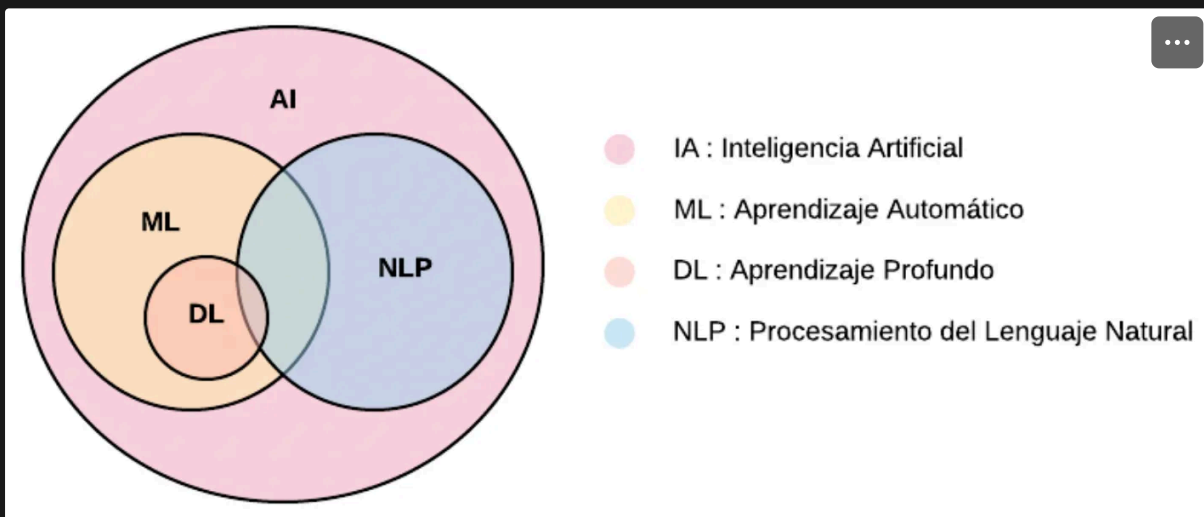


# Procesamiento del Lenguaje Natural



## Introducción

El Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) es una rama interdisciplinaria de la inteligencia artificial que se centra en la interacción entre las computadoras y el lenguaje humano. Su objetivo es permitir que las máquinas comprendan, interpreten y generen lenguaje de una manera que sea valiosa.



El Procesamiento del Lenguaje Natural o NLP es una disciplina que se encuentra en la intersección de varias ciencias, tales como las Ciencias de la Computación, la inteligencia artificial y Psicología Cognitiva. Su idea central es la de darle a las máquinas la capacidad de leer y comprender los idiomas que hablamos los humanos. La investigación del Procesamiento del Lenguaje Natural tiene como objetivo responder a la pregunta de cómo las personas son capaces de comprender el significado de una oración oral / escrita y cómo las personas entienden lo que sucedió, cuándo y dónde sucedió; y las diferencias entre una suposición, una creencia o un hecho.

Los elementos comunes de cualquier arquitectura completa de un sistema para el Procesamiento del Lenguaje Natural son:

- **Reconocimiento de voz:** Convertir una palabra hablada en un conjunto de palabras. Las palabras habladas se componen de una serie de parámetros relacionados con el sentido de la audición.

- **Comprensión del lenguaje:** El objetivo de este elemento es generar un significado para las palabras habladas, y ese significado será utilizado por el siguiente elemento (gestión del diálogo).
- **Gestión del diálogo:** La tarea principal de este elemento es coordinar y mantener unidas todas las partes del sistema y los usuarios, y conectarse con otros sistemas.
- **Comunicación con sistemas externos:** como sistemas expertos, sistemas de bases de datos u otras aplicaciones informáticas.
- **Generación de respuesta:** Establecer el mensaje que el sistema debe entregar.
- **Salida de voz:** Uso de diferentes técnicas para producir el mensaje desde el sistema.

En general, en Procesamiento del Lenguaje Natural se utilizan seis niveles de comprensión con el objetivo de descubrir el significado del discurso. Estos niveles son:

- **Nivel fonético:** Aquí se presta atención a la fonética, la forma en que las palabras son pronunciadas. Este nivel es importante cuando procesamos la palabra hablada, no así cuando trabajamos con texto escrito.
- **Nivel morfológico:** Aquí nos interesa realizar un análisis morfológico del discurso; estudiar la estructura de las palabras para delimitarlas y clasificarlas.
- **Nivel sintáctico:** Aquí se realiza un análisis de sintaxis, el cual incluye la acción de dividir una oración en cada uno de sus componentes.
- **Nivel semántico:** Este nivel es un complemento del anterior, en el análisis semántico se busca entender el significado de la oración. Las palabras pueden tener múltiples significados, la idea es identificar el significado apropiado por medio del contexto de la oración.
- **Nivel discursivo:** El nivel discursivo examina el significado de la oración en relación a otra oración en el texto o párrafo del mismo documento.
- **Nivel pragmático:** Este nivel se ocupa del análisis de oraciones y cómo se usan en diferentes situaciones. Además, también cómo su significado cambia dependiendo de la situación.

Todos los niveles descritos aquí son inseparables y se complementan entre sí. El objetivo de los sistemas de NLP es incluir estas definiciones en una *computadora* y luego usarlas para crear una oración estructurada y sin ambigüedades con un significado bien definido.

## El gran desafío

NLP es considerado uno de los grandes retos de la inteligencia artificial ya que es una de las tareas más complicadas y desafiantes: ¿cómo comprender **realmente** el significado de un texto? ¿cómo intuir neologismos, ironías, chistes ó poesía? Si la estrategia/algoritmo que utilizamos no sortea esas dificultades de nada nos servirán los resultados obtenidos.



Modelos, maquetas y el mundo

En NLP no es suficiente con comprender meras **palabras**, se deberá comprender al **conjunto de palabras** que conforman *una oración*, y al **conjunto de líneas** que comprenden *un párrafo*. Dando un sentido global al análisis del texto/discurso para poder sacar buenas conclusiones.

Nuestro lenguaje está lleno de ambigüedades, de palabras con distintas acepciones, giros y diversos significados según el contexto. Esto hace que el NLP sea una de las tareas más difíciles de dominar.

## Aplicaciones del Procesamiento del Lenguaje Natural

Los algoritmos de Procesamiento del Lenguaje Natural suelen basarse en algoritmos de aprendizaje automático. En lugar de codificar manualmente grandes conjuntos de reglas, el NLP puede confiar en el aprendizaje automático para aprender estas reglas automáticamente analizando un conjunto de ejemplos y haciendo una inferencia estadística. En general, cuanto más datos analizados, más preciso será el modelo. Estos algoritmos pueden ser utilizados en algunas de las siguientes aplicaciones:

- **Resumir texto:** Podemos utilizar los modelos de NLP para extraer las ideas más importantes y centrales mientras ignoramos la información irrelevante.
- **Crear chatbots:** Podemos utilizar las técnicas de NLP para crear chatbots que puedan interactuar con las personas.
- **Generar automáticamente etiquetas de palabras clave:** Con NLP también podemos realizar un análisis de contenido aprovechando el algoritmo de LDA para asignar palabras claves a párrafos del texto.
- **Reconocer entidades:** Con NLP podemos identificar a las distintas entidades del texto como ser una persona, lugar u organización.
- **Análisis de sentimiento:** También podemos utilizar NLP para identificar el *sentimiento* de una cadena de texto, desde muy negativo a neutral y a muy positivo.

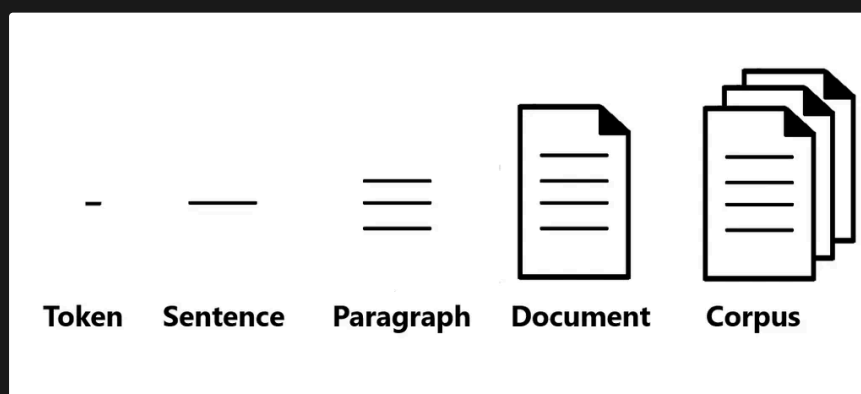


## Corpus lingüístico

Hoy en día, es indispensable el uso de buenos recursos lingüísticos para el desarrollo de los sistemas de NLP. Estos recursos son esenciales para la creación de gramáticas, en el marco de aproximaciones simbólicas; o para llevar a cabo la formación de módulos basados en el aprendizaje automático.

Un corpus lingüístico es un conjunto amplio y estructurado de ejemplos reales de uso de la lengua. Estos ejemplos pueden ser textos (los más comunes), o muestras orales (generalmente transcritas). Un corpus lingüístico es un conjunto de textos relativamente grande, creado independientemente de sus posibles formas o usos. Es decir, en cuanto a su estructura, variedad y complejidad, un corpus debe reflejar una lengua, o su modalidad, de la forma más exacta posible; en cuanto a su uso, preocuparse de que su representación sea real. La idea es que representen al lenguaje de la mejor forma posible para que los modelos de NLP puedan aprender los patrones necesarios para entender el lenguaje. Encontrar un buen corpus sobre el cual trabajar no suele ser una tarea sencilla; uno que se suele utilizar para entrenar modelos es la información de wikipedia.

En el ámbito del NLP, encontraremos habitualmente estas clasificaciones de textos, según su jerarquía:



El gráfico anterior presenta una jerarquía de unidades textuales utilizadas comúnmente en el procesamiento del lenguaje natural (NLP), ordenadas de menor a mayor complejidad:

1. Token: La unidad más pequeña, generalmente una palabra o signo de puntuación.
2. Sentence (Oración): Compuesta por varios tokens, forma una idea completa.
3. Paragraph (Párrafo): Agrupa varias oraciones relacionadas.
4. Document (Documento): Contiene múltiples párrafos, formando un texto completo.
5. Corpus: La unidad más grande, una colección de documentos.

Esta estructura jerárquica es fundamental en NLP, ya que permite analizar el lenguaje en diferentes niveles de granularidad. Los sistemas de NLP pueden procesar desde unidades básicas (tokens) hasta grandes volúmenes de texto (corpus), dependiendo de la tarea específica, como análisis sintáctico, semántico, o extracción de información a gran escala.

## Librerías de Python para Procesamiento del Lenguaje Natural

Actualmente, Python es uno de los lenguajes más populares para trabajar en el campo la inteligencia artificial. Para abordar los problemas relacionados con el Procesamiento del Lenguaje Natural, el ecosistema de Python nos proporciona diferentes librerías para trabajar. Algunas de ellas son:

- NLTK: Es la librería líder para el Procesamiento del Lenguaje Natural. Proporciona interfaces fáciles de usar a más de 50 corpus y recursos léxicos, junto con un conjunto de bibliotecas de procesamiento de texto para la clasificación, tokenización, el etiquetado, el análisis y el razonamiento semántico.
- Spacy: Es una librería relativamente nueva que sobresale por su facilidad de uso y su velocidad a la hora de realizar el procesamiento de texto.
- Gensim: Es una librería diseñada para extraer automáticamente los temas semánticos de los documentos de la forma más eficiente y con menos complicaciones posible.
- Stanza: Es una librería de Python para el procesamiento del lenguaje natural desarrollada por el Grupo de Procesamiento del Lenguaje Natural de Stanford. Ofrece herramientas precisas para el análisis lingüístico, incluyendo tokenización, lematización, etiquetado morfológico, reconocimiento de entidades nombradas y análisis de dependencias para múltiples idiomas.
- txtai: Es una biblioteca de IA centrada en el texto que proporciona bloques de construcción para crear flujos de trabajo de procesamiento de lenguaje natural y búsqueda semántica. Ofrece funcionalidades como indexación de vectores, incrustaciones de texto, y extracción de información, lo que permite construir aplicaciones avanzadas de PLN de manera eficiente.

- Transformers: Desarrollada por Hugging Face, esta librería proporciona miles de modelos preentrenados para realizar tareas de PLN. Es especialmente conocida por su implementación de arquitecturas de transformers como BERT, GPT, T5 y más. Ofrece una API fácil de usar para tareas como clasificación de texto, generación de lenguaje, respuesta a preguntas y traducción automática.

Actualmente el NLP se está sumando a la popularidad del Deep Learning, por lo que muchos de los frameworks que se utilizan en Deep Learning pueden ser aplicados para realizar modelos de NLP.

## Deep Learning y Procesamiento del Lenguaje Natural

Durante mucho tiempo, las técnicas principales de Procesamiento del Lenguaje Natural fueron dominadas por métodos de aprendizaje automático que utilizaron modelos lineales como las máquinas de vectores de soporte o la regresión logística, entrenados sobre vectores de características de muy alta dimensional pero muy escasos. Recientemente, el campo ha tenido cierto éxito en el cambio hacia modelos de deep learning sobre entradas densas.

Las redes neuronales proporcionan una poderosa máquina de aprendizaje que es muy atractiva para su uso en problemas de *lenguaje natural*. Un componente importante en las redes neuronales para el lenguaje es el uso de una capa de word embedding, una asignación de símbolos discretos a vectores continuos en un espacio dimensional relativamente bajo. Cuando se utiliza word embedding, se transforman los distintos símbolos en objetos matemáticos sobre los que se pueden realizar operaciones. En particular, la distancia entre vectores puede equipararse a la distancia entre palabras, facilitando la generalización del comportamiento de una palabra sobre otra. Esta representación de palabras como vectores es aprendida por la red como parte del proceso de entrenamiento. Subiendo en la jerarquía, la red también aprende a combinar los vectores de palabras de una manera que es útil para la predicción. Esta capacidad alivia en cierta medida los problemas de dispersión de los datos. Hay dos tipos principales de arquitecturas de redes neuronales que resultan muy útiles en los problemas de Procesamiento del Lenguaje