NLP

Curso de Especialización en Inteligencia Artificial y Big Data

Francisco Gallego Perona

Contenidos

C1. Procesamiento del Lenguaje Natural (NLP)

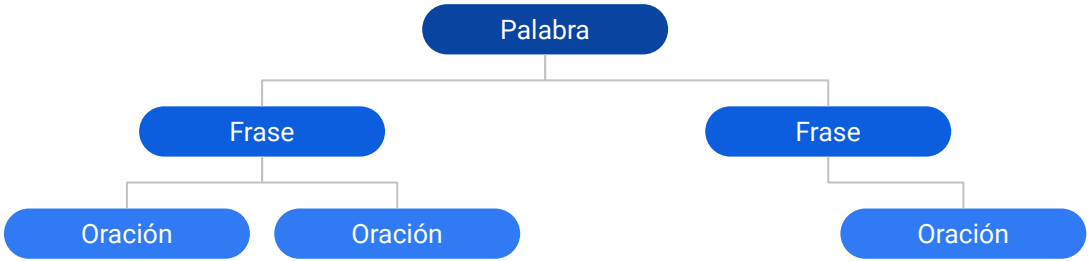
C1.1. Procesamiento del lenguaje natural

C1.2. Aplicaciones del procesamiento del lenguaje natural

¿Qué es NLP?

NLP (Procesamiento de lenguaje natural) desarrolla algoritmos para analizar y representar el lenguaje humano de forma automática.

Analizando la estructura del lenguaje, los sistemas de aprendizaje computacional pueden procesar largos conjuntos de palabras, frases y oraciones.



Obstáculos de NLP

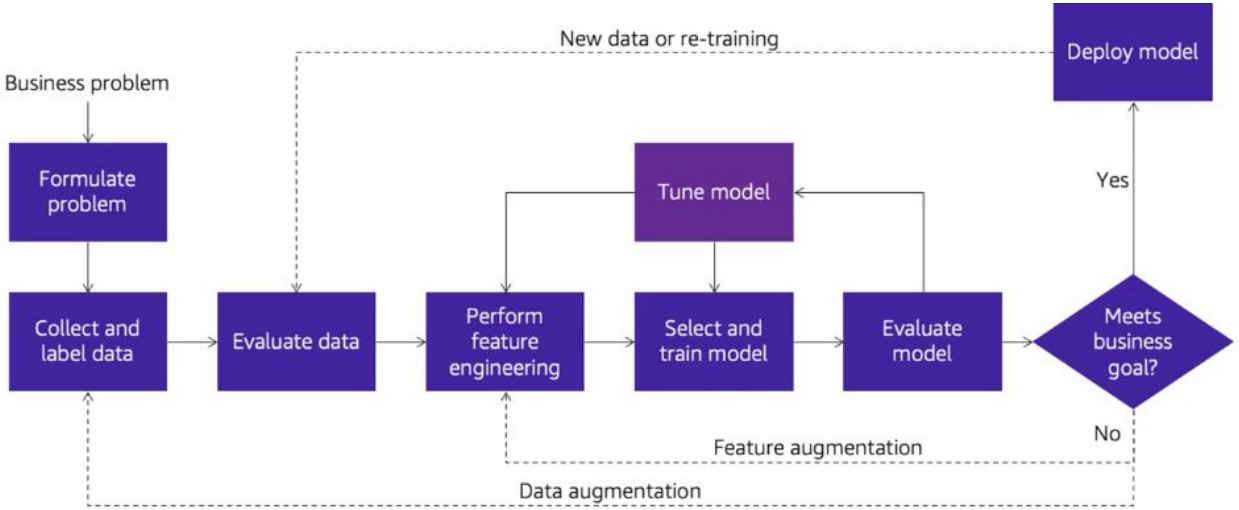
- El lenguaje suele ser **impreciso** → Depende de cada hablante/escritor

- Contiene una cantidad de **dependencias complejas** → El castellano es más complejo que el inglés en cuanto a las construcciones del lenguaje

- El significado de un texto está basado en el **contexto**

- Carece de **estructura** → No se puede insertar en una tabla

ML Pipeline adaptado a NLP



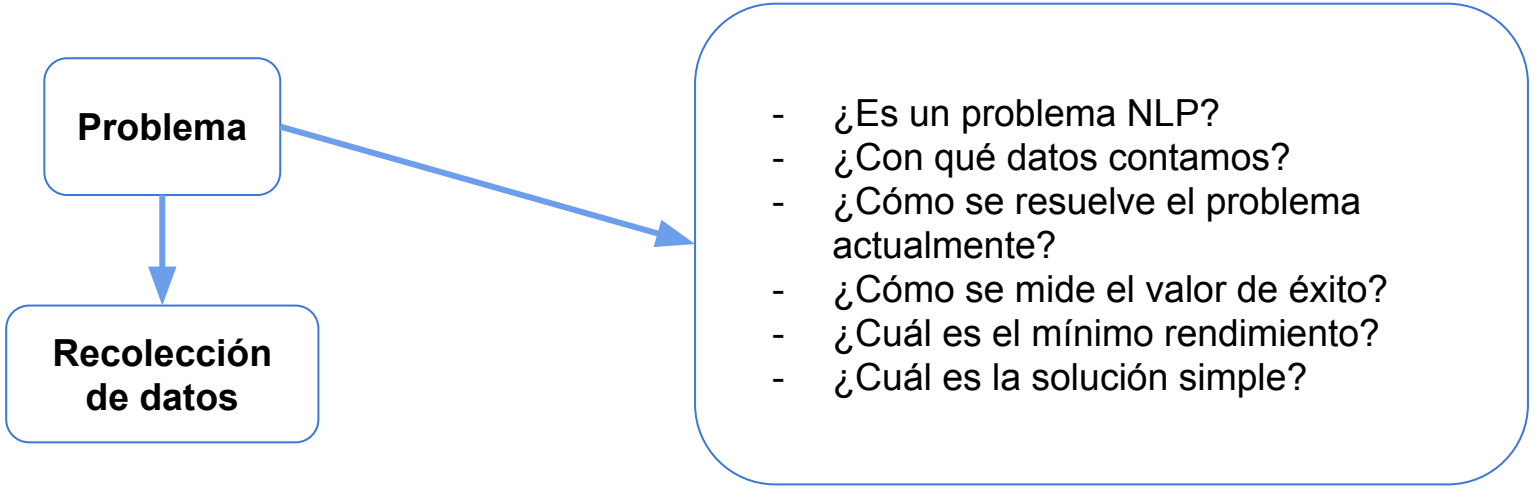
Algunos casos de uso de NLP

- Aplicaciones de búsqueda → Google, Bing
- Traducción
- Chatbots
- Interfaces Humano máquina

Fases de NLP

El texto se trabaja en su mayoría usando **modelos no supervisados** → Los datos de texto no suelen estar etiquetados.

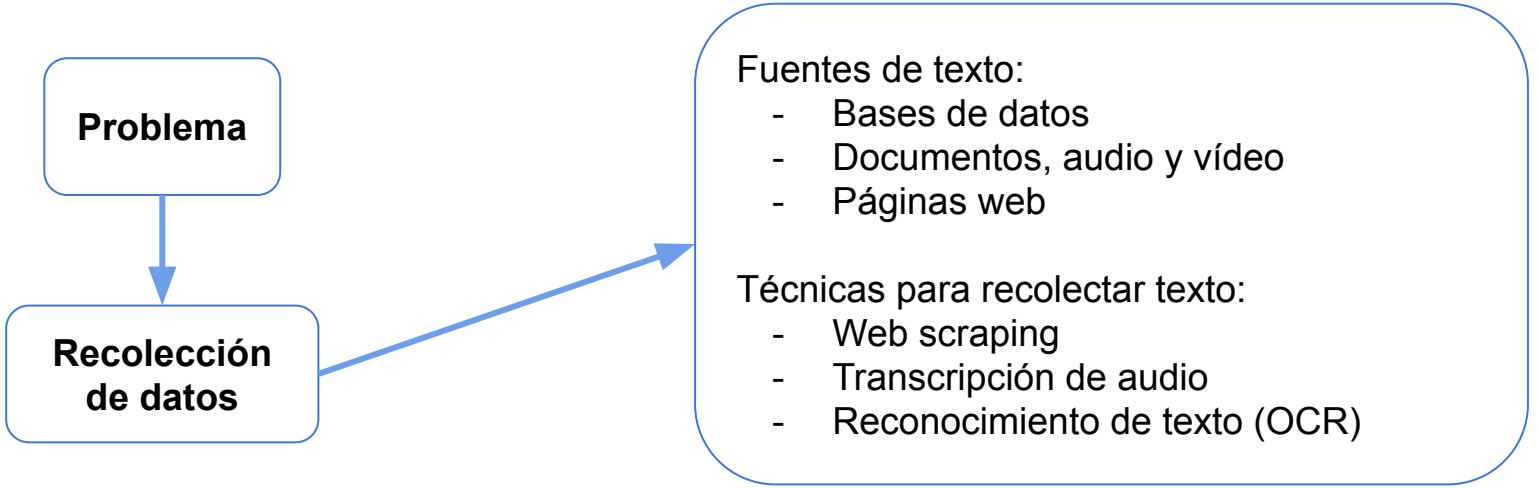
Formulación de un problema



Fases de NLP

El texto se trabaja en su mayoría usando **modelos no supervisados** → Los datos de texto no suelen estar etiquetados.

Recolección de datos



Fases de NLP

Evaluación del texto

1. ¿Qué **lenguajes** son utilizados en el texto? → La estructura del lenguaje cambia de un lenguaje a otro.
2. ¿Qué **estilo de escritura** se utiliza?
    a. Texto científico
    b. Un post en redes sociales
    c. Libro
3. ¿Hay **errores** en el texto? → Cuanto menos errores, mejores resultados al procesar el texto.
4. ¿Contamos con **caracteres especiales**? → Emojis, símbolos de puntuación, fórmulas matemáticas...
5. ¿Hay **términos de un dominio específico**?
6. Conversión de mayúsculas en minúsculas
7. Valores numéricos → Los podemos convertir en su representación textual o extraer su valor numérico.
8. Formas verbales → Conversión de palabras con la misma raíz a una única estándar.

Fases de NLP

Feature engineering (Ingeniería de características)

Después de evaluar los datos → Debemos procesarlos para extraer las características que necesitamos.

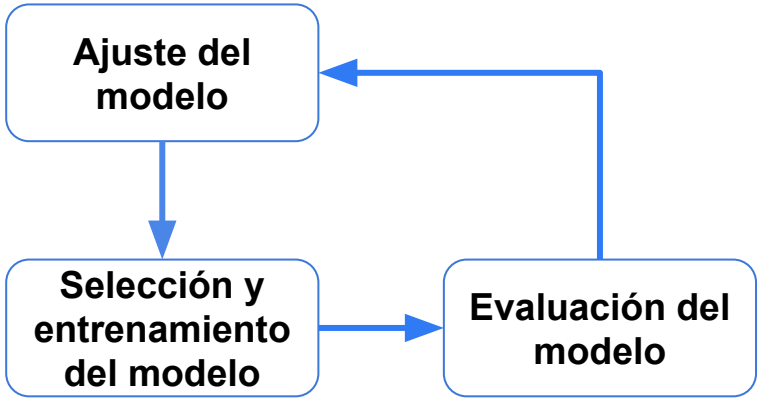
Preprocesamiento de texto:

- Normalización
- Stemming
- Lemmatization

Transformación de palabras en números → Si transformamos cada palabra en un número, tendremos cientos de columnas.

Fases de NLP

Selección, entrenamiento, evaluación y ajuste del modelo



División de los datos para:

- Training
- Testing
- Validation

Tareas comunes en NLP

- Extracción de texto de la web:
  - Herramientas: Beautiful Soup, Scrapy → En el lenguaje de programación python

- Obtención de texto desde bases de datos relacionales
- Datos a través de OCR
- Transcripción de audio o vídeo

Tareas comunes en NLP

- Eliminación de las stopwords
- Normalización
  - Eliminación de los signos de puntuación
  - Conversión de números a texto
  - Conversión de texto a lowercase

- Stemming
- Lemmatization
- Tokenización