SESIÓN I PROCESAMIENTO DE LENGUAJE NATURAL INTRODUCCIÓN

AGOSTO 08 DE 2023

FABIÁN SÁNCHEZ SALAZAR

INTRODUCCIÓN

INTRODUCCIÓN

PROCESAMIENTO DE LENGUAJE NATURAL

LENGUAJE

¿QUÉ ES UN LENGUAJE?

- Capacidad propia del ser humano para expresar pensamientos y sentimientos por medio de palabras
- Sistema de signos que utilizan las personas para comunicarse
- Sistema de comunicación estructurado para el que existe un contexto de uso y ciertos principios combinatorios formales.

INTRODUCCIÓN

¿QUÉ ES UN LENGUAJE?

Conjunto finito o infinito de oraciones, cada una de las cuales posee una extensión finita, construida a partir de un número finito de elementos.

(Chomsky 1957)

INTRODUCCIÓN

¿A QUÉ NOS REFERIMOS CUANDO DECIMOS LENGUAJE "NATURAL"?

- ✓ Adjetivo que refiere a la naturaleza del lenguaje
- ✓ Es la lengua o idioma hablado o escrito por humanos para propósitos generales de comunicación

INTRODUCCIÓN

LINGÜISTICA

Es el estudio del origen, la evolución y la estructura del lenguaje con el fin de reconocer reglas que rigen una determinada lengua.

LINGÜÍSTICA COMPUTACIONAL

- Campo interdisciplinar: lingüística y la computación.
- Busca desarrollar formalismos descriptivos del funcionamiento del lenguaje natural, que puedan ser transformados en programas ejecutables.

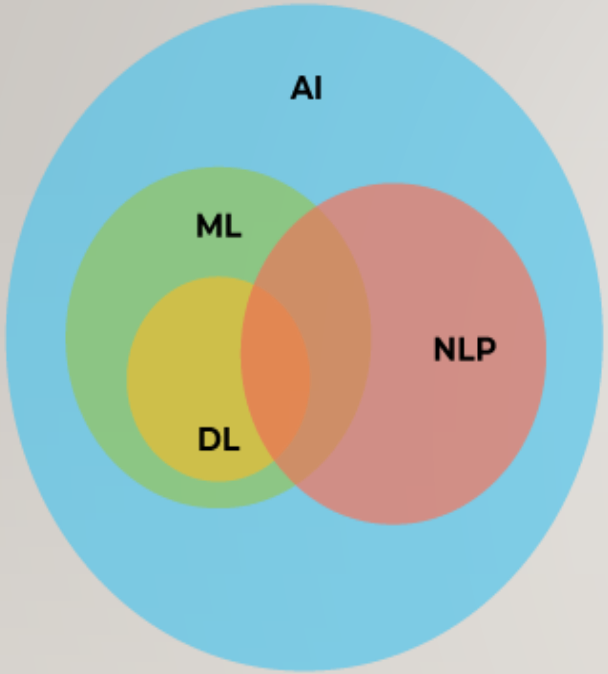
INTRODUCCIÓN

¿QUÉ ES EL PROCESAMIENTO DE LENGUAJE NATURAL?

El Procesamiento de Lenguaje Natural (PLN) es una subdisciplina de la Inteligencia Artificial que intenta resolver con computadoras tareas vinculadas al lenguaje humano, permitiendo la comunicación entre el humano y la computadora a través del lenguaje natural o resolviendo diferentes tareas que implican algún tipo de preprocesamiento de texto o habla.

(Jurasfsky and Martin, 2008)

INTRODUCCIÓN



- Artificial Intelligence
- Machine Learning
- Language Processing
- Deep Learning

Es una rama de la Inteligencia Artificial que junta las Ciencias de la Computación, las Matemáticas y **la Lingüística** para:

Construir sistemas (software/aplicaciones) que puedan **entender, analizar, manipular y generar lenguaje natural humano**, de forma parecida o mejor que como lo hacemos los humanos.

- Comprensión
- Generación

INTRODUCCIÓN

PROCESAMIENTO DE LENGUAJE NATURAL

Trata acerca de tareas que involucran el lenguaje:

- Traducción
- Resumen
- Reconocimiento de voz
- Generación de texto

LINGÜÍSTICA COMPUTACIONAL

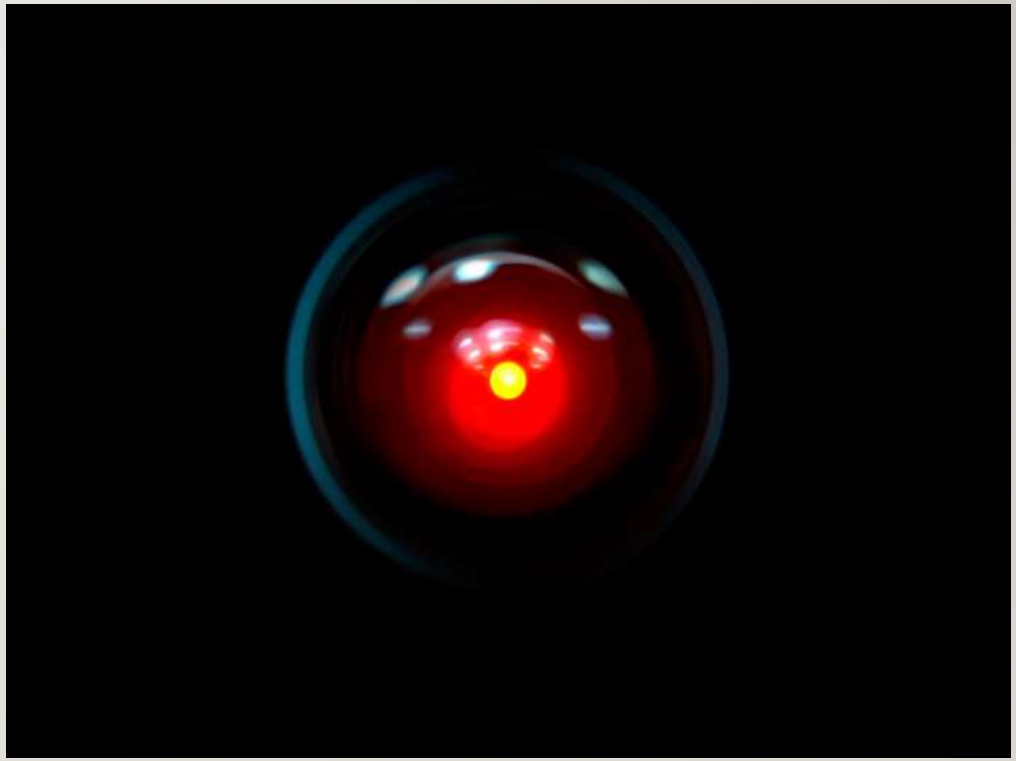
- Fundamentación teórica en los modelos y métodos computacionales propuestos
- Desarrollo de teorías lingüísticas computables.

INTRODUCCIÓN

HAL 9000

- Acrónimo: Heuristically Algorithmic Computer.
- Serie TV: Odisea Espacial (2001)
- Computadora a bordo de la nave espacial *Discovery*

https://www.youtube.com/watch?v=Hx6PC2id1U4



INTRODUCCIÓN

DOS CONCEPTOS CLAVES

1. Comprensión
2. Generación

- Dave: Open the pod bay doors, HAL
- HAL: I am sorry Dave. I am affraid I can't do that.

- Dave: Abre las compuertas, HAL
- HAL: Lo siento, Dave. Me temo que no puedo hacerlo

HAL 9000

- Comprensión de humanos por medio de:

  ✓ Reconocimiento del habla
  ✓ Comprensión del lenguaje natural.

- Comunicación con humanos vía

  ✓ Generación de lenguaje natural
  ✓ Síntesis del habla

INTRODUCCIÓN

HAL 9000

- Hal tiene que reconocer una señal sonora y generar una secuencia de palabras.
- Se requiere conocimientos de:
  - ✓ Fonética: Naturaleza física de los sonidos

HAL 9000

- Debe reconocer, entre otras cosas
  - ✓ Género de los sustantivos: Gallo vs Gallina. Gato-Gata
  - ✓ Plural: perro vs perros
- El asunto se complica un poco cuando:
- Casa no es el femenino de Caso
- Ni Luzs ni Luzes es el plural de Luz.

INTRODUCCIÓN

HAL 9000

- Si se agregan prefijos y sufijos a una palabra existente, se generan otras palabras:

  - ✓ Creíble – Increíble
  - ✓ Calmada - Calmadamente

Morfología: Estudio de la estructura interna de las palabras.

HAL 9000

- Se debe conocer el orden correcto de las palabras para que una oración tenga sentido.

- Hal: Lo siento, Dave. Me temo que no puedo hacerlo.

- - Dave, lo siento. Que no puedo hacerlo, me temo. (mismo sentido)
- - No Dave, lo siento. Temo que puedo hacerlo. (cambia el sentido)

Sintaxis: estudio de la estructuración (orden y agrupamiento) de las palabras.

INTRODUCCIÓN

HAL 9000

- Debe entender el significado de lo que le están diciendo.

Semántica Léxica: Significado de las palabras.

- ✓ Banco, Sirena, Hoja, Carta.

HAL 9000

- Hal: Lo siento, Dave. Me temo que no puedo hacerlo.

Uso apropiado y educado del lenguaje.

Pragmática: rama de la lingüística que estudia la influencia del contexto en la interpretación de un significado.

(Wikipedia)

NIVELES DE PROCESAMIENTO DE LENGUAJE NATURAL

- Fonética y Fonología: estudio de los sonidos lingüísticos.
- Morfología: estudio de la estructura interna de las palabras.
- Sintaxis: estudio de la estructuración (orden y agrupamiento) de las palabras.
- Semántica: estudio del significado.
- Pragmática: estudio de cómo el lenguaje se utiliza en un contexto y su significado.
- Discurso: estudio de las unidades mayores a la oración.

HITOS HISTÓRICOS

ALGUNOS HITOS DEL PLN

DÉCADA DEL 50

- El primer problema que aparece es el de la traducción automática.
- Experimento IBM-Georgetown
- Logro traducir de Ruso al Inglés más de 60 oraciones.
- Marco un hito en el inicio del PLN

ALGUNOS HITOS DEL PLN

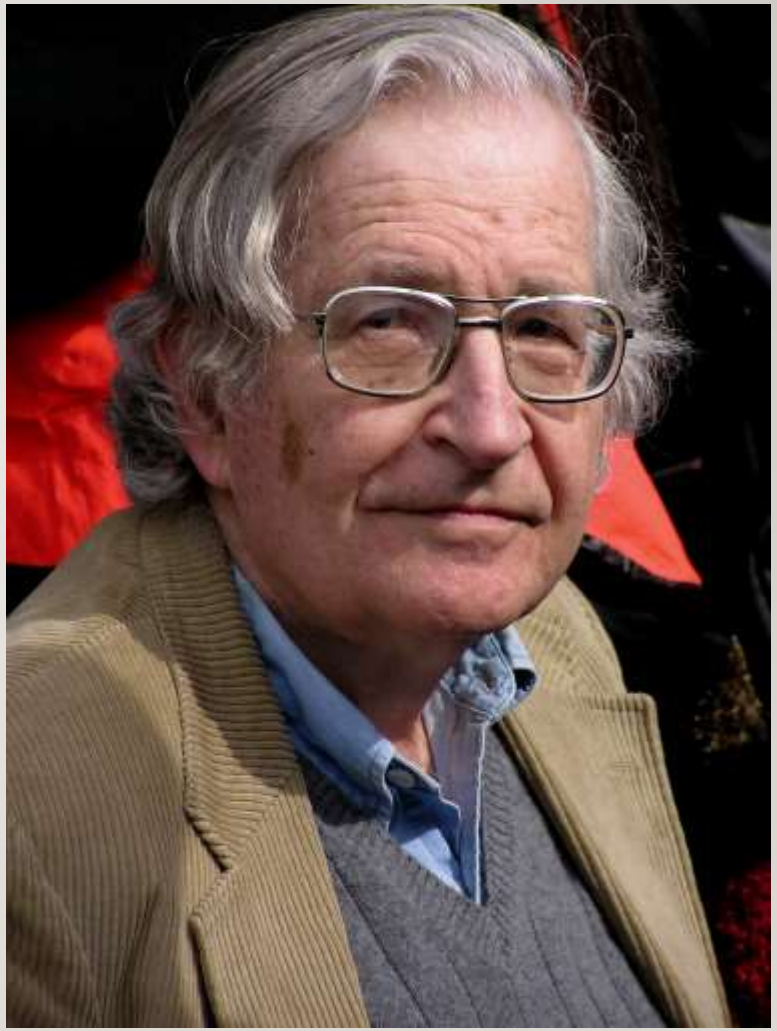
DÉCADA DEL 50

- Alan Turing: Computing Machinery and Intelligence.
- ¿Can machines think?
- Test de Turing: mide capacidad de una máquina para exhibir un comportamiento inteligente similar al de un ser humano.

ALGUNOS HITOS DEL PLN

DÉCADA DEL 50

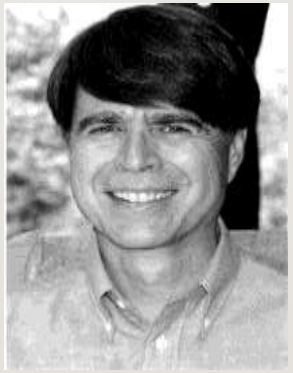
- En 1957 desarrollos notables en "Syntactic Structures" : Noam Chomsky.
- Los idiomas que usamos los seres humanos tienen características o principios comunes en su propia estructura.
- Aportes a la Lingüística.



ALGUNOS HITOS DEL PLN

DÉCADA DEL 60

- Jay Early (1970) : algoritmo no determinista de análisis sintáctico para gramáticas libres.



- Cocke-Kasami-Younger (1965): algoritmo que determina si una cadena puede ser generada por una gramática libre de contexto.

ALGUNOS HITOS DEL PLN

DÉCADA DEL 70

- Richard Montague

  ✓ "English as a Formal Lenguaje"
  ✓ Estudio el enfoque lógico de la semántica del lenguaje natural

DÉCADA DEL 70

- Alain Colmerauer: Prolog (1972)

  ✓ Lenguaje de programación basado en lógica.
  ✓ Trabaja cálculo de predicados

ALGUNOS HITOS DEL PLN

DÉCADA DEL 90

- Frederik Jeinek
  - ✓ Traducción estadística y reconocimiento de voz (IBM)
  - ✓ Se trabaja PLN basado en datos.

AÑOS 2000

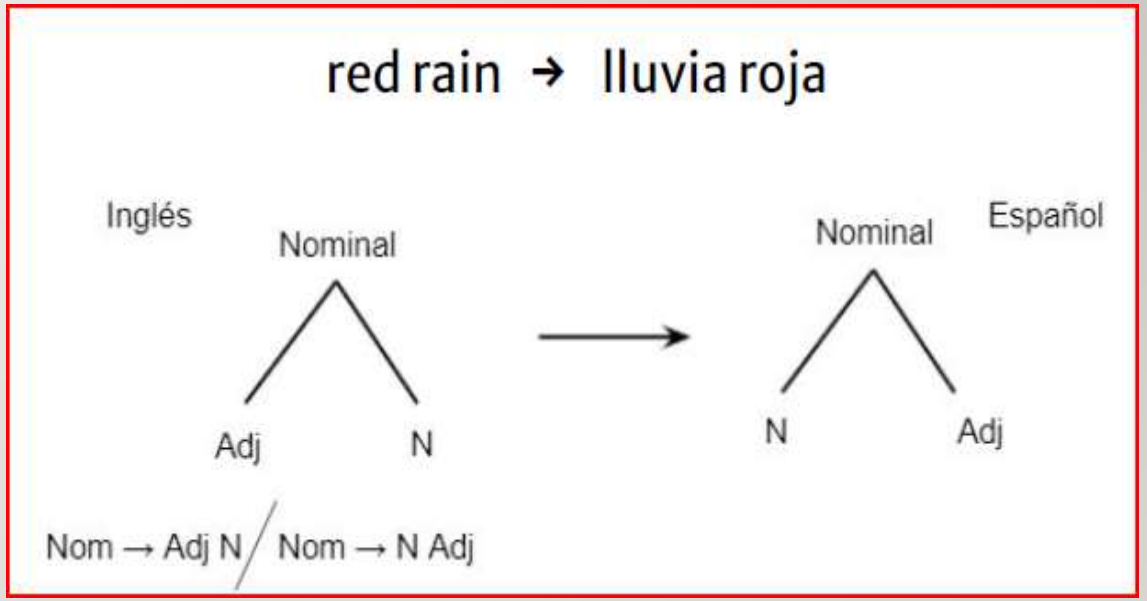
- Vladimir Vapnik
  - ✓ Support Vector Machines
  - ✓ Inteligencia Artificial
  - ✓ Facebook

PLN EN LA ACTUALIDAD

PLN EN LA ACTUALIDAD

TRADUCCIÓN AUTOMÁTICA

- De los primeros problemas de PLN
- Tiene dificultades importantes:
  - ✓ Divergencia Léxica: pata vs pierna vs leg
  - ✓ Sujetos Omitidos en algunos idiomas.
  - ✓ Triángulo de Vauquois : traducción basada en reglas.



PLN EN LA ACTUALIDAD

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica.La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

(2013) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs Calcium, and are playing a duel to stay with Diego Forlan, the Uruguayan striker was voted World Player of Sudáfrica.La Inter manages figure is far from the 36 million euros for the striker's release clause. But the club president Massimo Moratti propose an improvement in the player's salary, who according to the Italian daily receive about 4 million euros until 2013.

(2014) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the EPL, and are playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted the best player in the World Sudáfrica.La Inter manages figure is far from the 36 million euros of the termination clause of the scorer. But the club president Massimo Moratti propose an improvement in the player's salary, according to the Italian newspaper who will receive about 4 million euros until 2013.

PLN EN LA ACTUALIDAD

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica.La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

(2015) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the Calcio, and they are playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted best player of the World Sudáfrica.La Inter figure handles is far from the 36 million euros of the termination clause of the scorer. But the club president Massimo Moratti propose an improvement in the player's salary, who the Italian daily receive about 4 million euros until 2013.

(2016) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the Calcio, and they are playing a duel to stay with Diego Forlan, the Uruguayan striker who was chosen as the best player in the World Sudáfrica.La Inter manages figure is far from the 36 million euros of the termination clause scorer. But the club president Massimo Moratti propose an improvement in the player's salary, according to the Italian daily who will receive about 4 million euros until 2013.

PLN EN LA ACTUALIDAD

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica.La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

(2017) The Italian championship has not yet begun but Inter Milan and Juventus, two of Calcio's most powerful clubs, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player of the World Cup in South Africa. Inter's figure is far from the 36 million euros of the scorer's termination clause. But the club that presides Massimo Moratti will propose an improvement in the salary of the player, who according to the Italian newspaper will receive close to 4 million euros until 2013.

(2018) The Italian championship has not yet begun but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan forward who was chosen as the best player in the World Cup in South Africa. The amount handled by Inter is far from the 36 million euros of the rescission clause of the scorer. But the club chaired Massimo Moratti propose an improvement in the salary of the player, who according to the Italian newspaper will receive about 4 million euros until 2013.

PLN EN LA ACTUALIDAD

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica.La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

(2018) The Italian championship has not yet begun but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan forward who was chosen as the best player in the World Cup in South Africa. The amount handled by Inter is far from the 36 million euros of the rescission clause of the scorer. But the club chaired Massimo Moratti propose an improvement in the salary of the player, who according to the Italian newspaper will receive about 4 million euros until 2013.

(2019) The Italian championship has not yet started but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. Inter's figure is far from the 36 million euros of the termination clause of the scorer. But the club chaired by Massimo Moratti will propose an improvement in the salary of the player, who according to the Italian newspaper will receive about 4 million euros until 2013.

PLN EN LA ACTUALIDAD

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica.La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

(2020) The Italian championship has not yet started but Inter Milan and Juventus, two of Calcio's most powerful clubs, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. The figure that Inter manages is very far from the 36 million euros of the termination clause of the scorer. But the club chaired by Massimo Moratti will propose an improvement in the salary of the player, who, according to the Italian newspaper, will receive about 4 million euros until 2013.

(2021) The Italian championship has not yet started but Inter Milan and Juventus, two of Calcio's most powerful clubs, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. The figure that Inter manages is very far from the 36 million euros of the termination clause of the scorer. But the club chaired by Massimo Moratti will propose an improvement in the player's salary, who according to the Italian newspaper will receive about 4 million euros until 2013.

PLN EN LA ACTUALIDAD

RESUMEN AUTOMÁTICO

- Resumir contenido de la información de un documento.

- Primeros trabajos se realizaron sobre 1960.
  - ✓ Basado en métodos estadísticos
  - ✓ Extracción de oraciones principales.
  - ✓ Peso en las oraciones.

RESUMEN AUTOMÁTICO

- Existen dos paradigmas fundamentales:

  - ✓ Extracción: extraer oraciones o fragmentos de oraciones literales relevantes de un texto original

  - ✓ Abstracción: regeneración de los fragmentos relevantes del texto original.

PLN EN LA ACTUALIDAD

RECUPERACIÓN DE INFORMACIÓN

Representación, almacenamiento y acceso y recuperación de la información a partir de un conjunto de documentos.

- Aspectos representativos de los documentos.
- Recuperación de información
- Consultas
- Relevancia del documento.

RECUPERACIÓN DE INFORMACIÓN

- Métodos basados en modelos vectoriales.
- Evaluación de algoritmos con métricas como Precision y Recall.
- Comparación por grado de similitud entre documentos.

PLN EN LA ACTUALIDAD

EXTRACCIÓN DE INFORMACIÓN

Responder consultas

- Analizar texto
- Extraer: entidades, relaciones y eventos de los textos.

OTRAS APLICACIONES

- Análisis de discurso
- Categorización de documentos
- Respuestas a preguntas
- Análisis de sentimientos
- Chat boots
- Chat GPT y otras aplicaciones

TRANSFORMERS

Es una arquitectura de propósito general basada en las redes neuronales para el NLP en la cual está basada: BERT (Google), GPT (Open AI), RoBERTa, XLM, DistilBert, XLNet, etc.

LENGUAJE Y AMBIGÜEDAD

LENGUAJES

TIPOS DE LENGUAJES

FORMALES

- Definidos por reglas pre-establecidas.
  - ✓ Símbolos
  - ✓ Reglas
  - ✓ Por ejemplo: matemática-lógica

NATURALES

- Utilizados para la comunicación humana
- Evolucionan
- Las reglas se van generando en el tiempo.

¿Qué tiene el lenguaje natural que no tienen los lenguajes formales?

AMBIGÜEDAD

CASOS DE AMBIGÜEDAD

- Ambiguo: tiene distintas interpretaciones.
- Homofonía: Ola/Hola , As/Has .
- Polisemia: múltiples significados.
  - ✓ Sirena

  - ❖ El hombre desciende del mono
  - ❖ El mono desciende del árbol

Ambigüedad Fonética

- Ató dos palos // A todos palos
- Yo loco, loco, y ella loquita // Yo lo coloco y ella lo quita.
- El dulce lamentar de los pastores // El dulce lamen tarde los pastores

CASOS DE AMBIGÜEDAD

Ambigüedad Morfológica

Nosotros plantamos papas

Ambigüedad Sintáctica

Pedro vio a Juan con un telescopio

CASOS DE AMBIGÜEDAD

Ambigüedad Sintáctica

Los hombres y las mujeres que hayan cumplido 60 años pueden solicitar una pensión.

Ambigüedad Sintáctica
(Cuantificadores)

- Todos los hombres aman a una mujer

- Todos los estudiantes leyeron un libro

CASOS DE AMBIGÜEDAD

Ambigüedad Pragmática

Llego a las ocho. Espérame.

- Previsión

¿A que horas llegas?
Rta: Llego a las ocho. Espérame

- Promesa

Nunca llegas a tiempo
Rta: Llego a las ocho. Espérame

Ambigüedad en Discurso

Tomé el alfajor del escritorio y lo comí

a) Tomé el alfajor que estaba en el escritorio y comí el alfajor.

b) Tomé el alfajor que estaba en el escritorio y comí el escritorio.

ACERCA DEL PLN

DIFICULTADES

- Alta ambigüedad
- Complejo
- Tener en cuenta el entorno
- Evolucionar

SOLUCIONES A PARTIR DEL ALGORITMOS Y MODELOS

- Lógica
- Teoría de probabilidad
- Modelos basados en redes neuronales
- Aprendizaje automático
- Programación dinámica.

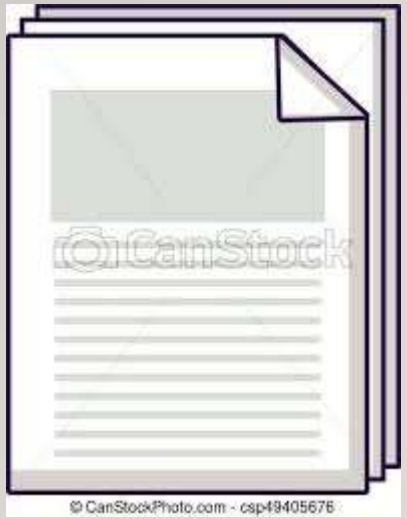
MODELOS COMPUTABLES

Nivel Morfológico, Léxico y Sintáctico

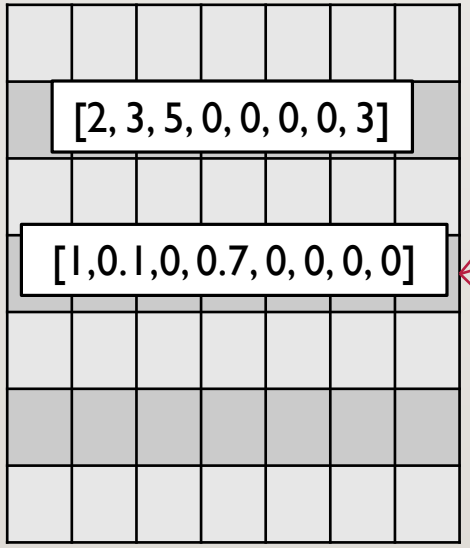
- Bag of words
- Terms Frequency
- TF-IDF
- N-Grams

Conjunto de Documentos
Corpus (contexto)

Representación computable
Dataset
(Vector de Característica)



Convertir



Nivel Semántico

Word Embeddings
- Word2vec
- GloVe
- FastText

Nivel de discurso y pragmático

Arquitectura Transformers y Modelos de Lenguaje
- OpenAI GPT
- BERT
- ELMo
- ERNIE

CONCEPTOS GRAMATICALES

CONCEPTOS

GRAMÁTICA

Estudia las unidades significativas del lenguaje y su combinatoria.

Es el estudio de la reglas y principios de una determinada lengua.

Dos grandes partes:

MORFOLOGÍA: Estudia la estructura interna de las palabras

SINTAXIS: Estudia la estructura de la oración, es decir, la combinatoria de las palabras.

CONCEPTOS

ORACIÓN

Es una unidad de predicación en donde se establece una relación entre dos constituyentes: sujeto y predicado.

Predicado: es lo que se dice del sujeto.

- El carro azul tiene pico y placa
- Aterrizó el avión

- Los artistas se ofrecieron a colaborar desde el principio.
- Voy a la Universidad todos los días
- La verdad siempre sale a la luz

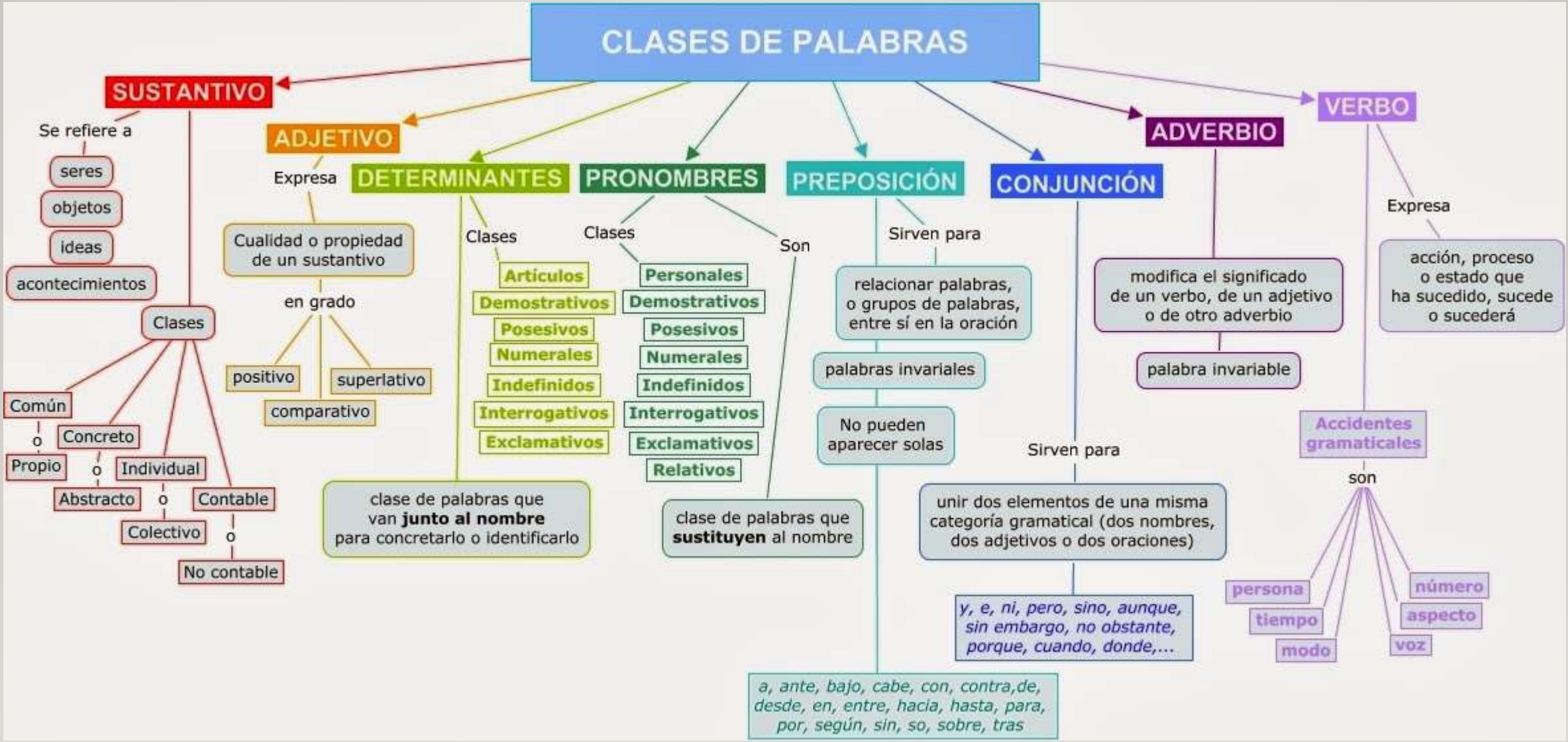
En un texto, la unidad es el enunciado, que va desde la mayúscula hasta el punto.

CATEGORIAS GRAMATICALES LÉXICAS

- Nombre/sustantivo
  - ✓ Casa, casas, felicidad,
- Verbo
  - ✓ Correr, pensar, estudiar
- Adjetivo
  - ✓ Alto, bajo, lindo, solo
- Adverbio
  - ✓ Solamente, ayer, rápidamente.

- Determinante
  - ✓ El, una, unos, esos, nuestro
- Preposición
  - ✓ A, de por, contra
- Pronombre
  - ✓ Yo, él, mí aquello, ese
- Conjunción
  - ✓ Y, o, si, que

CATEGORIAS GRAMATICALES LÉXICAS



¡GRACIAS!