UNIMINUTO
Corporación Universitaria Minuto de Dios
Educación de calidad al alcance de todos
Vigilada MinEducación

ESPECIALIZACIÓN EN INTELIGENCIA ARTIFICIAL

Procesamiento de Lenguaje Natural (NPL)

TEMARIO:

Semana 1 – Estructura del lenguaje natural

Semana 2 – Procesamiento léxico y morfológico

Semana 3 – Librerías para el NPL

Semana 4 – Aplicar redes neuronales al NPL

Semana 5 – Aplicación del NPL

Semana 6 – Programación de la inteligencia artificial

Semana 7 – Uso de las redes neuronales recurrentes

Semana 8 – Usos de GPT

METODOLOGÍAS ACTIVAS

Estudios de caso:

- Calentamiento sin código (conceptos clave)
- Tareas del PLN
- Enfoques

Estructura del lenguaje natural

Definición de Procesamiento de Lenguaje Natural (PLN / NPL)

El **Procesamiento de Lenguaje Natural (PLN)** es un área de la inteligencia artificial que busca que las computadoras **entiendan, interpreten y generen lenguaje humano** (escrito o hablado).

Se nutre de varias disciplinas:

Lingüística → Reglas gramaticales, estructura del lenguaje, significados.

Informática → Algoritmos, programación, estructuras de datos.

Psicología → Cómo los humanos procesan y comprenden el lenguaje.

Ingeniería → Creación de sistemas aplicados (chatbots, traductores, buscadores)

Analizar los problemas del PLN significa detectar las dificultades que surgen cuando intentamos que la máquina entienda lo que para los humanos es natural (ambigüedades, ironía, sinónimos, contexto, etc.).

Estructura del lenguaje natural

Estructura del lenguaje natural en PLN (Procesamiento de Lenguaje Natural)

En **Procesamiento de Lenguaje Natural (PLN)**, cuando hablamos de la **estructura del lenguaje natural**, nos referimos a cómo se organiza y analiza el lenguaje humano para que una máquina pueda procesarlo.

Esta estructura se adapta desde la lingüística, pero con un enfoque computacional.
Se suele dividir en **múltiples niveles**

Nivel	Qué analiza	Ejemplo
Fonético / Fonológico	Sonidos, pronunciación, entonación (aplica a voz).	“casa” vs “caza” → suenan parecido, significan distinto.
Morfológico	La forma de las palabras: raíces, afijos, conjugaciones.	estudiantes → raíz: estudiar + sufijo -antes.
Léxico	Palabras como unidades de un diccionario mental.	“banco” → puede ser institución financiera o asiento.
Sintáctico	Reglas gramaticales y estructura de oraciones.	“El perro muerde al gato” ≠ “El gato muerde al perro”.
Semántico	El significado de palabras y oraciones.	“Python” → ¿lenguaje de programación o serpiente?
Pragmático	El uso del lenguaje en contexto real.	“¿Puedes abrir la ventana?” → no pregunta por capacidad, es una petición.
Discursivo	Cómo se conectan varias oraciones en un texto coherente.	En una noticia: introducción, desarrollo, conclusión.

Estructura del lenguaje natural

Enfoques en PLN

Simbólico (basado en reglas)

Usa gramáticas, reglas lingüísticas y diccionarios.

Ejemplo: “Si la palabra termina en *-ar*, probablemente es un verbo en infinitivo.”

Problema: requiere miles de reglas → difícil de mantener.

Estadístico (basado en datos)

Usa probabilidades y frecuencia de palabras en grandes corpus.

Ejemplo: *bag of words* → contar cuántas veces aparece cada palabra.

Problema: pierde el orden y el contexto.

Hoy en día se combinan con redes neuronales y modelos de deep learning (lo veremos más adelante).

Concepto de *Corpus* y Tokenización

Corpus

Conjunto grande de textos (ej: Wikipedia, noticias, tuits) que se usan para entrenar modelos de lenguaje.

Tokenización

Dividir un texto en piezas manejables (palabras, frases, o incluso sub-palabras).

Ejemplo:
Texto = *“Los modelos de lenguaje son poderosos”*

Tokens: ["Los", "modelos", "de", "lenguaje", "son", "poderosos"]

Estructura del lenguaje natural

La Metáfora

Enseñar PLN es como enseñar a un extranjero a cocinar en nuestro país:

Lingüística → Le enseñas la receta y el idioma de los ingredientes.
Informática → Le das utensilios y técnicas de cocina (algoritmos).
Psicología → Le explicas *cómo* los humanos deciden qué plato cocinar.
Ingeniería → Lo llevas a una cocina real con clientes (aplicación práctica).

- El corpus sería su *libro de recetas* (miles de ejemplos).
- La tokenización serían los *ingredientes* (tomates, cebollas, arroz).
- Los enfoques **simbólico** y **estadístico** son como dos formas de aprender:

Memorizar reglas (simbólico).
Probar y contar cuántas veces un plato sale rico (estadístico).

Estructura del lenguaje natural

Estudio de caso 1

Texto:
“Los estudiantes de ingeniería usan Python para proyectos de inteligencia artificial en la universidad.”

Fases de preprocesamiento:

1.Segmentación: Detectar oraciones.
→ Aquí hay solo una.

2.Tokenización: ["Los", "estudiantes", "de", "ingeniería", "usan", "Python", "para", "proyectos", "de", "inteligencia", "artificial", "en", "la", "universidad"]

3.Normalización: pasar a minúsculas → "python“

4.Limpieza: quitar puntuación/stopwords → ["estudiantes", "ingeniería", "usan", "python", "proyectos", "inteligencia", "artificial", "universidad"]

5.Lematización: "usan" → "usar“

6.Vectorización (estadístico): contar frecuencia de cada palabra.

Estructura del lenguaje natural

Estudio de caso 2

Texto:

"Los algoritmos modernos de aprendizaje automático permiten clasificar grandes volúmenes de datos.“

Tokenización: ["Los", "algoritmos", "modernos", "de", "aprendizaje", "automático", "permiten", "clasificar", "grandes", "volúmenes", "de", "datos"]

Normalización: ["los", "algoritmos", "modernos", "de", "aprendizaje", "automático", "permiten", "clasificar", "grandes", "volúmenes", "de", "datos"]

Limpieza (sin stopwords): ["algoritmos", "modernos", "aprendizaje", "automático", "permiten", "clasificar", "grandes", "volúmenes", "datos"]

Lematización: ["algoritmo", "moderno", "aprendizaje", "automático", "permitir", "clasificar", "grande", "volumen", "dato"]

Problemas detectados:

1."automático" puede ser adjetivo o sustantivo según contexto.
2."volúmenes" → ¿se mantiene como "volumen" o se pierde el matiz de pluralidad?
3."aprendizaje automático" es un concepto técnico, no debería dividirse.

Estructura del lenguaje natural

Estudio de caso 3

Texto:
"El presidente anunció nuevas medidas económicas para apoyar a las pequeñas empresas.“

Tokenización: ["El", "presidente", "anunció", "nuevas", "medidas", "económicas", "para", "apoyar", "a", "las", "pequeñas", "empresas"]

Normalización: ["el", "presidente", "anunció", "nuevas", "medidas", "económicas", "para", "apoyar", "a", "las", "pequeñas", "empresas"]

Limpieza: ["presidente", "anunció", "nuevas", "medidas", "económicas", "apoyar", "pequeñas", "empresas"]

Lematización: ["presidente", "anunciar", "nuevo", "medida", "económico", "apoyar", "pequeña", "empresa"]

Problemas detectados:
"presidente" puede referirse a un cargo o a una persona concreta.
"medidas económicas" debe verse como unidad semántica.
"nuevas" → se pierde el matiz de plural/femenino al pasar a lema.

Estructura del lenguaje natural

Estudio de caso 4

Texto:
"Los estudiantes de ingeniería usan Python para proyectos de inteligencia artificial en la universidad."

Tokenización: ["Los", "estudiantes", "de", "ingeniería", "usan", "Python", "para", "proyectos", "de", "inteligencia", "artificial", "en", "la", "universidad"]

Normalización: ["los", "estudiantes", "de", "ingeniería", "usan", "python", "para", "proyectos", "de", "inteligencia", "artificial", "en", "la", "universidad"]

Limpieza: ["estudiantes", "ingeniería", "usan", "python", "proyectos", "inteligencia", "artificial", "universidad"]

Lematización: ["estudiante", "ingeniería", "usar", "Python", "proyecto", "inteligencia artificial", "universidad"]

Problemas detectados:
"Python" no debe traducirse a "pitón".
"inteligencia artificial" debe tratarse como concepto compuesto.
"usan" puede perder información temporal (tiempo verbal).

Estructura del lenguaje natural

Pautas de aprendizaje específico

Analizar problemas del PLN significa:

i. Detectar **ambigüedad semántica** (una palabra con varios significados).

ii. Identificar **dependencia del contexto** (qué significa “ellos” depende de la oración anterior).

iii. Notar **variabilidad morfológica** (correr, corría, corrimos).

iv. Reconocer **ruido en el texto** (errores ortográficos, emojis, hashtags).

Pautas para Lematizar

La **lematización** consiste en reducir una palabra a su forma base.

- **Identifica el verbo en infinitivo** → "usando" → *usar*, "corría" → *correr*.
- **Identifica el sustantivo en singular** → "estudiantes" → *estudiante*.
- **Adjetivos en su forma básica** → "mejores" → *mejor*.
- **Mantén nombres propios y acrónimos** → "Python", "ONU".
- **Cuida expresiones compuestas** → "inteligencia artificial" debe tratarse como unidad.

Estructura del lenguaje natural

Actividad práctica

Objetivo: Identificar problemas y fases de preprocesamiento.

Tomar uno de los 3 casos de análisis y responder el padlet

Realizar:

1. **Tokenización**: separa palabras.
2. **Normalización**: convierte a minúsculas.
3. **Limpieza**: elimina stopwords y signos de puntuación.
4. **Lematización**: intenta reducir palabras a su raíz
   (mientras vemos las librerías de python, hazlo intuitivamente).

Escribe 3 problemas detectados en el texto:

1. Palabras polisémicas (“banco” puede ser institución o asiento).
2. Nombres propios (“Python” no debería convertirse en “pitón”).
3. Expresiones compuestas (“inteligencia artificial” debe tratarse como un concepto, no dos palabras separadas).

Actividad práctica – casos de análisis

Texto 1 – Salud

“Los médicos recomiendan realizar actividad física tres veces por semana para mejorar la salud cardiovascular.”

Texto 2 – Economía

“El mercado financiero mostró una caída significativa debido a la inflación y la inestabilidad política en la región.”

Texto 3 – Tecnología

“La inteligencia artificial está revolucionando la forma en que las empresas gestionan datos y procesos de negocio.”

https://padlet.com/sergiopuertomo/tareas-del-pln-5hxgrlhxg5tmx8yd