INTELIGENCIA ARTIFICIAL CON DEEP LEARNING

ING. JORGE ALBERTO CASTELLANOS

Universidad de La Sabana

FACULTAD DE INGENIERÍA

Universidad de La Sabana
FACULTAD DE INGENIERIA

Preprocesamiento de datos en Lenguaje natural NLP

- **Qué es el leguaje:**

Un Lenguaje es un conjunto potencialmente infinito de oraciones y sentencias de palabras construidas mediante reglas gramaticales, fonéticas y de significación que rigen el propio lenguaje.

Tipos de Lenguaje

01 Nace de manera espontanea por la necesidad de comunicarse. (Idiomas, ingles, español, chino)

Lenguaje Natural

Lenguaje Formal

Lenguajes creados para una situación particular. (Matemático, lógico, musical, programación) 02

03 Lenguajes creados antes de ser usado por los parlantes, como una mezcla de natural y formal.

Lenguaje Artificial

Universidad de La Sabana
FACULTAD DE INGENIERÍA

¿Qué es el Preprocesamiento en NLP?

01

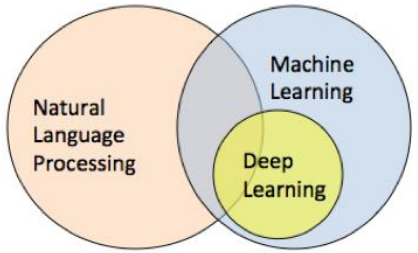
Definición

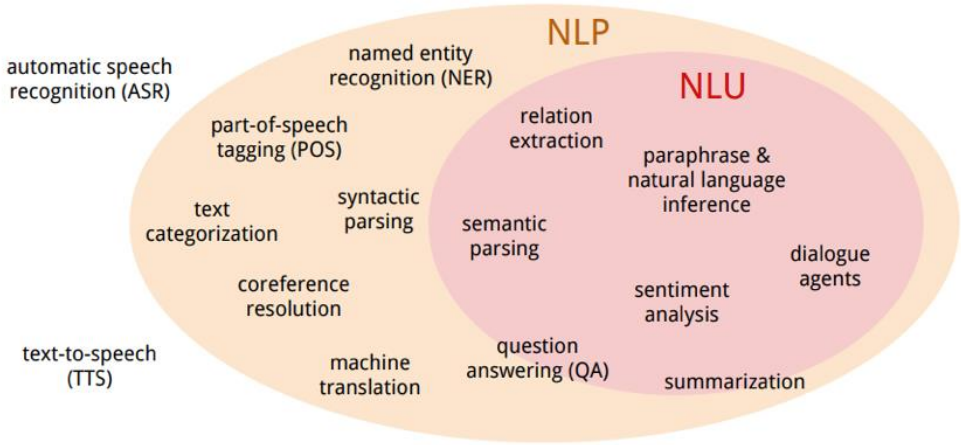
Transformación de texto en un formato adecuado para su análisis.

02

Objetivo

Tratar la interacción entre los lenguajes humanos (lenguajes naturales) y los dispositivos informáticos.





Campo que combina la ***Informática***, la ***Inteligencia Artificial*** y la ***Lingüística***;

Donde se puede encontrar

- Recuperación de información
- Extracción y categorización de información
- Análisis automático de texto subjetivo (Análisis de sentimientos)
- Traducción automática
- Generación del lenguaje
- Questions & Answering (Chatbots)

Conceptos para el NLP

- Corpus: Colección de textos como puede ser un conjunto de artículos científicos, libros, tweets, criticas.

- Bag of Words (BoW): modelo para simplificar el contenido de un documento(s) – Sin gramática, ni orden de palabras. +ocurrencias de palabras

- Normalización: Poner el texto en igual de condiciones:
  - Convertir Mayúsculas o minúsculas
  - Eliminar puntuación
  - Convertir números a palabras
  - Eliminar palabras que no aporten al texto (Stop Word)

Conceptos para el NLP

- Tokenización: dividir el texto en unidades mas pequeñas llamadas **tokens** que pueden ser palabras, frases, símbolos u otros elementos significativos.
  - Segmentación: dividir en oraciones o párrafos
  - Tokenización: dividir grandes cadenas de texto en palabras.
- Stemming: Proceso de eliminar los afijos (sufijos, prefijos, infijos, circunflejos) de una palabra para obtener un tallo de palabra.
  - Caminando → Caminar
- Lematización: Proceso lingüístico, sustituys palabra flexionada (Plurales, verbos conjugados o femeninos) por su lema; como una palabra valida en el idioma.

Conceptos para el NLP

- Stop Word: Son palabras que no aportan nada al significado de las frases como las preposiciones, determinantes, etc.
- Part of speech (POS) Tagging: Asignar una etiqueta de categoría a las partes tokenizadas de una oración. El etiquetado POS más popular sería identificar palabras como sustantivos, verbos, adjetivos, etc.

  - Artículo o determinante
  - Sustantivo o nombre
  - Pronombre
  - Verbo
  - Adjetivo
  - Adverbio
  - Preposición
  - Conjunción
  - Interjección

Conceptos para el NLP

- n-grammas: A diferencia de la representación sin orden de una bolsa de palabras (bag of words), el modelado de n-gramas está interesado en preservar secuencias contiguas de N elementos de la selección de texto.

Actividad

- https://colab.research.google.com/drive/1ayXlZpI0-SIHJRsLA4Z2XurPGIsVKhpA#scrollTo=MtnaMFuXHuSI
- https://colab.research.google.com/drive/1pZ0BsKXKekqlY-GVd1vEIK9BuxRU5Osn#scrollTo=WXv3dTcIFIFl