



STATE OF THE ART

Legal NLP



STATE OF THE ART

Introduction Legal NLP

John Snow Labs in 2022



Globally awarded

As best AI Specialist
of the 2022 year

Global 100

Most popular

NLP library in
the enterprise

O'Reilly Media
PyPI downloads

#1 Accuracy

on 20 benchmarks in
peer-reviewed papers

Papers with Code



A unified CV, OCR, and NLP approach for
scalable document understanding at DocuSign



Automated Classification and Entity Extraction
from Essential Clinical Trial Documents



Adverse Drug Event Detection using Spark NLP



Accelerating Biomedical Innovation by
Combining NLP and Knowledge Graphs



Spark NLP in action: intelligent, high-accuracy
fact extraction from long financial documents



Extracting what, when, why, and how from
Radiology Reports in Real World Data Projects



Text Classification into a Hierarchical Market
Taxonomy using Spark NLP at Bitvore



Lessons Learned De-Identifying 700 Million
Patients Notes with Spark NLP

Applying Spark NLP to Develop Multi-Modal
Prediction Models from EHR Records

Spark NLP

Community & models hub:
<https://nlp.johnsnowlabs.com>

downloads 40M

downloads/month 2M

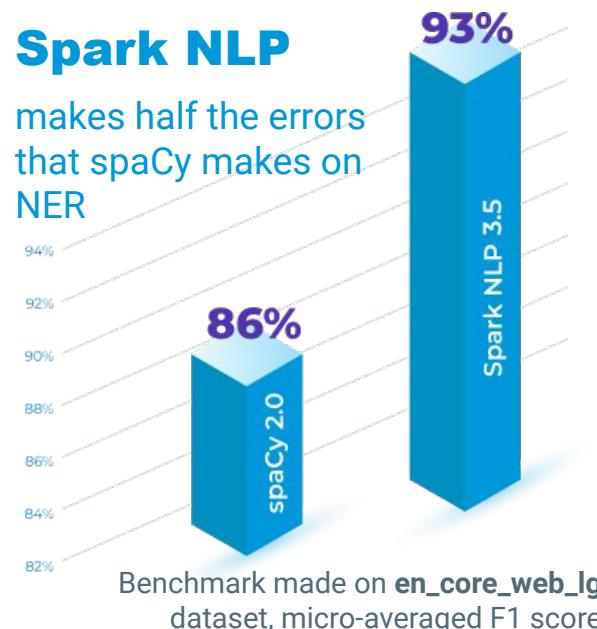
Entity Recognition	Information Extraction	Spelling & Grammar	Text Classification																					
I love Lucy PERSON	They met Last week DATE -> 29-04-2020	abc ✓ She become the first... -> She became the first																						
Translation	Summarization	Question Answering	Emotion Detection																					
 [je t'aime -> i love you]		 Q&A																						
Split Text <ul style="list-style-type: none"> Sentence Detector Tokenizer Normalizer nGram Generator Word Segmentation 		Clean Text <ul style="list-style-type: none"> Spell Checker Grammar Checker Writing Style Checker Stopword Cleaner Summarization 																						
Understand Grammar <ul style="list-style-type: none"> Stemmer Lemmatizer Part of Speech Tagger Dependency Parser Translation 		Find in Text <ul style="list-style-type: none"> Text Matcher Regex Matcher Date Matcher Chunker Question Answering 																						
Trainable & Tunable	Scalable to a Cluster	Fast Inference	Hardware Optimized																					
	APACHE Spark ML Pipelines		 NVIDIA																					
Community																								
																								
10,000+ Pre-trained Pipelines, Models & Transformers <table border="1"> <tr><td>BERT</td><td>ELMO</td><td>GloVe</td></tr> <tr><td>ALBERT</td><td>DeBERTa</td><td>USE</td></tr> <tr><td>Longformer</td><td>ELECTRA</td><td></td></tr> <tr><td>T5</td><td>NMT</td><td>LaBSE</td></tr> <tr><td>DistilBERT</td><td>RoBERTa</td><td></td></tr> <tr><td colspan="2">XLM-RoBERTa</td><td></td></tr> <tr><td>S-BERT</td><td>XLNet</td><td></td></tr> </table>				BERT	ELMO	GloVe	ALBERT	DeBERTa	USE	Longformer	ELECTRA		T5	NMT	LaBSE	DistilBERT	RoBERTa		XLM-RoBERTa			S-BERT	XLNet	
BERT	ELMO	GloVe																						
ALBERT	DeBERTa	USE																						
Longformer	ELECTRA																							
T5	NMT	LaBSE																						
DistilBERT	RoBERTa																							
XLM-RoBERTa																								
S-BERT	XLNet																							

Spark NLP

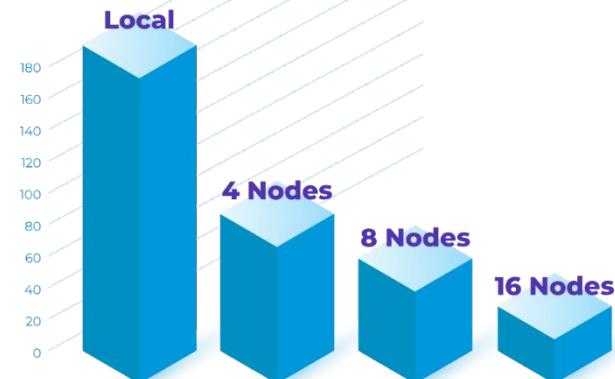
is natively scalable and production-ready

Spark NLP

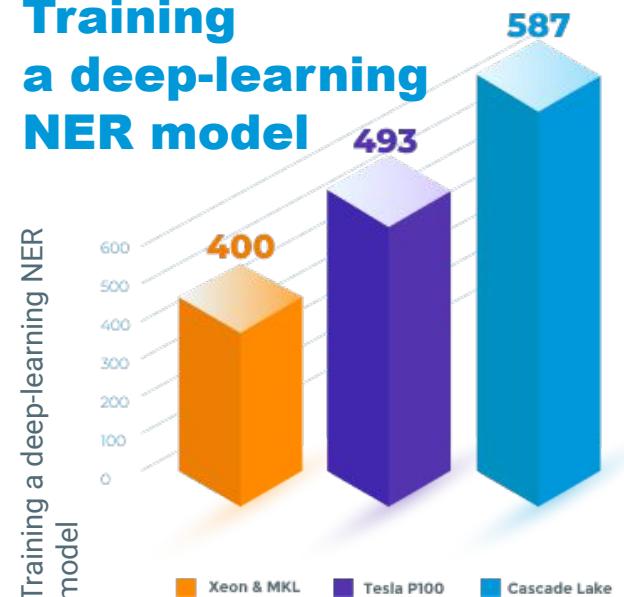
makes half the errors
that spaCy makes on
NER



Speedup on Cluster (less is better)



Training a deep-learning NER model



Accuracy

Scalability

Speed

Legal NLP Infrastructure



- Install **on-premises (locally)**
 - **Fully compliant, air-gapped environments**
- Use it in the **cloud** with ready-to-use images in Databricks, AWS and Azure
- **Automatic scalability** in environments Databricks, AWS EMR and Azure HDInsight

Install Guide

Tell us what you need and we'll guide you how to get it.

Choose Product NLP Libraries Annotation Lab

Choose Edition FREE Community Healthcare Finance FREE Legal Visual / OCR

Where to Install on Premise FREE TRIAL on AWS Marketplace FREE TRIAL on Azure Marketplace FREE TRIAL on Databricks

Autopilot Options
 Enable autoscaling ?
 Terminate after 120 minutes of inactivity ?

Worker Type ? Min Workers 2 Max Workers 8 Spot instances ?

New Configure separate pools for workers and drivers for flexibility. [Learn more](#)

Driver Type
Same as worker 14.0 GB Memory, 4 Cores, 0.75 DBU | ▾

Advanced Options 1.508 x 1.226

Legal NLP



Legal Entity Recognition	Legal Entity Linking	Assertion Status	Relation Extraction
<p>This Loan Agreement dated as of November 17, 2014 (this «Agreement»), is made by and among Auxilium Pharmaceuticals, Inc., a corporation incorporated under the laws of the State of Delaware («U.S. Borrower»), Auxilium UK LTD, a private company limited by shares registered in England and Wales («UK Borrower») and, collectively with the U.S. Borrower, the «Borrowers» and Endo Pharmaceuticals Inc., a corporation incorporated under the laws of the State of Delaware («Lender»).</p>	<p>Sector & Industry Finance (4800) Major Banks (4805)</p> <p>Fiscal Year End December</p> <p>Exchange/ISIN SIX Swiss/CH0244767585</p> <p>SEDOL BR0176</p> <p>Investor Relations Contact Martin A. Osinga</p> <p>LEI 5493005ZJ9VS85GXANS1</p>	<p>Designates to sign the form of 731 and other documentation: → TRUE</p> <p>Neither...nor...is subject to a denial... → FALSE</p> <p>...may require approval to... → POSSIBLE</p>	<p>Sauer Christopher</p> <p>↓ ↓</p> <p>designates works for has_power</p> <p>Michael Lin Sporton International Sign 731 form</p>
<p>UBS Group AG is a holding company, which engages in the provision of financial management solutions. It operates through the following segments: Global Wealth Management, Personal and Corporate Banking, Asset Management; Investment Bank and Corporate Center. The Global Wealth Management segment advises and offers financial services to wealthy private clients except those served by Wealth Management Americas which include banking and lending, wealth planning, and investment management. The Personal and Corporate segment offers financial products and services to private, corporate, and institutional clients in Switzerland. The Asset Management segment consists of investment management products and services, platform solutions and advisory support to institutions; wholesale intermediaries, and wealth management clients. The Investment Bank segment comprises investment advice, financial solutions, and capital markets access among corporate, institutional, and wealth management clients. The Corporate Center segment is involved in the services, group asset and liability management and non-core and legacy portfolio. The company was founded on June 29, 1998 and is headquartered in.</p>	<p>Legal Embeddings</p> <p>Document Splitting</p> <p>Knowledge Graphs</p> <p>Long Span Extraction with Question Answering</p>	<p>Zero-shot Relation Extraction</p> <p>Clause Extraction</p> <p>Pattern Matching and Text Mining</p> <p>Deidentification</p>	





STATE OF THE ART

Text Splitting Legal NLP

Splitting Legal texts

One of the first tasks when applying NLP to texts is **splitting**. Splitting means dividing the text into smaller chunks.

The main component to do that is **SentenceDetector**, a rule-based annotator, or **SentenceDetectorDL**, a pretrained, deep-learning based **Sentence Detector**. Don't get confused by the name, it could return whole **paragraphs or sections** as well using the setter `setCustomBounds()`. Other relevant setters: `setUseCustomBoundsOnly()` and `setCustomBoundsStrategy()`.

AGREEMENT

NOW, THEREFORE, for good and valuable consideration, and in consideration of the mutual covenants and conditions herein contained, the Parties agree as follows:

2. Definitions. For purposes of this Agreement, the following terms have the meanings ascribed thereto in this Section 1. 2.
Appointment as Reseller.

2.1 Appointment. The Company hereby [***]. Allscripts may also disc
Processing Services and facilitate procurement of Merchant Process
without limitation by references to such pricing information and Me

2.2 Customer Agreements.

a) Subscriptions. Allscripts and its Affiliates may sell Subscriptions for
years on a subscription basis to Persons who subsequently execute a
into Customer Agreements with terms longer than four (4) years with
each instance in writing in advance, which consent will not be unreas

```
text = """
4. GRANT OF KNOW-HOW LICENSE
4.1 Arizona Know-How Grant. Subject to the terms and conditions of this Agreement, Arizona hereby grants
4.2 Company Know-How Grant. Subject to the terms and conditions of this Agreement, the Company hereby grants
5. GRANT OF PATENT LICENSE
5.1 Arizona Patent Grant. Subject to the terms and conditions of this Agreement, Arizona hereby grants
"""


```

```
documentAssembler = nlp.DocumentAssembler()\n    .setInputCol("text")\n    .setOutputCol("document")
```

```
paragraphDetector = nlp.SentenceDetector()\n    .setInputCols(["document"])\n    .setOutputCol("paragraph")\n    .setCustomBounds([\n        "\n[\d.]+"
    ])\n    .setCustomBoundsStrategy('prepend')\n
```

Splitting Legal texts

Sentence *Splitting* is very important for all the NLP use cases we are going to comment later on:

- For **Text Classification**, Sentence Detector will decide how much information will be sent to the Classifier.
 - Missing text could retrieve bad predictions
 - Passing too much may make the model ignore due to *token restrictions*, or get the *information mixed or deluded* (where you miss the key information in an ocean of other stuff).

'The IC and SoC design excellence requires technologies for custom IC, digital IC design and signoff, and functional verification, and leverages pre-built semiconductor IP. These tools, IP and associated services are specifically designed to meet the growing requirements of engineers designing increasingly complex chips across analog, digital and mixed-signal domains, and perform the associated verification tasks, including validation of low-level software running on the silicon model, thereby enabling design teams to manage complexity and verification throughput without commensurately increasing the team size or extending the project schedule, while reducing technical risks.\nThe second layer of our strategy centers around system innovation. It includes tools and services used for system design of the packages that encapsulate the ICs and the PCBs, system simulation which includes electromagnetic, electro-thermal and other multi-physics analysis necessary as part of optimizing the full system's performance, radio frequency ("RF") and microwave systems, and embedded software.\nThe third layer of our strategy addresses pervasive intelligence in new electronics. It starts with providing solutions and services to develop AI-enhanced systems and includes machine learning and deep learning capabilities being added to the Cadence\n\n technology portfolio to make IP and tools more automated and to produce optimized results faster.\nOur software and emulation products also support cloud access to address the growing computational needs of our customers.Recent Acquisitions During fiscal 2021, we continued to execute our Intelligent System Design strategy and expanded our product offerings and solutions into computational fluid dynamics ("CFD") with our acquisitions of Belgium-based NUMECA International, a leader in CFD technology, and Pointwise, Inc, a leading provider of CFD meshing technology. The addition of these technologies and talent broadens our System Design and Analysis portfolio and expertise. Chief Executive Officer Transition: On December 15, 2021, Anirudh Devgan assumed the role of President and Chief Executive Officer of Cadence, replacing Lip-Bu Tan. Prior to his role as Chief Executive Officer, Dr. Devgan served as President of Cadence. Concurrently, Mr. Tan transitioned to the role of Executive Chair.'



```
is_acquisitions? NO  
is_work_experience? NO
```

Splitting Legal texts

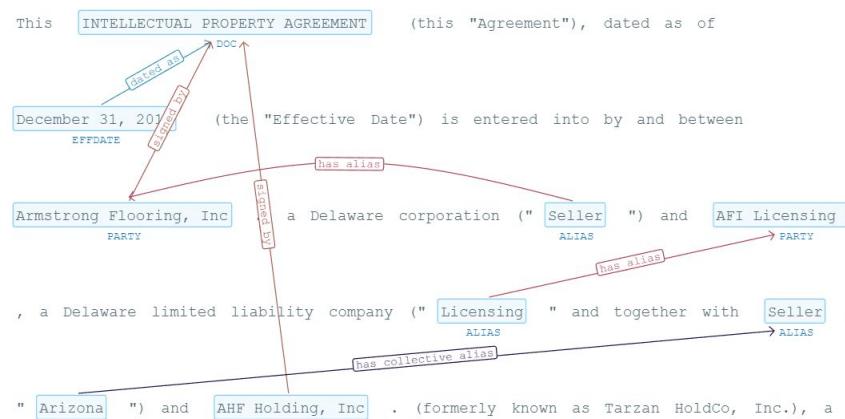
Sentence *Splitting* is very important for all the NLP use cases we are going to comment later on:

- For **NER**, in most cases, the information is contained in the **same sentence**, although in case of enumerations you may want to consider paragraph NER.

WHEREAS, the Corporation wishes to provide:
a) **investment advise;**
b) **management services;**
c) **administrative services**

...

- For **Assertion**, as with Text Classification, you may want to send the model more than just a sentence.
- For **Relation Extraction**, quite common entities are in different sentences, so you may want to split by paragraph



Splitting Legal texts

Sentence *Splitting* is very important for all the NLP use cases we are going to comment later on:

- For **NER**, in most cases, the information is contained in the **same sentence**, although in case of enumerations you may want to consider paragraph NER.

WHEREAS, the Corporation wishes to provide:

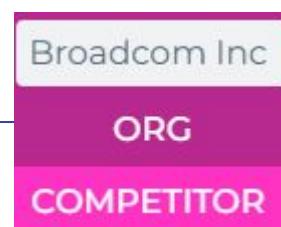
- a) **investment advise;**
- b) **management services;**
- c) **administrative services**

...

- For **Assertion**, as with Text Classification, you may want to send the model more than just a sentence.

Our **competitors** include legacy antivirus product providers. The most relevant ones are:

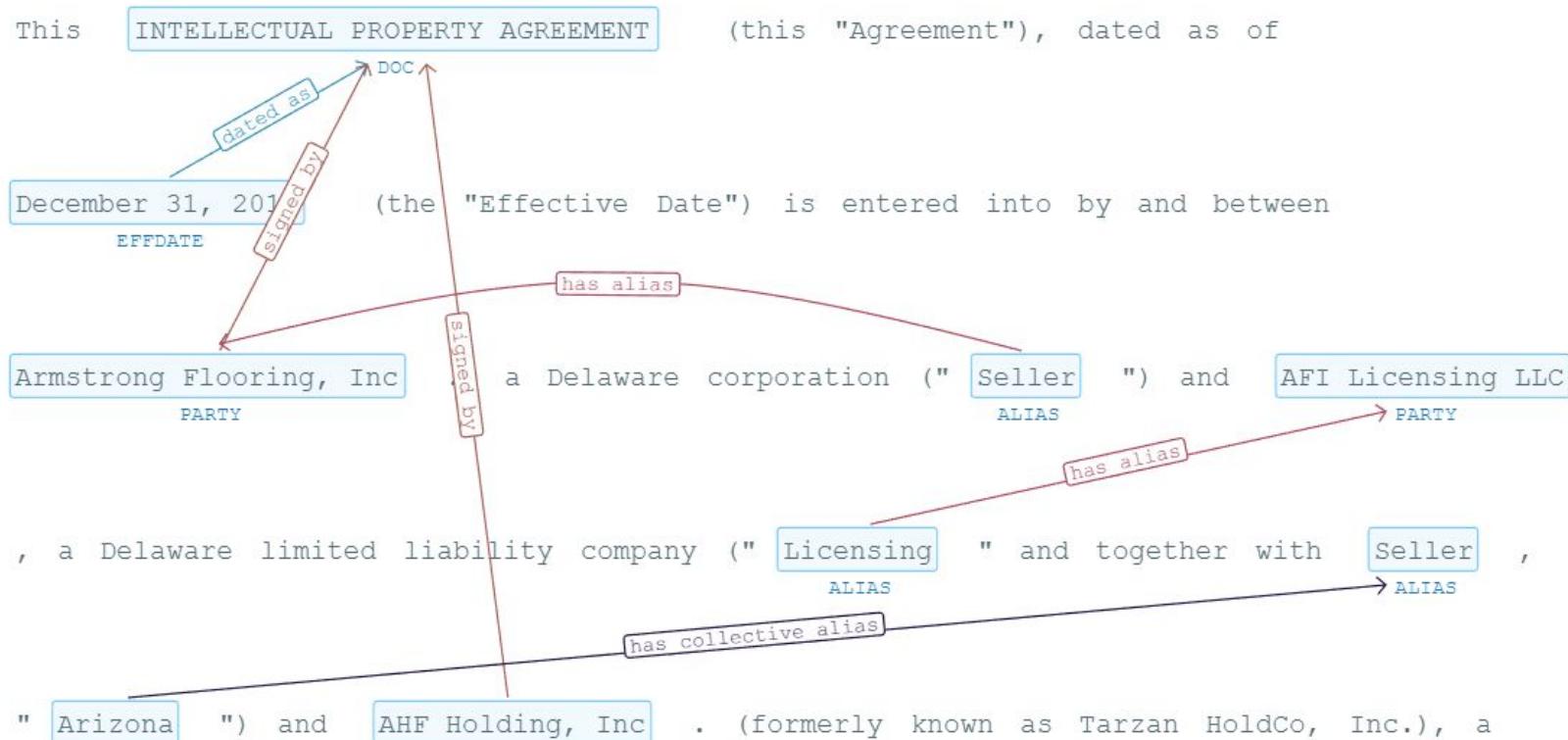
and



Splitting Legal texts

Sentence *Splitting* is very important for all the NLP use cases we are going to comment later on:

- For **Relation Extraction**, is quite common entities are in different sentences, so you may want to split by paragraph





STATE OF THE ART

Language Models Legal NLP

Language Models and Embeddings

Language Models are Deep Learning objects you will use to process your texts. They are based on **Fill-mask** and **next-token prediction**, which means they learn the texts they see in training time and are able to predict a word if you mask it.

What we use from Language Models is not the fill-mask or next-token prediction, but the **numerical representation of the words** (or sentences), also called as **Embeddings**.

These numerical representations of words store information of their meaning in context.

The screenshot shows a dictionary entry for the word "bank".

bank²
/bæŋk/
noun
noun: bank; plural noun: banks

1. a financial establishment that uses money deposited by customers for investment, pays it out when required, makes loans at interest, and exchanges currency.
"a bank account"

Similar: financial institution, commercial bank, savings bank, finance company, ▾

• the store of money or tokens held by the banker in some gambling or board games.
noun: the bank

• the person holding the bank in some gambling or board games; the banker.

• INFORMAL • US
a large amount of money.
"those entrepreneurs are raking in some serious bank"

2. a stock of something available for use when required.
"a blood bank"

Similar: store, reserve, accumulation, stock, stockpile, inventory, supply, ▾

• a site or receptacle where something may be deposited for recycling.
"a paper bank"

3. a set of similar things, especially electrical or electronic devices, grouped together in rows.
"the DJ had big banks of lights and speakers on either side of his console"

Similar: array, row, line, tier, group, series, panel, console, ▾

• a tier of oars.
"the early ships had only twenty-five oars in each bank"

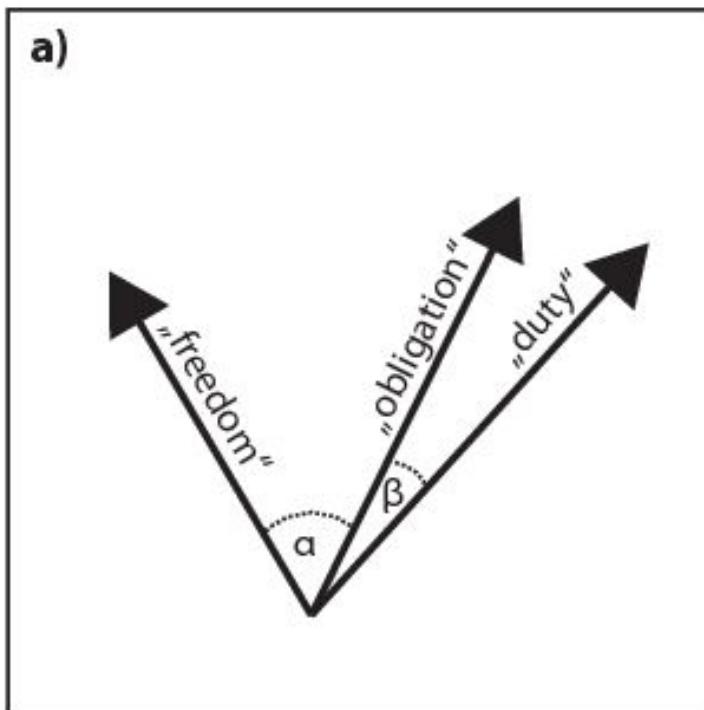
4. the cushion of a pool table.
"a bank shot"

All of these will have different embeddings (numerical representations) in context!

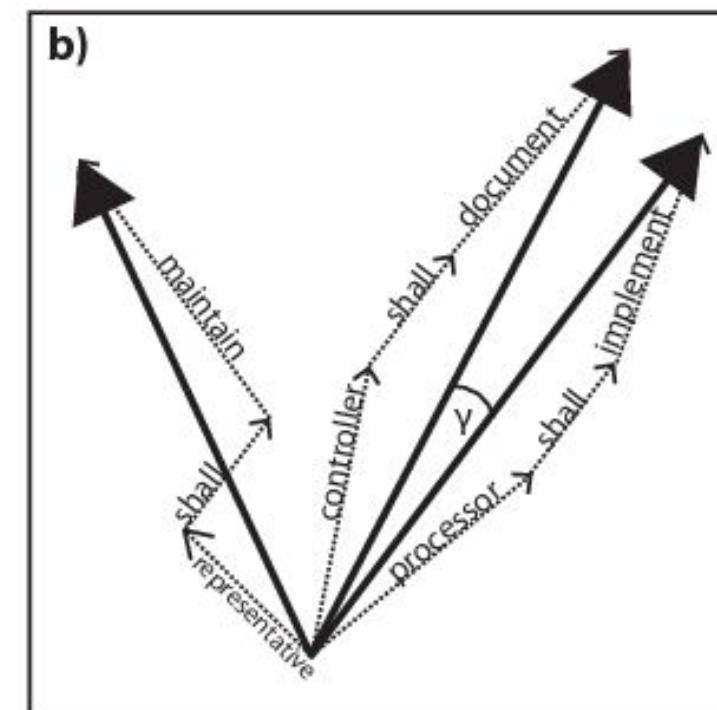
Language Models and Embeddings

We have two type of embeddings:

- **Word Embeddings**, for word-based NLP tasks, as:
 - Name Entity Recognition
 - Assertion Status
 - Relation Extraction, etc.
- **Sentence Embeddings**, for sentence/paragraph/document NLP tasks, as:
 - Text Classification
 - Entity Resolution



Legal Word Embeddings



Legal Sentence Embeddings

Language Models and Embeddings

Domain specificity

- As a consequence of their context-specificity, it's very important you use domain specific embeddings. Fortunately, we have **more than 30** Legal NLP Language Models in Models Hub, including English, Portuguese and Spanish.

Word vs Sentence

- If you don't find a proper Sentence Embeddings for you and you have a suitable Word Embeddings model, we provide with an **annotator called SentenceEmbeddings**, which will do the transformation for you.

Cased vs Uncased

- Please pay attention to the casing of the models. Some of them will require to lowercase the text first.

The screenshot shows the Hugging Face Model Hub interface. At the top, there is a search bar with the text "Show Embeddings × models & pipelines in All Languages for All versions". Below the search bar, the results are displayed under the heading "31 Models & Pipelines Results:". There is a checkbox labeled "Supported models only" which is unchecked. On the left, there is a sidebar with filters for "All", "Models", and "Pipelines", and dropdowns for "Assigned tags" and "Entities". The main area displays three cards for supported models:

- Legal BERT Base Uncased Embedding**: A BERT model for the legal domain. It was trained by PlanTL-GOB-ES. The card includes details like Date: 09.2021, task: Embeddings, Language: English, and Edition: Spark NLP 3.2.2.
- Legal BERT Sentence Base Uncased Embedding**: Another BERT model for the legal domain. It was trained by PlanTL-GOB-ES. The card includes details like Date: 09.2021, task: Embeddings, Language: English, and Edition: Spark NLP 3.2.2.
- Spanish Legal RoBERTa Embeddings**: A RoBERTa model for the Spanish legal domain. It was trained by PlanTL-GOB-ES/Im-legal-es. The card includes details like Date: 04.2022, task: Embeddings, Language: Spanish, and Edition: Spark NLP 3.4.2.



STATE OF THE ART

Text Classification Legal NLP

This agreement, or any term thereof, may be changed or waived only by written amendment, signed by the party against whom enforcement of such change or waiver is sought.

This sentence has been classified as : **Amendments**

Classification Confidence: **99.92%**

Classify clauses and whole documents

We anticipate the value of our company to continue to rise for over the next few years, increasing dividends for shareholders.

This sentence has been classified as : **Compensation**

Classification Confidence: **99.75%**

No loans shall be contracted on behalf of the company and no evidences of indebtedness shall be issued in its name unless authorized by a resolution of the board of managers

This sentence has been classified as : **Loans**

Classification Confidence: **100.0%**

Each party warrants and represents that it has full capacity and authority, all necessary licenses, permits and consents to enter into and perform its obligations under the agreement.

This sentence has been classified as : **Guarantee**

Classification Confidence: **99.76%**

The powers of the company shall be exercised by or under the authority of, and the business and affairs of the company shall be managed under the direction of, the member.

This sentence has been classified as : **Management**

Classification Confidence: **99.97%**

Legal NLP Classification

Text Classification is the NLP Task in charge of retrieving a **class/category** per input text.

- **Classification** require domain **Sentence Embeddings**. Remember, if you don't find proper sentence embeddings, you can use SentenceEmbeddings annotator to transform your word embeddings into SentenceEmbeddings.

We count on more than 400 Text Classifiers, which can be divided using 2 categorization systems:

- By **Input** type or type of **text splitting needed**

Sentences	Clauses / Paragraphs / Sections	Whole Documents
To do classification at sentence level. For example, detecting sentiment on a sentence, if a sentence talks about a specific topic , etc.	This is the most common type of classifiers in Legal NLP. They can be used to identify if a piece of texts bigger than a sentence (a paragraph) is of a specific class. Very useful to detect Legal clauses	To carry out Document Classification. Bear in mind current NLP Models are not able to process big texts. The biggest amount of text we can process is using Legal Longformers with 4096 tokens , or using Bert-based models with 512 . The rest of the text will be discarded. However, the good news is that in most cases, the information to classify a document is in the first page of it.

Legal NLP Classification

- By **output type or class assigned to the input text**

Binary Classifiers

Return *true* or *false* values. For example, our more than 300 Clause Binary classifiers, which return the **name of the clause** if it is classified as such, or **other** otherwise.

This agreement, or any term thereof, may be changed or waived only by written amendment, signed by the party against whom enforcement of such change or waiver is sought.

We anticipate the value of our company to continue to rise for over the next few years, increasing dividends for shareholders.

legclf_amendments

amendments

other

Multiclass classifiers

Returns 1 value from all the categories the model was trained on. Only works for models with a small number of categories (up to 100).

It's not suitable for:

- Big number of classes (more than 100)
- Non-disjoint classes (a text can be of several classes at the same time)

The Commission considered that a special feature in the present case was the fact that the applicant had chosen to express himself through poetry. However, even taking into account the prerogatives of a poet, it found that parts of the applicant's poems glorified armed rebellion against the Turkish State and martyrdom in that fight.

This text has been classified as : **COMMISSION/CHAMBER**
or **APPLICANT** or ...

Multilabel classifiers

Returns n value from all the categories the model was trained on. Only works for models with a small number of categories (up to 100).

It's not suitable for:

- Big number of classes (more than 100)

(a) No failure or delay by the Administrative Agent or any Lender in exercising any right or power hereunder shall operate as a waiver thereof, nor shall any single or partial exercise of any such right or power, or any abandonment or discontinuance of steps to enforce such a right or power, preclude any other or further exercise thereof or the exercise of any other right or power

This text has been classified as: **waivers, amendments**

Legal NLP Classification



We have **thousands of different legal clauses** in documents, and often times, , they **are not disjoint**: a paragraph could have information about **some different legal clauses at the same time**.

Contribution

66k

Indemnification

424k

Indemnification and Contribution

40k

Indemnification and Contribution. (a) The Company agrees to indemnify and hold harmless each Underwriter, the directors, officers, employees and agents of each Underwriter and each person who controls any Underwriter within the meaning of either the Act or the Exchange Act against any and all losses, claims, damages or liabilities, joint or several, to which they or any of them may become subject under the Act, the Exchange Act or other Federal or state statutory law or regulation, at common law or otherwise, insofar as such losses, claims, damages or liabilities (or actions in respect thereof) arise out of or are based upon any untrue statement or alleged untrue statement of a material fact contained in the registration statement for the registration of the Securities as originally filed or in any amendment thereof, or in the Basic Prospectus, any Preliminary Final Prospectus or the Final Prospectus, or in any amendment thereto or supplement thereto, or arise out of or are based upon the omission or alleged omission to state therein a material fact required to be stated therein or necessary to make the statements therein not misleading, and agrees to reimburse each such indemnified party, as incurred, for any legal or other expenses reasonably incurred by them in connection with investigating or defending any such loss, claim, damage, liability or action; provided, however, that the Company will not be liable in any such case to the extent that any such loss, claim, damage or liability arises out of or is based upon any such untrue statement or alleged untrue statement or omission or alleged omission made therein in reliance upon and in conformity with written information furnished to the Company by or on behalf of any Underwriter through the Representatives specifically for inclusion therein. This indemnity agreement will be in addition to any liability which the Company may otherwise have.

indemnification - TRUE

contribution - TRUE

amendment - FALSE

...

xxx - FALSE

Legal NLP Classification

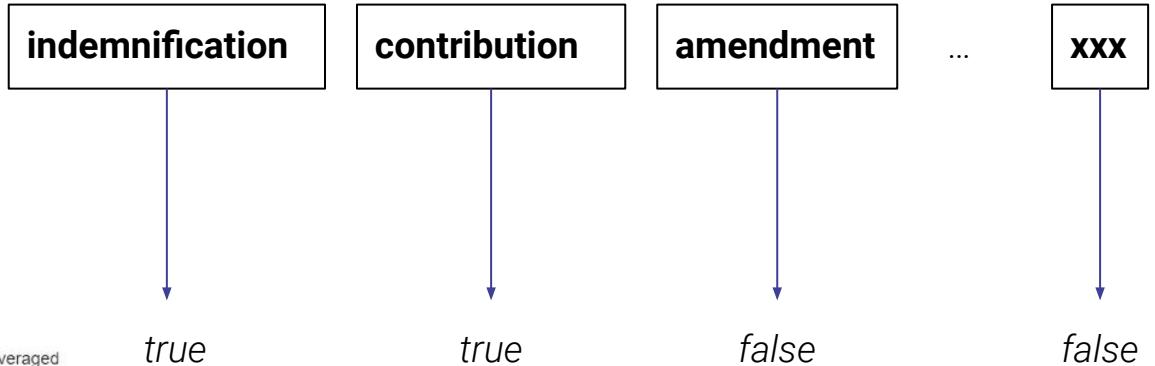
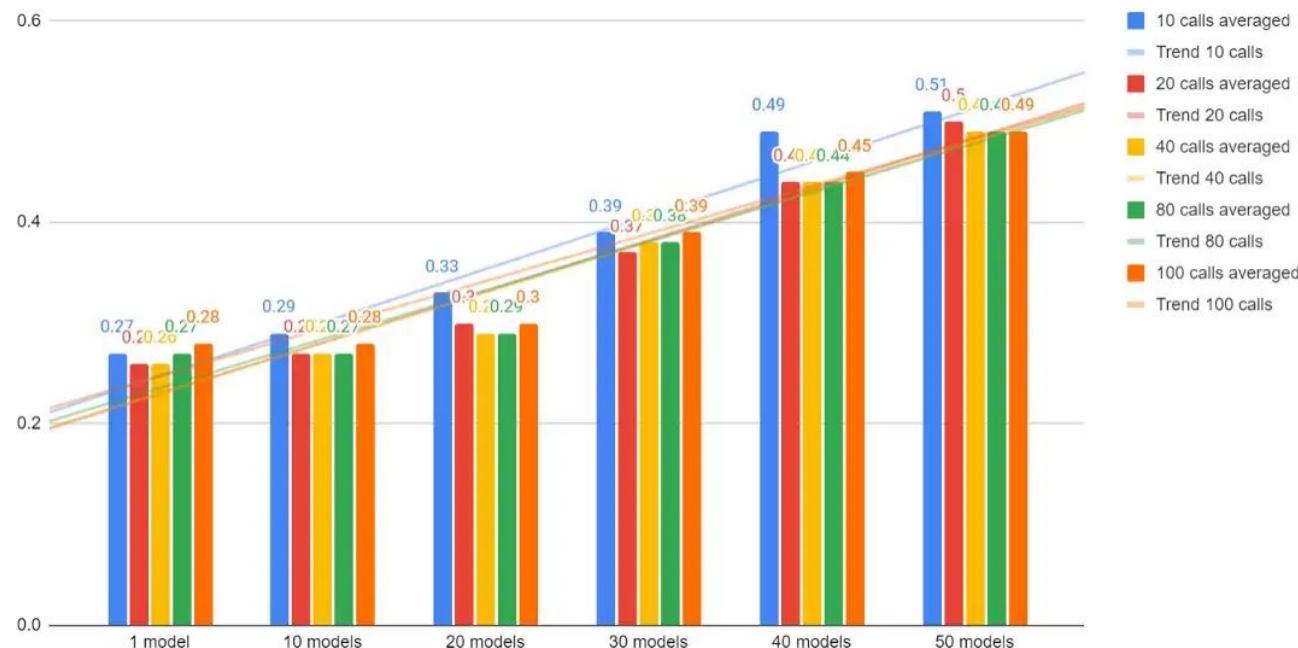


We have **thousands of different legal clauses** in documents, and often times, , they **are not disjoint**: a paragraph could have information about **some different legal clauses at the same time**.



paragraph splitting

Indemnification and Contribution. (a) The Company agrees to indemnify and hold harmless each Underwriter, the directors, officers, employees and agents of each Underwriter and each person who controls any Underwriter within the meaning of either the Act or the Exchange Act against any and all losses, claims, damages or liabilities, joint or several, to which they or any of them may become subject under the Act, the Exchange Act or other Federal or state statutory law or regulation, at common law or otherwise, insofar as such losses, claims, damages or liabilities (or actions in respect thereof) arise out of or are based upon any untrue statement or alleged untrue statement as to a material fact contained in any prospectus or any amendment thereto, or in any preliminary prospectus, any Preliminary Final Prospectus or the Final Prospectus, or in any amendment thereto or supplement thereto, or arise out of or are based upon the omission or alleged omission to state therein a material fact required to be stated therein or necessary to make the statements therein not misleading, and agrees to reimburse each such indemnified party, as incurred, for any legal or other expenses reasonably incurred by them in connection with investigating or defending any such loss, claim, damage or liability arises out of or is based upon any such untrue statement or alleged untrue statement or omission or alleged omission made therein in reliance upon and in conformity with written information furnished to the Company by or on behalf of any Underwriter through the Representatives specifically for inclusion therein. This indemnity agreement will be in addition to any liability which the Company may otherwise have.



Performance-wise, these models are super light. You can include hundreds of them and have them predicting on a paragraph in less than 0.5 seconds. Check <https://medium.com/p/a2f9b899de92>



STATE OF THE ART

**Named Entity Recognition
Legal NLP**

Recognize Legal Entities in Documents

Choose Sample Text

Exhibit 2.7 FORM OF TRADEMARK LICENSE AGREEMENT...

Text annotated with identified Named Entities

Exhibit 2.7 FORM OF TRADEMARK LICENSE AGREEMENT THIS TRADEMARK LICENSE AGREEMENT (this "Agreement"), made and entered into as of the July 1, 2020 (the "Effective Date"), by and between ARCONIC INC., a corporation organized under the laws of Delaware ("Licensee") and ARCONIC ROLLED PRODUCTS CORP., a corporation organized under the laws of Delaware ("Licensor")
and between ARCONIC INC. ("Party") and ARCONIC ROLLED PRODUCTS CORP. ("Party")
("ALIAS")
("ALIAS")

Choose Sample Text

WHEREAS, the Company desires...

Text annotated with identified Named Entities

WHEREAS, the Company desires to retain Nantz Communications and Nantz to provide certain promotional services and Nantz
WHEREAS SUBJECT WHEREAS ACTION WHEREAS OBJECT WHEREAS OBJECT WHEREAS SUBJECT
is willing to provide such services on the terms and conditions set forth herein;
WHEREAS ACTION WHEREAS OBJECT

Extract main actions in an agreement

Choose Sample Text

PPD may engage VS to perform imaging services

Text annotated with identified Named Entities

PPD may engage VS to perform imaging services
OBLIGATION SUBJECT OBLIGATION ACTION OBLIGATION INDIRECT OBJECT OBLIGATION

Extract who, what, to whom?

Legal NLP Named Entity Recognition



NER is the NLP task in charge of detecting relevant words / chunks in texts and categorize them.

- **NER** requires **Word embeddings**.
- As with Classification, **NER** also requires **splitting**. Usually, the split is done at the **sentence** level, but there may be cases where you would like to provide to the NER model more context than a sentence:

Sentences	Paragraphs
<p>To do NER at sentence level, after you split a text into sentences with SentenceDetector.</p> <p>Used in most of the cases, since the context of a relevant entity is found in the surroundings of its sentence.</p>	<p>To do NER at sentence level, after you split a text into paragraphs with SentenceDetector, not into sentences.</p> <p>We may need to do this in some exceptional cases:</p> <p><i>WHEREAS, the Corporation wishes to provide:</i></p> <p>a) <i>investment advise</i>;</p> <p>b) <i>management services</i>;</p> <p>c) <i>administrative services</i></p> <p>...</p>

Legal NLP Named Entity Recognition



We provide with **Legal NER** at **clause** and **document level**.

Clause Level

NER entities can be only found in some specific parts of the document. For example, **whereas**, **indemnification**, **termination**, **the introduction of the parties in an agreement**, etc.

WHEREAS, the Company
WHEREAS SUBJECT
desires to retain
WHEREAS ACTION
Nantz Communications
WHEREAS OBJECT
and Nantz to provide certain promotional
WHEREAS OBJECT

is willing to provide
WHEREAS ACTION
such services
WHEREAS OBJECT
on the terms and conditions set forth herein;

There is **no point** in applying these clause NER models to the whole document, since:

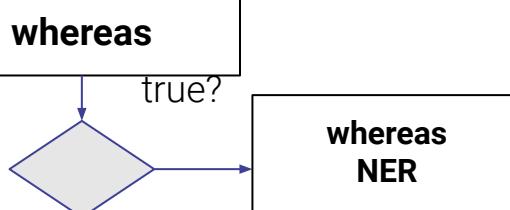
- It will affect drastically the performance
- It will retrieve more false positives / negatives

In order to carry out NER of specific clauses, please use first Text Classification, as described before, and if the specific class you have detected is relevant for your, apply its specific NER.



paragraph splitting

Indemnification and Contribution. (a) The Company agrees to indemnify and hold harmless each Underwriter, the directors, officers, employees and agents of each Underwriter and each person who controls any Underwriter within the meaning of either the Act or the Exchange Act against any and all losses, claims, damages or liabilities, joint or several, to which they or any of them may become subject under the Act, the Exchange Act or other Federal or state statutory law or regulation, at common law or otherwise, insofar as such losses, claims, damages or liabilities result from any untrue statement of a material fact contained in the registration statement or any amendment thereto or in the Base Prospectus, any Preliminary Final Prospectus or the Final Prospectus, or any statement amended thereto or supplement thereto, or in any document or communication filed with the Commission or otherwise required to be filed with the Commission or necessary to make the statements therein not misleading, and agrees to reimburse each such indemnified party, as incurred, for any legal expenses reasonably incurred by such party in connection with investigating or defending any such claim, loss, damage, liability or actions provided, however, that the Company will not be liable in any such case to the extent that any such loss, claim, damage, liability or actions arises out of or is based upon any statement or omission in any such document or communication which is true as of the time it was made therein or reliance upon and in conformity with written information furnished to the Company by or on behalf of any Underwriter through the Representatives specifically for inclusion therein. This indemnity agreement will be in addition to any liability which the Company may otherwise have.



Document Level

NER entities can be only found **all over** the document.

You can apply NER to the whole document.

For example, **PARTY**, **ALIAS** (how the company is mentioned through the document), **OBLIGATIONS** (sentences describing what a **PARTY** has agreed to do), **LAW**, **PERSON**, **DATE**, **COUNTRY**, etc.

Kiromic, Inc., a Delaware corporation (the "Company")
ORG ALIAS

Legal NLP Zero-shot NER



	Entity	Question
0	DATE	['When was the company acquisition?', 'When was the comp.
1	ORG	['Which company?', 'Which was the company acquisition?']
2	STATE	['Which state?']
3	AGREEMENT	['What kind of agreement?']
4	LICENSE	['What kind of license?']
5	LICENSE_REC	['To whom the license is granted?']
6	ALIAS	['Which is the alias?']

In February 2017, the Company entered into an asset purchase agreement with NetSeer, Inc..

DATE

AGREEMENT

ORG

The Company hereby grants to Seller a perpetual, non- exclusive, royalty-free license

LICENSE_RECIPIENT

LICENSE

LICENSE

LICENSE

On March 12, 2020 we closed a Loan and Security Agreement with Hitachi Capital America Corp. (also known as "Hitachi").

DATE

AGREEMENT

ORG

ALIAS

Usually, NLP models follow a **fit-transform** approach, where:

- 1) You **first** train a model, using what we call an Approach (*NerApproach* for NER), using training data.
- 2) And then, you **transform** (predict) on final data (*NerModel* for NER)

However, with the recent improvements in *Natural Language Inference*, we can use **Question Answering** models as well. The idea is quite simple:

- 1) You have a context document;
- 2) You have some *prompts* in form of *questions or examples*.

Using our **ZeroShotNER** annotator, those *questions (prompts)* can be asked to our NLI-based language model, and *retrieve the answers* in form of *predictions*, without a training step. And most importantly, **without any training data required**.



STATE OF THE ART

Relation Extraction Legal NLP

Understand Relationships in Legal Documents

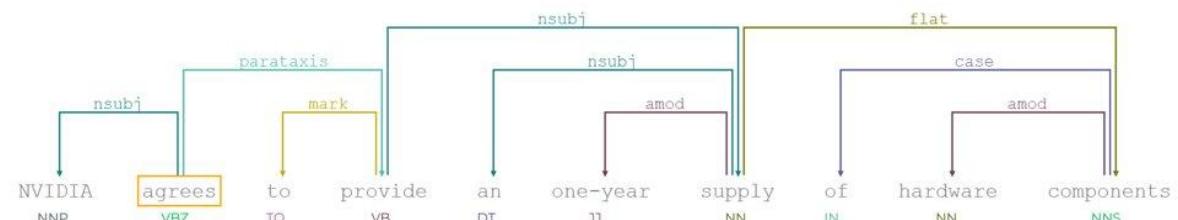
This INTELLECTUAL PROPERTY AGREEMENT (this "Agreement"), dated as of

December 31, 201_{EFFDATE} (the "Effective Date") is entered into by and between

Armstrong Flooring, Inc_{PARTY} a Delaware corporation ("Seller") and AFI Licensing LLC_{ALIAS}

, a Delaware limited liability company ("Licensing") and together with Seller_{ALIAS}

"Arizona") and AHF Holding, Inc. (formerly known as Tarzan HoldCo, Inc.), a



Relation Extraction

Relation Extraction is the NLP Task in charge of detecting if there is any relationship between 2 NER entities, and categorize them.

- Relation Extraction requires **Word embeddings**.
- As with Classification and NER, **Relation Extraction** also requires **splitting**. However, the main difference is that **entities may be in different sentences**, especially in Legal NLP, so it's recommended a bigger splitting than sentences. Usually **paragraph splitting** has good results, but you can also use **section** splitting.

Relation Extraction always goes after **Entity Recognition (NER)**, and tries to **categorize each pair of entities** retrieved in the same chunk,

What happens with texts with many entities?	What happens with long texts?
<p>Relation Extraction will try to understand if each combination of 2 entities is a category.</p> <p>This may be low performant or have undesired results, which you can prevent by:</p> <ul style="list-style-type: none">• Setting which combinations of entities may be checked.	<p>Relation Extraction will try to understand if each combination of 2 entities is a category.</p> <p>This may be low performant or have undesired results, which you can prevent by:</p> <ul style="list-style-type: none">• Set a maximum distance between entities.

Relation Extraction

```
"""
ONLY NEEDED IF YOU WANT TO FILTER RELATION PAIRS OR SYNTACTIC DISTANCE
pos_tagger = PerceptronModel()\
    .pretrained("pos_clinical", "en", "clinical/models") \
    .setInputCols(["document", "tokens"])\
    .setOutputCol("pos_tags")

dependency_parser = DependencyParserModel() \
    .pretrained("dependency_conllu", "en") \
    .setInputCols(["document", "pos_tags", "tokens"]) \
    .setOutputCol("dependencies")

Set a filter on pairs of named entities which will be treated as relation candidates
re_filter = RENerChunksFilter()\n    .setInputCols(["ner_chunks", "dependencies"])\n    .setOutputCol("re_ner_chunks")\n    .setMaxSyntacticDistance(7)\n    .setRelationPairs(['PARTY-ALIAS', 'DOC-PARTY', 'DOC-EFFDATE'])\n"""

re_model = legal.RelationExtractionDLModel.pretrained("legre_contract_doc_parties", "en", "legal/models")\
    .setPredictionThreshold(0.5)\n    .setInputCols(["ner_chunk", "sentence"])\n    .setOutputCol("relations")

pipeline = nlp.Pipeline(stages=[\n    documentAssembler,\n    sentenceDetector,\n    tokenizer,\n    embeddings,\n    nerModel,\n    nerConverter,\n    reModel\n])
```

For doing that, we have a helper annotator called

RENERCHUNKSFILTER.

You can use:

setMaxSyntacticDistance, to restrict the maximum distance between 2 entities.

- **setRelationPairs**, to allow only certain combination of entity types.

These steps are optional, as you can see in some examples they will just be commented out. In other cases it will be crucial due to false positives or negatives.

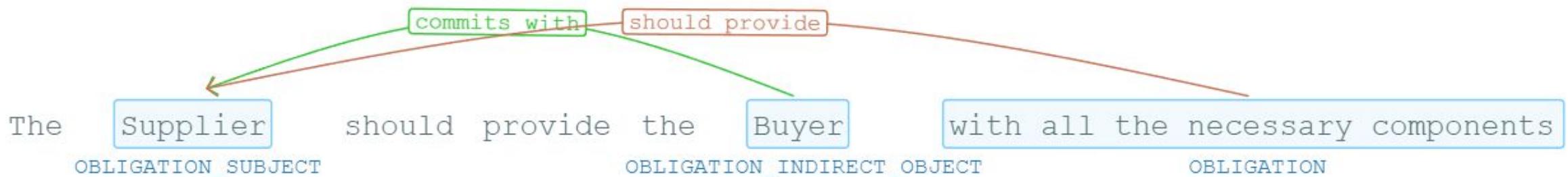
Zero-shot Relation Extraction

As with Zero-shot NER, we can carry out zero-shot Relation Extraction, using the following prompt syntax:

```
re_model = legal.ZeroShotRelationExtractionModel.pretrained("legre_zero_shot", "en", "legal/models")\
    .setInputCols(["ner_chunk", "document"]) \
    .setOutputCol("relations")

re_model.setRelationalCategories({
    "should_provide": ["{OBLIGATION SUBJECT} will provide {OBLIGATION}", "{OBLIGATION SUBJECT} should provide {OBLIGATION}"],
    "commits_with": ["{OBLIGATION SUBJECT} to {OBLIGATION INDIRECT OBJECT}", "{OBLIGATION SUBJECT} with {OBLIGATION INDIRECT OBJECT}"],
    "commits_to": ["{OBLIGATION SUBJECT} commits to {OBLIGATION}"],
    "agree_to": ["{OBLIGATION SUBJECT} agrees to {OBLIGATION}"],
})
```

setRelationalCategories requires a dictionary, having as keys the relationship name, and as values a list of possible prompts which model those relations. In **brackets {}** you need to put the **entity names** (from NER) involved in the relation.





STATE OF THE ART

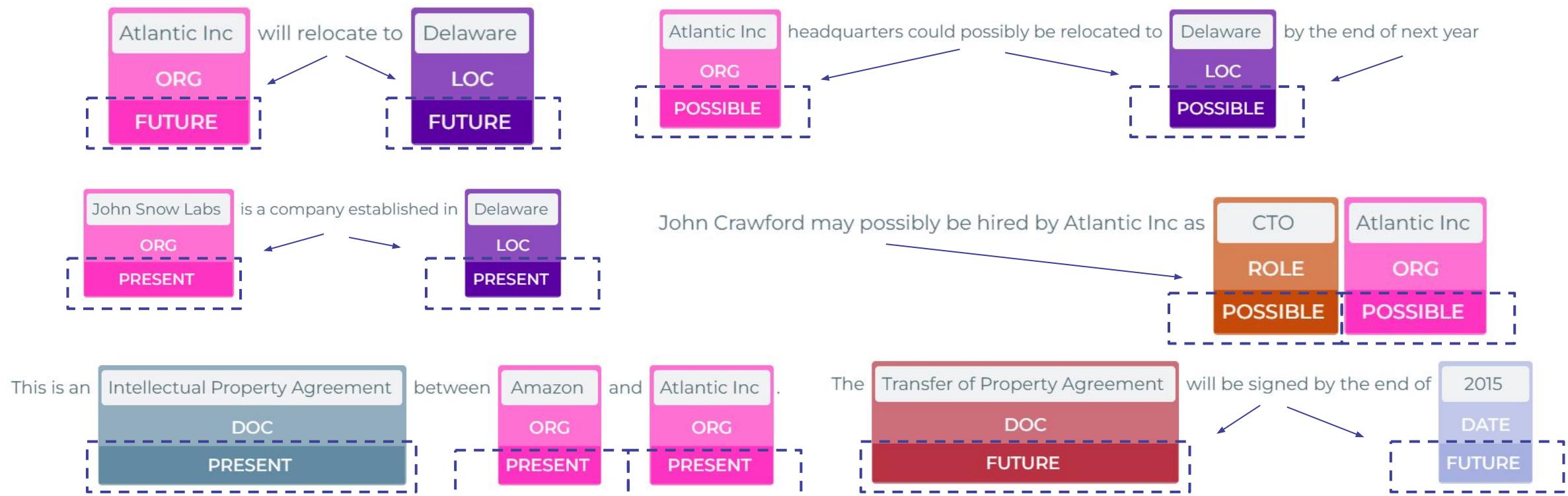
Assertion Status Legal NLP

Understanding Entities in Context: Assertion Status

Assertion Status is the NLP Task in charge of **understanding entities in context**, and categorize them base on it. For example, it can detect if an entity is mentioned in a *Past*, *Future*, *Present* or *Possible* context.

- **Assertion Status** requires **Word embeddings**.
- **Assertion Status** also requires **splitting**. However, the main difference is that will need to decide if the context of the sentence of the entity is enough or you want to provide with more. That should be taken into account to decide either to go with **sentence splitting** or with **paragraph splitting**. Usually, sentence splitting should suffice.

Assertion Status always goes after **Entity Recognition (NER)**.





STATE OF THE ART

Entity Resolution Legal NLP

Entity Resolution



Entity Resolution is the NLP Task in charge of, given an **NER chunk, retrieve the most semantically similar candidate** from a training set the model has been trained on. But it is much more than a *Text Similarity task*, **it can store unique IDs** so that, after the sentence similarity task, it retrieves not only the most similar **name, but also an ID**.

It requires **Sentence Embeddings**

This **LOAN AGREEMENT**, dated as of **November 17, 2014** (this “**Agreement**”), is made by
and among **Auxilium Pharmaceuticals, Inc.**, a corporation incorporated under the laws of the State of

Auxilium Pharmaceuticals, Inc (from NER)

- Normalized name (Edgar): **AUXILIUM PHARMACEUTICALS INC**
- Unique ID (Edgar): **0001182128**

Entity Resolution

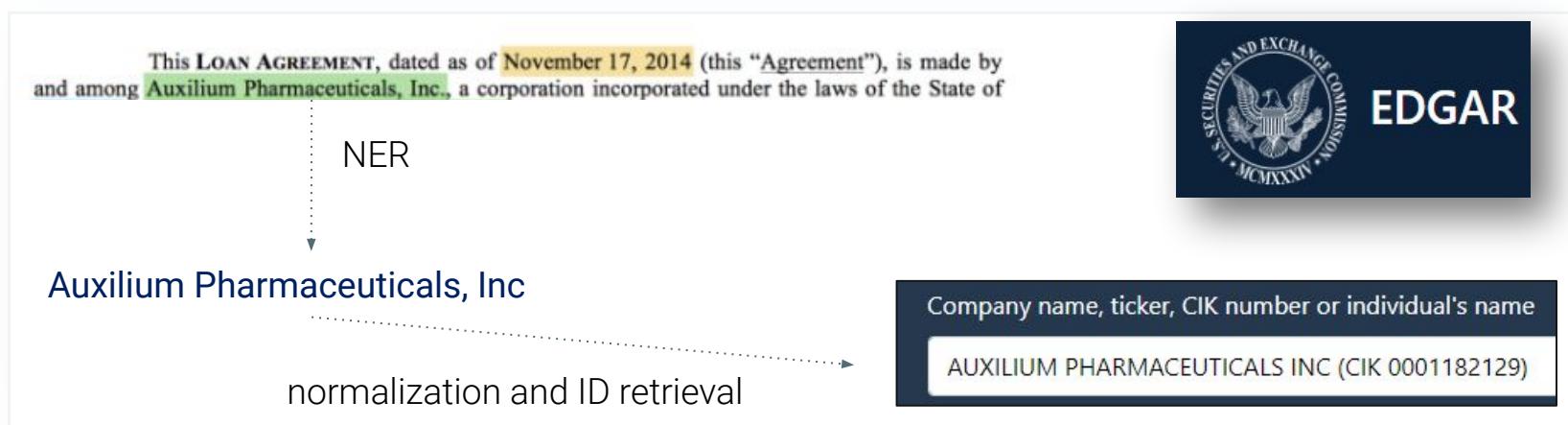


Entity Resolution is the NLP Task in charge of, given an **NER chunk, retrieve the most semantically similar candidate** from a training set the model has been trained on. But it is much more than a *Text Similarity* task, **it can store unique IDs** so that, after the sentence similarity task, it retrieves not only the most similar **name, but also an ID**.

It requires **Sentence Embeddings**.

This has been widely used for retrieving **normalized versions** of, for example, **company names** (which can have many version as *INC, Inc., inc.*, different punctuation, etc) and their **unique ID**, as for example, their CIK in Edgar Database

Entity Resolution goes always after **Name Entity Recognition (NER)**.





STATE OF THE ART

Data Augmentation with Chunk Mappers Legal NLP

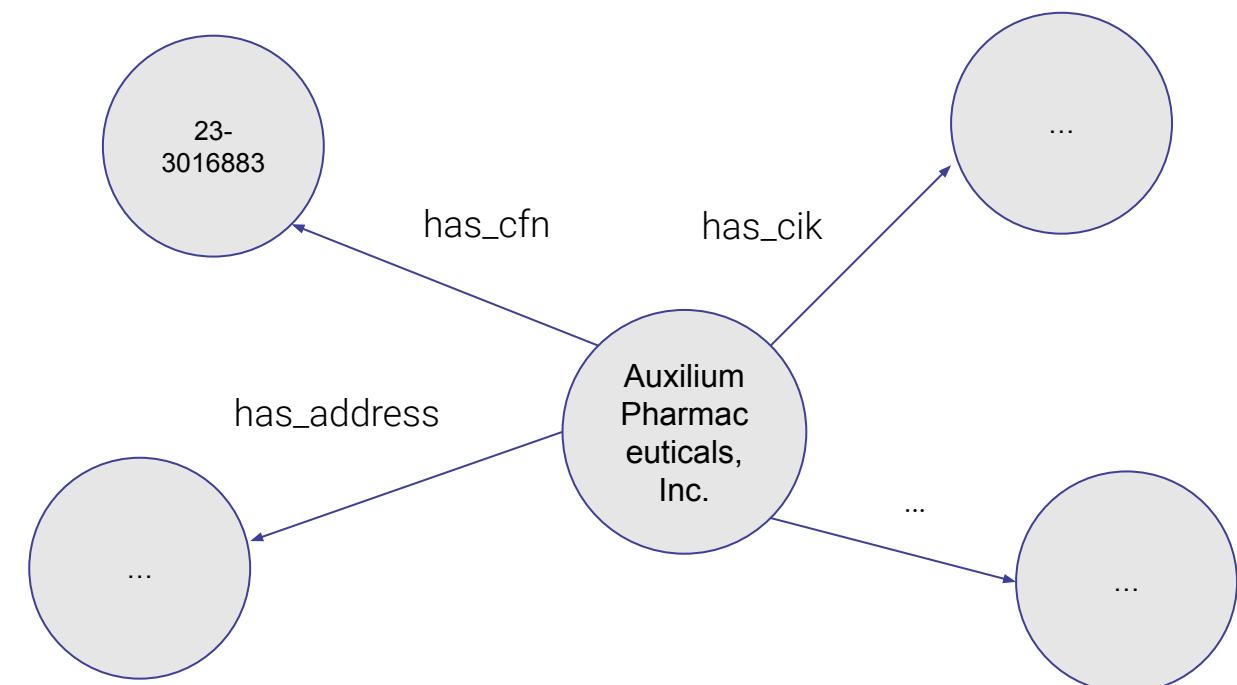
Data augmentation with Chunk Mappers

Given an **NER chunk extracted in NER**, and a **dictionary** in json format, you can use the NER chunks as a **key to retrieve the values from a dictionary** in form of relationships.

Example:

This LOAN AGREEMENT, dated as of November 17, 2014 (this "Agreement"), is made by
and among Auxilium Pharmaceuticals, Inc., a corporation incorporated under the laws of the State of

```
"mappings": [  
  {  
    "key": "Auxilium Pharmaceuticals, Inc.",  
    "relations": [  
      {  
        "key": "has_cfn",  
        "values" : ["23-3016883"]  
      },  
      ...  
    ]  
  }]
```



Data augmentation with Chunk Mappers

Given an **NER chunk extracted in NER, and a dictionary** in json format, you can use the NER chunks as a **key to retrieve the values from a dictionary** in form of relationships.

Chunk Mappers always go after **Entity Resolution**, because in your json file you should have unique keys. That means you should not save in a Chunk Mapper both *Auxilium Pharmaceuticals* and *Auxilium Pharmaceuticals Inc*, **you should only stored the normalized / official version** (AUXILIUM PHARMACEUTICALS INC, as per Edgar Database) in the json. And then, after NER, you carry out **normalization with Entity Resolvers**, and then Chunk Mapping to retrieve the rest of information.





STATE OF THE ART

**Question & Answering
Legal NLP**

Legal NLP Question Answering

	Entity	Question
0	DATE	['When was the company acquisition?', 'When was the company purchase agreement?']
1	ORG	['Which company?', 'Which was the company acquisition?']
2	STATE	['Which state?']
3	AGREEMENT	['What kind of agreement?']
4	LICENSE	['What kind of license?']
5	LICENSE_REC	['To whom the license is granted?']
6	ALIAS	['Which is the alias?']

In February 2017, the Company entered into an asset purchase agreement with NetSeer, Inc..

DATE

AGREEMENT

ORG

The Company hereby grants to Seller a perpetual, non-exclusive, royalty-free license

Seller
LICENSE_RECIPIENT

perpetual
LICENSE

non-exclusive
LICENSE

royalty-free
LICENSE

On March 12, 2020 we closed a Loan and Security Agreement with Hitachi Capital America Corp. (also known as "Hitachi").

DATE

AGREEMENT

ORG

ALIAS

Legal NLP Question Answering

Question Answering is the NLP Task in charge of, given a **question**, **retrieve an answer**. **There are two main groups of QA models:**

- **Open book**: We provide also with a context where to look.
- **Closed book**: The knowledge is stored in the Language Model and you don't give any example.

We use the *Open-book* approach, as **we want to retrieve answers in our specific documents**.

These models are **NLI**-based (*Natural Language Inference*). They use the question as a **hypotheses**, and try to find the maximum number of consequent tokens which maximize the probability to be an **answer** to that hypotheses.

Premise	Hypotheses	Inference Results
In February 2017, the Company entered into an asset purchase agreement with NetSeer, Inc.	The Agreement is an Asset Purchase Agreement.	Entailment
	The Company entered into agreement in March 2020.	Contradiction
	The Company is John Snow Labs, Inc.	Neutral

	Entity	Question
0	DATE	['When was the company acquisition?', 'When was the company purchase agreement?']
1	ORG	['Which company?', 'Which was the company acquisition?']
2	STATE	['Which state?']
3	AGREEMENT	['What kind of agreement?']
4	LICENSE	['What kind of license?']
5	LICENSE_REC	['To whom the license is granted?']
6	ALIAS	['Which is the alias?']

On March 12, 2020 we closed a Loan and Security Agreement with Hitachi Capital America Corp . (also known as "Hitachi".)
DATE AGREEMENT ORG ALIAS



STATE OF THE ART

Automatic Question Generation Legal NLP

Legal NLP NER QA-based: Automatic Question Generation

One of the main restrictions of Zero-shot NER or QA is that you need the question before hand. Fortunately, there is a way you can generate **questions (prompt) on the fly**.

1. Imagine you want to extract:

The Buyer shall use such materials and supplies only in accordance with the present agreement

2. **NerQuestionGenerator** annotator can generate questions for you if you have an NER with retrieves **SUBJECT** and **VERB**, which is much easier to train (for example, **legner Obligations**)

The Buyer shall use

SUBJECT VERB

```
qagenerator = legal.NerQuestionGenerator()\n.setInputCols(["ner_chunk"])\n.setOutputCol("question")\n.setQuestionMark(False)\n.setQuestionPronoun("What")\n.setEntities1(["OBLIGATION SUBJECT"])\n.setEntities2(["OBLIGATION ACTION"])
```

```
+-----+\n|result|\n+-----+\n|[What Buyer shall use ]|\n+-----+
```

we send it to QA

```
qa =nlp.BertForQuestionAnswering.pretrained("legqa_bert_large","en", "legal/models")\n.setInputCols(["question", "document"])\n.setOutputCol("answer") \n.setCaseSensitive(True)
```

such materials and supplies only in accordance with the present agreement



STATE OF THE ART

Deidentification Legal NLP

De-Identification



DATE: 2020-02-01 10:00:00 AM

AGREEMENT NUMBER: 1234567

26 Mar 2022
IN WITNESS WHEREOF, the Parties
have duly executed this Agreement as
of the date first written above.
ARMSTRONG FLOORING, INC.

By: /s/
Donald R. Maier Title: President and
Chief Executive Officer

Detect
sensitive
entities

DATE: 2020-02-01 10:00:00 AM

AGREEMENT NUMBER: 1234567

26 Mar 2022
IN WITNESS WHEREOF, the
Parties have duly executed this
Agreement as of the date first written
above. ARMSTRONG FLOORING, INC.

By: /s/
Donald R. Maier Title: President
and Chief Executive Officer

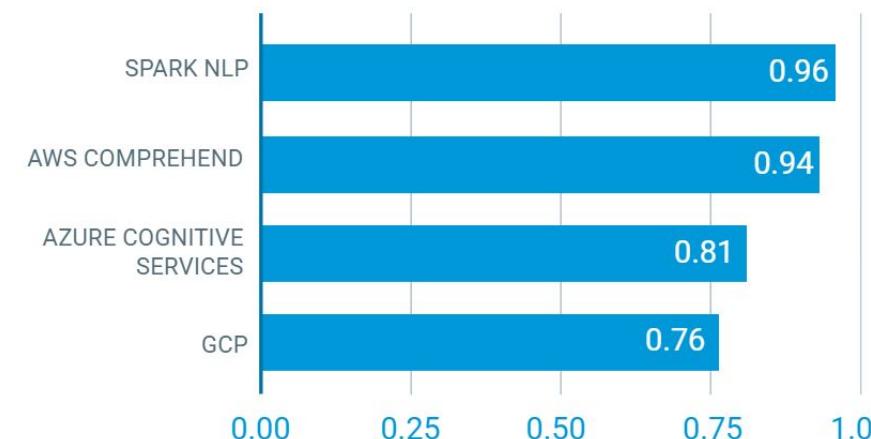
Transform
sensitive
entities

DATE: 2020-02-01 10:00:00 AM

AGREEMENT NUMBER: 1234567

16 Jan 2019
IN WITNESS WHEREOF, the
Parties have duly executed this
Agreement as of the date first written
above. AUXILIUM, INC.

By: /s/
David Bill Title: Director and Chief
Accounting Officer



Legal NLP Deidentification



Deidentification is the NLP task in charge of:

- 1) **Masking NER chunks or Obfuscating (faking) with synthetic data;**
- 2) **Returning an anonymized version** of the text;

It works on top of **NER** and **ContextualParser**, with specific **Deidentification** annotators which retrieve the NER chunks and mask / obfuscate them, all along with some other capabilities as *Language*, *Masking Technique*, *Date shift selection*, etc.

	Sentence	Masked	Masked with Chars	Masked with Fixed Chars	Obfuscated
0	CARGILL, INCORPORATED		[*****]	****	TURER INC
1	By: Pirkko Suominen	By:	By: [*****]	By: ****	By: SESA CO.
2	Name: Pirkko Suominen Title: Director, Bio Technology Development Center, Date: 10/19/2011	Name: : Center, Date:	Name: [*****]; [*****] Center, Date: [*****]	Name: ****: **** Center, Date: ****	Name: John Snow Labs Inc: Sales Manager Center, Date: 03/08/2025
3	BIOAMBER, SAS	,	[*****], [*]	****, ****	Clarus Ilc., SESA CO.
4	By: Jean-François Huc	By:	By: [*****]	By: ****	By: JAMES TURNER
5	Name: Jean-François Huc Title: President Date: October 15, 2011\n\nemail : jeanfran@gmail.com...	Name: : Date:\n\nemail :\n\ncphone : 0	Name: [*****]: [*****]Date: [*****]\n\nemail : [*****]\n\ncphone : ...	Name: ****: ****Date: ****\n\nemail :\n\ncphone : ****0	Name: MGT Trust Company, LLC: Business ManagerDate: 11/7/2016\n\nemail : Berneta@hotmail.com)\n...



STATE OF THE ART

**Text and Layout information
Legal NLP and Visual NLP**

Visual classification



Sometimes textual information is not enough to classify a document. For example, let's suppose you have to classify 2 types of document with the same content, but only differing in the layout disposition of the information.

If we just get the text from them, and the contents are the same, Legal NLP by itself may get confused. For this, we have 2 ways to go:

Legal NLP with Vision Transformers

Image Level

Don't use text at all. Use **Visual Transformers** (**ViT models**) to transform only at image-level.

Characters consist of pixels, so they will be taken into account. Not a **language-level**, but a **pixel-level**.

The image shows a grid of 20 document samples, each labeled with its category below it. The categories are: letter, memo, email, filefolder, form, handwritten, invoice, advertisement, budget, news article, presentation, scientific publication, questionnaire, resume, scientific report, and specification. The documents include various types of legal and administrative forms, contracts, invoices, and reports.

Inconvenient: If you need the text to do NLP afterwards, maybe it's quicker to use the previous approach

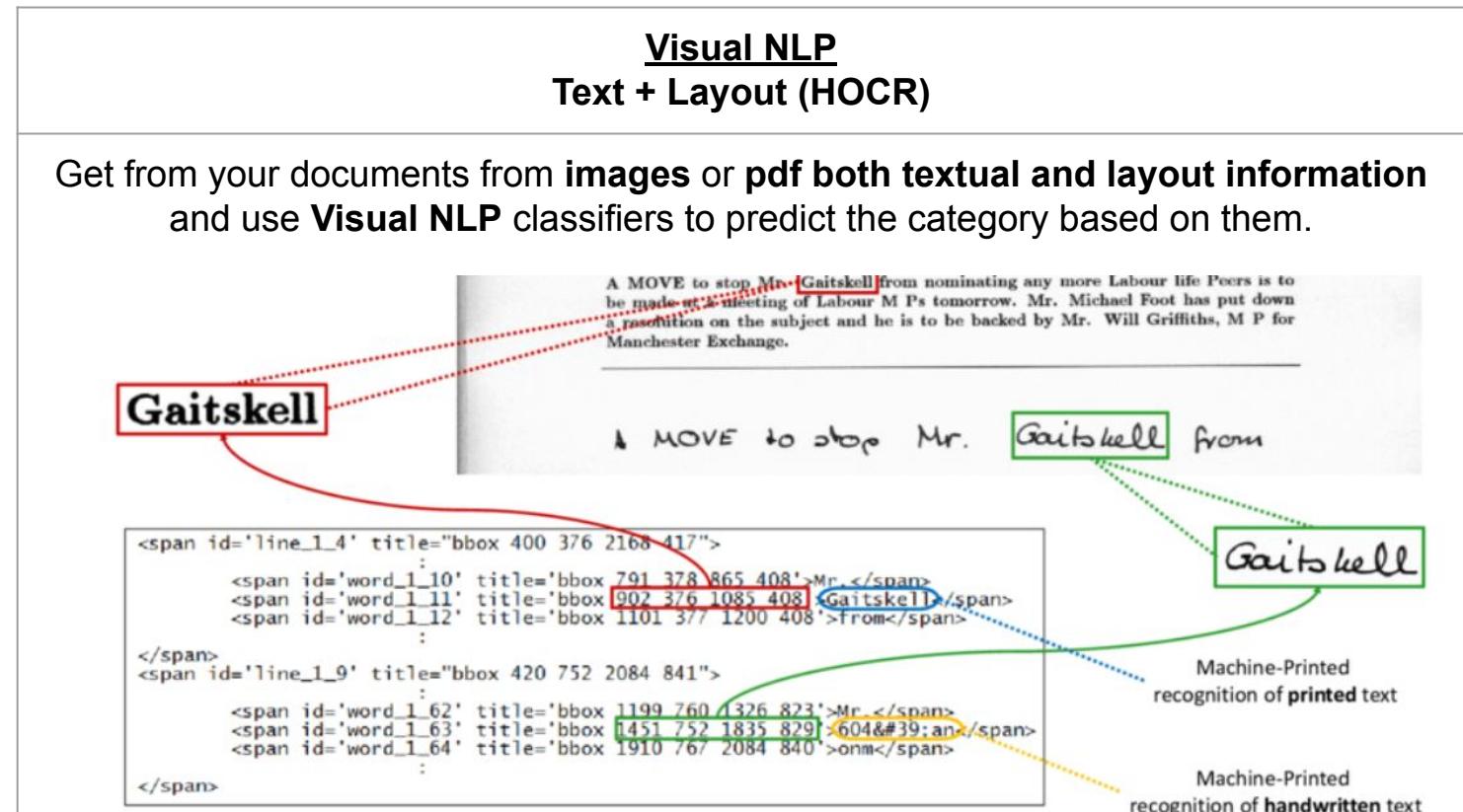
Visual classification

Sometimes textual information is not enough to classify a document. For example, let's suppose you have to classify 2 types of document with the same content, but only differing in the layout disposition of the information.

If we just get the text from them, and the contents are the same, Legal NLP by itself may get confused. For this, we have 2 ways to go:

Visual NLP
Text + Layout (HOCR)

Get from your documents from **images** or **pdf** both **textual and layout information** and use **Visual NLP** classifiers to predict the category based on them.



Inconvenient: This approach uses OCR and tables, handwritten text, images, etc. may be ignored

52

Other Visual NLP capabilities



Document Classification

Classified Image

This AGREEMENT (the "Agreement") is entered into effective as of _____
Date _____, by and between _____, a _____ corporation ("Party A").
In consideration of the mutual covenants herein contained and other good and valuable
consideration, the parties agree as follows:
1. [Statement of Business Relationship, Rights, Obligations] Subject to the terms and
conditions of this Agreement,
2. [Item Details etc.], Party A agrees to pay Party B:
3. [Obligation conditions]
4. [Other terms conditions]
5. [Term and Termination] The initial term of this Agreement will be for _____ months.
Thereafter, the term will be for one year unless otherwise specified in writing by either party. Either
party may terminate this Agreement at any time before its final term or any renewal term is
commenced, for any reason, but not less than at all, provided that at least _____ days' notice or written
notice of termination is given to the non-terminating party by the terminating party.
6. [Applicable Law, Disputes] This Agreement will be governed by and construed in
accordance with the laws of the State of _____, notwithstanding any conflict of
laws principles that might otherwise govern or construction of this Agreement in
the event of such a conflict. This Agreement will at all times and in all events be construed as a
whole, according to its meaning, and not merely by the separate parts.
7. [Cross-references] This Agreement is intended to be contemporaneous, with the same effect as
if both parties had signed the same document. All such contemporaneity will be deemed as original
with the content of original and will constitute one and the same document.
8. [Entire Agreement, Amendment] This Agreement constitutes the entire understanding
between the parties and supersedes all prior negotiations, understandings, writings, representations, and
understandings, oral and written, and all other communications between the parties relating to
the subject matter hereof. This Agreement cannot be amended or otherwise modified except in
writing that is countersigned by all of the parties.
9. [Party B Seal] This agreement will be binding upon, and have the benefit of, each of
the parties hereto in the manner applicable to them and their respective successors and assigns.
10. [Initial Understanding] Each party has read this entire Agreement fully and made the
contents thereof known to his/her/its attorney or legal advisor or to his/her/its wife, and is
relying on the advice of any such he or she or it. This Agreement reflects the mutual
understanding of the parties with respect to all subjects herein addressed herein and will be
construed so as to reflect.

Classification

This document has been classified as: **Agreement**
Classification Confidence: **99.6%**

From images, pdf, docx, ppt...

EXHIBIT 1A.1a

DRAFT (Amended 1/12/00 Rev 1) SUPPORT AND MAINTENANCE AGREEMENT

This Support and Maintenance Agreement ("Agreement") is entered into as of the _____ day of _____, 2000 (the "Effective Date") by and between XACCT Technologies, Inc., a Delaware corporation ("XACCT") with its principal place of business at 2900 Lakeside Drive, Suite 100, Santa Clara, California 95054 and _____, a _____ corporation ("Licensee") with its principal place of business at _____.

The Agreement is to bind the parties and their successors in interest which XACCT will provide its products, services, support and support services (as differentiated for the Product which is licensed by Licensee) on an as-needed basis. License Agreement ("License Agreement") is used herein to mean the License Agreement, all other terms and conditions of the License Agreement are incorporated by reference. Capitalized terms that are not defined in Section 1, below or elsewhere in this Agreement have the same meaning as in the License Agreement.

1. DEFINITIONS

1.1 "Approved License Contract" means a license agreement which is not required to contact the XACCT support center.

1.2 "Customer" means a single, discrete organization or other entity which is required to contact the XACCT support center.

... to plain text

EXHIBIT 10.16

DRAFT (Americas) 1/12/00 (Rev 1) SUPPORT AND MAINTENANCE AGREEMENT

This Support and Maintenance Agreement ("Agreement") is entered into and is effective as of the _____ day of _____, 2000 (the "Effective Date") by and between XACCT Technologies, Inc., a Delaware corporation ("XACCT") with its principal place of business at 2900 Lakeside Drive, Suite 100, Santa Clara, California 95054 and _____, a _____ corporation ("Licensee") with its principal place of business at _____.



Object detection



Thank you!