Inteligencia Artificial II

Juan Pablo Restrepo Uribe

Ing. Biomedico - MSc. Automatización y Control Industrial

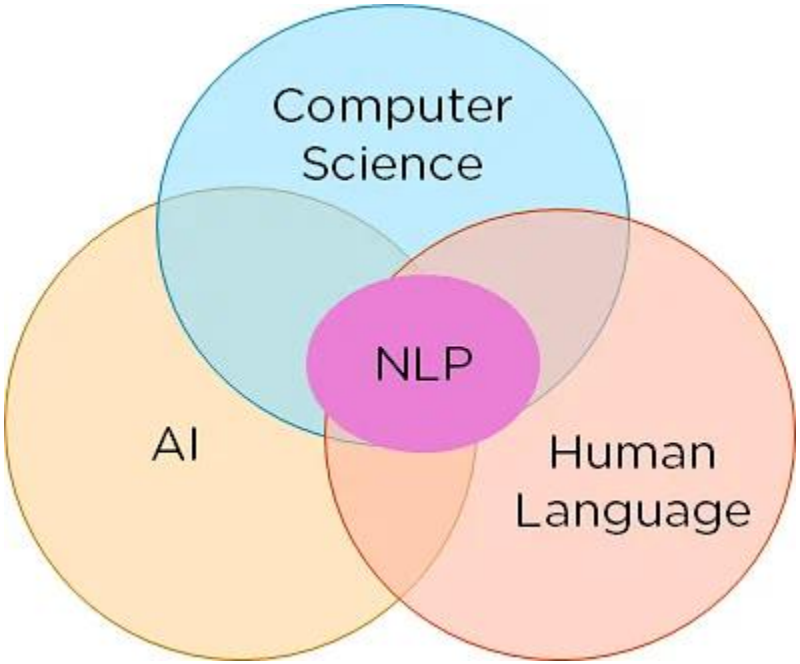
jprestrepo@correo.iue.edu.co

2023

Institución Universitaria de Envigado

NLP

El procesamiento de lenguaje natural (NLP) es una tecnología de machine learning que brinda a las computadoras la capacidad de interpretar, manipular y comprender el lenguaje humano



¿Por qué es importante la NLP?

El procesamiento de lenguaje natural es fundamental para analizar a profundidad los datos de texto y voz de manera eficiente. Puede resolver las diferencias en dialectos, jerga e irregularidades gramaticales típicas en las conversaciones cotidianas. Las empresas lo utilizan para varias tareas automatizadas, como:

- Procesar, analizar y archivar documentos grandes
- Analizar los comentarios de los clientes o las grabaciones de centros de atención telefónica
- Ejecutar chatbots para ofrecer un servicio al cliente automatizado
- Responder preguntas de quién, qué, cuándo y dónde
- Clasificar y extraer texto

¿Cómo funciona el NLP?

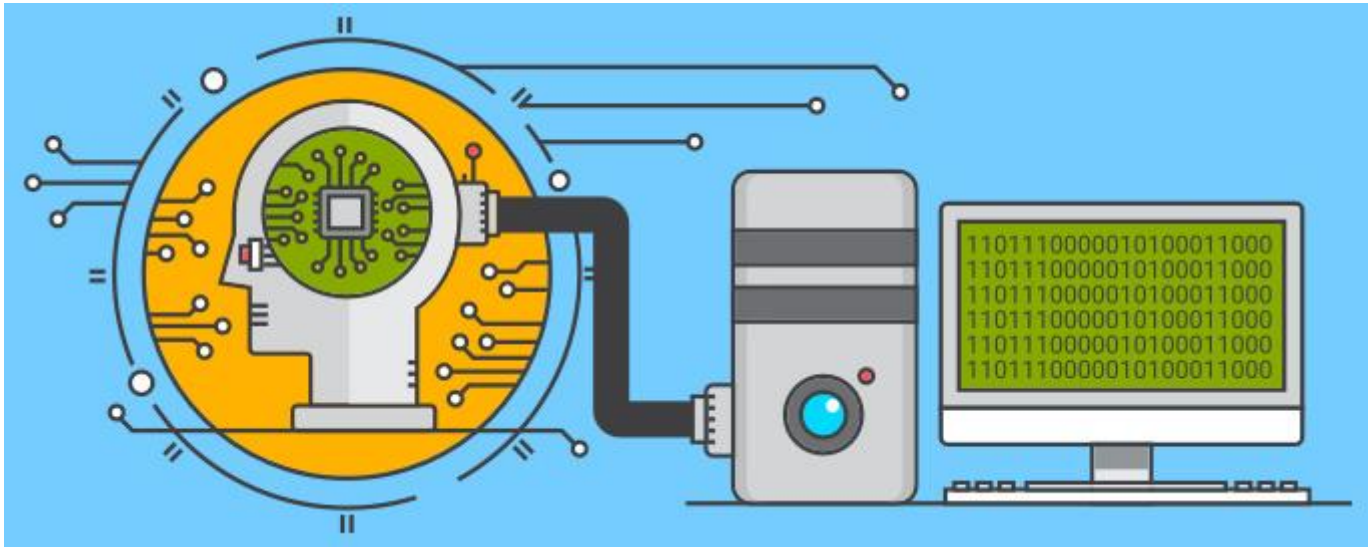
Lingüística computacional

La lingüística computacional es la ciencia de entender y crear modelos de lenguaje humano con computadoras y herramientas de software. Los investigadores utilizan métodos lingüísticos computacionales, como el análisis sintáctico y semántico, para crear marcos que ayuden a las máquinas a entender el lenguaje humano conversacional. Las herramientas como los traductores de idiomas, los sintetizadores de texto a voz y el software de reconocimiento de voz se basan en la lingüística computacional.

¿Cómo funciona el NLP?

Machine learning

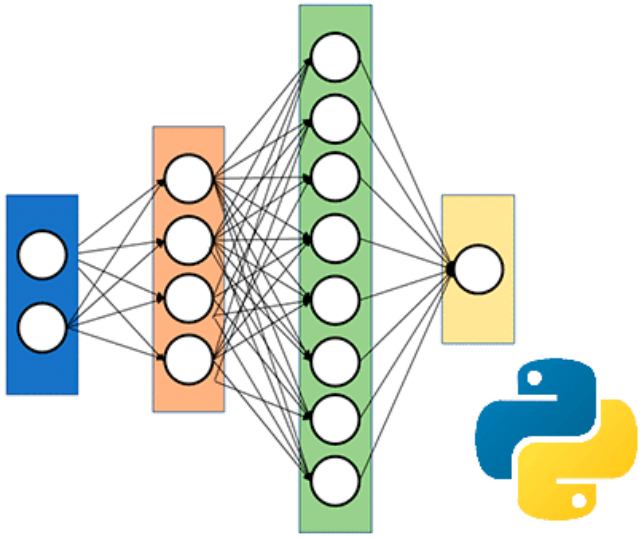
El machine learning es una tecnología que entrena a una computadora con datos de muestra para mejorar su eficiencia. El lenguaje humano tiene varias características, como sarcasmo, metáforas, variaciones en la estructura de las oraciones, además de excepciones gramaticales y de uso que los humanos tardan años en aprender. Los programadores utilizan métodos de machine learning para enseñar a las aplicaciones de NLP a reconocer y comprender con precisión estas características desde el principio.



¿Cómo funciona el NLP?

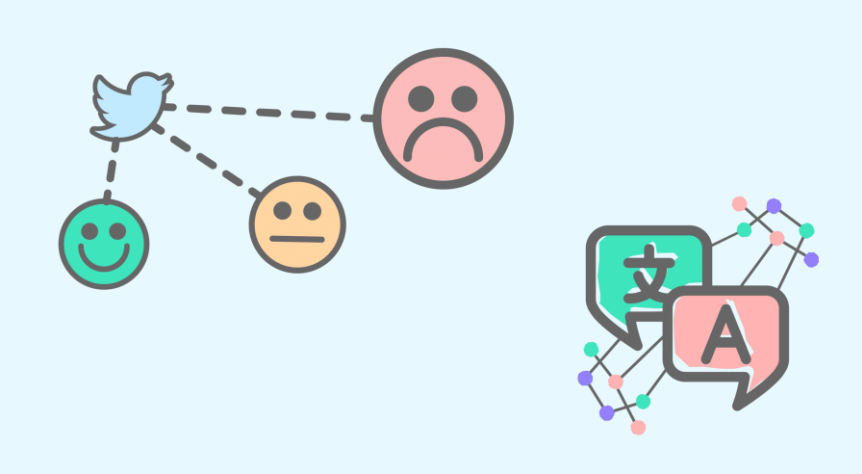
Aprendizaje profundo

El aprendizaje profundo es un campo específico del machine learning que enseña a las computadoras a aprender y pensar como humanos. Se trata de una red neuronal que consta de nodos de procesamiento de datos que se asemejan a las operaciones del cerebro humano. Con el aprendizaje profundo, las computadoras reconocen, clasifican y correlacionan patrones complejos en los datos de entrada.



Pasos de la implementación del NLP

- ***Preprocesamiento***
  - La creación de tokens divide una oración en unidades individuales de palabras o frases.
  - La derivación y la lematización simplifican las palabras en su forma raíz. Por ejemplo, estos procesos convierten *iniciar* en *inicio*.
  - La eliminación de palabras de parada garantiza que se eliminen las palabras que no añaden un significado relevante a una oración, como*por* y *con*.



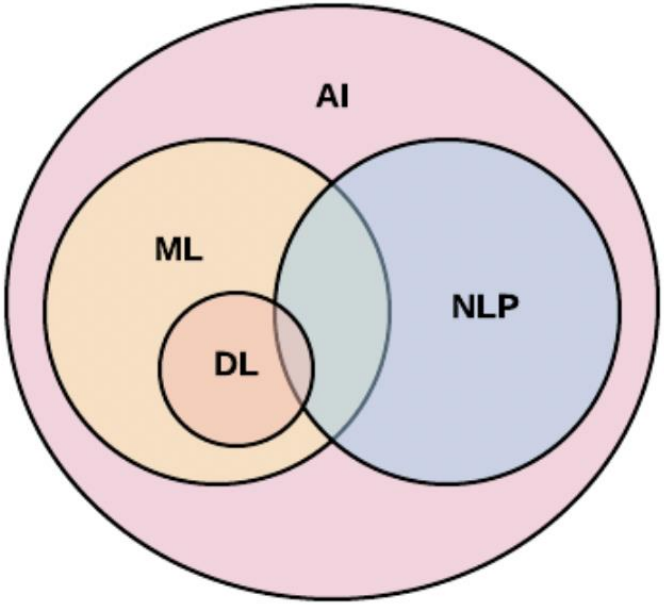
Pasos de la implementación del NLP

- ***Capacitación***

Los investigadores utilizan los datos preprocesados para entrenar modelos de NLP con machine learning para realizar aplicaciones específicas basadas en la información textual proporcionada. El entrenamiento de los algoritmos de NLP requiere suministrar al software con grandes muestras de datos para aumentar su precisión.

Enfoques para el NLP

NLP supervisado
NLP no supervisado
Comprensión de lenguaje natural
Generación de lenguaje natural



IA : Inteligencia Artificial

ML : Aprendizaje Automático

DL : Aprendizaje Profundo

NLP : Procesamiento del Lenguaje Natural

Elementos comunes del NLP

El procesamiento del lenguaje natural debe realizarse de forma sistemática, dividiéndolo en partes, agregando elementos gramaticales e identificando elementos interesantes:

- Tokenización: consiste en dividir el texto en oraciones y estas, en palabras.
- Normalización: estandariza todas las palabras.
- Eliminación de palabras redundantes: consiste en omitir o eliminar palabras redundantes.
- Etiquetado POS: consiste en asignar etiquetas para sustantivos, pronombres, verbos, adjetivos, adverbios, etc.
- Bolsa de palabras: una oración se considera como un conjunto de palabras
- N-gramas: son una secuencia continua de palabras adyacentes en una oración, necesarias para obtener el significado correctamente.
- TF (frecuencia del término): es el número de veces que aparece una palabra en un mensaje o una oración.
- Reconocimiento de entidades nombradas: identifica y etiqueta palabras que representan entidades de palabras reales.

Tareas de NLP

- Parte del etiquetado de voz: Este es un proceso en el que el software de NLP etiqueta palabras individuales en una oración de acuerdo con los usos contextuales.
- Desambiguación del sentido de las palabras: Algunas palabras pueden tener diferentes significados cuando se usan en diferentes escenarios.
- Reconocimiento de voz: El reconocimiento de voz convierte los datos de voz en texto. El proceso implica dividir las palabras en partes más pequeñas y superar desafíos como acentos, insultos, entonación y uso incorrecto de la gramática en la conversación cotidiana.
- Traducción automática El software de traducción automática utiliza el procesamiento de lenguaje natural para convertir texto o voz de un idioma a otro, manteniendo la precisión contextual.
- Reconocimiento de entidades nombradas: Este proceso identifica nombres únicos de personas, lugares, eventos, empresas y más.

Librerias

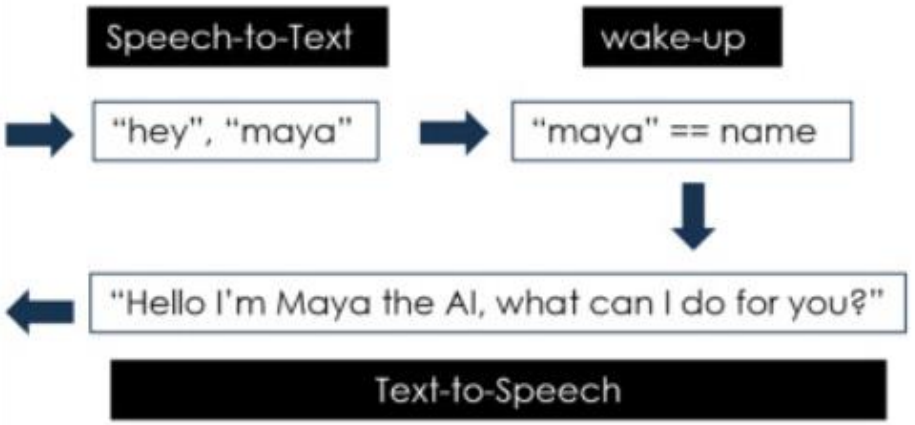
- **NLTK** (https://www.nltk.org/): librería Python para trabajar con lenguaje natural que proporciona interfaces fáciles de usar junto con cincuenta corpus y recursos léxicos como WordNet.
- **Polyglot** (https://github.com/aboSamoor/polyglot): juego de herramientas de lenguaje natural que soporta aplicaciones multilingües masivas.
- **TextBlob** (https://textblob.readthedocs.io/en/dev/): proporciona una API sencilla para sumergirse en tareas comunes de procesamiento del lenguaje natural (NLP), como el etiquetado de parte de la voz, la extracción de frases de nombres, el análisis de sentimientos.
- **spaCy** (https://spacy.io/): biblioteca de procesamiento de lenguaje natural diseñada específicamente con el objetivo de ser útil para implementar sistemas listos para la producción.
- **OpenNLP** (https://opennlp.apache.org/): kit de herramientas basado en el aprendizaje automático para el procesamiento de texto en lenguaje natural.

Modelos

Métrica de polaridad y subjetividad. La polaridad es el sentimiento mismo, que va de -1 a +1. La subjetividad es una medida del sentimiento siendo objetivo a subjetivo, y va de 0 a 1.

Modelos


Hello I'm Maya the AI, what can I do for you?"


Speech-to-Text
wake-up
"hey", "maya"
"maya" == name
"Hello I'm Maya the AI, what can I do for you?"
Text-to-Speech

Corpus lingüístico

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.