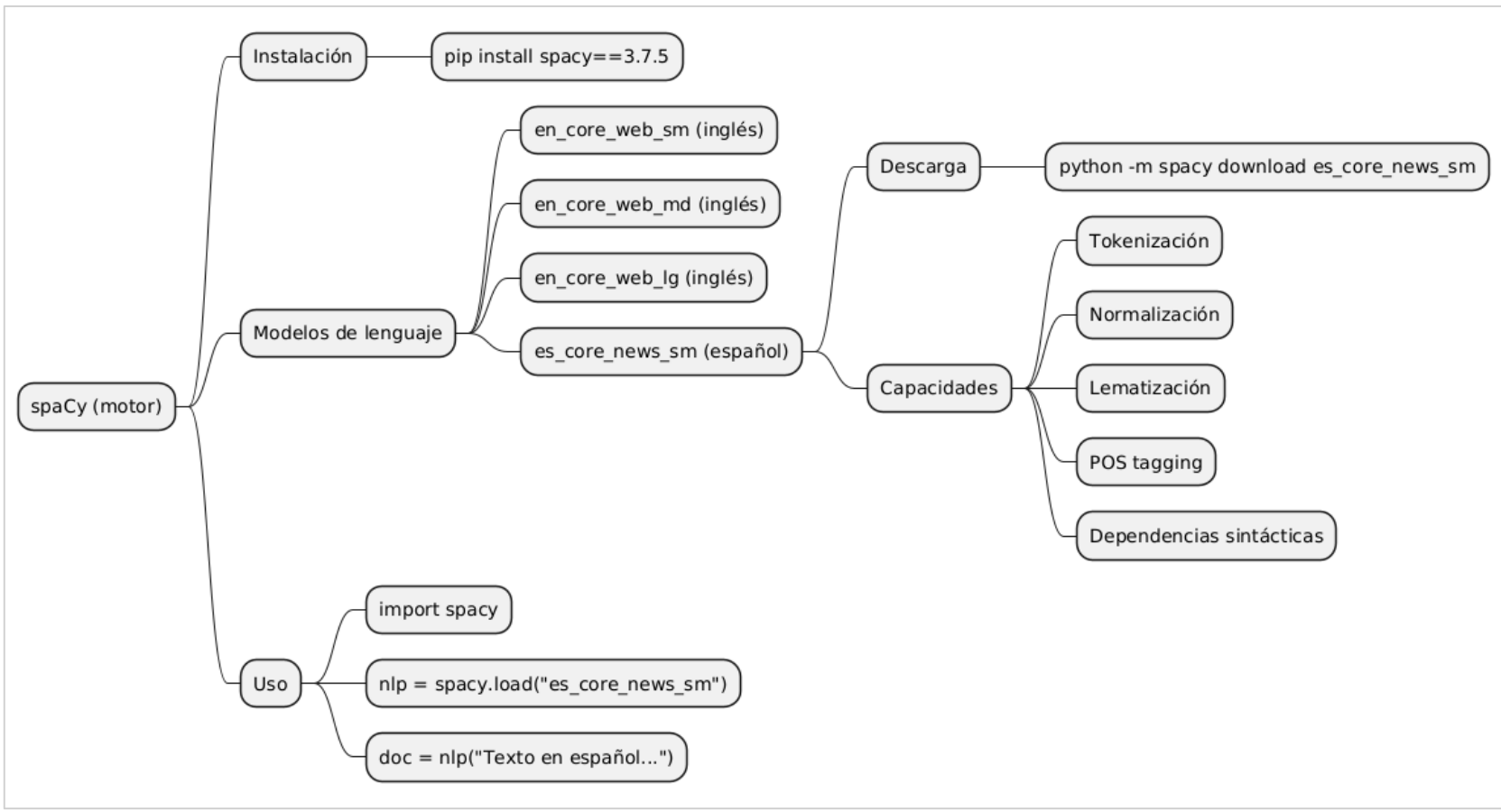
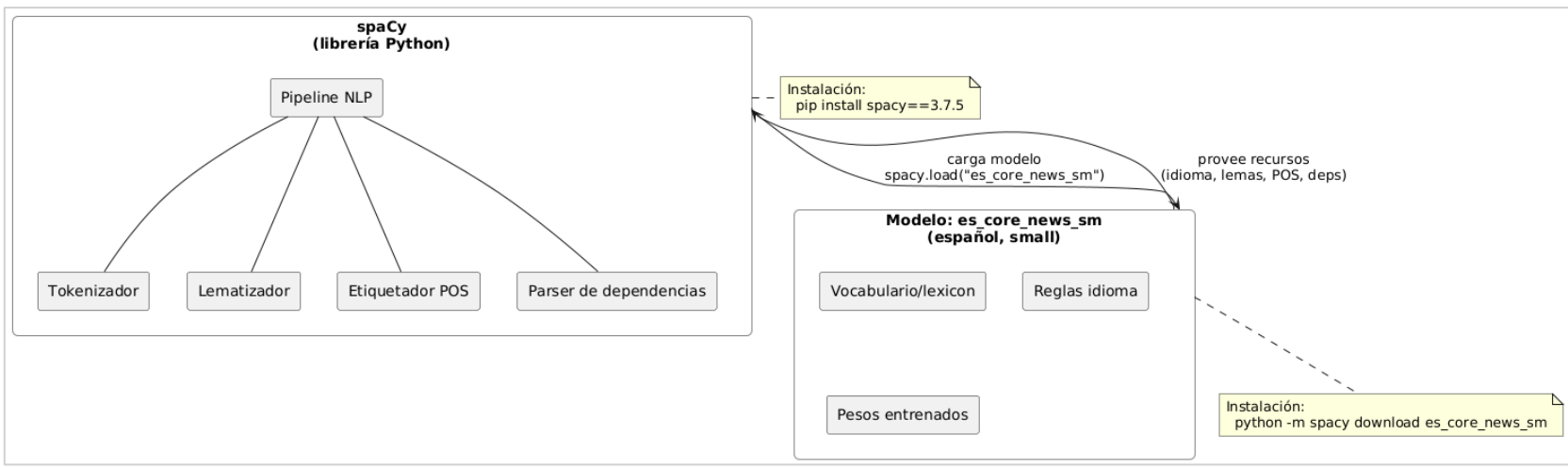
La librería **spaCy**





¿Qué es la librería spaCy?

spaCy es una **librería de Python especializada en Procesamiento de Lenguaje Natural (PLN)**, diseñada para que las computadoras puedan **entender, procesar y analizar textos escritos en lenguaje humano**.

Se caracteriza por ser **rápida, eficiente y lista para producción** (no solo para investigación).

Funciones principales de spaCy

- **Tokenización** → divide un texto en palabras, frases o símbolos.
- **Lematización** → reduce las palabras a su forma base (*corría* → *correr*).
- **Etiquetado gramatical (POS tagging)** → identifica categorías como sustantivo, verbo, adjetivo.
- **Análisis de dependencias** → detecta relaciones sintácticas entre palabras.
- **Reconocimiento de entidades (NER)** → reconoce nombres propios, lugares, fechas, organizaciones.
- **Vectorización** → representa palabras como vectores numéricos (para usarlas en modelos de IA).

Ventajas de spaCy

- Muy rápida (escrita en Cython, mezcla de C y Python).
- Incluye **modelos entrenados** para muchos idiomas (inglés, español, alemán, etc.).
- Orientada a proyectos **reales y en producción** (chatbots, buscadores, análisis de texto).
- Se integra fácil con librerías como **scikit-learn**, **TensorFlow** o **PyTorch**.

NER – Named Entity Recognition

NER (Reconocimiento de Entidades Nombradas) es una técnica del **Procesamiento de Lenguaje Natural (PLN)** que sirve para **identificar y clasificar automáticamente entidades en un texto**, es decir, **palabras o frases que representan cosas específicas del mundo real**.

Tipos comunes de entidades en NER

- **Personas** → "Gabriel García Márquez" → PER
- **Lugares** → "Colombia" → LOC
- **Organizaciones** → "ONU" → ORG
- **Fechas** → "10 de septiembre de 2025" → DATE
- **Cantidades / Dinero** → "$1000" → MONEY
- **Productos** → "iPhone 16" → PRODUCT

¿Para qué sirve NER?

- **Motores de búsqueda** (Google detecta nombres de lugares o personas).
- **Chatbots** (extraer información clave de una conversación).
- **Análisis de noticias** (detectar empresas, fechas, países).
- **Extracción de información** (encontrar medicamentos, síntomas en textos médicos).

NER es la tarea de encontrar "nombres propios e información clave" en un texto y clasificarlos en categorías como persona, lugar, organización, fecha, etc.

Significado de las etiquetas más comunes en NER (spaCy en español)

- **LOC** → **Location (Lugar)**
  Se usa para **lugares geográficos**.
  Ejemplo: *"Colombia"*, *"París"*, *"Andes"*.
- **DATE** → **Fecha**
  Detecta expresiones de **tiempo**.
  Ejemplo: *"1927"*, *"10 de septiembre de 2025"*, *"ayer"*.
- **MISC** → **Miscellaneous (Varios / Miscelánea)**
  Son entidades que no encajan en categorías clásicas pero que tienen importancia semántica.
  Ejemplo: *"Premio Nobel"*, *"COVID-19"*, *"Eurocopa"*.

Las instalaciones

!pip install -q spacy==3.7.5

Le dice a Colab que instale la librería **spaCy** (versión 3.7.5) en el entorno de trabajo.

- pip install = instala paquetes de Python.
- -q = "quiet", para que muestre menos texto en la instalación.
- spacy==3.7.5 = instala exactamente esa versión (asegura compatibilidad).

!python -m spacy download es_core_news_sm

Descarga e instala el **modelo de lenguaje en español pequeño (sm = small)** que necesita spaCy para entender textos en español.

- python -m spacy = ejecuta spaCy como un módulo.
- download es_core_news_sm = baja el modelo entrenado en español (tokenización, lematización, POS, etc.).

El primer comando instala el motor (**spaCy**).
El segundo instala el "cerebro" en español (**modelo lingüístico**).
Con ambos, ya puedes analizar textos en español (tokens, lemas, categorías gramaticales, etc.).

Desglose del nombre es_core_news_sm

- **es** → idioma (**español**).
- **core** → modelo “central” o base de spaCy (los más usados, no especializados).
- **news** → tipo de corpus usado para entrenar (noticias → lenguaje estándar, variado, formal).
- **sm** → tamaño del modelo (**small** = ligero, rápido, menos preciso).

Recordemos

spaCy: se usará para procesar el lenguaje natural (tokenización, lematización, etc.).

pandas: para organizar y analizar datos en tablas.

Matplotlib: para graficar resultados.

Counter: para contar ocurrencias de palabras, tokens u otros elementos.

Significado de las etiquetas (POS)

DET → **Determinante**
Palabra que acompaña a un sustantivo para precisar (los, la, un, este…).
Ejemplo en tu oración: **Los**.

NOUN → **Sustantivo**
Nombra personas, cosas, ideas o lugares.
Ejemplo: **médicos**, **actividad**, **veces**, **semana**.

VERB → **Verbo**
Expresa acción, estado o proceso.
Ejemplo: **recomiendan**, **realizar**.

ADJ → **Adjetivo**
Acompaña a un sustantivo para calificarlo o describirlo.
Ejemplo: **física**.

NUM → **Número**
Palabra que expresa cantidad o número.
Ejemplo: **tres**.

ADP → **Adposición**
Incluye **preposiciones** (a, de, en, por, con…) y **posposiciones** en otros idiomas.
Ejemplo: **por**.