UTAMED

Ruta de Aprendizaje Autónomo

Procesamiento del Lenguaje Natural

RUTA DE APRENDIZAJE:

Contexto

Los estudiantes participan en el proyecto ficticio “Smart Urban System”, cuyo objetivo es diseñar e implementar soluciones inteligentes para un sistema urbano conectado. En PLN, desarrollarán un módulo de interacción ciudadana mediante lenguaje natural, integrando sus resultados con los entregables de otras asignaturas.

En el marco del proyecto Smart Urban System, nuestra asignatura de Procesamiento de Lenguaje Natural dotará al sistema urbano conectado de capacidades avanzadas para normalizar, desambiguar y enriquecer semánticamente un corpus heterogéneo —que incluye foros vecinales, quejas ciudadanas y tickets de incidencia—, desarrollar modelos de clasificación temática y análisis de sentimiento capaces de categorizar y priorizar las demandas de la ciudadanía, generar representaciones vectoriales mediante técnicas como Word2Vec u otros adaptadas a los matices lingüísticos urbanos, e integrar un modelo de lenguaje conversacional (LLM) que ofrezca respuestas automáticas, aclare consultas ciudadanas y derive avisos críticos al sistema de control de la plataforma, asegurando así una comunicación bidireccional inteligente y mejorando la eficiencia en la gestión de incidencias y la calidad del servicio público.

Papel del Procesamiento Lenguaje Natural en Smart Urban System: dentro del proyecto, la asignatura de PLN desempeña un rol esencial al dotar al sistema urbano conectado de capacidades avanzadas de comprensión y generación de lenguaje natural para la interacción ciudadana y la toma de decisiones automatizadas.

Objetivos específicos

- Desarrollar un flujo de preprocesamiento avanzado capaz de limpiar, normalizar y enriquecer semánticamente un corpus heterogéneo (foros vecinales, quejas ciudadanas, tickets de incidencia).
- Diseñar, entrenar y validar modelos de clasificación temática y análisis de sentimiento que prioricen demandas ciudadanas y detecten tendencias emergentes en tiempo real.
- Generar y comparar representaciones vectoriales (embeddings) mediante técnicas como Word2Vec u otros, optimizadas para reflejar matices lingüísticos del entorno urbano.
- Integrar un modelo de lenguaje conversacional (LLM) que automatice respuestas, clarifique consultas ciudadanas y derive alertas críticas al sistema central, garantizando la coherencia y eficacia de la comunicación bidireccional.

Entrega

Eres un especialista en Procesamiento del Lenguaje Natural integrado al equipo de Smart Urban System, cuyo objetivo es facilitar la interacción ciudadana y la toma de decisiones automatizadas mediante un módulo de PLN. Se te pide entregar un informe técnico que incluya:

- La definición de las tuberías (pipelines) de preprocesamiento: limpieza, normalización, tokenización y enriquecimiento semántico del corpus ciudadano.
- El diseño y justificación de los modelos de clasificación temática y de análisis de sentimiento, con sus esquemas de entrenamiento y validación.
- La descripción de la generación de embeddings, así como la comparación de sus prestaciones si haces varios intentos y cobertura semántica.
- La arquitectura e implementación del componente conversacional (LLM), con el detalle de los prompts, flujos de diálogo y la lógica de derivación de alertas críticas.
- Las métricas e indicadores de evaluación utilizados para identificar configuraciones óptimas (precisión, recall, F1, coherencia de respuesta, tiempos de latencia).
- Las pruebas de concepto y simulaciones (demo interactivo, Jupyter notebooks) que validen la robustez, eficiencia y escalabilidad del módulo.

Para ello, es necesario que sigas estos pasos:

Tareas

Semana 1: Captura y preprocesamiento del corpus ciudadano

- **Actividad 1:** Identificación y extracción de datos
    - Seleccionar fuentes representativas (foros vecinales, quejas ciudadanas, tickets de incidencia, ...).
    - Documentar procedencia, volumen y formato de cada fuente.

- **Actividad 2:** Limpieza y normalización textual
    - Eliminar ruido (HTML, caracteres especiales, duplicados).
    - Tokenizar, lematizar y aplicar técnicas de enriquecimiento semántico (reconocimiento de entidades, corrección ortográfica).

- **Actividad 3:** Diseño y entrenamiento de modelos
    - Definir etiquetas temáticas y escalas de sentimiento.
    - Entrenar modelos supervisados sencillos si es necesario.

Semana 2: Construcción e integración conversacional

- **Actividad 4:** Generación de vectores
    - Entrenar Word2Vec (u otros) desde cero sobre el corpus.
- **Actividad 5:** Análisis comparativo
    - Comparar calidad semántica mediante analogías y clustering.
    - Visualizar distribuciones en 2D si fuese necesario.
- **Actividad 6:** Diseño de flujo de diálogo
    - Definir prompts y scripts de interacción automatizada.
    - Conectar con API o entorno local de LLM.
- **Actividad 7:** Pruebas y ajuste
    - Simular consultas ciudadanas, medir coherencia y latencia.
    - Ajustar prompts y parámetros según métricas de usabilidad.
- **Entrega final:** Documento técnico definitivo que incluya la introducción contextual, desarrollo matemático completo, simulaciones computacionales, interpretación práctica de resultados y recomendaciones concretas. Demo funcional de la implementación de código de esta asignatura PLN (en Jupiter Notebook o similares).

Directrices de la Entrega Final

El informe técnico deberá cumplir las siguientes normas de presentación: tipografía Calibrí 12 pt; interlineado de 1,5; márgenes de 2,5 cm en todas las caras; citación en formato APA 7ª edición. Se entregará un único documento en PDF (máximo 8 páginas, incluyendo gráficos, tablas y anexos de código), estructurado en:

1. Introducción y objetivos

2. Datos y preprocesamiento
3. Modelado (clasificación temática, análisis de sentimiento y embeddings)
4. Integración conversacional (LLM y demo)
5. Validación, conclusiones y rúbrica