Spark NLP for Healthcare Data Scientists

August 5-6, 2020

Veysel Kocaman
Sr. Data Scientist
veysel@johnsnowlabs.com

Welcome

Day-1	60 min	- Intro to John Snow Labs and Spark NLP - Intro to NLP and Clinical NLP Modules in Spark NLP
	10 min	Break
	50 min	- Clinical Named Entity Recognition with Spark NLP
	10 min	Break
	50 min	- Clinical Assertion Status Model in Spark NLP - Clinical Spell Checkers
Day-2	50 min	- Clinical Entity Resolvers
	10 min	Break
	60 min	- Deidentification and Obfuscation of PHI - Generic Classifier with TF - Keyword Extraction (YAKE)
	10 min	Break
	50 min	- Spark OCR

Setup

RUNNING CODE:

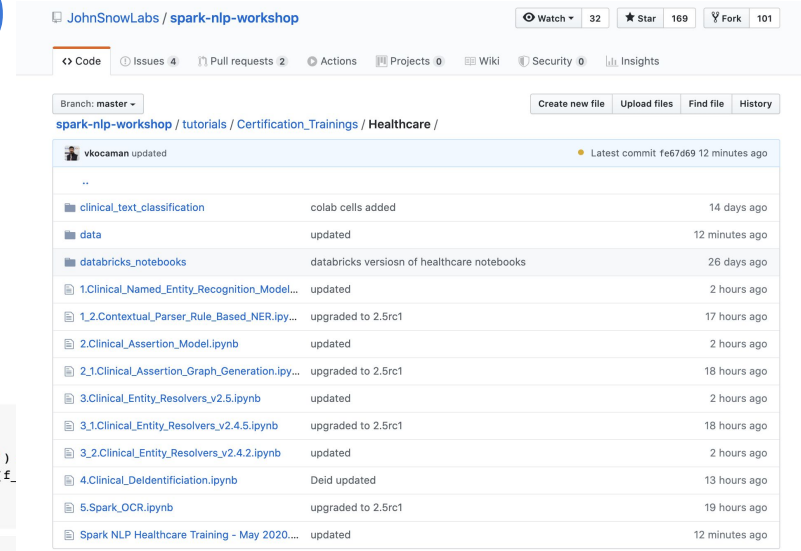
https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/Certification_Trainings/Healthcare

[How to set up Google Colab]

BOOKMARK:

nlp.johnsnowlabs.com/docs/en/concepts

spark-nlp.slack.com




```
[ ] import json

    with open('license_keys.json')
        license_keys = json.load(f_

    license_keys.keys()
```

```
[ ] import os

    # Install java
    ! apt-get install -y openjdk-8-jdk-headless -qq > /dev/null
    os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
    os.environ["PATH"] = os.environ["JAVA_HOME"] + "/bin:" + os.environ["PATH"]
    ! java -version

    # Install pyspark
    ! pip install --ignore-installed -q pyspark==2.4.4

    secret = license_keys['secret']
    os.environ['SPARK_NLP_LICENSE'] = license_json['SPARK_NLP_LICENSE']
    os.environ['JSL_OCR_LICENSE'] = license_json['JSL_OCR_LICENSE']
    os.environ['AWS_ACCESS_KEY_ID']= license_json['AWS_ACCESS_KEY_ID']
    os.environ['AWS_SECRET_ACCESS_KEY'] = license_json['AWS_SECRET_ACCESS_KEY']

    ! python -m pip install --upgrade spark-nlp-jsl==2.4.2  --extra-index-url https://pypi.johnsnowlabs.com/$secret

    # Install Spark NLP
    ! pip install --ignore-installed -q spark-nlp==2.4.5
```

Part - I

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

JOHN SNOW LABS

CIO Review ARTIFICIAL INTELLIGENCE SOLUTION PROVIDER OF THE YEAR - 2018

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

AI PLATFORM OF THE YEAR

CIO APPLICATIONS 2019

"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."

Spark NLP in Healthcare

John Snow LABS

"As this company and its award-winning innovations show us, the future was on display at this year's Strata Data Conference. Keep an eye on this company – as it represents where the industry and data science are headed."

Ben Lorica, chief data scientist at O'Reilly and Strata Data Conference program chair

Introducing Spark NLP

- Natural Language Toolkit (NLTK): The complete toolkit for all NLP techniques.
- TextBlob: Easy to use NLP tools API, built on top of NLTK and Pattern.
- SpaCy: Industrial strength NLP with Python and Cython.
- Gensim: Topic Modelling for Humans
- Stanford Core NLP: NLP services and packages by Stanford NLP Group.
- Fasttext: NLP library by Facebook's AI Research (FAIR) lab
- ...

- 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
  - Lack of any NLP library that's fully supported by Spark
  - 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.)

https://medium.com/spark-nlp/introduction-to-spark-nlp-foundations-and-basic-components-part-i-c83b7629ed59

TRUSTED BY


Microsoft


Google


amazon


intel


IBM


databricks


verizon


indeed


CapitalOne


viacom


McKESSON


MERCK


Roche


selectdata


UiPath
Reboot Work.


ASCO
CANCER·LINQ
DISCOVERY

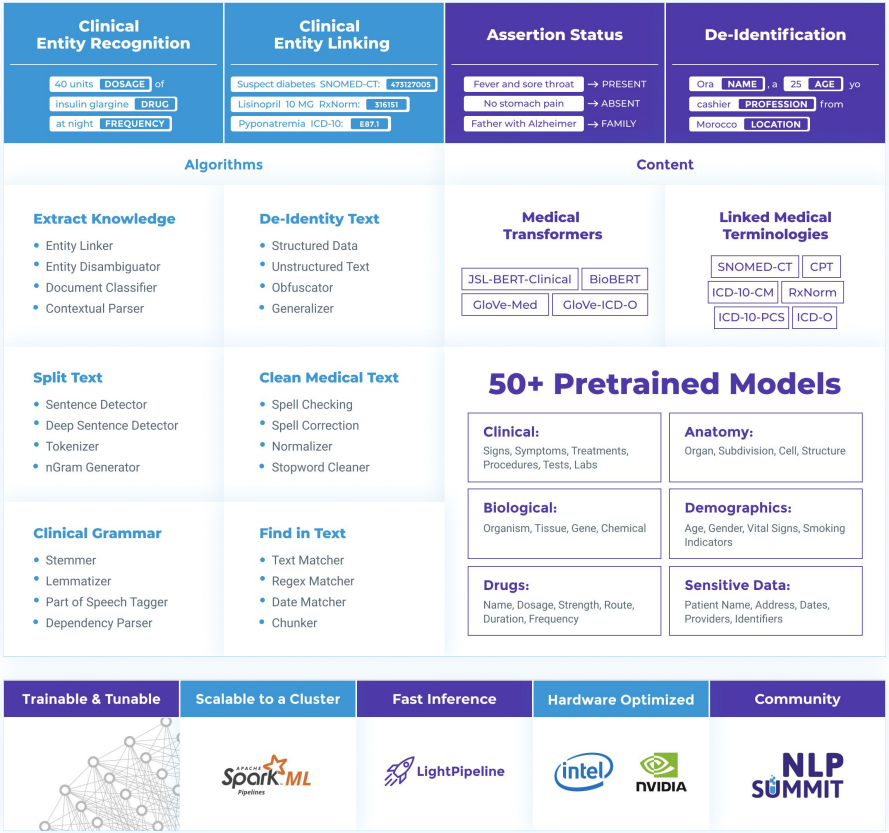

cnrs

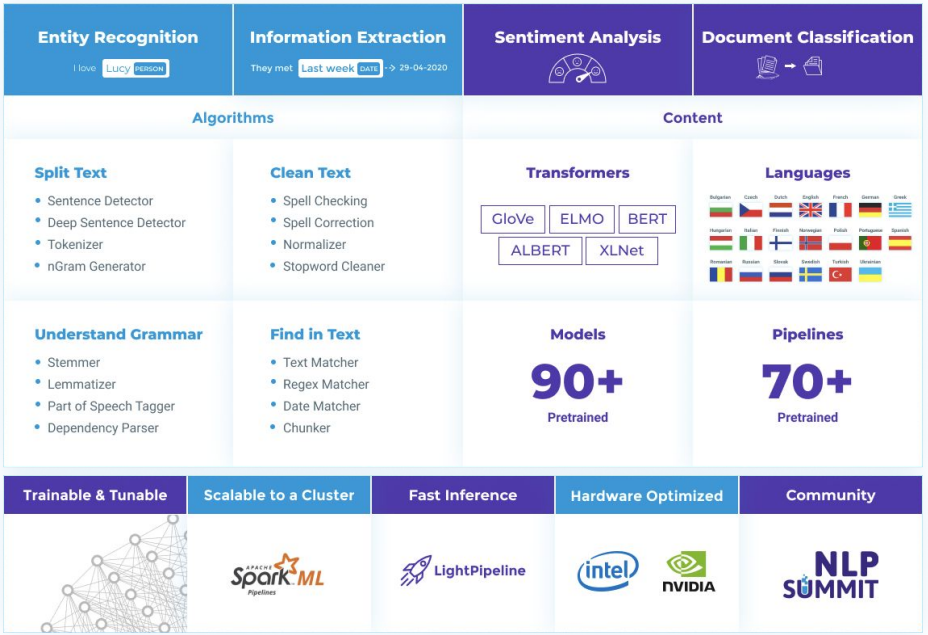

Imperial College
London


Georgia
Tech


STANFORD
UNIVERSITY

Spark NLP Modules (Enterprise and Public)

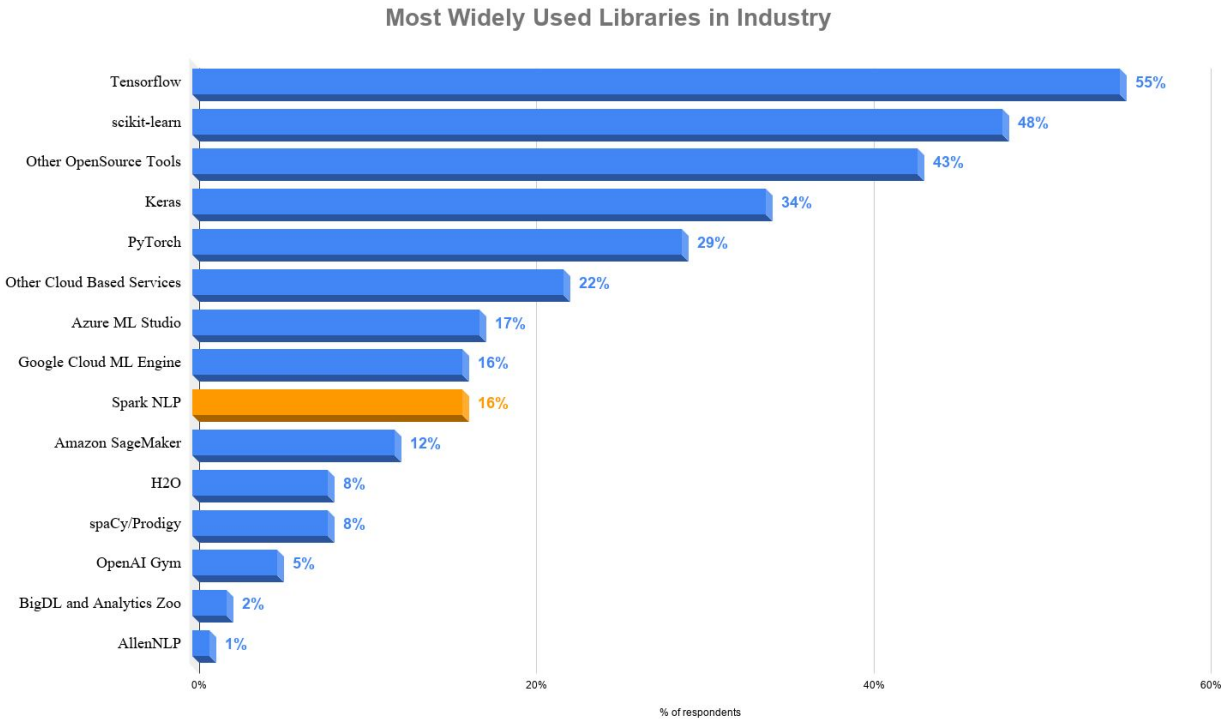




Introducing Spark NLP

Name	Spark NLP	spaCy	NLTK	CoreNLP
Sentence detection	Yes	Yes	Yes	Yes
Tokenization	Yes	Yes	Yes	Yes
Stemming	Yes	Yes	Yes	Yes
Lemmatization	Yes	Yes	Yes	Yes
POS tagger	Yes	Yes	Yes	Yes
NER	Yes	Yes	Yes	Yes
Dependency parse	Yes	Yes	Yes	Yes
Text matcher	Yes	**Yes**	**No**	Yes
Date matcher	Yes	**No**	**No**	Yes
Chunking	Yes	Yes	Yes	Yes
Spell checker	Yes	**No**	**No**	**No**
Sentiment detector	Yes	**No**	**No**	Yes
Pretrained models	Yes	Yes	Yes	Yes
Training models	Yes	Yes	Yes	Yes

Available in **Python**, **R**, **Scala** and **Java**

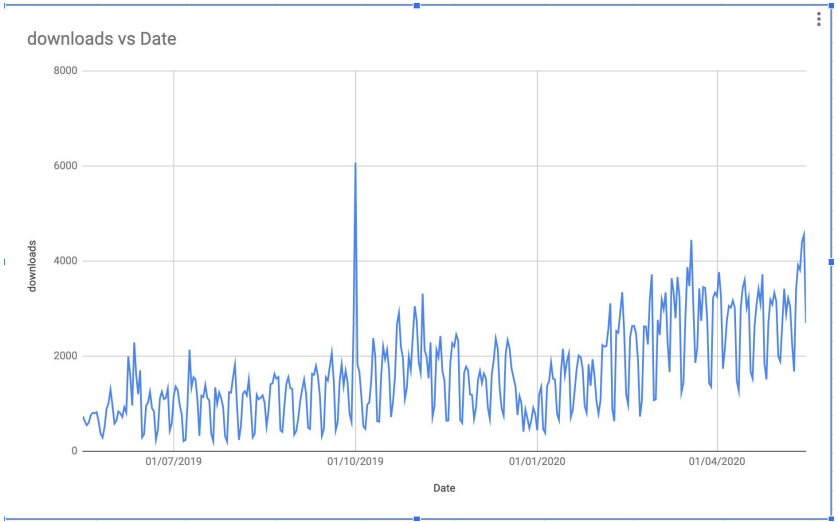


"AI Adoption in the Enterprise", February 2019
Most widely used ML frameworks and tools survey of 1,300 practitioners by O'Reilly

OFFICIALLY SUPPORTED RUNTIMES

Introducing Spark NLP

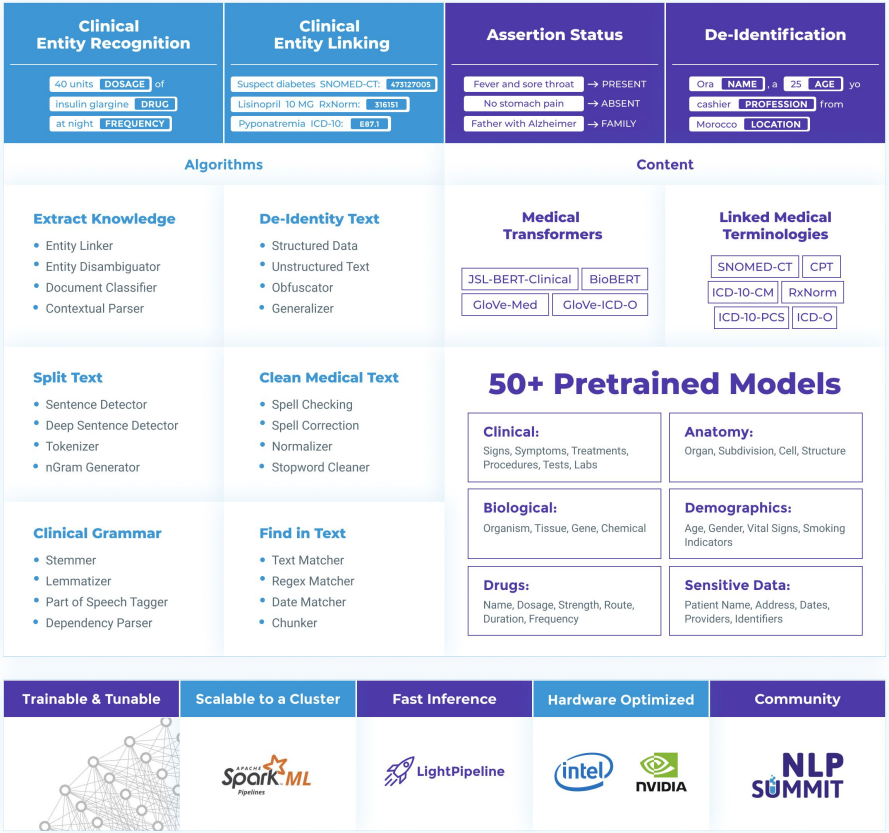
Summary	Daily ~ 5K Monthly > 100K
PyPI link	https://pypi.org/project/spark-nlp
Total downloads	1,012,260
Total downloads - 30 days	137,021
Total downloads - 7 days	32,245

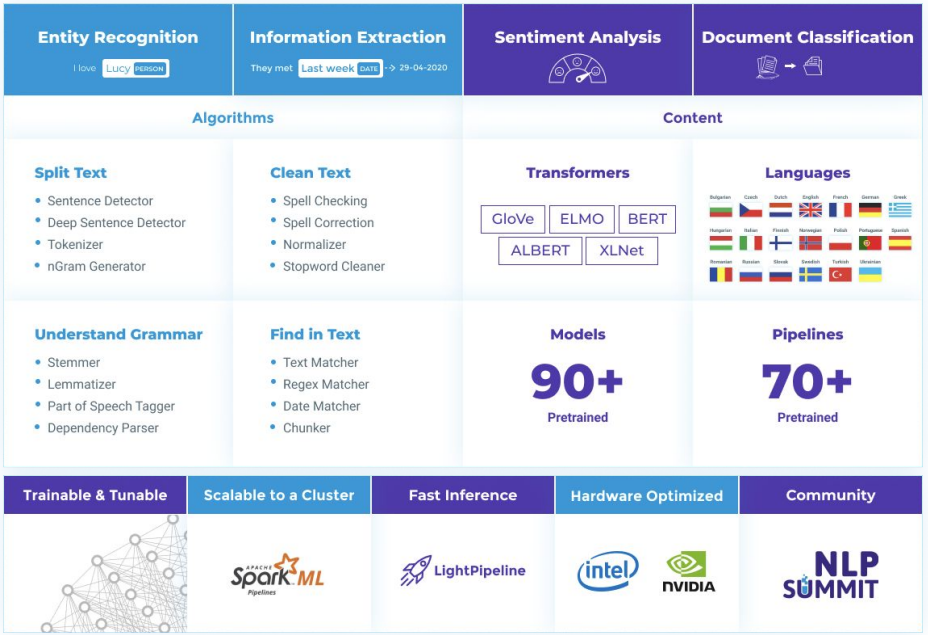


- Spark NLP is an open-source natural language processing library, built on top of Apache Spark and Spark ML. (initial release: Oct 2017)
  - A single unified solution for all your NLP needs
  - Take advantage of transfer learning and implementing the latest and greatest **SOTA** algorithms and models in NLP research
  - The most widely used NLP library in industry.
  - 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.)

https://medium.com/spark-nlp/introduction-to-spark-nlp-foundations-and-basic-components-part-i-c83b7629ed59

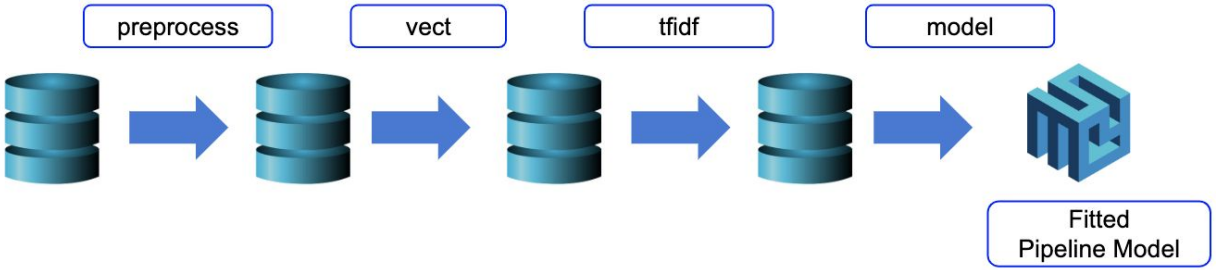
Spark NLP Modules (Enterprise and Public)



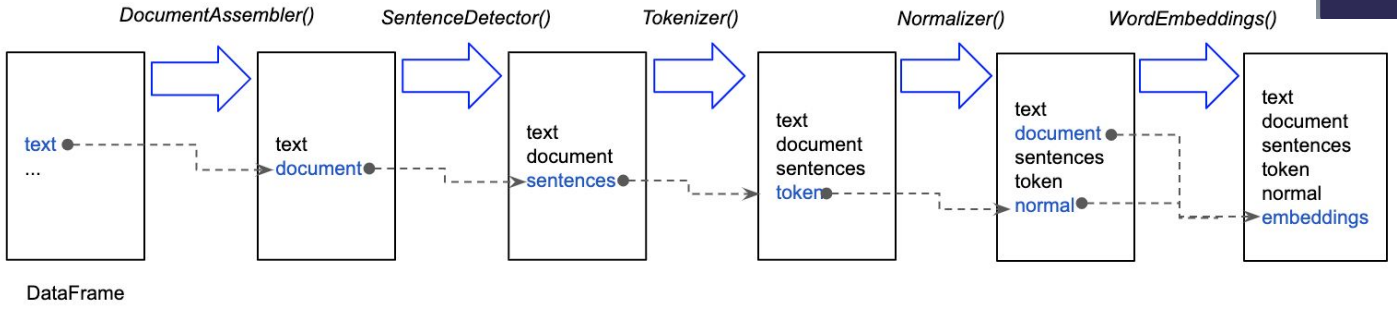


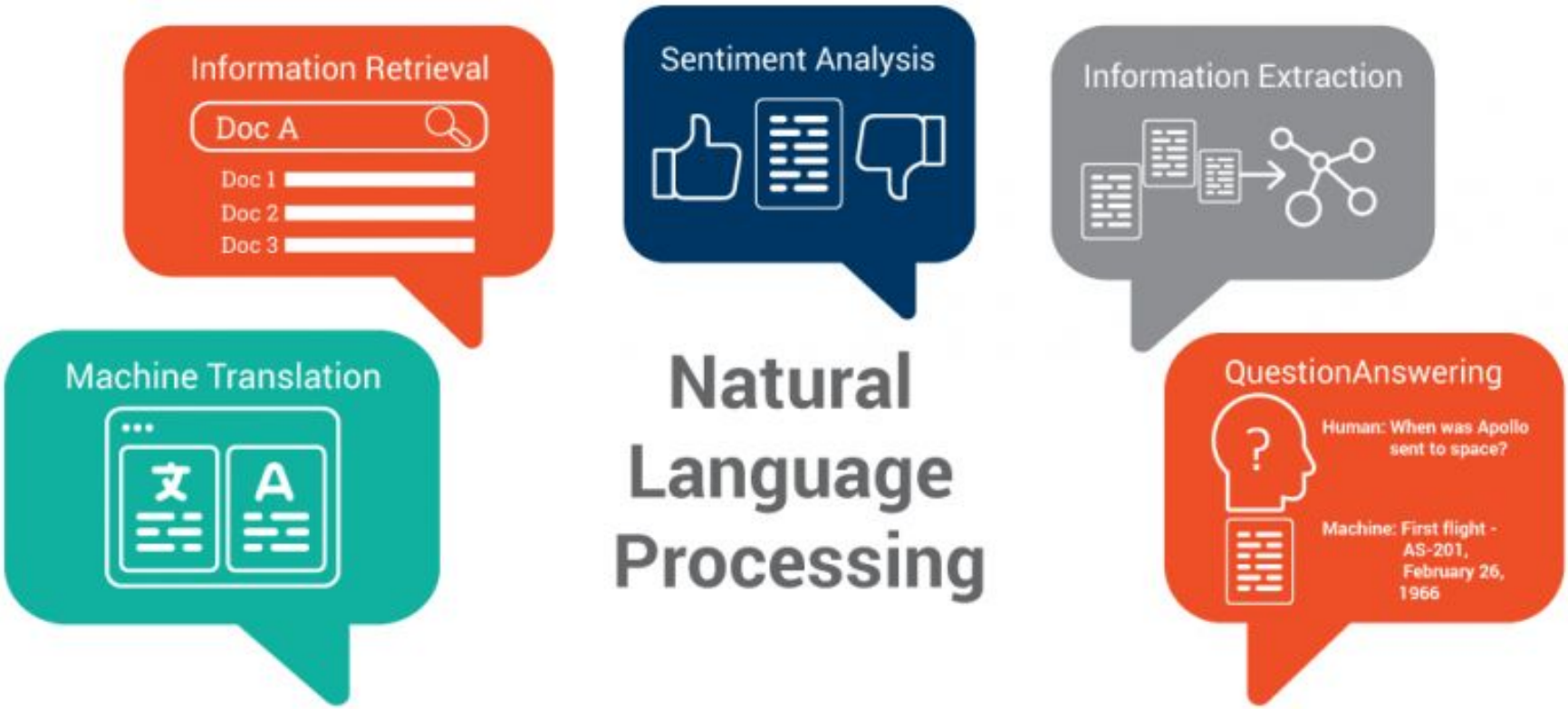
Introducing Spark NLP

Pipeline of annotators




```
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)
```




Information Retrieval
Doc A
Doc 1
Doc 2
Doc 3
Sentiment Analysis
Information Extraction
Machine Translation
文
A
Natural Language Processing
QuestionAnswering
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
studying	-ing	study
niñas	-as	niñ
niñez	-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

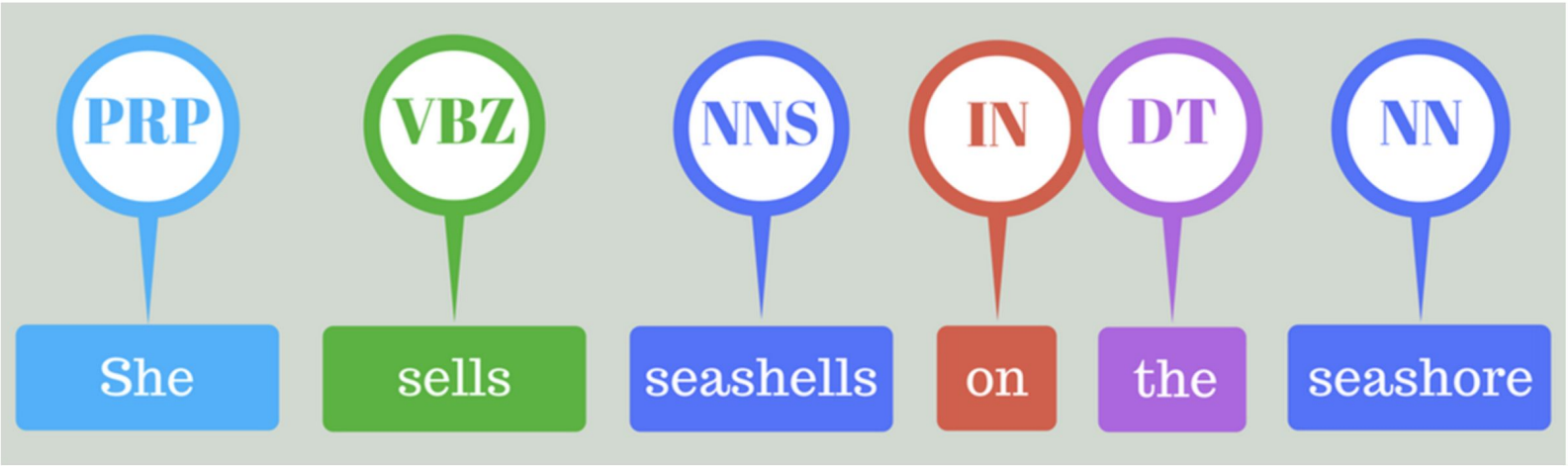
Remove or replace undesirable characters or regular expressions:

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

Spark NLP also comes with a Slang normalizer:

PART OF SPEECH TAGGING

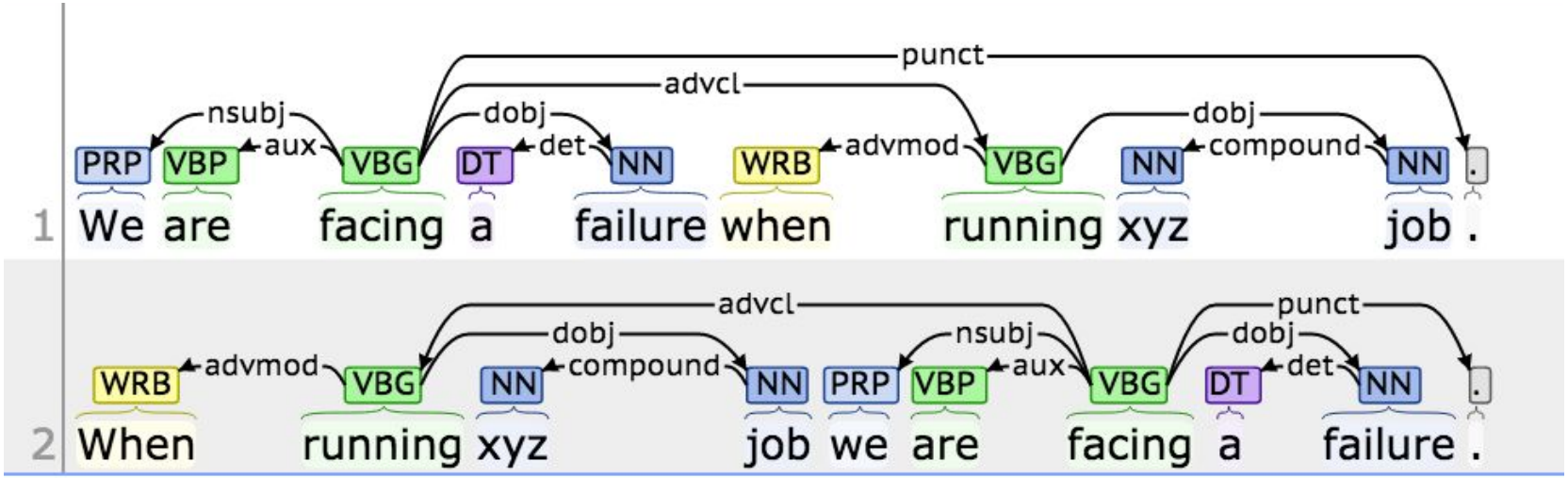
Often useful for recognizing named entities or word relationships.



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.

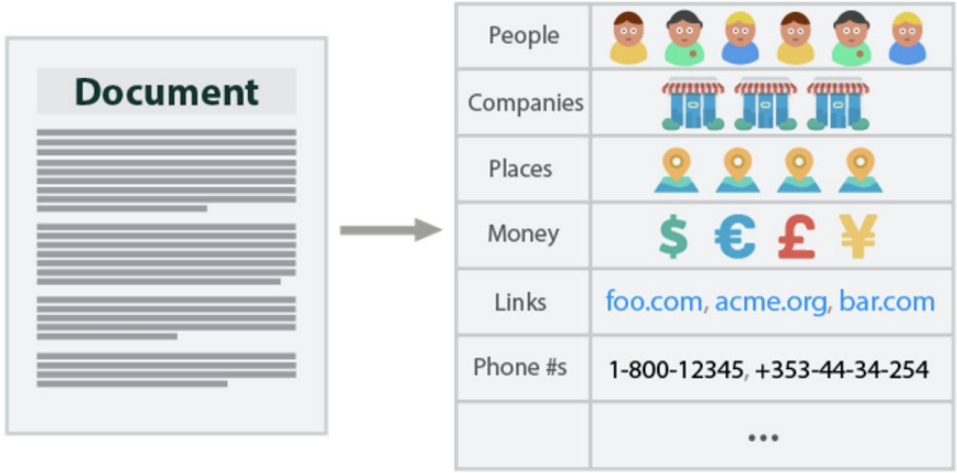
DEPENDENCY PARSING

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



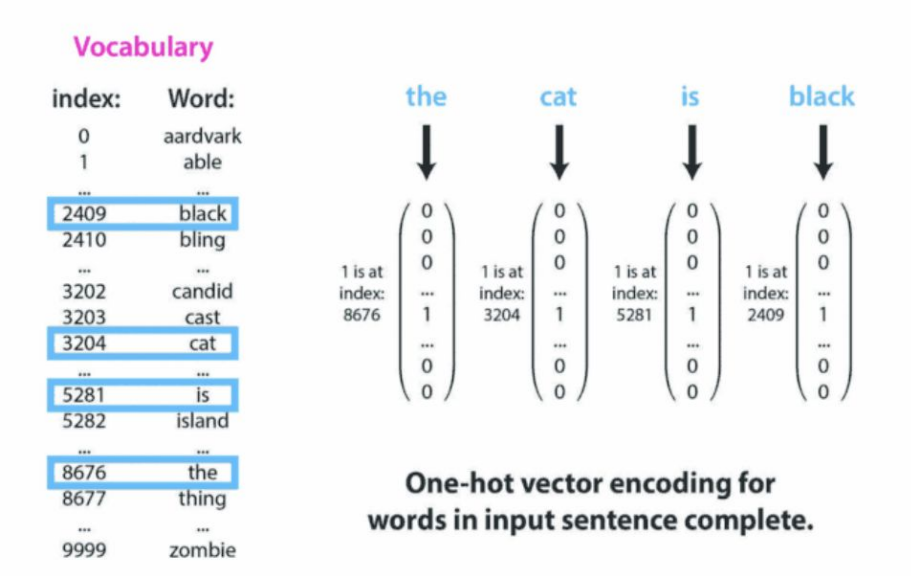
Named Entity Recognition (NER)

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.

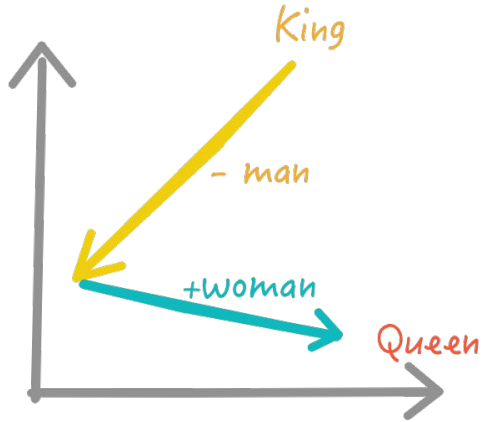
Word & Sentence Embeddings



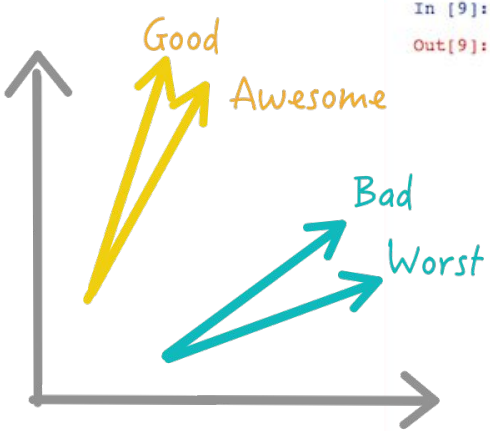

```
In [9]: doc[3].vector
Out[9]: array([ 0.037103 , -0.31259  , -0.17857  ,  0.30001  ,  0.078154 ,
                0.17958  ,  0.12048  , -0.11879  , -0.20601  ,  1.2849   ,
               -0.20409  ,  0.80613  ,  0.34344  , -0.19191  , -0.084511 ,
                0.17339  ,  0.042483 ,  2.0282   , -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  ,
```

- 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)

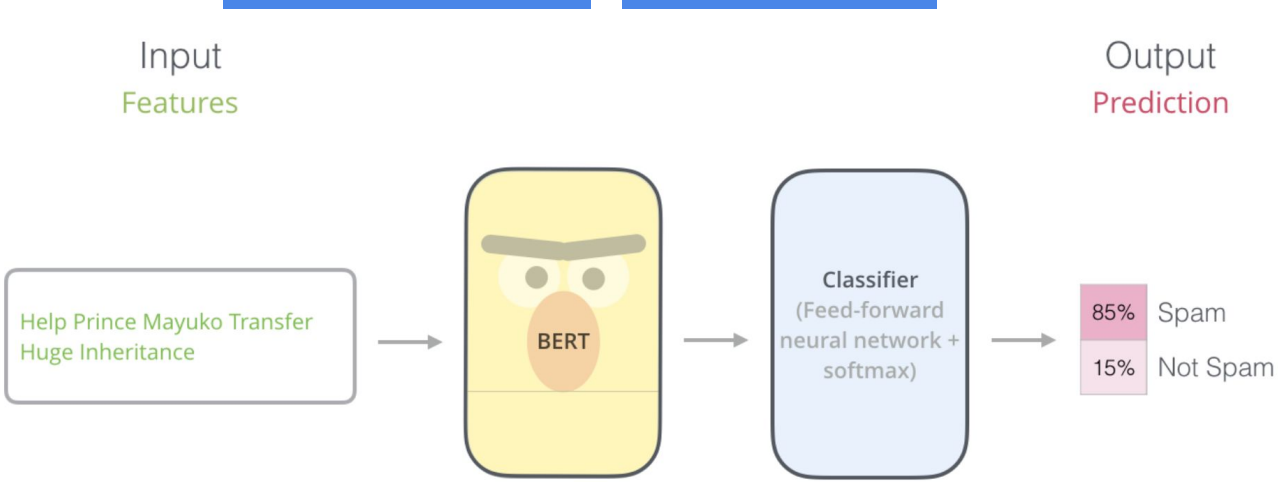
ELMO
(512, 1024)

BERT
(768d)

Albert
(768, 1024, 2048, 4096)

XLNet
(768, 1024)

Universal Sentence Encoders
(512)

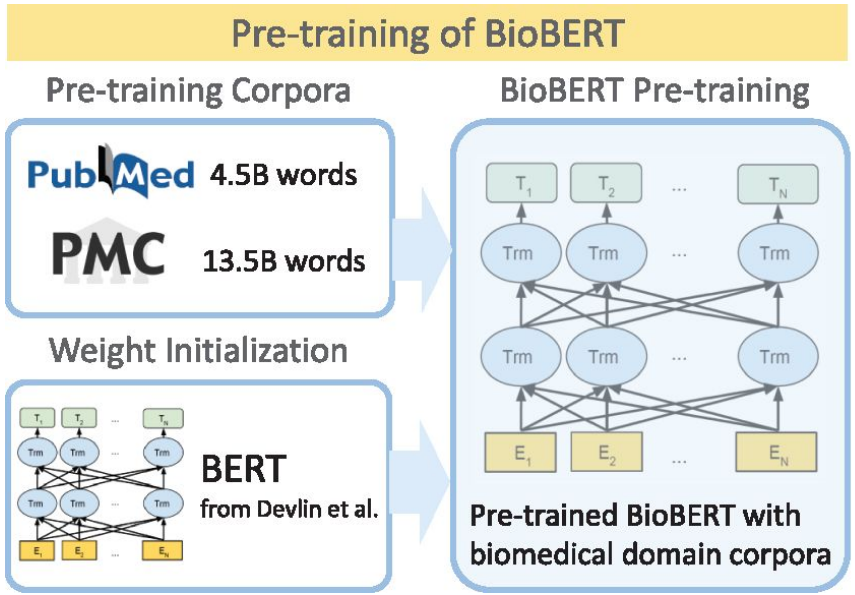


Clinical Word Embeddings

Clinical Glove (200d)	ICDO Glove (200d)	Bio BERT	Clinical BERT
PubMed + PMC	PubMed + ICD10 UMLS + MIMIC III	Pubmed + PMC	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

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

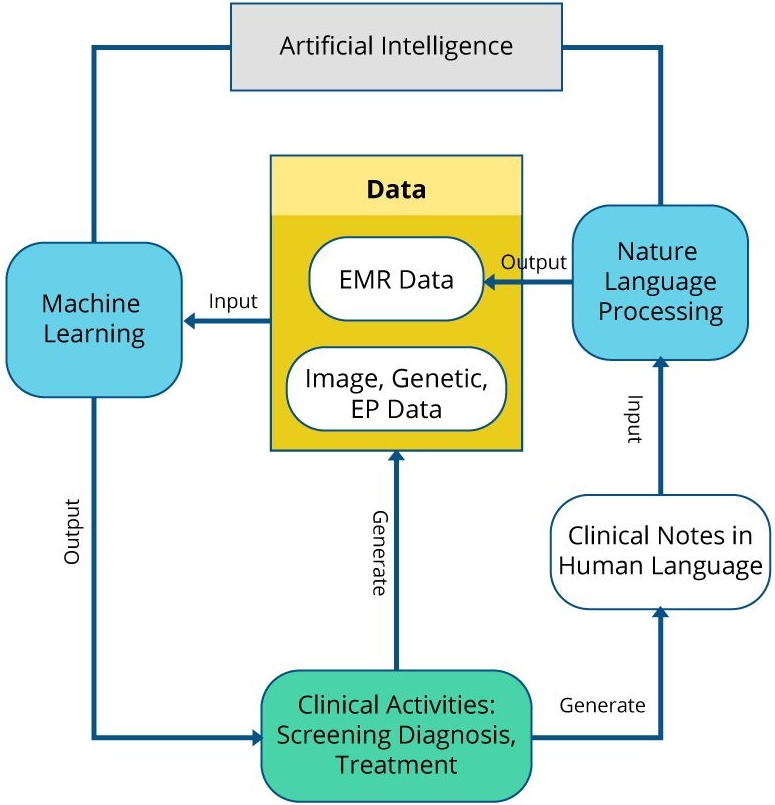
Healthcare specific needs:

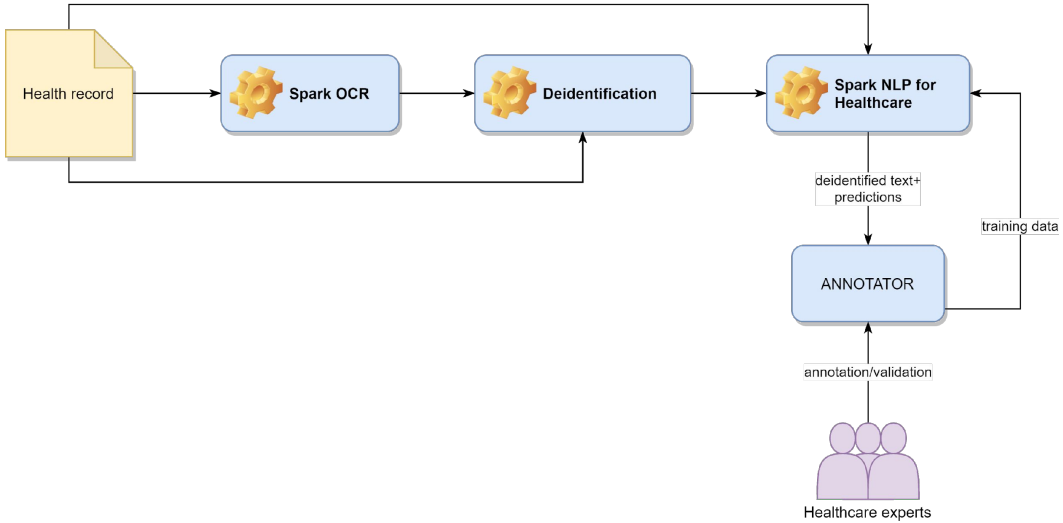
1. Core Annotators
   Part of speech, spell checking, …
2. Vocabulary
   Ontologies, relationships, word embeddings, …
3. ML & DL Models
   Named entity recognition, entity resolution, …

ED Triage Notes
states started last night, upper abd, took alka seltzer approx 0500, no relief. nausea no vomiting
Since yeatreday 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

NLP in Healthcare



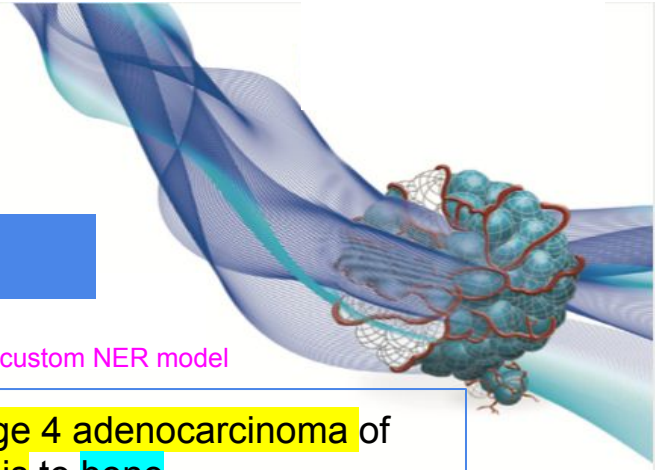


She returns today for ongoing evaluation of her EGFR mutated, stage 4 lung cancer with metastasis to her L2 vertebrae and her lungs bilaterally.

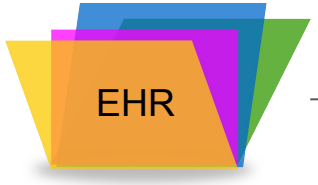
Bone negative for metastatic disease.

Patient denies any family history of cancer.

NLP in Healthcare



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



Extract metastasis related sentences with mets-classifier

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

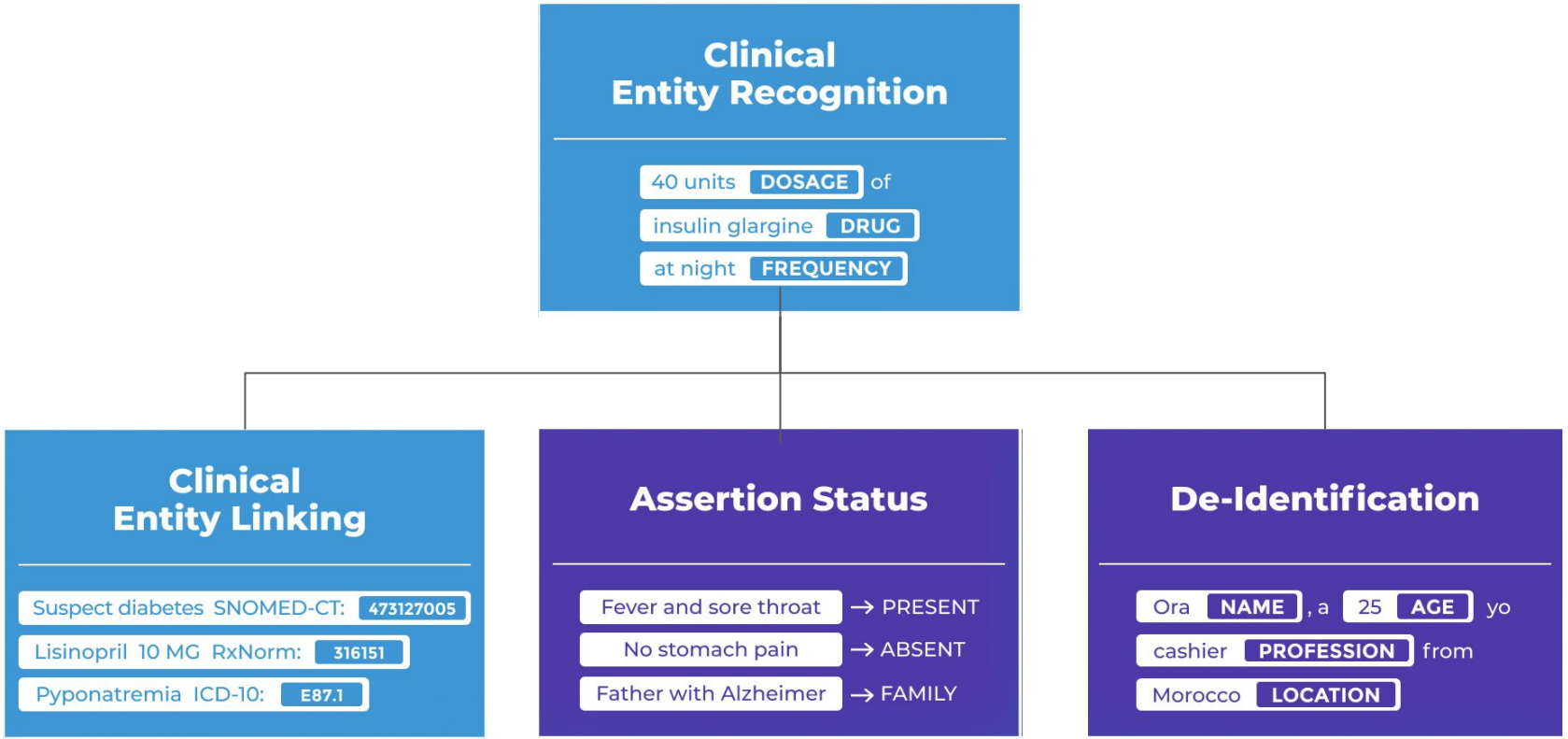
- *The frequency of clinical visits using document dates*
- *The number of positive site (organ) entities (hits by window)*
- *The number of Radiology/Oncology/Pathology reports in the last x days*
- *The number of tests applied in the last x days*
- *The number of diseases detected in the last x days*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

Text embeddings thru clinical word embeddings

Train a DL & ML Model

Prediction: Bone metastasis on June 2018

Named Entity Recognition (NER)



Healthcare extensions

NLP Library / Feature	State of the Art (SOTA) Research
Named Entity Recognition	"Entity Recognition from Clinical Texts via Recurrent Neural Network". *Liu et al., BMC Medical Informatics & Decision Making, July 2017.*
Word Embeddings	- "How to Train Good Word Embeddings for Biomedical NLP". *Chiu et al., In Proceedings of BioNLP'16, August 2016.* - "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *Devlin et. al. (Google Research), October 2018.*
Assertion Status Detection	- "Improving Classification of Medical Assertions in Clinical Notes". *Kim et al., In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.* - "Neural Networks For Negation Scope Detection" *Fancellu et al., In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.*
Entity Resolution	"CNN-based ranking for biomedical entity normalization". *Li et al., BMC Bioinformatics, October 2017.*

Not Secure | nlpprogress.com/eng...

CoNLL 2003 (English)

The CoNLL 2003 NER task consists of newswire text from the Reuters RCV1 corpus tagged with four different entity types (PER, LOC, ORG, MISC). Models are evaluated based on span-based F1 on the test set. ♦ used both the train and development splits for training.

Model	F1	Paper / Source	Code
CNN Large + fine-tune (Baevski et al., 2019)	93.5	Cloze-driven Pretraining of Self-attention Networks	
RNN-CRF+Flair	93.47	Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition	
LSTM-CRF+ELMo+BERT+Flair	93.38	Neural Architectures for Nested NER through Linearization	Official
Flair embeddings (Akbik et al., 2018)♦	93.09	Contextual String Embeddings for Sequence Labeling	Flair framework
BERT Large (Devlin et al., 2018)	92.8	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	
CVT + Multi-Task (Clark et al., 2018)	92.61	Semi-Supervised Sequence Modeling with Cross-View Training	Official
BERT Base (Devlin et al., 2018)	92.4	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	
BiLSTM-CRF+ELMo (Peters et al., 2018)	92.22	Deep contextualized word representations	AllenNLP Project AllenNLP GitHub
Peters et al. (2017) ♦	91.93	Semi-supervised sequence tagging with bidirectional language models	
CRF + AutoEncoder (Wu et al., 2018)	91.87	Evaluating the Utility of Hand-crafted Features in Sequence Labelling	Official
Bi-LSTM-CRF + Lexical Features (Ghaddar and Langlais 2018)	91.73	Robust Lexical Features for Improved Neural Network Named-Entity Recognition	Official
BiLSTM-CRF + IntNet (Xin et al., 2018)	91.64	Learning Better Internal Structure of Words for Sequence Labeling	
Chiu and Nichols (2016) ♦	91.62	Named entity recognition with bidirectional LSTM-CNNs	

NER-DL in Spark NLP

SYSTEM	YEAR	LANGUAGE	ACCURACY
Spark NLP v2.4	2020	Python/Scala/Java/R	**93.3 (test F1) - 95.9 (dev F1)**
Spark NLP v2.x	2019	Python/Scala/Java/R	**93**
Spark NLP v1.x	2018	Python/Scala/Java/R	92
spaCy v2.x	2017	Python/Cython	92.6
spaCy v1.x	2015	Python/Cython	91.8
ClearNLP	2015	Java	91.7
CoreNLP	2015	Java	89.6
MATE	2015	Java	92.5
Turbo	2015	C++	92.4

The best NER score in production

93.3 %
Test Set

Bert

NerDLApproach

John Snow LABS

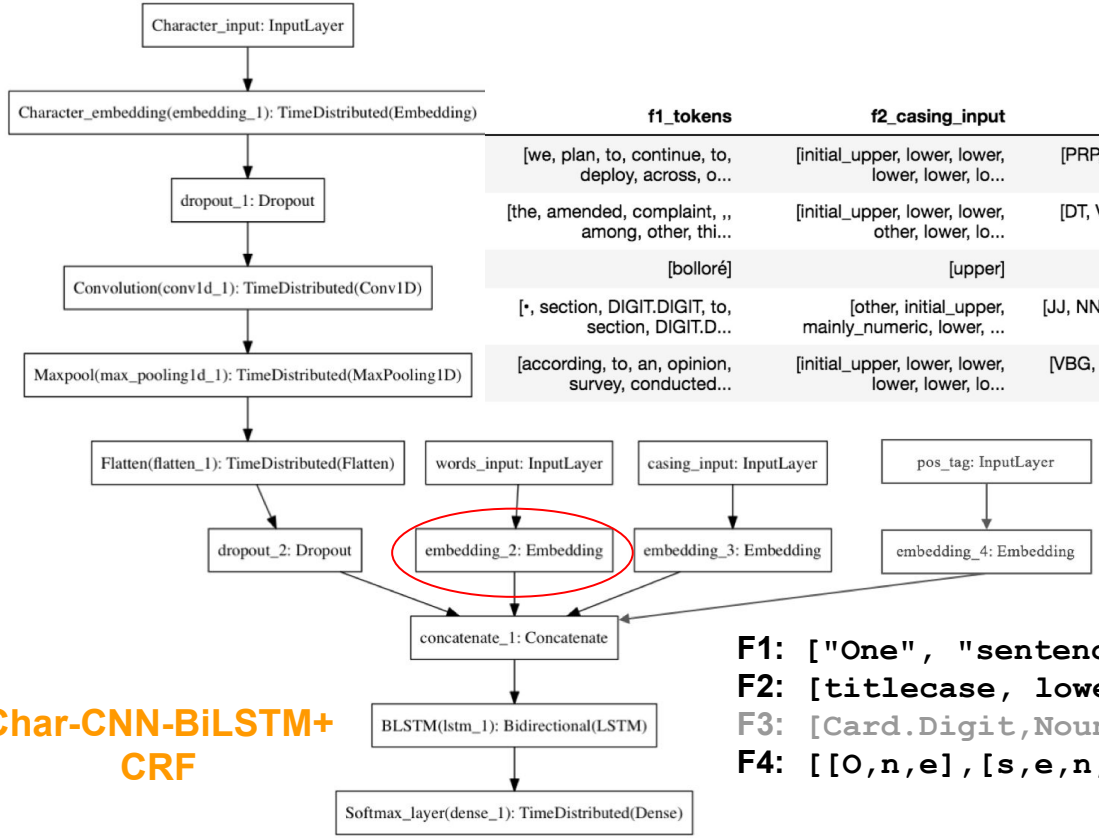
NER Systems

Feature-engineered machine learning systems	Dict	SP	DU	EN	GE
Carreras et al. (2002) binary AdaBoost classifiers	Yes	81.39	77.05	-	-
Malouf (2002) - Maximum Entropy (ME) + features	Yes	73.66	68.08	-	-
Li et al. (2005) SVM with class weights	Yes	-	-	88.3	-
Passos et al. (2014) CRF	Yes	-	-	90.90	-
Ando and Zhang (2005a) Semi-supervised state of the art	No	-	-	89.31	75.27
Agerri and Rigau (2016)	Yes	**84.16**	**85.04**	**91.36**	**76.42**
Feature-inferring neural network word models					
Collobert et al. (2011) Vanilla NN +SLL / Conv-CRF	No	-	-	81.47	-
Huang et al. (2015) Bi-LSTM+CRF	No	-	-	84.26	-
Yan et al. (2016) Win-BiLSTM (English), FF (German) (Many fets)	Yes	-	-	88.91	**76.12**
Collobert et al. (2011) Conv-CRF (SENNA+Gazetteer)	Yes	-	-	89.59	-
Huang et al. (2015) Bi-LSTM+CRF+ (SENNA+Gazetteer)	Yes	-	-	**90.10**	-
Feature-inferring neural network character models					
Gillick et al. (2015) – BTS	No	**82.95**	**82.84**	**86.50**	**76.22**
Kuru et al. (2016) CharNER	No	82.18	79.36	84.52	70.12
Feature-inferring neural network word + character models					
Yang et al. (2017)	Yes	85.77	**85.19**	91.26	-
Luo (2015)	Yes	-	-	91.20	-
Chiu and Nichols (2015)	Yes	-	-	**91.62**	-
Ma and Hovy (2016)	No	-	-	91.21	-
Santos and Guimaraes (2015)	No	82.21	-	-	-
Lample et al. (2016)	No	85.75	81.74	90.94	**78.76**
Bharadwaj et al. (2016)	Yes	**85.81**	-	-	-
Dernoncourt et al. (2017)	No	-	-	90.5	-
Feature-inferring neural network word + character + affix models					
Re-implementation of Lample et al. (2016) (100 Epochs)	No	85.34	85.27	90.24	78.44
Yadav et al. (2018)(100 Epochs)	No	86.92	87.50	90.69	78.56
Yadav et al. (2018) (150 Epochs)	No	**87.26**	**87.54**	90.86	**79.01**

1. Classical Approaches (rule based)
2. ML Approaches
   - Multi-class classification
   - Conditional Random Field (CRF)
3. DL Approaches
   - Bidirectional LSTM-CRF
   - Bidirectional LSTM-CNNs
   - Bidirectional LSTM-CNNS-CRF
   - Pre-trained language models (Bert, Elmo)
4. Hybrid Approaches (DL + ML)

John Snow LABS

NER-DL in Spark NLP



f1_tokens	f2_casing_input	f3_pos_tag	f4_chars	l1_classes
[we, plan, to, continue, to, deploy, across, o...	[initial_upper, lower, lower, lower, lower, lo...	[PRP, VBP, TO, VB, TO, VB, IN, PRP$, NN, CC, V...	[[W, e], [p, l, a, n], [t, o], [c, o, n, t, i,...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ...
[the, amended, complaint, ,, among, other, thi...	[initial_upper, lower, lower, other, lower, lo...	[DT, VBN, NN, ,, IN, JJ, NNS, ,, VBZ, DT, NNS,...	[[T, h, e], [a, m, e, n, d, e, d], [c, o, m, p...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ...
[bolloré]	[upper]	[NN]	[[B, O, L, L, O, R, É]]	[Group]
[•, section, DIGIT.DIGIT, to, section, DIGIT.D...	[other, initial_upper, mainly_numeric, lower, ...	[JJ, NN, CD, TO, NN, CD, IN, DT, NNP, NNP, ,, ...	[[•], [S, e, c, t, i, o, n], [5, ., 1], [t, o]...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ...
[according, to, an, opinion, survey, conducted...	[initial_upper, lower, lower, lower, lower, lo...	[VBG, TO, DT, NN, NN, VBN, IN, NNS, IN, CD, ,,...	[[A, c, c, o, r, d, i, n, g], [t, o], [a, n], ...	[O, O, O, O, O, O, O, Company, O, O, O, Gro...

Char-CNN-BiLSTM+ CRF

F1: `["One", "sentence", "with", "5", "WORDS"]`
F2: `[titlecase, lowercase, lowercase, numeric, uppercase]`
F3: `[Card.Digit,Noun,Preposition,Card.Digit,Noun plural]`
F4: `[[O,n,e],[s,e,n,t,e,n,c,e],[w,i,t,h],[5],[W,O,R,D,S]]`

NER-DL in Spark NLP

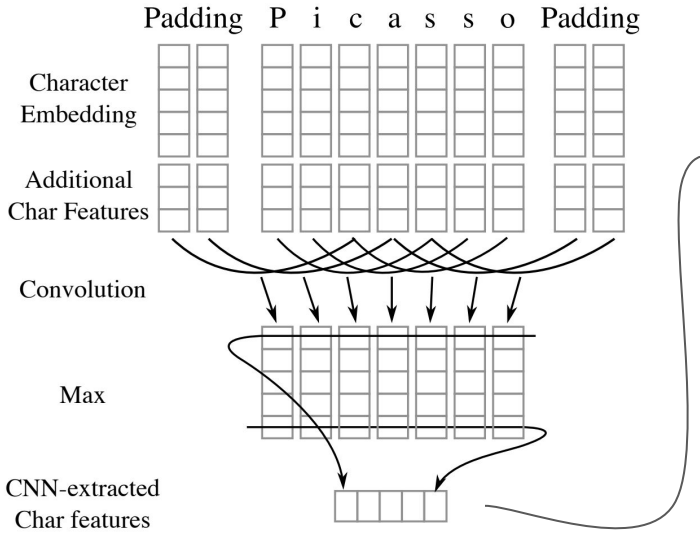


Figure 2: The convolutional neural network extracts character features from each word. The character embedding and (optionally) the character type feature vector are computed through lookup tables. Then, they are concatenated and passed into the CNN.

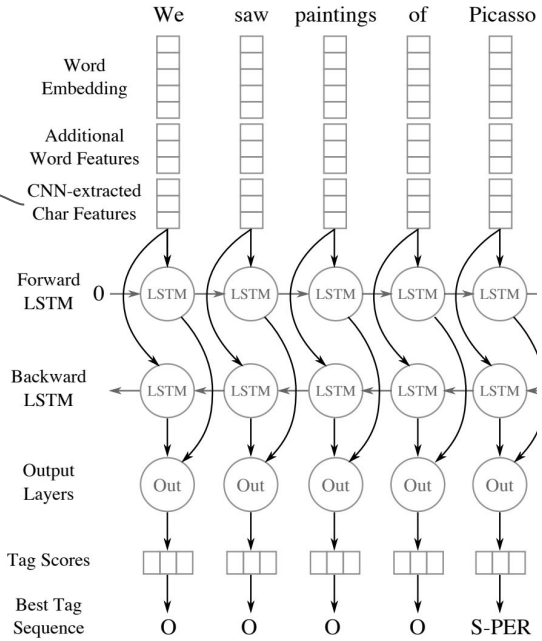


Figure 1: The (unrolled) BLSTM for tagging named entities. Multiple tables look up word-level feature vectors. The CNN (Figure 2) extracts a fixed length feature vector from character-level features. For each word, these vectors are concatenated and fed to the BLSTM network and then to the output layers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

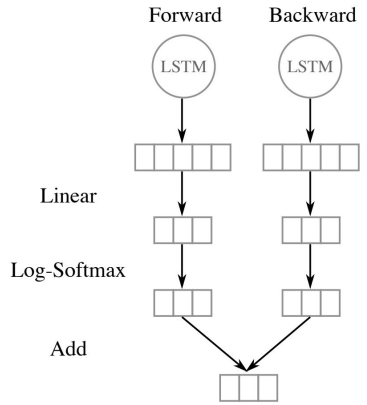


Figure 3: The output layers ("Out" in Figure 1) decode output into a score for each tag category.

Char-CNN-BiLSTM

John Snow LABS

Clinical Named Entity Recognition

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m2 , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . However , serum chemistry obtained six hours after presentation revealed her glucose was 186 mg/dL , the anion gap was still elevated at 21 , serum bicarbonate was 16 mmol/L , triglyceride level peaked at 2050 mg/dL , and lipase was 52 U/L . The β-hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L - the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again .

Clinical NER

Color codes:PROBLEM, TREATMENT, TEST,

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, Pathological_formation, 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-DL in Spark NLP

CoNNL2003 format

All data files contain one word per line with empty lines representing sentence boundaries. At the end of each line there is a tag which states whether the current word is inside a named entity or not. The tag also encodes the type of named entity. Here is an example sentence:

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

* Each line contains four fields: the word, its part-of-speech tag, its chunk tag and its named entity tag.

* **CoNLL**: Conference on Computational Natural Language Learning

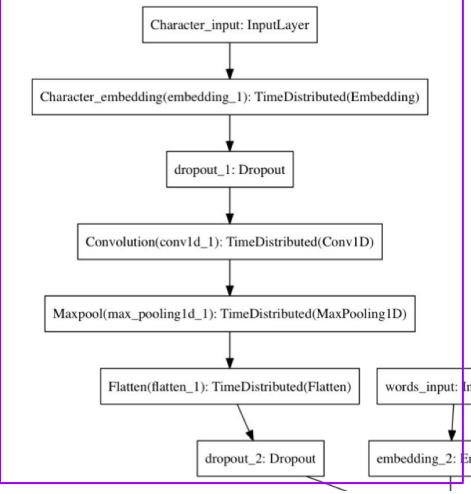
BIO schema

```
John      B-PER
Smith     I-PER
lives     O
in        O
New       B-LOC
York      I-LOC
```

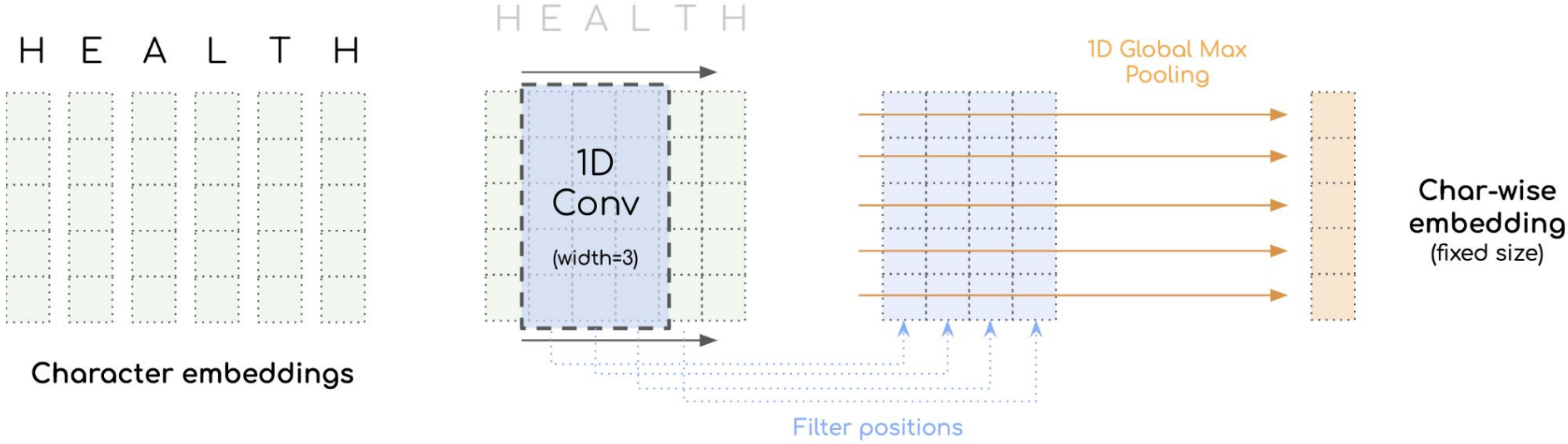
John Smith ⇒ PERSON
New York ⇒ LOCATION

John Snow LABS

NER-DL in Spark NLP

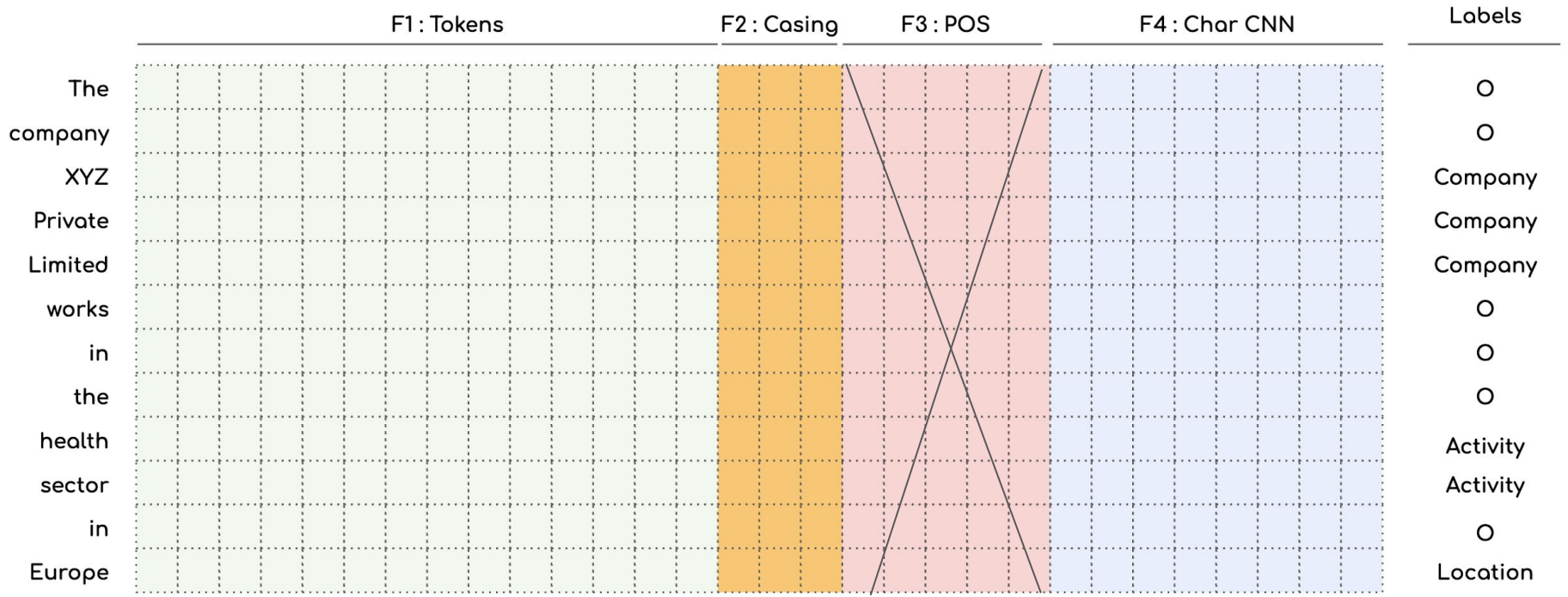


Char-CNN process, e.g. on the world “HEALTH”



NER-DL in Spark NLP

Char-CNN-BiLSTM



NER-DL in Spark NLP

Classification

	The	company	XYZ	Private	Limited	works	in	the	health	sector	in	Europe
GROUND TRUTH	O	O	Company	Company	Company	O	O	O	Activity	Activity	O	Location
PREDICTION	O	O	O	Company	Company	O	Group	O	O	Activity	O	Location

Class	O	Company	Location	Activity	Group
O	91969	546	295	1069	251
Company	569	3084	69	43	129
Location	137	48	1677	1	4
Activity	735	28	0	1329	0
Group	98	95	8	0	1185

Recall	Precision	F1
97,7 %	98,4 %	98 %
79,2 %	81,1 %	80,2 %
89,8 %	81,8 %	85,6 %
63,5 %	54,4 %	58,6 %
85,5 %	75,5 %	80,2 %

Introducing Spark NLP

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m2 , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . However , serum chemistry obtained six hours after presentation revealed her glucose was 186 mg/dL , the anion gap was still elevated at 21 , serum bicarbonate was 16 mmol/L , triglyceride level peaked at 2050 mg/dL , and lipase was 52 U/L . The β-hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L - the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again .

Clinical NER

Color codes:PROBLEM, TREATMENT, TEST,

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, Pathological_formation, Anatomical_system,

Anatomy NER

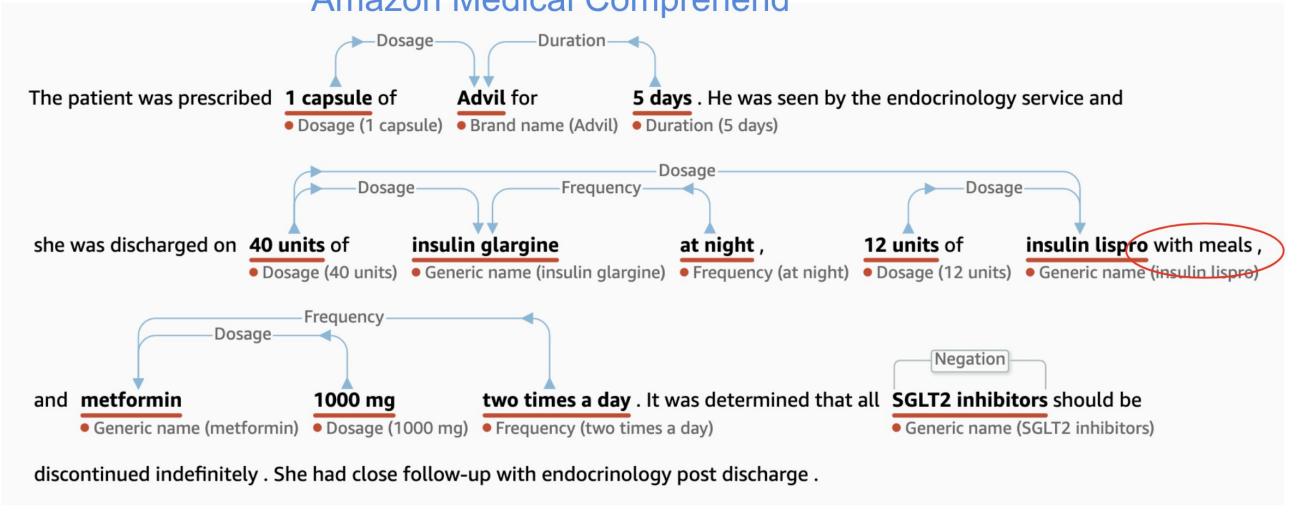
A . Record date : 2093-01-13 , David Hale , M.D . , Name : Hendrickson , Ora MR . # 7194334 Date : 01/13/93 PCP : Oliveira , 25 years-old , Record date : 2079-11-09 . Cocke County Baptist Hospital . 0295 Keats Street

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

PHI NER

NER Comparison with AWS Medical Comprehend

Amazon Medical Comprehend



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,

Clinical Named Entity Recognition (NER)

Dataset	Name	Entities
I2B2	ner_clinical	Problem, Test, Treatment
i2b2_med7+FDA	ner_posology	Drug, Dosage, Strength, Form, Route, Frequency, reason, ADE, Duration
BioNLP	ner_bionlp	Amino_acid, Anatomical_system, Cancer, Cell, Cellular_component, Developing_anatomical_structure, Gene_or_gene_product, Immaterial_anatomical_entity, Organ, Organism, Organism_subdivision, Organism_substance, Pathological_formation, Simple_chemical, Tissue, Multi-tissue_structure
n2c2	ner_deid_small	Name, Profession, Location, Age, Date, Contact, Id
n2c2+enriched	ner_deid	Patient, Hospital, Date, Bioid, Organization, Url, City, Street, Username, Device, Fax, Idnum, State, Location-other, Email, Zip, Medicalrecord, Profession, Phone, Country, Healthplan, Doctor, Age
risk_factors_2014	ner_riskfactors	PHI, Medication, CAS, Hypertension, Diabetes, Smoker, Hyperlipidemia, Obese, Family_Hist

NerDLModel	deidentify_dl
NerDLModel	ner_anatomy
NerDLModel	ner_bionlp
NerDLModel	ner_cellular
NerDLModel	ner_clinical
NerDLModel	ner_deid_enriched
NerDLModel	ner_deid_large
NerDLModel	ner_diseases
NerDLModel	ner_drugs
NerDLModel	ner_healthcare
NerDLModel	ner_jsl_enriched
NerDLModel	ner_jsl
NerDLModel	ner_posology_large
NerDLModel	ner_posology_small
NerDLModel	ner_posology
NerDLModel	ner_risk_factors

John Snow LABS

Clinical Named Entity Recognition (NER)

Table 1: Test Micro F1 scores (excluding O's) for medical NER data sets with clinical embeddings and Glove 6B embeddings under the same settings with our implementation. Strict (exact match) evaluation is used while comparing the metrics.

Dataset	SOTA	Clinical Embeddings	Glove 6B
2010 i2B2/VA	0.8684	**0.8693**	0.8546
2018 i2B2/Medication	0.8956	0.8931	0.8884
NCBI Disease	0.8860	0.8836	0.8533
BC5CDR Disease	0.8408	0.8383	0.8156
BC5CDR Chem	0.9331	0.9154	0.8945
BC4CHEMD	0.9114	**0.9352**	0.9202
Linnaeus	0.8702	0.8611	0.8156
Species800	0.7498	**0.8102**	0.7840
BC2GM	0.8169	**0.8725**	0.8487
JNLPBA	0.7858	**0.8039**	0.7894
AnEm	0.9161	0.8881	0.8620
BioNLP-CG	0.7674	**0.8571**	0.8337

Part - II

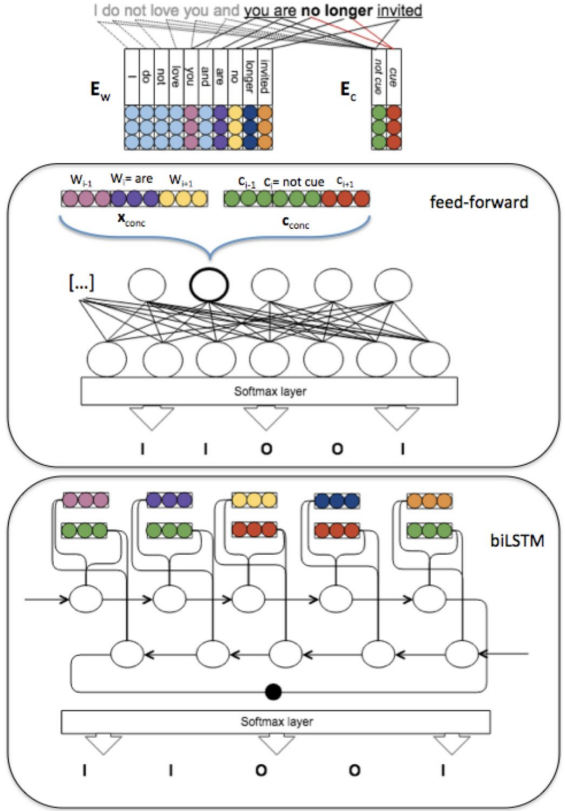
- Assertion status detection

Clinical Assertion Model

Prescribing sick days due to diagnosis of **influenza**.	*Present*
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.	*Conditional*
Jane's **RIDT** came back clean.	*Absent*
Jane is at risk for **flu** if she's not vaccinated.	*Hypothetical*
There was a **dense hemianopsia** on the left side.	*Present*

F-Score	Dataset	Task
94.17%	4th 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("xxxxx")

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

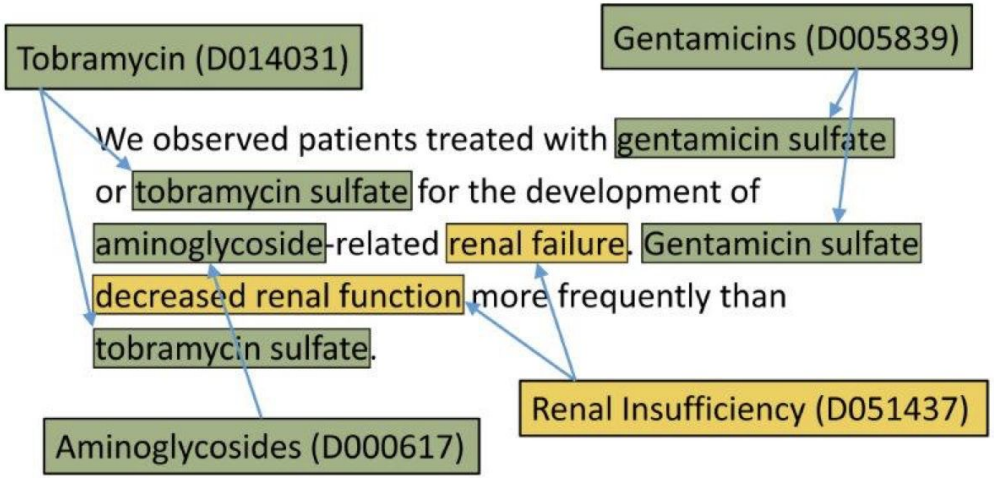
```
['A .',
 'Record date : <DATE>  , <DOCTOR> , M.D .',
 ', Name : <PATIENT> , <PATIENT> MR .',
 '# <MEDICALRECORD> Date : <DATE> PCP : <DOCTOR> , 25
month years-old , Record date : <DATE> .',
 '<HOSPITAL> .',
 '<STREET>']
```

```
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
- Keyword Extraction (YAKE)

Entity Resolution



"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

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

Drug Entities and RxNorm Codes

	ner	entity	code	resolved_text	alternative_codes
0	Advil	DRUG	352893	phoslo gelcap	1941952:|:1318187:|:827207:|:19711
1	SGLT2 inhibitors	DRUG	1431605	mao inhibitors	836:|:1431707:|:1430896:|:1431605
2	metformin	DRUG	607999	metformin and pioglitazone	614348:|:607999:|:1431025:|:729717
3	insulin lispro	DRUG	1652237	insulin lispro 2 0 0 unt / ml	343263:|:343663:|:343262:|:343264
4	insulin glargine	DRUG	1858994	insulin glargine and lixisenatide	1858994:|:1858994:|:1858994:|:1727493

A 72-year-old man with a history of diabetes mellitus , hypertension , and hypercholesterolemia self-palpated a left submandibular lump in 2012 . Complete blood count (CBC) in his internist's office showed solitary leukocytosis (white count 22) with predominant lymphocytes for which he was referred to a hematologist . Peripheral blood flow cytometry on 04/11/12 confirmed chronic lymphocytic leukemia (CLL)/small lymphocytic lymphoma (SLL): abnormal cell population comprising 63% of CD45 positive leukocytes , co-expressing CD5 and CD23 in CD19-positive B cells . CD38 was negative but other prognostic markers were not assessed at that time .

Problem Entities and ICD10 Codes

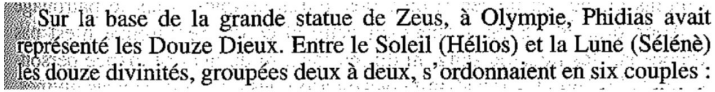
	ner	entity	code	resolved_text
0	CLL)/small lymphocytic lymphoma	PROBLEM	C880	Waldenstrom macroglobulinemia
1	solitary leukocytosis	PROBLEM	R911	Solitary pulmonary nodule
2	hypertension	PROBLEM	I150	Renovascular hypertension
3	abnormal cell population	PROBLEM	R978	Other abnormal tumor markers
4	diabetes mellitus	PROBLEM	P702	Neonatal diabetes mellitus
5	hypercholesterolemia	PROBLEM	E7801	Familial hypercholesterolemia
6	chronic lymphocytic leukemia	PROBLEM	E063	Autoimmune thyroiditis
7	a left submandibular lump	PROBLEM	L02422	Furuncle of left axilla
8	predominant lymphocytes	PROBLEM	C8107	Nodular lymphocyte predominant Hodgkin lymphoma, spleen

Part - VI

- Spark OCR

Maximizing Text Recognition Accuracy with Image Transformers in Spark OCR

Spark OCR



Uploaded Image.

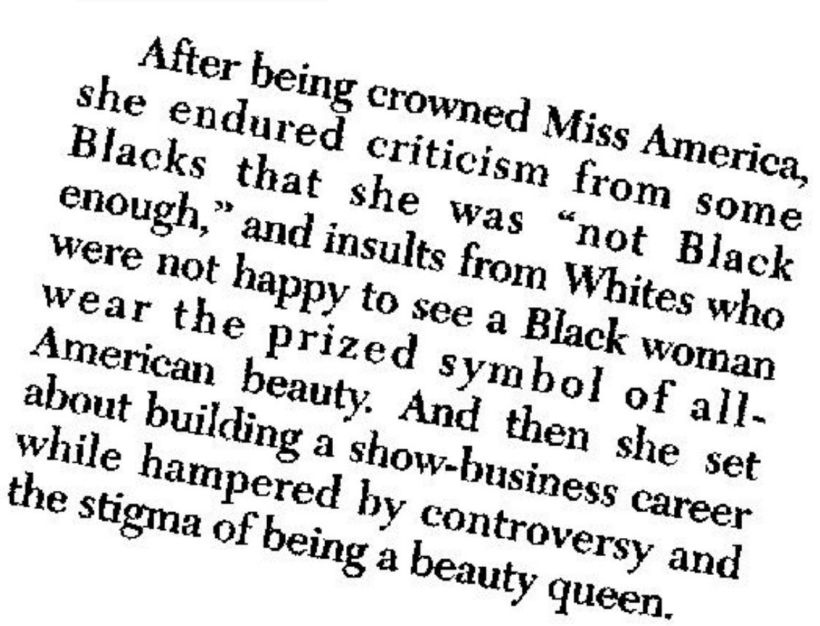
Converted Text:

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

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

Converted Text:

Digital ‘ Image Processing



Uploaded Image.

Converted Text:

After being crowned Miss America, she endured criticism from some Blacks that she was “not Black cnough,” and insults from Whites who were not happy to see a Black woman wear the prized Symbol of al]. 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:

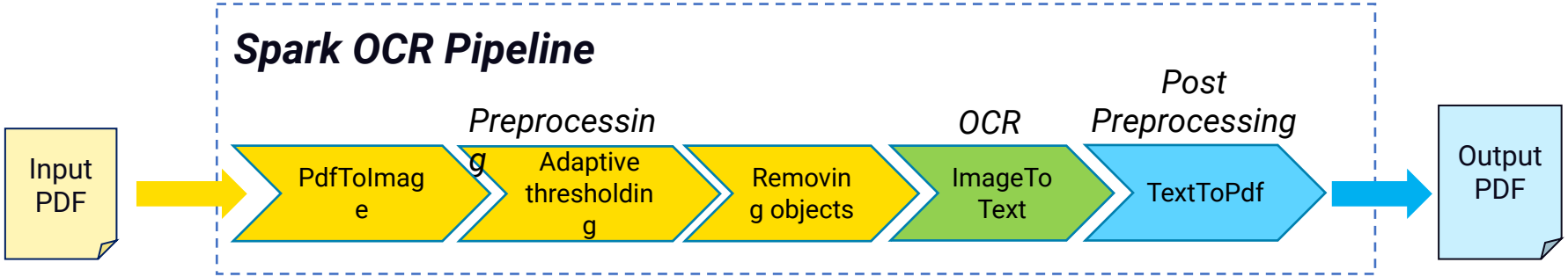
- **PdfToText** – extract text from selectable PDF
- **PdfToImage** – render each page as image
- **ImageToPdf** – store image to PDF format
- **TextToPdf** – render text with positions to PDF format
- **PdfDrawRegions** – draw regions to existing PDF

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



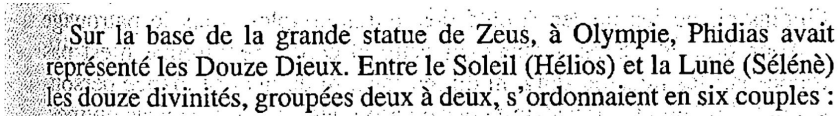
Some example of OCR pipelines

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

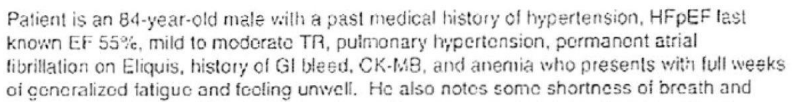
Image Preprocessing

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 (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

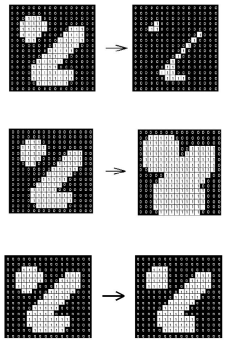
Bad quality of image


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

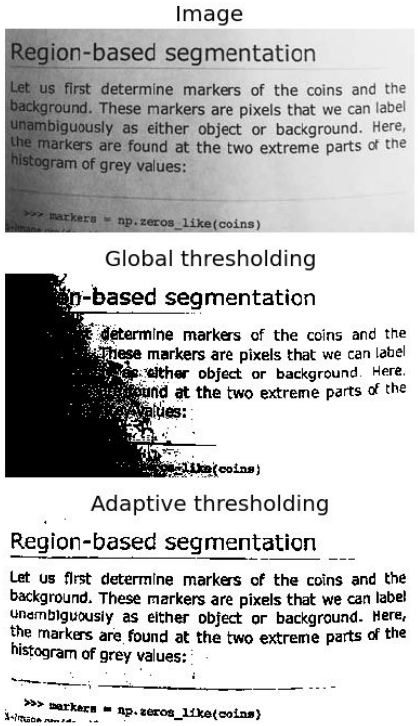


Thresholding

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

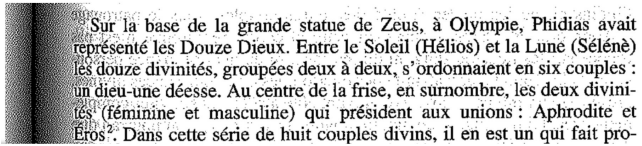
If that background is relatively ***uniform***, then you can ***use a global threshold*** value to binarize the image by pixel-intensity.

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.

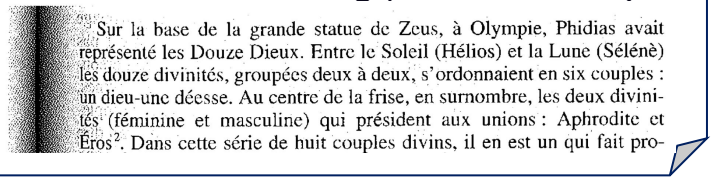


Thresholding with Spark OCR

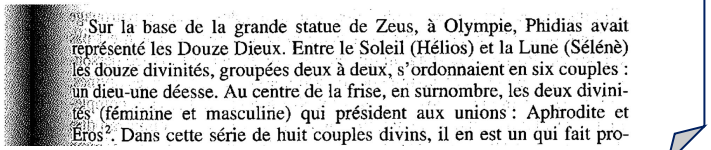
Source image


Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (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 Éros[2]. 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 (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 Éros[2]. Dans cette série de huit couples divins, il en est un qui fait pro-

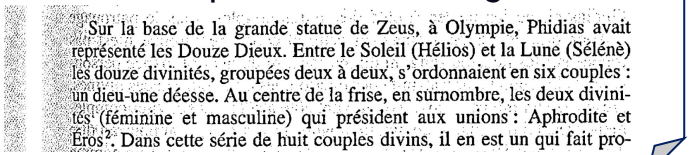
Global thresholding *(threshold = 120)*


Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (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 Éros[2]. 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)
```

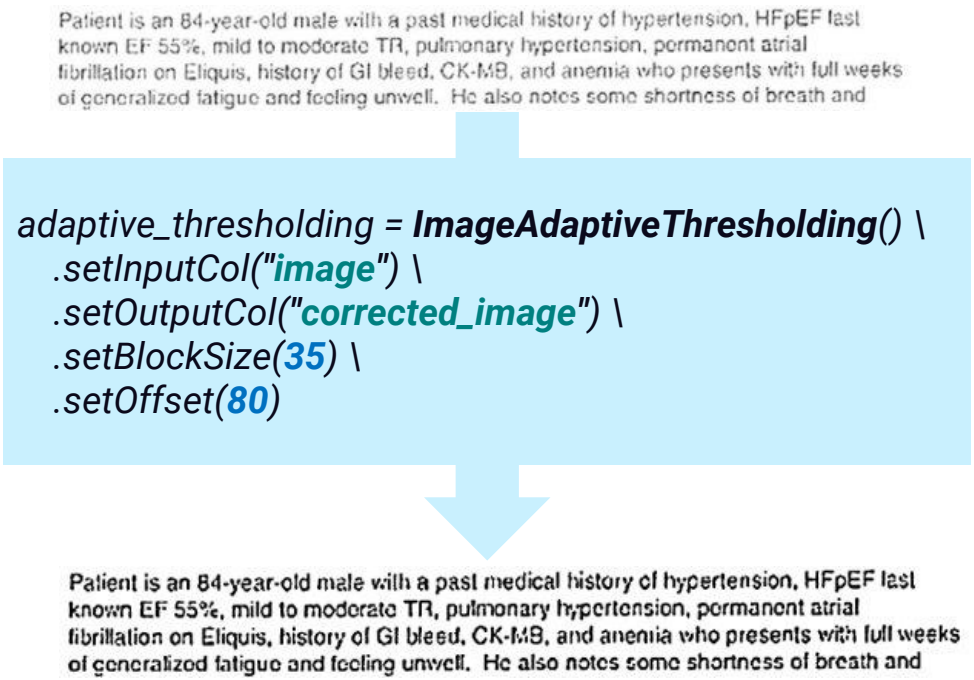
For various images usually provides better results

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 (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 Éros[2]. 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



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.

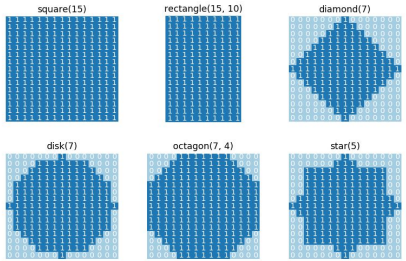
Methods:

- GAUSSIAN
- MEAN
- MEDIAN

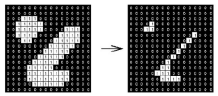
Morphological Operations

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.

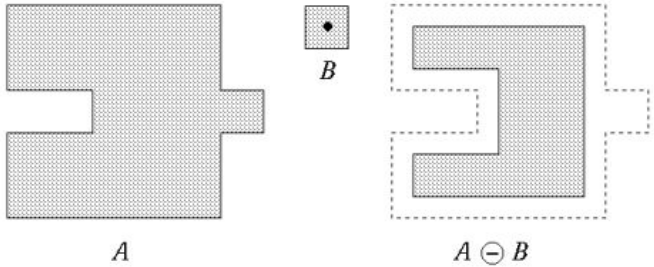
The structuring element used in practice is generally much smaller than the image, often a 3x3 matrix.



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.,

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

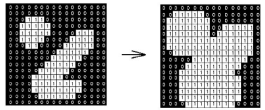
Source image

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

Image after applying erosion

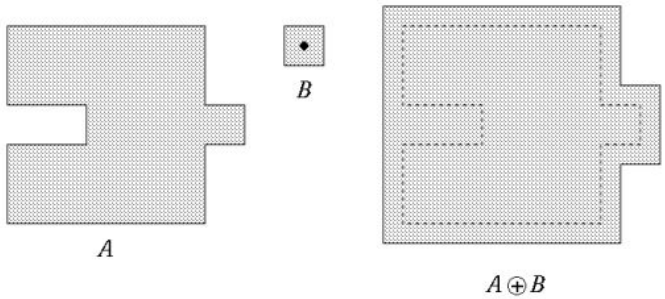
Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

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

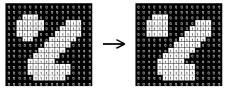
Patient is an 84-year-old male with a past medical history of hypertension, HFpEF last known EF 55%, mild to moderate TR, pulmonary hypertension, permanent atrial fibrillation on Eliquis, history of GI bleed, CK-MB, and anemia who presents with full weeks of generalized fatigue and feeling unwell. He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

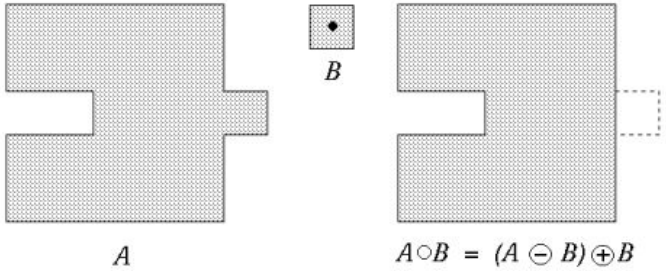
Image after applying dilation

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*.





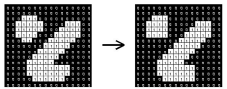
Source image

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

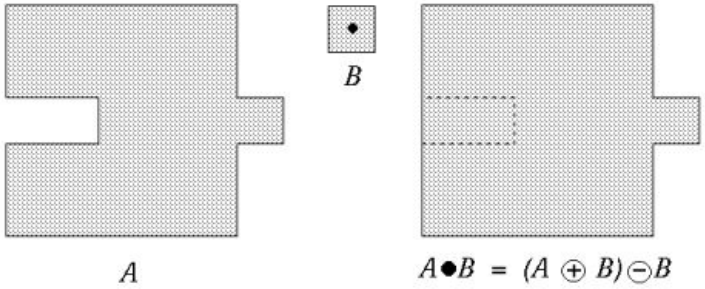
Image after applying opening

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

ImageMorphologyOperation transformer

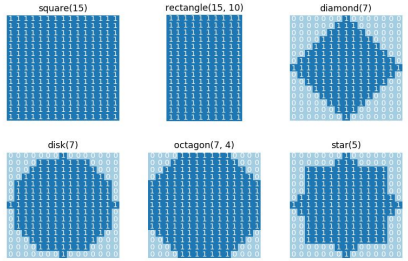
```
ImageMorphologyOperation() \
    .setKernelShape(KernelShape.SQUARE) \
    .setKernelSize(3) \
    .setOperation(MorphologyOperationType.EROSION) \
    .setInputCol("image") \
    .setOutputCol("corrected_image")
```

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

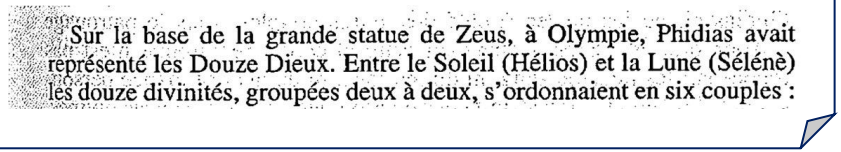
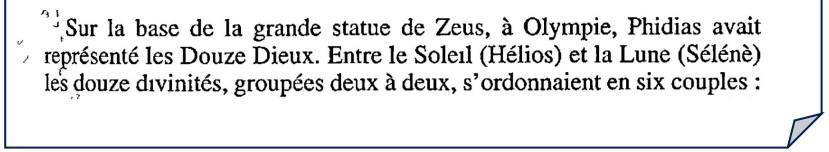
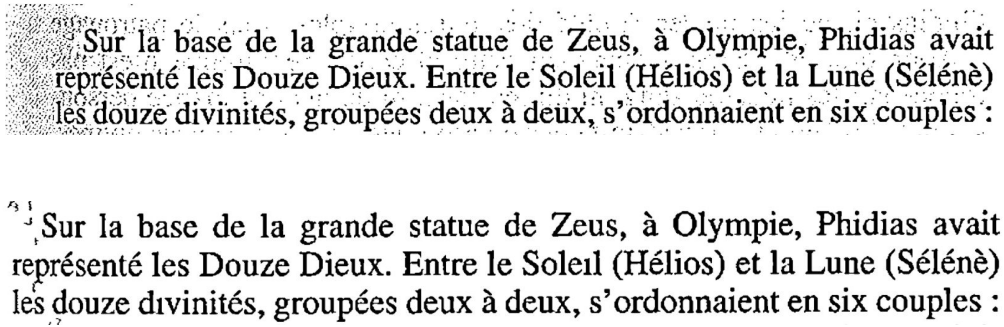

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

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 (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples :

ImageRemoveObject transformer

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

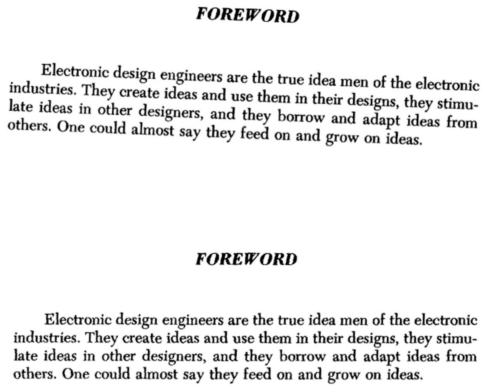


Param name	Type	Default	Description
minSizeFont	int	10	Min size of font in pt.
minSizeObject	int	None	Min size of object which will keep on image [*].
minSizeHole	int	None	Min size of hole which will keep on image [*].
maxSizeObject	int	None	Max size of object which will keep on image [*].

[*] : *None* value disables removing objects

Skew correction

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

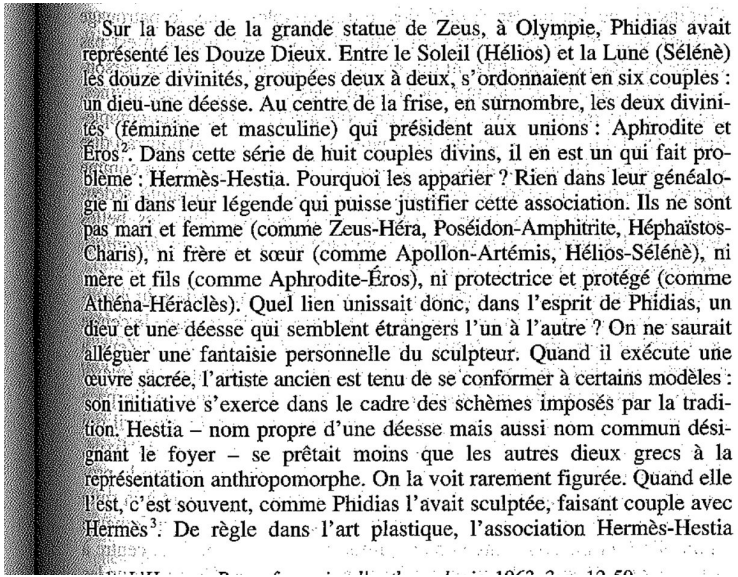


Param name	Type	Description
rotationAngle	double	rotation angle
automaticSkewCorrection	boolean	enables/disables adaptive skew correction

OCR in action

Tesseract

F1 score - 0.2610

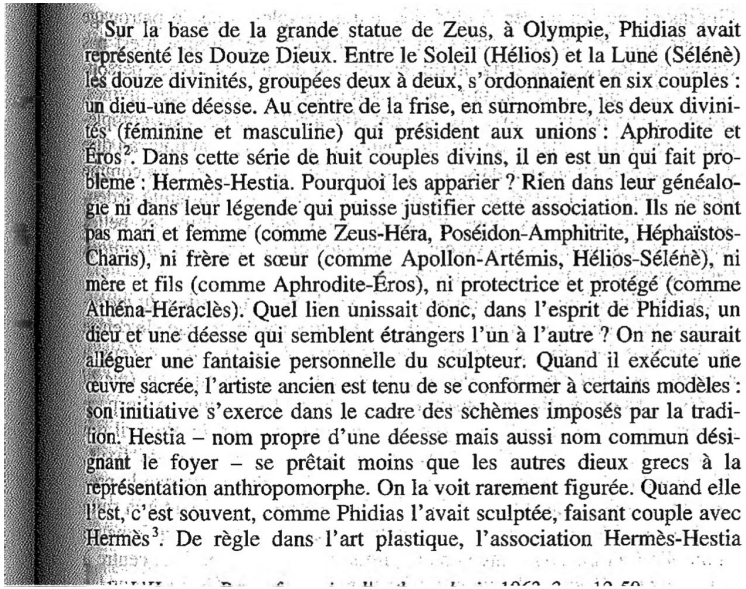


Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Éros[2]. 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-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 Hermès[3]. De règle dans l'art plastique, l'association Hermès-Hestia

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 sccur (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 prétait moins que les autres dieux grecs a la
Stésentation anthropomorphe. On la voit rarement figurée: Quand elle
est souvent, comme Phidias |' 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 sccur (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 prétait moins que les autres dieux grecs a la
Stésentation anthropomorphe. On la voit rarement figurée: Quand elle
est souvent, comme Phidias |' avait sculptée; faisant couple avec
3s°: De régle dans- l'art 'Plastique, Passociation Henmes-Hestia
—_ _ . wt a. wo et anhserh A an om *

ABBYY FineReader

F1 score - 0.8972

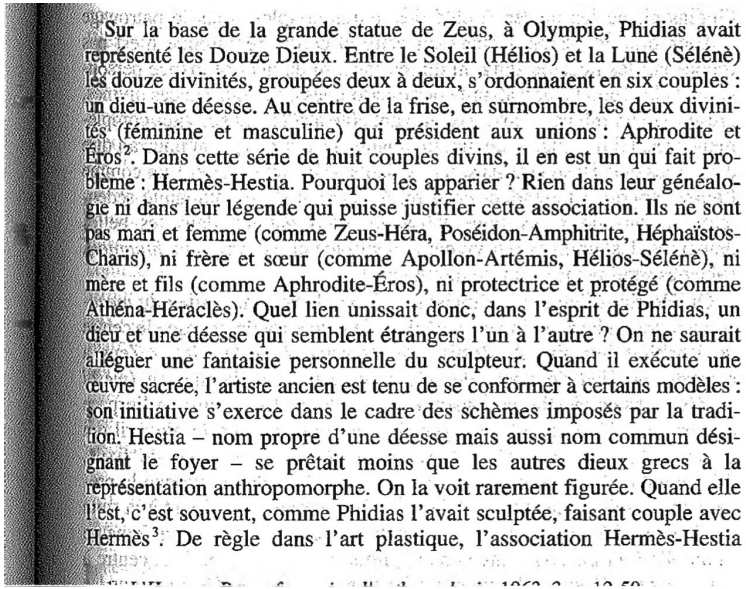


Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Éros[2]. 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 Hermès[3]. De règle dans l'art plastique, l'association Hermès-Hestia

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait
présenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sëlénè)
& douze divinités, groupées deux à deux, s'ordonnaient en six couples :
n dieu-une déesse. Au centre de la frise, en surnombre, les deux divini-
(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 généaloJfêfrïï dans
leur légende qui puisse justifier cette association: Ils rie sont
M m ari et femme (comme Zeus-Héra, Poséidon-Amphitrite,
HéphaïstosÊHaris), ni frère et sœur (comme Apollon-Artémis, Hélios-Séléribè),
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 :
|®ihitiative s'exerce dans le cadre des schèmes imposes 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
èprésentation anthropomorphe. On la voit rarement figurée. Quand elle
'est,5c'est souvent, comme Phidias l'avait sculptée, faisant couple avec
érmës3. De règle dans l'art plastique, l'association Hermès-Hestia
RpvÆ îi. -4'i 1 T I T i A r o a -i a

AWS Text Ract

F1 score - 0.9323



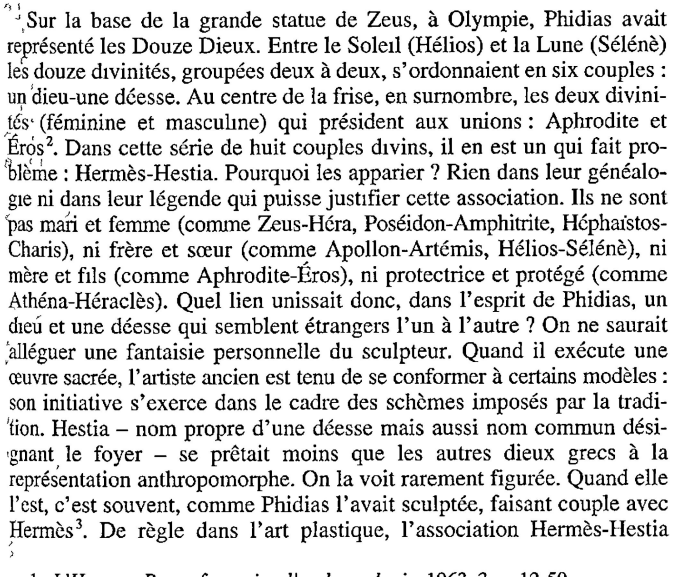
Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Éros[2]. 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-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 Hermès[3]. De règle dans l'art plastique, l'association Hermès-Hestia

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, groupees deux a deux, 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 nl 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 schemes 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

F1 score - 0.9812

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 (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 Éros[2]. 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-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 Hermès[3]. De règle dans l'art plastique, l'association Hermès-Hestia

Sur la base de la grande statue de Zeus, a Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux a deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de 1a 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 probléme : Hermés-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, Héphaistos-Charis), ni frére et sccur (comme Apollon-Artémis, Hélios-Séléné), ni mére et fils (comme Aphrodite-Eros), ni protectrice et protégé (comme Athéna-Héraclés). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un a l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une ceuvre sacrée, l'artiste ancien est tenu de se conformer a certains modeéles : son initiative s'exerce dans le cadre des schémes imposés par la tradition. Hestia — nom propre d'une déesse mais aussi nom commun désignant le foyer — se prétait moins que les autres dieux grecs a la représentation anthropomorphe. On la voit rarement figurée. Quand elle Vest, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermés*, De régle dans l'art plastique, l'association Hermés-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/winning-in-health-care-ai-with-small-data/#1b2fc2555664
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

John Snow LABS


NOW ANNOUNCING
NLP
SUMMIT
Applied Natural
Language Processing
Boston, Oct 27-28 | San Francisco, Nov 17-18