John Snow LABS

Spark NLP for Healthcare Data Scientists

April 9, 2020

Veysel Kocaman
Sr. Data Scientist
veysel@johnsnowlabs.com

Welcome !

60 min	Overview and key concepts in Spark NLP Cleaning medical text: Normalization, stop-words, clinical POS, spell checking Common medical NLP use cases Clinical named entity recognition Assertion status detection
30 min	Break / Q&A
60 min	Medical Entity Resolution (ICD, RxNorm, SNOMED) Healthcare data De-identification
30 min	Break / Q&A
60 min	Object Character Recognition (OCR) Building and configuring a Spark OCR pipeline Running OCR on PDF with text, PDF with images, and image files Image pre-processing and document enhancement Unifying Spark OCR & Spark NLP pipelines

Setup

HOUSEKEEPING:

[How to ask questions on Teams]
[How to mute / unmute / screen share]
[Notify that recording starts from next slide]

RUNNING CODE:

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

[How to set up Google Colab]

BOOKMARK:

DOCS : nlp.johnsnowlabs.com/docs/en/concepts
SLACK: spark-nlp.slack.com
CODE : github.com/johnsnowlabs/spark-nlp

Part - I

- ❖ Overview and key concepts in Spark NLP
- ❖ Cleaning medical text: Normalization, stop-words, clinical POS, spell checking
- ❖ Common medical NLP use cases
- ❖ Clinical named entity recognition
- ❖ Assertion status detection

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

"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

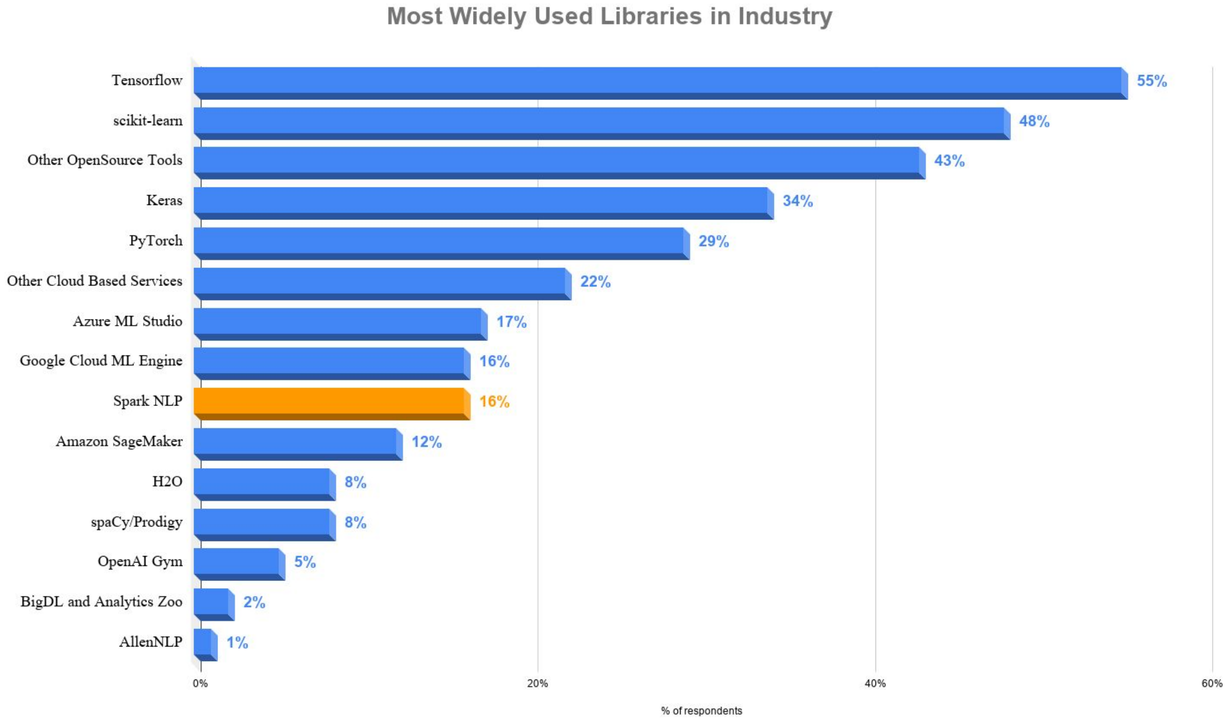
TRUSTED BY

Microsoft | Google | amazon | intel
IBM | databricks | verizon | indeed
Capital One | viacom | McKESSON | MERCK
Roche | selectdata | UiPath Reboot Work. | ASCO CANCER·LINQ DISCOVERY
cnrs | Imperial College London | Georgia Tech | STANFORD UNIVERSITY

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

Java

python™

Scala

R

NVIDIA®

intel® XEON® inside™

APACHE Spark™

databricks®

CLOUDERA

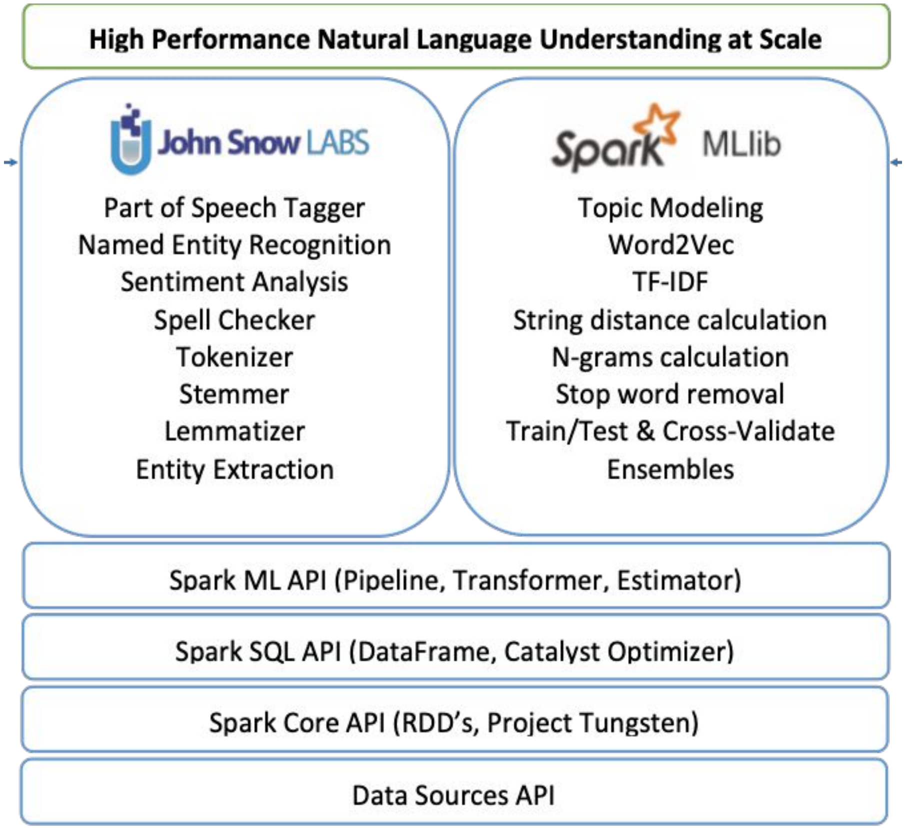
Azure

aws

BUILT ON THE SHOULDERS OF SPARK ML

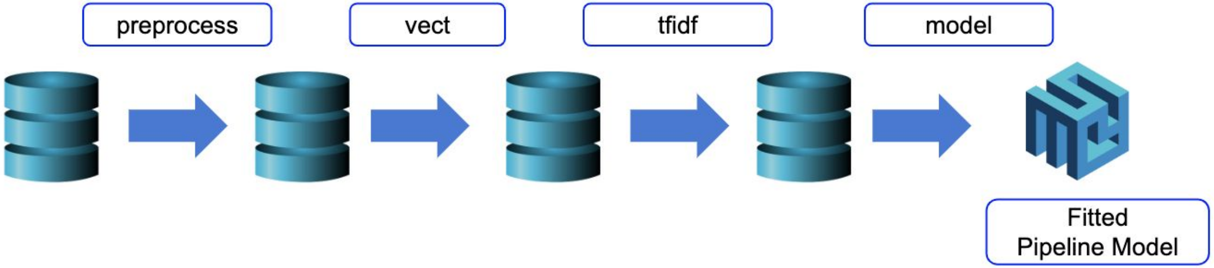
- Reusing the Spark ML Pipeline
  - Unified NLP & ML pipelines
  - End-to-end execution planning
  - Serializable
  - Distributable

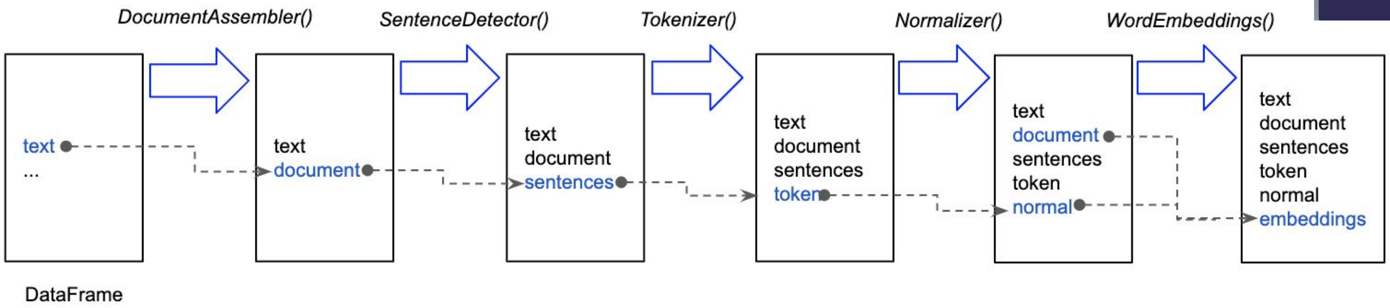
- Reusing NLP Functionality
  - TF-IDF calculation
  - String distance calculation
  - Topic modeling
  - Distributed ML algorithms



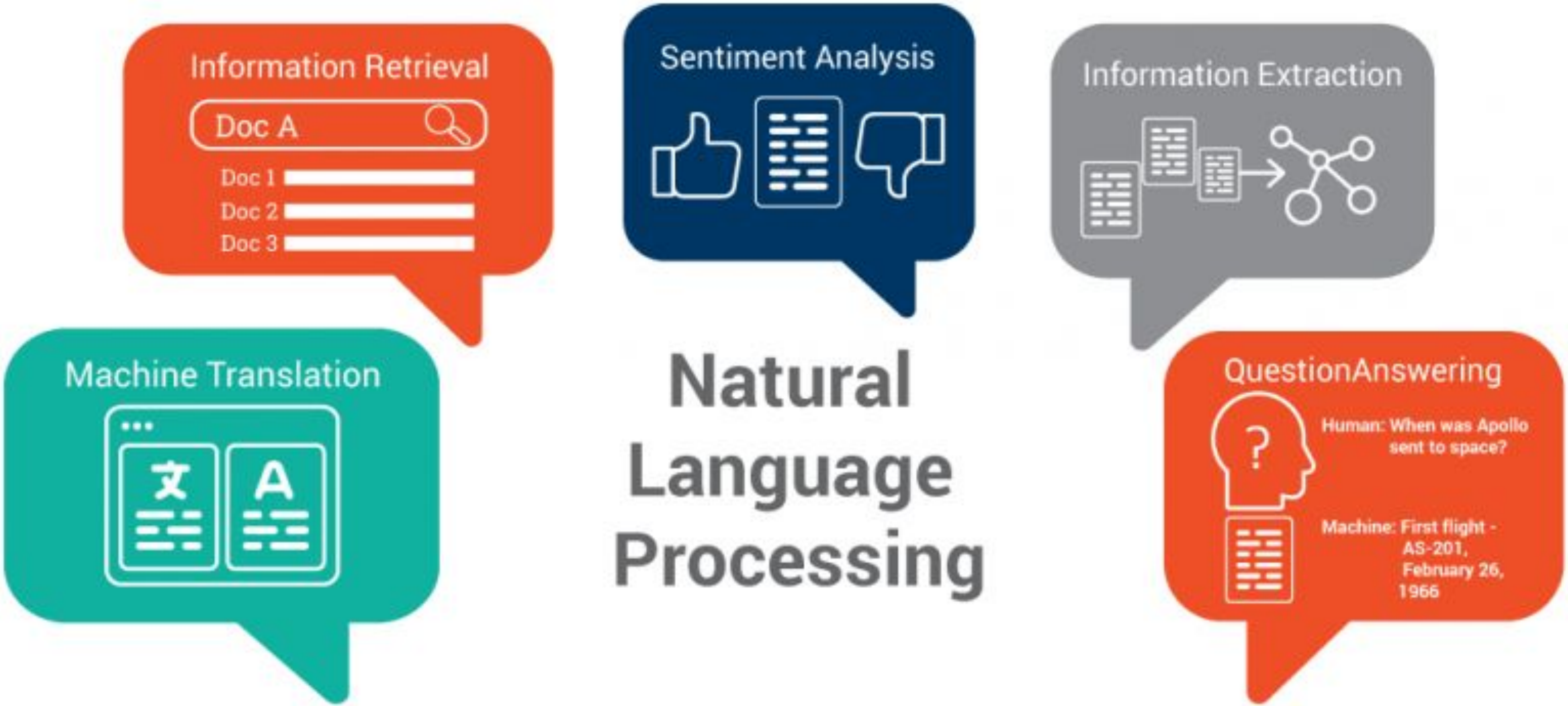
Introducing Spark NLP

Pipeline of annotators






```python
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

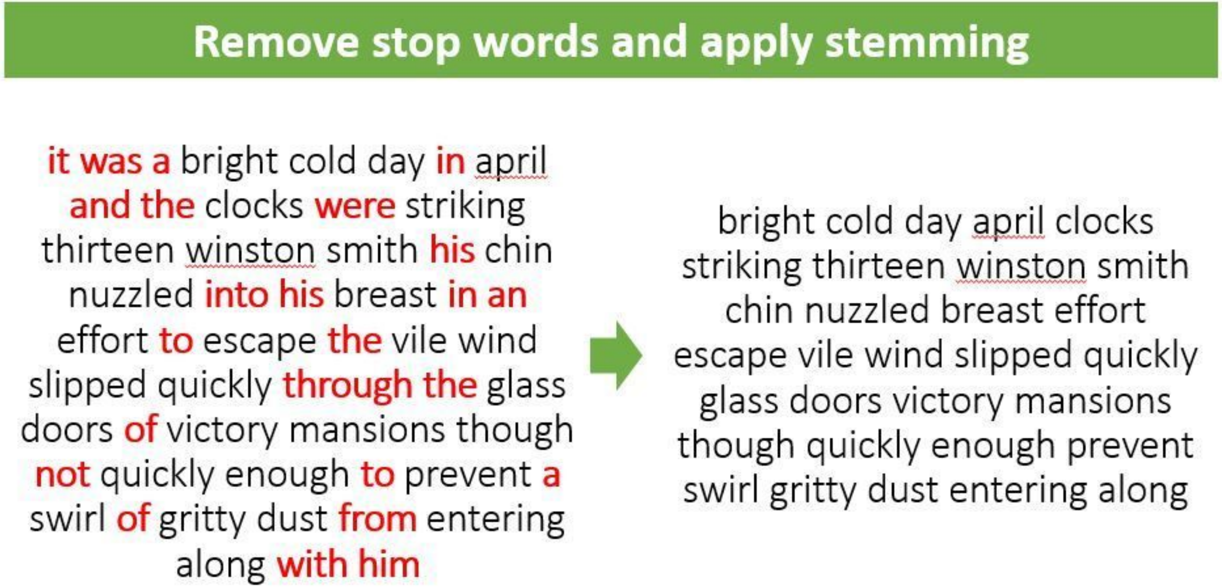
Find the **stem** of each word.

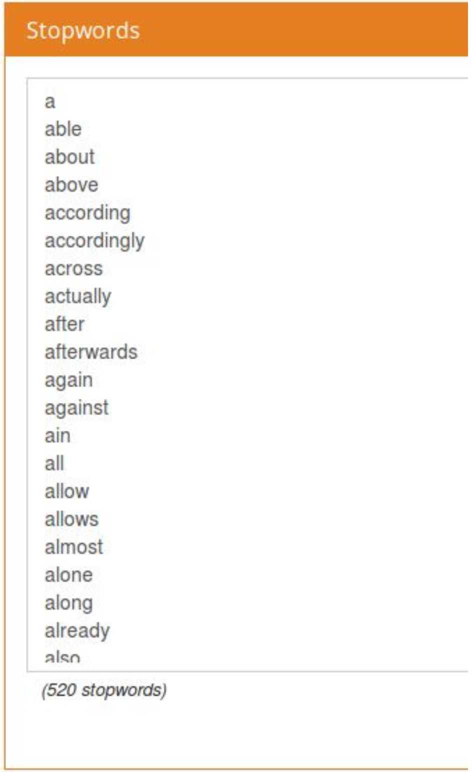
Uses rules: e.g, remove common suffixes.

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





- 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.

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

NORMALIZATION

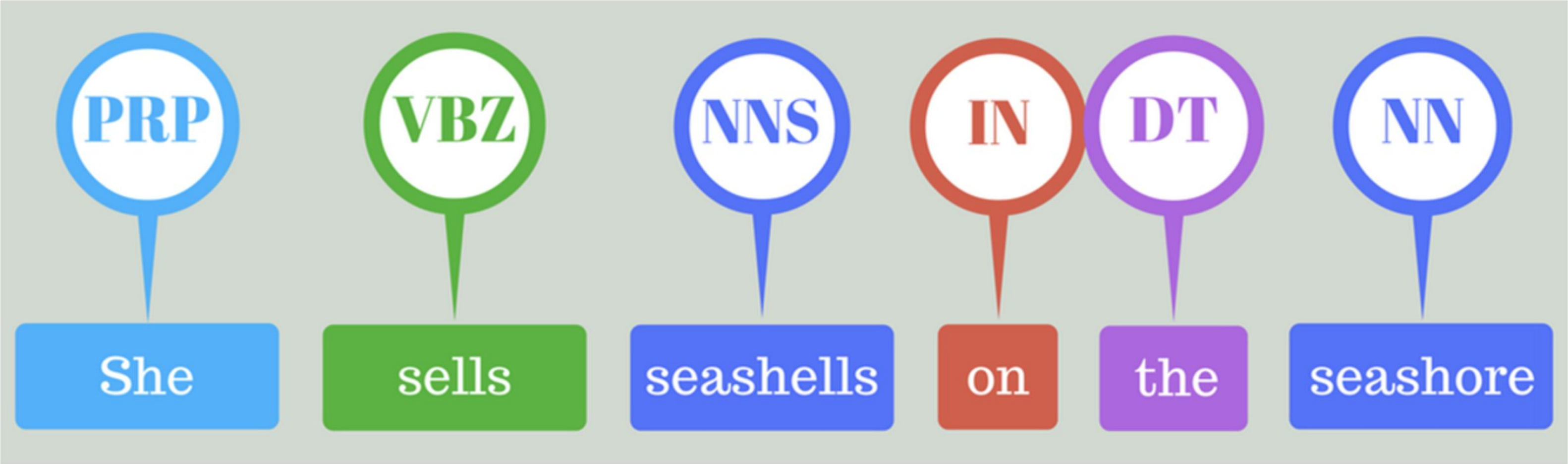
Remove or replace undesirable characters or regular expressions:

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

Spark NLP also comes with a Slang normalizer:

PART OF SPEECH TAGGING

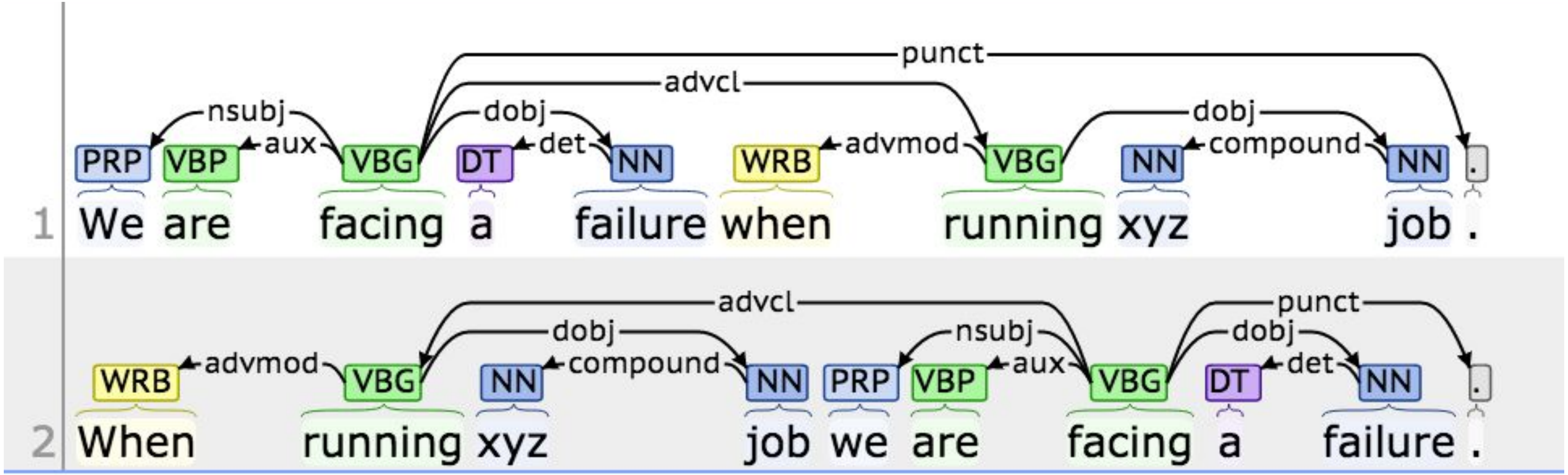
Often useful for recognizing named entities or word relationships.



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.

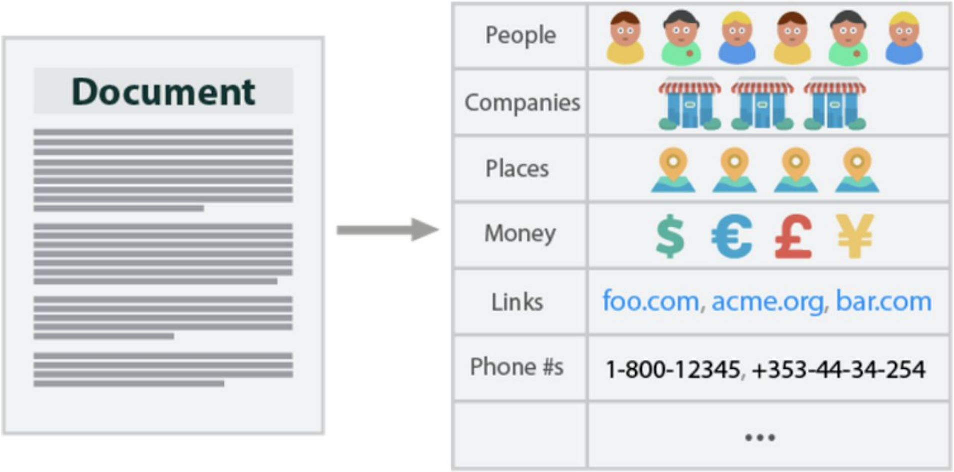
DEPENDENCY PARSING

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.

Coding ...

Open **1. Spark NLP Basics** and **2. Text Preprocessing with SparkNLP** notebooks in **Colab**

(click on Colab icon or open in a new tab)

https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/Certification_Trainings/Public/1.SparkNLP_Basics.ipynb

https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/Certification_Trainings/Public/2.Text_Preprocessing_with_SparkNLP(Annotators_Transformers).ipynb

John Snow LABS

Open in Colab

Spark NLP Basics

0. Colab Setup

```
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

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

```
openjdk version "1.8.0_242"
OpenJDK Runtime Environment (build 1.8.0_242-8u242-b08-0ubuntu3~18.04-b08)
OpenJDK 64-Bit Server VM (build 25.242-b08, mixed mode)
     |████████████████████████████████| 215.7MB 59kB/s
     |████████████████████████████████| 204kB 41.9MB/s
  Building wheel for pyspark (setup.py) ... done
     |████████████████████████████████| 112kB 2.7MB/s
```

John Snow LABS

Open in Colab

2. Text Preprocessing with Spark NLP

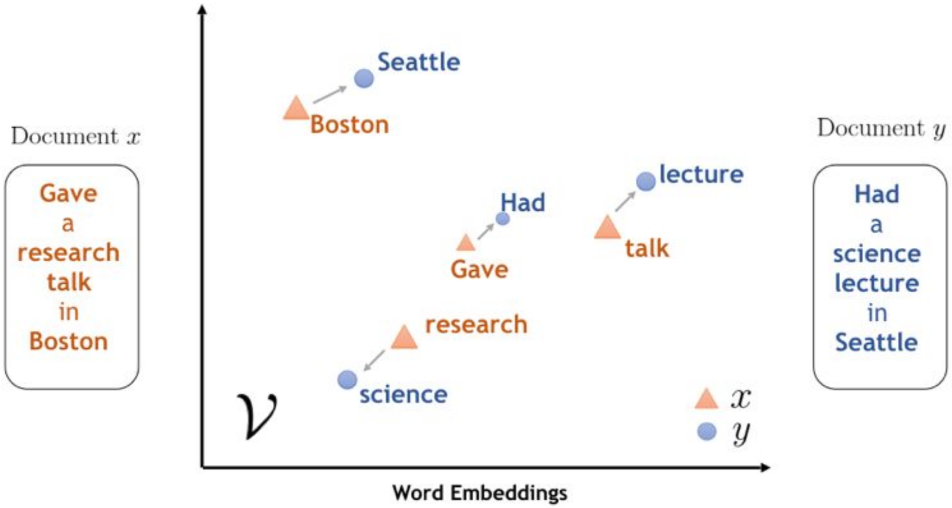
0. Colab Setup


```
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

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


```
openjdk version "1.8.0_242"
OpenJDK Runtime Environment (build 1.8.0_242-8u242-b08-0ubuntu3~18.04-b08)
OpenJDK 64-Bit Server VM (build 25.242-b08, mixed mode)
     |████████████████████████████████| 215.7MB 65kB/s
     |████████████████████████████████| 204kB 48.9MB/s
  Building wheel for pyspark (setup.py) ... done
     |████████████████████████████████| 112kB 2.8MB/s
```

1. Annotators and Transformer Concepts

Word & Sentence Embeddings

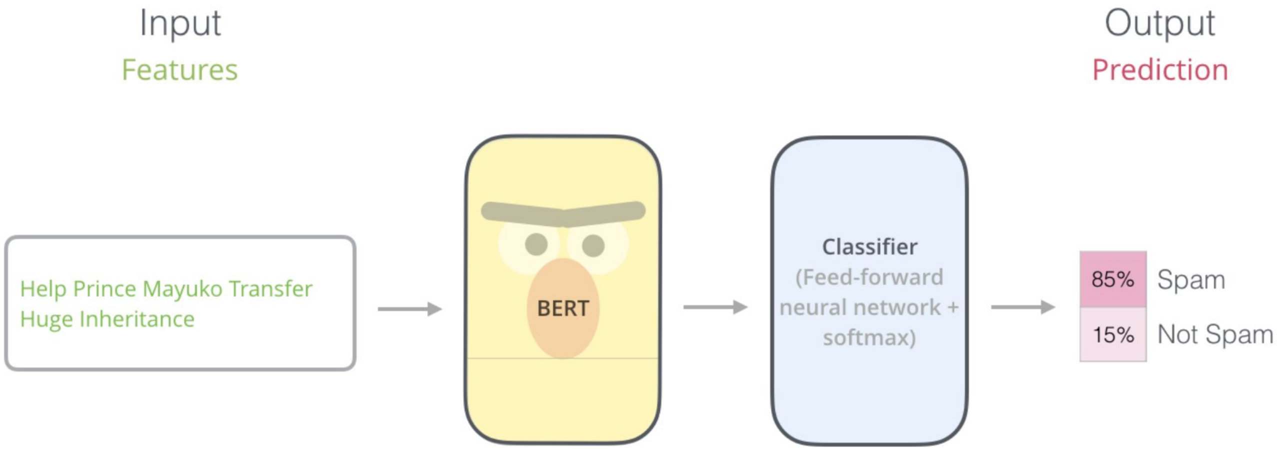



```
In [9]: doc[3].vector
Out[9]: array([ 0.037103 , -0.31259  , -0.17857  ,  0.30001  ,  0.078154 ,
         0.17958  ,  0.12048  , -0.11879  , -0.20601  ,  1.2849   ,
        -0.20409  ,  0.80613  ,  0.34344  , -0.19191  , -0.084511 ,
         0.17339  ,  0.042483 ,  2.0282   , -0.16278  , -0.60306  ,
        -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  ,
```

- 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.

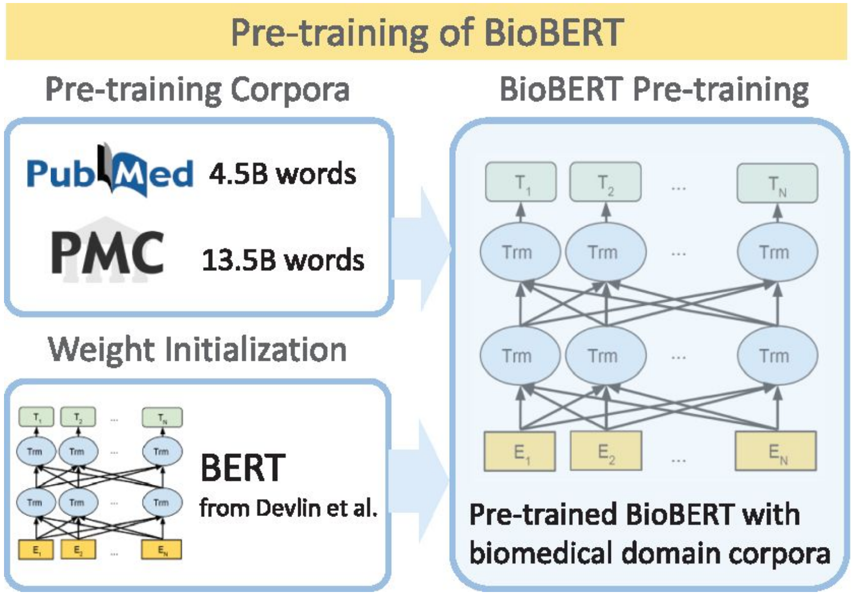
Word & Sentence Embeddings

Glove (100, 200, 300)	ELMO (512, 1024)	BERT (768d)	Universal Sentence Encoders (512)



Clinical Word Embeddings

Clinical Glove (200)	ICDO Glove (200)	Bio BERT (< 2.4)	Clinical BERT (< 2.4)
		Pubmed + PMC	Fine tuned Pubmed + PMC + Discharge summaries



PubMed abstracts and PMC full-text articles

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

Spark NLP in Healthcare

Output from one of the NLP libraries - MIMIC-III dataset

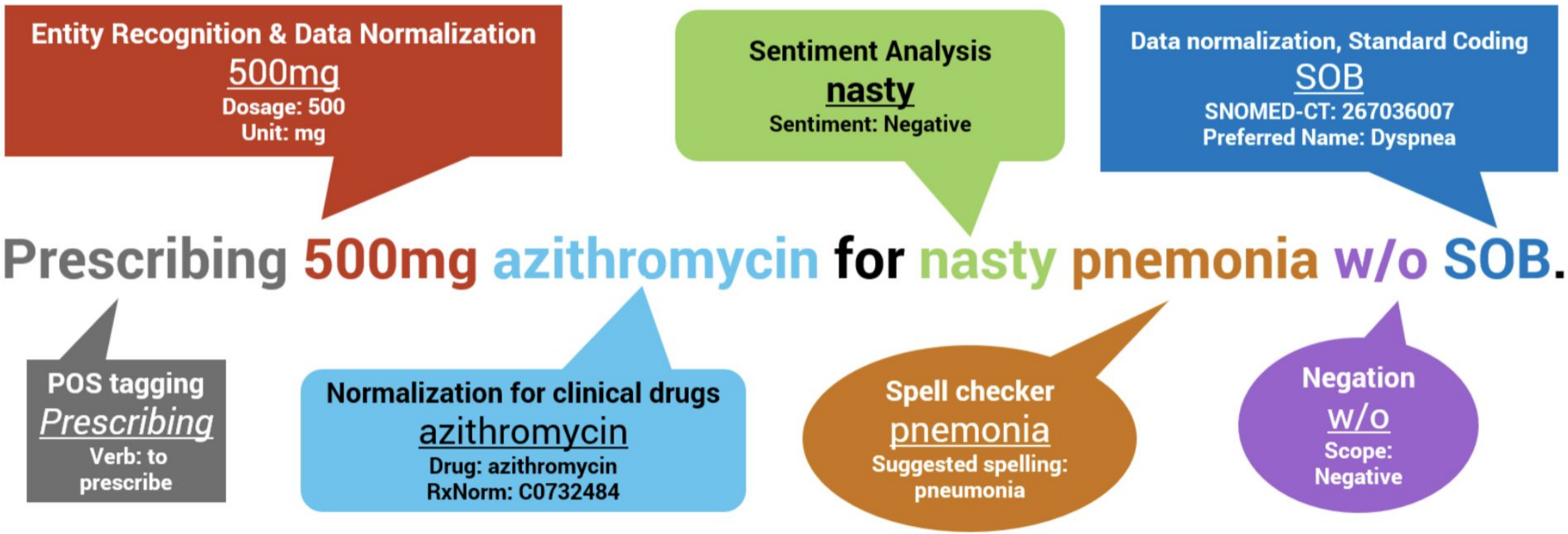
(an openly available dataset developed by the MIT Lab for Computational Physiology)

"(admission): 50.4 kg\\n Height: 61 Inch\\n ICP: 7 (1 - 14) mmHg\\n Total In:\\n 3,279 mL\\n 911 mL\\n PO:\\n Tube feeding:\\n 243 mL\\n 237 mL\\n IV Fluid:\\n 2,827 mL\\n 624 mL\\n Blood products:\\n Total out:\\n 2,333 mL\\n 370 mL\\n Urine:\\n 2,330 mL\\n 370 mL\\n NG:\\n Stool:\\n Drains:\\n 3 mL\\n Balance:\\n 946 mL\\n 541 mL\\n Respiratory support\\n O2 Delivery Device: None\\n SPO2: 97%\\n ABG: ///26/\\n Physical Examination\\n General Appearance: No acute distress, Non communicative due to\\n language barrier\\n HEENT: PERRL, EOMI\\n Cardiovascular: (Rhythm: Regular)\\n Respiratory / Chest: (Expansion: Symmetric), (Breath Sounds: CTA\\n bilateral :), (Sternum: Stable)\\n Abdominal: Soft, Non-distended, Non-tender, Bowel sounds present\\n Left Extremities: (Edema: Absent), (Temperature: Warm), (Pulse -\\n Dorsalis pedis: Present), (Pulse - Posterior tibial: Present)\\n Right Extremities: (Edema: Absent), (Temperature: Warm), (Pulse -\\n Dorsalis pedis: Present), (Pulse - Posterior tibial: Present)\\n Skin: (Incision: Clean / Dry / Intact)\\n Neurologic: (Awake / Alert / Oriented: x 2), Follows simple commands,\\n Moves all extremities, Limited due to language barrier\\n Labs / Radiology\\n 275 K/uL\\n 9.8 g/dL\\n 134 mg/dL\\n 0.4 mg/dL\\n 26 mEq/L\\n 3.5 mEq/L\\n 15 mg/dL\\n 102 mEq/L\\n 137 mEq/L\\n 30.3 %\\n 8.8 K/uL\\n [image002.jpg]\\n [**2140-7-23**] 03:30 PM\\n [**2140-7-24**] 02:51 AM\\n [**2140-7-24**] 03:03 AM\\n [**2140-7-24**] 08:13 AM\\n [**2140-7-24**] 10:07 AM\\n [**2140-7-25**] 02:45 AM\\n [**2140-7-26**] 01:15 AM\\n [**2140-7-27**] 03:09 AM\\n [**2140-7-27**] 10:58 AM\\n [**2140-7-28**] 02:58 AM\\n WBC\\n 9.7\\n 10.3\\n 11.2\\n 7.7\\n 7.1\\n 8.8\\n Hct\\n 31.8\\n 32.6\\n 34.3\\n 33.3\\n 31.4\\n 30.3\\n Plt\\n [**Telephone/Fax (3) 8785**]\\n Creatinine\\n 0.5\\n 0.5\\n 0.5\\n 0.5\\n 0.5\\n 0.5\\n 0.4\\n TCO2\\n 26\\n 28\\n 29\\n Glucose\\n 168\\n 253\\n 147\\n 180\\n 92\\n 160\\n 194\\n 134\\n Other labs: PT / PTT / INR:11.6/25.8/1.0, CK / CK-MB / Troponin\\n T:54//<0.01, ALT / AST:25/32, Alk-Phos / T bili:87/,\\n Differential-Neuts:93.0 %, Lymph:5.3 %, Mono:1.0 %, Eos:0.5 %, Lactic\\n Acid:1.5 mmol/L, Ca:7.9 mg/dL, Mg:1.8 mg/dL, PO4:2.5 mg/dL\\n Assessment and Plan\\n AIRWAY, INABILITY TO PROTECT (RISK FOR ASPIRATION, ALTERED GAG, AIRWAY\\n CLEARANCE, COUGH), CVA (STROKE, CEREBRAL INFARCTION), HEMORRHAGIC ,\\n HYPERTENSION, BENIGN, [**Last Name 12**] PROBLEM - ENTER DESCRIPTION IN COMMENTS\\n Assessment and Plan: 69 yo F w/ left cerebellar thrombotic stroke,\\n hemorrhage, transtentorial herniation s/p EVD placement, surgical\\n decompression on [**7-22**], now w/ improved neuro exams\\n Neurologic: ICP monitor, Pain controlled, s/p crani for cerebellar\\n CVA, moves all 4, EVD clamped.

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.*

Spark NLP in Healthcare



Clinical Named Entity Recognition (NER)

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 .

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,

Label	Concept	Description
DOSAGE	1-2, sliding scale, taper, bolus, thirty (30) ml	The total amount of a drug administered
DRUG	aspirin, lisinopril, prednisone, vitamin b, flagyl	Generic or brand name of the medication
DURATION	for 3 days, 7 days, chronic, x5 days, for five more days	The length of time that the drug was prescribed for
FORM	tablet, capsule, solution, puff, adhesive patch, disk with device	A particular configuration of the drug which it is marketed for use
FREQUENCY	once a day, b.i.d., prn, q6h, hs, every six (6) hours as needed	The dosage regimen at which the medication should be administered
ROUTE	iv, p.o. (by mouth), gtt, nasal canula, injection,	The path by which the drug is taken into the body
STRENGTH	5mg, 100 unit/ml, 50mg/2ml, 0.05%, 25-50mg	The amount of drug in a given dosage

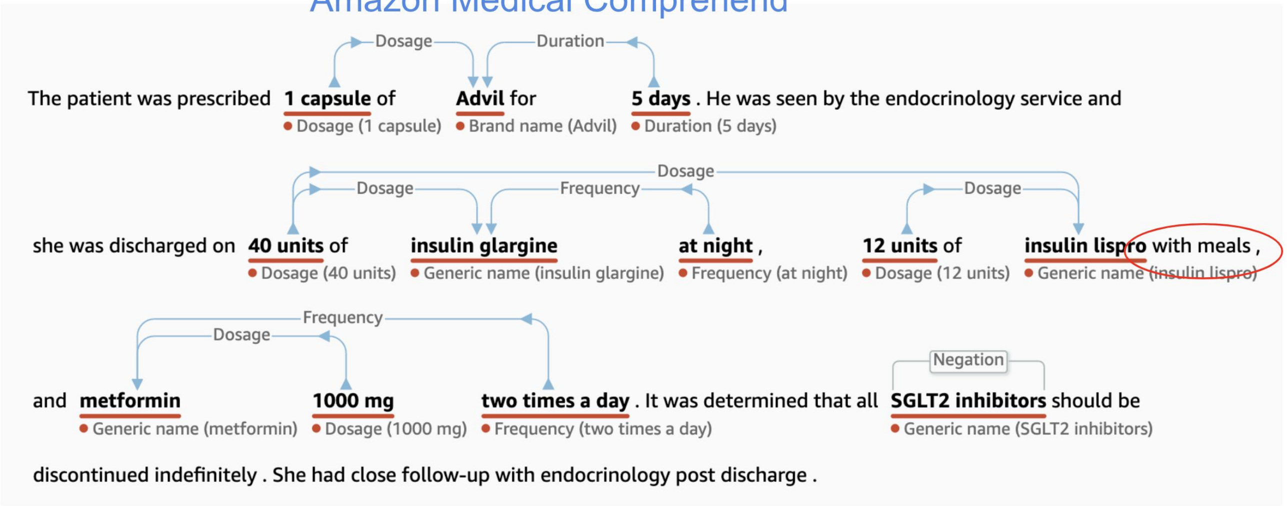
A . Record date : 2093-01-13 , David Hale , M.D . , Name : Hendrickson , Ora MR . # 7194334 Date : 01/13/93 PCP : Oliveira , 25 years-old , Record date : 2079-11-09 . Cocke County Baptist Hospital . 0295 Keats Street

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

John Snow LABS

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	State
I2B2	ner_clinical	Problem, Test, Treatment	Pretrained
i2b2_med7+FDA	ner_posology	Drug, Dosage, Strength, Form, Route, Frequency, reason, ADE, Duration	Pretrained
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	Pretrained
n2c2	ner_deid_small	Name, Profession, Location, Age, Date, Contact, Id	Pretrained
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	Pretrained
risk_factors_2014	ner_riskfactors	PHI, Medication, CAS, Hypertension, Diabetes, Smoker, Hyperlipidemia, Obese, Family_Hist	Pretrained

John Snow LABS

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 .

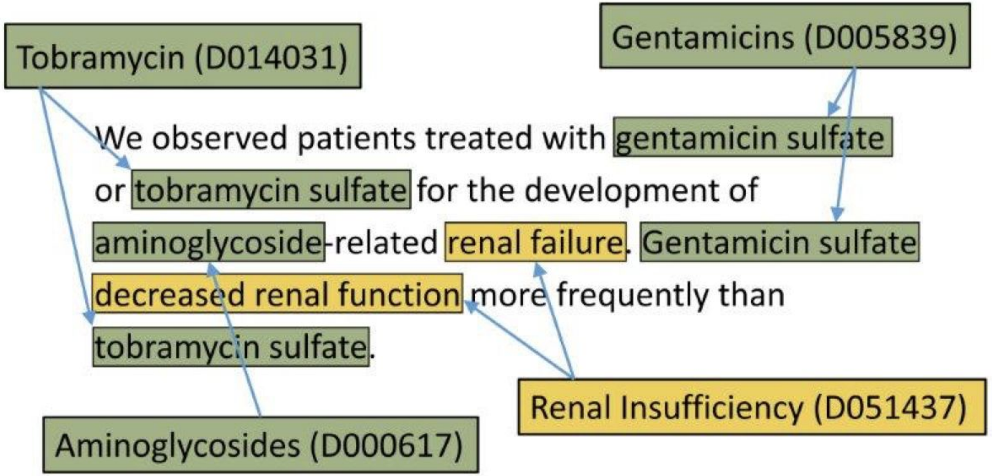
Color codes:PROBLEM, TREATMENT, TEST,

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)]
```

Entity Resolution




```
+---------+--------------------------------------------
|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
```

"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

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

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 : 2093-01-13 , <DOCTOR> , M.D .

, Name : <PATIENT> , <PATIENT> MR .

<MEDICALRECORD> Date : <DATE> PCP : <DOCTOR> , 25 month years-old, Record date : 2079-11-09 .

<HOSPITAL> .

<STREET>

```
# Call annotate() in order to test a sentence or a list of sentences
ori_str = "Name: Smith García, DOB: 23/07/1977 Dr. Suarez. 17 Main Street, Miami Hospital, USA"
light_data = light_pipeline.annotate(ori_str)
print(ori_str)
print("".join(light_data['deidentified']))
```

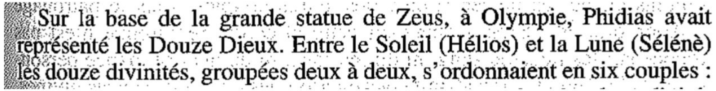
```
Name: Smith García, DOB: 23/07/1977 Dr. Suarez. 17 Main Street, Miami Hospital, USA
Name: <DOCTOR>, DOB: <DATE> Dr. <DOCTOR>.17 Main <STREET>, <HOSPITAL>, USA
```

Here we can how the NERDI for deidentification assigns the different NER classes to the tokens:

```
print("TOKEN (NER)")
print("===========")
for i in range(len(light_data['token'])):
    print(light_data['token'][i] + " (" + light_data['ner'][i]+")")
    print("-----------")
```

```
TOKEN (NER)
===========
Name (O)
-----------
: (O)
-----------
Smith (I-DOCTOR)
-----------
García (I-DOCTOR)
-----------
, (O)
-----------
DOB (O)
-----------
: (O)
-----------
23/07/1977 (I-DATE)
-----------
Dr (O)
-----------
. (O)
-----------
Suarez (I-DOCTOR)
-----------
```

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

Uploaded Image.

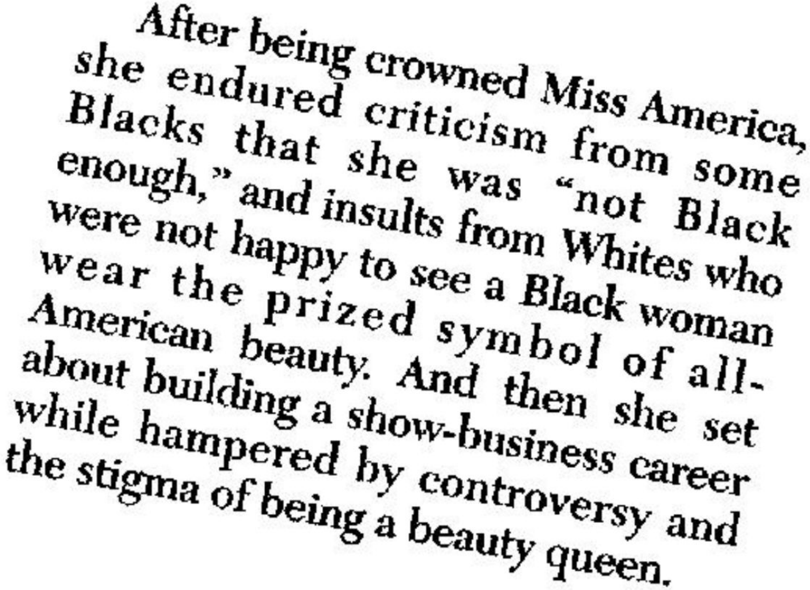
Converted Text:

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

Tepresénté 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


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 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,

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)
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