SPARK+AI SUMMIT 2018

State of the Art Natural Language Processing at Scale

Alex Thomas

Data Scientist @ Indeed

David Talby

CTO @ Pacific AI

#DD4SAIS

CONTENTS

- ✓ NLU REAL-WORLD EXAMPLES
- ✓ DOCUMENT CLASSIFICATION - WALKTHROUGH
- ✓ STATE OF THE ART NLU IN HEALTHCARE
- ✓ TRAIN YOUR OWN DEEP LEARNING NLU MODELS

INTRODUCING SPARK NLP

- Industrial Grade NLP for the Spark ecosystem
- Design Goals:
  1. **Performance & Scale**
  2. **Frictionless Reuse**
  3. **Enterprise Grade**
- Built on top of the Spark ML API's
- Apache 2.0 licensed, with active development & support

NATIVE SPARK EXTENSION

High Performance Natural Language Understanding at Scale

John Snow LABS

Part of Speech Tagger
Named Entity Recognition
Sentiment Analysis
Spell Checker
Tokenizer
Stemmer
Lemmatizer
Entity Extraction

Spark MLlib

Topic Modeling
Word2Vec
TF-IDF
String distance calculation
N-grams calculation
Stop word removal
Train/Test & Cross-Validate
Ensembles

Spark ML API (Pipeline, Transformer, Estimator)

Spark SQL API (DataFrame, Catalyst Optimizer)

Spark Core API (RDD's, Project Tungsten)

Data Sources API

FRICTIONLESS REUSE

```
pipeline = pyspark.ml.Pipeline(stages=[
                document_assembler,
                tokenizer,
                stemmer,
                normalizer,
                stopword_remover,
                tf,
                idf,
                lda])

topic_model = pipeline.fit(df)
```

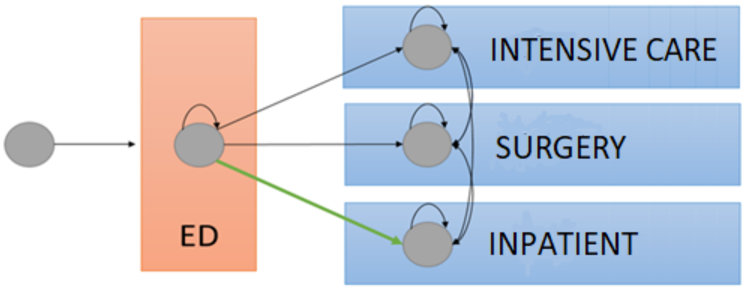
Spark NLP annotators

Spark ML featurizers

Spark ML LDA implementation

Single execution plan for the given data frame

Case study: Demand Forecasting of Admissions from ED



Features from Structured Data

- How many patients will be admitted today?
- Data Source: EPIC Clarity data

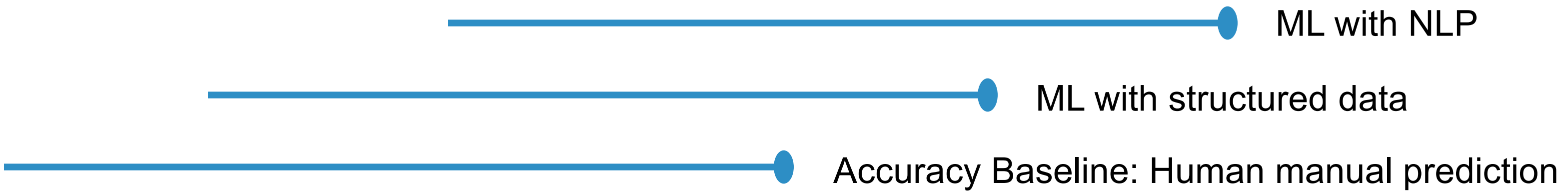
Reason for visit	Current wait time
Age	Number of orders
Gender	Admit in past 30 days
Vital signs	Type of insurance

Case study: Demand Forecasting of Admission from ED

Features from Natural Language Text

- A majority of the rich relevant content lies in unstructured notes that are contributed by doctors and nurses from patient interactions.
- Data Source: Emergency Department Triage notes and other ED notes

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



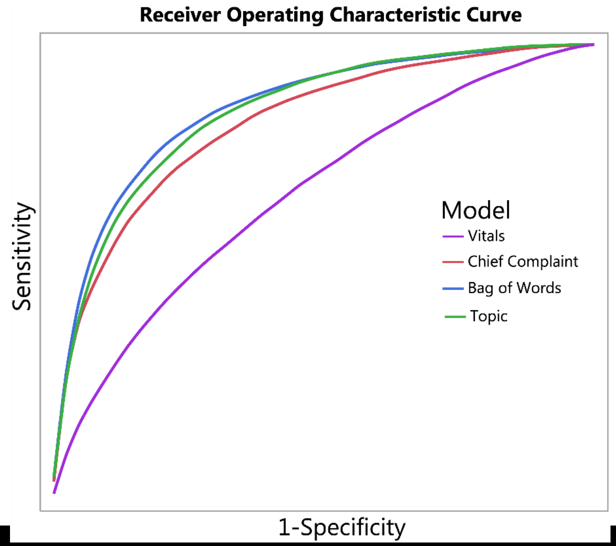
Risk prediction Case Study: Detecting Sepsis

Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning

Steven Horng, David A. Sontag, Yoni Halpern, Yacine Jernite, Nathan I. Shapiro, Larry A. Nathanson

Published: April 6, 2017 • https://doi.org/10.1371/journal.pone.0174708

"Compared to previous work that only used structured data such as vital signs and demographic information, utilizing free text drastically improves the discriminatory ability (increase in AUC from 0.67 to 0.86) of identifying infection."



Cohort selection Case Study: Oncology

"Using the combination of structured and unstructured data, 8324 patients were identified as having advanced NSCLC.

Of these patients, only 2472 were also in the cohort generated using structured data only.

Further, 1090 patients who should have been excluded based on additional data, would be included in the structured data only cohort."

Opportunities and challenges in leveraging electronic health record data in oncology

Marc L Berger*[,1], Melissa D Curtis[2], Gregory Smith[1], James Harnett[1] & Amy P Abernethy[2]

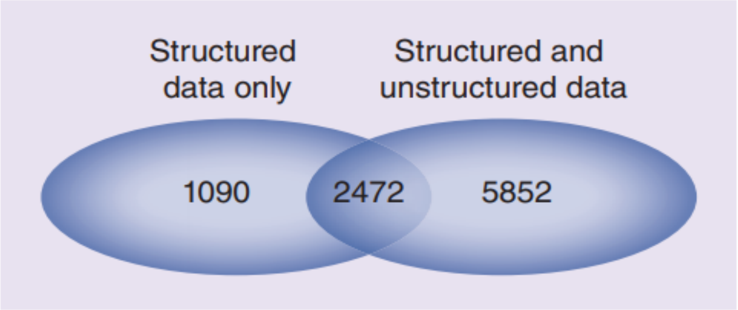


Figure 1. Comparison of patients selected for the analysis using structured data only versus structured and unstructured data.

CODE WALKTHROUGH: DOCUMENT CLASSIFICATION

- A combined NLP & ML Pipeline
- Word embeddings as features
- Training your own custom NLP models

github.com/melcutz/nlu_tutorial

Different Vocabulary	Tokenizer Lemmatizer	Normalizer Fact Extraction
Different Grammar	Part of Speech Tagger Coreference Resolution Sentence Splitting	Spell Checker Dependency Parser Negation Detection
Different Context	Named Entity Recognition Intent Classification Word Embeddings	Sentiment Analysis Summarization Emotion Detection
Different Meaning	Question Answering Best Next Action	Relevance Ranking Translation

Different Language Models

Healthcare Extensions

High Performance Natural Language Understanding at Scale

John Snow LABS

com.johnsnowlabs.nlp.clinical.*

Healthcare specific NLP annotators for Spark in Scala, Java or Python:

- Entity Recognition
- Value Extraction
- Word Embeddings
- Assertion Status
- Sentiment Analysis
- Spell Checking, …

John Snow LABS

Part of Speech Tagger
Named Entity Recognition
Sentiment Analysis
Spell Checker
Tokenizer
Stemmer
Lemmatizer
Entity Extraction

Spark MLlib

Topic Modeling
Word2Vec
TF-IDF
String distance calculation
N-grams calculation
Stop word removal
Train/Test & Cross-Validate
Ensembles

Spark ML API (Pipeline, Transformer, Estimator)

Spark SQL API (DataFrame, Catalyst Optimizer)

Spark Core API (RDD's, Project Tungsten)

Data Sources API

John Snow LABS

data.johnsnowlabs.com/health

1,800+ Expert curated, clean, linked, enriched & always up to date data:

- Terminology
- Providers
- Demographics
- Clinical Guidelines
- Genes
- Measures, …

Named Entity Recognition

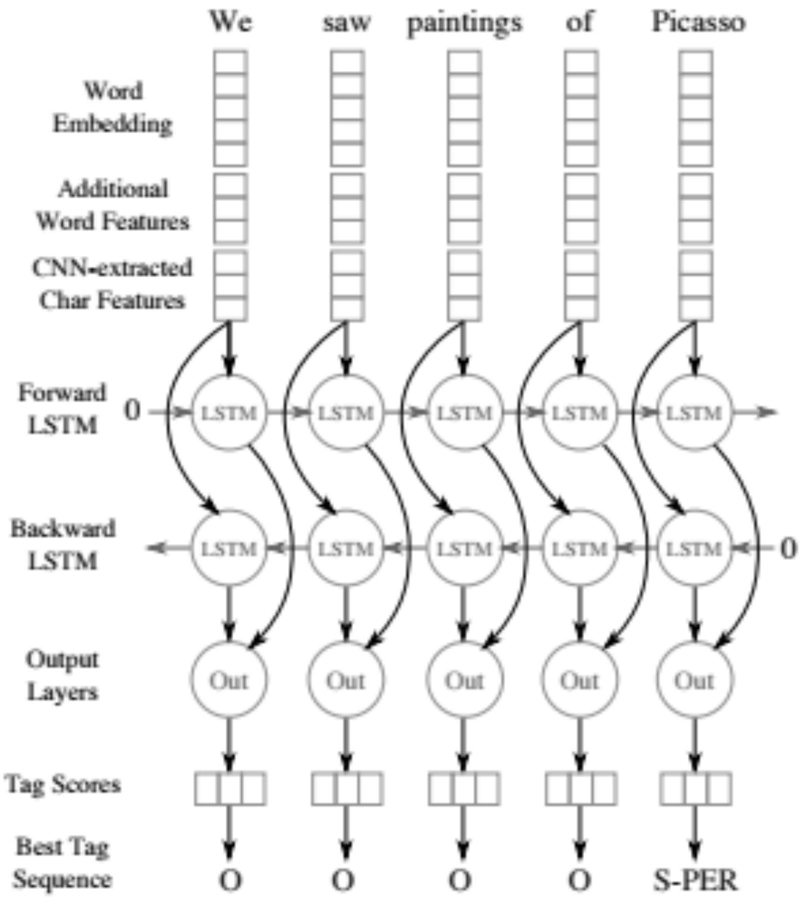
around the left eye . <test>CT of the brain</test> showed no <problem>acute changes </problem> , <problem>left periorbital soft tissue swelling </problem> . <test> CT of the maxillofacial area</test> showed no <problem>facial bone fracture </problem> . <test> Echocardiogram </test> showed normal left ventricular function , <test>ejection fraction</test> estimated greater than 65% . She was set up with a skilled nursing facility , which took several days to arrange , where she was to be given <treatment>daily physical therapy</treatment> and <treatment> rehabilitation </treatment> until appropriate .

Deep Learning for NER

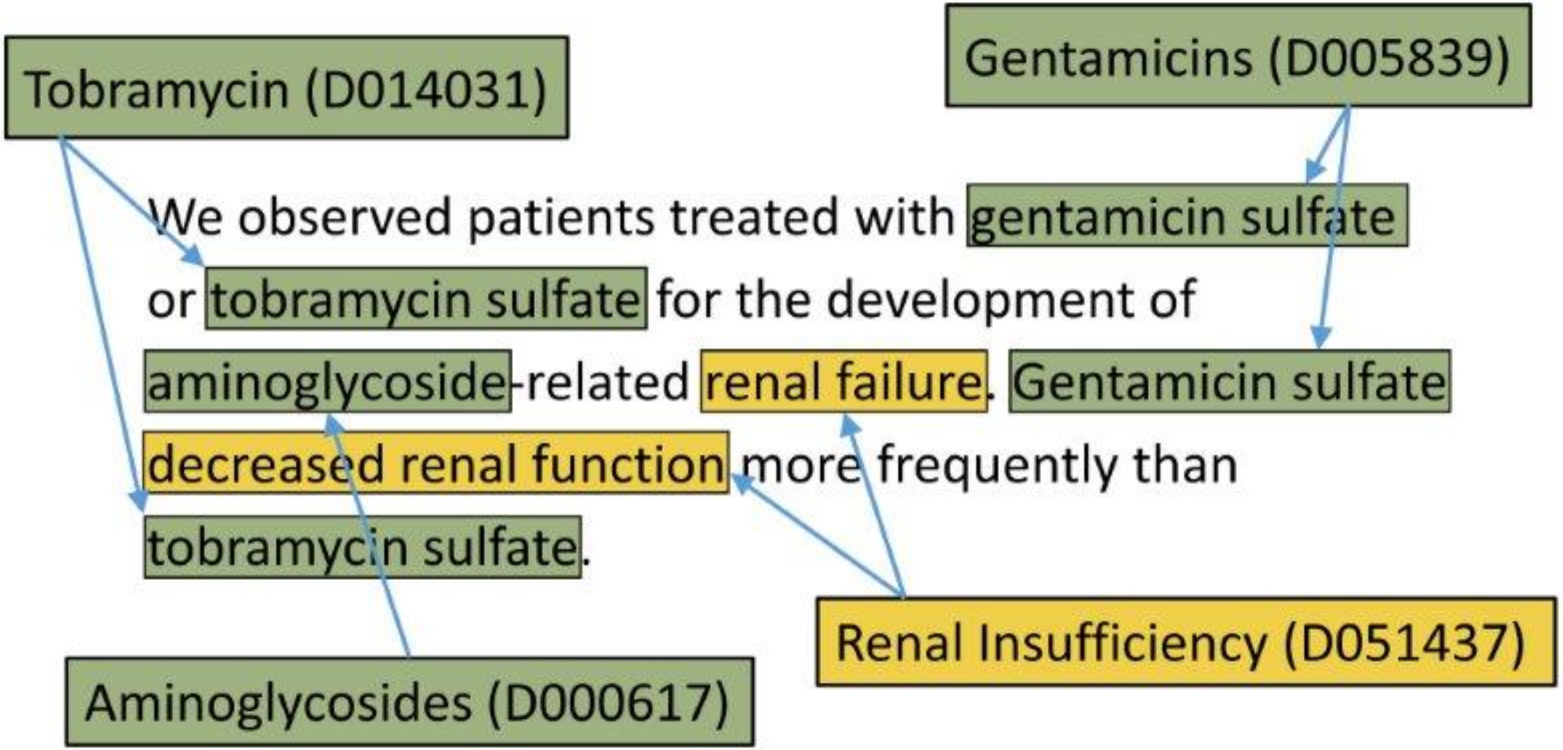
F-Score	Dataset	Task
85.81%	2010 i2b2	Medical concept extraction
92.29%	2012 i2b2	Clinical event detection
94.37%	2014 i2b2	De-identification

“Entity Recognition from Clinical Texts via Recurrent Neural Network”.

Liu et al., *BMC Medical Informatics & Decision Making*, July 2017.



Entity Resolution

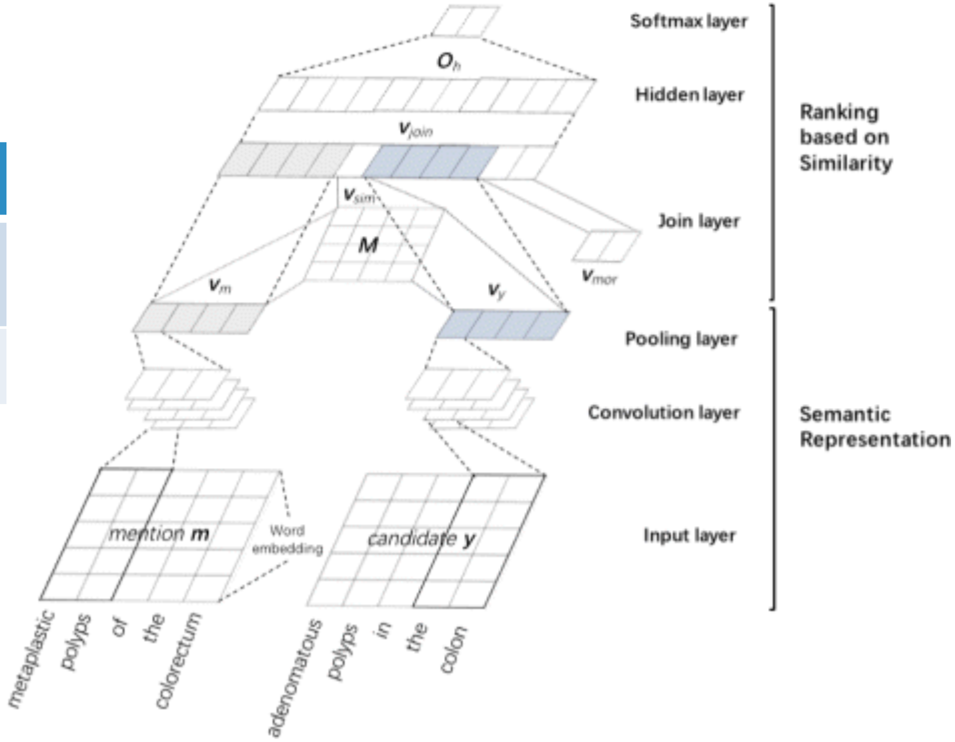


Deep Learning for Entity Resolution

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

“CNN-based ranking for biomedical entity normalization”.

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



Assertion Status Detection

Prescribing sick days due to diagnosis of influenza.	*Positive*
Jane complains about flu-like symptoms.	*Speculative*
Jane’s RIDT came back clean.	*Negative*
Jane is at risk for flu if she’s not vaccinated.	*Conditional*
Jane’s older brother had the flu last month.	*Family history*
Jane had a severe case of flu last year.	*Patient history*

Deep Learning for Assertion Status Detection

	Dataset	Metric
94.17%	4^{th} i2b2/VA	Mirco-averaged F_1
79.76%		Marco-averaged F_1

"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.

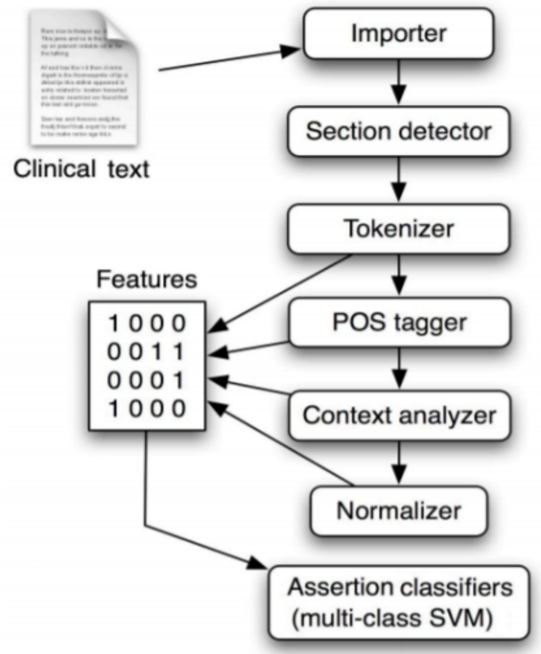


Figure 1: System Pipeline

USING SPARK NLP

- Homepage: https://nlp.johnsnowlabs.com
  - Getting Started, Documentation, Examples, Videos, Blogs
  - Join the Slack Community
- GitHub: https://github.com/johnsnowlabs/spark-nlp
  - Open Issues & Feature Requests
  - Contribute!
- The library has Scala and Python 2 & 3 API's
- Get directly from maven-central or spark-packages
- Tested on all Spark 2.x versions

THANK YOU!

althomas@indeed.com
in/alnith/

david@pacific.ai
in/davidtalby
@davidtalby