

Introducción © EDICIONES ROBLE, S.L.

Indice

Introducción

3

campusproyectosnebrija.imf.com © EDICIONES ROBLE, S.L.
JOAO MANUEL DA SILVA FONTES COELHO

campusproyectosnebrija.imf.com © EDICIONES ROBLE, S.L.
JOAO MANUEL DA SILVA FONTES COELHO

campusproyectosnebrija.imf.com © EDICIONES ROBLE, S.L.
JOAO MANUEL DA SILVA FONTES COELHO

campusproyectosnebrija.imf.com © EDICIONES ROBLE, S.L.
JOAO MANUEL DA SILVA FONTES COELHO

Introducción



Introducción. Manuel Lucania

De forma acelerada, cada día se consolida más el sueño de una máquina capaz de hablar con un humano de forma natural. El procesamiento del lenguaje natural (PLN) es un área de la Inteligencia Artificial que pretende emular las capacidades humanas para comprender y generar el lenguaje humano.

Por otro lado, la ingente cantidad de datos que proporciona en la actualidad la web representa una excelente fuente de información para extraer conocimiento mediante la aplicación de técnicas de minería de texto (minería de datos aplicada a los textos).

La aplicación de estas técnicas, por tanto, abre un mundo aún por desarrollar de posibilidades prácticas y de negocio, que se irán explorando a lo largo de este material.

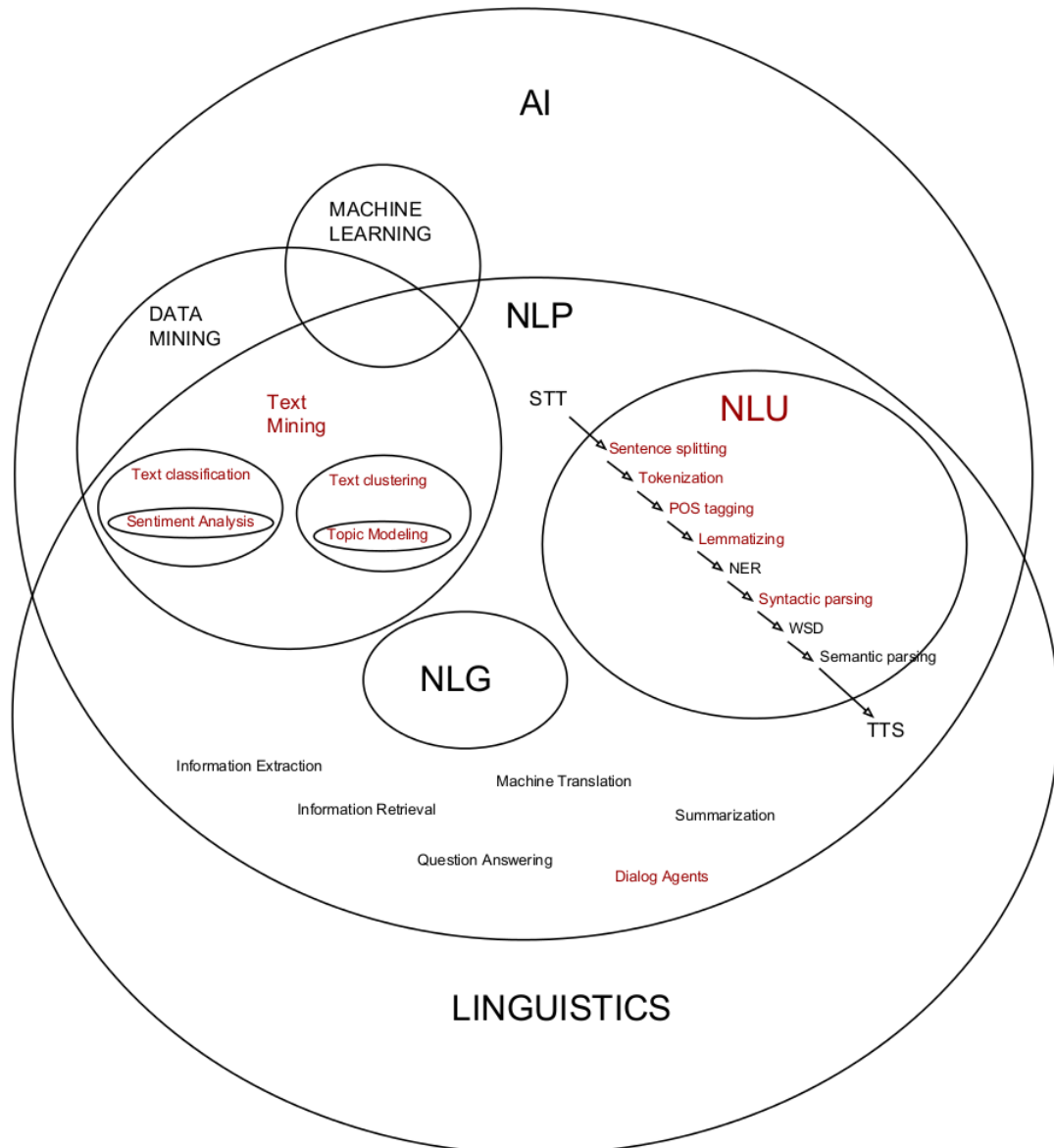


Figura 1.1. El PLM como rama de conocimiento. *Fuente:* elaboración propia.

En el diagrama anterior se puede observar cómo el NLP (Natural Language Processing) se conforma por la confluencia de dos ramas del conocimiento: la Inteligencia Artificial y la lingüística; es decir, $NLP = IA + \text{lingüística}$.

Por tanto, muchas de las técnicas de IA (minería de datos, aprendizaje automático) se aplican directamente al NLP para dotar a los sistemas de la inteligencia necesaria para resolver los problemas inherentes al manejo automático del lenguaje humano.

Dentro del NLP, se distinguen NLU (Natural Language Understanding) y NLG (Natural Language Generation). Es decir, $NLP = NLU + NLG$.

NLU

Parte del NLP que analiza textos para intentar extraer un entendimiento. Por tanto, es similar a nuestra capacidad de leer y entender textos.

NLG

Parte del NLP que, a partir de alguna representación del conocimiento, es capaz de expresar dicho conocimiento. Por tanto, es similar a nuestra capacidad de pensar una idea y de expresarla con palabras.



En este módulo se imparten los fundamentos mínimos para adquirir un conjunto de habilidades tecnológicas básicas, representadas en el diagrama de Venn anterior. Sin embargo, no se verán en detalle todas las tecnologías, ya que excedería los límites y la naturaleza de un curso introductorio, pero sí se verán las partes fundamentales —destacadas en rojo—. El resto de tecnologías se verán muy por encima en la última unidad didáctica. Se recomienda consultar el documento que contiene el glosario de este módulo para profundizar en los conceptos desconocidos, aunque se verán a lo largo del temario.

1

El módulo comienza con una **Introducción histórica y tecnológica al PLN**, el alumno deberá leer y entender este material teórico de introducción al PLN. Esta unidad será puramente teórica, por ello, es la única unidad que no contiene un caso o ejercicio práctico. No obstante, al finalizar su estudio, se deberá responder un cuestionario teórico de autoevaluación.

2

La segunda unidad, **Introducción a NLTK**, será la primera sesión práctica, donde se profundizará en la primera de las herramientas que se usará en la asignatura. Se mostrarán y practicarán las llamadas a funciones de la librería NLTK. Incluye dos casos prácticos solucionados.

Herramientas: NLTK, Jupyter y Anaconda.

3

La tercera unidad abarca dos herramientas nuevas: **Brat y Gate**.

- ➔ Brat es la segunda herramienta NLP de la asignatura con la que se aprenderá a anotar textos. Se practicará con un sencillo caso práctico de anotación.
- ➔ Gate es la tercera herramienta del módulo. Se verán todas las posibilidades que ofrece el Framework Gate, tanto para la programación como para la visualización de cadenas NLP. Se explicarán las gramáticas —fundamentalmente las basadas en autómatas de estados finitos— y se demostrará su utilidad mediante prácticas con gramáticas Jape de Gate. El alumno deberá realizar el caso práctico correspondiente a modo de práctica.

Herramientas: Para Brat, no es necesario realizar ninguna instalación. Se hará conexión mediante un browser a un servidor público de Brat para practicar anotando textos. Para Gate, se instalará la librería Gate.

4

En la cuarta unidad, **Text Mining I: clustering**, se detallarán las principales técnicas y se usarán las librerías de Anaconda para solucionar un caso práctico propuesto.

Herramientas: NLTK, Jupyter y Anaconda.

5

Y, en la quinta unidad, **Text Mining II: sentimiento y temas**, se profundizará en la segunda gran aplicación del Text Mining, que es la clasificación de textos, usada, por ejemplo, para localizar spam entre el e-mail entrante o para analizar el sentimiento expresado por un texto (Sentiment Analysis). Se estudiará también cómo se pueden descubrir los temas de los que habla un texto (Topic Classification y Topic Modeling).

Herramientas: NLTK, Jupyter y Anaconda.

6

Finalmente, la sexta unidad trata de ejemplos de **aplicaciones y técnicas de PLN** que resuelven problemas de nuestro mundo real: Extracción de información, asistentes conversacionales, buscadores... Algunas de estas aplicaciones se verán únicamente de forma teórica y otras —por ejemplo, los asistentes— se ejercitarán con una práctica correspondiente.

En todas las unidades —excepto en la primera—, se aportan casos prácticos con solución a modo de práctica, lo que preparará a los alumnos para la realización del caso práctico final del módulo evaluable.



Los **objetivos generales** que los alumnos alcanzarán tras el estudio de este módulo pueden resumirse en los siguientes:

1. Comprender los fundamentos teóricos, las técnicas y las aplicaciones prácticas del PLN.
2. Manejar de forma práctica distintas librerías y herramientas de PLN (NLTK, Brad, Gate).
3. Adquirir la capacidad de implementar, o al menos conocer de forma básica, distintas soluciones de PLN (análisis de sentimiento, Text Mining de datos web, asistentes virtuales...).