

K SCHOOL

Máster de *Data Science*

Módulo de Procesamiento de Lenguaje Natural

Elena del Olmo



- **Introducción a la Lingüística.**
- **NLTK:** *An amazing library to play with natural language.*
- **SpaCy:** *Industrial-strength Natural Language Processing.*
- **Textacy:** *Higher-level NLP built on spaCy.*
- **RasaNLU:** *Turn natural language into structured data.*
- **NeuralCoref:** *Coreference Resolution in spaCy with Neural Networks.*
- **Word2Vec:** *A group of related models used to produce word embeddings.*

Los niveles de conocimiento lingüístico

- Fonético
- Morfológico
- Sintáctico
- Semántico
- Pragmático
- Discursivo
- Conocimiento del mundo

La sintaxis, la semántica y la pragmática

- 📌 *El lenguaje es uno de los principales elementos del comportamiento humano y un componente crucial para nuestras vidas.*
- 📌 *Las ranas verdes tienen la nariz grande.*
- 📌 *Las ideas verdes tienen la nariz grande.*
- 📌 *La las tienen verdes grandes nariz ideas.*

Datos, información y conocimiento

- Los **datos** constituyen la unidad mínima de significado, pero por sí mismos son irrelevantes. No son más que conjuntos de valores. Los datos constituyen la unidad mínima de significado, pero por sí mismos son irrelevantes. No son más que conjuntos de valores.

Por ejemplo, cada transacción económica que realizamos deja numerosas “huellas” en forma de datos, pero un hipotético fichero que almacenara la ristra de cantidades intercambiadas no tendría ningún valor.

Datos, información y conocimiento

- Sin embargo, el procesamiento e interrelación de los datos es lo que genera la **información**. No hay información sin datos, pero la información debe ser potencialmente útil para un cierto objetivo.

Los bancos tienen la potencialidad de generar información, ya que pueden tomar los datos de las transacciones a los que tienen acceso, seleccionar solo las transacciones positivas o negativas, solo las de una determinada persona, las que superen una cierta cantidad, categorizarlas por tipos (compra de bienes inmuebles, compra de productos de lujo, donaciones, etc.), relacionar las transacciones de los miembros de una misma familia, de cada persona junto con quienes han ejercido de avales de la misma, etc.

Datos, información y conocimiento

- El **conocimiento** es fruto de la integración de la información a partir de la experiencia y es la base para la toma de decisiones. Para que exista conocimiento es necesario que exista información y que quien la interprete sea un profesional.

Los expertos de un banco, con el objetivo de evaluar el riesgo de préstamo en mente, podrían seleccionar el historial de compra de una persona y de sus posibles avales y organizarlo en función del tipo de bienes que consideren relevantes. También podrían decidir evaluar la frecuencia de sus ciclos de ingresos/gastos, con qué frecuencia cobran distintas cantidades en su nómina, cuántas veces han cambiado de trabajo en los últimos años, etc. Su experiencia les sirve tanto para seleccionar qué información considerar relevante como para decidir si se concederá un préstamo a una determinada persona.

Sistemas basados en conocimiento

- La descripción de los sistemas de IA en términos intencionales (por ejemplo decir que el robot sabe dónde están las piezas del juego y que quiere ganar la partida) es una metáfora de los sistemas basados en conocimiento, que manejan conjuntos de representaciones simbólicas llamadas **bases de conocimiento**.
- Cuando construimos un sistema basado en conocimiento, **queremos que actúe teniendo en cuenta las creencias que tiene sobre el mundo**, no solo sobre el conocimiento representado de manera explícita en su base de conocimiento.


```
imprimeColor(nieve) :- !, write("Es blanco.").
imprimeColor(hierba) :- !, write("Es verde.").
imprimeColor(cielo) :- !, write("Es amarillo.").
imprimeColor(Cosa) :- write("No lo sé :-/").
```

```
imprimeColor(Cosa) :- color(Cosa, Color), !, write("Es "),
                        write(Color), write(".").
imprimeColor(Cosa) :- write("No lo sé :-/").

color(nieve, blanco).
color(cielo, amarillo).
color(Cosa, Color) :- estaHechoDe(Cosa, Material),
                        color(Material, Color).

estaHechoDe(hierba, vegetacion).
color(vegetacion, verde).
```

Procesamiento del Lenguaje Natural

- Los ordenadores se las apañan bien para manejar **datos estructurados**, como las tablas de una Base de Datos. Pero, desafortunadamente, los humanos nos comunicamos con palabras.
- El Procesamiento del Lenguaje Natural es una subdisciplina de la Inteligencia Artificial cuyo objetivo es que las máquinas entiendan el **lenguaje humano**.

La ambigüedad

- **La desambiguación parece mucho más sencilla en el uso cotidiano de la lengua si las comparamos con el coste que conlleva su tratamiento computacional.**

Si por ejemplo tratamos de generar todo el conjunto de árboles posibles ante una sucesión de palabras dada basándonos en una gramática generativa con etiquetas morfológicas como piezas (del tipo “el sintagma nominal se reescribe como una secuencia de determinante más sustantivo”), nos daremos cuenta de que, como hablantes nativos que somos, no consideramos todas las opciones:

Nadie entendería que, cuando le dicen “está cansado de que siempre la coja de ahí arriba”, se estén refiriendo a “det(coja/NN, la), prep(coja, de), pcomp(de, ahí)”, pero un programa sin conocimiento superior al morfosintáctico generaría también este árbol.

Interacción entre capas

No dijo que comería == $ROOT(X, dijo),$
 $ccomp(dijo, comería), \underline{mark}(comería,$
 $que), \underline{neg}(dijo, no) ==$

	Come	Dice que come
Sí	?	✗
No	?	✓

Dijo que no comería == $ROOT(X, dijo),$
 $ccomp(dijo, comería), \underline{mark}(comería,$
 $que), \underline{neg}(comería, no) ==$

	Come	Dice que no come
Sí	?	✓
No	?	✗

No puede valer == $ROOT(X, valer),$
 $\underline{aux}(valer, puede), \underline{neg}(valer, no) ==$

	Vale	Puede valer
Sí	✗	✗
No	✓	✓

Puede no valer == $ROOT(X, valer),$
 $\underline{aux}(valer, puede), \underline{neg}(puede, no) ==$

	Vale	Puede no valer
Sí	?	✓
No	?	✗

Divide y vencerás

- Cuando queramos tratar fenómenos lingüísticos complejos, que requieren varios niveles de formalización, debemos dividir la tarea en partes que abordaremos individualmente.

¡Un problema salvaje apareció!

Madrid es la capital de España y de la Comunidad de Madrid. Es la ciudad más poblada de España. En su área metropolitana viven más de seis millones de personas, lo que la convierte en la tercera o cuarta área metropolitana de la Unión Europea, por detrás de las de París y Londres. La capital de España es también la tercera ciudad más poblada de la Unión Europea, por detrás de Berlín y Londres.

Madrid es la **capital de España y de la Comunidad de Madrid**. **Es** la **ciudad más poblada de España**. En su **área metropolitana viven más de seis millones de personas**, lo que **la** convierte en la **tercera o cuarta área metropolitana de la Unión Europea**, por detrás de las de París y Londres. La **capital de España** es también la **tercera ciudad más poblada de la Unión Europea**, por detrás de Berlín y Londres.

Paso 1: **segmentación en frases**

1. *Madrid es la capital de España y de la Comunidad de Madrid.*
2. *Es la ciudad más poblada de España.*
3. *En su área metropolitana viven más de seis millones de personas, lo que la convierte en la tercera o cuarta área metropolitana de la Unión Europea, por detrás de las de París y Londres.*
4. *La capital de España es también la tercera ciudad más poblada de la Unión Europea, por detrás de Berlín y Londres.*

Paso 2: **segmentación en *tokens***

[“Madrid”, “es”, “la”, “capital”, “de”, “España”, “y”,
“de”, “la”, “Comunidad”, “de”, “Madrid”, “.”]

Paso 3: etiquetado morfológico

“Madrid”



**Modelo predictivo de
PoS pre-entrenado**



NOMBRE_PROPIO

Paso 4: lematización

Madrid → **Madrid**

es → **ser**

la → **el**

capital → **capital**

de → **de**

España → **España**

y → **y**

de → **de**

la → **el**

Comunidad → **Comunidad**

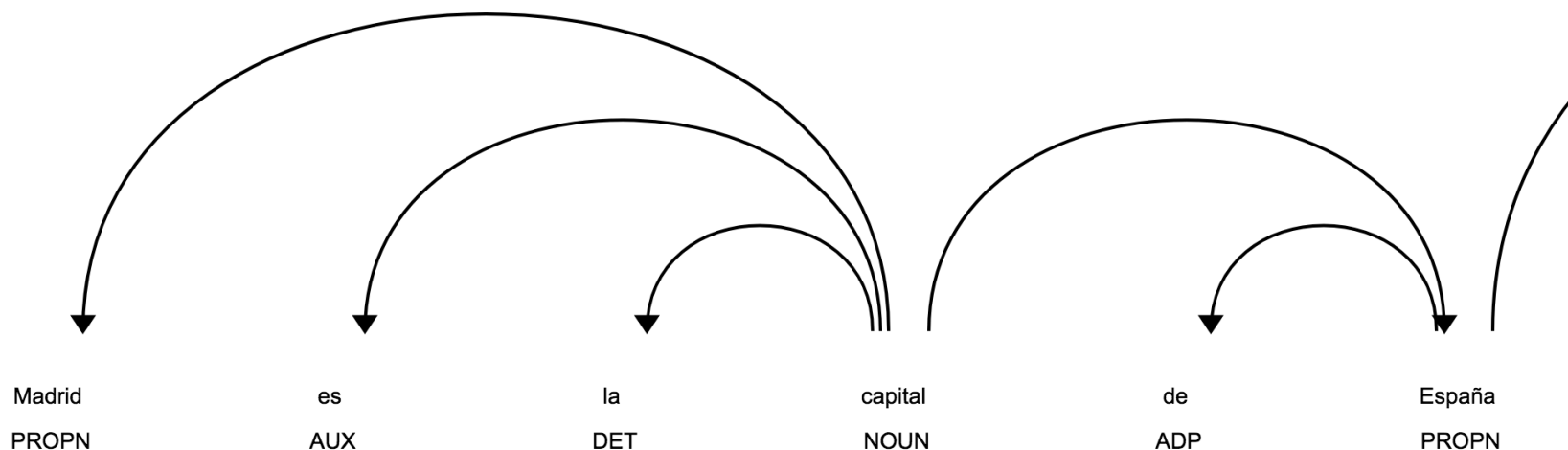
de → **de**

Madrid → **Madrid**

Paso 5: eliminación de *stop words*

["Madrid", "es", "la", "capital", "de", "España",
"y", "de", "la", "Comunidad", "de", "Madrid", "."]

Paso 6: etiquetado dependencial



Paso 6: **localización de sintagmas nominales**

1. Madrid es la capital de España y de la Comunidad de Madrid.
2. Es la ciudad más poblada de España.
3. En su área metropolitana viven más de seis millones de personas, lo que la convierte en la tercera o cuarta área metropolitana de la Unión Europea, por detrás de las de París y Londres.
4. La capital de España es también la tercera ciudad más poblada de la Unión Europea, por detrás de Berlín y Londres.

Paso 7: reconocimiento de entidades nombradas



```
displacy.render(doc2, style='ent', jupyter=True)
```



Madrid LOC es la capital de España LOC y de la Comunidad de Madrid LOC .

Paso 7: reconocimiento de entidades nombradas

Nombres de personas

Nombres de empresas

Localizaciones geográficas

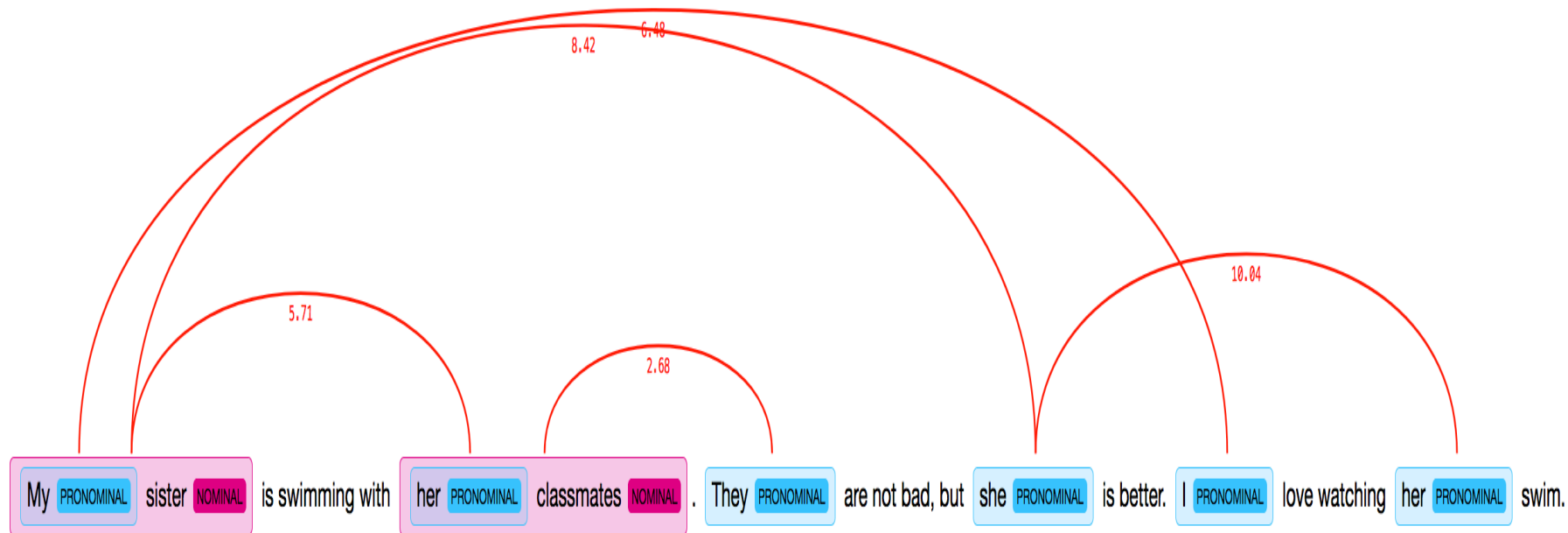
Marcas

Fechas y horas

Importes

Eventos

Paso 8: resolución de correferencias



¡Ayuda!

El Procesamiento del Lenguaje Natural se divide fundamentalmente en tres campos de conocimiento, en función de la tecnología utilizada por las herramientas:

- PLN basado en **reglas**.
- PLN **estadístico**.
- PLN basado en **redes neuronales**.

PLN basado en reglas

La metodología basada en reglas consiste en usar reglas (gramaticales, lógicas, etc.) para inferir información de textos. Casos típicamente resueltos mediante este tipo de enfoque son:

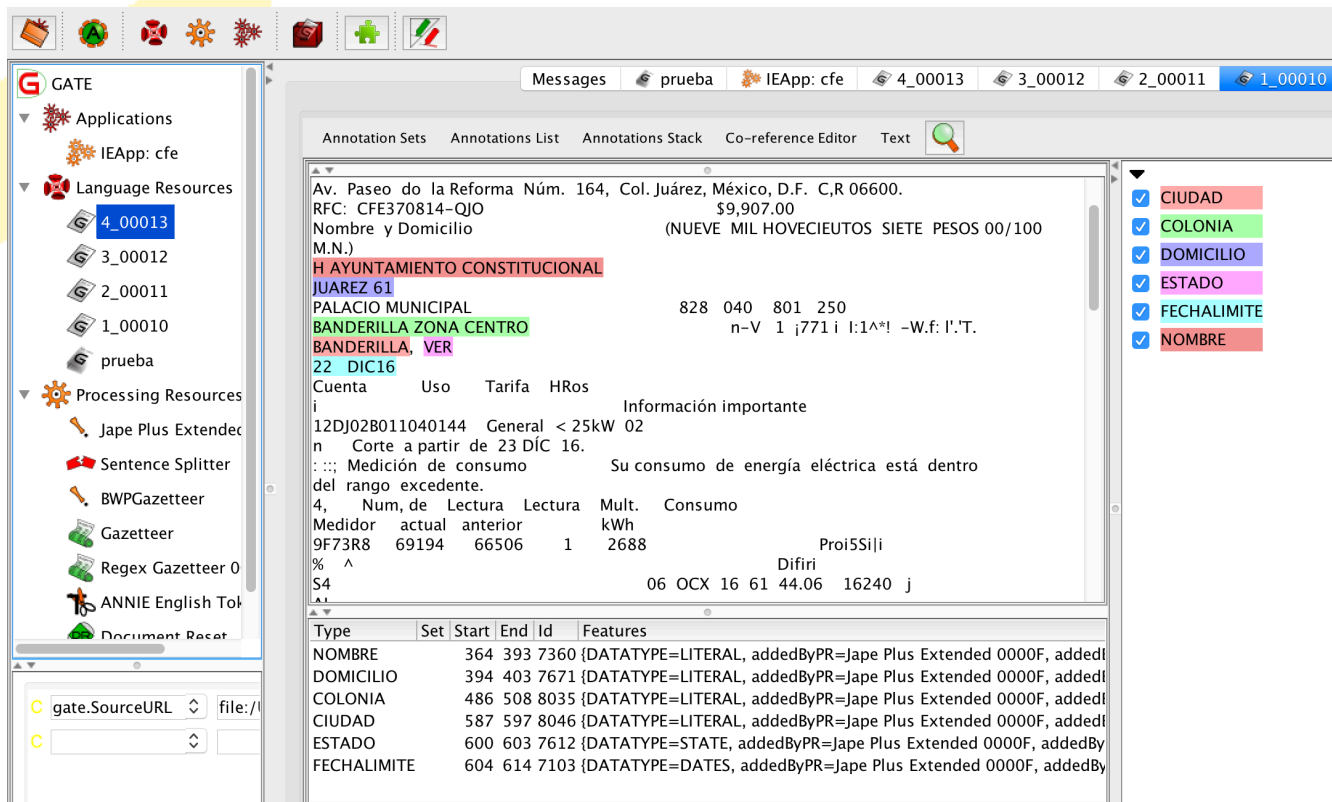
- Segmentación en párrafos
- Segmentación en frases.
- Tokenización
- *Mayusculado verdadero*
- Reconocimiento de entidades nombradas
- Linkado ortográfico
- Etiquetado morfológico
- Etiquetado dependencial
- Resolución de correferencias
- Análisis de sentimientos
- Lematización
- Detección de idioma
- Traducción automática
- Desambiguación

PLN basado en reglas

Hay varias herramientas de código abierto que permiten funcionalidades de este tipo, por ejemplo:

- [Apache UIMA](#), un entorno de trabajo para Java.
- [NLTK](#), una librería para Python.
- [CoreNLP](#), un conjunto de herramientas para Java.
- [GATE](#), con una versión de escritorio e incrustado en Java.

GATE



The screenshot shows the GATE software interface. The left sidebar lists the project structure, including 'Applications' (IEApp: cfe), 'Language Resources' (4_00013, 3_00012, 2_00011, 1_00010, prueba), and 'Processing Resources' (Jape Plus Extended, Sentence Splitter, BWP Gazetteer, Gazetteer, Regex Gazetteer, ANNIE English Tool, Document Reset). The main text area displays a document snippet with various annotations. The right sidebar shows a list of features with checkboxes.

Annotations List:

Type	Set	Start	End	Id	Features
NOMBRE	364	393	7360	{DATATYPE=LITERAL, addedByPR=Jape Plus Extended 0000F, addedl	
DOMICILIO	394	403	7671	{DATATYPE=LITERAL, addedByPR=Jape Plus Extended 0000F, addedl	
COLONIA	486	508	8035	{DATATYPE=LITERAL, addedByPR=Jape Plus Extended 0000F, addedl	
CIUDAD	587	597	8046	{DATATYPE=LITERAL, addedByPR=Jape Plus Extended 0000F, addedl	
ESTADO	600	603	7612	{DATATYPE=STATE, addedByPR=Jape Plus Extended 0000F, addedBy	
FECHALIMITE	604	614	7103	{DATATYPE=DATES, addedByPR=Jape Plus Extended 0000F, addedBy	

[Guía de usuario oficial](#)

PLN estadístico

- Los modelos estadísticos se basan en un conjunto de **datos de entrenamiento** y un **set de pruebas**.
- El **modelo devuelve la probabilidad** por la que el **input** de prueba se ajusta al conjunto de datos de entrenamiento, escogiéndose finalmente la opción que presente una **mayor probabilidad**.
- A los modelos de PLN estadístico subyace la asunción de que **no podrían listarse todas las reglas** que dieran cuenta de un modelo lingüístico, o bien la asunción de que sí es posible pero sería demasiado costoso.

PLN estadístico

Casos típicamente resueltos mediante este tipo de enfoque son:

- Clasificación de documentos
- Extracción de tópicos
- Generación automática de resúmenes
- Análisis de sentimientos
- Traducción automática

PLN estadístico

Existen varios algoritmos para estos modelos, basados en los modelos bayesianos (es decir, en grafos acíclicos dirigidos):

- Los modelos ocultos de Márkov.
- El modelo de Márkov de máxima entropía.
- Los campos aleatorios condicionales, especialmente útiles para la segmentación y etiquetado de datos estructurados.

Más información sobre PLN estadístico en [este libro](#).

PLN estadístico

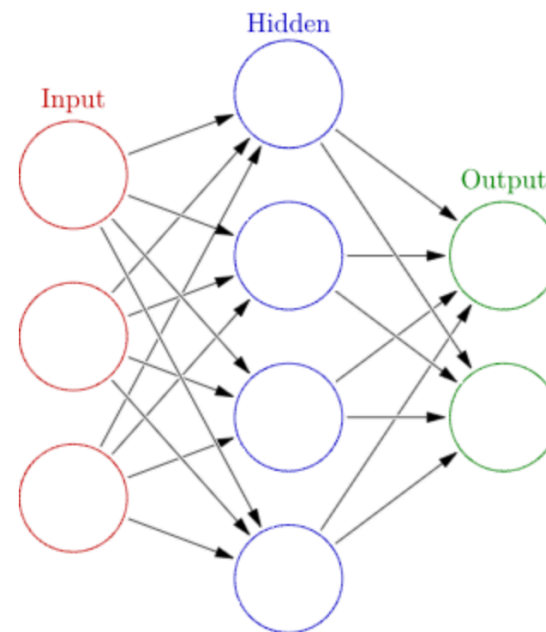
Las principales herramientas para entrenar modelos de PLN estadístico son:

- [MALLET](#), basado en Java ([tutorial](#) y [presentación](#)).
- [Apache UIMA](#), ataca de nuevo.
- [CoreNLP](#), ataca de nuevo.
- [KenLM](#), basado en Python.

PLN basado en redes neuronales

Una red neuronal es un **modelo computacional no lineal** basado en la estructura del **cerebro** que es capaz de llevar a cabo tareas complejas.

- Consiste en una serie de **neuronas artificiales** organizadas en **tres capas** interconectadas.
- Cada neurona artificial tiene **entradas ponderadas**, **funciones de activación** y **salidas**.
- Las **sinapsis** vendrían a ser lo que hacen del modelo un sistema parametrizado.



PLN basado en redes neuronales

Casos típicamente resueltos mediante este tipo de enfoque son:

- Clasificación de documentos
- Extracción de tópicos
- Generación automática de resúmenes
- Análisis de sentimientos
- Traducción automática

PLN basado en redes neuronales

Las principales herramientas para entrenar modelos neuronales en PLN son:

- [Keras](#), en Python.
- [TensorFlow](#), con API estable para varios lenguajes.
- [DL4J](#), en Java.