

Procesamiento de Lenguaje Natural

Semestre Enero-Mayo 2026

Mauricio Toledo-Acosta
`mauricio.toledo@unison.mx`

January 12, 2026

1 Temario

1. Módulo 1: Introducción

- Introducción y Estado del Arte
- Tareas Principales del Curso y métricas de evaluación
- Corpus de Referencia y setup del Curso
- **Actividad:** Exploración de corpus y acceso a Yuca.

2. Módulo 2: Preprocesamiento y Obtención de Datos

- **Actividad:** Web scraping de noticias de tópicos separados
- Limpieza, Normalización y Tokenización
- Expresiones Regulares
- **Actividad:** Extracción de patrones (emails, teléfonos, fechas, URLs, RFC, datos financieros)
- Análisis Lingüístico con spaCy
- **Actividad:** Extracción de acuerdos de actas.
- Almacenamiento y Visualización
- **Actividad:** Análisis exploratorio de corpus

3. Módulo 3: Modelos de Lenguaje Tradicionales

- Fundamentos de Modelos Probabilísticos
- Topic Modeling con Técnicas Clásicas: LSA y LDA
- **Actividad:** Topic Modeling
- Leyes de Conteo
- **Actividad:** Detección de textos anómalos
- n -gramas de palabras y caracteres, suavizado y aplicaciones

- Generación de Texto y perplejidad
- **Actividad:** Detector de idiomas
- Primeras Representaciones Vectoriales: Bag of Words y TF-IDF
- **Actividad:** Baseline en Corpus de Referencia

4. Módulo 4: Representaciones Vectoriales Densas

- Introducción
- Word2Vec, FastText y Doc2Vec
- **Actividad:** Implementar Skip-gram simplificado
- **Actividad:** Detección de Evasión Léxica y Exploración de embeddings
- Aplicaciones a Corpus de Referencia

5. Módulo 5: Transformers

- Introducción
- Arquitectura Transformer y Atención
- BERT y Hugging Face Transformers
- Fine-tuning BERT
- Embeddings Contextuales
- BERT para Otras Tareas: NER, Question Answering y BERTopic
- **Actividad:** BERT en los corpus de referencia

6. Módulo 6: Large Language Models

- Introducción y conceptos importantes
- **Actividad:** Comparar modelos actuales
- Prompt Engineering
- **Actividad:** Text Mining con LLMs
- Fine-tuning de LLMs
- **Actividad:** Fine-tuning
- RAG
- **Actividad:** Implementación de RAG
- Evaluación de LLMs
- Optimización y Deployment
- **Actividad:** Ollama
- Agentes
- **Actividad:** Agente ReAct
- Consideraciones Éticas y Sociales

7. **Actividad:** Comparación Final - Todas las Técnicas en los corpus de referencia

2 Información Adicional

- Repositorio del curso: <https://github.com/gmauricio-toledo/NLP-LCC>
- La evaluación será por medio de presentaciones, trabajos en clase y actividades a entregar. La calificación final será el promedio de cumplimiento de estas actividades.

