DIPLOMADO EN APRENDIZAJE MAQUINA Y PROFUNDO APLICADO A GRANDES VOLUMENES DE DATOS
(Machine Learning, Deep Learning y Big Data)

PROYECTO PAPIME PE103124 2024-2025

Módulo II. Deep Learning, del 29 de julio al 30 de agosto 2024

Procesamiento del lenguaje natural (NLP) - Antecedentes

Dr. José Luis Cendejas Valdez
Profesor – investigador

Procesamiento del lenguaje natural (NLP)

Según aws (2024):

- Hoy en día, las organizaciones tienen grandes volúmenes de datos de voz y texto de varios canales de comunicación, como correos electrónicos, mensajes de texto, fuentes de noticias en redes sociales, vídeo, audio y más.

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

- Utilizan software de NLP para procesar de forma automática estos datos, analizan la intención o el sentimiento del mensaje y responden en tiempo real a la comunicación humana.

- También puede integrar el NLP en aplicaciones orientadas al cliente para comunicarse de manera más eficaz con ellos. Por ejemplo, un chatbot analiza y ordena las consultas de los clientes, responde automáticamente a las preguntas comunes y redirige las consultas complejas al servicio de atención al cliente. Esta automatización ayuda a reducir los costos, evita que los agentes dediquen tiempo a las consultas redundantes y mejora la satisfacción del cliente.

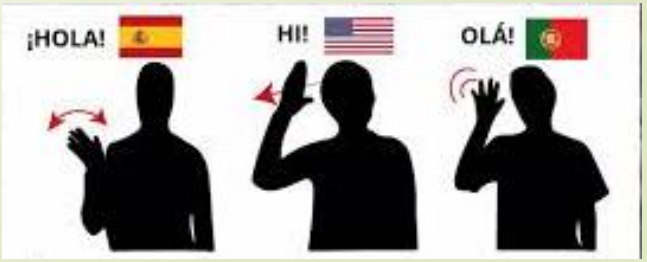
Procesamiento del lenguaje natural - PLN

- El PLN combina la **lingüística computacional** (modelado del lenguaje humano basado en reglas) con modelos estadísticos, de **machine learning** y de **deep learning**.

- Juntas, estas tecnologías permiten que las computadoras procesen el lenguaje humano en forma de texto o datos de voz y "comprendan" su significado completo, con la intención y el sentimiento de la persona que habla o escribe.

Lingüística computacional - Alcance

- Estudio del lenguaje desde una perspectiva computacional
- Desarrollar tecnología que trabaje con lenguaje natural
- Atender la complejidad del sistema linguistico



- Analiza colecciones de documentos escritos para decubrir patrones (Minería de textos)
- Desarrollo de software que analice texto

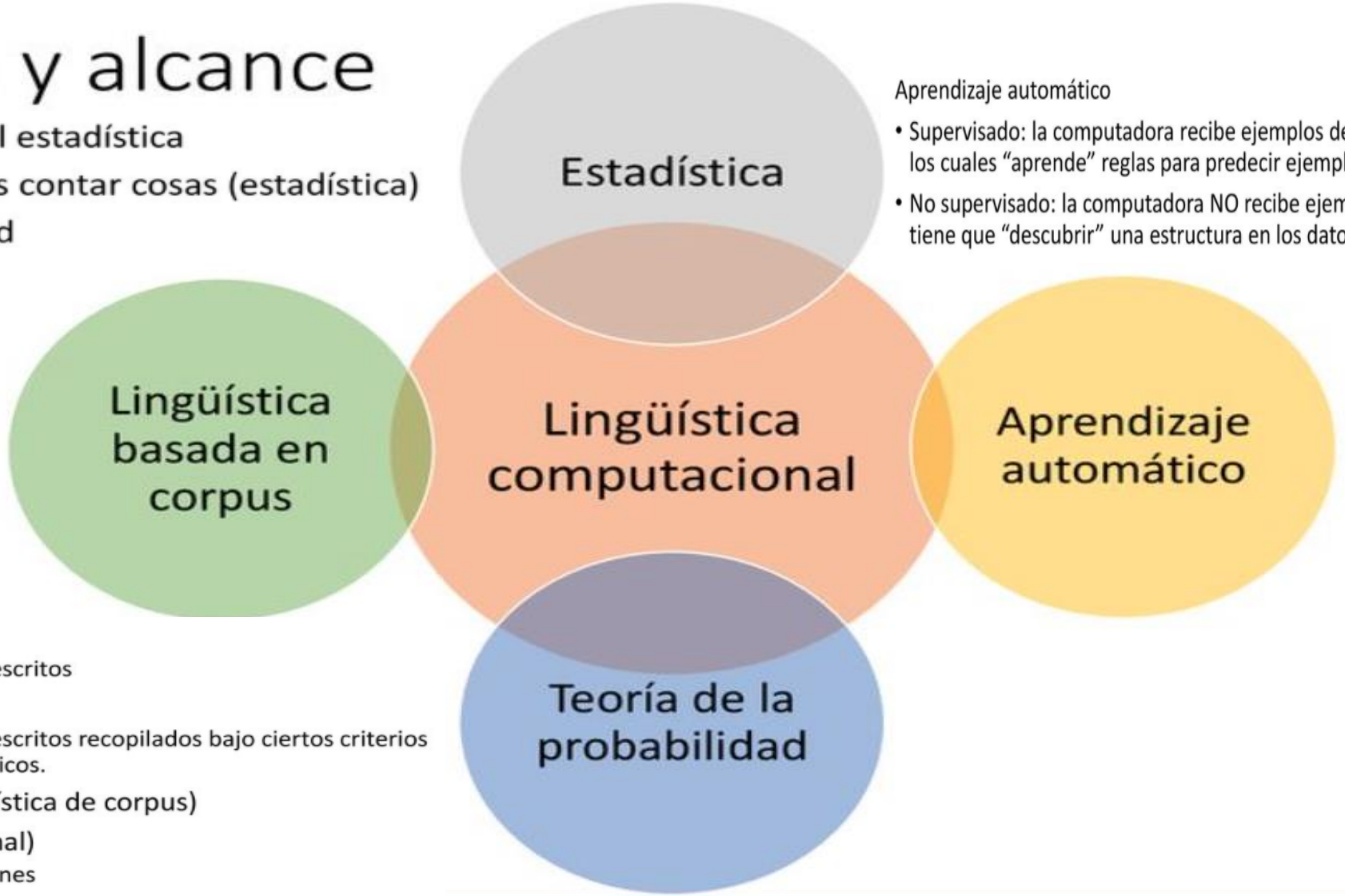
Procesamiento del lenguaje natural - Semiótica

Lingüística computacional - Alcance

Definición y alcance

Lingüística computacional estadística

- Principal herramienta es contar cosas (estadística)
- Teoría de la probabilidad



Aprendizaje automático

- Supervisado: la computadora recibe ejemplos de entrada a partir de los cuales "aprende" reglas para predecir ejemplos de salida.
- No supervisado: la computadora NO recibe ejemplos de entrada y tiene que "descubrir" una estructura en los datos.

- Corpus
  - Colección de materiales hablados o escritos
- Corpus lingüísticos
  - Colección de materiales hablados o escritos recopilados bajo ciertos criterios de investigación para análisis lingüísticos.
- Estudios basados en corpus (Lingüística de corpus)
- Corpus (en lingüística computacional)
  - Colección de documentos o grabaciones

El enfoque de la bolsa de palabras (bag of words) - corpus

Queremos analizar y comparar las primeras frases de dos de las fábulas de Esopo, 'La liebre y la tortuga' y 'La tortuga y los patos', las cuales son:

- Una Liebre se estaba burlando de la Tortuga un día por ser tan lenta.
- La Tortuga, ya sabes, lleva su casa a la espalda.

Vocabulario/Corpus
- Liebre
- Tortuga
- Casa
- Haciendo
- Siendo
- Lleva

Fábula	Liebre	Tortuga	Casa	Haciendo	Siendo	Lleva
La liebre y la tortuga	1	1	0	1	1	0
La tortuga y los patos	0	1	1	0	0	1

Vectores

hareAndTortoise = [1, 1, 0, 1, 1, 0]
tortoiseAndDucks = [0, 1, 1, 0, 0, 1]

Lenguaje natural vs Lenguaje Artificial

Lenguaje Natural

- Medio principal para la comunicación humana.
- Gran poder expresivo.
- Para conseguir un alto grado de comprensión del lenguaje natural es necesario que los algoritmos posean un completo conocimiento del idioma; desde los caracteres de una palabra hasta el contexto del diálogo.
- Propiedades:
  - Han sido desarrollados por enriquecimiento progresivo previo a cualquier teoría.
  - La importancia de su carácter expresivo se debe a la riqueza del componente semántico (polisemia).
  - Existe dificultad o imposibilidad de una formalización completa.

Lenguaje Artificial

- Se compone de símbolos y fórmulas, con el objetivo de formalizar la programación de ordenadores o representar simbólicamente el conocimiento científico.
- Las palabras y oraciones están perfectamente definidas, una palabra mantiene el mismo significado prescindiendo del contexto o su uso.
- Propiedades:
  - Se desarrollan a partir de una teoría preestablecida.
  - Componente semántico mínimo.
  - Posibilidad de incrementar el componente semántico de acuerdo con la teoría a formalizar.
  - La sintaxis produce oraciones no ambiguas.
  - Los números tienen un rol importante.
  - Poseen una completa formalización, posibilitando la construcción computacional.

Niveles del Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN) tiene por objetivo habilitar a las computadoras para que entiendan el texto, procesándolo por su sentido. Para llevar a cabo esta tarea, un sistema de procesamiento de lenguaje natural necesita conocer la estructura del lenguaje, la cual se analiza normalmente en 4 niveles:

1. Nivel Morfológico: Se estudia cómo se construyen las palabras. Detecta las relaciones que se establece entre las unidades mínimas que forman una palabra (sufijos, prefijos) y la relación con el léxico, siendo éste un conjunto de información sobre cada palabra que el sistema utiliza para el procesamiento.

A continuación se muestra un ejemplo de análisis morfológico para la oración "Un gato negro caza un ratón blanco":

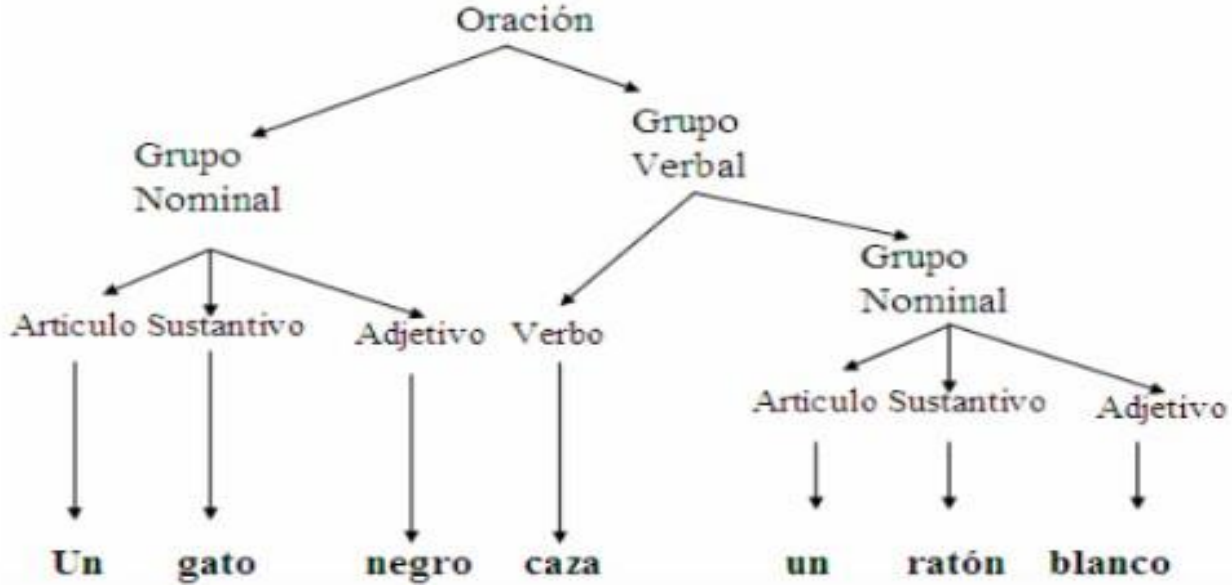
un	Artículo, singular, masculino
gato	Sustantivo, común, masculino, singular
negro	Adjetivo, singular, masculino
caza	Verbo *cazar*. Principal, indicativo, presente, tercera persona, singular
un	Artículo, singular, masculino
ratón	Sustantivo, común, masculino, singular
blanco	Adjetivo, singular, masculino

Niveles del Procesamiento de Lenguaje Natural

2. Nivel Sintáctico: Se estudia cómo combinar las palabras para formar oraciones. Su función es asignar etiquetas a cada uno de los componentes que aparecen en la oración a analizar, de manera que se sepa cómo se combinan las palabras y forman estructuras gramaticales correctas.

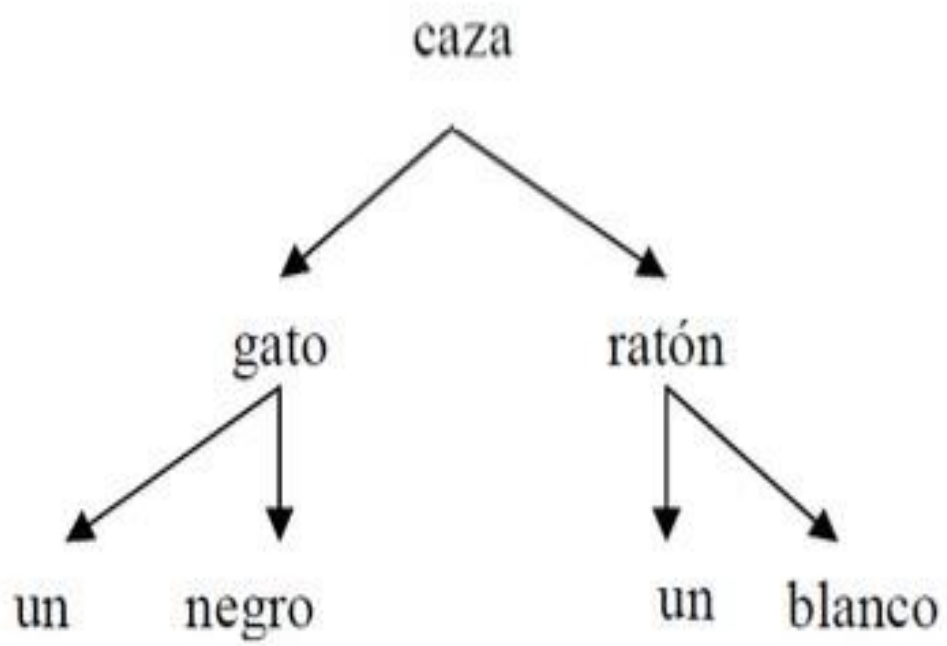
Existen dos enfoques para describir formalmente la gramaticalidad de las oraciones: los constituyentes y las dependencias:

a) El **enfoque de constituyentes** consiste en analizar la oración mediante un proceso de segmentación y clasificación, obteniendo como resultado un árbol como el siguiente:



Niveles del Procesamiento de Lenguaje Natural

b) El **enfoque de dependencias** consiste en el establecimiento de la relación entre pares de palabras, una de ellas tiene el rol de rectora y la otra el rol de dependiente o subordinada, obteniendo como resultado una estructura jerárquica, el árbol de dependencias, donde la única palabra que no tiene rectora es la raíz del árbol, tal y como se muestra a continuación:

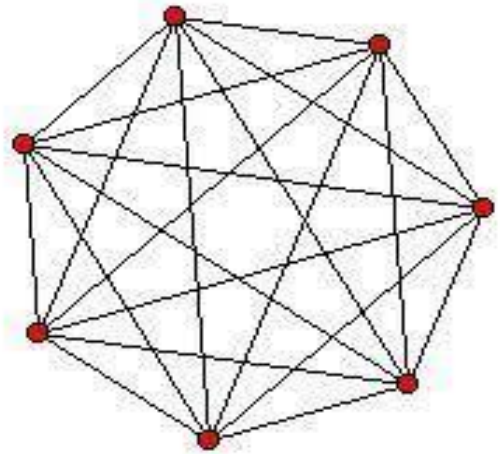


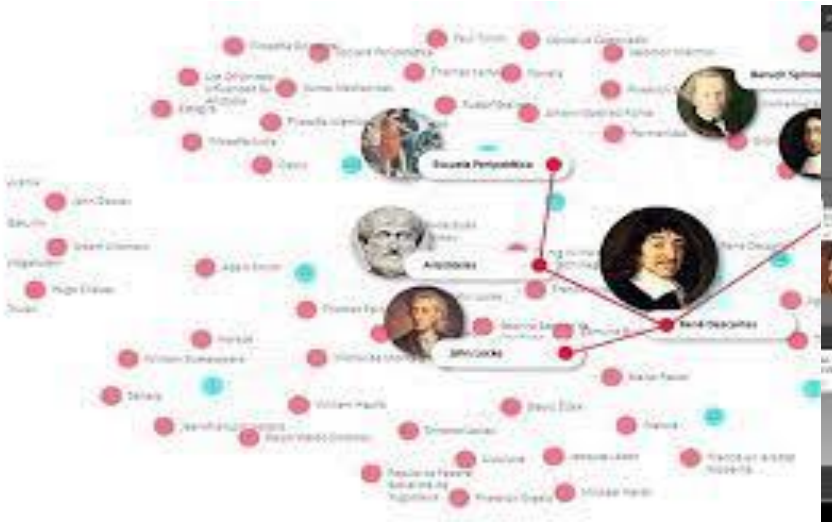
Niveles del Procesamiento de Lenguaje Natural

3. Nivel Semántico: Se estudia el significado de las oraciones, representándolo de manera formal.

Existen varias formas de representación formal semántica de las oraciones, tales como las siguientes:

- Lógica de primer orden
- Redes semánticas
- Grafos conceptuales



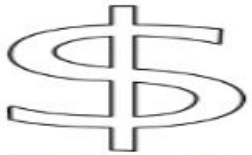


Niveles del Procesamiento de Lenguaje Natural

4. Nivel Pragmático: Se estudia cómo el contexto afecta a la interpretación de las oraciones.

Ésta es la cadena de análisis ideal con los cuatro niveles, que sin embargo no se corresponde con la realidad, dado que la mayoría de los sistemas no van más allá del análisis sintáctico.

Letra S



Símbolo de pesos

Símbolo de dolares

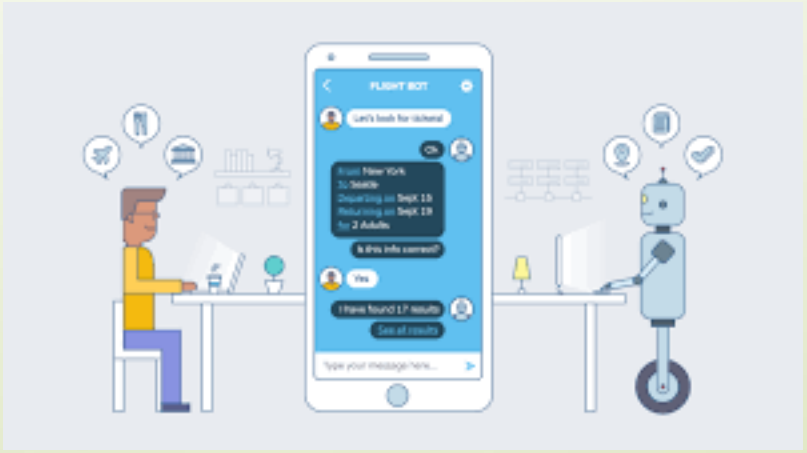
Símbolo de superheroe

Para qué sirve el procesamiento de lenguaje natural

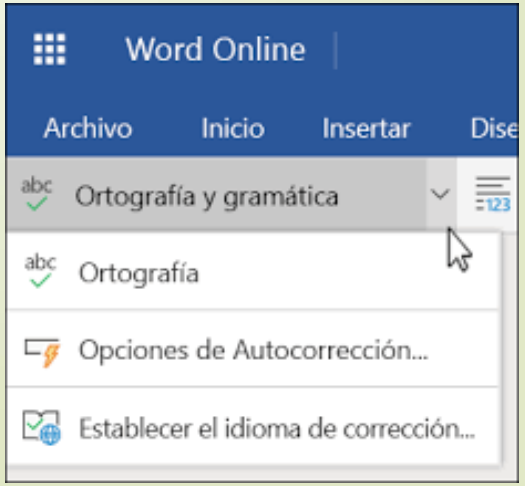
- Reconocimiento de patrones de lenguaje: procesar grandes cantidades de documentos, el reconocimiento de patrones permite filtrar datos importantes en cadenas de texto con un tiempo récord. Es el primer paso para que la recuperación de información y la clasificación de textos sea posible.
- Recuperación de información: los sistemas de PLN reconocen patrones de lenguaje, también hacen más sencilla la tarea de encontrar un fragmento en particular dentro de una gran cantidad de texto.
- Traducciones automáticas de idiomas, ya sea con voz o texto, estos sistemas son ideales para traducir discursos en cualquier lengua.
- Resumen de textos: al igual que con la clasificación, resumir un documento de gran extensión se apoya en ciertas palabras o frases clave.
- Detección de sentimientos y emociones: sirve para comprender ciertos mensajes con su intencionalidad, el procesamiento de lenguaje natural ya incursiona en el análisis de las emociones que se expresan a través de frases que aparecen en opiniones.

Aplicaciones del procesamiento del lenguaje natural

1.- Asistentes virtuales



2.- Chatbots

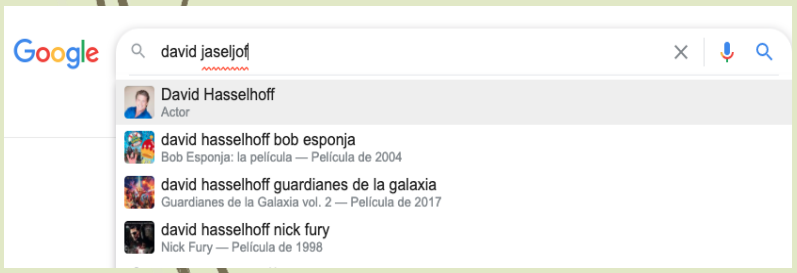


3.- Autocorrección

4.- Traductores

5.- Detección de spam en correo

6.- Dictado de voz



7.- Resultados de búsquedas

Casos de uso

1.- Análisis de sentimientos

2.- La minería de textos a traves de noticias – Caso de uso