# MIA – Unidad 3 Resumen

- **Procesamiento del Lenguaje Natural (PLN):** El PLN es el estudio de las relaciones entre el lenguaje humano y las máquinas. Implica el uso de la lingüística y requiere un *corpus* específico para cada idioma en el aprendizaje automático. El análisis del lenguaje se divide pedagógicamente en fases sintáctica, semántica y pragmática, requiriendo fases de análisis previo.
- **Corpus lingüísticos:** Los corpus son conjuntos de palabras usados para formar diccionarios y son la base del aprendizaje automático para el procesamiento del lenguaje. Se requiere un *corpus* específico para cada idioma. Algunos corpus lingüísticos importantes son:
    - British National Corpus (BNC) para inglés.
    - Corpus of Contemporary German para alemán.
    - Corpus PAISÀ para italiano.
    - NINJAL-LWP para japonés.
- **Fases del PLN:** El análisis del lenguaje se puede descomponer en distintas fases:
    - **Preprocesamiento:** Involucra la identificación del idioma, la eliminación de elementos no relevantes, la limpieza y normalización, la corrección de errores, la eliminación de palabras vacías (*stopwords*), la lematización, el *stemming* y la segmentación (*tokenización*). La segmentación divide el texto en unidades básicas llamadas *tokens*.
    - **Análisis léxico:** Implica determinar la función que desempeña un término en una frase.
    - **Análisis sintáctico:** Analiza las reglas y principios que gobiernan la combinación de los constituyentes sintácticos.
    - **Análisis semántico:** Permite entender el significado de los términos individuales y su función en la oración, utilizando recursos como diccionarios, tesauros y ontologías.
    - **Análisis pragmático:** Determina la intencionalidad de las frases, requiriendo técnicas avanzadas como la Computación Afectiva.
- **Potencial del PLN:** El PLN tiene diversas aplicaciones, incluyendo el reconocimiento del habla (ASR), la síntesis de texto a voz (TTS), la detección de entidades nombradas (NER), la traducción automática, la similitud de textos y el análisis del sentimiento.
    - **Reconocimiento del habla (ASR):** Convierte los fonemas emitidos por un humano en grafías escritas.
    - **Síntesis de texto a voz (TTS):** Convierte texto en voz.
    - **Detección de entidades nombradas (NER):** Detecta palabras clave y las clasifica en categorías como personas, organizaciones, lugares y cantidades.
    - **Traducción automática:** Traduce texto de un idioma a otro.
    - **Similitud de textos:** Etiqueta palabras aludiendo a conceptos semánticos.
    - **Análisis del sentimiento:** Capta la positividad o negatividad en un texto.
- **Limitaciones del PLN:** La ambigüedad lingüística es un desafío importante. Los tipos de ambigüedad incluyen la sintáctica, pragmática, fonológica y funcional.
- **Modelos y técnicas existentes:**
    - **BERT:** Analiza palabras ubicadas a la izquierda y a la derecha de cada término, considerando el contexto para la desambiguación.
    - **RoBERTa:** Difiere en la manera de llevar a cabo el enmascaramiento en el proceso de atención y tiene un entrenamiento más lento.

    - **XLNET:** Incrementa el tiempo de entrenamiento.
    - **MEGATRON:** Modelo con 8.3 millardos de parámetros.
    - **GPT:** No es bidireccional y su afinado es costoso. GPT-3.5 mejora GPT-3 con una arquitectura más refinada. GPT-4 maneja indicaciones complejas, se adapta a tonos y puede generar código e interpretar 26 idiomas.
- **Herramientas y librerías:**
    - **NLTK:** Librería de Python para tareas de PLN como tokenización, *stemming*, lematización y análisis morfosintáctico.
    - **SpaCy:** Biblioteca de Python para PLN, rápida y eficiente, con modelos preentrenados para varios idiomas. Permite construir sistemas avanzados de comprensión del lenguaje natural.
    - **Nemo**: Se emplea de manera global para NLP, TTS y ASR.
    - **Stanford NLP (Stanza):** Biblioteca de procesamiento del lenguaje natural que ofrece un conjunto completo de herramientas para analizar texto en múltiples idiomas.
    - **UDPipe:** Herramienta de procesamiento de lenguaje natural basada en el proyecto Universal Dependencies, capaz de analizar morfosintácticamente textos en una variedad de idiomas.
    - **NLP-Cube:** Herramienta de procesamiento del lenguaje natural diseñada para analizar de manera eficiente grandes volúmenes de texto en múltiples idiomas.
- **spaCy:**
    - El objeto nlp contiene el *pipeline* de procesamiento.
    - Los objetos Doc permiten acceder a la información sobre el texto de forma estructurada.
    - El *matcher* permite escribir reglas para encontrar palabras y frases en el texto.
    - El vocabulario (vocab) guarda todos los datos y codifica los *strings* a *hash IDs*.
    - El *pipeline* incluye componentes como el *tagger*, *parser* y *entity recognizer*.
    - Se pueden añadir componentes personalizados al *pipeline* para modificar el Doc.
    - Se pueden añadir atributos personalizados a los objetos Doc, Token y Span.
    - Es posible entrenar y actualizar los modelos estadísticos, como el *entity recognizer*.
- **Análisis de sentimientos con LLM (Cohere):** Se utiliza un LLM como Cohere para clasificar reseñas como positivas, negativas o neutras.

# MIA – Unidad 3 Test

**Opción Múltiple**

1. ¿Cuál es el objetivo principal del Procesamiento del Lenguaje Natural (PLN)?
    a) Estudiar la comunicación entre animales.
    b) Estudiar las relaciones del lenguaje entre humanos y máquinas.
    c) Crear diccionarios de idiomas.
    d) Analizar la estructura de las oraciones solamente.
2. Según Noam Chomsky, ¿qué plantea su teoría lingüística?
    a) El lenguaje es un producto del aprendizaje cultural.
    b) El lenguaje es producto de la interacción social.
    c) El lenguaje es producto de una facultad innata de la mente humana.
    d) El lenguaje se adquiere solo a través de la educación formal.
3. ¿Qué es un *corpus* en el contexto del PLN?
    a) Un diccionario de sinónimos y antónimos.
    b) Un conjunto de palabras de una lengua empleado para el aprendizaje automático.
    c) Un programa para traducir idiomas.
    d) Un método para corregir errores gramaticales.
4. ¿Qué estudia la morfología según D. Jurafsky?
    a) El significado de las palabras.
    b) La pronunciación de las palabras.
    c) Las reglas que rigen la flexión, composición y derivación de las palabras.
    d) El origen etimológico de las palabras.
5. ¿Cuál de las siguientes es una aplicación del PLN?
    a) Diseño de videojuegos.
    b) Análisis del sentimiento y de la opinión.
    c) Control de robots industriales.
    d) Predicción del clima.
6. ¿Qué significa ASR en el contexto del PLN?
    a) Análisis Semántico Regular.
    b) Automatic Speech Recognition (Reconocimiento Automático del Habla).
    c) Advanced Sentence Retrieval.
    d) Algoritmo de Similitud de Textos.
7. ¿Qué es NER (Named Entity Recognition)?
    a) Un sistema de traducción automática.
    b) Detección de entidades nombradas en un texto, como personas, organizaciones y lugares.
    c) Un método para corregir errores ortográficos.
    d) Un algoritmo para analizar el sentimiento en redes sociales.

8. ¿Qué tipo de arquitectura es común en modelos de traducción automática?
    a) Redes Neuronales Convolucionales.
    b) Máquinas de Vectores de Soporte.
    c) Arquitectura secuencia a secuencia de Transformer.
    d) Modelos de Markov.
9. ¿Cuál es una característica clave del modelo BERT?
    a) Es unidireccional.
    b) No tiene en cuenta el contexto.
    c) Es bidireccional.
    d) Requiere supervisión constante.
10. ¿Cuál de los siguientes modelos de lenguaje es de código cerrado (no Open Source)?
    a) BERT.
    b) RoBERTa.
    c) GPT-3.
    d) XLNET.
11. ¿Qué tipo de ambigüedad se presenta en la frase "María vio a Juan con un telescopio"?
    a) Ambigüedad léxica.
    b) Ambigüedad sintáctica.
    c) Ambigüedad semántica.
    d) Ambigüedad fonológica.
12. ¿Cuál es el primer paso recomendado para una persona experta en IA que desea adquirir conocimiento en PLN?
    a) Empezar directamente con nVidia NeMo.
    b) Formación teórica y clásica.
    c) Usar Spacy en lenguas inglesa y española.
    d) Programar un sistema de ideas clave o resumen.
13. ¿Qué tarea NO se puede realizar con NLTK-BOOK?
    a) Cargar un corpus.
    b) Tokenizar y etiquetar.
    c) Emplear un ASR (Automatic Speech Recognition).
    d) Programar un analizador de sentimientos.
14. ¿Qué permite la lematización en el preprocesamiento de textos?
    a) Eliminar errores ortográficos.
    b) Reducir el vocabulario sustituyendo palabras por su lema.
    c) Identificar el idioma del texto.
    d) Convertir mayúsculas a minúsculas.
15. ¿Qué es el *stemming*?
    a) Algoritmo que encuentra la raíz de cada término.
    b) Eliminar palabras vacías de un texto.
    c) Identificar el idioma del texto.
    d) Convertir mayúsculas a minúsculas.
16. ¿Cuál es el objetivo de la tokenización?
    a) Corregir errores gramaticales.
    b) Segmentar el texto en unidades mínimas (tokens).
    c) Traducir el texto a otro idioma.
    d) Resumir el contenido del texto.

17.¿Qué función realiza el etiquetado POS (Part of Speech tagging)?
    a) Traducir el texto a otro idioma.
    b) Identificar la función de cada palabra en la frase (nombre, verbo, adjetivo, etc.).
    c) Corregir errores ortográficos.
    d) Resumir el contenido del texto.
18.¿Qué es un *parser* en el análisis sintáctico?
    a) Un diccionario de sinónimos y antónimos.
    b) Una herramienta para analizar cada una de las sentencias que conforman el texto.
    c) Un programa para traducir idiomas.
    d) Un método para corregir errores gramaticales.
19.¿Qué son los word vectors?
    a) Representaciones multidimensionales de los significados de las palabras.
    b) Un diccionario de sinónimos y antónimos.
    c) Un programa para traducir idiomas.
    d) Un método para corregir errores gramaticales.
20.Según el curso de SpaCy, ¿qué contiene el objeto nlp?
    a) El pipeline de procesamiento.
    b) Los textos originales sin procesar.
    c) Las visualizaciones gráficas de los datos.
    d) Las estadísticas de uso del lenguaje.
21.¿Qué es el objeto Doc en spaCy?
    a) Un modelo estadístico para predicciones.
    b) Una estructura que permite acceder a la información sobre el texto en una forma estructurada.
    c) Un diccionario de sinónimos y antónimos.
    d) Una herramienta para corregir errores gramaticales.
22.¿Qué permiten los modelos estadísticos en spaCy?
    a) Solo analizar la estructura gramatical de las oraciones.
    b) Hacer predicciones dentro del contexto, como part-of-speech tags y entidades nombradas.
    c) Traducir texto a diferentes idiomas.
    d) Resumir el contenido de un documento.
23.¿Qué hace el *matcher* de spaCy?
    a) Corrige errores ortográficos en el texto.
    b) Permite escribir reglas para encontrar palabras y frases en el texto.
    c) Traduce el texto a otro idioma.
    d) Resumir el contenido del texto.
24.¿Qué guarda spaCy en el vocabulario (vocab)?
    a) Solo las palabras más comunes del idioma.
    b) Todos los datos, incluyendo palabras y esquemas de labels para tags y entidades.
    c) Solo los sinónimos y antónimos de las palabras.
    d) Las reglas gramaticales del idioma.

25.¿Qué es un Lexema en spaCy?
   a) Entradas en el vocabulario independientes del contexto.
   b) Solo las palabras más comunes del idioma.
   c) Solo los sinónimos y antónimos de las palabras.
   d) Las reglas gramaticales del idioma.

26.¿Qué es un Span en spaCy?
   a) Un slice de un Doc que está formado por uno o más tokens.
   b) Representaciones multidimensionales de los significados de las palabras.
   c) Un programa para traducir idiomas.
   d) Un método para corregir errores gramaticales.

27.¿Qué es el método .similarity en spaCy?
   a) Eliminar palabras vacías de un texto.
   b) Identificar el idioma del texto.
   c) Convertir mayúsculas a minúsculas.
   d) Comparar dos objetos y predecir qué tan similares son.

28.¿Qué debe tener un modelo spaCy para usar el método .similarity?
   a) Diccionario de sinónimos y antónimos.
   b) Un programa para traducir idiomas.
   c) Método para corregir errores gramaticales.
   d) Word vectors.

29.¿Qué hace el método nlp.pipe en spaCy?
   a) Eliminar palabras vacías de un texto.
   b) Identificar el idioma del texto.
   c) Convertir mayúsculas a minúsculas.
   d) Procesar los textos como un stream y usa yield para devolver objetos Doc.

30.¿Cuál de las siguientes opciones muestra un tipo de ambigüedad funcional?
   a) El hombre era brillante.
   b) Pedro y yo escribimos un cuento.
   c) Tomó su paraguas y salió a la calle.
   d) En Inglaterra hay menos contagios porque se hacen muchos más tés.

**Verdadero/Falso**

1. **Verdadero o Falso:** El Test de Georgetown (1954) consistió en la traducción automática de unas 60 oraciones del inglés al ruso.

2. **Verdadero o Falso:** La semántica estudia las reglas que rigen la flexión, la composición y la derivación de las palabras.

3. **Verdadero o Falso:** El análisis del sentimiento puede ser empleado en chatbots para averiguar el estado de ánimo del usuario.

4. **Verdadero o Falso:** Los modelos de lenguaje GPT son bidireccionales, analizando el contexto tanto a la izquierda como a la derecha de cada palabra.

5. **Verdadero o Falso:** La ambigüedad lingüística es un problema menor que no afecta la capacidad de un sistema automático para responder correctamente.

6. **Verdadero o Falso:** El *stemming* es un proceso que permite reducir el vocabulario con el que se trabaja.

7. **Verdadero o Falso:** En el preprocesamiento de textos, es irrelevante conocer el idioma en el que está escrito un texto.

8. **Verdadero o Falso:** La lematización es el proceso de convertir un texto a voz.

9. **Verdadero o Falso:** Según Noam Chomsky existen distintos tipos de gramáticas para describir los lenguajes formales.

10. **Verdadero o Falso:** El análisis semántico se centra en la estructura gramatical de las oraciones, sin considerar el significado de las palabras.

11. **Verdadero o Falso:** Un *parser* es una herramienta para analizar sintácticamente un lenguaje.

12. **Verdadero o Falso:** spaCy es principalmente útil para la traducción automática de textos.

13. **Verdadero o Falso:** Los modelos estadísticos de spaCy permiten hacer predicciones dentro del contexto.

14. **Verdadero o Falso:** En spaCy, los atributos que devuelven un string terminan con un guión bajo (_).

15. **Verdadero o Falso:** El Doc es una de las estructuras de datos centrales de spaCy y es creado manualmente.

16. **Verdadero o Falso:** El string store está disponible como nlp.vocab.strings.

17. **Verdadero o Falso:** Los hash IDs se pueden revertir.

18. **Verdadero o Falso:** Los Lexemas tienen part-of-speech tags.

19. **Verdadero o Falso:** Si quieres añadir extensiones de atributos en un span, casi siempre debes usar una extensión de propiedades con un getter.

20. **Verdadero o Falso:** El método nlp.pipe es más lento que solo llamar al objeto nlp sobre cada texto.

| | | | | |
|---|---|---|---|---|
| 1. b | 11. b | 21. b | 31. F | 41. V |
| 2. c | 12. b | 22. b | 32. F | 42. F |
| 3. b | 13. c | 23. b | 33. V | 43. V |
| 4. c | 14. b | 24. b | 34. F | 44. V |
| 5. b | 15. a | 25. a | 35. F | 45. F |
| 6. b | 16. b | 26. a | 36. F | 46. V |
| 7. b | 17. b | 27. d | 37. F | 47. F |
| 8. c | 18. b | 28. d | 38. F | 48. F |
| 9. c | 19. a | 29. d | 39. V | 49. V |
| 10. c | 20. a | 30. b | 40. F | 50. F |