. By mapping data from multiple sites to a single format based on FHIR, data are standardized into a homogeneous format. **c**, Sequencing. By temporally sequencing all data into a patient timeline, time-based deep-learning techniques can be applied on the entirety of EHR datasets for making predictions about single patients.

NER in Healthcare

Case: Predicting if a patient would develop a metastasis on certain sites.



Annotate your own data and train a custom NER model

Recently diagnosed, stage 4 adenocarcinoma of both lungs with metastasis to bone.
CT scan shows no indication of mets on brain.

Extract named entities with Spark NLP *NERDL* and assign assertion statuses with *AssertionDL* model

Feature extraction & engineering

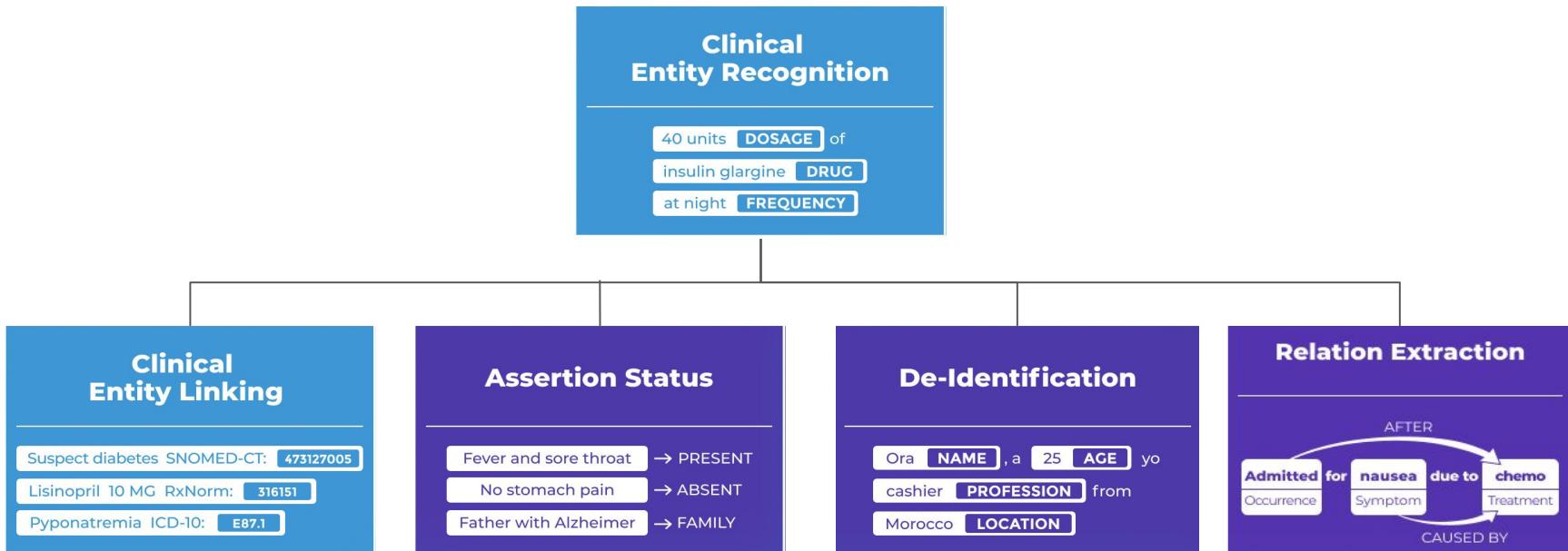
Prediction: Bone metastasis on June 2018



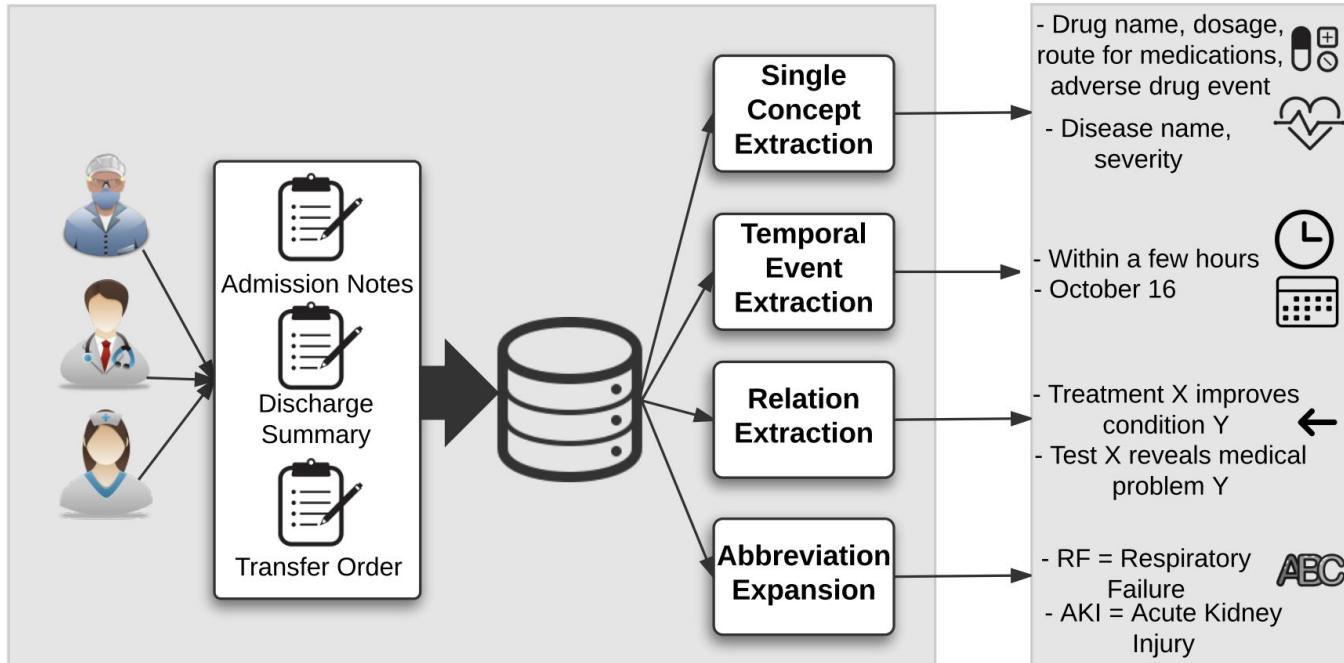
Text embeddings thru clinical word embeddings

- The frequency of clinical visits using document dates
- The number of positive site (organ) entities (hits by window)
- The number of Radiology/Oncology/Pathology reports in the last x days
- The number of tests applied in the last x days
- The number of diseases detected in the last x days
- Family history and social health determinants, etc.
- Date extraction and normalization

NER in Healthcare



NER in Healthcare



Care
Providers

Clinical
Notes

EHR

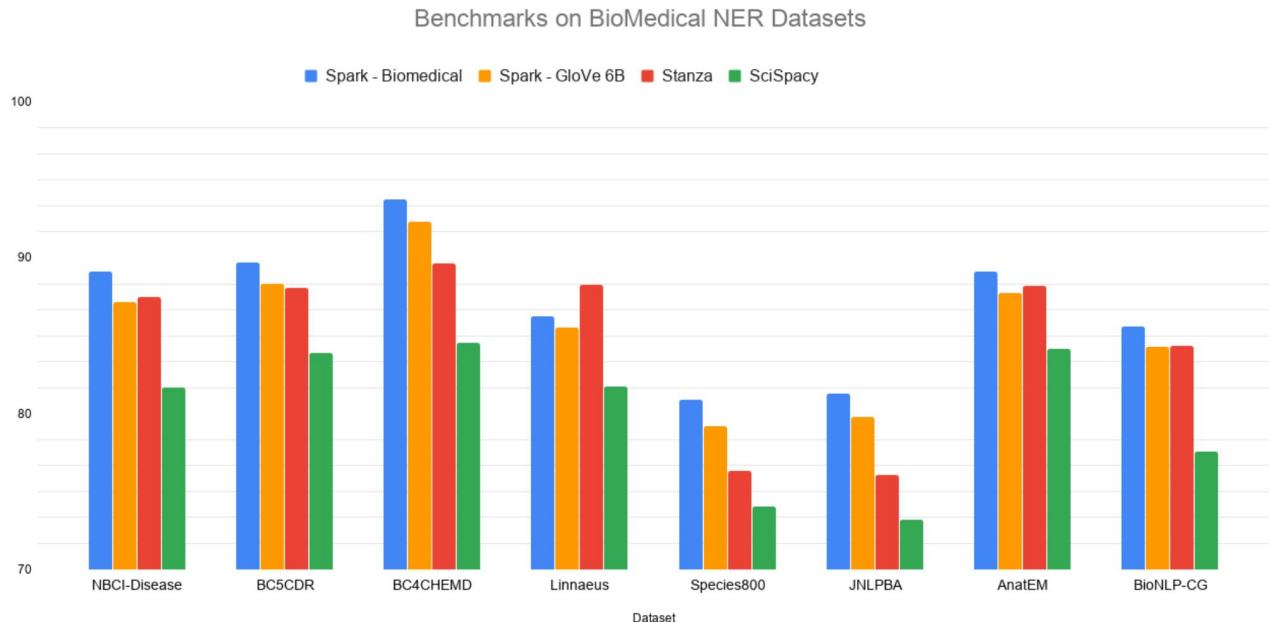
Information
Extraction

Example
Tasks

Spark NLP NerDL

The best NER
score in
production

Dataset	Entities	Spark - Biomedical	Spark - GloVe 6B	Stanza	SciSpacy
NBCI-Disease	Disease	89.13	87.19	87.49	81.65
BC5CDR	Chemical, Disease	89.73	88.32	88.08	83.92
BC4CHEMD	Chemical	93.72	92.32	89.65	84.55
Linnaeus	Species	86.26	85.51	88.27	81.74
Species800	Species	80.91	79.22	76.35	74.06
JNLPBA	5 types in cellular	81.29	79.78	76.09	73.21
AnatEM	Anatomy	89.13	87.74	88.18	84.14
BioNLP13-CG	16 types in Cancer Genetics	85.58	84.3	84.34	77.6



NER Comparison with AWS Medical Comprehend

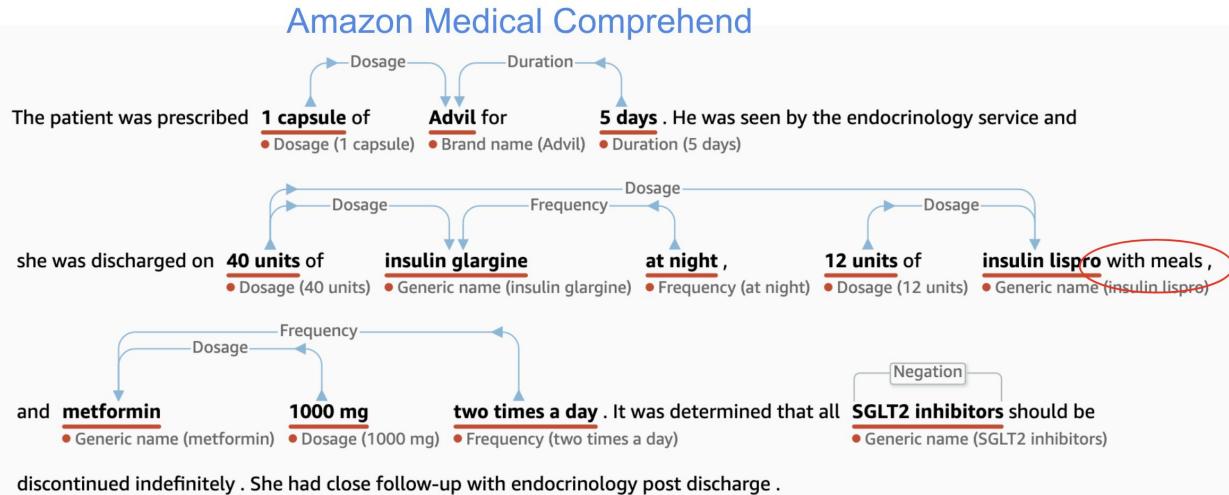
up to 20% better

AWS Medical Comprehend

	precision	recall	f1-score	support
0	0.7751	0.8951	0.8308	56048
PROBLEM	0.4882	0.4899	0.4890	4054
TEST	0.5466	0.7170	0.6203	5032
DRUG	0.8151	0.9102	0.8601	10284

Spark NLP

	precision	recall	f1-score	support
0	0.9116	0.9431	0.9271	56048
PROBLEM	0.7261	0.6626	0.6929	4054
TEST	0.7824	0.7331	0.7570	5032
DRUG	0.9463	0.8820	0.9130	10284



Spark NLP Posology NER

The patient was prescribed **1 capsule of Advil for 5 days**. He was seen by the endocrinology service and she was discharged on **40 units of insulin glargine at night, 12 units of insulin lispro with meals**, and **metformin 1000 mg two times a day**. It was determined that all **SGLT2 inhibitors** should be discontinued indefinitely. She had close follow-up with endocrinology post discharge.

Color codes: DURATION, FREQUENCY, STRENGTH, DRUG, DOSAGE, FORM,

* Tested on 1000 sample text from MIMIC-III

NER-DL in Spark NLP

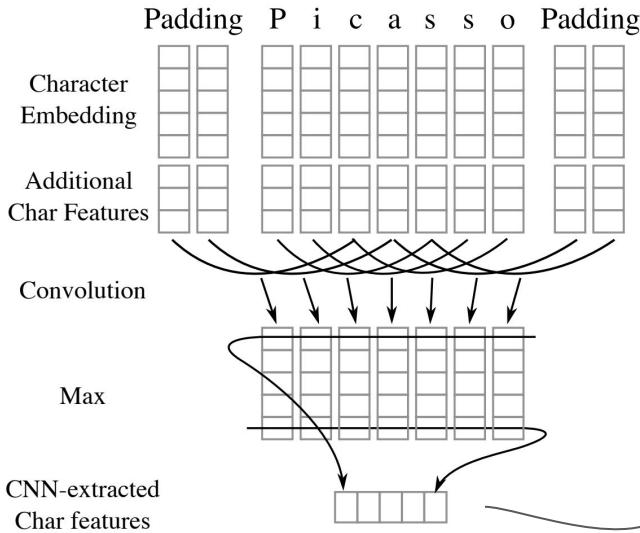


Figure 2: The convolutional neural network extracts character features from each word. The character embedding and (optionally) the character type feature vector are computed through lookup tables. Then, they are concatenated and passed into the CNN.

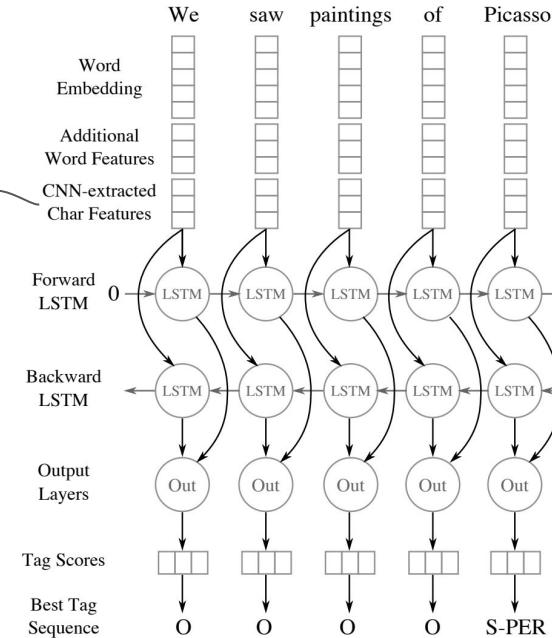


Figure 1: The (unrolled) BLSTM for tagging named entities. Multiple tables look up word-level feature vectors. The CNN (Figure 2) extracts a fixed length feature vector from character-level features. For each word, these vectors are concatenated and fed to the BLSTM network and then to the output layers (Figure 3).

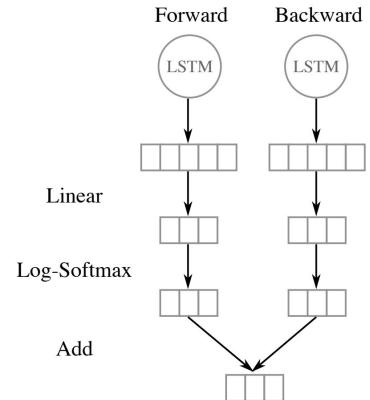


Figure 3: The output layers (“Out” in Figure 1) decode output into a score for each tag category.

Char-CNN-BiLSTM

NER-DL in Spark NLP

Char-CNN-BiLSTM

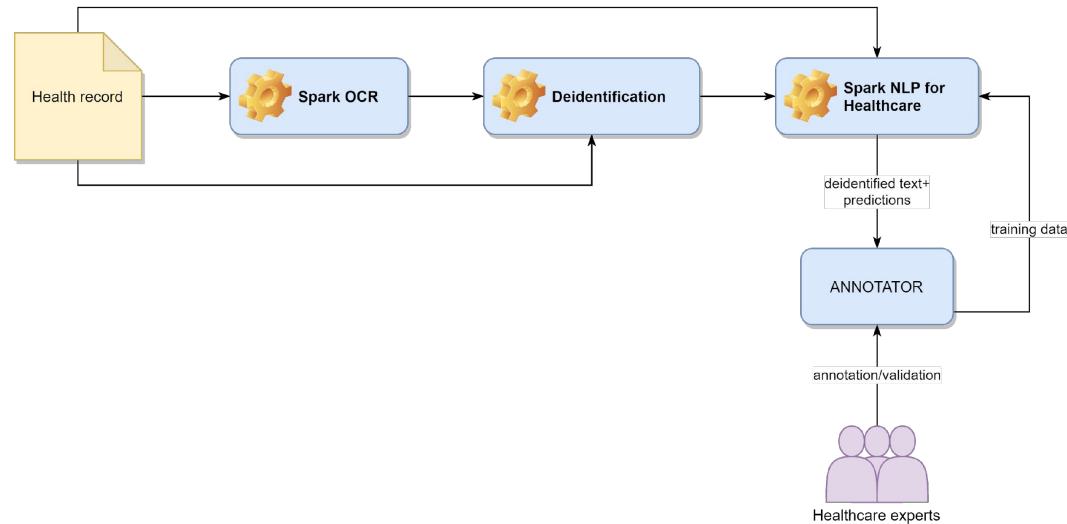
	F1 : Tokens	F2 : Casing	F3 : POS	F4 : Char CNN	Labels
The					O
company					O
XYZ					Company
Private					Company
Limited					Company
works					O
in					O
the					O
health					Activity
sector					Activity
in					O
Europe					Location

BIO schema

John	B-PER
Smith	I-PER
lives	0
in	0
New	B-LOC
York	I-LOC

John Smith ⇒ PERSON
New York ⇒ LOCATION

NER in Healthcare



- Close coordination with annotators
- Inter annotator agreements
- CoNLL preparation (tokenization and schema)

She returns today for ongoing evaluation of her EGFR mutated, stage 4 lung cancer with metastasis to her L2 vertebrae and her lungs bilaterally.

Bone negative for metastatic disease.

Patient denies any family history of cancer.

Part - II

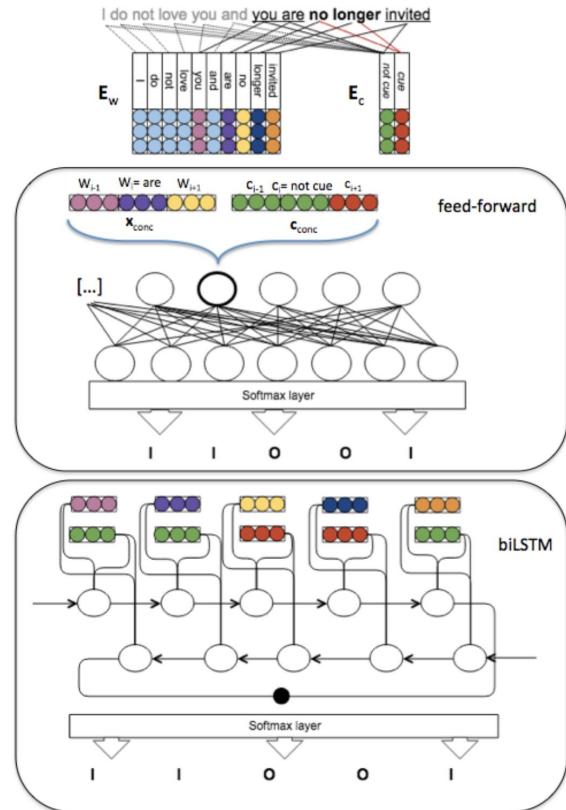
- ❖ Assertion status detection

Clinical Assertion Model

Prescribing sick days due to diagnosis of influenza .	<i>Present</i>
41 yo man with CRFs of DM Type II, high cholesterol, smoking history, family hx, HTN p/w episodes of atypical CP x 1 week , with rest and exertion.	<i>Conditional</i>
Jane's RIDT came back clean.	<i>Absent</i>
Jane is at risk for flu if she's not vaccinated.	<i>Hypothetical</i>
There was a dense hemianopsia on the left side.	<i>Present</i>

F-Score	Dataset	Task
94.17%	4 th i2b2/VA	Disease & problem norm.

"Neural Networks For Negation Scope Detection", Fancellu et al., In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.



scope of negation: given a negative instance, to identify which tokens are affected by negation

Clinical Assertion Model

Patient with **severe fever** and **sore throat**. He shows no **stomach pain** and he maintained on **an epidural** and **PCA** for pain control . He also became **short of breath** with climbing a flight of stairs . After **CT** , lung tumor located at the right lower lobe . Father with **Alzheimer**.

Color codes:**PROBLEM**, **TREATMENT**, **TEST**,

Entities

	chunks	entities	assertion
0	severe fever	PROBLEM	present
1	sore throat	PROBLEM	present
2	stomach pain	PROBLEM	absent
3	an epidural	TREATMENT	present
4	PCA	TREATMENT	present
5	pain control	PROBLEM	present
6	short of breath	PROBLEM	conditional
7	CT	TEST	present
8	lung tumor	PROBLEM	present
9	Alzheimer	PROBLEM	associated_with_someone_else

```
● ● ●  
import sparknlp_jsl  
  
spark = sparknlp_jsl.start("xxxx")  
  
from pyspark.ml import PipelineModel  
  
pretrained_model = PipelineModel.load("explain_clinical_doc_dl")  
  
from sparknlp.base import LightPipeline  
  
ner_lightModel = LightPipeline(pretrained_model)  
  
clinical_text = """  
Patient with severe fever and sore throat.  
He shows no stomach pain and he maintained on an epidural and PCA for pain control.  
He also became short of breath with climbing a flight of stairs.  
After CT, lung tumour located at the right lower lobe. Father with Alzheimer.  
"""  
  
result = ner_lightModel.fullAnnotate(clinical_text)  
  
entity_tuples = [(n.result, n.metadata['entity'], m.result, n.begin, n.end)  
                 for n,m in zip(result[0]['ner_chunk'],result[0]['assertion'])]  
  
print(entity_tuples)  
=>  
time: 270 ms  
[('severe fever', 'PROBLEM', 'present', 14, 25),  
 ('sore throat', 'PROBLEM', 'present', 31, 41),  
 ('stomach pain', 'PROBLEM', 'absent', 57, 68),  
 ('an epidural', 'TREATMENT', 'present', 91, 101),  
 ('PCA', 'TREATMENT', 'present', 107, 109),  
 ('pain control', 'PROBLEM', 'present', 115, 126),  
 ('short of breath', 'PROBLEM', 'conditional', 144, 158),  
 ('CT', 'TEST', 'present', 200, 201),  
 ('lung tumour', 'PROBLEM', 'present', 204, 214),  
 ('Alzheimer', 'PROBLEM', 'associated_with_someone_else', 261, 269)]
```

Part - III

- ❖ De-Identification and Obfuscation of PHI data

De-Identification

- * Identifies potential pieces of content with personal information about patients and remove them by replacing with semantic tags.

```
A . Record date : 2093-01-13 , David Hale , M.D . , Name : Hendrickson , Ora MR . # 7194334  
Date : 01/13/93 PCP : Oliveira , 25 month years-old , Record date : 2079-11-09 . Cocke  
County Baptist Hospital . 0295 Keats Street
```

Color codes: DOCTOR, HOSPITAL, DATE, STREET, MEDICALRECORD, PATIENT,

Deidentified Text

```
['A .',  
 'Record date : <DATE> , <DOCTOR> , M.D .',  
 ', Name : <PATIENT> , <PATIENT> MR .',  
 '# <MEDICALRECORD> Date : <DATE> PCP : <DOCTOR> , 25  
month years-old , Record date : <DATE> .',  
 '<HOSPITAL> .',  
'<STREET>']
```

```
def get_deidentify_model():  
  
    custom_ner_converter = NerConverter()\  
        .setInputCols(["sentence", "token", "ner"])\\  
        .setOutputCol("ner_chunk")  
        #.setWhiteList(entity_types)  
  
    deidentify_pipeline = Pipeline(  
        stages = [  
            documentAssembler,  
            sentenceDetector,  
            tokenizer,  
            word_embeddings,  
            clinical_ner,  
            custom_ner_converter,  
            deidentification_rules  
        ])  
  
    empty_data = spark.createDataFrame([[""]]).toDF("text")  
  
    model_deidentify = deidentify_pipeline.fit(empty_data)  
  
    return model_deidentify
```

Part - V

- ❖ Entity Resolution (ICD1-, RxNorm, Snomed)
- ❖ GenericClassifier with TF

Entity Resolution

Tobramycin (D014031)

Gentamicins (D005839)

We observed patients treated with gentamicin sulfate or tobramycin sulfate for the development of aminoglycoside-related renal failure. Gentamicin sulfate decreased renal function more frequently than tobramycin sulfate.

Aminoglycosides (D000617)

Renal Insufficiency (D051437)

"CNN-based ranking for biomedical entity normalization".

Li et al., *BMC Bioinformatics*, October 2017.

F-Score	Dataset	Task
90.30%	ShARe / CLEF	Disease & problem norm.
92.29%	NCBI	Disease norm. in literature

codes	description
17473003	Cecotomy
17473003	Cecotomy (procedure)
304587000	Excision of colonic pouch
304587000	Excision of colonic pouch (procedure)
87279008	Excision of lesion of colon
174117007	Excision of lesion of colon NEC
174117007	Excision of lesion of colon NEC (procedure)
87279008	Excision of lesion of colon (procedure)
276190007	Ileocolic resection
276190007	Ileocolic resection (procedure)
43075005	Partial resection of colon
43075005	Partial resection of colon (procedure)
428305005	History of partial resection of colon (situation)
428305005	History of partial resection of colon
444165004	Partial resection of colon and resection of terminal
738552004	Partial resection of colon with stoma (procedure)
738552004	Partial resection of colon with stoma
84952009	Resection of colon for interposition
84952009	Resection of colon for interposition (procedure)
445884009	Wedge resection of colon

only showing top 20 rows

Assigns a **ICD10** (International Classification of Diseases version 10) code to chunks identified as "PROBLEMS" by the NER Clinical Model

Entity Resolution - RxNorm

the patient was prescribed 1 capsule DRUG of advil DRUG for 5 days DURATION . he was seen by the endocrinology service and she was discharged on 40 units DRUG of insulin glargine DRUG at night FREQUENCY , 12 units DRUG of insulin lispro DRUG with meals FREQUENCY , and metformin 1000 mg DRUG two times a day FREQUENCY . it was determined that all sglt2 inhibitors DRUG should be discontinued indefinitely .

advil : DRUG

	rxnorm_code	description	distance
0	153010	advil	0
1	669348	advate	0.0417

insulin lispro : DRUG

	rxnorm_code	description	distance
0	86009	insulin lispro	0
1	1157461	insulin lispro injectable product	0.0743

insulin glargine : DRUG

	rxnorm_code	description	distance
0	274783	insulin glargine	0.0000
1	1157459	insulin glargine injectable product	0.0653

metformin 1000 mg : DRUG

	rxnorm_code	description	distance
0	316255	metformin 1000 mg	0.0000
1	860995	metformin hydrochloride 1000 mg	0.0445

Entity Resolution - Snomed / ICD-10

a 28-year-old female with a history of gestational diabetes mellitus PROBLEM diagnosed eight years prior to presentation and subsequent type two diabetes mellitus PROBLEM (t2dm PROBLEM), one prior episode of htg-induced pancreatitis PROBLEM three years prior to presentation , associated with an acute hepatitis PROBLEM , and obesity PROBLEM with a body mass index PROBLEM (bmi) of 33.5 kg/m² , presented with a one-week history of polyuria PROBLEM , polydipsia PROBLEM , poor appetite PROBLEM , and vomiting PROBLEM .

gestational diabetes mellitus : PROBLEM

	snomed_code	description	distance	athena_concept_id	domain_id	concept_class_id	ICD10CM_mapping
0	11687002	gestational diabetes mellitus	0.0000	4024659	Condition	Clinical Finding	024.429, 024.439, 024.414, 024.419, 024.4, 024.410
1	40791000119105	postpartum gestational diabetes mellitus	0.0423	45757789	Condition	Clinical Finding	024.4, 024.439

obesity : PROBLEM

	snomed_code	description	distance	athena_concept_id	domain_id	concept_class_id	ICD10CM_mapping
0	414916001	obesity	0.0000	433736	Condition	Clinical Finding	E66.9
1	414915002	obese	0.0264	4215968	Observation	Clinical Finding	Z68.41, E66.9, E66.8

Part - VI

❖ Relation Extraction

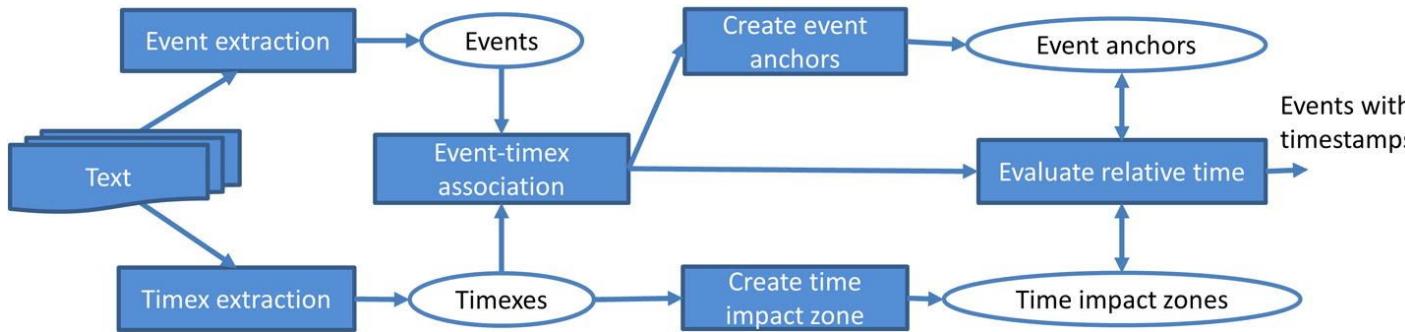
RelationExtractionModel	re_clinical	2.5.5
RelationExtractionModel	re_posology	2.5.5
RelationExtractionModel	re_temporal_events_clinical	2.6.0
RelationExtractionModel	re_temporal_events_enriched_clinical	2.6.0
RelationExtractionModel	re_human_phenotype_gene_clinical	2.6.0
RelationExtractionModel	re_drug_drug_interaction_clinical	2.6.0
RelationExtractionModel	re_chemprot_clinical	2.6.0

Relation Extraction



On 5/21/99, the infant received her 1st dose of vaccine A and her 2nd injection of vaccine B. The infant began vomiting and having diarrhea 5 days later. She was taken to the local ER where evaluation was ""non-diagnostic"" ...

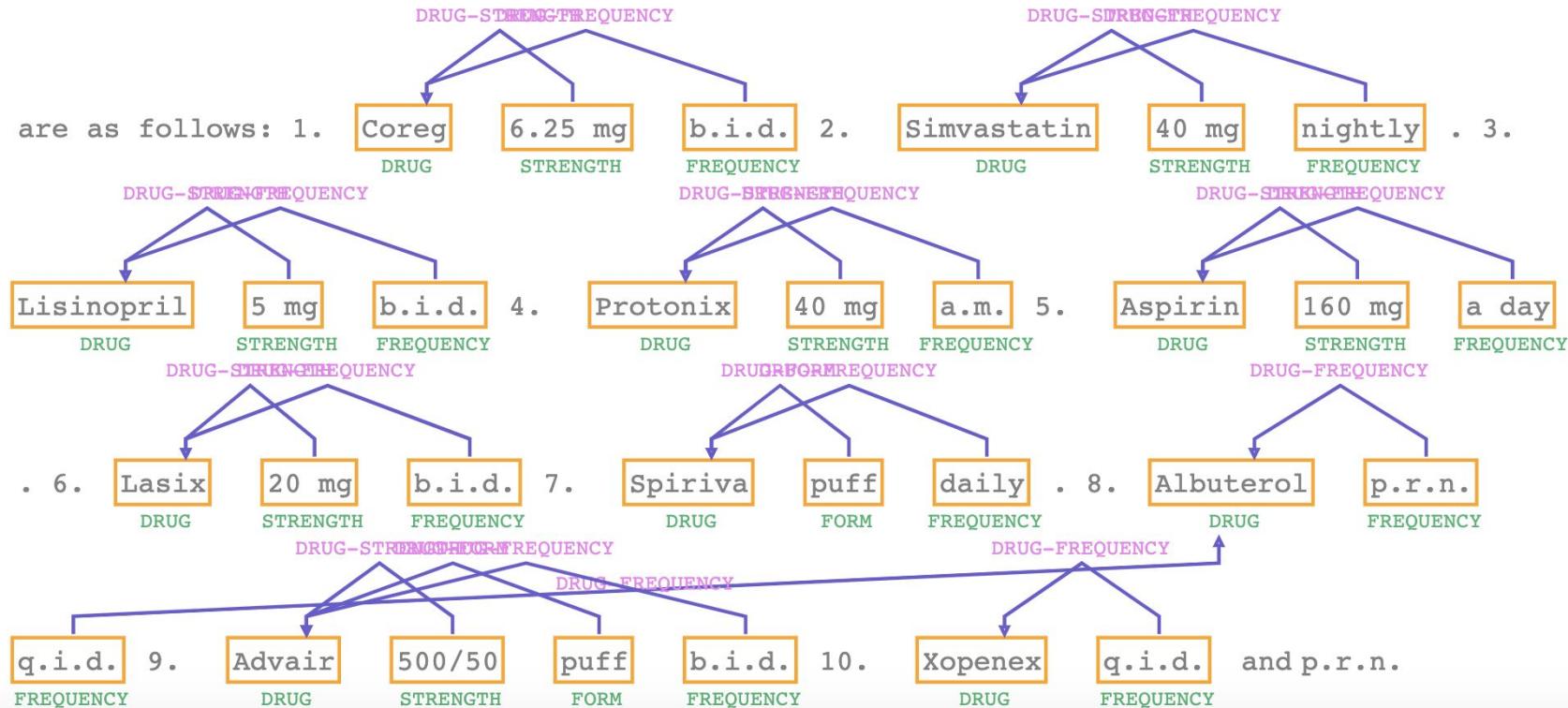
Feature	Type	Date
Vaccine A	Vaccine	1999-05-21
Vaccine B	Vaccine	1999-05-21
Vomiting	Symptom	1999-05-26
Diarrhea	Symptom	1999-05-26
...



Relation Extraction

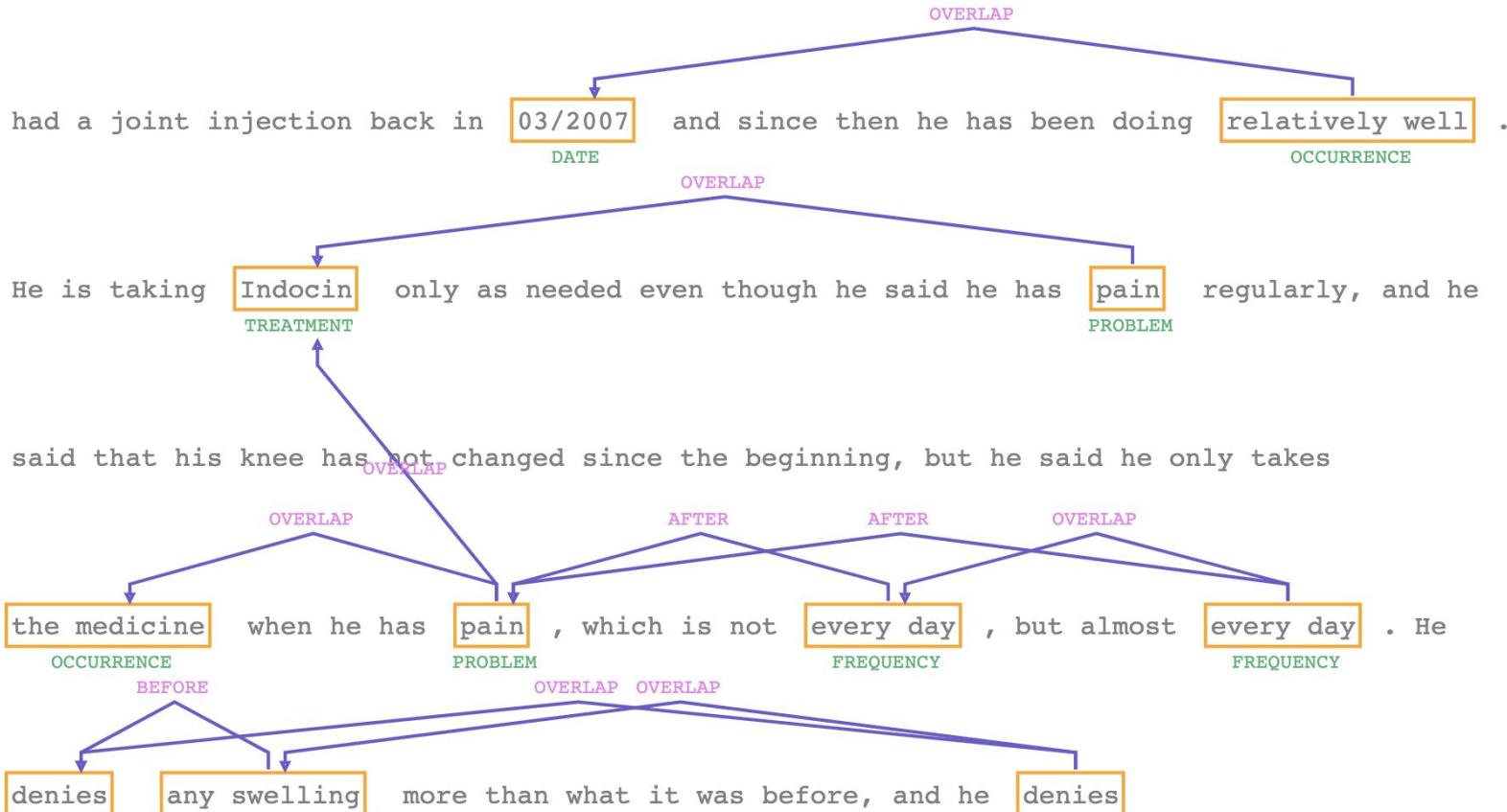
Posology

are as follows: 1. Coreg 6.25 mg b.i.d. 2. Simvastatin 40 mg nightly . 3.



Relation Extraction

Temporal Events



Part - VI

- ❖ Spark OCR



Maximizing Text Recognition Accuracy with Image Transformers in Spark OCR



Spark OCR

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléna) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Uploaded Image.

Converted Text:

; a 'Sur. la base de la grande statue de Zeus, a 'Olympie, Phidias avait

Présenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléna) Aes 'douze divinités, groupées deux a deux, Ss ordonnaient en six couples :



Converted Text:

Digital 'Image Processing

After being crowned Miss America, she endured criticism from some Blacks that she was "not Black enough," and insults from Whites who were not happy to see a Black woman wear the prized symbol of all-American beauty. And then she set about building a show-business career while hampered by controversy and the stigma of being a beauty queen.

Uploaded Image.

Converted Text:

After being crowned Miss America, she endured criticism from some Blacks that she was "not Black enough," and insults from Whites who were not happy to see a Black woman wear the prized symbol of all-American beauty. And then she set about building a show-business career while hampered by controversy and the stigma of being a beauty queen,

Content

10 mins	Introduction to Spark OCR
5 mins	Spark OCR pipeline examples
20 mins	Image preprocessing
10 mins	Image preprocessing with Spark OCR in action
10 mins	Q&A

Spark OCR

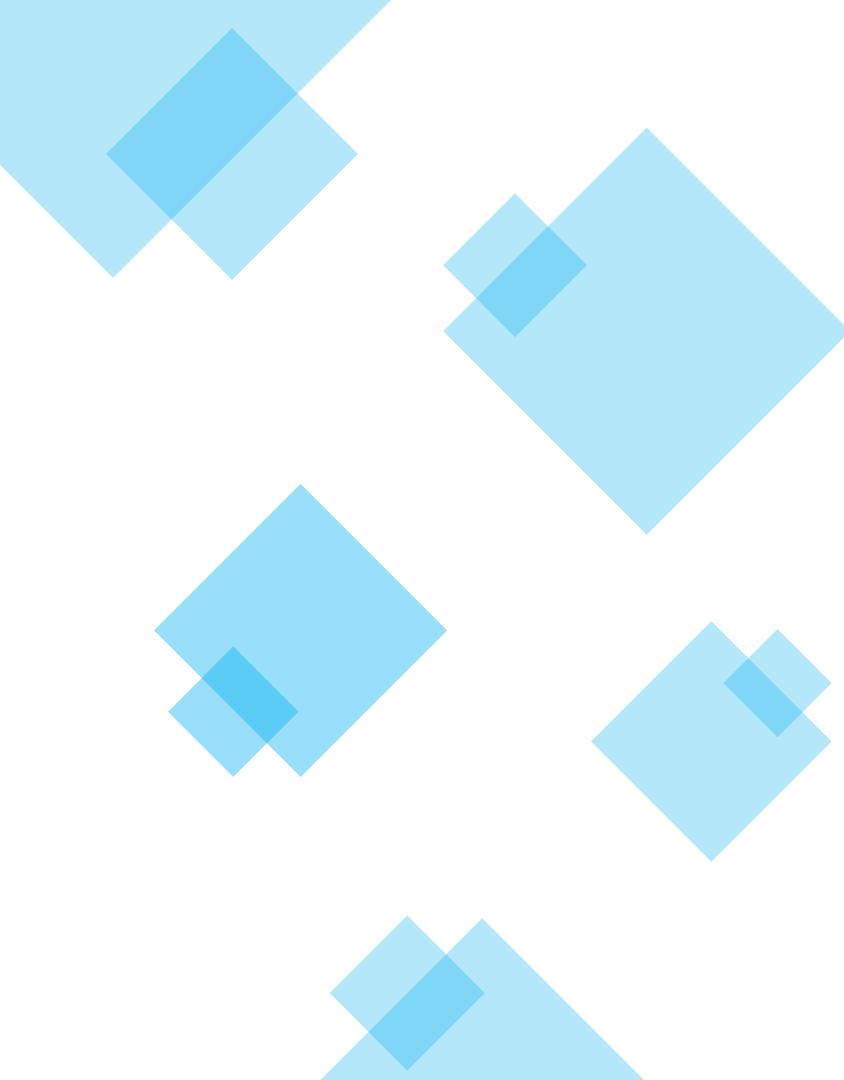
- Based on Spark ML.
- Offers various transformers for:
 - OCR
 - PDF processing
 - Image processing
 - Layout analysis
 - Some other utility transformers

Spark OCR Benefits

- Scalability guaranteed by Spark.
- Running in isolated environment for dealing with personal data.
- Out of the box support for different input/output formats (image, pdf, dicom)
- Existing opensource solutions provide low quality results.
- Compatibility with Spark NLP.

Common use cases

- Digitize scanned PDF's and images.
- De-identify PDF and image documents.
- Extract text from PDF, process it using Spark NLP and render back or highlight text.
- Process images using Spark.



Spark OCR Transformers

Transformers

ImageToText - OCR

Transformers for Layout analysis:

- ImageLayoutAnalyzer - Detect regions on page.
- ImageSplitRegions - Split image to regions.

Transformers for deal with text positions:

- PositionFinder
- UpdateTextPosition

PDF transformers

Transformers for dealing with PDF files:

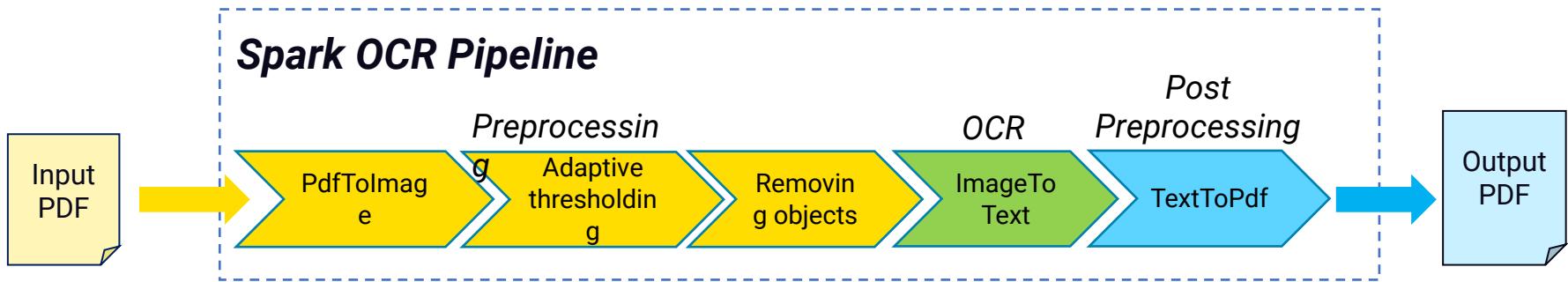
- **PdfToText** – extract text from selectable PDF
- **PdfToImage** – render each page as image
- **ImageToPdf** – store image to PDF format
- **TextToPdf** – render text with positions to PDF format
- **PdfDrawRegions** – draw regions to existing PDF

Image processing transformers

- [BinaryToImage](#) - Convert binary data to image
- [ImageBinarizer](#) – Image binarization by custom threshold.
- [ImageAdaptiveThresholding](#) – Image binarization using local thresholding.
Supports *Gaussian*, *mean*, *median* methods.
- [ImageScaler](#) – Scale image by custom scale factor.
- [ImageAdaptiveScaler](#) – Detects font size and scales image to have a desired font size.
- [ImageSkewCorrector](#) - Autocorrect skew for images with text.
- [ImageNoiseScorer](#) – Compute noise score for images.
- [ImageRemoveObjects](#) – Remove small/big objects from image.
- [ImageMorphologyOpening](#) – Apply morphology opening to image.
- [ImageDrawRegions](#) - Draw regions to image.

Spark OCR Pipelines

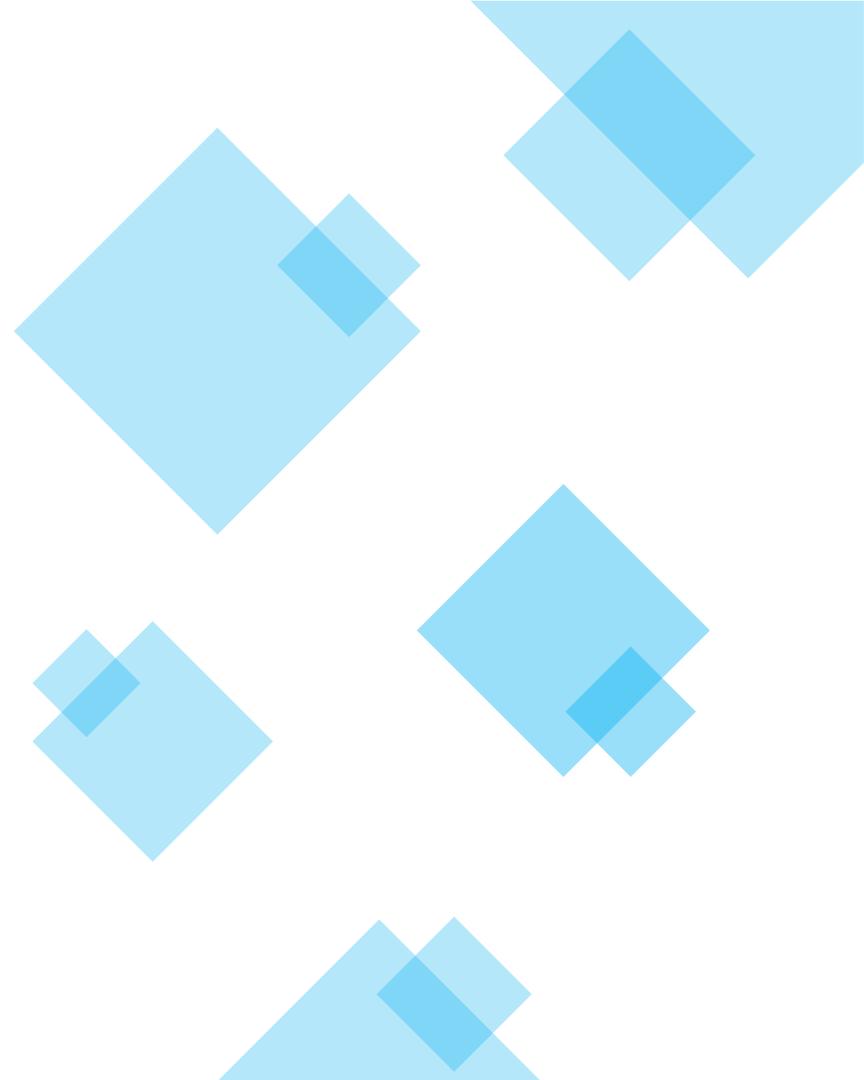
OCR workflow



Some example of OCR pipelines

- Processing images
- Processing PDF's
- Image de-identification

Image Preprocessing



Why do we need image preprocessing?

Background noise

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléne) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Bad quality of image

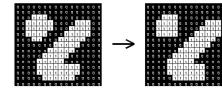
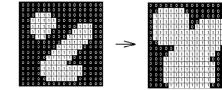
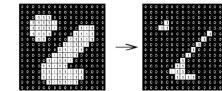
Patient is an 84-year-old male with a past medical history of hypertension, HFpEF last known EF 55%, mild to moderate TR, pulmonary hypertension, permanent atrial fibrillation on Eliquis, history of GI bleed, CK-MB, and anemia who presents with full weeks of generalized fatigue and feeling unwell. He also notes some shortness of breath and

Background in nature scene



Image preprocessing operations in Spark OCR

- Adaptive thresholding
- Morphological operations
 - Erosion
 - Dilation
 - Opening
 - Closing
- Removing objects
- Skew correction



Thresholding

Thresholding is the simplest way to segment objects from a background.

If that background is relatively ***uniform***, then you can ***use a global threshold*** value to binarize the image by pixel-intensity.

If there's ***large variation*** in the background intensity, however, ***adaptive thresholding*** may produce better results. We binarize an image using the `threshold_adaptive` function, which calculates thresholds in regions of size `block_size` surrounding each pixel (i.e. local neighborhoods). Each threshold value is the weighted mean of the local neighborhood minus an offset value.

Image

Region-based segmentation

Let us first determine markers of the coins and the background. These markers are pixels that we can label unambiguously as either object or background. Here, the markers are found at the two extreme parts of the histogram of grey values:

```
>>> markers = np.zeros_like(coins)
```

Global thresholding

Region-based segmentation

determine markers of the coins and the
These markers are pixels that we can label
as either object or background. Here,
the markers are found at the two extreme parts of the
histogram of grey values:

```
>>> markers = np.zeros_like(coins)
```

Adaptive thresholding

Region-based segmentation

Let us first determine markers of the coins and the background. These markers are pixels that we can label unambiguously as either object or background. Here, the markers are found at the two extreme parts of the histogram of grey values:

```
>>> markers = np.zeros_like(coins)
```

Thresholding with Spark OCR

Source image

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait pro-

Global thresholding (*threshold = 80*)



Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait pro-

For various images usually provides better results

Global thresholding (*threshold = 120*)



Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait pro-

```
gbinarizer = ImageBinarizer()
gbinarizer.setInputCol("scaled_image")
gbinarizer.setOutputCol("binarized_image")
gbinarizer.setThreshold(80)
```

Adaptive thresholding



Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait pro-

```
binarizer = ImageAdaptiveThresholding()
binarizer.setInputCol("scaled_image")
binarizer.setOutputCol("binarized_image")
binarizer.setBlockSize(91)
binarizer.setOffset(60)
```

ImageAdaptiveThresholding transformer

Patient is an 84-year-old male with a past medical history of hypertension, HFpEF last known EF 55%, mild to moderate TR, pulmonary hypertension, permanent atrial fibrillation on Eliquis, history of GI bleed, CK-MB, and anemia who presents with full weeks of generalized fatigue and feeling unwell. He also notes some shortness of breath and

```
adaptive_thresholding = ImageAdaptiveThresholding() \  
    .setInputCol("image") \  
    .setOutputCol("corrected_image") \  
    .setBlockSize(35) \  
    .setOffset(80)
```



Patient is an 84-year-old male with a past medical history of hypertension, HFpEF last known EF 55%, mild to moderate TR, pulmonary hypertension, permanent atrial fibrillation on Eliquis, history of GI bleed, CK-MB, and anemia who presents with full weeks of generalized fatigue and feeling unwell. He also notes some shortness of breath and

ImageAdaptiveThresholding transformer

Param name	Type	Default	Description
blockSize	int	170	Odd size of pixel neighborhood which is used to calculate the threshold value (e.g. 3, 5, 7, ..., 21, ...).
method	AdaptiveThresholdingMethod	GAUSSIAN	Method used to determine adaptive threshold for local neighborhood in weighted mean image.
offset	int	0	Constant subtracted from weighted mean of neighborhood to calculate the local threshold value.

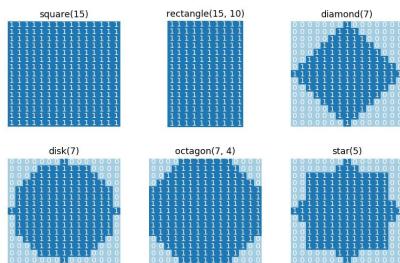
Methods:

- GAUSSIAN
- MEAN
- MEDIAN

Morphological Operations

Morphological operators often take a binary image and a structuring element (kernel) as input and combine them using a set operator (intersection, union, inclusion, complement). They process objects in the input image based on characteristics of its shape, which are encoded in the structuring element. The mathematical details are explained in Mathematical Morphology.

The structuring element used in practice is generally much smaller than the image, often a 3x3 matrix.



Erosion

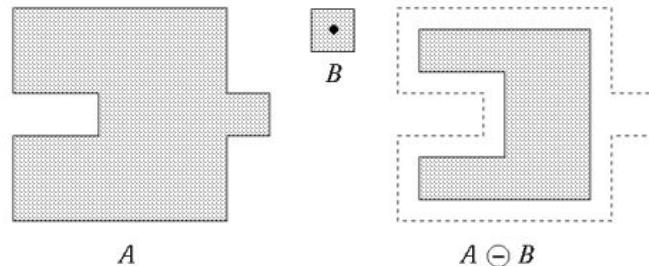
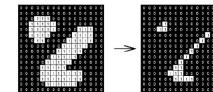
The erosion operation uses a structuring element for reducing the shapes contained in the input image.

Let E be a Euclidean space or an integer grid, and A a binary image in E . The **erosion** of the binary image A by the structuring element B is defined by:

$$A \ominus B = \{z \in E | B_z \subseteq A\}$$

where B_z is the translation of B by the vector z , i.e.,

$$B_z = \{b + z | b \in B\}$$



Source image

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

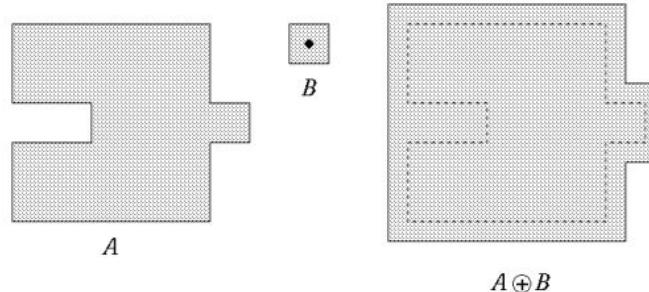
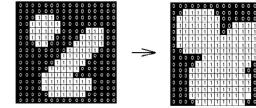
Image after applying erosion

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Dilation

The dilation operation uses a structuring element for expanding the shapes contained in the input image.

$$A \oplus B = \bigcup_{b \in B} A_b,$$



Source image

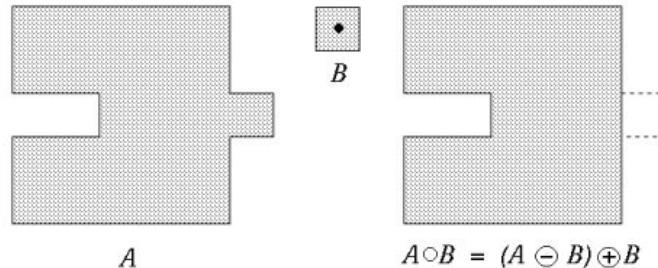
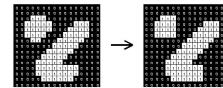
Patient is an 84-year-old male with a past medical history of hypertension, HFpEF last known EF 55%, mild to moderate TR, pulmonary hypertension, permanent atrial fibrillation on Eliquis, history of GI bleed, CK-MB, and anemia who presents with full weeks of generalized fatigue and feeling unwell. He also notes some shortness of breath and

Image after applying dilation

Patient is an 84-year-old male with a past medical history of hypertension, HFpEF last known EF 55%, mild to moderate TR, pulmonary hypertension, permanent atrial fibrillation on Eliquis, history of GI bleed, CK-MB, and anemia who presents with full weeks of generalized fatigue and feeling unwell. He also notes some shortness of breath and

Opening

Very simply, an opening is defined as an erosion followed by a dilation *using the same structuring element for both operations.*



Source image

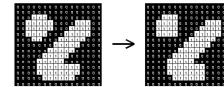
Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :



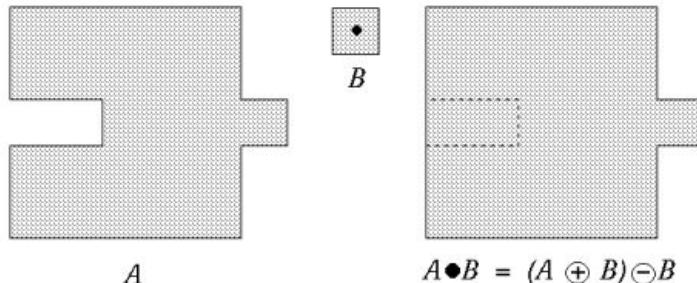
Image after applying opening

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Closing



Closing is opening performed in reverse. It is defined simply as a dilation followed by an erosion *using the same structuring element for both operations*.



Source image

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Luné (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Image after applying opening

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Luné (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

ImageMorphologyOperation transformer

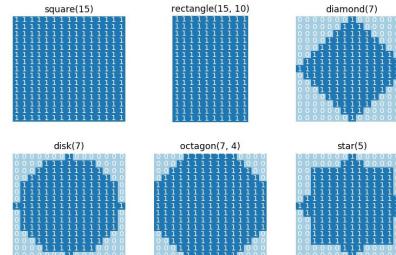
```
ImageMorphologyOperation() \  
.setKernelShape(KernelShape.SQUARE) \  
.setKernelSize(3) \  
.setOperation(MorphologyOperationType.EROSION) \  
.setInputCol("image") \  
.setOutputCol("corrected_image")
```

Kernel Shapes:

- **SQUARE**
- **DIAMOND**
- **DISK**
- **OCTAGON**
- **STAR**

Morphology Operations:

- **EROSION**
- **DILATION**
- **OPENING**
- **CLOSING**



Removing objects

- Removing small objects
- Remove small holes
- Removing small objects with specifying min font size
- Removing big objects (images, lines etc)

Source image

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Image after applying removing objects

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

ImageRemoveObject transformer

```
ImageRemoveObjects() \  
.setInputCol("binarized_image") \  
.setOutputCol("corrected_image") \  
.setMinSizeObject(50)
```

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

Param name	Type	Default	Description
minSizeFont	int	10	Min size of font in pt.
minSizeObject	int	None	Min size of object which will keep on image [*].
minSizeHole	int	None	Min size of hole which will keep on image [*].
maxSizeObject	int	None	Max size of object which will keep on image [*].

[*] : None value disables removing objects

Skew correction

```
ImageSkewCorrection() \  
.setInputCol("image") \  
.setOutputCol("corrected_image") \  
.setAutomaticSkewCorrection(True)
```

FOREWORD

Electronic design engineers are the true idea men of the electronic industries. They create ideas and use them in their designs, they stimulate ideas in other designers, and they borrow and adapt ideas from others. One could almost say they feed on and grow on ideas.

FOREWORD

Electronic design engineers are the true idea men of the electronic industries. They create ideas and use them in their designs, they stimulate ideas in other designers, and they borrow and adapt ideas from others. One could almost say they feed on and grow on ideas.

Param name	Type	Description
rotationAngle	double	rotation angle
automaticSkewCorrection	boolean	enables/disables adaptive skew correction

OCR in action

Tesseract

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaïstos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Séléné), ni mère et fils (comme Aphrodite-Éros), ni protectrice et protégé (comme Athéna-Héracles). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modèles : son initiative s'exerce dans le cadre des schèmes imposés par la tradition. Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe. On la voit rarement figurée. Quand elle l'est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermès³. De règle dans l'art plastique, l'association Hermès-Hestia

F1 score - 0.2610

on base de la Basie. statue de Zeus, a Olympie, Phidias 'avait "Das cette série de huit couples divins, il en est. un qui 'fait | pro- Hermés-Hestia. Pourquoi les apparier ?? Rien dans leur 'généalo- ati, et femme (comme Zeus-Héra, Poséidon-Amphittite, 'Héphaistos- ; ni ffére et occur (comme Apolion-Antémis, Hélios-Séléné), ni t une: déesse qui semblent étrangers r un a l'autre ?' On ne saurait ler une fantaisie personnelle du sculpteur. Quand il exécute' une t le foyer – se prêtait moins que les autres dieux grecs a la Stésentation anthropomorphe. On la voit rarement figurée: Quand elle est souvent, comme Phidias l' avait sculptée; faisant couple avec 3s°: De régle dans- l'art 'Plastique, Passociation Henmes-Hestia -_ . wt a. wo et anhserh A an om *

on base de la Basie. statue de Zeus, a Olympie, Phidias 'avait "Das cette série de huit couples divins, il en est. un qui 'fait | pro- Hermés-Hestia. Pourquoi les apparier ?? Rien dans leur 'généalo- ati, et femme (comme Zeus-Héra, Poséidon-Amphittite, 'Héphaistos- ; ni ffére et occur (comme Apolion-Antémis, Hélios-Séléné), ni t une: déesse qui semblent étrangers r un a l'autre ?' On ne saurait ler une fantaisie personnelle du sculpteur. Quand il exécute' une t le foyer – se prêtait moins que les autres dieux grecs a la Stésentation anthropomorphe. On la voit rarement figurée: Quand elle est souvent, comme Phidias l' avait sculptée; faisant couple avec 3s°: De régle dans- l'art 'Plastique, Passociation Henmes-Hestia -_ . wt a. wo et anhserh A an om *

ABYY FineReader

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait présenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaïstos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Sélénè), ni mère et fils (comme Aphrodite-Eros), ni protectrice et protégé (comme Athéna-Héraclès). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modèles : son initiative s'exerce dans le cadre des schèmes imposés par la tradition. Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe. On la voit rarement figurée. Quand elle l'est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermès³. De règle dans l'art plastique, l'association Hermès-Hestia

F1 score - 0.8972

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait présenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) & douze divinités, groupées deux à deux, s'ordonnaient en six couples : n dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et ms2; Dans cette série de huit couples divins, il en est un qui fait prolemé : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généralfefrii dans leur légende qui puisse justifier cette association: Ils rie sont M m ari et femme (comme Zeus-Héra, Poséidon-Amphitrite, HéphaïstosHaris), ni frère et sœur (comme Apollon-Artémis, Hélios-Sélériè), ni mère et fils (comme Aphrodite-Eros), ni protectrice et protégé (comme théna Héraclès). Quel lien unissait donc, dans T esprit de Phidias, un lieu et une dée s s e qui semblent étrangers l'un à l'autre ? On ne saurait "eguer une fantaisie personnelle du sculpteur. Quand il exécuté une livre sacrée, T artiste ancien est tenu de se conformer à certains modèles : l@initiative s'exerce dans le cadre des schèmes imposés par la tradiipnV Hestia - nom propre d'une déesse mais aussi nom commun désignant le foyer - se prêtait moins que les autres dieux grecs à la épresentation anthropomorphe. On la voit rarement figurée. Quand elle 'est,5c'est souvent, comme Phidias l'avait sculptée, faisant couple avec érmès3. De règle dans l'art plastique, l'association Hermès-Hestia RpvÆ ii. -4'i 1 T I T i A r o a -i a

AWS Text Ract

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaïstos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Sélénè), ni mère et fils (comme Aphrodite-Éros), ni protectrice et protégé (comme Athéna-Héraclès). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modèles : son initiative s'exerce dans le cadre des schèmes imposés par la tradition. Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe. On la voit rarement figurée. Quand elle l'est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermès³. De règle dans l'art plastique, l'association Hermès-Hestia

F1 score - 0.9323

Sur la base de la grande statue de Zeus, a Olympie, Phidias avait represente les Douze Dieux. Entre le Soleil (Helios) et la Lune (Selene) les douze divinites, groupées deux a deuix, s ordonnaient en six couples : un dieu-une deesse. Au centre de la frise, en surnombre, les deux divinites (feminine et masculine) que president aux unions : Aphrodite et Eros?. Dans cette serie de huit couples divins, il en est un qui fait probleme': Hermes-Hestia. Pourquoi les apparier ? Rien dans leur genealogre ni dans leur legende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Hera, Poseidon-Amphitrite, Hephaistos-Charis), ni frere et soeur (comme Apollon-Artemis, Helios-Selene), ni mere et fils (comme Aphrodite-Eros), ni protectrice et protege (comme Athena-Heracles): Quel lien unissait donc, dans l'esprit de Phidias; un dieu et une deesse qui semblent etrangers 'un a 1'autre ? On ne saurait alleguer une fantaisie personnelle du sculpteur: Quand il execute une ceuvre sacree, T'artiste ancien est tenu de se conformer a certains modeles : son initiative 'exerce dans le cadre des schemas imposes par la tradition. Hestia - nom propre d'une deesse mais aussi nom commun designant le foyer - se pretait moins que les autres dieux grecs a la representation anthropomorphe. On la voit rarement figuree. Quand elle Pest, c est souvent, comme Phidias 'avait sculptee, faisant couple avec Hermes3 De regle dans l'art plastique, l'association Hermes-Hestia TYy,,

Spark OCR

Image after preprocessing:

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Éros². Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaïstos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Sélénè), ni mère et fils (comme Aphrodite-Éros), ni protectrice et protégé (comme Athéna-Héraclès). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modèles : son initiative s'exerce dans le cadre des schèmes imposés par la tradition. Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe. On la voit rarement figurée. Quand elle l'est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermès³. De règle dans l'art plastique, l'association Hermès-Hestia

F1 score - 0.9812

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions: Aphrodite et Eros", Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaïstos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Sélénè), ni mère et fils (comme Aphrodite-Eros), ni protectrice et protégé (comme Athéna-Héraclès). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modeéles : son initiative s'exerce dans le cadre des schèmes imposés par la tradition. Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe. On la voit rarement figurée. Quand elle Vest, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermés*, De règle dans l'art plastique, l'association Hermès-Hestia

Coding ...

Spark NLP Resources

Spark NLP Official page

Spark NLP Workshop Repo

JSL Youtube channel

JSL Blogs

Introduction to Spark NLP: Foundations and Basic Components (Part-I)

Introduction to: Spark NLP: Installation and Getting Started (Part-II)

Named Entity Recognition with Bert in Spark NLP

Text Classification in Spark NLP with Bert and Universal Sentence Encoders

Spark NLP 101 : Document Assembler

Spark NLP 101: LightPipeline

<https://www.oreilly.com/radar/one-simple-chart-who-is-interested-in-spark-nlp/>

<https://blog.dominodatalab.com/comparing-the-functionality-of-open-source-natural-language-processing-libraries/>

<https://databricks.com/blog/2017/10/19/introducing-natural-language-processing-library-apache-spark.html>

<https://databricks.com/fr/session/apache-spark-nlp-extending-spark-ml-to-deliver-fast-scalable-unified-natural-language-processing>

<https://medium.com/@saif1988/spark-nlp-walkthrough-powered-by-tensorflow-9965538663fd>

<https://www.kdnuggets.com/2019/06/spark-nlp-getting-started-with-worlds-most-widely-used-nlp-library-enterprise.html>

<https://www.forbes.com/sites/forbestechcouncil/2019/09/17/why-spark-nlp-is-the-most-widely-used-nlp-library-enterprise/>

<https://medium.com/hackernoon/mueller-report-for-nerds-spark-meets-nlp-with-tensorflow-and-bert-part-1-32490a8f8f12>

<https://www.analyticsindiamag.com/5-reasons-why-spark-nlp-is-the-most-widely-used-library-in-enterprises/>

<https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-training-spark-nlp-and-spacy-pipelines>

<https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-accuracy-performance-and-scalability>

<https://www.infoworld.com/article/3031690/analytics/why-you-should-use-spark-for-machine-learning.html>



NOW ANNOUNCING

NLP SUMMIT

Applied Natural
Language Processing

Boston, Oct 27-28 | San Francisco, Nov 17-18