Creando un Sistema Básico de Procesamiento de Lenguaje Natural (PLN).

Objetivo.

El objetivo de este proyecto es que los alumnos y las alumnas apliquen sus conocimientos sobre Procesamiento de Lenguaje Natural (PLN) para diseñar y desarrollar un sistema básico de análisis de texto. Se busca que los estudiantes comprendan las herramientas y metodologías utilizadas por los investigadores en PLN y que implementen una solución funcional para una tarea específica.

Este proyecto permitirá a los alumnos y las alumnas a desarrollar habilidades prácticas en el campo del Procesamiento de Lenguaje Natural, fomentando el trabajo en equipo, la investigación y la innovación tecnológica.

Descripción de la Actividad.

1. Investigación Inicial

- El alumnado deberá investigar sobre las herramientas mencionadas.
- Se proporcionará un conjunto de textos para analizar y procesar como base, pudiendo utilizar corpus.

2. Selección de la Tarea a Resolver

Cada grupo elegirá una de las siguientes tareas de PLN:

- **Análisis de Sentimiento:**
  - Objetivo: Clasificar textos como positivos, negativos o neutros.
  - Utilizar librerías como **NLTK o SpaCy** para tokenizar y analizar palabras clave.
- **Reconocimiento de Entidades Nombradas (NER):**
  - Objetivo: Identificar nombres de empresas, leyes, números de pedido, lugares, etc.
  - Implementar modelos preentrenados de **SpaCy o Hugging Face Transformers** para extraer entidades.
- **Clasificación de Textos:**
  - Objetivo: Diferenciar entre tipos de documentos (noticias, reseñas, documentos legales, conversaciones de chatbots, etc.).
  - Utilizar **Hugging Face Transformers** con modelos como BERT para entrenar un clasificador.
- **Chatbots:**
  - Objetivo: Diseñar un bot básico que responda preguntas automáticamente.
  - Usar herramientas como **NLTK o SpaCy** para procesar las preguntas y generar respuestas predefinidas.

3. Implementación

- Programar un sistema en **Python** utilizando las librerías seleccionadas.
- Utilizar **Jupyter Notebook** o **Google Colab** para documentar y probar el código.

4. Evaluación y Mejora

- Aplicar métricas básicas de evaluación (precisión, recall, F1-score).
- Ajustar hiperparámetros o cambiar modelos según los resultados obtenidos.

5. Presentación de Resultados

- Cada grupo expondrá su sistema en clase, explicando **qué modelo usaron, cómo funciona y qué mejoras podrían implementarse**.

Evaluación

Cada criterio será calificado de forma que la suma total máxima sea **10 puntos**, considerando el trabajo en equipo y la contribución individual:

- **Claridad y organización del código** (2 puntos): Se evaluará la correcta estructura del código, su legibilidad, uso adecuado de comentarios y modularidad.
- **Precisión y funcionamiento del modelo** (3 puntos): Se medirán los resultados obtenidos utilizando métricas como precisión, recall y F1-score, asegurando que el modelo cumpla con los objetivos de la tarea seleccionada.
- **Presentación y explicación de los resultados** (2 puntos): Se calificará la claridad en la exposición, el análisis de los resultados y la justificación de las decisiones tomadas en el desarrollo del proyecto.
- **Creatividad en la implementación y mejoras propuestas** (2 puntos): Se valorará la originalidad de la solución, las optimizaciones realizadas y las mejoras adicionales implementadas en el modelo.
- **Trabajo en equipo y colaboración** (1 punto): Se tendrá en cuenta la distribución equitativa de tareas, la cooperación entre los integrantes del grupo y la comunicación efectiva durante el desarrollo del proyecto.

Anexos.

Conjunto de Textos para Analizar

Categoría 1: Noticias (Análisis de Sentimiento / Clasificación)

1. "El mercado bursátil experimentó una fuerte caída hoy debido a la incertidumbre económica global. Los analistas prevén una recuperación en los próximos meses, pero los inversionistas están preocupados."
2. "La selección nacional de fútbol ganó el campeonato con una actuación espectacular. Los aficionados celebraron en las calles hasta altas horas de la madrugada."
3. "Un nuevo estudio revela que el cambio climático está afectando la biodiversidad de los océanos, con consecuencias impredecibles para el ecosistema marino."

Categoría 2: Reseñas de productos (Análisis de Sentimiento / Opinión de Usuarios)

4. "Compré este teléfono hace un mes y ha sido una gran decepción. La batería dura muy poco y la cámara tiene una calidad pésima. No lo recomendaría."
5. "Este libro es increíble. La historia es atrapante y los personajes están muy bien desarrollados. ¡Uno de los mejores que he leído este año!"
6. "El restaurante tenía un ambiente agradable, pero el servicio fue muy lento. La comida estaba bien, aunque esperaba algo mejor por el precio."

Categoría 3: Extractos de documentos legales (Reconocimiento de Entidades - NER)

7. "El contrato entre Empresa XYZ y el proveedor ABC establece que la entrega de los bienes deberá realizarse en un plazo no mayor a 30 días desde la firma del acuerdo."
8. "Según la Ley 25/2018 de Protección de Datos, todas las empresas deben garantizar la privacidad de la información de sus clientes y empleados."

Categoría 4: Diálogos de atención al cliente (Chatbots / NLP Conversacional)

9. **Cliente:** "Hola, quisiera saber el estado de mi pedido número 12345." **Bot:** "Déjame verificar... Tu pedido ha sido enviado y llegará en 3 días."
10. **Cliente:** "Necesito cancelar mi suscripción al servicio." **Bot:** "Para proceder con la cancelación, dime tu número de cuenta o correo electrónico asociado."

Herramientas a Utilizar (todas gratuitas y de acceso libre).

1. **NLTK** (Natural Language Toolkit) - https://www.nltk.org/
    - Librería en Python especializada en procesamiento de texto.
    - Permite realizar tokenización, stemming, lematización y análisis sintáctico.
    - Ideal para procesamiento de lenguaje natural en proyectos pequeños.
2. **SpaCy** - https://spacy.io/
    - Alternativa más rápida y eficiente que NLTK.
    - Incluye modelos preentrenados para diferentes idiomas y tareas como reconocimiento de entidades (NER).
    - Su uso es recomendado para grandes volúmenes de texto.
3. **Hugging Face Transformers** - https://huggingface.co/
    - Plataforma con modelos avanzados como BERT y GPT-2.
    - Facilita la implementación de modelos preentrenados en tareas como traducción automática, clasificación de texto y generación de lenguaje.
4. **NVIDIA NeMo** - https://developer.nvidia.com/nemo
    - Especializada en modelos de PLN basados en redes neuronales profundas.
    - Permite entrenar modelos avanzados de reconocimiento de voz y síntesis de texto.
5. **Google Colab** (para programación en la nube sin instalaciones) - https://colab.research.google.com/
    - Entorno de desarrollo basado en Jupyter Notebook que permite ejecutar código en la nube.
    - Ofrece acceso gratuito a GPUs y facilita la ejecución de modelos de IA sin necesidad de instalación local.

Bibliografía Recomendada

- Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing* (3rd ed.). Stanford University.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
- Vaswani, A., et al. (2017). *Attention Is All You Need*. arXiv.